AI-Based Classroom Observation Evaluation: A Multimodal Research Pipeline

Matt Krasnow
Swiftscore | *Harvard*
matt@swiftscore.org

June 18, 2025

Abstract

This document outlines a comprehensive research pipeline for developing and evaluating AI-based classroom observation systems. The study investigates multimodal versus single-modal approaches using TEACH framework data, implementing novel diversified chain of thought methodologies, and exploring synthetic data augmentation techniques. The research aims to determine the viability of AI systems for consistent classroom evaluation compared to human evaluators.

Contents

1 Introduction

The evaluation of classroom teaching effectiveness remains a critical challenge in educational research and practice. Traditional human-based evaluation systems, while thorough, are resource-intensive and subject to inter-rater variability. This research investigates the potential for artificial intelligence systems to provide consistent, reliable classroom observation evaluations using multimodal inputs including audio recordings and transcriptions.

The primary research questions addressed in this study are:

1. Can AI consistently evaluate classroom observations compared to human evaluators?
2. How do multimodal approaches compare to single-modal inputs in evaluation accuracy?
3. Does fine-tuning on domain-specific data improve evaluation performance?
4. Can diversified chain of thought methods enhance evaluation quality?
5. Do synthetic data augmentation techniques boost model performance?

2 Methodology

2.1 Phase 1: Data Preparation and Exploration

2.1.1 Exploratory Data Analysis (EDA)

The initial phase involves comprehensive examination of existing datasets to understand:

- Data structure and format characteristics
- Quality metrics and completeness assessment
- Pattern identification and anomaly detection
- Documentation of limitations and constraints

2.1.2 Data Cleaning and Dataset Construction

Two distinct datasets will be constructed corresponding to TEACH1 and TEACH2 frameworks. Each dataset will contain the following standardized column structure:

1. **Identifier**: Unique ID for each classroom evaluation/observation
2. **Audio File 1**: First audio recording file or link
3. **Audio File 2**: Second audio recording file or link
4. **Transcription 1**: Automated transcription of first audio
5. **Transcription 2**: Automated transcription of second audio
6. **Evaluation Columns**: All TEACH framework evaluation metrics and scores
7. **Metadata**: Language, country of origin, and contextual information

2.1.3 Audio Transcription Generation

An automated transcription pipeline will be implemented using Whisper or alternative state-of-the-art models. This process includes:

- Processing all audio recordings for text generation
- Quality validation and accuracy assessment
- Organized storage in accessible formats

2.2 Phase 2: Data Splitting and Visualization

2.2.1 Train-Test Split

Given the independence of evaluation observations, a random splitting approach will be employed:

- Optimal split percentage determination based on dataset size
- Balanced representation across evaluation categories
- Methodology documentation and rationale

2.2.2 Feature Extraction and Data Visualization

Comprehensive visualization suite creation includes:

- Data quality and quantity visualizations
- Statistical summaries of evaluation scores
- Distribution pattern analysis across metrics
- Organized codebase folder structure for retrieval