Developing an AI Model to Identify Math & Literacy Instruction in Early Childhood Education Classrooms

Technical White Paper

December 2025

Prepared by:
Aravind Sundaresan, Leigh Ann DeLyser, Sarah Gerard, Gullnar Syed, Nancy Perez, John Niekrasz, and Claire Christensen

SRI

Suggested citation:
Sundaresan, A., DeLyser, L. A., Gerard, S., Syed, G., Perez, N., Niekrasz, J., & Christensen, C. (2025). *Developing an AI Model to Identify Math and Literacy Instruction in Early Childhood Education Classrooms* [Technical report]. SRI.

Contents

Abstract

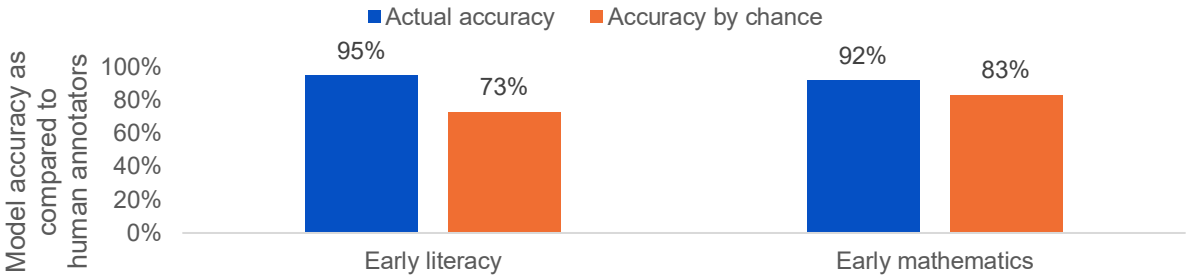
A high-quality early childhood classroom provides children with a safe and nurturing environment to develop their physical, social, emotional, and academic capabilities. Children may receive literacy instruction during a morning circle story read-aloud, and then later break into small-group instruction focused on comparing quantities of dinosaur toys. Additional instructional moments also arise during informal play and learning, such as children counting off in line to go out to recess. Many early childhood providers use video to support teacher development, classroom observations, or for parents to check in on their children during the day. Researchers from SRI, supported by the Gates Foundation, explored opportunities to leverage AI and video from early childhood classrooms to explore AI-supported approaches to identify when math and literacy instruction happens. The research team used previously developed models that label academic content in YouTube videos, repurposing the models to measure their ability to detect instruction content in classroom settings. This technical white paper describes the research team's approach to automatically identifying academic content, the data, initial model performance, and implications of this work.

Introduction

To automatically detect instructional moments in a Pre-K classroom, the study team used SRI's existing AI capabilities to identify math and literacy content in videos of Pre-K classroom instruction. APPROVE Literacy and Math [1] is an AI tool developed by SRI to automatically detect standards-aligned early math and literacy content—and associated pedagogical quality—in online videos for prekindergarten and kindergarten children. Using a multimodal machine learning (MML) 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 [2]. The tool was designed to help researchers, parents, and educators understand the content children encounter when using online video platforms such as YouTube, where videos are released so rapidly that manual review cannot keep pace. The model has previously been shown to detect early math and literacy content with 92%–95% accuracy in YouTube videos with respect to manual annotations and to outperform chance across both curated and naturalistic datasets of children's YouTube viewing (Figure 1) [3]. Figure 1 shows the accuracy of the APPROVE model, 95% for early literacy and 92% for math instruction detection, compared to chance. Chance is higher than 50% because the data used to test the model does not have evenly distributed examples of all the annotated standards. For example, one of the annotations may only exist in 10% of the videos, and therefore a strategy of always guessing "no content" would yield a 90% accuracy rate. Therefore, we compare the accuracy of the model to chance for clarity in terms of its performance.

In the current study, the research team analyzed how effectively the same MML framework could automatically classify moments of math and literacy instruction in preschool classroom videos. In this paper, we detail the dataset used, how it differed from the original data that trained the model, our application of the previous model, and its performance against classroom videos.

Figure 1. APPROVE accurately detects early math and literacy content in YouTube videos



Math and Literacy Codebook Development

Dataset

Video recordings were collected from four preschool classrooms and totaled 113 minutes, or 2 hours. Full-length recordings were segmented into 1-minute segments. Each segment was then analyzed for instructional content. Segments captured different types of instructional moments, such as whole-group book reading, teacher-led small-group math instruction, and classroom activities such as morning meeting, lunch, and transitions. Videos often contained multiple students moving around the classroom, students receiving 1-on-1 behavior support, and time spent in centers such as the blocks area, library, or dramatic play kitchen area. In many classrooms, 2–3 educators (often a lead teacher, co-teacher, and/or paraeducator) led or monitored groups of children in concurrent activities.

Research team members collected videos using an iPad on 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 lanyard microphone markers clipped to their collars to record audio. The tripod was assembled so that the iPad was at chest height with the teacher and the Swivl rotated to track the movements of the lead teacher around the classroom. This allowed for the field of view to include the lead teacher's interactions with students and the classroom area adjacent to the teacher. Data collectors were present during the recording and did not move the tripod during the recording, unless the device became an obstacle for classroom movements. In these few cases, recordings were stopped, and the camera was repositioned. The research team obtained permission from teachers for the video to be used for this purpose, and parents had the opportunity to "opt out" their child from participation. The dataset will not be released for public use to ensure privacy of participating teachers and children.

A second dataset, containing an additional 20 minutes of video from 10 classrooms, was added to increase the variety of classroom settings in the existing dataset. These recordings captured a range of classroom activities, with an emphasis on small-group interactions. The research team established a data-sharing agreement with an external researcher to obtain permission to access these videos, which included releases from both teachers and families allowing their use for research purposes [4]. In total, the combined dataset included two hours of video from 14 classrooms.

Framework

The research team adapted the APPROVE Pre-K & Kindergarten Mathematics and Literacy codebooks for use with preschool classroom video (see Appendices A and B). As this project focuses on Pre-K classrooms, the team retained only Pre-K-relevant codes and simplified the more complex codebook by reducing the number of codes. The team 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. See Table 1 for a sample of codes and their descriptions.

Table 1. Codebook examples

Indicator	Description	Subject
Numerals	Teacher highlights one written numeral AND either highlights the corresponding quantity of objects OR verbalizes the number.	Math
ShapeID	Teacher highlights and names a shape.	Math
LetterSound	Teacher verbalizes the sound of a letter and highlights either the letter, or objects that illustrate the letter sound.	Literacy
Rhyming	Teacher says the word "rhyme" or "rhyming" AND says at least two rhyming words.	Literacy

Annotator Training and Reliability

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 2 presents interrater reliability for the final dataset, which consisted of 130 one-minute video segments. Raters demonstrated high percent agreement for both math and literacy codes. Kappa was similarly high for literacy, but moderate for mathematics. 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 2. Human interrater reliability for math and literacy codes

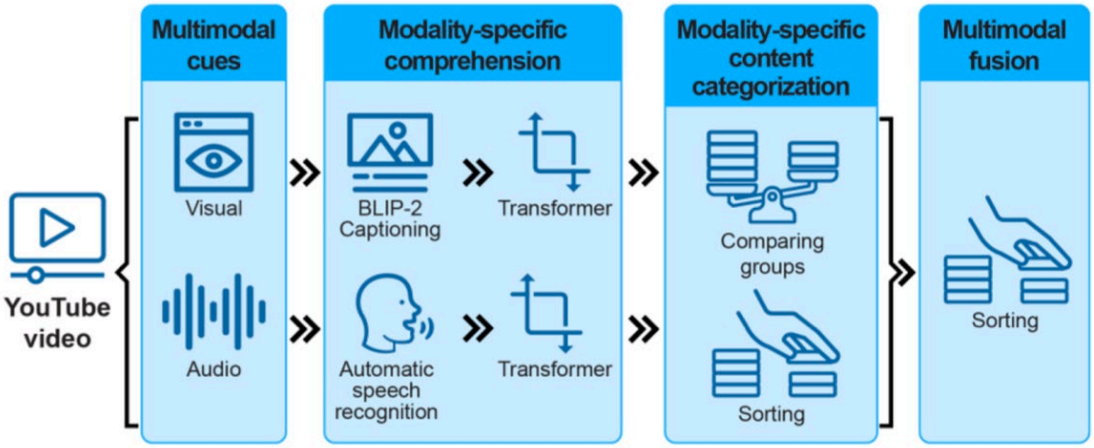
Codes	Percent Agreement	Kappa
Mathematics	93.3%	.60
Literacy	98.1%	.92

Video Classification and Evaluation

Model Description

We applied a pre-trained transformer-based [5] deep neural network (DNN) model on the videos. In our previous work [6], we trained a pair of transformer models for audio and video streams for each type of content (TROVE [math] and APPROVE [literacy]). The audio from the sequence is transcribed to text, using automatic speech recognition (ASR) on SRI's Open Language Interface for Voice Exploitation (OLIVE) platform and OpenAI Whisper, and input to the audio model. The audio model then identifies if each instructional content category is present in the segment through a multilabel classifier output. Key frames are extracted from the video and converted to captions using the Bootstrapping Language-Image Pre-training (BLIP-2) [7] model. These captions are input to the video transformer model, which produces the same type of output as the audio model. Finally, we perform late fusion to combine the outputs of the audio and video models. The framework is displayed in Figure 1. Full details are described in the original AI tool [6]. We note that the captions obtained from the key frames in our dataset are not sufficiently descriptive of the instructional content and, as a result, the video model performance is barely above chance. Consequently, late fusion does not improve overall accuracy, and we therefore present the results of only the audio data.

Figure 1. Model framework



Results

Overall

Summary performance metrics, including the true positive rate (TPR) and precision, are displayed in Table 3. The TPR is the fraction of actual positive codes that are identified as positive codes (same as recall). Precision is the fraction of codes identified as positive that are correct. The false positive rate (FPR) is set at 0.2 and is the fraction of actual negative codes that are identified as positive codes. This threshold was set to allow for a higher number of segments with instructional content to be identified rather than missing segments; the potential users of this model could more easily manually delete extraneous segments rather than identify additional segments. Therefore, a higher recall number is desired. For the literacy model, most

codes achieve at least a 40% recall, except for *OneSyllable*. Precision ranges between 0.65 and 0.83 for literacy codes, showing that more than half of the segments identified by the model were accurate. For math, *Sorting* and *CompGroup* perform the best, although they have very few occurrences. *Subitizing*, *Counting*, and *Numerals* perform well and have significant number of occurrences. *SpatialLang* performs the worst. Recall is also strong for most codes in math, showing that high proportions of math content were identified by the math model compared to the literacy model. For more details about the performance of each of these models, see the receiver operating characteristic (ROC) curves in Appendix C.

Table 3. 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**	

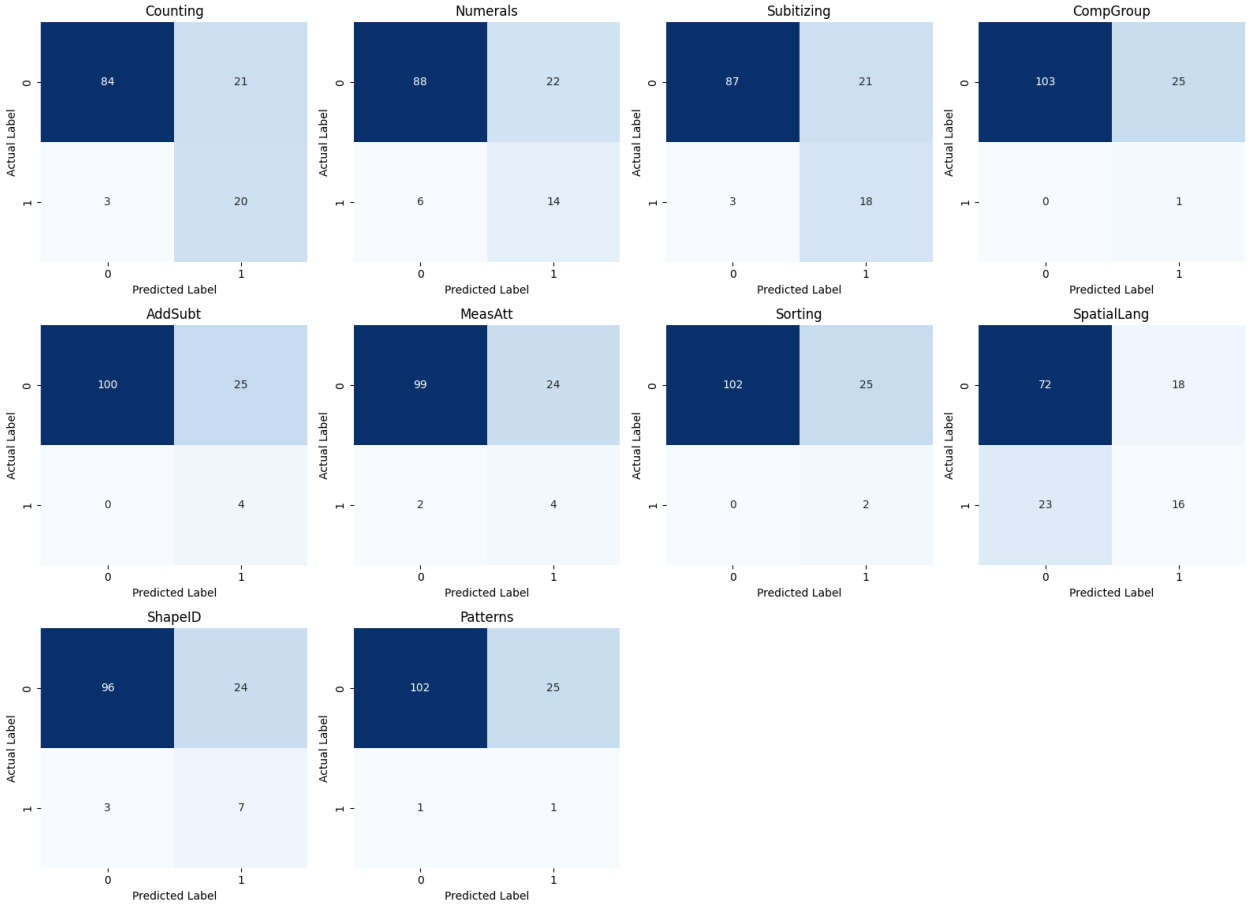
Note: All codes have an FPR of 0.2 and a true negative rate (TNR) of 0.8. *N* = number of occurrences in dataset.
* = few occurrences.

Confusion matrices, which display counts of false positives, false negatives, true positives, and true negatives, are also plotted for each class for FPR = 0.2 (Figures 3 and 4). The y-axis indicates the actual label of the segments, and the x-axis indicates the predicted labels. In the ideal scenario, in which all predictions are correct, only the diagonal from the upper left to the bottom right will have nonzero values.

Math

Across codes, a high number of positive cases (instances where a code is present) are correctly identified (at a fixed FPR of 0.2; Figure 3). Spatial language (*SpatialLang*), or language related to the directionality, order, and position of objects (e.g., up, down, in front of, behind), has a higher proportion of false negatives, with 23 cases classified as not present by the model when the human coded them as present in the video clip. This may be because the audio model did not capture visual demonstrations of these concepts, such as gesturing to a set of three teddy bears and pointing to the bear that is behind the others.

Figure 3. Math confusion matrices

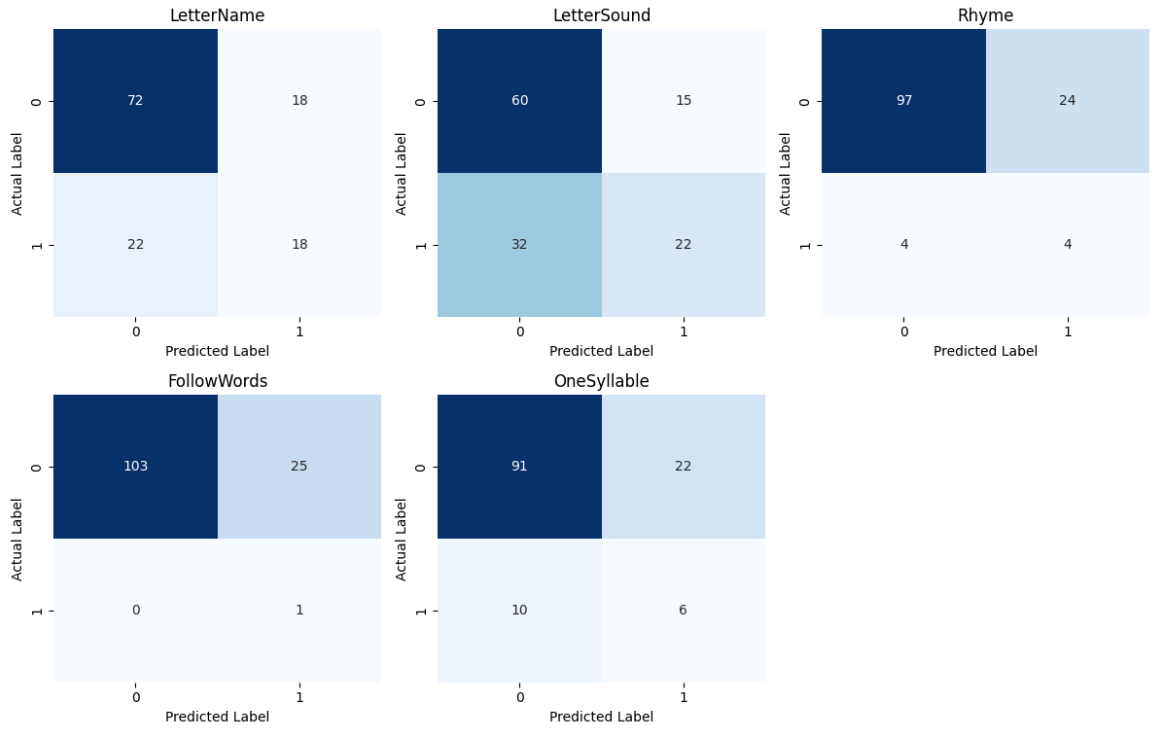


Note: Matrices are displayed with an FPR of 0.2. Number of occurences for each code: Counting = 23; Numerals = 20; Subitizing = 21; CompGroup = 1; AddSubt = 4; MeasAtt = 6; Sorting = 2; SpatialLang = 39; ShapeID = 10; Patterns = 2.

Literacy

The literacy confusion matrices (Figure 4) depict variation in performance across codes, fixed at an FPR of 0.2. *LetterSound*, in which the teacher verbalizes the sound of a letter and highlights the letter or object that illustrates the letter sound, has a higher rate of a false negative (32), in which the human annotator identified the code but the model did not. This particular code requires two components: 1) the teacher verbalizes the sound and 2) the teacher gestures to or highlights the letter associated with the sound. In our analyses, the audio model alone would not have captured the gesture of highlighting the letter or object associated with the sound, as that component of the code is primarily visual.

Figure 4. Literacy confusion matrices



Note: Matrices are displayed with an FPR of 0.2. Nunber of occurences of each code: LetterName = 40; LetterSound = 54; Rhyme = 8; FollowWords = 1; OneSyllable = 16.

Limitations

The research detailed in this white paper has limitations to its generalizability. The data used to test the model was small, including 14 classrooms, which does not fully represent the diversity of classroom settings or teacher variability. Additionally, the time represented in dataset is relatively small (130 minutes). We offer the approach and framework 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 develop the most efficacious MML framework that integrates visual and audio analysis. Therefore, while we utilized an existing MML with minimal iterations, we did not deeply examine mechanisms to improve the result. Potential future improvements include exploring additional models and continuing to train the model. Alternatively, one could develop a prompt-based multimodal large language model (LLM) allowing for regular revisions of prompt language and framework.

The dataset did not have an equal number of observations of each class. For example, there were many more instances of *SpatialLang* (39 instances) than there were of *Sorting* (2 instances). Sorting appears to have high reliability, but this could be due to the small number of instances. Spatial language has much more variability and may be a truer measure of what is happening. As the MML was not trained on the data, there would be no bias related to the frequency of class

observation in the data. In future work, the research team intends to select additional instances of infrequently observed classes to better understand the performance of the MML.

Finally, this approach is designed to be multimodal, but our results only employ the audio transcript as we observed that the video captioning framework was not sufficiently descriptive in a classroom video, possibly due to the high clutter (e.g., background objects, posters), leading to poor performance of the video model. We believe that improving the video captioning will help improve the overall accuracy of the model. Additionally, the video recorded by cameras positioned at the side of a classroom is very different from the type of YouTube video that contains such academic content; creators are often facing the camera in a framed shot with graphics or other on-screen features used to highlight the concepts being taught. In a classroom recording, the teacher may be seen in profile and pointing to a board that is not visible, may be working at a table with physical manipulatives that are not clearly in view, or may be obstructed from view by children or furniture during informal instructional moments.

Implications for Future Use

We selected the identification of math and literacy content as a classification task because these subjects are commonly taught and easy to identify, and because it gave us the opportunity to use existing models from SRI. The ability to identify where students are receiving formal and informal instruction in these two subjects could form a first step in many approaches to support educators, assess classroom environments, and support quality curriculum at scale. The approach used in this study is meant to be complemented with use case-specific educator, researcher, curriculum developer, or administrative-facing tools.

Supporting teacher development is an evidence-based approach to improving student outcomes from classroom settings. Identifying when math or literacy content is addressed in the classroom could be useful to reduce preparation for coaching sessions, support teacher reflection, or help teachers learn from one another by identifying informal instructional moments. In addition to supporting teacher development, researchers and educators can use the output from these models to identify the amount of time, or number of instances, spent on math and literacy within a classroom, in a school, or across a larger set of instructional settings.

Curriculum developers, researchers, and administrators also have a vested interest in identifying when math and literacy instruction happen to assess new curricular rollouts, teacher implementation, and alignment between instruction and student outcomes. Identifying when math and literacy instruction happens is a first step to evaluating what is taught, how it is taught, and whether educators are using high quality materials and practices in those moments.

Together with application and audience specific use cases, the identification of when (and how often) math and literacy instruction happens can be an important first step in understanding the Pre-K classroom and providing teachers with feedback for high-quality outcomes.

References

[1] C. Christensen, A. Roy, M. Cincebeaux, and K. Ramneet, "An AI tool to detect educational YouTube content for kids, presented at the Digital Media and Developing Minds International Scientific Congress, Washington, DC, USA, July 13–16, 2025.

[2] Office of Head Start, "Head Start Early Learning Outcomes Framework: Ages Birth to Five," U.S. Department of Health and Human Services, Administration for Children and Families, 2015. Available: https://eclkc.ohs.acf.hhs.gov/school-readiness/article/head-start-early-learning-outcomes-framework

[3] C. Christensen, M. Cincebeaux, A. Roy, and S. Kim, "A machine learning model to detect early math content in YouTube videos," *Artificial Intelligence in Education*, vol. 1, no. 1, pp. 75-92, Dec. 2025, doi: 10.1108/AIIE-12-2024-0050.

[4] S. Curenton, *Conversation Compass: A Teacher's Guide to High-Quality Language Learning in Young Children*. Databrary, 2018. [Online]. Available: https://databrary.org/volume/732

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, and R. Fergus, Ed., Long Beach, CA, USA, Dec. 2017, pp. 6000–6010, doi: 10.5555/3295222.3295349.

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

[7] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," ArXiv, 2023, doi: 10.48550/arXiv.2301.12597.

Appendix A. Prekindergarten Math Codebook

The Pre-K math codebook is based on the Head Start Early Learning Outcomes Framework (HSELOF) Mathematics Development (60 Months) [2]. Specifically, the codebook focuses on the Mathematics Development Preschooler Domain within the Cognition central domain, which includes four sub-domains: counting and cardinality, operations and algebraic thinking, measurement, and geometry and spatial sense. Each code within a section corresponds to a single goal (learning standard) or a group of related standards that are likely to be taught similarly. Some standards are covered in more than one code. Codes in this codebook include the learning goals within all four sub-domains, as well as one goal from the Scientific Reasoning domain (Scientific Inquiry). An initial edition of this codebook was created to classify children's YouTube videos for TROVE (Technology to Review Online Videos for Education) [3]; this codebook has been modified for a Pre-K classroom context.

Table A1. TROVE Pre-K math codes and corresponding HSELOF goals

Code	Pre-K: Head Start Early Learning Outcomes Framework Goals
	Counting and Cardinality
Counting	**Goal P-MATH 1.** Counts verbally or signs to at least 20 by ones. **Goal P-MATH 3.** • When counting objects, says or signs the number names in order, pairing one number word that corresponds with one object, up to at least 10. • Counts and answers "How many?" questions for approximately 10 objects. • Accurately counts as many as 5 objects in a scattered configuration. • Understands that each successive number name refers to a quantity that is one larger. • Understands that the last number said represents the number of objects in a set.
Numerals	**Goal P-MATH 5.** Recognizes and, with support, writes some numerals up to 10.
Subitizing	**Goal P-MATH 2.** Instantly recognizes, without counting, small quantities of up to 5 objects and says or signs the number.
Comparing groups	**Goal P-MATH 4.** Identifies whether the number of objects in one group is more than, less than, or the same as objects in another group for up to at least five objects.
	Operations and Algebraic Thinking
Addition or subtraction	**Goal P-MATH 6.** • Represents addition and subtraction in different ways, such as with fingers, objects, and drawings. • Solves addition and subtraction word problems. Adds and subtracts up to 5 to or from a given number. • With adult assistance, begins to use counting on from the larger number for addition. For example, when adding a group of 3 and a group of 2, counts "One, two, three…" and then counts on "Four, five!" (keeping track with fingers). When counting back for subtraction such as taking away 3 from 5, counts, "Five, four, three…two!" (keeping track with fingers).

Code	Pre-K: Head Start Early Learning Outcomes Framework Goals
Patterns	**Goal P-MATH 7.** • Fills in missing elements of simple patterns. • Duplicates simple patterns in a different location than demonstrated, such as making the same alternating color pattern with blocks at a table that was demonstrated on the rug. Extends patterns, such as making an eight block tower of the same pattern that was demonstrated with four blocks. • Identifies the core unit of sequentially repeating patterns, such as color in a sequence of alternating red and blue blocks.
	Measurement and Data
Measurable attributes	**Goal P-MATH 8.** • Measures using the same unit, such as putting together snap cubes to see how tall a book is. • Compares or orders up to 5 objects based on their measurable attributes, such as height or weight. • Uses comparative language, such as shortest, heavier, or biggest.
Sorting	**Goal P-SCI-3.** Categorizes by sorting observable phenomena into groups based on attributes such as appearance, weight, function, ability, texture, odor, and sound.
	Geometry and Spatial Sense
Shape ID	**Goal P-MATH 9.** • Names and describes shapes in terms of length of sides, number of sides, and number of angles. • Correctly names basic shapes regardless of size and orientation. • Analyzes, compares and sorts two- and three-dimensional shapes and objects in different sizes. Describes their similarities, differences, and other attributes, such as size and shape. • Creates and builds shapes from components.
Spatial language	**Goal P-MATH 10.** • Understands and uses language related to directionality, order, and the position of objects, including up/down, and in front/behind. • Correctly follows directions involving their own position in space, such as "Stand up" and "Move forward."

Instructions

- Throughout, we use the word "**highlight**" to mean that the teacher draws students' attention to one specific visual, when multiple visuals are present. This is most relevant when multiple letter or numbers are displayed at once. Highlighting helps students to identify the focal letter. Examples include pointing at an object, holding an object, or telling children where to look (e.g., "Look at that last letter"). Highlighting can also include displaying only one letter or number at a time. If the camera does not show the visual, it is OK to infer highlighting from the audio.
- Throughout, if a visual is required, it is OK to infer its presence from audio if it is not on camera.
- **Check every box that applies to any part of the video**.
- **Written numerals** refer to symbols like 1, 2, 3. They do not refer to written number names, like "one."
    - Written numerals do not include Roman numerals or tally marks.
    - Written numerals do not include the minute hand on an analog clock when the numeral does not match the time (e.g., hand points to six, audio says "thirty").
    - Written numerals can be handwritten or typewritten.

    - The whole numeral must be visible onscreen.
- Verbalizations (e.g., naming shapes or numbers) can be **spoken** or **sung**.
- Who should you focus on coding as the "teacher": A "teacher" can be any adult who is wearing a microphone in the recording. "Teacher" can also refer to an actor in a video the teacher is showing, if the video can be clearly heard/seen in the recording. If there are multiple teachers on camera, focus on the teacher wearing the microphone and visible. If multiple teachers are wearing the microphone and visible, code all visible teachers with microphones on. This may mean coding separate streams of audio in the same clip.

Table A2. TROVE math codebook - modified for pre-K classroom use

Counting and Cardinality			
Indicator	**Description**	**Includes / Examples**	**Does NOT include**
Counting. Does the teacher **verbalize more than 1 number** in the standard sequence? 1. At least once 0. Never	• **Standard sequence:** Starting point can be any whole number.	• Counting parts of a shape, such as counting sides or vertices. • Counting with a few words between numbers, like "one little, two little, three little pumpkins," or "1-2 buckle my shoe, 3-4, shut the door." • Using objects or numerals to illustrate while counting verbally.	• Counting backwards, counting by any number except 1, or counting only to 1. • Counting with a story between the numbers (e.g., in 5 Little Ducks, there are several sentences between each number; 5 Little Ducks also does not count because it is counting backwards).
Numerals. Does the teacher **highlight one written numeral** AND either highlight the corresponding **quantity of objects** OR **verbalize** the number? 1. At least once 0. Never	• **Written numeral:** Teacher directs students' attention to a single written numeral. The numeral may be displayed on its own or as part of a count sequence. If more than one numeral is displayed, teacher directs students' attention to at least one specific numeral. • **Quantity of objects**: Similar objects, visually grouped together. Can include tally marks.	• The teacher may do all three (highlight a numeral + show quantity of objects + verbalize the number).	
Subitizing. Does the teacher name a **number 2-5** and show a corresponding **quantity of objects**? 1. At least once	• **Number:** Teacher must say at least one number 2 to 5. • **Quantity of objects**: Similar objects, visually grouped together or visually distinct.	• Can include tally marks • Teacher may ask how many or tell how many	• Teacher highlights a subset of objects in a larger group, but subset is not visually grouped other than the teacher highlighting. • Teacher counting in order, like "1, 2, 3."

Counting and Cardinality			
Indicator	**Description**	**Includes / Examples**	**Does NOT include**
Never			Counting can be used after subitizing to confirm quantity.
CompGroup. Does the teacher **display** and verbally **compare** two or more **groups of objects**? 1. At least once 0. Never	**Verbally compare** • Any language that compares quantities between two or more groups. • **Count** "bigger" or "smaller" numbers, **but not** "bigger" or "smaller" groups, as this risks conflating object size (e.g., a group of 3 basketballs is bigger than a group of 5 golf balls) with quantity of objects. If "bigger" or "smaller" is used it must be paired with the word "number" or a numeral. **Display objects** • Groups of objects should be visually distinguished from one another.	• **Verbally compare:** greater than, less than, equal to, same size, more, or less.	

Operations and Algebraic Thinking			
Indicator	**Description**	**Includes / Examples**	**Does NOT include**
AddSubt. The teacher **verbalizes** at least **two numbers** and the number that results when they are **added** or **subtracted** AND highlights corresponding **numerals** or **quantities of objects** to represent addition or subtraction. 1. At least once 0. Never	**Verbalization** • Teacher names at least two numbers and the number that results when they are added or subtracted. These statements should be concise and clear. • Does not need to use the word "add" or "subtract." **Quantities of objects** • Showing the addition or removal of objects, including fingers or drawings. **Numerals** • Numerals must represent subsets and the result. For example "1 + 2= 3." • Mathematical symbols (e.g., +, =) are not required.	• "Four and five makes eight" or "Ten minus three makes seven." • Adding by counting on. For example, "We have three, let's add two. Three, four, five. Three and two make five." • Decomposing sets of objects into two or more sets.	• **Does NOT include** counting from 1 (code as counting). • **Does NOT include** narratives with many words between the three numbers involved. For example, "Five Little Ducks" is technically subtraction, but the original lyrics are too wordy to make that clear for a young child. To count, the story would need to say something like "Five ducks minus one makes four ducks."

**Numerals code typically co-applies with this code.*

Measurement and Data			
Indicator	**Description**	**Includes/Examples**	**Does NOT include**
MeasAtt. Teacher **highlights object(s)** and describes them using a **measurable attribute.** *(Select all that apply.)* 1. At least once 0. Never	**Verbal description** • Teacher must name the measurable attribute, such as length, width, breadth, height, volume, mass. • Teacher does not need to state the numeric measurement (e.g., 12 pounds). Describing the attribute (e.g., heavy) is sufficient. **Showing the object** • The video clip does not need to show the act of measuring, only the object being described. • If other objects are displayed, focal object should be highlighted.	• "Two sides are long and two are short" counts: short/long are length measurements. • Can describe a class of objects rather than a specific object (e.g., "2D shapes have two measurements: length and breadth"). • Can compare multiple objects (e.g., the height of two people) using words such as shortest, heavier, or biggest.	• The phrase "four equal sides" does not count because it does not name a measurable attribute (length). The phrase "sides of the same length" would count because it includes the word length.
Sorting. Teacher **sorts** objects into categories and **names** the categories. 1. At least once 0. Never	• Teacher must demonstrate the act of sorting – that is, physically grouping objects, such as physically moving them to piles or locations. • Two ways to name categories: ○ Naming the attribute the objects are being sorted by and using that attribute to describe objects (e.g., "Let's make a pile of fat objects and flat objects. This one is flat"). ○ AND/OR mentioning that like objects are being grouped together.	• Categories include but are not limited to color, shape, object type, purpose, pattern, or species.	• Labeling groups of pre-sorted items without showing the act of sorting (e.g., "I have a bowl of red sprinkles and a bowl of green sprinkles").
Patterns. Teacher **uses the word pattern** and **describes** or **shows** a **pattern.** 1. At least once 0. Never	• **Pattern description:** The word "pattern" • **AND at least one of these:** ○ **Visual pattern:** with at least two repetitions of the pattern completed during the segment. For example, red block, blue block, red block, blue block. • **Audio description of pattern:** Naming at least two repetitions verbally (e.g., "The		• **Does NOT include** audio patterns, such as music, a series of tones, or clapping patterns.

Measurement and Data			
Indicator	**Description**	**Includes/Examples**	**Does NOT include**
	pattern on your shirt goes green white green white").		

Geometry and Spatial Sense			
Indicator	**Description**	**Includes/Examples**	**Does NOT include**
ShapeID. Teacher **highlights** and **names** a **shape**. 1. At least once 0. Never	• **Highlights:** This counts if one shape is displayed on its own, OR if the teacher draws students' attention to a particular shape among several displayed. • **Shape name:** Includes any geometric name for a shape.	• Can include naturally occurring shapes (e.g., "This pie is a circle"). • Can include drawing a shape or building it from other objects or shapes, either on their own or as part of a larger object or drawing.	• Nonstandard variations on shape names like a shape character named "Squarey." • Negating statements (e.g., "That's not a square").
SpatialLang. Teacher uses **words** and **visuals** to describe **position** or **movement**. 1. At least once 0. Never	• **Key Words:** Audio of any of these words to describe **position** or **movement**: Up, down, front, back, backward, behind, between, through, forward, right, left, in, above, below, beside, on, behind, next to, beneath, under, underneath, over, near, far, middle, bottom, top, across, side (if referring to position). • **Visuals:** The teacher shows or highlights/points to demonstrate positionality.	• Can include describing shapes, for example: A cone is wide at the bottom and narrow at the top, or a cylinder has a round middle.	• Describing an object's orientation (e.g., "Upside-down, right-side up"). • Non-specific words like "here" or "there" (e.g., "There's a circle *here*, and a square *there*"). • Using key words to describe something other than position or movement, or without illustrating position or movement. Examples: **In** the summer, sorting **in** different ways. Put it **down**. Put it **back**. Clean it **up**. Sit **down**. • Describing a spatial relationship without a visual indication (e.g., "Go look on the stool" if the teacher does not point to the stool).

Key Focus		
Indicator	**Description**	**Does NOT include**
KeyFocus. Which of these is the **key focus** of these clip? *(Select all that apply.)* *[Display each option only if it was selected above.]* • Counting • Cardinality • Numerals • Subitizing • Comparing groups • Addition or subtraction • Measurable attributes • Sorting • Spatial language • Shape identification • Patterns • None of the above	• Key focus is the main content or task the teacher is focused on. Examples include teaching an academic subject, transitioning, or supporting social interaction. If math is not the key focus, select "none of the above."	• If the teacher is focused on something else – for example, teaching another academic subject, or transitioning – and uses incidental math content (e.g., I'm going to count you off to go to the bathroom), do NOT code math content as the key focus.

Learning Connections		
Indicator	**Description**	**Does NOT include**

[Display only if any items above apply.]
LearningConnections. Does the teacher encourage students to relate learning content **to their own lives**?
1. At least once
0. Never
	• The teacher asks or tells children to think about how the literacy content they are learning relates to something outside the lesson, such as:
	  ○ Something else they have done at school (e.g., "Do you remember when we sorted buttons?")
	  ○ Something they may have done at home or with their family ("Let's count how many people are in your family!")
	  ○ A familiar object (e.g., "Can you think of anything that's shaped like a triangle?")	

Suggested Citation: Christensen, C., Gerard, S., Cincebeaux, M., and Perez, N. (2025). *TROVE math codebook - modified for pre-K classroom use*. SRI.

Appendix B. Prekindergarten Literacy Codebook

The Pre-K literacy codebook is based on the Head Start Early Learning Outcomes Framework (HSELOF) Literacy domain (60 Months) [2]. Specifically, the codebook is focused on the Literacy preschooler domain within the Language and Literacy central domain. This domain includes four sub-domains: phonological awareness, print and alphabet knowledge, comprehension and text structure, and writing. Codes in this codebook include the phonological awareness and print and alphabet knowledge sub-domains. The comprehension and text structure and writing sub-domains are not included; future iterations of the model could include them. Each code within a section corresponds to a single HSELOF goal or group of related goals likely to be taught similarly. Some goals are covered in more than one code. An initial edition of this codebook was created to classify children's YouTube videos for APPROVE (Assisting Parents to Review Online Videos for Education; [1]); this codebook has been modified for a Pre-K classroom context.

Table B1. APPROVE Pre-K literacy codes and corresponding HSELOF goals

Code	Head Start Early Learning Outcomes Framework Goals
	Phonological Awareness
Rhyming	**Goals P-LIT 1.** Provides one or more words that rhyme with a single given target, such as "What rhymes with log?"
	Print and Alphabet Knowledge + Phonological Awareness
Letters	**Goal P-LIT 1.** Provides a word that fits with a group of words sharing an initial sound, with adult support, such as "Sock, Sara, and song all start with the /s/ sound. What else starts with the /s/ sound?" **Goals P-LIT 1.** Produces the beginning sound in a spoken word, such as "Dog begins with /d/." **Goal P-LIT 2.** Understands that written words are made up of a group of individual letters. **Goal P-LIT 3.** Names 18 upper- and 15 lower-case letters. **Goal P-LIT 3.** Knows the sounds associated with several letters.
Words	**Goal P-LIT 2.** *(48 to 60 months)* Shows a growing awareness that print is a system that has rules and conventions, such as holding a book correctly or following a book left to right. **Goal P-LIT 2.** Begins to point to single-syllable words while reading simple, memorized texts.

Instructions

- Throughout, we use the word "**highlight**" to mean that the teacher draws students' attention to one specific visual, when multiple visuals are present. This is most relevant when multiple letters or numbers are displayed at once. Highlighting helps students to identify the focal letter. Examples include pointing at an object, holding an object, or telling children where to look (e.g., "Look at that last letter"). Highlighting can also include displaying only one letter or number at a time. If the camera does not show the visual, it is OK to infer highlighting from the audio.
- On the Qualtrics scoring form, check every option that applies to any part of the video.
- Who should you focus on coding as the "teacher": A "teacher" can be any adult who is wearing a microphone in the recording. "Teacher" can also refer to an actor in a video the

teacher is showing, if the video can be clearly heard/seen in the video recording. If there are multiple teachers on camera, focus on the teacher who is wearing the microphone and visible. If multiple teachers are wearing the microphone and visible, code all visible teachers with microphones on. This may mean coding separate streams of audio in the same clip.

- Verbalizations (e.g., naming letters) can be **spoken** or **sung**.
- Where written words are referred to, word must be written (picture alone does not suffice).

Table B2. APPROVE literacy codebook – modified for pre-K classroom use

Rhyming			
Code	**Description**	**Includes/Examples**	**Does NOT include**
Rhyme. Teacher says the word "**rhyme**" or "rhyming" AND says at least 2 **rhyming words**. 1. At least once 0. Never	• Teacher should say rhyming words within about 60 seconds of the word "rhyme" or "rhyming." Teacher could say rhyme first and rhyming words second, or vice versa.	• Rhyming words do not need to be spoken one after the other (e.g., "cheese, please"); they could have words between them, such as a poem or song (e.g., cat jumped over the hat).	
Letters			
LetterName. Teacher **highlights** and **names** a letter 1. At least once 0. Never	• **Letter:** Any upper- or lowercase letter. • Does not need to say whether the letter is upper- or lowercase.	• Letter may be displayed on its own, as part of the alphabet, as part of a list of letters, or as part of a word.	• Phonetic letter sounds that are not letter names. • Letter names used as words (e.g., A in "A cat").
LetterSound. Teacher verbalizes the **sound** of a letter and highlights EITHER the letter, or objects that illustrate the letter sound. 1. At least once 0. Never	• Can include letter names if they are also letter sounds (e.g., long forms of vowels).	• Sounds of letters on their own. • Sounds of letters within words. The teacher should highlight the individual letter within the word. • Sound may occur anywhere within a word, including the beginning sound. • Example: /a/ /a/ apple.	• Using the letter name instead of the sound the letter makes within the word. For example, naming the letter T as part of tree. Code as letter names.

Words			
Code	**Description**	**Includes/Examples**	**Does NOT include**
FollowWords. The teacher displays more than one word at once and **highlights each word as it is read.** 1. At least once 0. Never	• Must show a passage containing multiple words. Only one word displayed at a time would not count. • Words must be highlighted left to right, top to bottom, and/or page to page. Highlighting can include one word in a passage appearing at a time. • It's OK if more than one word is highlighted at a time, so long as highlighting moves left to right or top to bottom. For example, it's OK to highlight a few words at a time moving left to right, or a sentence at a time moving top to bottom.	• It's OK if words aren't highlighted exactly as they are spoken (e.g., highlighting words at a constant pace that doesn't totally line up with audio), so long as the highlighting generally moves left to right or top to bottom as the words are spoken. • If the teacher shows a video in the classroom, this code includes sing-along style videos that highlight words as they are sung.	• Showing a video with captions or lyrics that don't highlight one word at a time.
OneSyllable. Teacher **highlights and reads** a **one-syllable** word. 1. At least once 0. Never	• Teacher points to or otherwise highlights a word as they read it; the word is one syllable.	• Examples: cat, bed, run, she • Includes pointing to words in a passage as they are read, so long as they are one syllable.	

Key Focus		
Indicator	**Description**	**Does NOT include**
KeyFocus. Which of these is the **key focus** of these clip? *(Select all that apply.)* *[Display each option only if it was selected above.]* • Letters • Rhyming • Words • None of the above	• Key focus is the main content or task the teacher is focused on. Examples include teaching an academic subject, transitioning, or supporting social interaction. If literacy is not the key focus, select "none of the above."	• If the teacher is focused on something else – for example, teaching another academic subject, or transitioning – and uses incidental literacy content (e.g., sit on the letter that starts your name for circle time), do NOT code literacy content as the key focus.

Learning Connections		
Indicator	**Description**	**Does NOT include**

[Display only if any items above apply.]
LearningConnections. Does the teacher encourage students to relate learning content to their own lives?
1. At least once
0. Never
	• The teacher asks or tells children to think about how the literacy content they are learning relates to something outside the lesson, such as:
	  ○ Something else they have done at school (e.g., "We just read a book that rhymed! What was it?")
	  ○ Something they may have done at home or with their family ("Do you know anyone whose name has the 'en' sound in it?")
	• A familiar object (e.g., "Can you think of anything in this room that starts with B?")
		• Using familiar objects to illustrate letter names/sounds, UNLESS the teacher talks about how those objects may be familiar to children (e.g., "A is for apple" does not count unless a physical or personal connection to apples is obvious).

Suggested Citation: Christensen, C., Gerard, S., Cincebeaux, M., and Perez, N. (2025). *APPROVE literacy codebook – modified for pre-K classroom use* SRI.

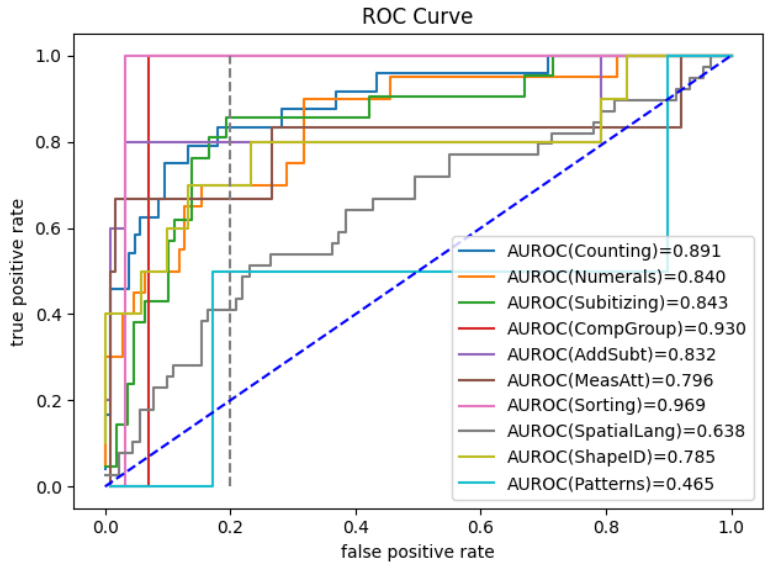
Appendix C. Receiver Operating Characteristic (ROC) Curves

For each class, the model outputs a value in [0, 1] that indicates the model's confidence for whether the instructional content is present or not. This is shown in the receiver operating characteristic (ROC) curves in Figures C1 and C2, in which we highlight the threshold based on a false positive rate (FPR) of 0.2 (or 20%) used for the error matrices and analysis in the white paper. A ROC curve is a graphical plot that illustrates the performance of a binary classifier model at varying threshold values. The ROC curve is the plot of the true positive rate (TPR) against the FPR at each threshold setting. The threshold value itself is not shown in the graph. During inference, a threshold is chosen to make the binary choice. For lower thresholds, both the TPR and FPR are high (i.e., the recall is high) corresponding to the points on the top right. As the threshold increases, both the TPR (recall) and FPR decrease.

Math

In this ROC curve (Figure C1), we note that the TPRs for most of the codes are in 0.74±0.8 and 0.8 for an FPR of 0.2, which indicates good performance. The gray dashed line intersects with the ROC curves for each code at FPR = 0.2. The area under the ROC (AUROC) is a measure of how well the model performs across different thresholds. A random number generator would have an AUROC of 0.5, shown as the blue dashed line. A higher value for AUROC indicates better performance. As shown in the ROC plot, *Subitizing* and *Counting* have the highest AUROCs and *SpatialLang* the lowest (for codes that have more than two instances).

Figure C1. Math ROC curve



Literacy

In this ROC curve (Figure C2), we note that the TPRs for most of the codes are in 0.32±0.10 for an FPR of 0.2, which indicates moderate performance. The gray dashed line intersects with the ROC curves for each code at FPR = 0.2. For codes that have more than once instance, *Rhyming* has the highest AUROC, while *LetterSound* and *OneSyllable* have the lowest.

Figure C2. Literacy ROC curve