AI-Based Classroom Observation Evaluation: A Multimodal Research Pipeline

Matt Krasnow
Swiftscore | Harvard
matt@swiftscore.org

May 26, 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 **3**

2 Methodology **3**
2.1 Phase 1: Data Preparation and Exploration 3
2.1.1 Exploratory Data Analysis (EDA) 3
2.1.2 Data Cleaning and Dataset Construction 3
2.1.3 Audio Transcription Generation 4
2.2 Phase 2: Data Splitting and Visualization 4
2.2.1 Train-Test Split 4
2.2.2 Feature Extraction and Data Visualization 4
2.3 Phase 3: Model Development and Training 4
2.3.1 Base Model Pipeline Creation 4
2.3.2 Model Fine-Tuning 5
2.3.3 Novel Diversified Chain of Thought Method 5
2.4 Phase 4: Advanced Techniques 5
2.4.1 Synthetic Data Generation 5
2.5 Phase 5: Model Evaluation and Analysis 6
2.5.1 Comprehensive Evaluation Suite 6
2.5.2 Performance Metrics Documentation 6

3 Expected Results 6

4 Timeline and Deliverables 7

4.1 Phase 6: Research Documentation 7

4.1.1 Final Report Preparation 7

4.1.2 Deliverables Organization 7

5 Success Criteria 7

6 Conclusion 8

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

2.3 Phase 3: Model Development and Training

2.3.1 Base Model Pipeline Creation

Three distinct modeling approaches will be developed:

Audio-Only Model Processes raw audio recordings directly for evaluation prediction.

Transcript-Only Model Analyzes text transcriptions using natural language processing techniques.

Multimodal Model Combines both audio and transcript inputs through fusion architectures.

2.3.2 Model Fine-Tuning

Each base model will undergo domain-specific fine-tuning:

- Training dataset adaptation
- Cross-validation hyperparameter optimization
- Performance comparison across modality combinations
- Configuration documentation

2.3.3 Novel Diversified Chain of Thought Method

A multi-agent discussion system will be implemented featuring specialized AI agents:

Rhetorician Agent Analyzes communication effectiveness and clarity

Principal Agent Evaluates from administrative and leadership perspective

Teacher Agent Provides pedagogical assessment and instructional quality

Student Agent Offers learner experience and engagement viewpoint

The system architecture includes:

1. Agent group chat discussion mechanism
2. Final evaluation synthesis AI processing all agent inputs
3. Various system prompt and model configuration testing
4. Fine-tuning application to ensemble approach

2.4 Phase 4: Advanced Techniques

2.4.1 Synthetic Data Generation

When time permits, synthetic data augmentation will be explored:

- Random sampling of real human transcriptions
- AI-generated classroom evaluation transcript creation
- Synthetic evaluation generation using fine-tuned models
- Expanded training dataset creation combining real and synthetic data
- Performance improvement assessment on augmented datasets

2.5 Phase 5: Model Evaluation and Analysis

2.5.1 Comprehensive Evaluation Suite

Multiple evaluation metrics will be employed:

Inter-Rater Reliability Cohen's kappa and related statistical measures for consistency assessment.

Accuracy Assessment Alignment measurement with human evaluation standards.

Error Analysis Identification of specific areas where AI performance degrades.

Score Proximity Analysis Determination of AI score ranges relative to human evaluations.

Comparative Analysis Performance evaluation across:

- Single modal vs. multimodal approaches
- Base models vs. fine-tuned models
- Traditional models vs. diversified chain of thought
- Real data only vs. synthetic data augmentation

2.5.2 Performance Metrics Documentation

Systematic documentation includes:

- Comprehensive evaluation result compilation
- Comparison tables and performance visualizations
- Statistical significance analysis
- Best-performing configuration identification

3 Expected Results

The research anticipates demonstrating the viability of AI-based classroom evaluation systems with particular emphasis on:

1. High inter-rater reliability between AI and human evaluators
2. Consistent performance across TEACH1 and TEACH2 frameworks
3. Clear identification of optimal modeling approaches
4. Actionable insights for practical implementation

4 Timeline and Deliverables

4.1 Phase 6: Research Documentation

4.1.1 Final Report Preparation

A comprehensive academic document will be prepared including:

- Standard research paper formatting in LaTeX
- Complete methodology documentation
- Results compilation with statistical analysis
- Discussion of implications and limitations
- Future research direction recommendations

4.1.2 Deliverables Organization

Final deliverables include (but are not limited to):

- Complete documented codebase
- Reproducible result frameworks
- Supplementary materials and appendices
- Literature review

5 Success Criteria

The research success will be measured by:

1. Achievement of statistically significant inter-rater reliability
2. Demonstration of multimodal approach superiority
3. Successful implementation of novel chain of thought methodology
4. Clear performance improvement through fine-tuning
5. Practical applicability of findings to educational settings

6 Conclusion

This research pipeline represents a comprehensive approach to investigating AI-based classroom evaluation systems. Through systematic exploration of multimodal approaches, novel methodological innovations, and rigorous evaluation protocols, the study aims to advance the field of educational technology and provide practical solutions for classroom observation challenges.

The integration of traditional machine learning approaches with innovative chain of thought methodologies, combined with potential synthetic data augmentation, positions this research at the forefront of educational AI applications.