AI Framework for Scalable Automated Continuous Formative Assessment

Arjun Rajasekar
RCTS, IIIT
Hyderabad, India
arjun.rajasekar@research.iiit.ac.in

Sakshi Mallenahalli
RCTS, IIIT
Hyderabad, India
m.sakkshi@research.iiit.ac.in

Inzela Mirza
RCTS, IIIT
Hyderabad, India
inzela27@gmail.com

Praveen Kumar Palaboyina
RCTS, IIIT
Hyderabad, India
praveen.palaboyina
@research.iiit.ac.in

Sai Kumar Pola
RCTS, IIIT
Hyderabad, India
saikumar.pola@research.iiit.ac.in

Syed Falahuddin Quadri
RCTS, IIIT
Hyderabad, India
sd.falahuddin@research.iiit.ac.in

Aravind Gondi
RCTS, IIIT
Hyderabad, India
aravind.gondi
@students.iiit.ac.in

Ramesh Loganathan
RCTS, IIIT
Hyderabad, India
ramesh.loganathan@iiit.ac.in

***Abstract*—There has been a great push towards evidence based learning across the globe. However, the level of progress made towards this goal has been greatly varied. One of the main bottle necks to this progress has been the increased administrative requirements for implementation of evidence based learning. Continuous formative assessments (CFA) a key metric for implementation of evidence based learning, is a work load intensive process in its structure which has led to poor or incorrect implementations of continuous formative assessments in many schools. We present a framework that uses natural language processing and computer vision based tools to perform automated continuous formative assessment in classrooms.**

The framework analyses the video and conversation streams that occur during a classroom session, to assess the engagement between teachers and students and generate insights into the student learning and behaviour. The framework builds upon existing models for automatic speech recognition, speaker diarization, facial recognition, body pose detection, and large language models, to promise a scalable automated assessment tool capable of providing standardised continuous formative assessments. The study covers both an online or digital classroom scenario as well as an offline or physical classroom scenario. We present the framework as it currently stands and present key improvements to be made before the framework is viable for field use.

***Index Terms*—Large Language Models, Evidence-Based Learning, Continuous Formative Assessment**

I. Introduction

A major tenet of evidence-based learning (EBL) is the emphasis on continuous improvement and the adaptation of teaching methods based on student learning outcomes. Continuous Formative Assessments (CFA) have evolved as a critical mechanism to facilitate the required monitoring for successful EBL implementation. Unlike traditional summative assessments, CFA integrates ongoing monitoring and frequent evaluation of student learning into the daily instructional process, supported by an immediate feedback loop that allows for timely adjustments to teaching strategies and learning activities. Additionally, CFA encourages student participation, focusing on self-reflection rather than grades.

However, the implementation of CFA in many educational systems has often been nominal, with some institutions merely rebranding monthly summative tests as CFA without meaningful changes to instructional practices. Such superficial implementations not only fail to support the shift towards EBL but also undermine the core objectives of EBL for teachers, students, and parents.

Several factors contribute to these inadequate implementations, including the significant administrative burden on teachers and ineffective teacher retraining due to logistical and financial constraints. Technology is seen as a potential solution to these challenges. Learning Management Systems (LMS), such as Moodle, Canvas, and Google Classroom, help streamline administrative processes, administer varied assessments, solicit student feedback, and generate detailed analytics and reports. Student Response Systems (SRS) like Kahoot, Socrative, and Plickers facilitate real-time formative assessments and enhance student engagement in class [1].

Despite these technological advancements, current solutions primarily capture student learning data through specific questions, overlooking the broader spectrum of student-teacher interactions during teaching sessions. These interactions provide valuable insights into student understanding and are crucial for effective CFA. Manually recording these interactions is impractical and disruptive. However, the advent of large language models (LLMs) offers a promising solution by analyzing classroom conversations to extract meaningful insights, provided that classroom recordings and appropriate digital infrastructure are in place.

The COVID-19 pandemic, impacting over 1.6 billion students globally, prompted a significant shift towards digital classrooms, accelerating the adoption of LMS and other educational technologies. This shift, despite its challenges, has led to a lasting investment in digital classrooms, enabling hybrid education models that offer flexibility and resilience.

The widespread availability of LMS and digital platforms

now provides an opportunity to introduce automation tools into education at scale, addressing administrative burdens and freeing up teachers to focus more on teaching.

In the following sections, we present a framework that leverages LLMs to analyze teacher-student conversations during teaching sessions to enhance CFA.

II. Methodology

LLMs are advanced AI systems trained on vast natural language datasets to understand, analyze, and generate human language. In recent years, they have shown explosive potential in tasks such as summarization, sentiment analysis, and conversational understanding. These models leverage deep learning techniques to identify patterns and contextual meanings within text, enabling them to extract nuanced insights from complex language data at speeds and scales exceeding human capability. LLMs can analyze teacher-student interactions, capturing subtle cues and feedback often missed in traditional assessments. By processing and interpreting these interactions, LLMs provide a more comprehensive and dynamic understanding of student learning, facilitating more effective and responsive teaching strategies.

In the current study, we examine two scenarios: one in a fully online digital classroom and the other in a more traditional physical classroom. This approach allows us to divide the system architecture (see Fig. 1 and 2) into two main components: conversation extraction and conversation assessment. We'll begin by discussing the conversation extraction process in both online and physical classroom settings before delving deeper into the conversation assessment aspect, which remains consistent regardless of the teaching session type.

A. Conversation Extraction

The goal here is to produce clear transcripts of the conversations ongoing during a teaching session.

1) Online Classrooms Scenario: An online classroom is a teaching session conducted over the internet, typically using video conferencing applications such as Zoom, Microsoft Teams, Google Meet, or Webex. A key feature of these platforms is their capability to capture participant-level audio and video streams for the entire classroom session, enabling a detailed reconstruction of the conversation transcript, see Fig. 1.

2) Physical Classroom Scenario: In traditional physical classrooms, obtaining clear, participant-level labeled audio and video streams is challenging. Typically, only a single combined audio-video stream is accessible, making the reconstruction of conversation transcripts more complex. For our study, we'll narrow the scope of our proposed framework by making several assumptions: first, participants (both students and teachers) do not speak simultaneously, second, all participants' speech is clear and audible in the recording, and finally third, the student speaker performs a certain pose indicating that they are the ones speaking, such as standing up or raising their hand.

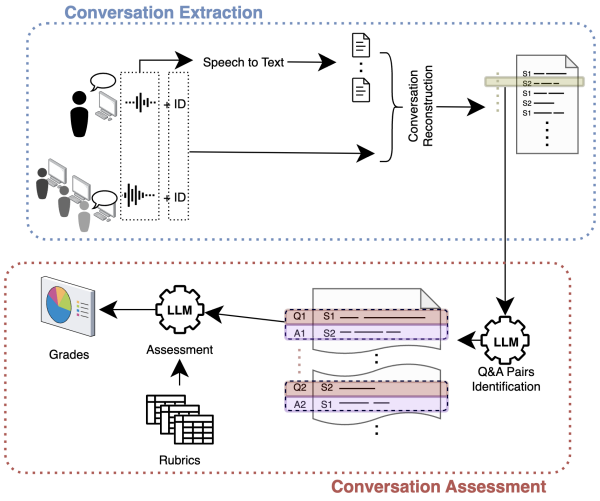

Fig. 1. Online assessment framework.

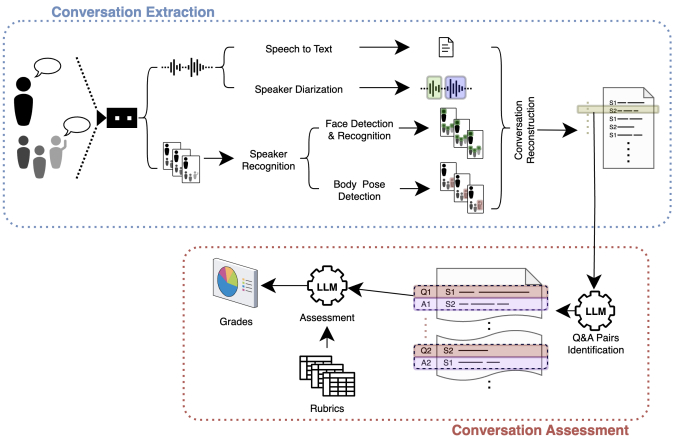

Fig. 2. Offline assessment framework.

The conversation extraction consists of two main additions compared to the online scenario (see Fig. 2), speaker diarization as well as speaker recognition. The speaker diarization is process of labelling sections of the audio spoken by the same speakers [2] and speaker recognition is to utilize the video stream (facial recognition [3] and body pose), to identify who is the speaker. By combining the above information, we can reconstruct the conversation transcript for the offline classroom session. The accuracy of the speaker diarization process and consequently the conversation reconstruction can also be improved using LLMs to process and clean the corresponding outputs [4].

B. Conversation Assessment

Once the conversation transcripts have been reconstructed, they can be analyzed using LLMs to extract insights into the student learning during the classroom session (see Fig. 1 and 2), such as:

- Evaluation of questions and answers (Q&A) within the conversation,

- Student behavior during the session.

The evaluation of Q&As within the conversation can act as a highly informative addition to the students CFA, providing a good indicator for their engagement and learning during the session. The evaluation consists of two main logical steps - first, the identification of Q&A pairs, and second, the assessment of the identified Q&As based on a set rubric.

The assessment of student behavior is based on the interpreting the tone of the ongoing conversation and participation. The behavioral information extracted from the conversation extraction can be used to improve upon the attention assessment made from video based on the student body postures [5].

Beyond CFA and behavior assessment, the conversation transcripts can also be further analyzed by the LLMs to output:

- Topics covered,
- Alignment with lesson plans,
- Suggested reading materials for students,
- Recommended content for the next teaching session,

Such reports would help teacher easily document classes, as well as quickly generate informative follow-up instruction to their students, improving the overall learning process.

III. Results and Upcoming Work

The models used during the current work were off the shelf commercial or open-source solutions. The audio processing models utilized, namely the automatic speech recognition (ASR) and the speaker diarization services, were offered as part of the Speech-to-Text API from Google Cloud Platform (GCP) [6]. The service reports high accuracy for well-formed speech in supported languages, achieving word error rates (WER) below 5% under ideal low noise conditions. However, given the context of the input data to the current system, further explorations are needed to handle the classroom noises and recording quality limitations that the field conditions impose on the system.

The developed framework was tested on synthetic conversations as well as sample conversations recorded from classroom sessions, to fine tune the prompts used by the LLMs to perform the conversation assessment. Below is a sample classroom conversation for a Physics lecture on "Understanding Newton's First Law of Motion", used to test the framework's functionality

Sample Classroom Conversation Transcript

Teacher: Good morning, class! Today, we're going to talk about Newton's First Law of Motion. Who can tell me what that law is?

Emily: Is it the one about objects staying in motion?

Teacher: That's right, Emily! Newton's First Law states that an object at rest stays at rest, and an object in motion stays in motion, unless acted upon by an external force. Can anyone give me an example of that in real life?

David: A soccer ball won't move until someone kicks it?

Teacher: Exactly! The ball is at rest, and it stays that way until a force, like your kick, makes it move. What happens after you kick it, Sarah?

Sarah: Um, it keeps rolling?

Teacher: Yes, it does. It keeps rolling because of inertia, which is an object's resistance to changes in motion. But does the ball roll forever?

Michael: No, it stops after a while.

Teacher: Right, and why does it stop, David?

David: Because of friction?

Teacher: Perfect! Friction is the force that acts opposite to the motion and eventually stops the ball. Now, let's try something simple. Everyone, take a pencil and place it on your desk.

Teacher: Notice that your pencil isn't moving. Why not, Emily?

Emily: Because nothing is pushing it?

Teacher: Correct! It's at rest, and it'll stay that way unless we apply a force. Now, gently push the pencil. What happens, Sarah?

Sarah: It moves!

Teacher: That's because you applied a force. Once you stop pushing, the pencil stops too. That's friction at work again, slowing it down. Now, what if there were no friction? What would happen to the pencil, Michael?

Michael: I guess it would keep going?

Teacher: Exactly! In space, where there's no friction, an object can keep moving indefinitely. That's why spacecraft can travel huge distances. Does anyone have any questions so far?

David: What about when a car stops? Is that also friction?

Teacher: Great question! Yes, the car's brakes apply friction to the wheels, which slows it down. But there's also air resistance—another type of friction—working against the car as it moves. Anything moving through air experiences that.

Sarah: So, without friction, cars wouldn't stop?

Teacher: That's right. Friction is crucial for things like driving, walking, and even writing. Without it, we'd be sliding everywhere! Now, who can tell me what Newton's First Law is also called? Michael?

Michael: Um... the Law of Inertia?

Teacher: Spot on! It's called the Law of Inertia. It describes how objects resist changes in their state of motion. Alright, let's quickly recap: what keeps an object at rest, Emily?

Emily: No forces acting on it?

Teacher: Exactly. And what causes an object to stop moving, Sarah?

Sarah: Friction!

Teacher: Excellent! Well, that's it for today's lesson. Newton's First Law is something you'll notice in everyday life now, so keep an eye out. Class dismissed!

The assessment of the conversation using the current framework and prompt configurations, generates the following analysis reports:

Alignment with Lesson Plans:

Topic 1: Understand Newton's First Law of Motion
Score: 2
Summary: The teacher explained Newton's First Law as 'an object at rest stays at rest, and an object in motion stays in motion, unless acted upon by an external force.'

Topic 2: Explain the concept of inertia.
Score: 2
Summary: The teacher defined inertia as 'an object's resistance to changes in its state of motion' and provided examples such as a soccer ball rolling due to inertia.

Topic 3: Identify real-life examples of inertia and forces such as friction.
Score: 2
Summary: The teacher provided several real-life examples, including a soccer ball being kicked, a pencil moving when pushed, and cars stopping due to friction.

Topic 4: Recognize the importance of friction in everyday life.
Score: 2
Summary: The teacher highlighted the crucial role of friction in various daily activities, such as driving, walking, and writing, and explained how it affects motion.

Summary of General Behavioral Engagement:

The students in this physics class demonstrate varying levels of engagement throughout the lesson. Emily and David are actively participating, answering questions correctly, and showing enthusiasm for the topic. Sarah is also contributing to the discussion, but seems a bit uncertain at times. Michael is quiet initially but starts to participate more towards the end.**Overall Student Attentiveness:**I would rate the overall student attentiveness as 4 out of 5. The students are generally engaged and focused on the lesson, with only occasional moments of distraction or uncertainty.

Individual Student Summaries:

Emily
Score: 4
Review: "Emily is a strong participant in the class discussion, answering questions correctly and showing enthusiasm for the topic. She's well-prepared and engaged throughout the lesson."

David
Score: 4
Review: "David is also actively participating, providing correct answers and asking thoughtful questions. He seems confident in his understanding of the material."

Sarah
Score: 3
Review: "Sarah is contributing to the discussion, but sometimes seems uncertain or hesitant. She could benefit from more confidence in her responses."

Michael
Score: 2
Review: "Michael starts off quiet, but becomes more engaged towards the end of the lesson. He still has room for improvement in terms of participation and confidence in his answers."

The scores above are subjective continuous formative assessment ratings generated based on the student's engagement and participation during the lesson, based on predefined rubrics.

However, testing revealed variability in the quality of analysis generated by the LLM for the different conversations on repeated runs. However, in order to make the framework be deployable in the field, further fine-tuning would be needed, both in terms of the reproducibility of results and the consistency of result quality for different conversation inputs.

The following activities are planned or currently underway to address these issues:

A. Custom Dataset Creation

There are 1000s of hours of recordings of lectures sessions available online, however to the best of our knowledge, there are no open standard datasets that also record the student engagement during these lecture sessions. To address this we are exploring the collection of audio/video for lecture sessions with focus on the student-teacher conversations as well as student behavior. The dataset would allow us better evaluate on-the field performance in the lab. The presence of a standardized dataset would also allow for us compare the performance of different LLM for the intended application scenario.

B. Framework Optimization

We performed our testing using the base Lama 3 [7] as well as ChatGPT 3.5 [8] models, with the contextual information regarding the lecture session such as the topic, syllabus, lecture plans, etc. provided alongside the prompt. Further explorations are needed for the training of custom LLM models for this specific assessment and analysis task, to allow for more efficient resource handling for computation for easy scaling and possible on-the edge application.

C. Multi-language Conversations

Not all teaching sessions are in one language. Especially in countries where English is not the primary language, it is common practice that the teacher uses a local language in addition to English to better explain to and engage with the students. As such, the framework must be improved to accommodate multi-language conversation.

Certain LLMs have demonstrated the ability to handle multi-language inputs. However, a significant issue remains regarding subject domains, as most language models are trained on general language usage and often lack the technical and specialized terminology frequently used in teaching scenarios.

D. Data Privacy and Other Ethical Considerations

We recognize that the proposed framework's requirement for continuous audio and visual monitoring of students and teachers during classroom activities raises significant privacy concerns. To ensure ethical compliance, any data capture must be preceded by informed consent from all parties involved, including teachers, students, and parents. This consent process should involve clear communication about the purpose of the monitoring, the types of data being collected, how this data will be used, and the measures in place for data security and privacy protection. Transparency is key to fostering trust among all stakeholders.

To safeguard personal information, the overall data handling systems will be designed to meet local regulatory requirements, such as the General Data Protection Regulation (GDPR) in the European Union. These regulations stipulate stringent guidelines for data processing, including data minimization, storage limitation, and the right to access and rectify personal data. Our commitment to adhering to these regulations will help mitigate potential privacy violations and promote responsible data stewardship.

Moreover, our team acknowledges the ethical and psychological implications of implementing constant monitoring systems in educational environments. Continuous observation can create an atmosphere of surveillance that may induce undue mental stress for both students and teachers. In an environment intended for free thinking, creativity, and open dialogue, such monitoring might inadvertently lead to behavioral changes, where individuals alter their actions to conform to perceived expectations or fear of judgment.

To address these concerns, we propose to conduct further exploratory research in the field. This research will focus on understanding the potential negative effects of constant monitoring on the psychological well-being of both students and teachers. We aim to assess how such systems can be designed and implemented in a way that supports learning rather than hinders it. This includes investigating alternative monitoring strategies that respect privacy while still allowing for the collection of valuable educational data.

Additionally, we will explore ways to tailor the system to enhance its positive contributions to the learning journey. For example, we could consider implementing features that allow for selective monitoring based on specific educational goals or time frames, rather than continuous observation. Providing users with agency over their data—such as options to opt in or out of monitoring during certain activities—could further alleviate concerns and enhance the system's acceptance.

By prioritizing ethical considerations and the psychological impact of monitoring, we strive to create a balanced approach that leverages technology to improve educational outcomes while safeguarding the well-being and privacy of all participants involved.

IV. Conclusions

We have presented a framework for performing automated continuous formative assessment in both offline and online classrooms. We see the integration of LLMs into CFA to be a significant step towards scalable evidence-based learning practices.

References

[1] V. V. Mshayisa, "Students' perceptions of plickers and crossword puzzles in undergraduate studies," *Journal of Food Science Education*, vol. 19, no. 2, pp. 49–58, 2020.

[2] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, p. 101317, 2022.

[3] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[4] Q. Wang, Y. Huang, G. Zhao, E. Clark, W. Xia, and H. Liao, "Diarizationlm: Speaker diarization post-processing with large language models," *arXiv preprint arXiv:2401.03506*, 2024.

[5] P. Pendyala, S. I. Reddi, A. Rajasekar, S. F. Quadri, N. Jaisankar, and R. Loganathan, "Framework for automated attendance & attention tracking to address learning gaps due to pandemic," in *2022 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*. IEEE, 2022, pp. 645–650.
[6] G. Cloud, "Cloud speech-to-text," https://cloud.google.com/speech-to-text/docs/reference/rest, n.d., accessed: 2024-10-14.
[7] L. Team, "Lama: Language models of modern ai," 2021, version 3. [Online]. Available: https://github.com/openlm-research/lama
[8] OpenAI, "Chatgpt: Chat generative pre-trained transformer 3.5," 2023, model version 3.5. [Online]. Available: https://www.openai.com/research/chatgpt