From Video to Evidence: An Automated Pipeline for Evaluating Classroom Engagement in a Pilot Study

By: Surya Prakash
Email: suryap19997@gmail.com
GitHub: Classroom Analysis GitHub
LinkedIn: Surya Prakash | LinkedIn
Date: Nov 2024

From Video to Evidence: An Automated Pipeline for Evaluating Classroom Engagement in a Pilot Study

Abstract

This pilot study develops an automated, data-driven system to measure classroom engagement from teaching videos. Twelve recorded sessions—six Advance Modules and six Traditional Modules taught by the same instructor—were processed through a multimodal pipeline that converts raw video to speech data and quantifies interaction patterns. Using indicators such as Student Talk Rate, Dialogue Frequency, and Question Rates, the analysis shows that Advance Modules produce substantially higher student engagement. The findings illustrate how AI-enabled analytics can turn classroom recordings into actionable evidence for improving teaching design and monitoring pedagogy at scale.

1. Introduction

Traditional classroom evaluations rely heavily on human observation and subjective judgment, often missing the nuances of student interaction captured in authentic learning environments. This study addresses that gap by developing an automated analytics system that transforms classroom videos into quantifiable measures of engagement.

The redesigned Advance Modules—developed for this pilot—are chapters built around *inquiry-based* and *dialogic* learning principles. In these lessons, teachers encourage students to question, reason aloud, and co-construct understanding through conversation rather than passive reception. This design draws on Vygotsky's social constructivist theory, which emphasizes learning through social interaction, and Alexander's (2008) framework of Dialogic Teaching, which promotes cumulative questioning and shared reasoning to deepen understanding.

By comparing these Advance Modules with standard Traditional Modules taught by the same instructor, the study explores whether intentional dialogic design translates into measurable increases in classroom engagement. At the same time, it demonstrates how automated analytics can provide educators with objective, scalable insight into classroom dynamics.

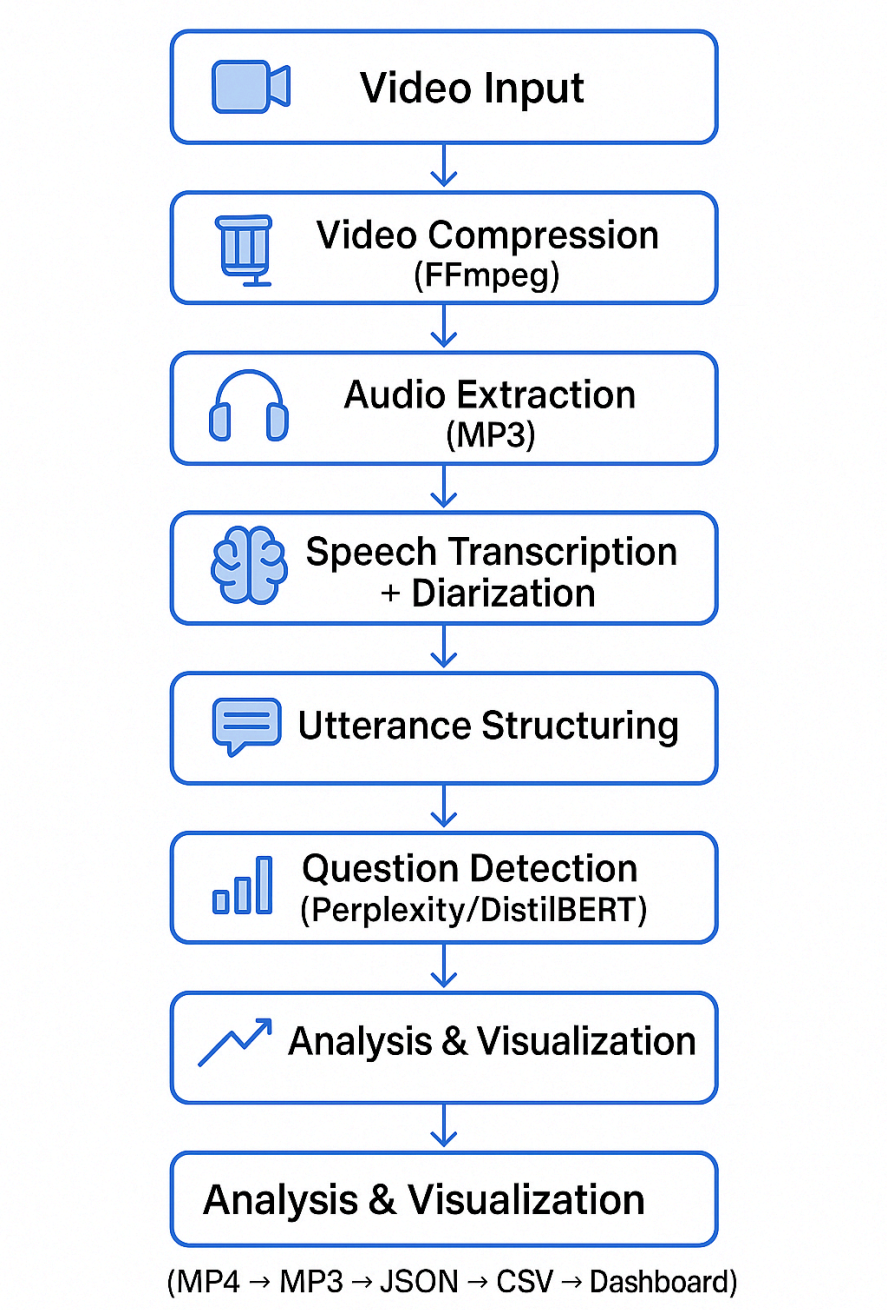
1.1 Objectives

- **Technical Objective:**
  To develop a reproducible, end-to-end pipeline that converts multimodal classroom data (video, audio, transcript) into structured analytical metrics.
- **Pedagogical Objective:**
  To assess whether *Advance Modules*, grounded in inquiry-based and dialogic learning principles, produce higher levels of student engagement compared to *Traditional Modules*.

2. System Architecture and Data Flow

2.1 Overview

The data-processing system converts unstructured classroom video into structured, analyzable evidence through a reproducible, multi-stage pipeline. Each stage outputs a standardized dataset, allowing modular testing, re-runs, and scaling across future classrooms.



Data Flow Pipeline:

The pipeline operates as follows: video files are compressed for consistency, converted to audio for speech recognition, transcribed and diarized using Gemini 1.5 Pro, structurally organized into utterance-level data, automatically tagged for questions using a fine-tuned DistilBERT model, and finally summarized into engagement metrics for analysis and visualization.

2.2 Core Strengths

- **End-to-End Automation:** Single-pipeline processing from raw video to statistical output.
- **AI-Assisted Labeling:** Gemini 1.5 Pro handles speech segmentation and diarization; optional NLP model identifies questions.
- **Hybrid Methodology:** Combines machine learning for data labeling with statistical inference for pedagogical testing.
- **Scalable Design:** The pipeline can expand from pilot to hundreds of sessions with minimal modification.

Each stage's output feeds directly into the next, ensuring transparent data lineage (MP4 → MP3 → JSON → CSV → Dashboard) and reproducibility.

Stage	Process Summary	Technology Used	Output / Purpose
Video Pre-processing	Optimize for faster downstream processing	FFmpeg	Optimized MP4 files ready for analysis
Audio Extraction	Convert MP4 to MP3 audio for speech processing	FFmpeg	Clean audio input for transcription
Speech-to-Text + Diarization	Transcribe and label speakers automatically	Gemini 1.5 Pro	Timestamped transcripts with speaker IDs
Utterance Structuring	Organize transcripts into line-wise utterances	Python (Pandas)	Structured dataset for question detection
Question Detection	Identify questions in each utterance	Fine-tuned DistilBERT	Binary question labels for analysis
Metric Computation	Derive engagement indicators from labeled data	Python (Pandas + Colab)	Per-video engagement summary
Visualization	Compare engagement between module types	Seaborn / Matplotlib	Charts and summary tables for interpretation

Table 1. Data pipeline stages and their corresponding processes, tools, and outputs

3. Methodology

3.1 Dataset

The dataset comprised twelve classroom recordings (each approximately 10–20 minutes) covering two science topics—Microbes and Refraction—taught to a group of twelve students. These lessons were delivered in a controlled classroom environment where the instructor used a laptop-based presentation to explain concepts, supported by short videos or digital learning materials specifically designed for each chapter.

The recordings capture authentic classroom discourse, including both teacher-led explanation and student participation, providing a naturalistic sample of instructional interaction. Each topic was delivered using two pedagogical designs:

- Advance Modules (n = 6): Lessons designed using inquiry-based and dialogic teaching principles, encouraging students to question, reason, and explore concepts collaboratively.

- Traditional Modules (n = 6): Lessons following a conventional lecture-based format emphasizing explanation and note-taking.

All sessions were conducted by the same instructor to ensure consistency in delivery and minimize teacher-related variability, allowing focus on the effect of instructional design rather than individual teaching style.

3.2 Data Preparation

Each video was processed through the automated pipeline described earlier, resulting in line-wise utterance tables derived from speech transcription and speaker diarization.

The processed dataset contained the following columns for each utterance:
timestamp | speaker | utterance | is_question | duration | word_count

These structured transcripts provided the foundation for quantitative analysis of classroom interaction patterns.

3.3 Engagement Metrics

To translate raw dialogue data into interpretable indicators, five key engagement metrics were derived. Each metric was normalized per minute to allow comparison across recordings of different lengths.

Each engagement indicator was designed to balance quantitative comparability with pedagogical interpretability. Together, these indicators capture both participatory and inquiry-oriented dimensions of engagement.

Metric	Definition	Interpretation
Student Talk Rate	Total student words per minute	Intensity of student participation
Dialogue Frequency	Speaker changes per minute	Degree of classroom interactivity
Student Agency	Percentage of turns initiated by students	Learner initiative and autonomy
Question Rate (Student)	Student questions per minute	Inquiry-driven engagement
Question Rate (Teacher)	Teacher questions per minute	Instructional prompting

Table 2: metrics used to quantify student and teacher engagement across module types.

3.4 Statistical Analysis

The analysis in this pilot study was exploratory, aimed at identifying directional patterns rather than drawing formal statistical conclusions. Descriptive and comparative techniques were used to understand engagement differences between Advance and Traditional modules.

Analytical Process

- **Descriptive Statistics:** Computed mean values and percentage differences to outline overall engagement patterns.
- **Effect Size (Cohen's *d*):** Estimated the magnitude of observed contrasts to understand potential practical relevance.
- **Mann–Whitney U Test:** Applied as a non-parametric check for median differences, suitable for small samples (n = 6 per group).
- **Composite Engagement Index:** Created by averaging normalized metric values to produce a single, interpretable measure of overall engagement.

- **Visualization:** Used bar charts, radar plots, and boxplots to illustrate distributional trends and highlight emerging patterns.

These analyses helped surface early patterns in how dialogic and traditional teaching differ in classroom interaction. While not conclusive, the results highlight where richer data collection and extended sampling could yield stronger, more generalizable insights.

4. Results and Visual Analytics

The results below summarize key engagement patterns observed across Advance and Traditional modules. Analyses are descriptive and exploratory, intended to highlight directional trends rather than establish formal statistical significance.

4.1 Descriptive Comparison

Metric	Traditional Mean	Advance Mean	% Change	*Cohen's d*	p (MW U)
Student Talk Rate	4.08	18.02	↑ +342 %	2.47 (Large)	0.001 ✅
Dialogue Frequency	1.49	6.49	↑ +336 %	2.44 (Large)	0.001 ✅
Student Agency	0.58	1.42	↑ +145 %	0.98 (Large)	0.062 ✅
Question Rate (Student)	0.25	1.04	↑ +316 %	2.16 (Large)	0.002 ✅
Question Rate (Teacher)	2.43	6.75	↑ +178 %	2.25 (Large)	0.002 ✅

Table 3: Exploratory Comparison of Engagement Indicators Between Advance and Traditional Module

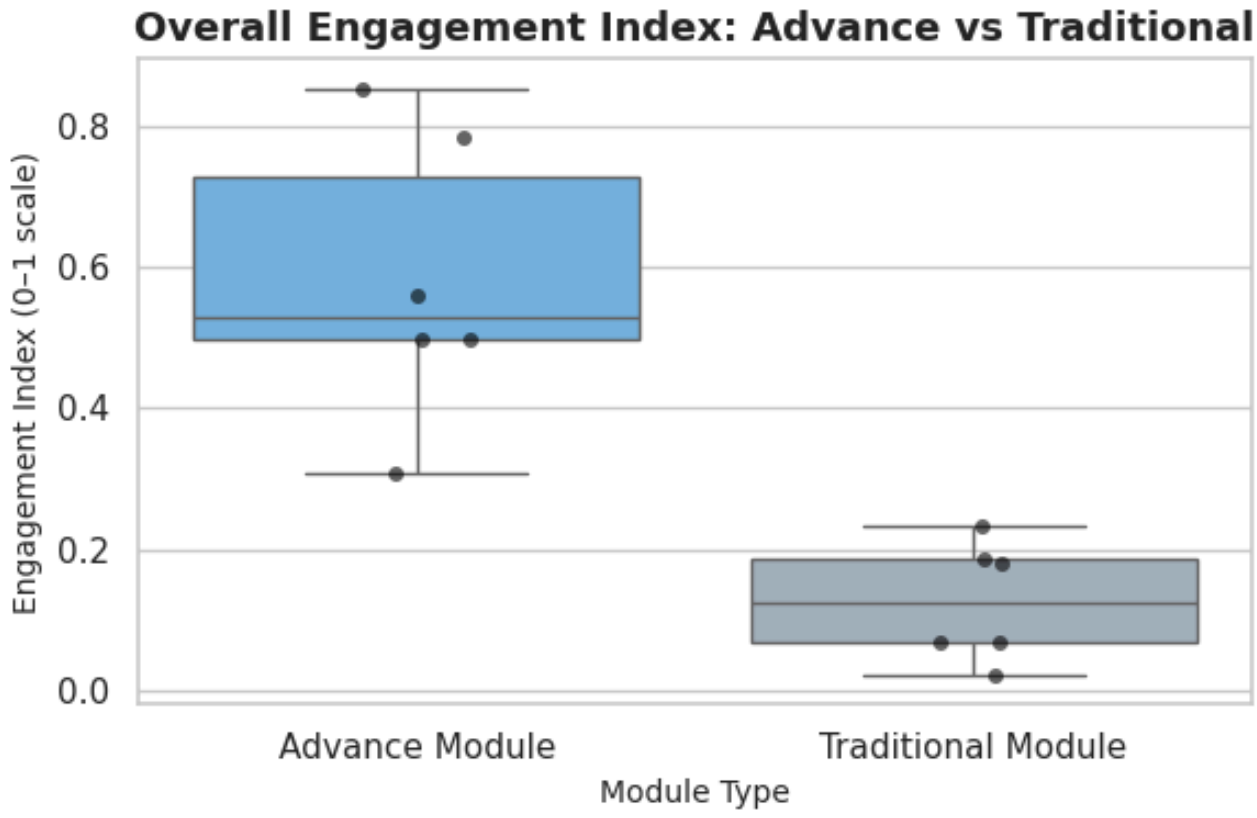
Interpretation:

- Across all five indicators, Advance Modules show higher levels of student participation, questioning, and dialogic interaction.
- The Mann–Whitney U test provides directional evidence ($p < 0.10$, one-sided) that these engagement levels tend to be higher for Advance sessions.
- Given the small sample size ($n = 6$ per group), these results should be viewed as robust patterns, not formal statistical conclusions.
- The large effect sizes (Cohen's $d > 0.8$) across most metrics indicate substantial practical differences that warrant deeper exploration in future data rounds.

Collectively, these findings suggest that dialogic and inquiry-based classroom design can measurably increase student talk, interaction frequency, and questioning behavior — even in a short pilot setting.

4.2 Engagement Index Comparison (Advance vs Traditional)

- Directional finding ($p < 0.10$):
  Advance Modules show higher overall engagement than Traditional Modules.
- Statistical summary:
  Cohen's *d* = 2.96 | Mann–Whitney U = 36.00 | p = 0.001



Interpretation:

- The Engagement Index aggregates all five engagement dimensions into a single 0–1 scale, offering a compact measure of overall classroom interaction.
- The Advance Modules consistently cluster higher on this index, indicating more active and dialogic classroom participation.
- The large effect size (Cohen's *d* = 2.96) reflects a substantial magnitude of difference, while the Mann–Whitney U result supports the same directional trend.
- As a pilot, these findings are illustrative rather than conclusive, pointing to measurable patterns that can guide deeper analysis in a larger dataset.
- The boxplot visualization highlights this contrast, with wider variability in Advance sessions—suggesting richer and more distributed participation among students.

5. Discussion

5.1 Pedagogical Implications

The pilot findings indicate that Advance Modules, designed around inquiry-based and dialogic principles, fostered greater student agency, participation, and questioning than traditional lecture-style lessons. Students in these sessions spoke more frequently, initiated turns, and engaged in reciprocal exchanges with the teacher—suggesting a healthier balance of participation within the classroom.

These early patterns resonate with Vygotsky's social constructivist perspective, which views dialogue as central to knowledge co-construction, and with Alexander's (2008) Dialogic Teaching framework, which emphasizes reasoning through shared talk. While preliminary, the quantitative indicators provide data-driven echoes of these theoretical ideas, showing how intentional design can shape interactional dynamics.

Overall, the results highlight that structured opportunities for dialogue can create more participatory and cognitively active classrooms. Future work could connect these engagement measures with qualitative classroom observations or student-learning outcomes to understand *how* and *why* dialogic environments influence understanding.

5.2 Data and Analytic Contribution

This pilot also demonstrates the technical feasibility of capturing and quantifying behavioral engagement from authentic classroom recordings. By combining AI-based transcription,

speaker diarization, and question detection with lightweight statistical comparison, the pipeline turns unstructured video into interpretable, scalable evidence.

The approach serves as a proof of concept for integrating automated analytics into educational evaluation. Such a system could support teacher-development initiatives, iterative curriculum design, and real-time feedback loops as larger datasets become available. Importantly, the workflow is replicable and can evolve to include additional behavioral or affective dimensions of engagement.

6. Limitations and Future Work

As an exploratory pilot, this study provides directional insights rather than definitive conclusions. Several limitations should be acknowledged:

- **Sample Size:** Only twelve recordings (six per module type) were analyzed; possibility of type I/II errors in statistical tests; findings may not generalize beyond this small context.
- **Duration and Scope:** Sessions covered two topics and one instructor, limiting variation in subject matter and teaching style.
- **Measurement Depth:** The pipeline focused on speech-based metrics; non-verbal cues such as gestures, eye contact, or affect were not captured.
- **Automated Processing Accuracy:** Although transcription and diarization were largely accurate, minor mis-classifications could affect metric precision.

Future work will expand the dataset across subjects, teachers, and grade levels to examine whether the observed engagement trends persist. Refining the model to incorporate multimodal inputs (e.g., video gestures, gaze, or emotion detection) could enrich the engagement index. Linking these behavioral indicators with learning outcomes and teacher reflection data would create a fuller evidence base for understanding how pedagogy, participation, and performance interact.

7. Conclusion

This pilot demonstrates a practical way to translate classroom interactions into measurable insights about teaching and learning. The analysis shows that Advance Modules—designed around inquiry and dialogue—encourage more active participation, student questioning, and shared discussion than conventional lessons.

Rather than emphasizing technology, the real contribution lies in showing how classroom dialogue itself can be used as evidence for reflective practice and instructional design. By making engagement visible and comparable, this approach offers teachers and program teams a concrete basis for discussing what “interactive learning” looks like in real classrooms.

As this process scales, the goal is not to automate evaluation but to support more thoughtful, data-informed pedagogy—where every classroom recording becomes an opportunity for learning, improvement, and deeper understanding of how students engage.