Education and Information Technologies (2025) 30:24649–24693
https://doi.org/10.1007/s10639-025-13755-7

# Towards reducing teacher burden in Performance-Based assessments using aivaluate: an emotionally intelligent LLM-Augmented pedagogical AI conversational agent

**Habeeb Yusuf**[1] · **Arthur Money**[1] · **Damon Daylamani-Zad**[1]

Received: 29 January 2025 / Accepted: 13 August 2025 / Published online: 3 September 2025


## Abstract

**Background** Performance-based assessments (PBAs), such as viva voce exams and oral presentations, offer comprehensive evaluations of student knowledge and skills but place substantial burdens on teachers. The integration of emotionally intelligent, LLM-augmented AI conversational agents presents a potential solution to alleviate teacher burden while maintaining the integrity and effectiveness of PBAs. This study investigates the use of AIvaluate, a pedagogical AI conversational agent designed to support teachers during oral PBAs by offering emotionally intelligent insights and streamlining the assessment process. A counterbalanced mixed-methods study design was employed with 35 teachers and students participating in both traditional face-to-face and AIvaluate-supported assessments. Data was collected through teacher-assigned grades, System Usability Scale (SUS) questionnaires, and qualitative open-response surveys. Quantitative and qualitative analyses were conducted to compare grading outcomes, system usability, and teacher preferences between the two assessment formats. Teachers issued significantly higher grades to students in AIvaluate-supported assessments ($p = 0.033$), attributed to more structured, consistent questioning and emotional state reporting. The overall SUS score for AIvaluate indicated "acceptable" usability, surpassing the face-to-face format. Thematic analysis revealed key strengths of AIvaluate, including automated question prompts, real-time emotional insights, and the convenience of remote operation. However, teachers noted limitations, such as occasional technical issues and the lack of a personal connection compared to traditional face-to-face interactions. AIvaluate demonstrates the potential to reduce teacher burden in PBAs while maintaining usability and assessment quality. Its emotionally intelligent features and automated functionalities enhance the assessment process, offering a scalable, technology-driven solution for modern education. While AIvaluate shows promise in reducing teacher burden during PBAs, technical limitations, emotional disconnection, and variability in assessment impact emphasise the need for further investigation before large-scale adoption.

Extended author information available on the last page of the article

Future research should explore building further functionality to address the diverse needs of teachers, while focusing on addressing technical limitations and assessing long-term impacts on teacher satisfaction and student outcomes.

**Keywords** Artificial intelligence · Conversational agent · Chatbot · Assessment · Education · Performance based assessment · Teacher burden · Assessor burden · Workload · Generative AI · LLM

## 1 Background

### 1.1 Performance-based assessments in education

Performance-based assessment (PBA) is a key pedagogical method employed by teachers to evaluate students' learning and academic progress. These assessments include activities such as oral presentations, debates, exhibitions, portfolios, performances, viva voce exams, technical demonstrations, and any other tasks that require students to apply their knowledge or skills in real time to produce a response, product, or outcome (Archbald et al., 1988; Darling-Hammond, 1994). PBAs are considered an effective means of assessing the extent to which students have retained and understood subject knowledge, offering insights into whether meaningful learning has occurred and to what extent. According to Baker (1997), PBAs prioritise the evaluation of cognitive abilities that synthesise and demonstrate comprehension across multiple disciplines, as opposed to traditional exams, which tend to focus on rote memorisation of factual information. Incorporating PBAs into a programme of study can foster students' interdisciplinary understanding, enabling them to identify connections between subjects and apply diverse knowledge to solve complex problems. Consequently, PBAs are increasingly utilised not only in arts and humanities disciplines but also in science, technology, engineering, and mathematics (STEM) subjects, where they are recognised as a more reliable measure of students' ability to apply what they have learnt (Ernst, 2008; Potter et al., 2017). Owing to their pedagogical benefits, many education systems in Europe, Asia, Australia, and Latin America are adapting their curricula to embed PBAs at various levels of education, ranging from primary to tertiary (Gallardo, 2020). However, despite their advantages over traditional summative assessments, the adoption of PBAs presents certain challenges, notably an increase in the workload for teachers.

#### 1.1.1 PBAs in the age of generative AI

Generative AI has revolutionised the educational landscape, offering unprecedented advancements in the generation of sophisticated text and multimedia outputs (Eke, 2023; Tzirides et al., 2023) While this technology has the potential to transform teaching and learning practices, it also places additional demands on teachers, particularly in ensuring the integrity and authenticity of assessments. Traditional coursework submissions now require much more scrutiny than before, as generative AI applications, and large language models, can produce high-quality, human-like out-

puts that challenge teachers' ability to verify the originality of student work (Alser & Waisberg, 2023; Cotton et al., 2023; Dwivedi et al., 2023; Eke, 2023). PBAs offer a promising solution to alleviate these burdens, as they inherently focus on authenticity and reduce the risks posed by AI-assisted plagiarism (Cotton et al., 2023; Pearce & Chiavaroli, 2023; Souppez et al., 2023). By requiring students to demonstrate their knowledge and skills through practical tasks, oral examinations, or real-time problem-solving, PBAs provide teachers with a robust framework for evaluating genuine student contributions. This reduces the need for extensive plagiarism checks or detailed authorship verification, which can be both time-consuming and labour-intensive (Kabbar & Barmada, 2024; Newell, 2023). However, while PBAs address challenges associated with generative AI, they can introduce new pressures for teachers, such as the increased time and effort required to design, administer, and evaluate these assessments. As educators adapt to the demands of ensuring academic integrity in a rapidly evolving technological environment, PBAs serve as a valuable tool in providing a safeguard against AI misuse, however they also simultaneously add to the complexity of teachers' workloads. Considering the dual challenges of ensuring assessment authenticity and managing teacher workload, PBAs are increasingly seen as productive methods for AI interventions. However, the complexity of these tasks extends beyond logistics. Teachers must also navigate students' emotional states and digital systems, emphasising the need to draw from broader theoretical models such as affective computing and emotional intelligence to ensure supportive, emotionally aware assessment environments.

### 1.1.2 Affective computing and emotional intelligence

Affective computing is a subfield of artificial intelligence concerned with the design of systems that can detect, interpret, and simulate human emotions (Picard, 2000). In the context of PBAs, affective computing enables AI systems to respond to student stress and emotional cues, potentially improving assessment quality and support. Emotionally intelligent systems, that integrate models of emotional awareness and regulation, can adapt their behaviour based on learners' affective states (D'mello & Graesser, 2013). Emotional intelligence itself, defined as the capacity to perceive, understand, and manage emotions (Mayer et al., 2008), plays a key role in educational contexts. Teachers often rely on emotional cues to scaffold learning, build rapport, and manage classroom dynamics. Integrating these principles into AI conversational agents supports not only cognitive engagement but also affective support, which is particularly valuable in high-stakes settings such as PBAs (Petrovica & Ekenel, 2016; Zhang et al., 2016).

### 1.1.3 Emotion theories

Emotion theories offer foundational insights into how students experience and express emotions during assessments. Appraisal theories (Lazarus, 1991) suggest that learners' emotional responses are tied to how they evaluate assessment demands and their capacity to respond. The Control Value Theory (Pekrun, 2006) emphasises how emotions, like anxiety and enjoyment, impact academic performance, particularly

in assessments. These models inform the development of AI systems that detect and respond to emotions, helping to personalise assessment conditions and reduce anxiety (Calvo & D'Mello, 2010). The inclusion of these theories enhances the capacity of AI agents to deliver responsive, human-centred assessment interactions.

### 1.2 Teacher burden in performance-based assessments

The implementation of PBAs is not without its challenges, and one of the most significant issues relates to the increased burden placed on teachers. A study by Szulewski et al. (2023) identifies three key burdens that educators face when employing performance and competency-based assessment methods. Firstly, there are notable difficulties in interpreting and applying assessment scales, which can compromise consistency and fairness in the evaluation process. Secondly, logistical challenges arise, including the need for sustained observation, the timely completion of assessments, and the provision of high-quality, actionable feedback to learners. Lastly, teachers often struggle with the ongoing monitoring of student progress and the necessity of making fair, well-informed decisions regarding their competence. These themes are echoed across the literature, highlighting the multifaceted nature of the challenges. For example, (Popham et al., 2001) iterates the issue of subjectivity in grading, which can lead to variability and potential bias in assessments. (Wiggins et al., 2005) describes PBAs as being inherently time and resource intensive, placing additional strain on educators who may already be operating under significant workload pressures. (Arter & McTighe, 2000) discuss the logistical complexities inherent in managing PBAs, which often demand a high level of organisation and coordination. Moreover, the process frequently requires the involvement of subject matter experts (Barber & Phillips, 2000; Ellis et al., 2015; Gallardo, 2020), further complicating the implementation and scaling of such assessments. As teacher burden presents a key challenge in delivering oral PBAs, there is a need to explore new approaches to delivering PBAs that reduce teacher burden but maintain the benefits of these assessment methods.

### 1.3 AI conversational agents in education

The integration of AI conversational agents into education represents a transformative step towards addressing the increasing demands on teachers. By supporting various pedagogical processes, including assessment administration, these systems have the potential to alleviate some of the workload associated with modern teaching practices. Through personalised and emotionally intelligent interactions, AI conversational agents can streamline assessment processes, enabling teachers to focus more effectively on critical aspects of their role.

Across the globe, educational frameworks are adapting to incorporate technology-enhanced and personalised curricula. Among the most influential advancements in this regard is artificial intelligence (AI), which educators increasingly recognise for its potential to enhance the efficiency and effectiveness of teaching and assessment processes (Alshumaimeri & Alshememry, 2024). Within the domain of human-computer interaction (HCI), AI conversational agents have emerged as a key development,

providing tools that simulate natural conversation through image, text, or spoken language (Laranjo et al., 2018). Since the development of the first chatbot in 1966, the field has evolved significantly. Modern AI conversational agents can now go beyond rule-based dialogue systems to fully interactive platforms capable of understanding user intent, accessing diverse data sources, and generating conclusions, assumptions, or even actions (Adamopoulou & Moussiades, 2020; Weizenbaum, 1966). In education, these agents are used in numerous ways, such as facilitating assessments, increasing dialogue among peers, offering role-play scenarios, and answering student queries (Gonda & Chu, 2019; Maryadi et al., 2017; Shorey et al., 2019; Wang et al., 2020). Their flexibility in supporting interactions, whether between individuals or between humans and computers, has led to their widespread adoption and growing acceptance. The conceptual framework developed as a result of a systematic review of pedagogical AI conversational agents by (Yusuf et al., 2025) stipulates that contemporary agents can be used in a wide array of contexts within education, such as improving students' cognitive ability, providing them with pastoral care and giving instruction. For teachers, the use of AI conversational agents presents opportunities to reduce workload by automating repetitive tasks, such as answering routine queries or managing assessment logistics. This enables educators to dedicate their efforts to higher-order responsibilities, ultimately fostering a more balanced and manageable workload.

The emergence of generative AI, driven by advancements in large language models (LLMs), has further enhanced the capabilities of AI conversational agents. LLMs, which use mathematical models to predict the likelihood of word sequences, form the foundation for many applications involving natural language processing (Devlin et al., 2019). The introduction of neural networks and transformer architectures has significantly improved these models' ability to interpret and generate human-like text, capturing nuances such as tone, context, and emotional subtleties (Vaswani et al., 2017). These improvements have expanded the potential applications of LLMs, from language translation and question-answering systems to multimedia generation and conversational agents. LLM-augmented AI conversational agents go beyond conventional systems, offering open-dialogue capabilities that allow users to engage across a wide range of topics (Abbasian et al., 2024; Cherakara et al., 2023). Within education, these advanced agents provide an effective means of facilitating assessments, not only by managing complex interactions but also by adapting to the needs of individual learners. For teachers, this reduces the effort required to design and administer assessments manually, offering a scalable solution to the growing demands of contemporary educational practices. By incorporating emotionally intelligent LLM-augmented conversational agents, educators can enhance the assessment experience while mitigating the administrative burden, ensuring that academic integrity and effective learning remain central in an increasingly technology-driven environment.

### 1.3.1 Conversational agents to reduce teacher burden in PBAs

There have been many conversational agents that have been designed and incorporated into learning situations, which have resulted in reduced teacher burden. This section provides of overview of AI conversational agents that have been used to

reduce teacher burden by highlighting some examples that have been used for this purpose.

Reducing cognitive load is a challenge in PBAs, especially where teachers often manage intricate planning and assessment tasks (Richmond & Regan, 2023). AI conversational agents, such as Deakin Genie, Jill Watson and the agents developed by Teng et al. (2024), represent a range of tools that actively address this issue. The Deakin Genie aids in task organisation and provides immediate answers to complex queries, serving as a virtual assistant that simplifies logistical and instructional burdens (Saihi et al., 2024). Meanwhile Jill Watson, an LLM-augmented AI conversational agent excels in automating responses to routine and context-specific instructional questions, freeing educators from repetitive tasks (Taneja et al., 2024). The intelligent agents developed by Teng et al. (2024) utilise retrieval-augmented generation to dynamically create educational resources tailored to individual classroom needs. Together, these tools showcase how AI conversational agents alleviate cognitive demands by enabling real-time processing of instructional data and generating actionable insights for teachers, allowing them to concentrate on high-value teaching activities. Furthermore, the cognitive load of teachers is also linked to teacher workloads and increasing pressure on educators (Longmuir & McKay, 2024). Rudolph et al. (2024) explores the potential of AI to optimise routine tasks and reduce educators' operational burdens. These tools also demonstrate the transformative potential of AI in enabling educators to focus on strategic teaching priorities while delegating routine tasks to intelligent systems.

In addition to the cognitive load of teachers, providing timely and detailed feedback in PBAs often overwhelms educators due to the sheer volume of submissions and the need for personalisation (Brown, 2023). AI conversational agents, such as Jill Watson, GraderTA and AutoR, demonstrate how AI can transform feedback mechanisms. Jill Watson not only answers students' questions but also analyses their assignments to deliver specific, constructive feedback (Taneja et al., 2024). GraderTA enhances this by automating evaluation processes, leveraging AI conversational agents to identify areas of improvement and ensure consistency in grading criteria Teng et al. (2024). The AutoR system takes this a step further by synthesising assessment data into easily interpretable reports (Latif et al., 2024). This system, combined with findings from Le Cunff et al. (2024) on cognitive load management in online environments, emphasises the importance of structured feedback to reduce educator workload and enhance student understanding. These agents collectively reduce the burden of feedback provision by automating repetitive tasks and ensuring personalised, high-quality responses.

Another role of teachers is to establish standardisation and fairness in PBAs to ensure equity in assessment outcomes. AI conversational agents, such as GraderTA, EthicsTA and Jill Watson play, important roles in this regard. GraderTA has the ability to apply consistent grading rubrics, ensuring uniformity across diverse student submissions, and similarly EthicsTA evaluates materials against ethical guidelines, reducing the risk of bias in instructional content Teng et al. (2024). The Jill Watson agent contributes by providing unbiased feedback through its reliance on machine learning algorithms trained on diverse datasets, minimising human error and subjective bias (Taneja et al., 2024). Rudolph et al. (2024) further emphasises the necessity

of ethical considerations and data transparency to ensure equitable outcomes in AI-driven educational processes. These agents highlight the importance of AI in promoting fairness and standardisation, ultimately reducing the burden of teachers in having to personally conform to unbiased standardisation.

### 1.3.2 The potential of emotionally intelligent, LLM-augmented AI conversational agents to teacher burden in PBAs

PBAs are widely regarded as offering pedagogical improvements over traditional examination methods, delivering a more comprehensive evaluation of student understanding and skills (Gallardo, 2020). However, PBAs also present significant challenges for teachers, particularly in terms of the increased workload required to design, facilitate, and evaluate these assessments, and expert knowledge required by teachers (Richmond & Regan, 2023). In this context, AI conversational agents offer a promising solution to support teachers in managing these demands. When enhanced with LLM capabilities, these agents can evolve into emotionally intelligent systems, capable of dynamically interacting with students and adapting to their needs. While existing research has explored the use of AI conversational agents in assessment contexts (Gonda et al., 2018; Lee & Fu, 2019; Maryadi et al., 2017), much of the focus has been on improving the student experience, with limited investigation into how these systems can directly address teacher burden. The potential of emotionally intelligent, LLM-augmented agents to assist in streamlining assessment processes, such as facilitating real-time interactions, automating routine tasks, and providing contextual subject-specific and emotional support to teachers is an area that warrants further exploration. Such systems could reduce the time and effort required by teachers, allowing them to focus on more complex and pedagogically impactful aspects of their role. By mitigating the pressures associated with assessment administration and subject knowledge requirement in these assessment processes, LLM-augmented conversational agents hold the potential to significantly alleviate the workload of educators, making PBAs more sustainable in practice.

### 1.3.3 Technology acceptance in educational AI

Understanding how educators accept and integrate AI systems requires a theoretical framework, such as the Technology Acceptance Model. Originally proposed by Davis (1989), the model posits that perceived usefulness and perceived ease of use are critical determinants of users' acceptance of technology. Extensions of the model (Venkatesh & Davis, 2000) have shown that social influence and facilitating conditions also impact adoption, particularly in educational settings (Teo, 2011). In parallel, pedagogical frameworks such as the TPACK model (Technological Pedagogical Content Knowledge) and AI-specific instructional models highlight the importance of aligning technology tools with pedagogical goals and teacher capacities (Holmes et al., 2019; Mishra & Koehler, 2006). The successful deployment of emotionally intelligent AI agents requires not only technical functionality but also alignment with these pedagogical and acceptance frameworks to ensure trust, usability, and educational impact.

### 1.4 Aims of this study

This study presents "AIvaluate", an emotionally intelligent, LLM-augmented pedagogical AI conversational agent developed with the aim of supporting teachers by facilitating oral PBAs. The AIvaluate system consists of innovative functionalities designed to address teacher burden in PBAs. Specifically, the following research questions are addressed in this study:

RQ1. To what degree does AIvaluate influence teacher grading outcomes during PBAs compared with equivalent traditional face-to-face assessments?
RQ2. How satisfied, in terms of usability, are teachers with the AIvaluate assessment compared with equivalent the traditional face-to-face assessment format?
RQ3. What factors influence teachers' preferences for AIvaluate assessments compared with traditional face-to-face assessments?

The remainder of this paper presents the AIvaluate system architecture and application walkthrough in Sect. 2, the methods used for empirical evaluation of the system in Sect. 3, followed by the results based on the data collected in Sect. 4. Furthermore, a discussion of the results and future recommendations is presented in Sect. 5 and conclusion are drawn in Sect. 6.

## 2 The aivaluate system

This section presents an overview of the AIvaluate system, detailing a step-by-step walkthrough of the application from the perspective of teacher users, demonstrating how they interact with the system during the AI conversational agent-based assessments. Furthermore, the system architecture, workflow, and operational mechanics is also presented in this section.

### 2.1 System walkthrough

Upon launching the AIvaluate system, which is rendered within a full-screen browser environment, the teacher is welcomed with an introduction screen, which is presented in Fig. 1.

The introduction *launch screen* has two functionalities; firstly, it welcomes the user into the system and secondly, it allows the user to select their role the using *user type selection* menu, based on the two defined user groups, namely teachers and students. Upon selecting the *teacher* user type in the introduction screen, the student will be transitioned to the *teacher interface* screen, which is presented in Fig. 2.

The *teacher interface* hosts a *chat history panel* where the student can observe the inputs received from the teacher. The chat panel employs test-to-speech (TTS) functionality to convert the student's messages into audio-based verbal prompts, enhancing the interactive and anthropomorphic qualities of the interface. Positioned beneath this is the *teacher input selection area*, which enables the teacher to compose and dispatch messages to the student. This input area is equipped with *controller buttons*,


**Fig. 1** AIvaluate introduction screen

including options to send a completed message or clear the text field. The *dictation functionality* allows the teacher to utilise the speech-to-text (STT) feature via microphone which allows verbal input of responses. In addition to this interface, the teacher also has access to a *suggested replies section*, wherein the system generates three emotionally intelligent, suggested adaptive responses for the teacher, based on three variables; 1) the student's last input, 2) the conversation history for context and 3) the student's emotional state based on the self-reporting tool. The teacher also has access to a second display, which provides them with the results of the emotional state, as presented in Fig. 3.

The *emotional state reporting* screen provides the teacher with a real-time report of the student's emotional state. This display hosts two emotional state reports; 1) *facial attribute analysis results* and 2) *self-reporting result*. The former report includes data derived from the analysis of the student's facial attributes, which features a *face detection live feed* of the student's face which is captured by a web camera on the student's device, allowing the teacher to assess body language. Additionally, the extent of each of the seven primary emotional states is presented to the teacher in both statistical and visual (percentage-bars) formats, in the *emotional state results* area. These elements enable the teacher to discern the emotions exhibited and their intensity as conveyed through the student's facial expressions. The latter report in the *self-report result area* consists of a real-time update of the student's anxiety levels, as self-reported by the student. This display shows a scale (1–10) for the teacher to analyse, accompanied by a descriptor and visual representation of the emotion (emotion icon) corresponding to

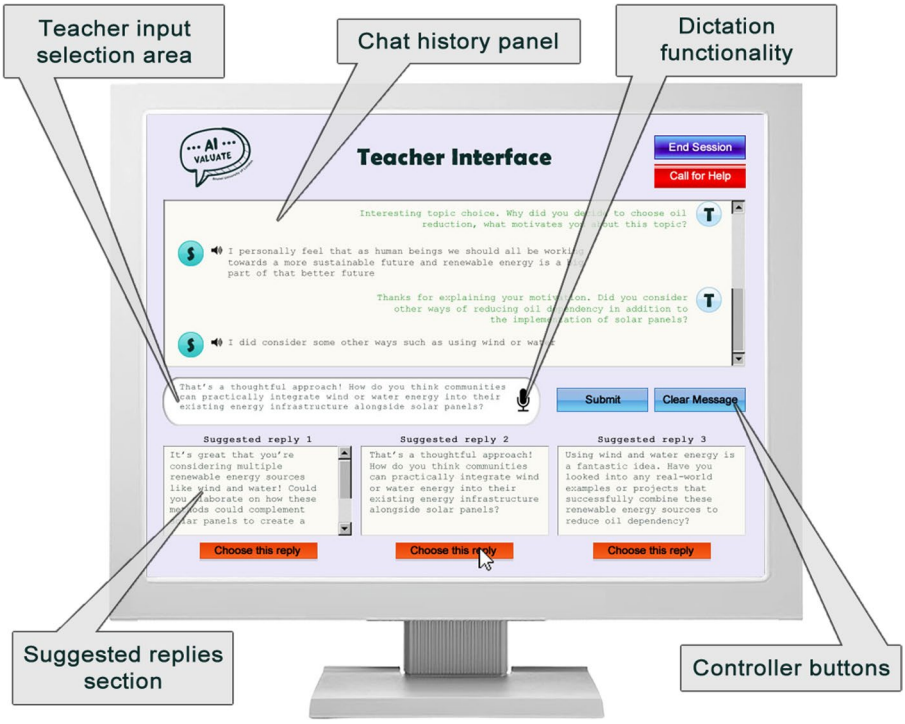


**Fig. 2** AIvaluate teacher interface screen

each level. From the *launch screen* as shown in Fig. 1, if the user selects the *student* user type, the teacher will be transitioned to the *student interface* screen, which is presented in Fig. 4.

The *student interface* hosts a *chat history panel* where the student can observe the inputs received from the teacher. Like the teacher interface, the chat panel employs TTS functionality to convert the student's messages into audio-based verbal prompts. Positioned beneath this is the *student input selection area*, which enables the student to compose and dispatch messages to the teacher. This input area is equipped with *controller buttons*, including options to send a completed message or clear the text field. Similar to the *teacher interface*, the *dictation functionality* allows the student to utilise the STT feature via microphone which allows verbal input of responses, thereby augmenting the naturalistic communication experience. In addition to this interface, the student also has access to the *emotional state reporting tool*, which is available to the student as a secondary display on a touch-screen tablet. The *emotional state reporting tool* is presented in Fig. 5.

The *emotional state reporting tool* hosts the student anxiety level *self-reporting input*, a critical component that allows the student to actively monitor and report their live emotional state throughout the assessment session. The input features a *slider bar*, enabling the student to self-assess and indicate their current level of anxiety on a scale from 1 to 10, with descriptors and semiotic *emoticons* to support their understanding of the levels. This input operates in real-time, allowing the student to make


**Fig. 3** AIvaluate teacher emotional state results screen

continuous adjustments to reflect their emotional state at any given moment, thereby providing a dynamic and immediate self-reflective assessment that informs the ongoing interaction. Lastly, the *selection view* shows the student their current-selected emotion at any given point during the assessment for reiterative purposes.

## 2.2 System architecture and functionality

AIvaluate is an emotionally intelligent LLM-augmented pedagogical AI conversational agent, deployed as a disembodied chatbot which has three key functionalities; 1) a *conversational interface* wherein dialogue can take place between a teacher and student during oral PBAs, 2) *emotional state detection* of the student for use by the teacher and 3) *LLM-augmentation* to generate automated suggested replies for the teacher. Figure 6: AIvaluate system architecture and workflow model presents the high-level AIvaluate system architecture and data flow model.

### 2.2.1 The conversational interface

The *conversational interface* is hosted in a browser environment through a two-tier construction; the *student-view* and *teacher-view*. The first feature that the AIvaluate system hosts is a chatbot interface, wherein students and teachers, in their respective views, can converse through a series of inputs and outputs. The input can be typed

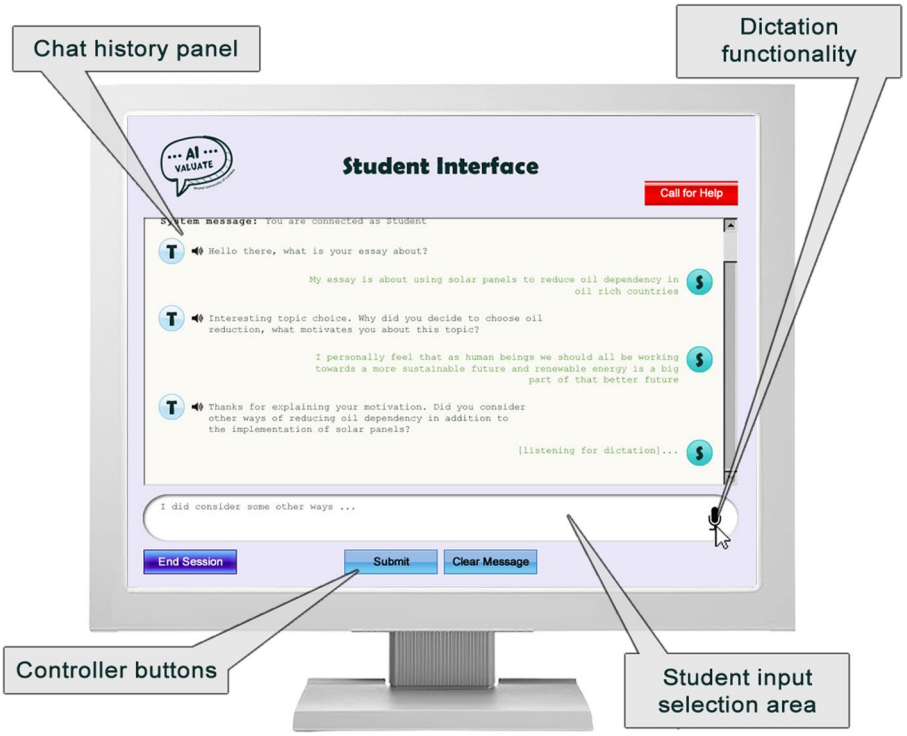


**Fig. 4** AIvaluate student interface screen

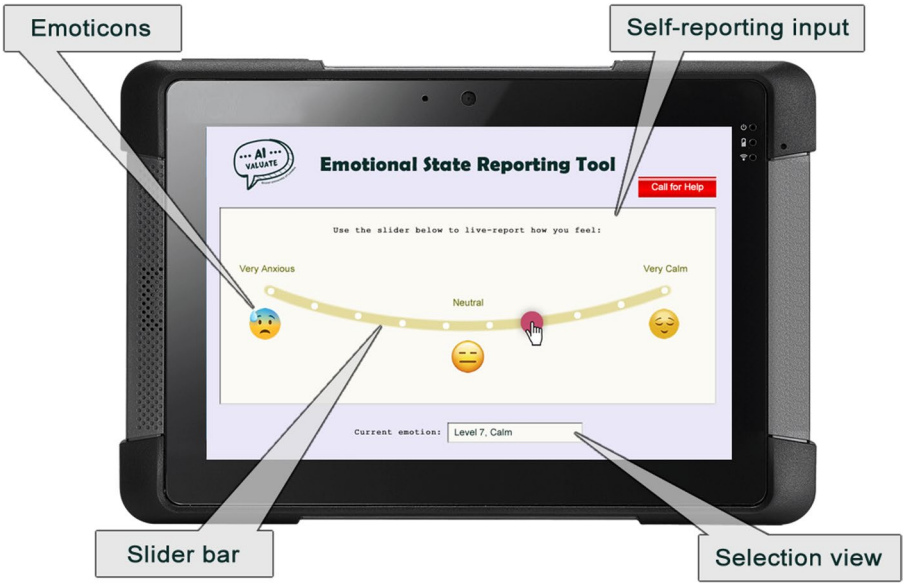


**Fig. 5** AIvaluate emotional state self-reporting tool

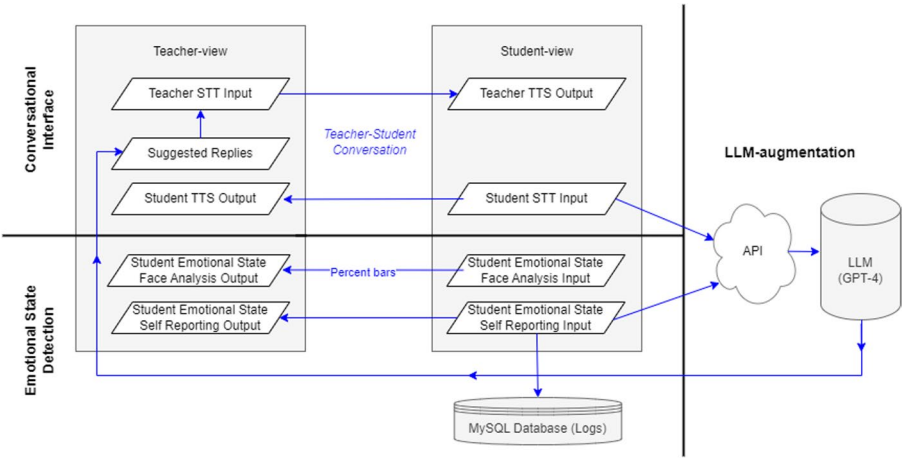


**Fig. 6** AIvaluate system architecture and workflow model

or verbal, in which case the application utilises a dictation STT library to ensure the input at terminal level is sting datatype. Outputs are also displayed in textual format within the interface, in addition to utilising a TTS library to convert the text into audio (spoken form). This dual text-speech method is used to present both teachers and students with an anthropomorphic conversational experience. The second feature of AIvaluate is a dual-reporting method for the students' *emotional state detection*, including a *self-reporting* tool and a *facial analysis* recognition feature.

### 2.2.2 Emotional state detection

AIvaluate has two methods of assessing a students' emotional state. Firstly, the emotional state *self-reporting* tool wherein the student reports their live-emotional state during the face-to-face and virtual meetings. Secondly, the *facial analysis* system which detects the students' emotional state based on their facial expressions during the virtual meeting.

The *self-reporting* tool is presented as a slider to the student which during an active session, allows them to input their emotional state from a scale-slider between 1 and 10, indicating their level of anxiety. The teacher is able to the see a real-time feed of the students' emotional state self-report. Additionally, a session log is produced for prospective analysis. The real-time emotional state data from the *self-reporting* tool is utilised by the *LLM-augmentation* integration using a custom JSON-based Application Programming Interface (API). The *self-reporting* tool was developed using web-based programming for server-side operations with MySQL as the *database* used for logging data, and scripting programming for browser-side behaviour manipulation.

The *facial analysis* system within AIvaluate is a refined web-based tool that captures live video, detects faces, and analyses emotions in real-time, providing immediate feedback through a web interface in the *teacher-view*. The *facial analysis* system did not contribute to the *LLM augmentation* of AIvaluate, it was solely for the use of teachers to gain an insight into the students' emotional state, to support teacher

understanding. This system utilises several key technologies: OpenCV for face detection, Flask for creating the web interface, and DeepFace for emotion recognition. The process begins with face detection using OpenCV's Haar Cascade Classifier, which identifies faces by analysing pixel intensity patterns. Each video frame is converted to grayscale, simplifying the data by reducing it to a single intensity value per pixel, enabling faster processing. The face detection algorithm employs a sliding window approach, analysing small sections of the grayscale image for patterns resembling facial features, such as the arrangement of the eyes, nose, and mouth. This is achieved by comparing pixel intensity sums across regions and matching them against predefined thresholds established during the classifier's training. Once a face is detected, the system analyses its emotions using the DeepFace library. DeepFace applies a deep learning model to evaluate facial features and assign probabilities to emotions such as happiness, sadness, and anger. These probability scores represent the confidence levels for each emotion, ensuring the total probabilities always sum to 1, providing a complete and normalised emotional assessment. The analysis operates continuously on each detected face in the video feed, with results transmitted to the user interface in real-time. SocketIO manages the server-client communication, enabling instantaneous updates. Processed video frames are streamed sequentially, creating a seamless and continuous video display on the webpage for the user.

To enhance reliability, the facial analysis component of AIvaluate is built using the DeepFace library, an open-source facial recognition and emotion classification framework. DeepFace employs deep convolutional neural networks (CNNs) and supports models such as VGG-Face and Facenet. Independent validation studies have reported high emotion classification accuracy across multiple datasets, with high scored performance benchmarks in standard emotion categories (Kollias & Zafeiriou, 2018; Mollahosseini et al., 2017). To ensure consistent emotion detection, our system uses grayscale conversion and region-specific feature extraction (e.g., eyes, mouth, eyebrows), which are fed into a trained CNN model for real-time classification. The system outputs probability scores across seven Ekman-based emotions, namely; happy, sad, angry, fearful, surprised, disgusted, neutral (Ekman & Friesen, 2003).

### 2.2.3 LLM-augmentation

The *LLM-augmentation* functionality is used by the system in the *teacher-view* interface to generate emotionally intelligent and automated suggestions for replying to student inputs. The LLM utilised is OpenAI's Generative Pre-trained Transformer (GPT-4) which is incorporated into the agent using their API integration library, allowing requests to take advantage of OpenAI's trained neural network to produce relevant, coherent, and contextually appropriate responses. In addition to the LLM producing a suggested response for the teacher based solely on the students' input, the *LLM-augmentation* codebase includes a custom variable to report the students' real-time emotional state based on the live feed of the *self-reporting* tool. The emotional state data is then used by the LLM to generate an emotionally appropriate, automated suggested response for the teacher based on both, the students' conversational input and their emotional state. The underlying algorithms for the generational

of emotionally intelligent responses are presented in Table 1, in accordance with the standardised software–engineering format of pseudocode with inclusion of interest points (◁).

The pseudocode presented in Table 1 outlines the methodological framework for the emotionally intelligent responses for the teacher. The core function of this appli-

**Table 1** Emotionally intelligent responses (with numbered references)

**PSEUDO-CODE:** GenerateEmotionallyIntelligentSuggestions <Method>, <Data Stream>

**INPUT:** student_message <string> emotional_state_data <JSON> conversation_history <list of strings>

**OUTPUT:** *suggested_responses* <list of strings>

**ACTIVATION:** *student_input* <string>

```
BEGIN
  //Retrieve Emotional State                                                    ◁(1)
  CALL get_emotional_state()
  SET emotional_state = {"level": emotion.level, "description": emotion.description}
  //Validate Emotional State Data                                               ◁(2)
  IF emotional_state.level IS NULL OR emotional_state.level IN ["NOT_STARTED", "SESSION_START",
"SESSION_END"] THEN
      RETURN {"level": NULL, "description": "No valid emotional data available"}
  END IF
  //Determine Tone Based on Emotional State                                     ◁(3)
  IF emotional_state.level > 5 THEN
      SET tone = "gentle and supportive"
  ELSE
      SET tone = "clear and academically focused"
  END IF
  //Construct Refined Prompt for LLM                                            ◁(4)
  SET refined_prompt = (
      "The student mentioned their work is about " + student_message +
      "Generate a response in tone " + tone +
      "Considering the student feels " + emotional_state.description +
      "Incorporate the context of the ongoing conversation " + conversation_history +
      "Keep responses concise and contextually appropriate."
  )
  //Generate Suggestions Using OpenAI's GPT-4                                   ◁(5)
  CALL openai.ChatCompletion.create(
      model="gpt-4",
      messages=conversation_history + [{"role": "user", "content": refined_prompt}],
      max_tokens=50,
      n=3
  )
  SET suggested_responses = GPT_output.choices
  //Ensure Uniqueness of Suggestions                                            ◁(6)
  SET unique_responses = REMOVE_DUPLICATES(suggested_responses)
  WHILE LENGTH(unique_responses) < 3 DO
      CALL openai.ChatCompletion.create(...) ◁ Repeat API Call for Additional Suggestions
      ADD NEW_RESPONSES TO unique_responses
  END WHILE
  //Return Suggested Responses                                                  ◁(7)
  RETURN unique_responses[0:3]
END
```

cation begins with retrieving the student's emotional state from the JSON-based API, which is exported from the self-reporting tool ◁(2). The API returns data that includes an anxiety level (a numerical value from 1 to 10) and a corresponding descriptive label, such as "slightly anxious" or "extremely anxious". This emotional data serves as the foundation for the system's decision-making process. When the app receives this data, it first checks whether the response includes valid information, such as an anxiety level, and not a status update like "SESSION_START" or "SESSION_END". If valid data is available, it proceeds to the next steps; otherwise, it reports that no emotional data is available. The emotional state data, specifically the anxiety level of the student, is used to determine the tone of the suggested responses that the system will generate through the LLM-augmentation using generative AI ◁(3s). Algorithmically, this can be seen as a conditional decision-making process based on comparing the anxiety level against a threshold value. If the anxiety level is greater than a neutral level of 5, which indicates a higher state of anxiety, the system adjusts the response tone to be gentler and more supportive. Conversely, if the anxiety level is 5 or below, suggesting a lower or manageable state of anxiety, the response tone is set to be clear and academically focused, without additional emotional supports in terms of the language used. This decision process can be expressed as follows:

$$Tone = \begin{cases} "gentle\ and\ supportive" & if\ anxiety\ level > 5, \\ "clear\ and\ academically\ focused" & if\ anxiety\ level \leq\ 5. \end{cases}$$

Once the appropriate tone is determined, the application uses OpenAI's language model to generate suggested replies for the teacher. The replies are crafted by combining the input messages of the student, their emotional state, and a refined prompt that incorporates the specified tone ◁(4). The language model then generates multiple potential responses, each phrased as a question or comment that aligns with the determined emotional tone ◁(5). The generation of these suggested replies uses probability principles. The model predicts the most likely sequence of words based on the input data it receives, which includes both the most recent message of the student, the historic inputs of the student (chat history) and the emotional context from the self-reporting tool. It selects words and phrases with the highest probability of fitting the intended tone and context, similar to statistical models that predict outcomes based on given conditions. For example, if the model receives a prompt indicating that the student is "very anxious," it prioritises responses that include reassuring language and gentle questioning, reflecting a higher probability of suitability given the anxiety context. To maintain uniqueness and appropriateness, the system generates several potential responses and checks them for duplications. If the responses are too similar, the model is prompted to generate additional replies until three distinct suggestions are available ◁(6). This process ensures that the teacher receives a varied set of options, each tailored to the student's emotional state, which is repeated every time the function is called with new student inputs ◁(7).

## 3 Methods

This study employed a mixed methods approach for data collection and analysis to comprehensively address the specific research aims, details of which are presented in this section. Figure 7 provides an overview of the protocol.

### 3.1 Participants

A total of 35 teachers were recruited from a British international school to conduct assessments with a student-teacher ratio of 1:1. The inclusion criteria for teacher participants were: 1) qualification (must be a qualified teacher with minimum 3 years of teaching experience), 2) assessment experience (must have experience in conducting the PBAs), 3) computer literacy, and 4) native-level proficiency in the English language. All teachers met the required criteria. The teachers assessed 35 students, also recruited from the same institution. These students were in their final year of the International Baccalaureate Diploma Programme (IBDP) and were engaged in completing their Extended Essay (EE), which includes a summative PBA known as the *viva voce*. Prior to participation, comprehensive informed consent was obtained from all participants. For individuals requiring proxy consent, informed consent was additionally secured from their legal guardians. Participants were explicitly apprised of their right to withdraw from the study at any point, without prejudice or adverse consequences. The number of participants required was estimated using G*Power

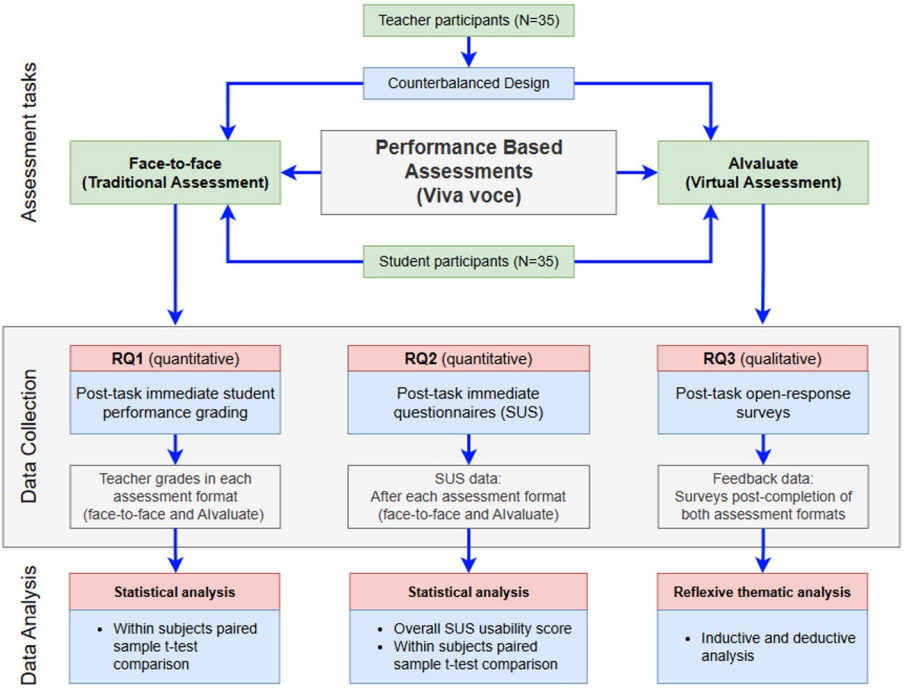


**Fig. 7** Mixed methods data collection and analysis procedure

3.1 software. To ensure a power of 0.80 with a medium effect size of 0.5 (dz), the calculation determined that a minimum total of $N = 34$ participants would be necessary. No participants reported any functional impairments that would affect their ability to engage with the AI agent, although minor adjustments to software settings were made to accommodate user preferences (e.g., adjusting audio output levels and screen size/scale levels). Participants were randomly allocated to condition sequences using a computer-generated block randomisation method to ensure balanced exposure to each assessment type (AIvaluate-first vs. control-first). Group equivalence was ensured by stratifying participants according to course level and having similar levels of prior engagement with digital learning and environments. Informal feedback from participants indicated that students had no previous experience with emotionally responsive AI assessment tools, thereby minimising the likelihood of significant bias. Teachers were blinded to the randomisation order during scheduling to reduce expectancy effects.

### 3.2 Protocol and instrumentation

Interactive PBA-based viva voce assessment sessions were held with all participants as part of the study on the effectiveness of the AIvaluate system. Each teacher participated in two distinct assessment sessions, one face-to-face and one virtual, each conducted with a student who was not their direct EE supervisee, hence the teachers were not familiar with the students' work. A researcher was present during all sessions, and written consent was obtained from all participants, as well as from guardians where student participants were under 18 years of age. On arrival at each session, any questions from participants were addressed, and a brief presentation was provided detailing the research process. Each teacher participated in two assessment meetings with the same student, separated into face to face and virtual formats, and a counterbalanced design was employed (i.e. alternating order of meetings for student-participants) to mitigate order and carryover effects and increase the generalisability of the research (Brooks, 2012). To reduce bias and support internal validity, a set of standardised control measures was implemented across both the AIvaluate and control assessment conditions. Task prompts were identical in wording, length, and cognitive demand to ensure uniformity in student challenge. Sessions were scheduled to occur at consistent times of day (between 9:00 a.m. and 11:00 a.m.) to minimise variation in student fatigue or alertness, and were conducted in the same physical room with matched environmental conditions (lighting, seating, acoustics). Teachers were instructed to follow a scripted protocol when introducing the activity, offering clarification, and concluding the session, thereby limiting variations in teacher behaviour or tone. Collectively, these measures were designed to preserve the internal validity of the experiment and isolate the effects of the AI intervention on learner experience and teacher workload.

#### 3.2.1 Face-to-Face assessment meeting

The face-to-face assessment involved a conventional viva voce session conducted in person, lasting 10–15 min, where the teacher and student engaged in a discus-

sion about the student's EE work. During this session, the students were required to explain, justify, and reflect upon various aspects of their essay, responding to questions posed by the teacher. To maintain consistency and standardisation of interactions, the teacher was provided with a simple script to guide the initial stages of the conversation and ensure comparability across sessions. Throughout the face-to-face assessment, students were asked to self-report their anxiety levels using an emotional state self-reporting tool, which is the same tool used for the virtual assessment as presented in Fig. 5, designed to capture real-time emotional states during the interaction.

### 3.2.2 Virtual assessment meeting

The virtual assessment was conducted using the AIvaluate system, designed to simulate an AI-augmented interaction through the Wizard of Oz technique. In this setup, students were unaware that the teacher was the interlocutor; instead, the assessment was framed as being conducted by the AIvaluate system, a fully automated artificially intelligent conversational agent (Steinfeld et al., 2009). The sessions were hosted in separate, soundproof rooms to prevent teachers from hearing the student during the interaction. Figure 8 illustrates the room layout and hardware setup, showing how the student and teacher were isolated, and the equipment arranged.

Given the emotionally sensitive nature of facial expression and self-reported anxiety data, several ethical safeguards were implemented. All emotional state data collected was anonymised and stored on encrypted institutional servers with restricted access. Students (and guardians where appropriate) were informed in advance of the use of real-time emotion analysis and were allowed to opt out without penalty. No emotional data was used for individual student evaluation; it served only to support the teacher during the session. All data streams involving facial attribute analysis were deleted following session completion. This approach aligns with best practices recommended for emotion AI applications in education and adheres to data protection principles on biometric data sensitivity (Gasiokwu et al., 2025; Khan, 2024).

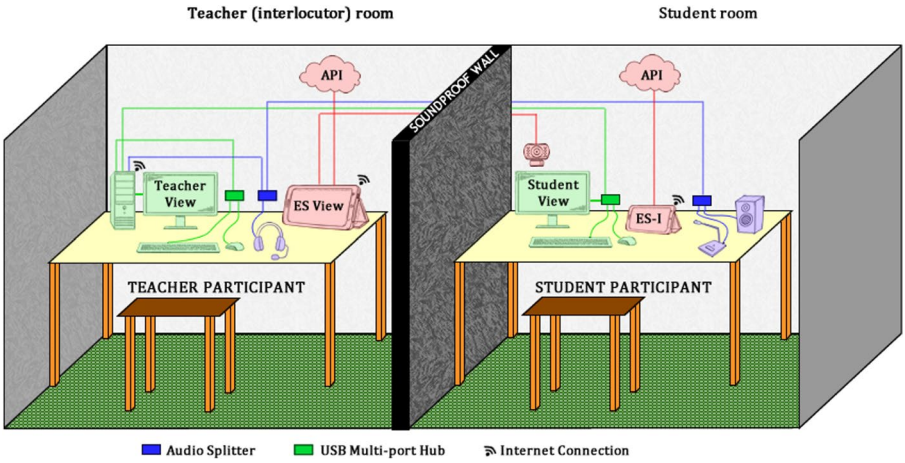


**Fig. 8** AIvaluate hardware setup

In the *Teacher (interlocutor) room*, the teacher had access to two screens: one displaying the AIvaluate interface labelled *Teacher View* as presented in Fig. 2, used for guiding the interaction and another screen providing real-time feedback on the student's emotional state, called the *ES View*, as presented in Fig. 3. The teacher utilised a keyboard, mouse, and headset to interact with the system and manage the flow of the session. In the *Student room*, the student interacted with AIvaluate using a similar setup with two screens: one labelled *Student View* as presented in Fig. 4 displaying the AIvaluate interface, and another screen enabling the student to input their emotional state, labelled *ES-I* as presented in Fig. 5. The AIvaluate system was designed to track and log emotional states in real time, providing critical data for analysing anxiety levels during the virtual assessment. Teachers were provided with suggested conversational prompts and a live feed of the student's emotional state to guide the interaction. The virtual assessment, like the face-to-face session, lasted 10–15 min.

### 3.2.3 Data collection

The study collected data in three phases; 1) teacher-assessed grades in both assessment formats, 2) quantitative data from post-assessment SUS questionnaires, and 3) qualitative insights from open-response questions collected after both assessment formats.

Teachers assessed student performance during both the face-to-face and virtual AIvaluate assessment sessions using a structured grading rubric. This rubric, developed in alignment with the IBDP assessment criteria, provided a consistent framework for evaluating articulation, clarity of argument, and subject-specific understanding. The EE is a core requirement of the IBDP graded on a 34-point scale, encompassing six distinct criteria such as research focus, analysis, and communication. In this study, teacher-assigned grades for the viva component were recorded as percentages out of 100, mapped directly onto the IB criteria to reflect the depth, coherence, and analytical rigour expected at this academic level. These percentage-based scores were not based on subjective perception but derived from a rubric that mirrors the formal marking standards used in the school. This ensured comparability across assessment formats and provided a curriculum-grounded measure of student learning outcomes. The use of a common assessment rubric enabled a direct, valid comparison of performance between the AI-supported and traditional assessment formats, ensuring that the data reflected actual academic achievement rather than perceived engagement or participation alone. By examining the results, the study aims to uncover potential disparities in achievement or consistency between the face-to-face and virtual modalities, offering valuable insights into the impact of assessment delivery on student outcomes.

The post-assessment questionnaires completed by teachers directly after the completion of each assessment formats were based on the System Usability Scale (SUS). The SUS is a widely recognised and validated tool for assessing subjective usability and user satisfaction (Brooke, 2013), and hence was used to gauge participants' perceptions of the assessment method. The questionnaire comprised ten standard items, plus four additional bespoke statements related to evaluating the teacher burden, rated on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly

agree). Each SUS item was modified in accordance with the practitioner guidelines to reflect the specific context of the assessment by replacing the word "system" with either "face-to-face assessment" or "AIvaluate virtual assessment" depending on the participants assessment type. This adaptation ensured that the scale accurately measured the participants' subjective satisfaction and acceptance of the assessment format. Furthermore, qualitative feedback was gathered from students through post-task open-response questions. These questions aimed to explore participants' perceptions of both assessment formats, capturing insights into the usability, acceptability, and emotional impact of the AIvaluate system compared to traditional face-to-face PBA's.

#### 3.2.4 Ethical safeguards

This study received ethical approval from the host university's research ethics committee (Approval Reference: 47921-MHR-Jul/2024-52220-3), and all procedures followed institutional and GDPR data protection guidelines. Informed consent was obtained from all participants prior to data collection, as detailed in Sect. 3.1. Participants were provided with an information sheet outlining the study's aims, procedures, and their right to withdraw at any time without penalty. For student participants, this also included specific details regarding the use of facial emotion recognition and the self-reporting of emotional states. To address concerns related to bias, data misuse, and emotional manipulation, several safeguards were implemented. Emotional state data were anonymised at the point of capture and stored securely on encrypted AES-256 compliant servers. Facial attribute analysis data were streamed only during the session and were neither stored nor saved. The emotion recognition system operated in passive mode, meaning it did not provide real-time feedback or adapt its behaviour based on detected emotions. This design was intended to prevent any potential psychological influence or manipulation during the assessment process.

Furthermore, the AI system did not collect or process any demographic or sensitive personal data (e.g., race, gender, age), thereby reducing the risk of algorithmic bias in emotional inference. All emotion data were used exclusively for post-assessment research analysis and had no bearing on grading outcomes or qualitative feedback. These comprehensive safeguards were adopted to preserve participants' privacy, ensure ethical data handling, and protect the emotional integrity of students and teachers within a high-stakes educational environment.

### 3.3 Data analysis

This section is ordered according to the three primary data collection phases mentioned in Sect. 3.2.3, which includes analysis of the teacher-assessed grades data, SUS questionnaire data, and qualitative survey response data.

#### 3.3.1 Teacher-assessed grades data analysis (RQ1)

Tacher-assessed grades were assigned based on student performance during both the face-to-face and virtual assessments. These grades provided a structured and stan-

dardised evaluation, ensuring consistency across the two formats. For each assessment, the grades were recorded as a percentage out of 100. The mean and standard deviation of the grades from all sessions were calculated to provide an overarching comparison of performance between the face-to-face and virtual assessments. Furthermore, a paired-sample t-tests was applied to these grades to evaluate differences in outcomes between the two formats, determining the statistical significance of any observed variations and providing robust evidence on the comparative effectiveness of each assessment mode.

#### 3.3.2 SUS questionnaire data analysis (RQ2)

Overall SUS scores were calculated and interpreted according to the acceptability range, and the adjective and school grading scales (Brooke, 1996). This involved calculating a mean SUS representative value on a 100-point rating scale for each sample. These scores were then mapped to descriptive adjectives (Best imaginable, Excellent, Good, OK, Poor, Worst Imaginable), an acceptability range (Acceptable, Marginal-High, Marginal-Low, Not acceptable) and a school grading scale (i.e. 90–100 = A, 80–89 = B etc.). The baseline adjective and acceptability ranges are derived from a sample of over 3000 software applications (Bangor et al., 2009). To further analyse the usability of each assessment format, paired-sample t-tests were conducted to determine whether there were statistically significant differences between the mean SUS scores for face-to-face and virtual assessments on each individual SUS item. IBM SPSS statistics package version 29.0.1 was used for this analysis. This approach allowed for a detailed item-by-item comparison, examining how each usability aspect, as represented by the SUS items, differed between the two assessment types. By using paired-sample t-tests, we accounted for the fact that the same participants provided SUS ratings for both formats, thus controlling for individual variability and increasing the precision of the comparisons. Each t-test assessed whether the mean difference in scores for a specific SUS item, such as ease of use, consistency, and confidence using the assessment, was statistically significant, thereby offering insights into specific usability strengths or weaknesses unique to each format.

#### 3.3.3 Qualitative open-response survey data analysis (RQ3)

Reflexive thematic analysis, using NVivo software package version 12.6.1.97 was conducted to analyse the open-response survey data (Braun & Clarke, 2006, 2012, 2013, 2019, 2021). Analysis adopted both an inductive and deductive approach to analysis as coding and analysis often do not fit neatly into a single approach; instead, they typically involve a blend of both methods Byrne (2022). The integration of top-down (deductive) and bottom-up (inductive) coding processes in this approach facilitates a comprehensive and rigorous analysis (Proudfoot, 2023). Themes in deductive analysis were a priori researcher-led, while the inductive analysis allowed for new themes to emerge from the data (D'Amore et al., 2023; Tafazoli & Meihami, 2023). This process involved iterative and collaborative coding of survey responses, allowing themes and sub-themes to emerge reflexively across repeated analysis cycles, providing valuable contextual insights into the emotional and usability-related dimensions

of both assessment types, which allowed for an in-depth exploration of students' subjective experiences and perceptions of each assessment condition. While inter-coder reliability was not calculated, this is consistent with the principles of reflexive thematic analysis, which conceptualises coding as a subjective, interpretative, and contextually situated process rather than a reliability test (Braun & Clarke, 2021). Two researchers independently engaged with the data, generating initial codes and noting patterns of meaning. Through iterative cycles of discussion, disagreement, and refinement, the team collaboratively developed a robust thematic framework. This reflexive dialogue enhanced the depth and transparency of interpretation. Data saturation was assessed by tracking the emergence of new codes and sub-themes across responses. No new substantive categories emerged in the final set of responses, and thematic coverage was observed across participants, indicating sufficient saturation for the scope of this analysis. Memo writing and cross-case comparisons supported thematic consistency.

## 4 Results

### 4.1 Teacher-assessed grades

The first research question was to compare PBAs conducted through AIvaluate with traditional face-to-face assessments to identify the degree of influence of AIvaluate on teacher-assessed grading outcomes. The results of the comparison of the teacher-assessed grades between AIvaluate and the face-to-face assessments are presented in Table 2, which shows the descriptive statistics and paired-sample comparisons of teacher assessed grades for face-to-face (F2F) versus virtual (AI) assessments. The table includes means, standard deviations, pooled standard deviations, mean difference, Cohen's d effect sizes, the 95% confidence intervals (CI) for the mean differences and corresponding significance values. The pooled standard deviation was used to compute Cohen's d, and confidence intervals were based on the standard error derived from the pooled standard deviation.

**Table 2** Teacher-assessed grades for aivaluate and face-to-face assessments

| | F2F | | AI | | Mean Diff | Df | t | *p* (1-tail) |
|---|---|---|---|---|---|---|---|---|
| | Mean | St. Dev. | Mean | St. Dev. | | | | |
| | 70.571 [a] | 14.738 | 75.571 [a] | 14.419 | 5.00 | 34 | −1.893 | 0.033* |
| Format | Min Statistic | Max Statistic | Skewness Statistic | Std. Error | Kurtosis Statistic | Std. Error | Pooled SD | 14.579 |
| Face-to-face | 40.00 [a] | 96.00 [a] | −0.033 | 0.398 | −0.680 | 0.778 | Cohen's d | 0.343 |
| AIvaluate | 50.00 [a] | 100.00 [a] | −0.271 | 0.398 | −1.187 | 0.778 | CI | [−0.008, 10.008] |

[a] Grade percentage out of 100

* Statistically significant < 0.05 level

The paired samples t-test comparing student performance grades between AIvaluate and face-to-face assessments reveals that students achieved significantly higher grades during the AIvaluate assessments ($M = 75.57$, $SD = 14.42$) compared to the face-to-face assessments ($M = 70.57$, $SD = 14.74$), $t\,(34) = -1.89$, $p = 0.033$ (*1-tailed*). The gap score ($M = 5.00$) highlights this difference. Skewness and kurtosis statistics indicate relatively normal distributions for both formats, with AIvaluate showing a slightly negative skew (−0.27) and face-to-face assessments presenting an almost neutral skew (−0.03). These findings suggest a statistically significant improvement in performance outcomes when students completed assessments in the AIvaluate-hosted virtual environment. However, the effect size for this difference was small ($d = 0.343$), suggesting that while statistically significant, the practical magnitude of the grade improvement may be modest. The 95% confidence interval for the mean difference [−0.008, 10.008] borders zero, indicating that the true effect may vary and should be interpreted with caution. In conclusion, although statistically significant, the effect size was small and the confidence interval bordered zero, suggesting that the observed improvement may not be practically meaningful across all contexts. These findings should therefore be interpreted with caution and validated through larger-scale replications.

### 4.2 System usability

The second research question aimed to evaluate and compare the usability and user experience of AIvaluate and face-to-face assessment formats using the System Usability Scale (SUS). The overall SUS scores are illustrated in Fig. 9, which shows that the score for AIvaluate was 74.79. According to the SUS evaluation framework, this corresponds to "good" (descriptive adjective), "acceptable" (acceptability range), and a "grade C" (school grading scale). In comparison, the face-to-face assessment scored slightly lower, with an overall SUS score of 71.64, also rated as "good", "acceptable", and a "grade C". These findings suggest that while both formats are deemed usable and acceptable, AIvaluate demonstrates slightly higher usability in comparison to the face-to-face assessment format.

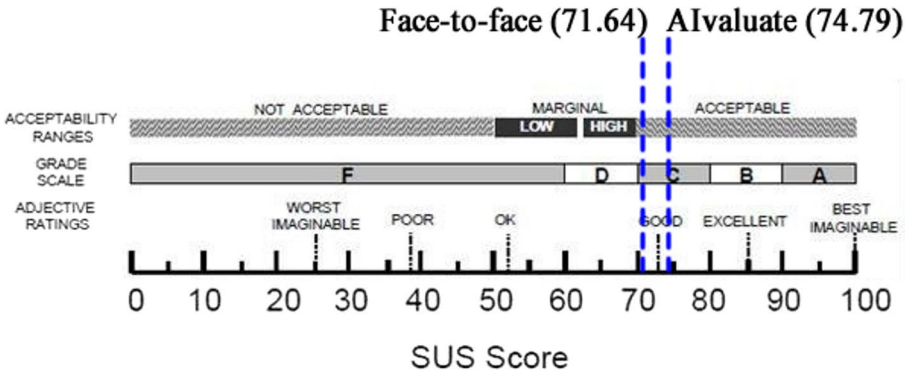


**Fig. 9** SUS rating scale results for AIvaluate and the face-to-face assessment

Follow-up analysis of individual SUS items were conducted to evaluate the usability of face-to-face assessments and AIvaluate assessments, paired-sample t-tests were used to compare the mean SUS scores for each format across individual SUS items. Table 3 presents the descriptive statistics and paired-sample comparisons of participant ratings for face-to-face (F2F) versus virtual (AI) assessments. The table includes means, standard deviations, pooled standard deviations, mean differences, Cohen's d effect sizes, the 95% confidence intervals (CI) for the mean differences and corresponding significance values. The pooled standard deviation was used to compute Cohen's d, and confidence intervals were based on the standard error derived from the pooled standard deviation.

The negative SUS items (S2, S4, S6, S8, and S10) were reversed so that higher scores consistently indicated a positive response, facilitating comparability between items. Across the ten standard SUS items (S1–S10), analysis revealed statistically significant differences in two items between face-to-face and AIvaluate assessments. For item S6, participants reported significantly higher scores for the AIvaluate format ($M=4.00$, $SD=0.68$) compared to face-to-face ($M=3.54$, $SD=0.91$), $t$ (34) = −2.85, $p=0.004$, indicating that AIvaluate assessments were perceived as less inconsistent. Similarly, item S10 showed a significant difference, with AIvaluate rated higher ($M=4.00$, $SD=0.89$) than face-to-face ($M=3.40$, $SD=1.29$), $t$ (34) = −2.18, $p=0.018$, suggesting that participants found AIvaluate assessments required less preparatory learning, reflecting its ease of use. Both S6 and S10 showed moderate-to-large effect sizes ($d=0.578$ and $d=0.549$, respectively), with their 95% confidence intervals not crossing zero, further supporting the robustness of these findings. The confidence intervals for S6 [0.186, 0.729] and S10 [0.225, 0.975] further support meaningful perceived differences in assessment consistency and learning burden between formats. For the remaining eight SUS items, no statistically significant differences were observed. For example, item S1 showed similar mean scores between face-to-face ($M=4.17$, $SD=0.81$) and AIvaluate ($M=3.89$, $SD=1.19$), $t$ (34)=1.12, $p=0.135$, indicating comparable participant interest in the frequency of each format. Although item S5 (integration of features) did not reach statistical significance ($p=0.066$), it showed a medium effect size ($d=0.314$) with a confidence interval [0.039, 0.589] that narrowly included zero, suggesting a possible trend worth further exploration in future studies.

The additional items (A1–A4), designed to evaluate participants' subjective experience with the assessments, also revealed significant findings. For item A4, AIvaluate ($M=4.37$, $SD=0.68$) scored significantly higher than face-to-face ($M=3.83$, $SD=1.03$), $t$ (34) = −2.80, $p=0.004$, indicating that participants perceived AIvaluate as having less impact on their workload. This result was supported by a moderate-to-large effect size ($d=0.636$) and a narrow confidence interval [0.250, 0.836], reinforcing the practical significance of this difference. Although item A4 indicated that AIvaluate did not significantly add to teachers' workloads, it is important to note that this finding is based on a single-item perception scale. The result should be interpreted with caution, as it reflects subjective impressions rather than objective workload metrics. The study did not collect direct data on workload components such as grading time, stress levels, or cognitive load. As such, while the result provides preliminary insights, it is not sufficient to conclude a definitive reduction in teacher

**Table 3** AIvaluate and face-to-face assessments comparison of SUS scores

| SUS Items | Mean F2F | Mean AI | Mean Diff | Pooled SD | 95% CI | Df | t | *p* (1-tail) | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|
| **S1**: I think I would like to have face-to-face/AIvaluate assessments frequently. | 4.171 | 3.886 | −0.286 | 1.000 | [−0.629, 0.058] | 34 | 1.122 | 0.135 | −0.286 |
| **S2**: I found the face-to-face/AIvaluate assessment unnecessarily complex. [a] | 4.000 | 4.057 | 0.057 | 0.948 | [−0.268, 0.383] | 34 | −0.251 | 0.402 | 0.060 |
| **S3**: I thought the face-to-face/AIvaluate assessment was easy to participate in/use. | 4.171 | 4.371 | 0.200 | 0.789 | [−0.071, 0.471] | 34 | −1.000 | 0.162 | 0.254 |
| **S4**: I think that I would need the support of a teacher/technical person to be able to understand/use the face-to-face/AIvaluate assessment.[a] | 3.286 | 3.314 | 0.029 | 1.235 | [−0.396, 0.453] | 34 | −0.087 | 0.466 | 0.023 |
| **S5**: I found the various aspects/functions of the face-to-face/AIvaluate assessment were well integrated. | 3.914 | 4.229 | 0.314 | 0.800 | [0.039, 0.589] | 34 | −1.540 | 0.066 | 0.393 |
| **S6**: I thought there was too much inconsistency in the face-to-face/AIvaluate assessment.[a] | 3.543 | 4.000 | 0.457 | 0.791 | [0.186, 0.729] | 34 | −2.847 | 0.004* | 0.578 |
| **S7**: I would imagine that most students would learn to participate in/use the face-to-face/AIvaluate assessment very quickly. | 4.057 | 4.000 | −0.057 | 0.937 | [−0.379, 0.265] | 34 | 0.243 | 0.405 | −0.061 |
| **S8**: I found the face-to-face/AIvaluate assessment very cumbersome.[a] | 4.086 | 4.229 | 0.143 | 0.817 | [−0.138, 0.423] | 34 | −0.682 | 0.250 | 0.175 |
| **S9**: I felt very confident during the face-to-face/AIvaluate assessment. | 4.029 | 3.829 | −0.200 | 0.963 | [−0.531, 0.131] | 34 | 0.793 | 0.216 | −0.208 |
| **S10**: I needed to learn a lot of things before I could get going with the face-to-face/AIvaluate assessment.[a] | 3.400 | 4.000 | 0.600 | 1.093 | [0.225, 0.975] | 34 | −2.177 | 0.018* | 0.549 |
| **Students' experience (additional items)** | | | | | | | | | |
| **A1**: The face-to-face/virtual meeting using AIvaluate was helpful in assessing the student's work. | 4.000 | 4.143 | 0.143 | 0.839 | [−0.146, 0.431] | 34 | −0.588 | 0.280 | 0.170 |
| **A2**: I felt I had a good understanding of the student's emotional state during the face-to-face/virtual meeting using AIvaluate. | 4.114 | 4.200 | 0.086 | 0.855 | [−0.208, 0.379] | 34 | −0.386 | 0.351 | 0.100 |

**Table 3** (continued)

| SUS Items | Mean F2F | Mean AI | Mean Diff | Pooled SD | 95% CI | Df | t | *p* (1-tail) | Co-hen's d |
|---|---|---|---|---|---|---|---|---|---|
| **A3**: The face-to-face/virtual meeting using AIvaluate was effective in providing pastoral care. | 3.886 | 3.629 | −0.257 | 1.116 | [−0.641, 0.126] | 34 | 1.055 | 0.149 | −0.230 |
| **A4**: The face-to-face/virtual meeting using AIvaluate did not add to my workload significantly. | 3.829 | 4.371 | 0.543 | 0.854 | [0.250, 0.836] | 34 | −2.801 | 0.004* | 0.636 |

A1–A4 bespoke items presented in addition to the 10 standard SUS items to evaluate students' experience

[a] Responses of negative items were reversed to align with positive items, higher scores indicate positive responses

* Indicates statistically significant <= 0.05 confidence level

workload. For items A1–A3, no statistically significant differences were found, with both formats receiving comparable ratings. However, for items A1 and A2, AIvaluate did score slightly higher indicating a marginally more positive experience for teachers in assessing student work and understanding students' emotional states. Notably, item A3 (pastoral care effectiveness) showed a small-to-moderate effect size ($d = -0.230$), and although not statistically significant, the wide confidence interval [−0.641, 0.126] suggests variability in participant views on AIvaluate's suitability for pastoral support.

Overall, while these results suggest a slightly more favourable perception of AIvaluate, most differences were not statistically significant. Moreover, usability ratings do not equate to reduced workload or improved pedagogical outcomes and should be considered indicative rather than definitive.

### 4.3 Teacher preferences

The third research question aimed to identify teachers' preferences for AIvaluate assessments over traditional face-to-face assessments, or vice versa. The thematic analysis of the interviews revealed a number of high-level themes and associated sub-themes as illustrated in Fig. 10.

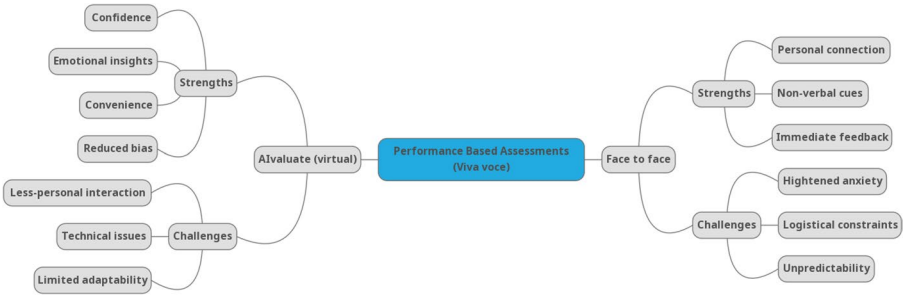


**Fig. 10** Thematic mind map of face-to-face and AIvaluate themes and sub-themes

#### 4.3.1 AIvaluate: strengths

The following sub-themes illustrate the perceived strengths of the AIvaluate system, derived from patterns in participant reflections on usability, emotional support, and fairness. These codes clustered into a broader theme around how the AI-enhanced viva format positively shaped the assessment experience.

Confidence.

This sub-theme captures how AI-generated prompts enhanced both teacher and student confidence. Coded segments consistently highlighted how automation supported questioning depth, especially when teachers felt less secure in subject knowledge. Participants appreciated the system's capacity to generate targeted prompts and questions, which supported teachers during the assessment process. By offering a structured way to probe students' understanding, AIvaluate helped teachers delve deeper into the student's responses without having to devise all questions on the spot.

> "It made me ask questions which I may not have thought of…" (Participant 1)

> "Generated responses to the student's reply were mostly useful and time saving."(Participant 2)

System-generated prompts also proved especially beneficial in cases where teachers felt less confident in a subject area. This capability indicated that AIvaluate's prompts not only guided discussions but also supported teacher confidence, ensuring a more structured and comprehensive assessment experience.

> "Some of the suggested questions could be good for the evaluator if they are not confident in the subject area." (Participant 10)

> "The suggested replies were excellent! I would have struggled to perform this assessment without these…" (Participant 22)

**Emotional insights** This sub-theme reflects codes that described the system's ability to detect and share student emotional states. Participants interpreted this as enhancing emotional awareness and enabling more sensitive or supportive interactions during the assessment. A notable strength of AIvaluate was its ability to provide insights into a student's emotional state. Participants found this feature useful in gauging student wellbeing, which, in a face-to-face environment, could sometimes be difficult to interpret with any degree of certainty.

> "The system was able to read the expression of the student which was surprisingly accurate." (Participant 8)

> "The emotional insight of the student was great, I didn't get this from the student in the face to face meeting." (Participant 20)

Such insights allowed teachers to tailor their follow-up questions, or offer supportive comments, during the assessment. These reflections illustrate how AIvaluate can serve as a tool that addresses not just academic performance but also emotional wellbeing during assessments.

"Having an insight into the student emotions was also great…" (Participant 31)

"The feedback on the student's emotional state was interesting…" (Participant 34)

**Convenience** This theme groups responses related to the system's practicality and ease of use. Codes frequently referenced the platform's intuitive design, speed of deployment, and flexibility, indicating that these affordances contributed to overall efficiency. Participants reported that the virtual format of AIvaluate offered practical advantages that streamlined the assessment process. Quick setup times, minimal training requirements, and greater flexibility in scheduling were among the benefits mentioned.

"It was very easy to use…" (Participant 1)

"It didn't require any training and can be conducted by anyone after a short briefing." (Participant 19)

Teachers also noted that the ability to conduct assessments remotely made the process more efficient and accessible. Hence, AIvaluate's usability and flexibility allowed teachers to focus on qualitative feedback rather than administrative or logistical constraints.

"The virtual felt quicker, could do it from anywhere." (Participant 16)

"The interface was intuitive and user-friendly; there were minimal delays…" (Participant 25)

"It should save time, increasing the opportunity to provide more contact time." (Participant 17)

**Reduced bias** This sub-theme captures perspectives on AIvaluate's perceived objectivity. Participants associated its standardised format with a reduction in emotional and cognitive bias, which emerged as a strong sub-code within equity- and fairness-focused comments. AIvaluate was seen as offering a more standardised assessment environment, particularly useful in reducing subjective biases. Some participants appreciated how the system's 'clinical' approach ensured that a teacher's emotional state or preconceived notions would not influence the evaluation. This aspect high-

lighted AIvaluate's potential to deliver more consistent and fair evaluations, an important consideration for maintaining assessment objectivity.

> "AIvaluate offers an efficient, standardised approach that minimises subjective biases." (Participant 25)

> "The virtual system was positive in that it was clinical, therefore emotional state does not impair or twist assessment bearings." (Participant 11)

#### 4.3.2 AIvaluate: challenges

While AIvaluate was seen as innovative and supportive, several challenges were noted. These sub-themes reflect tensions around emotional connection, technical limitations, and responsiveness, drawn from participant narratives coded under 'empathy', 'functionality', and 'adaptability'.

**Less-personal interaction** This sub-theme encapsulates the emotional distance some participants felt in the AI-based setting. Codes clustered here frequently referenced lack of warmth, human presence, and interpersonal rapport. Some criticism of the system included the less personalised nature of the virtual assessment setting. Teachers mentioned that they could not immediately comfort students who appeared nervous or adapt to subtle emotional cues in the same way they might face-to-face.

> "It didn't feel personal and didn't allow me to calm an individual if they looked flustered." (Participant 4)

> "Lack of interaction. Felt impersonal." (Participant 5)

Others felt that the system's clinical nature and automated responses made the assessment feel more methodical and disconnected. These comments highlight how, despite the usefulness of automated prompts, AIvaluate could not fully replicate the empathy and rapport-building commonly associated with in-person evaluations.

> "It is impersonal and does not have that human touch." (Participant 17)

> "It felt quite robotic and lacked that human warmth." (Participant 33)

**Technical issues** Participants identified a number of technical challenges during assessment delivery, including delays, confusing outputs, and awkward pacing. These were coded consistently as disrupting engagement and conversational flow. Teachers also cited various technical issues that undermined the fluidity of the assessment. Delays in the system's responsiveness and the time it took to input or generate replies could create disruptions that hindered the natural flow of conversation.

"There is quite a bit of delay in responding as it takes time to type questions…" (Participant 3)

"Some open ended generated answers were confusing… The waiting time… had a disruptive character." (Participant 2)

Participants noted that waiting for students to type or the AI to process responses occasionally elongated sessions more than anticipated. Such pauses often broke engagement and required teachers to carefully manage the pacing of the assessment.

"The length of time it took the student to respond to questions." (Participant 6)

"I was waiting for the student to input data during some parts of the assessment." (Participant 14)

**Limited adaptability** This sub-theme synthesises codes related to the system's inflexibility when shifting topics or addressing spontaneous content. Comments reflected the system's tendency to stay within predefined structures rather than allowing organic discourse. Teachers expressed frustration with AIvaluate's limitations in steering conversations dynamically. Once a topic had been introduced, the system sometimes struggled to transition to new areas or adapt appropriately to more nuanced shifts.

"The AI responses carried on with the same topic… It would be good if one of the responses asked a different question to move the topic on." (Participant 18)

"It felt like being stuck in a bit of a loop. It was like the system didn't know when to move on…" (Participant 30)

Furthermore, teachers noted that branching off to explore related but distinct areas was less intuitive within AIvaluate's automated structure. These observations underline the need for adaptive functionalities that allow for more spontaneity and comprehensive exploration of the student's knowledge.

"If you want to branch off to a new area… they weren't appropriate suggestions." (Participant 21)

"… if the system relies solely on automated scoring or lacks flexibility for open-ended responses." (Participant 9)

#### 4.3.3 Face-to-face: strengths

Participants reflected positively on traditional face-to-face assessment formats, with recurring codes related to emotional presence, immediacy, and interpersonal connec-

tion. These themes reflect how in-person assessments leveraged human warmth and spontaneity.

**Personal connection** This sub-theme describes the sense of relational closeness teachers experienced when engaging students in person. Codes emphasised rapport, familiarity, and emotional visibility as key advantages. Teachers valued the opportunity to establish rapport with students, which they felt led to more open and trusting communication. Observing a student in a shared physical space was viewed as beneficial in fostering a supportive and empathetic environment.

> "Face to face was nice because it was more personal and you could see the reactions of the student." (Participant 1)

> "I preferred the face to face as it lets me see the student's reaction… more personal…" (Participant 4)

Building such rapport not only helped teachers understand students better but also made it easier to tailor feedback in real time. The face-to-face setting was seen as delivering this advantage for those educators that prioritised creating a comfortable assessment context for their students.

> "I can build/maintain a rapport with the student." (Participant 11)

> "I felt like I could get to know the student better and build a rapport due to the close proximity." (Participant 12)

**Non-verbal cues** This theme was constructed from codes referencing body language, facial expression, and tone. These were seen as crucial to dynamic conversation and allowed real-time adaptation in a way that participants felt AI could not yet replicate. Another strength was the enhanced ability to read a student's body language and facial expressions during the assessment, which provided real-time indicators of comprehension and comfort.

> "Being able to read the body language more and gain eye contact when asking questions." (Participant 4)

> "I get to read the face of the student and they tend to express and explain themselves more." (Participant 8)

These cues were also seen as contributing to a more natural conversational flow and helped teachers adjust their approach as needed. This immediate visual feedback enabled a more responsive and dynamic dialogue than what is typically possible with virtual formats.

"It was easier to see facial expressions… The conversation was more flowing…" (Participant 12)

"I love meeting face to face. I feel like I can register emotions more efficiently…" (Participant 24)

**Immediate feedback** This sub-theme includes reflections on the fluidity and spontaneity of in-person interactions. Coded content highlighted the ability to give real-time clarification, adjust tone, and co-create the pace of the conversation. Face-to-face assessments were deemed advantageous for the real-time feedback loop they provided. Teachers could spontaneously elaborate on, or clarify, questions, and students were able to respond or ask for guidance on the spot.

"Allows for a flow of conversation… Able to communicate more effectively as it involves body language and tone." (Participant 10)

"When meeting in person, I can provide personalized feedback and encouragement… immediate responses." (Participant 9)

It was felt that the ability to react spontaneously improved efficiency and also deepened the level of engagement that was p[possible between the assessor and the student. Overall, in-person conversations were regarded as lively, efficient, and more conducive to nuanced back-and-forth discussion.

"It was faster than the AI meeting as there was instant response." (Participant 18)

#### 4.3.4 Face-to-face: challenges

Despite the perceived emotional richness of face-to-face assessments, participants also raised concerns. These sub-themes encompass the cognitive load, scheduling challenges, and emotional unpredictability that can complicate in-person formats.

**Heightened anxiety** This sub-theme aggregates participant observations about how live, high-stakes interactions can elevate student anxiety. Coding frequently revealed concerns about students feeling rushed, judged, or exposed. Many participants observed that being physically present in front of a teacher could cause or exacerbate student anxiety. The urgency to speak coherently and respond without delay often made students more nervous than in a virtual format.

"Student felt under pressure to give answers quickly without thinking things through." (Participant 5)

"If I feel nervous or anxious… it may have affected how openly I could communicate." (Participant 9)

This anxiety was seen to sometimes compromise the quality of students' responses and overshadowed their true capabilities. Hence, teachers acknowledged that while face-to-face settings could foster personal connection, they also risked amplifying performance anxiety.

> "Students may be more nervous with face to face." (Participant 10)

> "It might make some people nervous meeting face to face if they struggle with social situations." (Participant 24)

**Logistical constraints** Codes under this theme addressed practical barriers, including scheduling, time pressure, and mental effort required by teachers. These constraints were especially pronounced in larger cohorts or multi-session formats. Teachers emphasised the practical challenges of arranging and conducting face-to-face sessions for all students, particularly regarding scheduling and time constraints.

> "Scheduling and conducting in-person meetings for every student can be inefficient…" (Participant 19)

> "It took time which is a factor to consider. Perhaps it also felt a bit rushed…" (Participant 30)

They also noted that in-person interactions required immediate thinking, on the part of the assessor, to sustain the conversation's momentum. Such constraints highlighted the logistical effort and mental demands that teachers must navigate during face-to-face assessments.

> "You can be in a vulnerable position… You cannot control how the other person interprets your body language." (Participant 17)

> "Student felt under pressure to give answers quickly… had to think on my feet about what to ask and steer the conversation." (Participant 5)

**Unpredictability** This sub-theme gathers insights on the emotional and conversational volatility of live assessments. Codes emphasised the need for quick teacher judgment in sensitive or off-script situations, often framed as taxing or risky. A challenge teachers encountered was centered on the unpredictable elements of in-person interactions. Teachers noted that spontaneous emotional responses or unexpected questions could derail an assessment plan or introduce discomfort.

> "Whether it was for summative or formative review, I do not know what was being assessed… Possibly more intimidating." (Participant 11)

> "It was more personal… could lead to further responses that might be unknown. Emotion can be involved… can overrun." (Participant 17)

Some teachers worried that certain sensitive topics, once raised, might require immediate and delicate handling. Thus, while in-person assessments often feel more natural, they also carry a heightened risk of emotional complexity or sudden shifts in direction, emphasising the need for adaptable facilitation skills on the part of the teacher.

> "Some of the content was potentially a little sensitive… student feel uncomfortable in the meeting." (Participant 27)

## 5 Discussion

The findings of this study provide valuable insights into the effectiveness, acceptability, and teacher preferences when using AIvaluate, an emotionally intelligent, LLM-augmented pedagogical AI conversational agent for performance-based assessments (PBAs). This section contextualises the results within the broader educational literature, discusses implications for reducing teacher burden, and identifies areas for future research.

A key aim of this study was to ascertain whether AIvaluate impacts teacher grading outcomes compared to traditional face-to-face PBAs. The results relating to *RQ1* revealed that students were graded significantly higher when assessed via AIvaluate, suggesting that a virtual, AI-augmented environment can positively influence teacher assessments. One justification could be AIvaluate's features that promote structured questioning and procedural consistency, potentially guiding teachers to evaluate a student's knowledge more systematically. These findings align with prior work indicating that advanced technological tools can enhance assessment fairness and reliability by reducing the influence of personal biases (Abbasian et al., 2024; Cherakara et al., 2023). However, while AIvaluate's automated prompts may increase consistency, there remains a need to ensure that teachers maintain critical oversight. PBAs rely on nuanced judgment for determining the depth and breadth of a student's mastery (Szulewski et al., 2023), and AI-driven suggestions must be used as a scaffold rather than a prescriptive determinant of final grades. Teachers in this study expressed appreciation for the structured prompts, however, they also highlighted the importance of professional expertise in evaluating performance. These observations are in line with existing literature, which emphasises the need for teacher autonomy and domain knowledge when using technology-based assessments (Maryadi et al., 2017; Wang et al., 2020).

The findings from *RQ2*, the System Usability Scale (SUS) indicated that AIvaluate achieved a slightly higher overall SUS score than traditional face-to-face PBAs, reflecting an overall positive teacher experience. Particularly, AIvaluate was rated significantly higher on items relating to inconsistency and the preparation effort required. In particular, teachers felt that AIvaluate's structured assessment interface, and emotionally intelligent suggested prompts, reduced the mental load associated with real-time questioning, confirming prior research that highlights the potential for AI tools to alleviate the administrative and logistical burdens of PBAs (Adamopoulou & Moussiades, 2020; Alshumaimeri & Alshememry, 2024). The additional

bespoke items in the SUS suggested that AIvaluate added less to teachers' workloads compared to face-to-face assessments. While the additional SUS item (A4) suggests that teachers perceived AIvaluate to add less to their workload than traditional assessments, this interpretation must be approached cautiously. The data derives from a single-item subjective scale, which limits its capacity to offer a holistic view of workload. Future research should further investigate this perception with additional metrics such as time taken for assessment preparation and grading, physiological or self-reported stress indicators, and validated cognitive load measures (Paas et al., 2003). Such exploration would provide a more nuanced and evidence-based understanding of teacher workloads in AI-supported PBAs. Further, while PBAs are often praised for their pedagogical strengths, such as authenticity and real-time demonstration of skills, many teachers report increased administrative overhead (Arter & McTighe, 2000; Wiggins et al., 2005). In this study, teachers appreciated how AIvaluate centralised tasks (e.g., generating questions, displaying student emotions, automating partial feedback), thereby minimising repeated or routine work. However, technical issues, such as response delays or speech-to-text inaccuracies, were sometimes mentioned as drawbacks, consistent with literature emphasising the importance of robust system performance in technology-enhanced assessments (Padmasiri et al., 2023).

The thematic analysis relating to *RQ3* provided additional insights around the multifaceted reasons for teachers' preferences for AIvaluate or face-to-face PBA formats. Teachers perceived AIvaluate as advantageous due to its helpful, automated prompts (particularly where they lacked subject expertise), real-time emotional insights, and the convenience of remote assessment. Consistent with the existing literature, tools that can adapt or extend teachers' capacities appear pivotal in reducing workload stressors without compromising the integrity of PBAs (Barber & Phillips, 2000; Ellis et al., 2015). Despite these strengths, the less-personal interaction and occasional technical glitches served as key limitations. Some teachers voiced concerns that AIvaluate lacked a certain human element critical for building rapport with students, an aspect that remains central to teaching and learning (Wiggins et al., 2005). These perceptions echo prior findings that while AI-based tools can streamline certain tasks, they cannot fully replicate the empathy and adaptability offered by a human assessor (Gonda & Chu, 2019). Similarly, participants noted AIvaluate's limited adaptability when attempting to pivot to new conversation topics or address unexpected student responses. Face-to-face assessments, on the other hand, were commended for enabling deeper emotional engagement, reading of non-verbal cues, and immediate responsiveness, factors crucial for nuanced formative feedback and pastoral support. Nonetheless, teachers also highlighted the anxiety-induced environment in live, in-person settings and the logistical complexities. These findings mirror prior research that reaffirms the two-sided nature of face-to-face assessments; while beneficial for real-time learning, the format can burden teachers with higher stress and resource demands (Gallardo, 2020; Szulewski et al., 2023).

Despite encouraging findings, several limitations warrant consideration. First, the sample size was modest and drawn from a single institutional context, limiting generalisability. Second, workload reductions were inferred from perception data rather than objective workload indicators such as grading time, teacher stress, or task complexity. Third, while AIvaluate supported structured interactions, it lacked the adapt-

ability and emotional nuance of face-to-face communication, which some teachers found impersonal. Furthermore, technical delays occasionally disrupted assessment flow, highlighting the need for improved system responsiveness. These limitations suggest that while AIvaluate may offer partial relief from certain aspects of teacher burden, it should be viewed as a complement rather than a replacement for traditional assessment formats. Future studies should include mixed methods measures of teacher cognitive load, stress levels, and assessment design effort to assess actual reductions in burden. Additionally, several participants highlighted the lack of human warmth and relational presence in the AIvaluate assessments, describing the experience as robotic or impersonal. While the tool facilitated structured questioning and emotional data reporting, it was not capable of replicating the dynamic empathy, spontaneity, or rapport-building commonly associated with face-to-face interactions. This perceived emotional distance may reduce teacher-student connection, a factor critical to assessment engagement and authenticity.

These findings carry important theoretical implications that can be interpreted through the lens of emotion theories, the Technology Acceptance Model, and pedagogical frameworks for AI integration. From an emotion-theoretic perspective, the observed benefits of real-time emotional insight align with appraisal-based models (Lazarus, 1991) and Control-Value Theory (Pekrun, 2006), both of which highlight how learners' perceived control and emotional value influence cognitive performance. AIvaluate's emotion-aware design allowed teachers to better modulate affective dynamics during assessment, potentially enhancing student composure and cognitive engagement. From a technology adoption viewpoint, the positive usability scores and reduced workload perception reflect key constructs of the Technology Acceptance Model, namely; perceived usefulness and perceived ease of use (Davis, 1989; Venkatesh & Davis, 2000). Teachers' acceptance of AIvaluate appeared rooted not only in its functional efficacy but also in its emotional intelligence and contextual adaptability, suggesting an evolved form of technology acceptance that encompasses affective affordances. Finally, the findings reinforce the need for AI systems to be pedagogically aligned, supporting core principles of the TPACK model (Mishra & Koehler, 2006) and newer frameworks for AI in education (Holmes et al., 2019). By scaffolded automation, affective responsiveness, and alignment with assessment goals, AIvaluate demonstrates how emotionally intelligent AI can complement rather than replace the pedagogical agency of teachers.

The findings suggest several implications for utilising emotionally intelligent, LLM-augmented AI conversational agents to support teachers. Practical considerations for institutions and educators considering the integration of an AI conversational agents into PBAs to lessen teacher burden are presented in Table 4, exploring future recommendations to build more enhanced agents.

RR1. **Streamlined logistics**: By automating aspects of assessment design, such as providing context-specific question suggestions, AIvaluate can help teachers focus on higher-order tasks like evaluating depth of understanding. This automation partially addresses the logistical burdens often noted in PBAs, reducing the cognitive load of constantly devising new prompts or consistently tracking student performance. Future agents having additional tools to automate logistical

**Table 4** Future recommendations

| # | Research Recommendation | Source |
|---|---|---|
| RR1 | Streamlined logistics | SUSA, ORSA: TI, LA |
| RR2 | Enhanced feedback and monitoring | TAG, ORSA: LI, IF, PC |
| RR3 | Reduced expert involvement | TAG, SUSA, ORSA: IF, PC, RB |
| RR4 | Authenticity | TAG, SUSA, ORSA: LI, IF, PC |
| RR5 | Training and infrastructure | SUSA, ORSA: TI, LA, LC |

RQ1 – Teacher-Assessed Grading (TAG)

RQ2 – Systems Usability Scale Analysis (SUSA)

RQ3 – Open Response Survey Analysis (ORSA): Technical Issues (TI), Limited Adaptability (LA), Less-personal Interaction (LI), Immediate Feedback (IF), Personal Connection (PC), Reduced Bias (RB), Logistical Constraints (LC)

and administrative burden on teachers, such as task automation, scheduling and integration with other in-school data systems, would allow teachers to further benefit from reduced workloads.

RR2. **Enhanced feedback and monitoring**: Teachers in this study benefited from AIvaluate's real-time emotional state detection, gaining additional insight into student wellbeing without having to rely solely on subjective observations. This function could help teachers tailor support or scaffolding in the moment, potentially lessening the burden of extensive post-assessment intervention. However, teachers must still exercise discretion, as the detection of anxiety or discomfort does not replace the nuanced understanding that comes from personal rapport and professional judgment. Future agents having feedback and marking capabilities, including the ability to assess uploaded student work before the PBA would give agents an in-depth ability to cater for student needs, thereby reducing teacher burden.

RR3. **Reduced expert involvement**: Many PBAs require consultation with specialists or collaboration among multiple educators, especially in cross-curricular or competency-based evaluations. The LLM-augmented prompts in AIvaluate can offer baseline subject guidance, relieving teachers from having to source expert input for every student query. Although this can reduce some of the coordination overhead, it is vital that educators validate AI-provided content to maintain academic rigor. Future agents allowing the upload of mark schemes, rubrics, metrics, curriculum grading mechanisms and syllabus data would allow agents to provide a more specific and tailored experience, reducing the need for teachers to have wider curricula knowledge.

RR4. **Authenticity**: While AIvaluate's prompts and grading support can standardise certain aspects of PBAs, authenticity must remain a priority, particularly given concerns around generative AI's potential to blur authorship boundaries. Teachers should leverage AI-driven suggestions as supportive scaffolds rather than definitive judgments, ensuring that professional expertise guides the final evaluation. Future agents with the capability to link student responses with specific reading

literature related to the curriculum would further support teachers to make specific recommendations, such as texts and further reading, to students.

RR5. **Training and infrastructure:** Successful adoption of AIvaluate requires adequate training so that teachers can confidently interpret AI-generated insights and adapt them to differing subject contexts. Schools and institutions should also ensure reliable technical infrastructures, such as bandwidth for video-based emotion recognition and stable API connections to LLM services, to avoid disruptions that could undermine instructional goals. Where possible, pilot studies and phased rollouts can help educators refine local practices before broad-scale implementation. Future agents with in-system guidance and training would further support teachers to operate and ultimately accept new AI-driven ways of working.

A key limitation is the single-institution context and relatively small teacher-student sample, which may curtail the generalisability of the findings. Future studies could replicate the methodology across diverse educational settings (e.g., primary vs. tertiary, large vs. small class contexts) to evaluate the scalability and broader applicability of AI conversational agents for PBAs. Additionally, while teachers' workload perceptions were positive overall, more granular metrics, such as time audits and cognitive load analyses, could generate deeper insights into precisely how and where workload reductions occur. Moreover, although teachers appreciated the automated prompt generation, the risk of over-reliance on AI suggestions raises questions about sustaining teacher autonomy and professional judgment. Investigations into the interplay between human expertise and AI guidance are warranted to ensure that PBAs remain robust, rigorous, and student-centred. Finally, longitudinal research examining whether improvements in teacher satisfaction and reduced burden are sustained over multiple PBA cycles is recommended to establish the long-term viability of AI-driven assessment tools. Future implementations of emotionally responsive AI in education must continue to foreground ethical design, especially in relation to fairness, transparency, and psychological safety.

## 6 Conclusion

This study examined the ability of AIvaluate, an emotionally intelligent LLM-augmented pedagogical AI conversational agent, to reduce teacher burden in performance-based assessments (PBAs). The findings indicate that AIvaluate provides some advantages over traditional face-to-face assessments in terms of reducing teacher burden, enhancing procedural consistency, and supporting the evaluation process through structured, emotionally intelligent interactions. Teachers reported that AIvaluate streamlined assessment logistics, reduced the mental effort of on-the-spot questioning, and offered real-time emotional insights into students' states, which enhanced their ability to provide tailored support during assessments. These benefits translated into higher student grades and improved teacher satisfaction compared to traditional methods, as reflected in the higher System Usability Scale (SUS) scores for AIvaluate.

Key strengths of AIvaluate included its capacity to generate targeted, adaptive prompts, offer insights into student emotions, and deliver a remote, flexible assessment solution. These features addressed some of the inherent challenges of PBAs, such as the complexity of assessment design and the cognitive demands placed on educators. Teachers valued the system's potential to reduce subjectivity, enhance procedural fairness, and provide a structured framework for conducting assessments. However, limitations were also identified, including the impersonal nature of AI-driven interactions and occasional technical issues, such as response delays or limited adaptability to unexpected conversational shifts. In contrast, face-to-face assessments were praised for their ability to foster personal connections, leverage non-verbal cues, and facilitate dynamic, interactive dialogues. These features allowed teachers to build rapport and provide immediate, nuanced feedback, which was particularly beneficial for formative assessments. However, the heightened anxiety reported by students in face-to-face settings, coupled with the logistical and cognitive demands on teachers, highlighted the need for alternative or hybrid approaches to PBAs.

The findings suggest that AIvaluate holds potential as a complementary tool for modern educational practices. By alleviating some of the logistical, administrative, and emotional burdens associated with PBAs, AI conversational agents like AIvaluate can help educators focus on higher-order tasks, such as evaluating depth of understanding and fostering meaningful student engagement. The study highlights the importance of balancing the strengths of AI-driven and traditional assessment formats, advocating for hybrid frameworks that integrate the procedural efficiency of AI with the empathetic, adaptive qualities of human interaction. Future research should explore the scalability and adaptability of emotionally intelligent, LLM-augmented conversational agents across diverse educational contexts and disciplines. Investigating the long-term effects of such systems on teacher workload, student performance, and educational equity will be critical for understanding their broader impact. The development of more robust, adaptive features and comprehensive training programs will further support the integration of AI-driven tools into the assessment landscape. Ultimately, AIvaluate exemplifies the transformative potential of innovative technology to enhance the sustainability, inclusivity, and effectiveness of performance-based assessments in education. While AIvaluate shows early promise as a supportive tool in PBA settings, its effectiveness must be corroborated through longitudinal research that explores actual workload metrics, teacher well-being, and student learning outcomes across varied educational environments. Furthermore, the use of AIvaluate comes with trade-offs in relational immediacy and human connection. As such, it may serve best as a supplementary tool rather than a full replacement for traditional assessments, pending further development to enhance emotional responsiveness.

**Acknowledgements** The authors would like to thank all of the participants that took part in this study and Brunel University for providing the facilities and materials necessary to carry out the interactive sessions with participants.

**Authors' contributions** AM, HY and DDZ conceived the study. HY and AM planned the study and formulated the experiment design. HY developed the application. HY performed the user trials. AM and HY were responsible for data analysis and writing the manuscript. HY, AM and DDZ reviewed and edited subsequent drafts of the manuscript.

**Funding** No funding was obtained for this study.

**Data Availability** The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Ethics approval and consent to participate** This research was approved by Brunel University of London Research Ethics Committee. Informed consent to participate was obtained from each participant through signing of informed consent forms. Participants were informed of their right to withdraw from the study at their own will without any repercussions at any point during or after the study.

**Consent for publication** Not applicable.

**Competing interests** The authors declare that they have no competing interests.



## References

Abbasian, M., Yang, Z., Khatibi, E., Zhang, P., Nagesh, N., Azimi, I., Jain, R., & M Rahmani, A. (2024). Knowledge-Infused LLM-Powered conversational health agent: A case study for diabetes patients. ArXiv Preprint arXiv: 2402 10153. https://doi.org/10.48550/arXiv.2402.10153

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Alser, M., & Waisberg, E. (2023). Concerns with the usage of ChatGPT in academia and medicine: A viewpoint. *American Journal of Medicine Open*, *9*, 100036.

Alshumaimeri, Y. A., & Alshememry, A. K. (2024). The extent of AI applications in EFL learning and teaching. *IEEE Transactions on Learning Technologies*, *17*, 653–663. https://doi.org/10.1109/tlt.2023.3322128

Archbald, D. A., Newmann, F. M., I., M. W., & A., R. V. (1988). *Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School*. https://go.exlibris.link/zjTKDF7d. Accessed 18 Apr 2025.

Arter, J. A., & McTighe, J. (2000). Scoring rubrics in the classroom. Using Performance Criteria for Assessing and Improving Student Performance. https://uk.sagepub.com/en-gb/eur/book/scoring-rubrics-classroom

Baker, E. L. (1997). Model-based performance assessment. *Theory into Practice*, *36*(4), 247–254. https://doi.org/10.1080/00405849709543775

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *J Usability Studies*, *4*(3), 114–123.

Barber, M., & Phillips, V. (2000). Big change questions: Should Large-Scale assessment be used for accountability?? The fusion of pressure and support. *Journal of Educational Change*, *1*(3), 277–281. https://doi.org/10.1023/a:1010064308012

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2012). *Thematic analysis*. American Psychological Association.

Braun, V., & Clarke, V. (2013). *Successful qualitative research. A practical guide for beginners*. https://uk.sagepub.com/en-gb/eur/successful-qualitative-research/book233059

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport Exercise and Health*, *11*(4), 589–597.

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238

Brooke, J. (1996). *SUS: A quick and dirty usability scale. Usability evaluation in industry*. https://www.taylorfrancis.com/chapters/edit/10.1201/9781498710411-35/sus-quick-dirty-usability-scale-john-brooke

Brooke, J. (2013). SUS: A retrospective. *J Usability Studies*, *8*(2), 29–40.

Brooks, J. L. (2012). Counterbalancing for serial order carryover effects in experimental condition orders. *Psychological Methods*, *17*(4), 600.

Brown, M. (2023). *Teachers' experiences of participation in performance appraisal in an English Academy School: navigating bureaucratic compliance and professional accountability*. University of Sussex.

Byrne, D. (2022). A worked example of Braun and clarke's approach to reflexive thematic analysis. *Quality & Quantity*, *56*(3), 1391–1412. https://doi.org/10.1007/s11135-021-01182-y

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*(1), 18–37.

Cherakara, N., Varghese, F., Shabana, S., Nelson, N., Karukayil, A., Kulothungan, R., Farhan, M. A., Nesset, B., Moujahid, M., & Dinkar, T. (2023). Furchat: An embodied conversational agent using llms, combining open and closed-domain dialogue with facial expressions. *ArXiv Preprint arXiv: 2308 15214*. https://doi.org/10.48550/arXiv.2308.15214

Cotton, D., Cotton, P., & Shipway, J. (2023). Chatting and cheating. Ensuring academic integrity in the era of ChatGPT. EdArXiv. Preprint posted online January, 10. https://doi.org/10.1080/14703297.2023.2190148

D'Amore, S., Maurisse, A., Gubello, A., & Carone, N. (2023). Stress and resilience experiences during the transition to parenthood among Belgian lesbian mothers through donor insemination. *International Journal of Environmental Research and Public Health*, *20*(4). https://doi.org/10.3390/ijerph20042800

D'mello, S., & Graesser, A. (2013). AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems, 2*(4), 23. https://doi.org/10.1145/2395123.2395128

Darling-Hammond, L. (1994). Performance-Based assessment and educational equity. *Harvard Educational Review*, *64*(1), 5–31. https://doi.org/10.17763/haer.64.1.j57n353226536276

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319–340. https://doi.org/10.2307/249008

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/N19-1423

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., & Ahuja, M. (2023). Opinion paper: So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology*, *13*, 100060.

Ekman, P., & Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk.

Ellis, L., Marston, C., Lightfoot, J., Sexton, J., Byrnes, J., Ku, H. Y., & Black, K. (2015). Faculty Professional Development in Student Learning Assessment: The Assessment Leadership Institute. *Research and Practice in Assessment*, *10*. https://www.proquest.com/docview/1829513087

Ernst, J. V. (2008). Analysis of cognitive and performance assessments in an engineering/technical graphics curriculum. *Journal of sTEm Teacher Education*, *45*, 7.

Gallardo, K. (2020). Competency-Based assessment and the use of Performance-Based evaluation rubrics in higher education: Challenges towards the next decade. *PROBLEMS OF EDUCATION IN THE 21ST CENTURY*, *78*(1), 61–79. https://doi.org/10.33225/pec/20.78.61

Gasiokwu, P. I., Oyibodoro, U. G., & Nwabuoku, M. O. I. (2025). GDPR Safeguards for Facial Recognition Technology: A Critical Analysis. https://doi.org/10.47857/irjms.2025.v06i01.02025

Gonda, D. E., & Chu, B. (2019). Chatbot as a learning resource? Creating conversational bots as a supplement for teaching assistant training course. *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*. https://doi.org/10.1109/tale48000.2019.9225974

Gonda, D. E., Luo, J., Wong, Y. L., & Lei, C. U. (2018). Evaluation of Developing Educational Chatbots Based on the Seven Principles for Good Teaching. 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE).

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education promises and implications for teaching and learning*. Center for Curriculum Redesign.

Kabbar, E., & Barmada, B. (2024). Assessment validity in the era of generative AI tools. http://www.unitec.ac.nz/epress/wp-content/uploads/2024/07/05-CITRENZ2023-Proceedings-Kabbar-Barmada.pdf

Khan, W. N. (2024). Ethical challenges of AI in education: Balancing innovation with data privacy. *Journal of AI Integration in Education*, *1*(1), 1–13.

Kollias, D., & Zafeiriou, S. (2018). Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm. *2018 International Joint Conference on Neural Networks (IJCNN)*. https://doi.org/10.1109/IJCNN.2018.8489340

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, *25*(9), 1248–1258. https://doi.org/10.1093/jamia/ocy072

Latif, E., Chen, Y., Zhai, X., & Yin, Y. (2024). Human-Centered Design for AI-based Automatically Generated Assessment Reports: A Systematic Review. arXiv preprint arXiv:2501.00081.

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Le Cunff, A. L., Giampietro, V., & Dommett, E. (2024). Neurodiversity and cognitive load in online learning: A systematic review with narrative synthesis. *Educational Research Review*, 100604. https://doi.org/10.1016/j.edurev.2024.100604

Lee, Y. C., & Fu, W. T. (2019). Supporting peer assessment in education with conversational agents. *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*. https://doi.org/10.1145/3308557.3308695

Longmuir, F., & McKay, A. (2024). Teachers workload strain: Considering the density as well as the quantity of teachers work. *Curriculum Perspectives*, *44*(4), 561–565.

Maryadi, J. A., Santoso, H., & Isa, Y. K. (2017). Development of personalized pedagogical agent for student-centered e- learning environment. *2017 7th World Engineering Education Forum (WEEF)*. https://doi.org/10.1109/weef.2017.8467028

Mayer, J. D., Salovey, P., & Caruso, D. R. (2008). Emotional intelligence: New ability or eclectic traits? *American Psychologist*, *63*(6), 503.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, *108*(6), 1017–1054.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18–31.

Newell, S. J. (2023). Employing the interactive oral to mitigate threats to academic integrity from ChatGPT. *Scholarship of Teaching and Learning in Psychology*. https://psycnet.apa.org/doi/10.1037/stl0000371

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4.

Padmasiri, P., Kalutharage, P., Jayawardhane, N., & Wickramarathne, J. (2023). AI-driven user experience design: Exploring innovations and challenges in delivering tailored user experiences. *2023 8th International Conference on Information Technology Research (ICITR)*. https://doi.org/10.1109/ICITR61062.2023.10382802

Pearce, J., & Chiavaroli, N. (2023). Rethinking assessment in response to generative artificial intelligence. *Medical Education*, *57*(10), 889–891.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*, 315–341.

Petrovica, S., & Ekenel, H. K. (2016). *Emotion Recognition for Intelligent Tutoring*. BIR Workshops.

Picard, R. W. (2000). Affective Computing. https://doi.org/10.7551/mitpress/1140.001.0001

Popham, W. J., Association for, S., Curriculum Development AVA. (2001). *The Truth about Testing: An Educator's Call to Action*. Association for Supervision and Curriculum Development.

Potter, B. S., Ernst, J. V., & Glennie, E. J. (2017). Performance-based assessment in the secondary STEM classroom: Performance-based assessments provide a vehicle to demonstrate student procedural knowledge as well as higher-order thinking abilities.(feature article). *Technology and Engineering Teacher*, *76*(6), 18. https://go.exlibris.link/06tjXt9s

Proudfoot, K. (2023). Inductive/Deductive hybrid thematic analysis in mixed methods research. *Journal of Mixed Methods Research*, *17*(3), 308–326. https://doi.org/10.1177/15586898221126816

Richmond, T., & Regan, E. (2023). Examining exams. https://smartthinking.org.uk/report/examining-exams/

Rudolph, J., Ismail, M. F. B. M., & Popenici, S. (2024). Higher education's generative artificial intelligence paradox: The meaning of chatbot mania. *Journal of University Teaching and Learning Practice*, *21*(6), 1–35.

Saihi, A., Ben-Daya, M., Hariga, M., & As' ad, R. (2024). A structural equation modeling analysis of generative AI chatbots adoption among students and educators in higher education. *Computers and Education: Artificial Intelligence*, *7*, 100274.

Shorey, S., Ang, E., Yap, J., Ng, E. D., Lau, S. T., & Chui, C. K. (2019). A virtual counseling application using artificial intelligence for communication skills training in nursing education: Development study. *Journal of Medical Internet Research*, *21*(10), e14658. https://doi.org/10.2196/14658

Souppez, J. B. R., Goswami, D., & Yuen, J. (2023). Assessment and feedback in the generative AI era: Transformative opportunities, novel assessment strategies and policies in higher education. *International Federation of National Teaching Fellows Symposathon 2023*. https://research.aston.ac.uk/en/publications/assessment-and-feedback-in-the-generative-ai-era-transformative-o

Steinfeld, A., Jenkins, O. C., & Scassellati, B. (2009). The oz of wizard: simulating the human for interaction research. *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. https://doi.org/10.1145/1514095.1514115

Szulewski, A., Braund, H., Dagnone, D. J., McEwen, L., Dalgarno, N., Schultz, K. W., & Hall, A. K. (2023). The assessment burden in Competency-Based medical education: How programs are adapting. *Academic Medicine*, *98*(11), 1261–1267. https://doi.org/10.1097/acm.0000000000005305

Tafazoli, D., & Meihami, H. (2023). Narrative inquiry for CALL teacher Preparation programs amidst the COVID-19 pandemic: Language teachers' technological needs and suggestions. *Journal of Computers in Education*, *10*(1), 163–187. https://doi.org/10.1007/s40692-022-00227-x

Taneja, K., Maiti, P., Kakar, S., Guruprasad, P., Rao, S., & Goel, A. K. (2024). Jill Watson: A virtual teaching assistant powered by ChatGPT. *International Conference on Artificial Intelligence in Education*. https://doi.org/10.48550/arXiv.2405.11070

Teng, D., Wang, X., Xia, Y., Zhang, Y., Tang, L., Chen, Q., Zhang, R., Xie, S., & Yu, W. (2024). Investigating the utilization and impact of large Language model-based intelligent teaching assistants in flipped classrooms. *Education and Information Technologies*, 1–34. https://doi.org/10.1007/s10639-024-13264-z

Teo, T. (2011). Factors influencing teachers' intention to use technology: Model development and test. *Computers & Education*, *57*(4), 2432–2440.

Tzirides, A. O., Saini, A., Zapata, G., Searsmith, D., Cope, B., Kalantzis, M., Castro, V., Kourkoulou, T., Jones, J., & da Silva, R. A. (2023). Generative AI: Implications and applications for education. arXiv preprint arXiv:2305.07605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. https://doi.org/10.48550/arXiv.1706.03762

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186–204.

Wang, Q., Jing, S., Camacho, I., Joyner, D., & Goel, A. (2020). Jill Watson SA: Design and evaluation of a virtual agent to build communities among online learners. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3334480.3382878

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural Language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wiggins, G., McTighe, J., Association for, S., & Curriculum, D. (2005). *Understanding by Design, Expanded 2nd Edition*. Association for supervision and curriculum development. https://www.pearson.com/en-us/subject-catalog/p/understanding-by-design-expanded-edition/P200000002022/9780131950849

Yusuf, H., Money, A., & Daylamani-Zad, D. (2025). Pedagogical AI conversational agents in higher education: A conceptual framework and survey of the state of the Art. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-025-10447-4

Zhang, B., Essl, G., & Mower Provost, E. (2016). Automatic recognition of self-reported and perceived emotion: Does joint modeling help? *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. https://doi.org/10.1145/2993148.2993173

## Authors and Affiliations

**Habeeb Yusuf[1] · Arthur Money[1] · Damon Daylamani-Zad[1]**

✉ Arthur Money
Arthur.Money@brunel.ac.uk

Habeeb Yusuf
Habeeb.Yusuf@brunel.ac.uk

Damon Daylamani-Zad
Damon.Daylamani-zad@brunel.ac.uk

[1] Department of Computer Science, Brunel University, Kingston Lane, Uxbridge UB8 3PH, UK