Opening the Classroom Door: Can AI Improve Understanding of Pre-K Experiences?

Leigh Ann DeLyser, Sarah Gerard,
Gullnar Syed, and Jennifer Nakamura

SRI Education

SRI

December 2025

Authors

Leigh Ann DeLyser
Sarah Gerard
Gullnar Syed
Jennifer Nakamura

Suggested Citation

DeLyser, L. A., Gerard, S., Syed, G., & Nakamura, J. (2025). *Opening the classroom door: Can AI improve understanding of pre-K experiences* [Final report]. SRI.

Acknowledgments

The Gates Foundation supported this study (INV-078962) to advance the development of accurate, equitable, useful, and innovative tools to support coaching focused on the quality of center-based pre-kindergarten classrooms that serve children ages 3 to 5 years in the United States. Thank you to our program officer Snow Li for your thoughtful guidance.

The SRI study team would like to thank the teachers and instructional coaches for sharing their perspectives, and educational technology developers and education researchers for sharing their experiences and challenges. We would also like to thank Dr. Stephanie Curenton of Boston University Wheelock College of Education & Human Development for the use of select videos of prekindergarten classroom observations from their video library to evaluate the AI approaches.

We would also like to thank Aravind Sundaresan, John Niekrasz, and Twumasi Mensah-Boateng for their leadership of AI model development and testing. Thanks to study team members Nancy Perez, Claire Christensen, Mallory Scott, and Maricela Morales, who contributed to interviews, video data coding, analysis, and synthesis. Additional thanks to Todd Grindal, Mallory Rousseau, Charles Harding, and Jennifer Medeiros.

SRI Project Number: 101883

Contents

Introduction

A high-quality early childhood classroom is alive with activity. Children might be rotating from whole-group morning meeting to free play centers and then to math small-group instruction, all before lunch. Pre-kindergarten (pre-K) teachers looking to make the most of these instructional groupings—as well as those who support them, such as instructional coaches—can benefit from understanding how classroom time is spent to best focus instructional goals. SRI researchers set out to explore how artificial intelligence (AI) and classroom video can provide insights into pre K learning environments and support high-quality learning experiences.

Our work focused on developing insights into the potential use cases, barriers to adoption, and feasibility of AI-enabled tools for early childhood settings. The use of AI in any educational setting requires that data be a key part of determining its value. In this project, we used classroom video as the primary data source for our exploration of AI-enabled tools. In Phase 1 of the project, we interviewed educators, program leaders, educational technology (edtech) companies, curriculum developers, nonprofit leaders, and venture capitalists to best understand the use cases and barriers. After conducting those interviews, we identified high-priority opportunities to test the feasibility of using classroom video to produce fundamental models and codebooks that could be applied in early childhood settings in a multitude of ways. In consultation with the Gates Foundation, we tested:

- The use of AI to determine instructional groupings within the pre-K classroom
- Automatic identification of video segments that contain math or literacy instruction

In this report, we share our findings, highlighting key takeaways and decisions, performance of the AI models, and recommendations for future work by the field.

Use Case Interview Themes

To better understand the desires, needs, and concerns faced by participants in the early childhood education ecosystem, SRI engaged pre-K teachers, instructional coaches, program leaders, edtech and curriculum developers, and venture capitalists to share their perspectives of what tools and resources are needed, how video may contain opportunities, what applications currently exist, and what concerns they have about the use of AI in early childhood settings. These stakeholders' efforts vary in maturity but reflect the growing momentum and interest in AI-enabled tools across the field.

This section summarizes the findings from interviews and convenings with individuals and organizations currently using classroom video in early childhood education or developing new tools. These conversations highlight the promise and limitations of current approaches, especially regarding how emerging

technologies such as generative AI might reduce burdens on educators or improve the quality of instructional support. In addition to surfacing innovative use cases, this phase of work has helped identify persistent challenges that are likely to shape the feasibility of future investments.

Primary use cases for video analysis included automated scoring and feedback on classroom video observations using established classroom measures, which would increase capacity to provide teacher feedback. Other video use cases included the ability to generate a highlight reel of adult–child interactions for coaching, analyze peer dynamics, and examine curriculum implementation in real-world contexts. Primary use cases for work sampling included integrating teacher notes and observation data of child skills into a tool that could be aligned with curriculum. Additionally, improving child data displays can be especially helpful in early childhood contexts with high teacher turnover or limited teacher training.

Key challenges included operationalizing AI tools to "view" classroom dynamics and images, rather than only relying on audio transcripts, child-directed speech recognition, privacy in streaming and data storage, and ability to detect nuanced gestures and gazes. Some participants also noted challenges resulting from variation in curriculum and assessment use across programs and from intellectual property restrictions.

Key Emerging Themes. The study team identified the following four cross-cutting themes that are shaping the feasibility of AI-enabled tools in early childhood contexts.

- **Trust:** Concerns about the security of classroom video and image data, privacy of students, use of the data (by whom), and accuracy of AI-enabled analyses are shaping the feasibility of future innovations. These concerns, articulated by educators, parents, community members, and edtech investors, reflect real limitations of current technologies and perceptions that may be influenced by broader skepticism about AI. Addressing these concerns is a prerequisite to deeper investment in new tools and wide adoption and use in classrooms.
- **Productive Engagement:** Accurately assessing classroom dynamics requires detecting nonverbal indicators of student engagement, including interactions with teachers, peers, and materials. Current tools primarily rely on teacher audio and miss key visual cues such as body language, sustained attention, and indicators of dysregulation. Capturing these signals will enable a more nuanced assessment of children's experiences in the classroom and support more effective support of teachers.
- **Instructional Grouping:** Identifying whether children and teachers are engaged in whole-group, small-group, self-directed, one-on-one, or transition time provides essential context for interpreting classroom interactions. While models for detecting instructional grouping exist in K–12 settings, early childhood classrooms present unique challenges. Tailored approaches to grouping detection can serve as critical first steps in narrowing video analysis to specific segments and can permit the identification of the variety of micro-behaviors that are of interest for different use cases. These capabilities would reduce the computational demands of full-video processing and enable faster, more efficient analysis.

- **Seamless Documentation of Student Learning:** Teachers reported that documenting student progress for systems like Teaching Strategies GOLD or the Desired Results Developmental Profile (DRDP) is burdensome and often incomplete. AI offers the potential to align captured images, short video clips, or student work sample artifacts with the developmental criteria used in these systems. One group is piloting tools that translate teacher narratives into readiness indicators. These innovations could reduce documentation demands, preserve instructional time, and catalyze the development of new tools for capturing what pre-K children know and are able to do.

After conducting the interviews, we identified priority areas for fundamental algorithm or model development. Together, the study team and the foundation focused on identifying instructional groupings and identifying the presence of math or literacy content within classroom video.

Teacher and Instructional Coach Perspectives

The study team conducted five interviews with pre-K instructional coaches and four interviews with pre-K teachers to hear their perspectives on potential applications of AI in early childhood classrooms and better understand current classroom documentation practices and challenges. These interviews helped to further clarify where AI-enabled tools might offer the greatest value. The 45-minute interviews were audio-recorded. Participants received a $100 gift card. Transcripts were reviewed and cleaned for accuracy and uploaded to ChatGPT 5.0 for analysis support based on a list of key themes.

Teachers and coaches emphasized three overarching themes: (1) the essential but burdensome nature of data collection; (2) the potential for AI to reduce administrative load and reveal classroom patterns not easily observed; and (3) the need for technology that enhances—not replaces—teacher judgment, relationships, and developmentally appropriate practices. Teachers and coaches were largely aligned in valuing simple, actionable outputs and leveraging technology to support instructional practice.

How Educators Collect and Use Data

Interviewed teachers and coaches rely heavily on observational notes, child portfolios, DRDP assessments,[1] small-group documentation, and artifacts such as drawings, cutting samples, and photos. Many use tools to upload photos, videos, and notes; others rely on sticky notes and handwritten logs. One teacher, when describing the burden of documentation, said that "I'm drowning in sticky notes and then rewriting the same thing on [assessment measure]." Coaches take detailed observation notes to inform feedback cycles. Video is used selectively because of privacy restrictions, logistical barriers, or limited time. Across settings, educators noted that capturing spontaneous child interactions remains difficult. One teacher commented that "by the time I grab my Chromebook and get across the room, the moment is gone."

[1] The Desired Results Developmental Profile (DRDP) is a formative assessment tool for young children and is required for many early childhood programs in California.

Use of Technology in Classrooms and Coaching

Interviewed teachers and coaches want technology to be simple, fast, and reliably integrated into existing tools. Tools used include Learning Genie, DRDP online, Chromebooks, iPads, smart boards, First 5–supported apps such as Footsteps2Brilliance and MathShelf, and Google platforms. Teachers prefer simple tools like class phones for capturing photos and primarily use technology for documentation, planning, and communication with families. Coaches often travel across classrooms or districts, using technology for virtual professional development, modeling, or reviewing video clips.

Challenges and Pain Points

Interviewed teachers and coaches described significant burdens: time-consuming child assessment documentation, transferring media from phones, managing data for Individualized Education Programs (IEPs), tracking small-group participation, and observing multiple classroom areas simultaneously. Limited staffing exacerbates challenges. Transitions and behavioral incidents are difficult to monitor in real time. AI can provide visibility into classroom areas and moments that teachers cannot monitor while meeting other demands. One coach said that "if I could see exactly when behaviors spike—like always after transitions—I could coach so much more effectively." One teacher noted their desire for nonintrusive technology, "something that just sits there and captures what kids say during play, without me running across the room."

Reactions to Proposed AI Approaches for Instructional Grouping and Content Detection

Interviewed teachers and coaches saw clear value in AI that detects transitions, small-group time, whole-group time, and embedded content such as literacy or math during centers or play. One teacher commented, "I embed literacy all day without even realizing it. It would be amazing to see where it's actually happening." Educators emphasized that AI must provide brief summaries, graphs, and short video snippets rather than long recordings or reports. One teacher said, "Give me a graph or bullet points. I don't have time for a 10 minute video." Coaches noted that this information could support coaching cycles and identify patterns—such as long transitions or inconsistent instructional pacing—while teachers viewed it as a tool for self-reflection and lesson planning.

Desired Features and Educator Wish Lists

Interviewed teachers and coaches expressed interest in automated captioning and assessment-aligned tagging of videos/photos, AI-generated summaries for parent conferences, segmentation of classroom videos, identification of behavior patterns, lesson plan suggestions, and bilingual communication tools. One teacher commented that "if AI could write the summary for parent conferences—oh my gosh, that would save hours." High-quality implementation requires attention to privacy, consent, and developmental appropriateness. Educators worry about technology becoming evaluative rather than supportive; transparency and control are key. One teacher summed this point up by saying, "I want AI to help me—not evaluate me."

Identifying Instructional Groupings

Motivation

The potential value of grouping detection becomes especially clear when considering curriculum use at scale. As outlined in the 2024 National Academies of Sciences, Engineering, and Medicine report, *A New Vision for High-Quality Preschool Curriculum*, many early childhood curricula are designed around structured activity formats such as small-group exploration, whole-group discussions, and guided play. The report emphasizes that curriculum fidelity involves the delivery of specified content and attention to the intended context for that content, including group size and instructional format. Misalignment between the intended and actual use of materials, such as delivering a small-group curriculum activity in a whole-group setting, may compromise the developmental appropriateness and effectiveness of instruction. Automated detection of instructional groupings could provide actionable insights for curriculum developers, coaches, and program administrators who seek to understand whether materials are being used as designed across varied sites.

Methods

Framework Development

Two early childhood experts on the study team reviewed early childhood classroom observation tools to identify approaches to coding instructional groupings. Reviewed tools included Teacher Observation in Preschools (TOP; Bilbrey et al., 2007); Child Observation in Preschools (COP; Farran, 2014); Classroom Assessment Scoring System, 2nd Edition (CLASS; Pianta & Hamre, 2022); and Early Childhood Environment Rating Scale, 3rd Edition (ECERS; Harms et al., 2014). We synthesized and modified these elements and generated an annotation framework to classify video segments for the type of instructional grouping shown in classroom video recordings.

Early childhood teachers use a variety of instructional groupings for different purposes. For instance, a teacher may introduce a new concept (data analysis with a basic block graph) during morning meeting by asking children to select their favorite food from three options—pizza, cheeseburger, or mac and cheese—and then graphing responses on the whiteboard. Next, the children move into small groups and center times, in which one teacher works with a small group of children to practice block graphs with new data (e.g., favorite animals and favorite vehicles) for 10 minutes. Concurrently, the remaining children are in centers—some working with Unifix cubes to match cube quantities on a poster, some discussing favorite foods in the dramatic play kitchen, and some perusing books in the library center focused on counting and comparing objects.

Many of these activities can look similar to an outside observer. Distinguishing between the groupings can be helpful for an instructional coach or program director looking to see how a classroom spends its time and whether children are spending sufficient time in groupings such as small groups, when high-impact instruction is often best delivered. Identifying the groupings requires considering the teacher activity and the teacher–student interactions, and assessing what is happening across multiple student learning groups.

Therefore, the annotation framework included definitions of each grouping and precise start and end cues for each grouping. The study team iteratively modified the framework following practice annotation to refine and tighten definitions as well as to collapse codes that were too similar or could not be accurately distinguished. For instance, pre-K students commonly engage in independent work, either directly with a teacher (one-on-one) or individually completing activities with a teacher monitoring. Originally, independent work was split into three subcodes: one-on-one, independent work at tables, and independent work not at tables. After practice coding, we removed the subcode of independent work not at tables, as this code's definition was too similar to the definition of centers/free play and would be frequently mis-annotated. The original framework also included separate subcodes for meal times (breakfast, lunch, and snack). However, after discussion, the subcodes were determined to be difficult to distinguish and less important than the larger identification that meal time was occurring. The last modification occurred at the suggestion of the development team; rather than identifying rest time, recess, or no people visible as subcodes to "other," these codes were brought to the level of primary codes. The final framework contains 10 codes, each with a three-letter abbreviation for annotations (Table 1).

Table 1. Framework and Codes

Code	Definition	Abbreviation	Applicable Subcodes (Abbreviation)
Whole group	Students standing or seated near teacher in circle or non-circle	WGR	Circle time (CIR); carpet time (CAR)
Small group	Students engaged in the same activity at different areas in the classroom with at least one teacher near each group	SGR	–
Centers/free play	Students at different areas engaged in different activities, with or without other students	CFP	–
Independent	Teacher working with an individual student or students completing an activity in parallel	IND	One-on-one (UNO), independent work at tables (TAB)
Transitions	Students moving from one instructional grouping to another	TRN	Within classroom (INS), to outside (OUT), to inside (INS)
Meal time	In-class breakfast, lunch, snack	MEL	–
Rest time	Students lying down in different areas of the classroom	RST	–
Outside/recess	Students are outside	REC	–
No people visible	No students or teachers visible	NOP	–
Other	Anything that does not fit in other groupings	OTH	–

Data

The data used in this study came from two datasets and encompassed 20 hours of video across 20 classrooms. The study team recorded 13 hours of video in four preschool classrooms using a Swivl, a robotic camera tripod mount that rotates 360 degrees to track a person wearing a microphone marker. The lead teacher and assistant teacher both wore the lanyard microphone markers clipped to their collars to record audio. The field of view included the lead teacher's interactions with students and the classroom area adjacent to the teacher. The second dataset was videos recorded by another research group and primarily consisted of group activities (small group or whole group). The field of view included at least one teacher and the group of children with which the teacher was interacting (Curenton, 2015). Both datasets included releases from both teachers and families allowing their use for research purposes. For all videos, the data were only used to test, not to train, the model. See the Modeling Approach section for details regarding the protection of these data.

Annotations

Following framework development, the study team created a template spreadsheet for annotating classroom videos (Figure 1). The template allows annotators to identify the start and end times for episodes of up to three distinct instructional groupings (and their subcodes) occurring concurrently. As an annotator watched a video, they would enter the primary grouping (code1) during a time period and, if applicable, any additional groupings co-occurring (code2 and code3). Primary groupings were those with which most students (~75%) were engaged. For example, Figure 1 shows that from 2:00 to 3:30, students are transitioning into the classroom. At 3:31, most students are still transitioning into the classroom, but some students begin independent work at tables. This approach of annotating continuous blocks of classroom video was necessary because instructional groupings and transitions tend to be in larger chunks of time (e.g., 3–20 minutes). Human experts can more efficiently detect transitions in this manner rather than reviewing short snippets of video for labeling. Annotators tested and refined the template with practice videos. The development team used completed spreadsheets containing timestamps and codes for all instructional grouping segments, to represent correct "ground-truth" information during model evaluation.

Figure 1. Instructional Grouping Annotation

A	B	C	D	E	F	G	H	I
start	**end**	**code1**	**sub1**	**code2**	**sub2**	**code3**	**sub3**	**Duration**
0:02:00	0:03:30	trn	int					0:01:30
0:03:31	0:05:20	trn	int	ind	tab			0:01:49
0:05:21	0:05:50	cfp		ind	tab	trn	int	0:00:29
0:05:51	0:09:32	ind	tab	trn	int			0:03:41
0:09:33	0:34:00	cfp		ind	tab			0:24:27
0:34:01	0:39:06	cfp						0:05:05
0:39:07	0:40:41	trn	ins					0:01:34

Annotator Training and Reliability

Three human annotators with experience in early childhood classrooms attended a 1.5-hr training and practiced coding six 10-minute video segments. The annotator team compared their annotations to those of the lead annotator, and annotators were given feedback to refine their codes to align with the lead annotator. Once the reliability threshold of 80% exact agreement was achieved on both content and timestamp on the training videos, annotators coded the remaining video data across two weeks. For ease of annotation, videos assigned to annotators were portions of longer recordings, approximately 20 minutes each. Annotators were instructed to watch the video at 1- or 2-times speed and note the start time, end time, and instructional grouping(s) present. Two annotators coded 10 segments that were compared for reliability. Annotations were first split into common time segments (with the same start time and end time) to allow for meaningful comparison. Exact agreement was also calculated between annotators and weighed by timestamp duration length; 83.5% of annotations were matched exactly, demonstrating high interrater agreement.

Modeling Approach

The study team determined the best model approach as **analysis of videos using multimodal large language models** (MLLMs; Tang et al., 2025) and utilizing prompts that describe the instructional grouping rubric. This approach relies on the capabilities of an MLLM to summarize images or videos as text and analyze the text based on user prompts. Contrasting with an analysis of videos using a suite of pretrained machine learning models, this approach requires no training data for model supervision or fine-tuning. For this work, we determined that the MLLM would be a better fit because it is faster to develop, is more general, and does not require further training and a suitable dataset to train on.

We applied the MLLM approach to 20 hours of preschool classroom video data using models such as Google Gemini (Gemini Team Google, 2025) and the OpenAI GPT family of models (OpenAI, 2024). While either model could be used for the video analysis, we used only OpenAI GPT models for evaluation on the dataset. SRI has an Enterprise license for GPT, which ensures that images uploaded into the model will not be used to train models outside of SRI's closed system. Given that images of teachers and students would be included in the classroom images, we selected the highest security model to protect their data privacy and ensure that GPT users outside of SRI would not be able to see their images.

We used a prompt-based approach to instructing the MLLM how to label video segments. The prompt to the MLLM consists of two parts: the instructions to the MLLM and the classification indices and descriptions for instructional grouping categories. These can be used with minor modifications across different families of LLMs. For technical details, see Sundaresan et al. (2025a).

Instructional Grouping Results

The study team evaluated two OpenAI GPT models, 4.1 and 5, on our annotated dataset. Overall, GPT-5-mini had a higher score than GPT-4.1-mini (Table 2). We used (1) a **full** classification for instructional grouping that includes all 10 categories and seven subcategories, and (2) a **simple** prompt that includes only the 10 categories to evaluate if the MLLM performed better with fewer categories. The "Primary Category" column in Table 2 shows the percentage of segments where the system-produced primary category matched any of the human annotations. The model also outputted a secondary category, and the "Either Category" column shows the accuracy when **either** the primary or secondary system-produced category matched any of the human annotations. Overall, the GPT-5-mini model using the simple framework prompt had the highest accuracy of the four model and prompt combinations, with an accuracy of 85.4% in identifying either category.

Table 2. Summary of OpenAI GPT Models with Full and Simple Prompts

Model	Prompt	Primary Category	Either Category
GPT-4.1-mini	Full	71.7%	75.6%
GPT-4.1-mini	Simple	56.7%	62.6%
GPT-5-mini	Full	74.0%	81.9%
GPT-5-mini	Simple	75.6%	85.4%

A noticeable trend is that all the models consistently underestimated the centers/free play (CFP) grouping while overestimating the whole group (WGR) grouping. Small group (SGR) was commonly misclassified as CFP and independent (IND); early childhood education experts on the team noted that this misclassification may be due to similar teacher/student makeup in the videos and the descriptive rubric. During SGR, students are seated or arrayed near at least one teacher who is delivering instruction. When students are similarly arrayed near a teacher but not receiving direct instruction, the class is annotated as CFP. The IND class is similar in that teachers deliver instruction to students but are not required to be near them throughout the period, although they may be. Another misclassification trend is that transitions (TRN) was commonly confused for CFP. Both TRN and CFP contain similar elements of movement, with both teachers and students moving separately or together in the classroom. Misclassification tends to occur when there are similar visual fields of the grouping, but additional context (e.g., transition cue, teacher instruction) may not have been considered by the model.

Limitations

The dataset used to test the instructional grouping model was small and therefore does not fully represent the diversity of classroom settings or teacher variability. We offer the approach, framework, and prompt language to encourage further testing and refinement with a larger variety of classrooms and teachers. Additionally, the goal of the project was to test the feasibility of video as a data source, not to identify the most efficacious prompt and MLLM combination. Therefore, while we developed a prompt with minimal iterations, we did not deeply examine mechanisms to improve the result. Potential future improvements

include exploring additional MLLMs and continuing to revise the language of the prompt and framework with iterative approaches and more diverse data.

Identifying Literacy and Math Content

Motivation

The study team recognized an opportunity to leverage SRI's existing AI capabilities to identify math and literacy content in videos of pre-K classroom instruction. APPROVE (Assisting Parents to Review Online Videos for Education) Literacy and Math[2] is an AI tool developed by SRI to automatically detect standards-aligned early literacy and math content—and associated pedagogical quality—in online videos for pre-K and kindergarten children. Using a multimodal machine learning framework that integrates visual and audio analysis, APPROVE identifies features aligned with the Common Core State Standards and the Head Start Early Learning Outcomes Framework (Gupta et al., 2023). The tool was designed to help researchers, parents, and educators understand the content children encounter when using online video platforms such as YouTube, where content is released so rapidly that manual review cannot keep pace. The model detects early literacy and math content with high precision as compared with humans, outperforming chance across both curated and naturalistic datasets of children's YouTube viewing (Christensen et al., 2025).

While the APPROVE Literacy and Math model was designed to identify literacy and math content in children's media, employing a similar approach to identify literacy and math content in videos of pre-K classroom instruction can provide valuable insights into teaching practices and curriculum implementation. Further, it can support teachers in authentically integrating standards-aligned literacy and math content throughout the school day outside of designated instructional periods (e.g., counting to 10 during a transition, asking a child to sit down on a letter on the carpet). By automatically detecting when and how concepts aligned with the Common Core State Standards and the Head Start Early Learning Outcomes Framework (Office of Head Start, 2015) are introduced, educators and researchers can analyze instructional quality and ensure alignment with research-based standards. These capabilities can also allow instructional coaches to more efficiently review and provide teachers with targeted feedback and professional development resources around academic content during coaching sessions.

[2] APPROVE's development has been funded by SRI internal research grants as well as by the National Science Foundation (Award No. 2139219) and the U.S. Department of Education (Award No. R305J250019). APPROVE was patented in 2024 (U.S. Patent No. 18/737,444).

Methods

Data

The datasets used for this approach were a subset of those used for the instructional groupings approach. This subset included 130 minutes of video across 14 classrooms. See the Data section under Identifying Instructional Groupings above for additional details.

Annotations

The study team adapted the APPROVE Pre-K & Kindergarten Literacy and Math codebooks for use with preschool classroom video. As this project focuses on pre-K classrooms, we retained only pre-K-relevant codes and simplified the more complex codebook by reducing the number of codes. We revised the codebooks for classroom contexts by replacing references to on-screen characters or narration with teacher behaviors and by narrowing pedagogical quality codes to two dimensions: whether the learning content was the primary focus and whether the teacher connected it to children's everyday experiences. The lead codebook developer and one annotator tested the revised codebook on a small sample of videos, clarifying ambiguous definitions and resolving discrepancies.

After finalizing the codebook, the researcher who modified the codebook trained two annotators, who achieved at least 80% agreement with the researcher on all content codes across 15 test videos. Annotators and the researcher then coded classroom videos over three weeks, with 16% overlap to assess reliability. The team met weekly to review one shared video, discuss discrepancies, and prevent drift. Table 3 presents interrater reliability for the final dataset, which consisted of 130 one-minute video segments. Raters demonstrated high percent agreement for both literacy and codes. Kappa was similarly high for literacy, but moderate for math. Given that there are five literacy codes and 10 math codes, it may have been more challenging for human annotators to reach the same degree of reliability.

Table 3. Interrater Reliability for Literacy and Math Codes

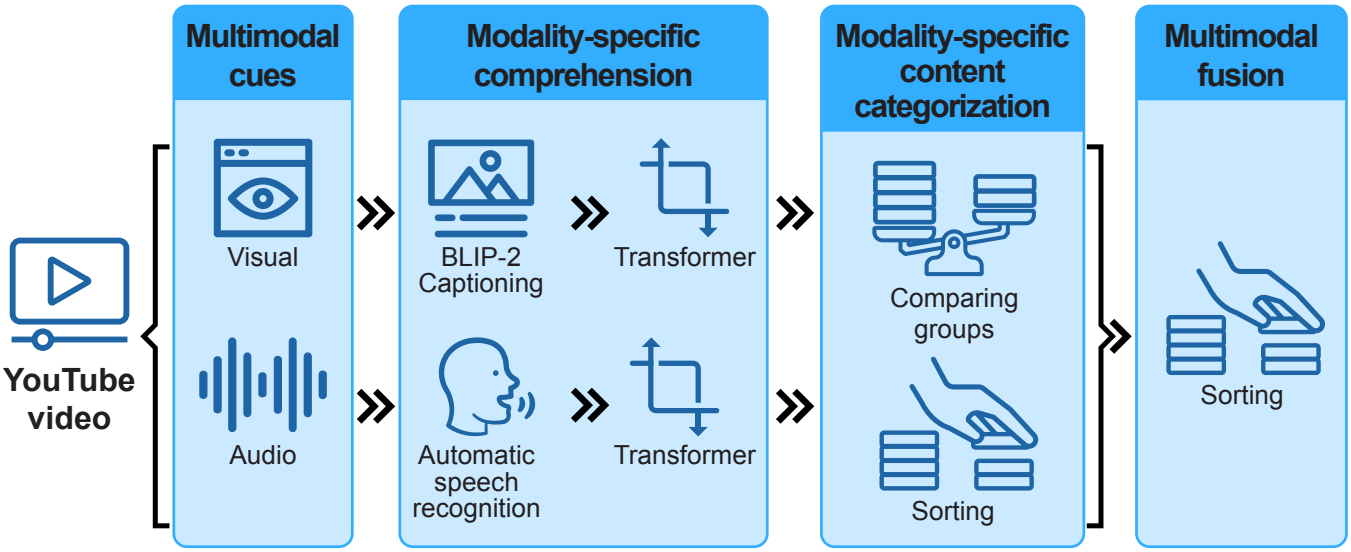
Codes	Percent Agreement	Kappa
Literacy	98.1%	.92
Math	93.3%	.60

Video Classification and Evaluation

The study team applied an existing AI model—a pretrained transformer-based deep neural network (Vaswani et al., 2017) —to analyze instructional videos. In prior work, SRI researchers developed two separate AI models, one that processed the audio stream and another that analyzed the visual stream, for both literacy and math content (Gupta et al., 2023). For the audio, the model first converts the spoken language into text using automatic speech recognition. The audio model then predicts which instructional categories are present in each video segment. For the visuals, the video model extracts key frames from each video and uses another

AI model, Bootstrapping Language-Image Pre-training (BLIP-2; Li et al., 2023), to generate short text captions describing the visual content. These captions are subsequently analyzed by the video model using the same classification approach as the audio model. Finally, a late fusion approach combines the outputs of the audio and video models (Figure 2). However, for this project, integrating the two modalities did not improve accuracy, primarily because the captions generated from the video frames lacked sufficient detail about the instructional content. Consequently, we report only the results derived from the audio analysis.

Figure 2. Model Framework



Math and Literacy Results

Overall

A summary of the model's performance results are shown in Table 4. The table includes key measures such as the true positive rate (TPR) and precision. The TPR—also called recall—shows the fraction of actual positive codes that are identified as positive codes. Precision is the fraction of codes identified as positive that are correct. We set the false positive rate (FPR) at 0.2, meaning that up to 20% of the noninstructional segments could be incorrectly labeled as instructional. This threshold was chosen so the model would capture more possible instructional moments, even if it included some extra ones. In practice, it is easier for users to remove extra segments than to find ones the model might have missed.

In the math model, *Sorting* and *CompGroup* performed best, although they appeared only a few times. *Subitizing*, *Counting*, and *Numerals* also performed well and occurred more often, while *Spatial Language* (*SpatialLang*) showed the weakest results. For the literacy model, most codes reached at least 40% recall, except for *LetterSound*. Precision for literacy ranged from 0.65 to 0.83, meaning that more than half of the identified segments were accurate. Overall, recall was higher for the math model, suggesting that it captured a greater share of relevant math content compared to the literacy model. Detailed performance graphs for each model are provided in the technical white paper (Sundaresan et al., 2025b).

Table 4. Summary Performance for Math and Literacy

Mathematics			
Code	**TPR/Recall**	**Precision**	**N**
AddSubt	0.80	0.80	4*
CompGroup	1.00	0.83	1*
Counting	0.83	0.81	23
MeasAtt	0.67	0.77	6*
Numerals	0.70	0.78	20
Patterns	0.50	0.71	2*
ShapeID	0.78	0.78	10
Sorting	1.00	0.83	2*
SpatialLang	0.41	0.67	39
Subitizing	0.86	0.81	21
Average	**0.75**	**0.71**	

Literacy			
Code	**TPR/Recall**	**Precision**	**N**
FollowWords	1.00	0.83	1*
LetterName	0.45	0.69	40
LetterSound	0.41	0.67	54
OneSyllable	0.38	0.65	16
Rhyme	0.50	0.71	8
Average	**0.55**	**0.71**	

Note: *All codes have a false positive rate of 0.2 and a true negative rate (TNR) of 0.8. N = number of occurrences in dataset. * = few occurrences.*

Math

Across codes, the model correctly identified a high number of positive cases, or instances where a code was present, at a fixed FPR of 0.2. The *SpatialLang* code, which involves words that describe direction, order, and position (e.g., up, down, in front of, behind), showed a higher number of false negatives. In 23 cases, the model classified the code as not present even though human annotators identified it in the video. This may be because the audio model did not capture visual demonstrations of these concepts, such as a teacher gesturing to a group of three teddy bears and pointing to the one that is behind the others. Detailed results and confusion matrices are provided in the technical white paper.

Literacy

Results show variation in model performance across different codes, also at an FPR of 0.2. The *LetterSound* code, where a teacher says the sound of a letter and highlights the corresponding letter or object, had a relatively high number of false negatives. In these cases, the human annotator marked the code as present, but the model did not. This may be because the code involves two components: verbalizing the sound and visually indicating the letter or object. While the model may have detected the spoken sound, the audio model would not have recognized the associated gesture, particularly in classroom videos where teachers do not always face the camera as directly as in instructional YouTube videos.

Limitations

We conducted these analyses to determine if an existing machine learning model trained to identify math and literacy content in YouTube videos could identify similar content in pre-K classroom videos. We made only minor adjustments to the existing model and did not focus on improving its performance. The model was tested with a small dataset (20 classrooms) that does not fully capture the variety of pre-K classroom environments or curricula, and the video data cover a short amount of time (130 minutes). Our goal is to share the literacy and math content frameworks informing our models to encourage others to test and refine them with a wider range of classrooms. In the future, researchers could explore other models, continue training the model, or develop a prompt-based model that allows easy updates to prompts and framework.

Although the machine learning model was designed to be multimodal, current results rely only on audio transcripts. The video captions were not detailed enough to capture classroom activities accurately, likely because of the typical visual clutter of a pre-K classroom (e.g., posters, objects, child movement). Classroom videos also differ from typical educational videos on platforms like YouTube, in which content creators typically directly face the camera and use a limited number of visuals. In classroom videos, teachers may not directly face the camera or may use small physical materials, which make classroom videos more challenging to analyze automatically. Improving video captioning should help increase the model's accuracy.

Recommendations and Future Steps

Early childhood education is one of the fastest growing segments of the edtech market (HolonIQ, 2024), and AI is likely to be an embedded technology feature in most new products. Based on the use case research as well as the development and feasibility testing of two models for using classroom video with AI, SRI recommends the following next steps for the field and investment:

- Continue to engage with tool developers, curriculum providers, program providers, and venture capitalists to maintain an updated view of the uptake of new technologies, models for implementation, and teacher- and coach-led creative thinking for high-impact use cases.
- Carefully consider how human–AI pairings can be more than the sum of their parts. AI has been touted as a valuable, time-saving replacement for monotonous work, but AI's power lies in the recognition of patterns and data and ability to provide insightful summaries of large quantities of information. Consider how AI can not only free up time so that educators and coaches can focus more on the human element of early childhood education, but also provide opportunities for ingestion and distillation of large amounts of data for insights that focus and improve outcomes.
- Invest in building blocks—models like identification of instructional grouping—that can be applied in a wide variety of ways. The field could use reliable API-enabled models that are public goods to build novel and responsive applications to support educators, coaches, and program administrators.
- Ensure transparency in model approaches, decisions, and outputs to engender trust. Require

generative, prompt-based models to explain the reasoning in output (see the instructional grouping examples in this report), and use confusion matrices in addition to simple accuracy rates to understand how the models get it wrong (see Sundaresan et al., 2025a and 2025b for example confusion matrices). Support the exploration of errors, even at small rates, to identify outlying cases and ensure explainability for future use.

- Look for opportunities to cross original use cases for fundamental models. SRI had originally developed academic content labeling models for YouTube videos, but when applied to the classroom dataset, the model performed much better than expected.
- Adopt an abundance mindset over scarcity. The field is rich with edtech solutions and research on AI in education, and the use of AI to evaluate video. Look for opportunities to leverage and expand existing work or combine partners for amplifying effects. At the same time, abundance does not mean completeness—use needs-sensing, use case analysis, and landscaping to identify critical gaps in need of investment.

Enrollment and investment in early childhood education continues to grow in the United States, particularly with the recent expansion of universal pre-K in several states. As this growth accelerates, early childhood education is ripe for research and innovation aimed at improving outcomes for both educators and children. While states invest more in early childhood education, many educators and pre-K providers struggle to expand capacity and provide all children with high-quality classroom experiences. AI can help meet this demand and create a better understanding of the routines and patterns in early childhood classrooms. Pairing AI with classroom video offers powerful applications for coaching, student documentation, and assessments.

In addition, AI is showing signs of improving efficiency across many professions. Thoughtfully applying AI tools in early childhood settings may also help educators reduce administrative burdens to focus their time on where it is needed most: building strong relationships with the children in their classroom. As with any new technology, questions remain about the safety and quality of early AI applications. It is essential that innovation is coupled with rigorous research and educator perspectives. Investing in capabilities that address multiple use cases and work across platforms can help ensure that AI tools designed for early childhood education are scalable, actionable, and aligned with the field's broader goals.

References

Bilbrey, C., Vorhaus, E., & Farran, D. C. (2007). *Teacher Observation in Preschools (TOP)*. Peabody Research Institute, Vanderbilt University.

Christensen, C., Cincebeaux, M., Roy, A., & Kim, S. (2025). A machine learning model to detect early math content in YouTube videos. *Artificial Intelligence in Education, 1*(1), 75–92. https://doi.org/10.1108/AIIE-12-2024-0050

Curenton, S. (2018). *Conversation Compass: A teacher's guide to high-quality language learning in young children*. Databrary.

Farran, D. C. (2014). *Child Observation in Preschools (COP)*. Peabody College, Vanderbilt University.

Gemini Team Google. (2025). *Gemini: A family of highly capable multimodal models*. arXiv. https://arxiv.org/abs/2312.11805

Gupta, R., Roy, A., Christensen, C., Kim, S., Gerard, S., Cincebeaux, M., Divakaran, A., Grindal, T., & Shah, M. (2023). Class prototypes based contrastive learning for classifying multi-label and fine-grained educational videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 19923–19933. IEEE. https://openaccess.thecvf.com/content/CVPR2023/html/Gupta_Class_Prototypes_Based_Contrastive_Learning_for_Classifying_Multi-Label_and_Fine-Grained_CVPR_2023_paper.html

Harms, T., Clifford, R. M., & Cryer, D. (2014). *Early Childhood Environment Rating Scale–third edition manual*. Teachers College Press.

HolonIQ. (2024). *2025 global education outlook*. https://www.holoniq.com/notes/2025-global-education-outlook

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). *BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models*. arXiv. https://doi.org/10.48550/arXiv.2301.12597

National Academies of Sciences, Engineering, and Medicine. (2024). *A new vision for high-quality preschool curriculum* [Consensus study report]. The National Academies Press. https://doi.org/10.17226/27429

Office of Head Start. (2015). *Head Start early learning outcomes framework: Ages birth to five*. U.S. Department of Health and Human Services, Administration for Children and Families. https://headstart.gov/school-readiness/article/head-start-early-learning-outcomes-framework

OpenAI. (2024). *GPT-4 technical report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Pianta, R. C., & Hamre, B. K. (2022). *CLASS 2nd edition: Pre-K–3rd reference manual*. Teachstone.

Sundaresan, A., DeLyser, L. A., Syed, G., Gerard, S., & Niekrasz, J. (2025a). *Developing an AI-supported approach to identify instructional groupings in early childhood education classrooms* [Technical report]. SRI. https://www.sri.com/publication/education-learning-pubs/developing-an-ai-supported-approach-to-identify-instructional-groupings-in-early-childhood-education-classrooms/

Sundaresan, A., DeLyser, L. A., Gerard, S., Syed, G., Perez, N., Niekrasz, J., & Christensen, C. (2025b). *Developing an AI model to identify math & literacy instruction in early childhood education classrooms* [Technical report]. SRI. https://www.sri.com/publication/education-learning-pubs/developing-an-ai-model-to-identify-math-and-literacy-instruction-in-early-childhood-education-classrooms/

Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., Vosoughi, A., Huang, C., Zhang, Z., Liu, P., Feng, M., Zheng, F., Zhang, J., Luo, P., Luo, J., & Xu, C. (2025). Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*. https://doi.org/10.1109/TCSVT.2025.3566695

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds)., *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010). Association for Computing Machinery.

SRI Education, a division of SRI, is helping federal and state agencies, school districts, major foundations, nonprofit organizations, and international and commercial clients tackle some of the most complex issues in education to help students succeed. Our mission is **to reduce barriers and optimize outcomes for all children, youth, and families**. We do this by conducting high-quality research, supporting use of data and evidence, helping to strengthen state and local systems, and developing tools that improve teaching and accelerate and deepen learning. Our work covers a range of topics, including early learning and development, student behavior and well-being, teaching quality, digital learning, STEM and computer science, literacy and language arts, and college and career pathways.

SRI is a nonprofit research institute whose innovations have created new industries, extraordinary marketplace value, and lasting benefits to society.

Silicon Valley
(SRI Headquarters)
333 Ravenswood Avenue
Menlo Park, CA 94025
+1.650.859.2000
education@sri.com

Washington, DC
1100 Wilson Boulevard
Suite 2700
Arlington, VA 22209
+1.703.524.2053
www.sri.com/education-learning/

STAY CONNECTED