AI-Powered Workflows for Understanding and Evaluating Instructional Discourse

Julian Bernado[†*1], Kirk Vanacore[†2], Ana Ribeiro[1], René F. Kizilcec[2], and Susanna Loeb[1]

[1]SCALE Initiative, Stanford University
[2]National Tutoring Observatory, Cornell University

[†]Workshop Instructors

April 1, 2026

1 Description of the Session

Artificial intelligence (AI) is changing the evaluation landscape, creating new possibilities for how researchers and evaluators can understand not only which instructional programs are effective, but also which underlying components of instruction drive impact. This workshop explores how AI-powered workflows can support rigorous, scalable analysis of conversational educational data while preserving human oversight, methodological validity, and public trust. As tutoring and other dialogue-rich learning environments generate increasingly large datasets, researchers have new opportunities to study instructional processes at scale. At the same time, using AI for annotation raises important questions about bias, transparency, privacy, and accountability. Bringing together complementary approaches from the SCALE Initiative and the National Tutoring Observatory, this session shows how AI can be used not simply to automate analysis, but to strengthen construct development, improve reproducibility, and support more trustworthy research workflows.

The session combines demonstration, hands-on engagement, and discussion. Participants will be introduced to workflows for rapidly developing coding schemas, refining annotations, and extending human judgment across larger corpora of conversational data. The workshop will also highlight privacy-preserving de-identification, AI-assisted annotation, and applications to tutoring interactions linked to student outcomes. Designed for education researchers and learning scientists, the session emphasizes responsible use of AI in the production of evidence and offers practical strategies for studying teaching and tutoring at scale.

[*]Corresponding author: jbernado@stanford.edu

The workshop will be broken into two major sections led by the workshop organizers from SCALE and NTO respectively. We request **four** hours for the workshop, with the approximate breakdown described in Table 1.

Section	Duration
Beginning remarks	0:00 - 0:20
Rapid schema development demo	0:20 - 1:50
Break	1:50 - 2:00
Annotation and Evaluation at Scale	2:00 - 3:30
Closing remarks	3:30 - 4:00

Table 1: Workshop agenda.

1.1 Rapid Schema Development Demo

The first portion of the workshop consists of a look into the SCALE Initiative's current work on text-based tutoring measures, then a hands-on demonstration of a tool enabling rapid development of qualitative coding schemas. Virtual tutoring generates a trove of rich interaction data between instructors and learners. However, understanding which tutoring behaviors drive student outcomes requires measures of all behaviors of interest across large volumes of text data. Given the wide range of tutoring behaviors potentially linked to student outcomes, many measures of tutoring must be developed to understand which pedagogical practices most affect student outcomes. After hearing about SCALE's efforts developing and validating such measures, participants will try their own hand at developing novel measures of tutor and student behaviors using a purpose-built AI-powered tool.

The tool, shown in Figure 1, is a web-app interface for iteratively refining a schema. Its structure is based on the following human-time-intensive iterative codebook development process used by the SCALE team:

1. Come up with an initial coding schema for a behavior of interest
2. Conduct a round of annotation between multiple human annotators
3. Compare annotations and, if agreement is insufficient, make revisions to the coding schema
4. Repeat until sufficient agreement is reached

The schema development tool allows researchers to focus on the definitional aspects of the above process by replacing the human annotation round with an LLM annotation round. Then, by validating a limited number of LLM annotations, the user is able to understand the degree to which the current schema encodes their behavior of interest. Since LLM annotations happen in seconds, users may iterate their schema and immediately see its impact on the resulting annotations. In addition to annotation, the schema development tool accelerates construct definition by allowing the user to generate AI-suggested additions to the current schema that are grounded in the set of human-validated utterances.

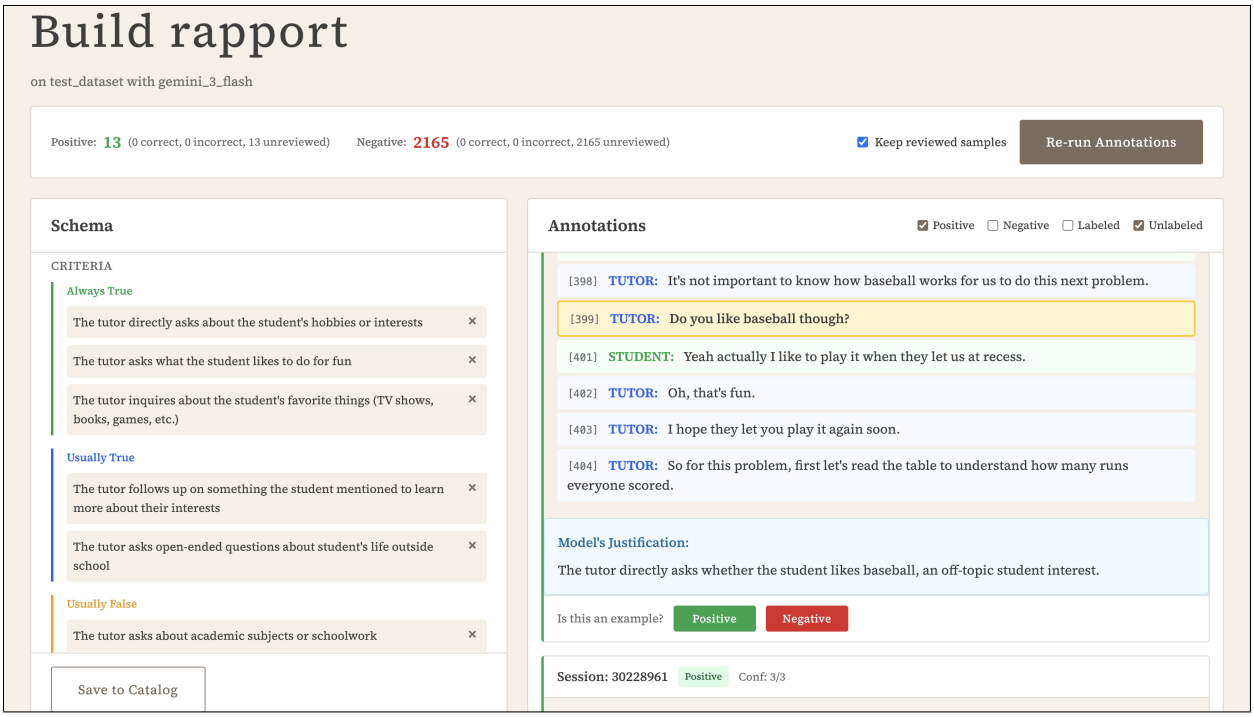

Figure 1: The schema development tool interface being used to develop a schema for a building rapport measure. The interface mainly features the current state of the schema (left), and a list of annotations derived from the schema (right), allowing the user to immediately see within the data which behaviors are encoded by a given version of a schema. Text is synthetic for display purposes only.

In the demo of the tool, workshop participants will be divided into several groups. Each group will discuss a tutor or student behavior of interest for which they'd like to derive a measure. After each group picks a measure to schematize, each participant will use the schema development tool on open-access tutoring transcripts to independently develop their own schemas for their chosen construct. Workshop instructors will circulate and help troubleshoot while participants develop their schemas. Halfway through schema development, we will do a whole-workshop check-in.

After finishing, groups will compare results to see whether schemas of their measures differed. We will then lead a whole-group discussion aiming to understand how the tool can be further developed to support rapid scoping and validation of conversational constructs of interest. Participants will leave this portion of the workshop with a readily usable classification pipeline for their chosen measure.

1.2 Annotation and Evaluation at Scale

The National Tutoring Observatory (NTO) portion of the workshop will focus on how researchers can use AI tools to study institutional dialogue interactions at scale while maintaining standards of rigor, privacy, and human oversight. The NTO is building shared research infrastructure to advance the science of tutoring and teaching, including large-scale repos-

itories of tutoring interactions linked to student outcomes and open workflows for secure de-identification, annotation, and analysis. One such infrastructure is *Sandpiper*, an AI-annotation tool designed to support researchers in developing transparent, theory-informed, and empirically grounded analyses of conversational data (Hedley et al. 2026). This open-source tool allows for transparent, human-in-the-loop annotation at scale.

This portion of the session will begin with a brief overview of the NTO's goals, data infrastructure, and research agenda. We will introduce participants to how tutoring data from multiple providers and contexts can be organized into interoperable, privacy-conscious resources that support both descriptive and causal questions about instructional effectiveness. We will then demonstrate the NTO's annotation workflow, including tools for de-identifying conversational data, applying and refining coding schemas, and using AI-assisted pipelines to extend human coding to larger corpora. Throughout, we will highlight how these workflows preserve a central role for researcher judgment through codebook design, disagreement review, validation, and iterative refinement.

Participants will then engage in a hands-on activity using either their own data or open-access tutoring transcripts, provided by the NTO. In small groups, they will explore how to define constructs of interest, test AI-assisted annotation approaches, and examine resulting outputs for validity, usefulness, and potential bias. The goal is not only to expose participants to new technical capacities but also to create space for reflection on the conditions under which such methods should be trusted in education research. The NTO portion will conclude with a discussion of challenges and future directions, including responsible data sharing, reproducibility, interoperability across tutoring datasets, and the role of human oversight in AI-supported research workflows.

2 Topic Significance and Support of Theme

Large language models (LLMs) have made it feasible to extend human annotations to a scale that would have been too costly and time-intensive to code by hand. In tutoring and other dialogue-rich educational settings, this creates an opportunity to study conversation-level measures of pedagogical behavior and their relationships to various student outcomes. While prompt-based workflows have become a popular method for analyzing conversational data (Wang and Chen 2025; Tran, Litman, and Godley 2025; Shin 2025), there remains a lack of domain-specific tooling for applying common LLM-based workflows. This workshop presents two complementary tools for developing and annotating constructs of interest. The workshop also includes discussions around rigor and validity of LLM-based annotations.

The workshop directly supports the 2026 SREE conference theme, *Education and Public Trust: Evidence and Accountability in a Changing Landscape*. It is especially well aligned with the theme area on data science, artificial intelligence, and emerging risks, while also supporting the theme area methods, transparency, and credible evidence. Across both the SCALE and NTO portions of the session, participants are introduced to AI-supported workflows that keep humans in the loop and center researcher judgment. The workshop outlines responsible workflows and guidelines for using AI in conversational analysis in light of concerns around rigor, privacy, and bias.

This workshop is particularly valuable for SREE because it demonstrates novel tooling

for AI-assisted research workflows. Participants will familiarize themselves with concrete workflows that can be readily applied to their own research. Researchers, evaluators, and tutoring providers may benefit from guidance and tooling for trustworthy AI-assisted analyses of conversational data. In doing so, we expand the methodological tool set available to researchers studying instructional discourse. To ensure researchers have a sober understanding of how to responsibly use these new AI tools, we incorporate guidance on proper usage and documentation of AI pitfalls at every step of the workshop. We aim to empower education researchers with AI-powered tools for conversational analysis designed to address concerns about validity, privacy, and bias. In doing so, the workshop advances discussion of how seemingly black-box, probabilistic AI models can be used in ways that make the resulting research *more* transparent, reviewable, and reproducible.

3 Target Audience

This workshop is intended for education researchers, evaluators, and providers who work with tutoring, classroom, or other dialogue-based educational data. It will be especially relevant for researchers interested in instructional processes and mechanisms, construct development, and annotation validity. The workshop is also designed to be useful for tutoring providers and other organizations that want to analyze their own sessions at scale.

The workshop does not require expertise or prior familiarity with AI or NLP-based annotation approaches. It is designed to be accessible to attendees who are new to AI-assisted annotation as well as to more experienced researchers. Participants who work with transcript-like data are likely to find the session especially useful as they will gain familiarity with various tools and analyses that can be readily applied to their own data. Attendees will leave with experience rapidly developing a coding schema, reviewing and validating AI-assisted annotations along with an understanding of the role these workflows play in education research. The session is best suited for participants seeking transparent, rigorous methods to understand educational conversational data.

References

Hedley, Daryl et al. (2026). *Sandpiper: Orchestrated AI-Annotation for Educational Discourse at Scale*. arXiv: 2603.08406 [cs.HC]. URL: https://arxiv.org/abs/2603.08406.

Shin, Eunhye (2025). "Co-Coding Classroom Dialogue: A Single Researcher Case Study of ChatGPT-Assisted Analysis in Science Education". In: *Journal of Computer Assisted Learning* 41.4. e70089 JCAL-24-1773.R2, e70089. DOI: https://doi.org/10.1111/jcal.70089. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcal.70089. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jcal.70089.

Tran, Nhat, Diane Litman, and Amanda Godley (Oct. 2025). "Using Large Language Models to Analyze Students' Collaborative Argumentation in Classroom Discussions". In: *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*. Ed. by Joshua Wilson, Christopher Ormerod, and Magdalen Beiting Parrish. Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United

States: National Council on Measurement in Education (NCME), pp. 111–125. ISBN: 979-8-218-84228-4. URL: https://aclanthology.org/2025.aimecon-main.13/.

Wang, Deliang and Gaowei Chen (2025). "Evaluating the use of BERT and Llama to analyse classroom dialogue for teachers' learning of dialogic pedagogy". In: *British Journal of Educational Technology* 56.6, pp. 2671–2704. DOI: https://doi.org/10.1111/bjet.13604. eprint: https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13604. URL: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13604.