SINERGI KAWULA MUDA

Sinergi International Journal of Education

Check for updates

OPEN ACCESS

RECEIVED 09 January 2026
ACCEPTED 19 February 2026
PUBLISHED 31 August 2026

CITATION
Ramadhan FH, Hikmawan R, (2026). Development and Initial Validation of a Chatbot-Based Conversational Assessment Model for Post-Learning Reflective Affective Assessment. Sinergi International Journal of Education. 4 (3), 7-13.
doi: 10.61194/education.v4i3.1051

TYPE Original Research
PUBLISHED 31 August 2026
DOI 10.61194/education.v4i3.1051
VOL 4 Issue 3 August 2026

Development and Initial Validation of a Chatbot-Based Conversational Assessment Model for Post-Learning Reflective Affective Assessment

Fikriansyah Haikal Ramadhan[1], Rizki Hikmawan[2]
[12]Universitas Pendidikan Indonesia, West Java, Indonesia
Correspondence: hikmariz@upi.edu[2]

Abstract

Affective assessment remains challenging in educational practice due to its subjective nature and the difficulty of systematically documenting students' attitudes and values. Although affective learning outcomes are essential for holistic education, assessment practices often rely on teacher observations that may be influenced by personal judgment. This study aimed to develop and conduct an initial validation of Affectra, a chatbot-based conversational assessment model for reflective affective assessment following learning activities. Unlike most educational chatbots designed for learning support or cognitive assessment, Affectra integrates conversational assessment, adaptive probing, Large Language Model (LLM)-based analysis, and Krathwohl's affective taxonomy into a unified assessment framework. A modified Research and Development (R&D) approach was adopted, consisting of needs analysis, model and instrument design, expert validation, alpha testing, and system refinement. The model was developed for secondary school learning, with validation limited to expert review and researcher-controlled alpha testing. Content validity was evaluated using the Index of Item–Objective Congruence (IOC) with six experts in educational assessment and educational technology, while system performance was assessed through functional and scenario-based testing. The instrument achieved an overall IOC score of 0.81, indicating satisfactory content validity, and demonstrated consistent classification across Krathwohl's five affective levels under predefined simulated scenarios. These findings provide preliminary evidence supporting the initial development and internal validation of Affectra. Further classroom validation involving actual students is required to evaluate its practical effectiveness.

KEYWORDS

affective assessment; conversational assessment; chatbot; large language model; krathwohl's affective taxonomy.

Introduction

Following classroom instruction, teachers are expected to evaluate not only students' cognitive achievement but also their affective development, including attitudes, values, responsibility, and other internal dispositions. However, assessing these affective outcomes remains particularly challenging because they are not always directly observable once learning activities have concluded. In secondary school settings, teachers therefore often rely on classroom observations, students' participation, and professional judgment to infer affective development, despite these indicators not always reflecting students' underlying attitudes and values (Abrahams et al., 2019; Stephens & Ormandy, 2019). This challenge is also evident in Indonesian schools, where teachers continue to require practical and systematic affective assessment instruments to support more objective and consistent evaluation of students' attitudes (Meyliasari, 2024). Recent studies have further highlighted that teacher assessment remains vulnerable to inconsistency and variations in judgment due to differences in information processing, accountability, and assessment practices (Pit-ten Cate et al., 2020). These findings are

10.61194/ijjm.v4i3.1080

consistent with the earlier work of Jacob & Lefgren (2008), which demonstrated that evaluators experience greater difficulty distinguishing students with moderate rather than extreme performance levels, potentially resulting in inconsistent affective judgments. Consequently, more structured, transparent, and evidence-based approaches are needed to support reliable documentation and evaluation of students' affective development.

Despite its recognized role as an essential component of holistic education, affective assessment continues to receive considerably less emphasis than cognitive assessment in routine classroom practice (Kraft, 2019). Consequently, affective assessment is frequently conducted incidentally and remains difficult to monitor systematically over time. In many classrooms, teachers depend on indirect behavioral indicators, such as attendance, classroom participation, and observable conduct, which may not adequately represent students' genuine attitudes or values and may reinforce inconsistencies across evaluators (Aldrup et al., 2022; Ferreira et al., 2020; Stephens & Ormandy, 2019). These challenges are further intensified by administrative workload, limited instructional time, and the lack of practical assessment instruments capable of capturing students' authentic affective experiences in a consistent manner (Vistorte et al., 2024). Furthermore, previous research suggests that reflective accounts of learning experiences provide richer evidence of affective development than observable behaviors alone, highlighting the importance of creating opportunities for students to articulate their thoughts, values, and personal learning experiences (Rogers et al., 2017) therefore, there is a clear need for assessment approaches that not only assist teachers in monitoring affective development more systematically but also facilitate meaningful student reflection as part of the assessment process.

Recent advances in Artificial Intelligence (AI), particularly conversational chatbot technologies, offer new opportunities to address these challenges in educational assessment (P. Wang et al., 2025; Yusuf et al., 2025). Through sustained dialogic interactions, chatbots enable students to articulate their learning experiences, perspectives, and self-reflections in a more natural and continuous manner than conventional closed-ended assessment instruments (Allouch et al., 2021; Boyd et al., 2022; van der Schaaf et al., 2013). Studies on human–chatbot interaction further suggest that individuals are often more willing to engage in meaningful self-disclosure when interacting with chatbots due to perceived anonymity, reduced social judgment, and supportive conversational responses (Ho et al., 2018). These characteristics make conversational chatbots particularly suitable for facilitating reflective dialogue, allowing students to express underlying attitudes, values, and personal experiences that may not be readily observable through conventional classroom assessments. Such reflective interactions provide opportunities to capture evidence of students' affective development beyond observable behaviors and align closely with Krathwohl's taxonomy, which conceptualizes affective learning as a progressive internalization of attitudes and values through the levels of receiving, responding, valuing, organization, and characterization by value (Krathwohl et al., 1964; Stephens & Ormandy, 2019; Yun & Cho, 2022).

Despite these potential advantages, the use of AI-based chatbots for educational assessment also introduces important ethical and methodological considerations (UNESCO, 2023). Because reflective conversations may involve students' personal experiences, attitudes, values, and potentially sensitive disclosures, issues related to data privacy, student vulnerability, algorithmic bias, explainability of AI-generated interpretations, and appropriate teacher oversight should be carefully considered when integrating AI into assessment practices (UNESCO, 2022). Consequently, AI should be positioned as a decision-support technology that complements, rather than replaces, teachers' professional judgment in evaluating students' affective development (Holmes et al., 2022). Furthermore, existing studies indicate that affective assessment research and practice, particularly in K–12 education, remain largely concentrated on lower affective levels and continue to lack assessment approaches capable of comprehensively capturing reflective processes and the internalization of values across higher levels of Krathwohl's taxonomy (Yun & Cho, 2022). These limitations highlight the need for conversational assessment models capable of eliciting authentic student reflection while systematically interpreting affective responses across multiple levels of Krathwohl's taxonomy within a teacher-supervised assessment process.

Previous studies have demonstrated the growing adoption of conversational technologies to support educational processes. Conversational agents and AI-based chatbots have been widely implemented as instructional tools for tutoring, mentoring, personalized feedback, and learner support across diverse educational settings (Allouch et al., 2021; Yusuf et al., 2025). Complementing these findings, a recent systematic review of AI-based learning tools identified assessment and evaluation, personalized feedback, and intelligent tutoring as the three predominant educational roles of AI, with reported learning outcomes largely focusing on cognitive achievement, learner motivation, engagement, attitudes, and user perceptions (Luo et al., 2025). Meanwhile, research on dialogic reflection has shown that technology-supported dialogue enables learners to articulate multiple perspectives, engage in collaborative meaning-making, and develop deeper reflective thinking (Boyd et al., 2022; P. Wang et al., 2025). Collectively, these studies demonstrate the educational value of conversational AI and reflective dialogue but primarily emphasize instructional support, learning enhancement, and general educational outcomes rather than structured reflective affective assessment.

Despite these advances, empirical research on chatbot-based assessment has yet to integrate conversational interaction with a structured framework for evaluating students' reflective affective development. Existing studies generally assess affective outcomes in terms of attitudes, motivation, engagement, or learner perceptions, without systematically interpreting reflective responses according to progressive levels of Krathwohl's affective taxonomy. Furthermore, although AI has increasingly been employed for educational assessment, current applications mainly focus on evaluating learning performance, providing automated feedback, or supporting instructional decision-making rather than assisting teachers in conducting reflective affective assessment. Addressing these gaps, this study develops a chatbot-based conversational assessment model for post-learning reflective affective assessment in secondary education. Grounded in Krathwohl's affective taxonomy, the proposed model is designed to facilitate authentic reflective dialogue while supporting teachers through a structured interpretation of students' reflective responses across affective levels. Accordingly, this study extends the application of conversational AI beyond instructional support and general educational assessment by providing an initial validated model for reflective affective assessment in secondary education.

Methods

This study adopted a Research and Development (R&D) approach to develop and conduct an initial validation of a web-based chatbot for post-learning reflective affective assessment. The study followed the general principles of educational product development proposed by Borg et al. (2007), while the operational development procedure was adapted from the Educational Design Research (EDR) framework described by Q. Wang (2024). Borg et al. (2007) provided the overarching R&D philosophy of iterative product development, validation, and refinement, whereas Wang's EDR framework was selected because it is well suited to early-stage educational technology development through preliminary research, prototyping, formative evaluation, and iterative refinement without requiring large-scale field implementation. Consistent with the scope of this study, the research focused on the development and initial validation of the proposed assessment model through expert validation and simulated testing rather than implementation involving students. Qualitative descriptive analysis was employed to interpret expert feedback and scenario-based testing results in order to refine the conversational assessment model (Kim et al., 2017).

Following Wang's EDR framework, the development process was operationalized into five stages: (1) needs analysis, (2) model and instrument design, (3) expert validation, (4) alpha and scenario-based testing, and (5) system refinement. The needs analysis was conducted through a targeted narrative literature review, in which literature was purposively selected to inform the design of the conversational assessment model rather than to provide an exhaustive synthesis of existing evidence (Sukhera, 2022). The literature search was conducted between January 2017 and December 2025 using peer-reviewed publications and educational guidelines retrieved from Scopus-indexed journals, SpringerLink, ScienceDirect, ERIC, and Google Scholar. Representative search terms included "affective assessment," "Krathwohl taxonomy," "reflective learning," "educational chatbot," "conversational AI," "AI-assisted assessment," and "large language model" in various combinations. The review included peer-reviewed studies published in English that addressed educational assessment, affective learning, conversational AI, or educational technology, while editorials, opinion papers, duplicate publications, and studies unrelated to educational contexts were excluded. Retrieved publications were initially screened based on titles and abstracts, followed by full-text review to determine their relevance to the six predefined thematic areas.

The review covered six thematic areas: affective assessment and Krathwohl's affective taxonomy, reflective learning, educational chatbots and conversational AI, AI-assisted educational assessment, AI ethics in education, and educational product development. Evidence synthesized from these themes was translated into specific pedagogical,

$$IOC = \frac{\sum R}{N}$$

Equation 1. Index of Item–Objective Congruence Formula

Table 1. Translation of Literature Review Findings into Design Requirements

Literature Theme	Key Findings	Design Requirement	System Implementation
Affective assessment	Need for structured affective evidence	Affective assessment rubric	Krathwohl-based assessment rubric
Reflective learning	Reflection facilitates expression of attitudes and values	Open-ended reflective prompts	Reflection stage
Conversational AI	Natural dialogue improves engagement and self-disclosure	Multi-stage conversational flow	Conversational workflow
AI-assisted assessment	Context-aware interpretation is required	LLM-based affective interpretation	Gemini-based analysis
AI ethics	Human oversight and transparency are essential	Teacher-supervised assessment	Teacher review of AI-generated assessment
Educational product development	Iterative validation improves educational products	Expert validation and refinement	Iterative EDR development process

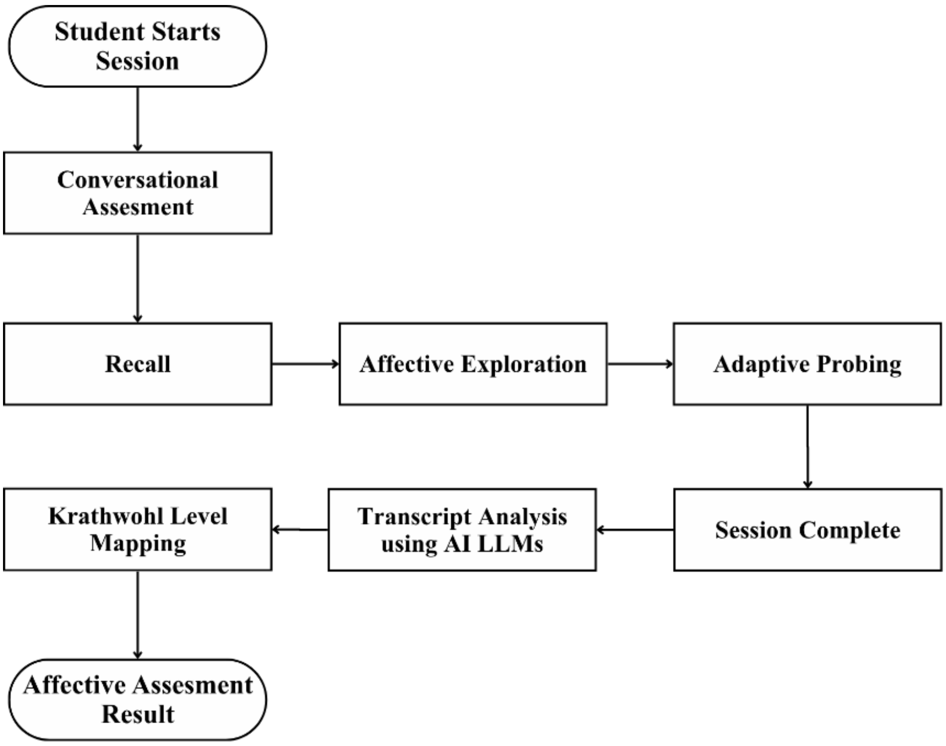

Figure 1. Conversational Assessment Workflow of the Chatbot System

assessment, conversational, ethical, and technical design requirements that guided the development of the proposed system. The relationship between the synthesized literature and the resulting design requirements is summarized in Table 1.

As the needs analysis was conducted through a targeted narrative literature review rather than a formal systematic literature review, the synthesis was intended to support educational product development rather than provide a comprehensive review of the literature. Consequently, the review may be subject to selection bias because studies were purposively selected to address specific design needs rather than identified through a formal systematic review protocol.

Based on the synthesized design requirements summarized in Table 1, a conceptual model of the chatbot and its conversational assessment workflow was designed. The assessment framework was grounded in Krathwohl's affective domain taxonomy, with reflective questions developed to elicit students' affective responses through natural dialogue (Rogers et al., 2017). Open-ended prompts and probing mechanisms were designed to facilitate self-reflection and encourage students to articulate their learning experiences and attitudes. In addition, an affective assessment rubric was developed to support the systematic interpretation of student responses according to the five levels of Krathwohl's affective taxonomy, ranging from receiving to characterization by value. The rubric considered both the depth of reflection and the degree of value internalization demonstrated in students' responses (Krathwohl et al., 1964).

The developed conversational assessment model and its accompanying assessment instruments were subsequently evaluated through expert validation to establish their content validity prior to system testing. A total of six experts participated in the validation process, comprising three experts in educational assessment and three experts in educational technology, whose expertise included educational assessment, post-learning assessment, artificial intelligence for STEM education, and AI in educational technology. Four validators were university lecturers holding master's degrees, while two held doctoral degrees in education. Experts were purposively selected based on their academic qualifications, professional experience, and expertise relevant to educational assessment or educational technology. Content validity was assessed using the Index of Item–Objective Congruence (IOC) proposed by Rovinelli & Hambleton (1977). Each assessment item was rated using a three-point scale of +1 (relevant), 0 (requires revision or uncertain), and -1 (not relevant) according to its alignment with the intended affective indicator. (Equation 1. Index of Item–Objective Congruence Formula) was used to calculate the IOC value for each assessment item, and an average IOC value of 0.80 or higher was adopted as the minimum acceptable threshold for content validity. In addition to assigning numerical ratings, the experts provided qualitative feedback regarding the conversational flow, probing strategy, assessment indicators, and scoring rubric. Their recommendations were analyzed descriptively and incorporated into the system refinement stage to improve the assessment model before alpha and scenario-based testing.

The primary research product was a web-based conversational assessment system developed using the Laravel framework with Blade as the presentation layer and MySQL as the database management system. The chatbot integrated the Google Gemini API using the gemini-3-flash-preview model as the primary large language model, with gemini-2.5-flash and gemini-2.0-flash configured as fallback models to maintain service continuity. The API was accessed during the system development and testing period in July 2026 through the Google Generative Language API (v1beta). The conversational module was configured with a temperature of 0.7, a top-p value of 0.9, and a maximum output length of 1,024 tokens to generate context-sensitive and coherent responses. The system prompt instructed the LLM to function as a reflective affective assessment assistant by following the predefined six-stage conversational workflow, generating adaptive probing questions based on students' responses, interpreting responses according to Krathwohl's affective taxonomy, and maintaining a neutral, non-leading conversational style. Following each assessment session, the

Table 2. Examples of Reflective Assessment Prompts and Targeted Affective Indicators

Example Prompt	Targeted Affective Indicator	Krathwohl Level
"Hello! I'm Affectra, your reflection companion today. Before we start talking about today's lesson, how are you feeling?" "While recalling today's lesson, which part caught your attention the most, and why?"	Serves as an introductory recall prompt to activate students' reflection before affective exploration rather than being directly scored as an affective indicator.	Receiving
"How did you feel while participating in today's lesson?" "What did you do when you encountered something difficult to understand?"	Demonstrates active participation, engagement, and emotional responses toward the learning experience.	Responding
Do you think today's lesson is useful for your daily life?" "What value or lesson from today's learning experience stands out most to you?"	Demonstrates appreciation of learning values and the ability to connect learning experiences with personal life and experiences.	Valuing
"What would you do if the value you learned today conflicted with your usual habits?" "What factors would you consider when choosing between two different courses of action?"	Demonstrates the ability to compare, integrate, and organize personal values when confronted with conflicts or dilemmas.	Organization
"Can you provide an example of a real action you have taken based on what you learned?" "How consistently do you apply this in your daily life?"	Demonstrates that a value has been internalized and is reflected in consistent behavior as part of the individual's character.	Characterization by Value

conversation transcript was processed using a separate analysis prompt to generate structured assessment outputs consisting of the identified affective level, confidence rating, analytical summary, and supporting conversational evidence.

The system implemented role-based authentication to separate teacher and student access, while conversation histories, assessment results, and system records were stored in a MySQL database to support assessment management and teacher review. During the development and validation stages, no real student data were collected or processed; all conversational data originated from researcher-generated scenarios used for alpha and scenario-based testing. Accordingly, the safeguards implemented in the prototype were limited to role-based authentication, controlled database access, server-side input validation, secure password hashing, and session management. Additional safeguards required for future classroom implementation, including informed consent procedures, data anonymization, privacy protection, secure data retention, and institutional ethical oversight, were beyond the scope of the present prototype. The conversational assessment workflow consisted of six sequential stages: Opening, Recall, Reflection, Affective Trigger, Probing and Self-Audit, and Closing. Responses were subsequently interpreted using an assessment rubric derived from Krathwohl's affective taxonomy to classify students' reflective responses across the five affective levels from Receiving to Characterization by Value, thereby supporting teachers in conducting structured reflective affective assessment (Følstad & Brandtzæg, 2017; Wollny et al., 2021).

The assessment prompts were formulated as open-ended and reflective questions rather than direct attitude statements. This indirect elicitation approach was adopted to reduce evaluative pressure and minimize socially desirable responses, allowing students to express their perspectives more naturally (Machost & Stains, 2023; Mohamad & Tasir, 2023). Rather than asking students to self-rate predefined attitudes, the prompts encouraged them to describe authentic learning experiences, explain their decisions, and reflect on the values underlying their actions. The prompts were designed to capture indicators associated with the five levels of Krathwohl's affective taxonomy, namely receiving, responding, valuing, organization, and characterization by value. The questioning strategy progressively elicited increasingly complex forms of affective expression, ranging from awareness and participation to value integration and internalized commitment.

Table 3. Examples of Reflective Responses and Classification Criteria Across Krathwohl's Affective Levels

Aspect	Low	Medium	High
Krathwohl Level	Receiving–Responding (A1–A2)	Valuing (A3)	Organization–Characterization (A4–A5)
Key Characteristics	Brief or passive responses; participates in discussion but does not demonstrate commitment to the underlying value.	Expresses intention or commitment to a value, but the commitment has not been tested or weakens when confronted with a dilemma.	Maintains commitment to the value under dilemma or pressure and/or provides evidence of long-term consistent behavior.
Example Response (Recall)	*"Yes, today we learned about photosynthesis. It was about plants needing sunlight."*	*"Today we learned about photosynthesis. It made me realize how important plants are because they produce the oxygen we breathe every day."*	*"Today we learned about photosynthesis. It strengthened my belief that protecting trees is everyone's responsibility, and I have been planting trees in my yard since last year."*
Example Response (Reflection)	*"It was okay. Not very exciting, but not boring either."*	*"At first I was confused, but after the teacher explained it using animations, I understood it better. I felt happy because I finally understood the process."*	*"I found the topic challenging because it was new to me, but that actually motivated me. After school, I even searched for additional videos on YouTube to learn more."*
Example Response (Affective Commitment)	*"I'm not really sure. I guess it's important, but that's about it."*	*"After learning this, I want to care more about the environment. Maybe I should join the school's tree-planting activities."*	*"I am determined to join an environmental conservation community in my neighborhood. To me, protecting the environment is more important than sacrificing some of my free time."*
Example Response (Dilemma Probe)	*(Not exposed to a dilemma because the conversation ended before this stage.)*	*"Hmm... if it costs me too much of my free time, maybe I'll postpone it. I guess that's okay."*	*"Even if I have to sacrifice my break time, I will continue participating because the environmental benefits are much greater. I don't want to be someone who only talks about protecting the environment without taking action."*
Example Response (Self-Audit)	*(Did not reach this stage.)*	*(Did not reach this stage because the commitment weakened during the dilemma probe.)*	*"I would rate myself 8 out of 10 because I have consistently reduced my plastic use and taken care of the plants in my yard for the past two years. My friends know that I am consistent about this."*
Classification Rationale	The student demonstrates passive acceptance of the learning content (Receiving) or participates in discussion (Responding) without showing commitment to the underlying value.	The student recognizes and values the importance of the value (Valuing) and expresses an intention to act, but the commitment is not yet stable when challenged by conflicting situations.	The student integrates the value into a coherent value system and prioritizes it over competing interests (Organization), and may also demonstrate long-term consistency in behavior that reflects internalized values (Characterization).

As illustrated in Figure 1, the conversational assessment followed a sequential workflow consisting of learning recall, affective exploration, adaptive probing, transcript analysis, and Krathwohl-level mapping. These prompts were initially developed from the indicators of Krathwohl's affective taxonomy and designed as open-ended, experience-based questions to encourage authentic reflection while reducing socially desirable responses and cognitive contamination. Prior to implementation, the prompts were iteratively refined through expert feedback to improve their clarity, conversational naturalness, and alignment with the intended affective indicators.

During the conversation, adaptive probing was governed by predefined prompting rules embedded within the system prompt rather than unrestricted LLM generation. Follow-up questions were generated only when students' responses required further clarification or deeper affective reflection. The conversational workflow incorporated conditional branching whereby students demonstrating passive or non-committal responses proceeded directly to the closing stage, whereas students expressing value commitments were presented with contextual probing and self-audit questions to explore the consistency of their responses. To maintain a consistent conversational structure, probing was confined to the designated Probing and Self-Audit stage, after which the conversation proceeded to session closure. The system prompt further instructed the LLM to maintain a neutral, empathetic, and non-leading conversational style in order to encourage authentic self-disclosure while avoiding evaluative language. Conversational outputs were further constrained by the predefined workflow and prompt instructions, such that responses deviating from the intended assessment sequence or pedagogical purpose were redirected to the next valid conversational stage rather than allowing unrestricted generation.

After the conversation was completed, the transcript was analyzed using predefined analysis prompts. The identified affective evidence was then mapped to the corresponding level of Krathwohl's affective taxonomy, after which the system generated a structured assessment output consisting of the identified affective level, confidence rating, analytical summary, and supporting conversational evidence. Both the adaptive probing strategy and the transcript analysis prompts, together with the overall assessment workflow, were reviewed during expert validation to ensure their pedagogical appropriateness and alignment with the intended affective indicators before system implementation.

To illustrate how reflective responses were interpreted, Table 2 presents representative examples of responses corresponding to different levels of Krathwohl's affective taxonomy. These examples served as conceptual references for rubric development rather than fixed response templates.

A scenario-based testing approach was adopted by constructing five predefined conversational scenarios, each representing one of the five levels of Krathwohl's affective taxonomy (Receiving, Responding, Valuing, Organization, and Characterization by Value) (see Table 3). Each scenario was intentionally designed to represent the defining characteristics of a single affective level based on the validated assessment rubric. The use of these predefined scenarios enabled a systematic evaluation of the chatbot's functionality, conversational consistency, contextual probing capability, affective indicator recognition, and assessment generation under controlled testing conditions, consistent with recommendations for iterative software testing and conversational AI evaluation in educational settings (Obigbesan et al., 2024; Pit et al., 2025; Yusuf et al., 2025).

The testing was conducted internally by the researchers through simulated interactions rather than by students or teachers. Data collected during this phase included conversation logs, identified affective indicators, generated assessment outputs, and records of system issues encountered during testing. These data were used to evaluate whether the chatbot maintained coherent conversational flow, accurately identified affective indicators, and generated assessment results consistent with the predefined assessment framework (Obigbesan et al., 2024; Pit et al., 2025). The objective of this alpha testing phase was to verify the chatbot's conversational processes, affective indicator recognition, and assessment generation mechanisms before subsequent refinement rather than to evaluate learning outcomes or user satisfaction (Allouch et al., 2021; Yusuf et al., 2025). Accordingly, this phase was intended solely to verify system functionality prior to classroom implementation and was not designed to evaluate educational effectiveness or generalize learner performance.

Conversation logs and assessment outputs were analyzed using directed qualitative content analysis, in which the coding framework was established deductively from an existing theoretical framework (Assarroudi et al., 2018). In this study, Krathwohl's affective taxonomy served as the predetermined coding framework, while the operational indicators and response criteria presented in Table 2 guided the classification process. Two researchers independently reviewed and classified the chatbot-generated responses according to the five affective levels. Any differences in interpretation were discussed until consensus was reached to improve the consistency of the classifications. The analysis focused on examining whether the chatbot correctly identified affective indicators and produced assessment outputs aligned with the intended affective levels. Responses demonstrating awareness and attention to learning experiences were classified as Receiving, responses reflecting active participation as Responding, responses expressing personal meaning or commitment as Valuing, responses integrating multiple values into a coherent value system as Organization, and responses demonstrating consistent internalized values through attitudes and behaviors as Characterization (Krathwohl et al., 1964). This deductive analysis enabled a systematic evaluation of both the chatbot's assessment performance and its alignment with the theoretical framework underpinning the conversational assessment.

Result and Discussion

The developed affective assessment instrument was evaluated by six experts using the Index of Item–Objective Congruence (IOC) method to examine the alignment between the assessment prompts and the intended affective indicators.

As presented in Table 4, the initial overall IOC score of the developed assessment instrument was 0.81, indicating satisfactory overall alignment between the assessment items and their intended affective indicators. Following the recommended interpretation criterion that IOC values of 0.80 or higher indicate acceptable alignment, most affective domains met the predefined criterion (Yusoff, 2019). However, the Receiving domain obtained an IOC score of 0.75 and

Table 4. IOC Validation Results by Affective Domain

Affective Domain	Number of Items	Total Score Obtained	Maximum Score	IOC Score
Receiving	10	45	60	0.75
Responding	10	48	60	0.80
Valuing	10	49	60	0.82
Organization	10	53	60	0.88
Characterization by Value	10	49	60	0.82
Overall	**50**	**244**	**300**	**0.81**

therefore required further examination and revision. It should be noted that IOC reflects the degree of alignment between assessment items and their intended objectives rather than the overall validity or effectiveness of the assessment instrument.

Among the five affective domains, the Organization domain achieved the highest initial IOC score (0.88), indicating strong agreement among experts regarding the suitability of prompts designed to explore value prioritization, conflict resolution, and value integration. The Valuing and Characterization by Value domains each obtained an IOC score of 0.82, while the Responding domain reached 0.80. These domains met the predefined acceptance criterion. In contrast, the Receiving domain obtained an initial IOC score of 0.75, which was below the 0.80 criterion and prompted an item-level examination of the assessment prompts.

Item-level analysis of the Receiving domain showed that five of the ten prompts obtained an IOC score of 0.67, while the remaining five prompts obtained 0.83. The lower-scoring prompts primarily involved wording and construct-alignment issues, particularly where the questions emphasized cognitive recall or reconstruction of learning content rather than students' awareness and attention toward learning experiences. Accordingly, the five affected prompts were revised based on expert recommendations, while the remaining prompts were retained and subsequently revalidated. The summary of expert feedback and revisions is presented in Table 5, the item-level initial and post-revalidation IOC results are presented in Table 6, and representative examples of the revised prompts are presented in Table 7.

In addition to the quantitative validation results, experts provided qualitative suggestions for improving the assessment instrument. Several prompts in the Responding domain were revised to focus more explicitly on students' emotional reactions and engagement rather than evaluative judgments. Similarly, minor revisions were made to selected prompts in the Valuing, Organization, and Characterization by Value domains to improve clarity and strengthen their alignment with the corresponding affective indicators. These revisions, together with the revalidation of the Receiving domain, were incorporated into the revised version of the conversational assessment model before proceeding to the implementation and testing stages.

The expert validation process resulted in a revised version of Affectra, a web-based conversational assessment system designed to support reflective affective assessment following learning activities. The validated assessment prompts, conversational workflow, and affective assessment rubric were integrated into a unified system capable of collecting, interpreting, and classifying affective evidence through conversational interaction. The resulting system provides an alternative to conventional affective assessment approaches by combining reflective dialogue with AI-assisted interpretation of students' responses.

Affectra consists of three interconnected modules: (1) a conversational interface, responsible for managing user interactions; (2) an LLM-based conversational analysis module, which interprets responses and generates contextual follow-up questions using predefined prompting rules; and (3) an affective assessment module, which maps conversational evidence to Krathwohl's affective taxonomy using the validated assessment rubric. Together, these modules process conversational data sequentially from response collection, contextual interpretation, adaptive probing, transcript analysis, affective classification, and assessment report generation. The operational interaction among these modules is summarized in Table 8.

A distinguishing feature of the developed model is its adaptive probing mechanism, which dynamically generates follow-up questions based on the semantic content and contextual meaning of students' responses rather than following a fixed questioning sequence (Wollny et al., 2021). The prompting process follows predefined conversational rules that encourage students to elaborate on their experiences, justify their responses, or provide behavioral examples whenever the initial response does not provide sufficient affective evidence. At the same time, prompt templates constrain the LLM to remain consistent with the intended affective indicator, thereby reducing the generation of irrelevant or leading questions. Representative examples of adaptive probing are presented in Table 9.

Following the conversational interaction, the complete transcript was analyzed using the validated affective assessment rubric and mapped to the five levels of Krathwohl's affective taxonomy. The generated assessment report summarizes the identified affective level, supporting affective indicators, and representative conversational evidence to provide teachers with interpretable information for formative assessment and monitoring of students' affective development (Gardner et al., 2021; Heritage, 2018; Rogers et al., 2017; Shute & Rahimi, 2017).

Alpha testing was conducted to evaluate the functionality and reliability of the developed conversational assessment

Table 5. Summary of Expert Feedback and Instrument Revisions by Affective Domain

Domain	Items Reviewed	Revised Items	Primary Revision Focus
Receiving	10	5	Open-ended wording, reducing cognitive recall, eliminating redundancy
Responding	10	3	Emotional engagement and construct alignment
Valuing	10	2	Clarity and conversational flow
Organization	10	2	Scenario wording and conflicting values
Characterization	10	3	Behavioral evidence and language consistency

Table 6. Item-Level IOC and Revalidation Results for the Receiving Domain

Prompt ID	Initial IOC	Status	Post-Revision IOC
PR-Q1	0.67	Revised	0.83
PR-Q2	0.83	Retained	1.00
PR-Q3	0.83	Retained	1.00
PR-Q4	0.83	Retained	1.00
PR-Q5	0.67	Revised	0.83
PR-Q6	0.83	Retained	1.00
PR-Q7	0.67	Revised	0.83
PR-Q8	0.67	Revised	0.83
PR-Q9	0.83	Retained	1.00
PR-Q10	0.67	Revised	0.83
Receiving Domain	0.75	Revalidated	0.92

system prior to implementation (Allouch et al., 2021; Obigbesan et al., 2024). The evaluation focused on two complementary aspects: system functionality and assessment performance. Functional testing was first performed to verify that all system components operated according to the intended workflow, while scenario-based testing was subsequently conducted to examine the capability of the assessment model in identifying affective levels based on predefined conversational scenarios (Pit et al., 2025).

Since the primary objective of this study was to develop and validate an affective assessment instrument rather than measure instructional effectiveness, the results of these evaluations provide evidence regarding the operational feasibility and assessment capability of the developed system. As shown in Table 10, functional testing was conducted to verify the operational performance of each core component of

Table 7. Representative Prompt Revisions Following Expert Validation

Domain	Original Prompt	Expert Concern	Revised Prompt
Receiving – Question 1	*How did you feel while learning this topic?* (delivery example: *"Could you tell me how you felt?"*)	The prompt required more natural conversational wording to encourage authentic reflection and reduce evaluative pressure. Experts recommended using supportive language while maintaining alignment with the intended affective indicator.	*Can you tell me how you felt during today's learning activity? What part of the lesson made you feel that way?*
Receiving – Question 2	*What is your attitude toward this issue?* (delivery example: *"What do you think about it?"*)	The wording was considered too broad and potentially elicited evaluative judgments rather than reflective engagement. Experts recommended encouraging students to relate their responses to personal learning experiences.	*After today's lesson, what do you think about this issue? Could you explain why you feel that way?*
Reflection – Question 3 (Inactive)	*Have you ever experienced something similar?* (delivery example: *"Would you like to share a little?"*)	Experts suggested that this question should not be presented unconditionally because it may be irrelevant to some students. It was therefore retained as an **adaptive probing prompt**, activated only when additional contextual evidence is required.	*Could you share an experience that relates to what we discussed today? If so, how did it influence your perspective? (Adaptive probing)*
Affective – Question 1	*How important do you think this topic is in everyday life?* (delivery example: *"How relevant do you think this topic is to your daily life?"*)	The prompt focused primarily on judging the topic's importance rather than encouraging students to explain the personal meaning they attributed to it. Experts recommended strengthening the value-reflection component.	*Why do you think this topic is important (or not important) in your daily life? Could you give an example?*
Affective – Question 2	*Do you think there is a value you can learn from this lesson?* (delivery example: *"Is there any life lesson you learned from today's lesson?"*)	The original wording could be answered briefly with "yes" or "no." Experts recommended using an open-ended formulation to encourage deeper reflection and value articulation.	*What value or lesson did you gain from today's learning experience, and why is it meaningful to you?*
Affective – Question 3	*How has this lesson changed your perspective?* (delivery example: *"Has learning this topic changed the way you think?"*)	Experts noted that the original prompt combined multiple ideas and did not sufficiently explore the student's commitment to the identified value. They recommended a clearer, sequential questioning strategy.	*Has today's lesson changed the way you think about this issue? If yes, what has changed, and how might it influence your future actions?*

Table 8. Workflow of the Affectra Conversational Assessment System

Stage	System Process	Output
Response Collection	Student submits reflective response	Conversation log
Contextual Analysis	LLM interprets response using prompt template	Semantic interpretation
Adaptive Probing	Generate follow-up question when evidence is insufficient	Additional reflective evidence
Transcript Analysis	Analyze complete conversation	Affective indicators
Krathwohl Mapping	Match rubric	Affective level
Assessment Report	Generate structured summary	Final assessment result

Table 9. Examples of Adaptive Probing During Conversational Assessment

Initial Response	Detected Situation	Adaptive Probe
"It was okay."	Reflection lacks affective evidence	"Could you explain which part of the lesson influenced you the most?"
"I think honesty is important."	Value identified but behavioral evidence absent	"Can you describe a situation where you applied this value?"
"Maybe I would do it."	Commitment unclear	"What factors would encourage or prevent you from acting on that decision?"

the Affectra conversational assessment system. The evaluation comprised multiple test cases with repeated executions according to the complexity of each component. Conversational features, including the chatbot interface, recall stage, affective exploration, adaptive probing, transcript analysis, Krathwohl level mapping, and assessment result generation, were evaluated through repeated conversational scenarios, whereas supporting system functions such as session initialization, data storage, and session completion were tested through multiple execution cycles. A component was classified as Pass when all repetitions completed successfully, produced the expected outputs, and no functional or workflow errors were observed.

The results indicate that all evaluated components satisfied their respective pass criteria and operated consistently with the designed conversational assessment workflow. Across all testing iterations, the system successfully managed conversational interactions, generated contextually appropriate adaptive probing questions, processed transcripts using the LLM, mapped responses to the intended Krathwohl affective levels, generated assessment reports, and stored assessment records without functional failures or workflow interruptions. Minor refinements were subsequently made to improve conversational pacing and response timing without affecting the assessment logic. Following the successful completion of functional testing, scenario-based testing was conducted using predefined conversational scenarios representing the five levels of Krathwohl's affective taxonomy to evaluate the model's capability to classify affective responses under different conversational contexts (see Table 11).

As shown in Table 11, all predefined scenarios were successfully classified into their corresponding affective levels, resulting in a 100% matching rate between the expected and detected classifications. Because the scenarios were intentionally developed from the validated affective assessment rubric, this result should be interpreted as evidence of the internal consistency of the developed assessment mechanism under controlled testing conditions rather than predictive validity or real-world classification accuracy. The findings indicate that the model consistently distinguished representative responses across the five levels of Krathwohl's affective taxonomy, including Receiving, Responding, Valuing, Organization, and Characterization by Value, according to their intended affective indicators.

The scenario-based evaluation also demonstrated that the

Table 10. Functional Testing Results

No.	Component	Test Cases	Repetitions	Pass Criteria	Result	Status
1	Session Initialization	Initiate a new assessment session	3	Assessment session is created successfully and initialized without errors.	Passed all 3 executions. The assessment session was initialized correctly in every trial.	Pass
2	Conversational Assessment Interface	Send and receive chatbot messages	3	User messages are transmitted and chatbot responses are returned successfully.	Passed all 3 executions. Messages and chatbot responses were exchanged correctly throughout the conversation.	Pass
3	Recall Stage	Deliver recall prompts	3	Recall prompts are presented correctly and recorded.	Passed all 3 executions. Recall prompts were delivered successfully as part of the assessment workflow.	Pass
4	Affective Exploration Stage	Conduct reflection and affective questioning	3	Reflective questions are generated and conversation flow proceeds correctly.	Passed all 3 executions. Reflection and affective exploration stages operated according to the designed workflow.	Pass
5	Adaptive Probing Mechanism	Generate contextual follow-up questions	3	Adaptive probing follows predefined branching rules based on student responses.	Passed all 3 executions. Contextual probing questions and branching logic operated correctly.	Pass
6	Transcript Analysis Using LLM	Analyze conversation transcript	3	Transcript is processed successfully and structured analysis is generated.	Passed all 3 executions. Conversation transcripts were analyzed successfully without processing failures.	Pass
7	Krathwohl Level Mapping	Classify responses into affective levels	3	Responses are mapped to the intended Krathwohl affective level.	Passed all 3 executions. Affective levels were identified correctly according to the assessment rubric.	Pass
8	Assessment Result Generation	Generate assessment report	3	Assessment summary and supporting evidence are generated automatically.	Passed all 3 executions. Assessment reports were generated successfully after conversation completion.	Pass
9	Data Storage	Store assessment records	3	Conversation logs and assessment results are stored successfully.	Passed all 3 executions. All assessment data were stored completely without data loss.	Pass
10	Session Completion	Complete assessment session	3	Session is terminated correctly and assessment process is finalized.	Passed all 3 executions. Sessions were completed successfully without workflow interruption.	Pass

adaptive probing mechanism and the LLM-based contextual analysis functioned consistently throughout the assessment process. Rather than relying solely on keyword matching, the LLM interpreted contextual meaning and supporting evidence obtained through adaptive probing to classify affective responses in accordance with the validated assessment rubric (Min et al., 2023; Wollny et al., 2021). To illustrate the assessment workflow, the following section presents a representative conversational example that was selected because it contains sufficient conversational evidence to demonstrate the complete process of adaptive probing, affective indicator identification, and affective level classification. Although only one representative example is presented for illustration, the same analytical procedure was applied consistently across all predefined testing scenarios.

Overall, the alpha testing results indicate that the developed conversational assessment model operated as intended from both functional and analytical perspectives. The successful execution of all system components, combined with the accurate classification of affective levels across the testing scenarios, provides preliminary evidence supporting the feasibility of Affectra as a chatbot-based instrument for reflective affective assessment.

To further illustrate how the developed assessment mechanism operates, a representative conversational example is presented in the following section. The example was selected because it demonstrates the complete assessment workflow, including reflective dialogue, adaptive probing, contextual transcript analysis, affective indicator identification, and final classification. Although only one example is presented for illustration, it represents the same analytical procedure applied across all predefined testing scenarios.

Table 12 illustrates how the system interprets student responses and maps them to the corresponding level of Krathwohl's affective taxonomy (Krathwohl et al., 1964).

During the assessment process, the chatbot analyzes the contextual meaning of students' responses and identifies indicators associated with affective development (Min et al., 2023).

Table 11. Scenario-Based Internal Validation Results Using Representative Krathwohl Scenarios

Scenario	Intended Indicator	Expected Level	Detected Level	Match	Summary
S1	Demonstrates awareness and attention toward the learning experience.	Receiving	Receiving	Yes	The response reflects receptiveness to the learning material without evidence of active engagement.
S2	Demonstrates active participation and meaningful response to the learning content.	Responding	Responding	Yes	The response shows active engagement and relates the learning material to real-life understanding.
S3	Demonstrates appreciation of learning values and willingness to apply them.	Valuing	Valuing	Yes	The response expresses personal appreciation and an intention to act based on the learned values.
S4	Demonstrates value prioritization and resolution of competing values.	Organization	Organization	Yes	The response shows the ability to organize and prioritize values when facing conflicting situations.
S5	Demonstrates consistent value-driven behavior integrated into daily life.	Characterization by Value	Characterization by Value	Yes	The response provides evidence of sustained value-based behavior as part of the student's character.

Table 12. Example Conversation and Assessment Result

Speaker	Conversation	Assessment Phase	Affective Indicator
Bot	Hi Andi! 👋 I'm Affectra, your learning companion. Today we're going to talk about Photosynthesis. How are you today?	Opening	-
Student	Hi Affectra! I'm doing well. Earlier, we learned about photosynthesis in class.	Opening	-
Bot	That's great! Can you tell me what you remember from today's lesson about photosynthesis?	Recall	-
Student	Photosynthesis is the process by which plants produce their own food using sunlight, water, and carbon dioxide. It produces glucose and oxygen.	Recall	-
Bot	Excellent! You remember the material well. How did you feel during the lesson? Was there anything that made you curious or confused?	Reflection	-
Student	At first, I was confused about the chemical reactions, but after the teacher explained them using animations, I understood them better and became more curious about the process.	Reflection	Responding
Bot	That's interesting! Curiosity often encourages deeper learning. In your opinion, why is photosynthesis important in everyday life?	Affective Assessment	-
Student	Now I understand why photosynthesis is important in our daily lives. Without photosynthesis, there would be no oxygen and no food.	Affective Assessment	Responding
Bot	Thank you for sharing your thoughts, Andi. Our session is now complete. Keep up the good work in your learning journey!	Closing	-

The classification process is supported by adaptive probing and LLM-based transcript analysis, allowing the system to examine not only explicit statements but also the reasoning and contextual evidence provided by students (Min et al., 2023; Wollny et al., 2021).

The assessment generated by the LLM classified the student's affective level as Responding, with the identified indicator being Active Participation and Engagement. This classification was supported by the student's statement, "Now I understand why photosynthesis is important in our daily lives. Without photosynthesis, there would be no oxygen and no food." The response demonstrates that the student not only participated actively in the conversation but also connected the learning material to its broader relevance in everyday life. Based on the conversational evidence, the system determined that the student was actively responding to the learning experience rather than merely recalling factual information, which is consistent with the characteristics of the Responding level in Krathwohl's affective taxonomy.

Table 12 presents an example of a conversational interaction stored in the system database, in which the chatbot explored a student's understanding and personal reflection regarding photosynthesis. Through a sequence of structured conversational phases and adaptive probing questions, the system guided the student from recalling factual information toward expressing personal interpretations of the learning content. The conversational evidence revealed that the student actively engaged with the discussion and was able to relate the concept of photosynthesis to its broader significance for human life. Based on the validated assessment rubric, the response was classified as belonging to the Responding level of Krathwohl's affective taxonomy (see Table 12).

This example demonstrates how Affectra transforms conversational interactions into structured assessment outputs. Rather than relying solely on direct questioning or predefined response options, the system utilizes reflective dialogue, contextual probing, and LLM-based analysis to capture affective evidence and generate interpretable assessment results (Følstad & Brandtzæg, 2017; Min et al., 2023; Wollny et al., 2021). The resulting output provides teachers with transparent information regarding the identified affective level and the supporting conversational evidence, thereby supporting the use of conversational assessment as a complementary approach for evaluating students' affective development (Rogers et al., 2017; Stephens & Ormandy, 2019).

The findings of this study provide preliminary evidence supporting the content validity and operational feasibility of the proposed conversational assessment model for reflective affective assessment. The expert validation process resulted in an overall IOC score of 0.81, indicating that the assessment prompts were appropriately aligned with the intended affective indicators. In addition, the alpha testing demonstrated that all system components operated according to the designed workflow and consistently generated assessment outputs under predefined testing scenarios. These findings suggest that a conversational assessment approach can provide a structured mechanism for collecting and documenting affective evidence, potentially addressing some of the limitations associated with traditional affective assessment practices that often rely on subjective observation and limited documentation (Jacob & Lefgren, 2008). However, these findings should be interpreted as evidence of the model's initial feasibility and content validity rather than its effectiveness or reliability in authentic classroom settings, as the current study did not involve real student implementation or comparisons with teacher judgments.

The findings are consistent with previous studies demonstrating that conversational agents can facilitate reflective dialogue and encourage learner self-disclosure through natural language interaction (Følstad & Brandtzæg, 2017; Wollny et al., 2021). Similarly, the chatbot developed in this study was able to elicit contextual responses and identify affective indicators through adaptive dialogue, supporting the argument that conversational approaches may provide richer affective evidence than conventional closed-ended assessment instruments (Allouch et al., 2021). However, whereas previous studies have primarily positioned chatbots as instructional support tools, virtual tutors, or learning companions, the present study extends this line of research by operationalizing conversational interaction as an assessment mechanism explicitly grounded in Krathwohl's affective taxonomy (Krathwohl et al., 1964). Consequently, the contribution of this study is methodological rather than instructional, demonstrating how LLM-supported reflective dialogue can be systematically used to collect interpretable affective evidence.

A key contribution of this study is therefore not merely the technical integration of conversational assessment, adaptive probing, and LLM-based transcript analysis, but the development of a structured assessment methodology that links reflective conversations with validated affective indicators. Unlike conventional affective assessment methods that primarily depend on observation checklists, rating scales, or self-report questionnaires, the proposed model systematically elicits contextual evidence through adaptive reflective dialogue and maps that evidence to the hierarchical levels of Krathwohl's affective taxonomy. This methodological framework directly addresses the research gap identified in the Introduction, where previous chatbot studies rarely provided theoretically grounded procedures for conducting reflective affective assessment. Accordingly, the proposed model offers an initial framework that demonstrates the potential of AI-supported conversational systems to assist teachers in formative affective assessment while maintaining a transparent connection between conversational evidence and assessment criteria.

Despite these encouraging findings, several limitations should be acknowledged. The evaluation was limited to expert validation and internal alpha testing using predefined conversational scenarios, representing an early-stage evaluation of the proposed system. Consequently, the reported classification performance reflects the behavior of the assessment mechanism under controlled testing conditions rather than authentic classroom interactions. Furthermore, this study did not examine psychometric reliability, compare chatbot-generated assessments with teacher judgments, or investigate the effectiveness of the system in improving students' reflective abilities or affective development. As the proposed model relies on a large language model, its analyses may also be affected by potential hallucination, bias in affective interpretation, and inconsistencies arising from future model updates (Chen et al., 2024). In addition, issues related to student privacy, informed consent, transparency of AI-assisted scoring, and the need for human teacher oversight should be carefully considered before implementation in real educational settings. Therefore, the present findings should be interpreted as evidence of the model's initial feasibility and validity rather than proof of its effectiveness or reliability in educational practice.

Future research should evaluate the proposed model through beta testing and implementation in authentic classroom environments involving students and teachers across different educational contexts. Such studies should examine system usability, user acceptance, and practical applicability while comparing chatbot-generated affective classifications with teacher assessments to establish criterion-related validity and inter-rater agreement. Larger and more diverse conversational datasets would also enable psychometric evaluation of the assessment model and improve the robustness, fairness, and consistency of LLM-based affective analysis. Additionally, future studies should investigate the influence of adaptive probing on assessment quality and evaluate the stability of assessment outcomes across different LLM versions, thereby providing stronger evidence for the practical deployment of conversational affective assessment in educational settings.

Conclusion

This study aimed to develop and conduct an initial evaluation of Affectra, a web-based conversational assessment model designed to support reflective affective assessment following learning activities. Grounded in Krathwohl's affective taxonomy, the proposed model integrates conversational assessment, adaptive probing, and LLM-based analysis to facilitate the systematic collection and interpretation of students' affective responses through reflective dialogue.

The findings provide preliminary evidence supporting the content validity and operational feasibility of the proposed assessment model. The assessment instrument achieved an initial overall Index of Item–Objective Congruence (IOC) score of 0.81, indicating appropriate alignment between the assessment prompts and the intended affective indicators, with the below-threshold Receiving domain subsequently revised and revalidated. Internal alpha testing confirmed that the core system components operated according to the designed workflow. In addition, the scenario-based testing functioned as an internal consistency check under predefined researcher-generated scenarios, in which the expected affective levels were specified according to the assessment rubric. The system consistently mapped the predefined scenarios to their intended Krathwohl levels under these controlled conditions. Therefore, these findings should not be interpreted as evidence of classification accuracy, predictive validity, classroom effectiveness, student acceptance, or equivalence with teacher assessments, as no independent human raters, blinded comparison, agreement coefficient, or external criterion were used.

A key contribution of this study is the development of a theoretically grounded conversational assessment framework that extends the role of educational chatbots beyond instructional support toward reflective affective assessment. Rather than relying solely on observation checklists or self-report questionnaires, the proposed framework systematically combines reflective dialogue, adaptive probing, and LLM-based contextual analysis to elicit and interpret affective evidence based on validated affective indicators. From a practical perspective, Affectra has the potential to serve as a complementary formative assessment tool that assists teachers in documenting students' affective reflections more systematically while supporting curriculum developers in exploring technology-enhanced approaches to affective assessment. Nevertheless, the assessment outcomes should continue to be interpreted with appropriate human teacher oversight.

Several limitations should also be acknowledged. The evaluation was limited to expert validation and internal alpha testing using predefined conversational scenarios, and the model has not yet been evaluated in authentic classroom environments involving actual students and teachers. In addition, psychometric reliability, agreement with teacher judgments, and the potential influence of LLM-related issues, such as hallucination, bias, and model-version inconsistency, were beyond the scope of the present study. Future research should therefore conduct beta testing in authentic classroom settings involving students at the secondary-school level and their teachers, compare chatbot-generated affective classifications with teacher ratings using appropriate agreement and reliability measures, evaluate system usability and user acceptance, and examine ethical considerations related to AI-assisted affective assessment, including transparency, privacy, informed consent, and human oversight. Such investigations will provide stronger evidence regarding the validity, reliability, and practical applicability of conversational assessment models in educational settings.

Author contributions

Fikriansyah Haikal Ramadhan conceived the research idea, conducted the literature review, designed the conversational assessment model, developed the chatbot system, designed the prompting strategy, carried out data collection and analysis, interpreted the findings, and prepared the original manuscript. Rizki Hikmawan supervised the research, provided methodological guidance, reviewed and validated the research design and manuscript, contributed to the interpretation of the findings, and approved the final version of the manuscript

Funding

This research was fully self-funded by the authors and did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors confirm that no external organization had any role in the study design, system development, data collection, analysis, interpretation of the findings, manuscript preparation, or the decision to submit the manuscript for publication.

Acknowledgements

The authors would like to express their sincere gratitude to Universitas Pendidikan Indonesia, particularly the Department of Information Systems and Technology Education (PSTI), for providing academic support throughout this research. The authors also sincerely thank the expert validators for their valuable feedback and constructive suggestions during the content validation process, which significantly contributed to the refinement of the developed conversational assessment model.

References

Abrahams, L., Pancorbo, G., Primi, R., Santos, D., Kyllonen, P. C., John, O. P., & De Fruyt, F. (2019). Social-Emotional Skill Assessment in Children and Adolescents: Advances and Challenges in Personality, Clinical, and Educational Contexts. *Psychological Assessment*, *31*(4), 460–473. https://doi.org/10.1037/pas0000591

Aldrup, K., Carstensen, B., & Klusmann, U. (2022). Is Empathy the Key to Effective Teaching? A Systematic Review of Its Association With Teacher-Student Interactions and Student Outcomes. *Educational Psychology Review*, *34*, 1177–1216. https://doi.org/10.1007/s10648-021-09649-y

Allouch, M., Azaria, A., & Azoulay, R. (2021). Conversational Agents: Goals, Technologies, Vision and Challenges. *Sensors*, *21*(24), 8448. https://doi.org/10.3390/s21248448

Assarroudi, A., Heshmati Nabavi, F., Armat, M. R., Ebadi, A., & Vaismoradi, M. (2018). Directed Qualitative Content Analysis: The Description and Elaboration of Its Underpinning Methods and Data Analysis Process. *Journal of Research in Nursing*, *23*(1), 42–55. https://doi.org/10.1177/1744987117741667

Borg, W. R., Gall, M. D., & Gall, J. P. (2007). *Educational Research: An Introduction* (8th ed.). Pearson.

Boyd, V. A., Woods, N. N., Kumagai, A. K., Kawamura, A. A., Orsino, A., & Ng, S. L. (2022). Examining the Impact of Dialogic Learning on Critically Reflective Practice. *Academic Medicine*, *97*(11 Suppl 2), S71–S79. https://doi.org/10.1097/ACM.0000000000004916

Chen, J., Liu, Z., Huang, X., & others. (2024). When Large Language Models Meet Personalization: Perspectives of Challenges and Opportunities. *World Wide Web*, *27*, 42. https://doi.org/10.1007/s11280-024-01276-1

Ferreira, M., Martinsone, B., & Talić, S. (2020). Promoting Sustainable Social Emotional Learning at School Through Relationship-Centered Learning Environment, Teaching Methods and Formative Assessment. *Journal of Teacher Education for Sustainability*, *22*(1), 21–36. https://doi.org/10.2478/jtes-2020-0003

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the New World of HCI. *Interactions*, *24*(4), 38–42. https://doi.org/10.1145/3085558

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial Intelligence in Educational Assessment: "Breakthrough? Or Buncombe and Ballyhoo?" *Journal of Computer Assisted Learning*, *37*(5), 1207–1216. https://doi.org/10.1111/jcal.12577

Heritage, M. (2018). Assessment for Learning as Support for Student Self-Regulation. *The Australian Educational Researcher*, *45*(1), 51–63. https://doi.org/10.1007/s13384-018-0261-3

Ho, A., Hancock, J. T., & Miner, A. S. (2018). Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication*, *68*(4), 712–733. https://doi.org/10.1093/joc/jqy026

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education*, *32*(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Jacob, B. A., & Lefgren, L. (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, *26*(1), 101–136. https://doi.org/10.1086/522974

Kim, H., Sefcik, J. S., & Bradway, C. (2017). Characteristics of Qualitative Descriptive Studies: A Systematic Review. *Research in Nursing & Health*, *40*(1), 23–42. https://doi.org/10.1002/nur.21768

Kraft, M. A. (2019). Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies. *Journal of Human Resources*, *54*(1), 1–36. https://doi.org/10.3368/jhr.54.1.0916.8265R3

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of Educational Objectives: Handbook II: Affective Domain*. David McKay Company.

Luo, J., Zheng, C., Yin, J., & Teo, H. H. (2025). Design and Assessment of AI-Based Learning Tools in Higher Education: A Systematic Review. *International Journal of Educational Technology in Higher Education*, *22*(1), 42. https://doi.org/10.1186/s41239-025-00540-2

Machost, H., & Stains, M. (2023). Reflective Practices in Education: A Primer for Practitioners. *CBE-Life Sciences Education*, *22*(1), es2. https://doi.org/10.1187/cbe.22-07-0148

Meyliasari, A. R. (2024). Penyusunan Instrumen Penilaian Afektif di Sekolah [Development of Affective Assessment Instruments in Schools]. *Muaddib: Jurnal Pendidikan Agama Islam*, *2*(2), 430–441. https://ejournal.insuriponorogo.ac.id/index.php/muaddib/article/view/6590

Min, B., Ross, H., Sulem, E., Veyseh, A. P. Ben, Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. *ACM Computing Surveys*, *56*(2), 1–40. https://doi.org/10.1145/3605943

Mohamad, S. K., & Tasir, Z. (2023). Exploring How Feedback Through Questioning May Influence Reflective Thinking Skills Based on Association Rules Mining Technique. *Thinking Skills and Creativity*, *47*, 101231. https://doi.org/10.1016/j.tsc.2023.101231

Obigbesan, O., Graham, K., & Benzies, K. (2024). Software Testing of eHealth Interventions: Existing Practices and the Future of an Iterative Strategy. *JMIR Nursing*, *7*, e56585. https://doi.org/10.2196/56585

Pit, S. W., Hamiduzzaman, M., Schneider, C. R., & Barraclough, F. (2025). Evaluation Framework for Conversational AI Agents in Pharmacy Education: A Scoping Review of Key Characteristics and Outcome Measures. *Research in Social and Administrative Pharmacy*, *21*(10), 729–742. https://doi.org/10.1016/j.sapharm.2025.05.006

Pit-ten Cate, I. M., Hörstermann, T., Krolak-Schwerdt, S., Gräsel, C., Böhmer, I., & Glock, S. (2020). Teachers' Information Processing and Judgement Accuracy: Effects of Information Consistency and Accountability. *European Journal of Psychology of Education*, *35*(3), 675–702. https://doi.org/10.1007/s10212-019-00436-6

Rogers, G. D., Mey, A., & Chan, P. C. (2017). Development of a Phenomenologically Derived Method to Assess Affective Learning in Student Journals Following Impactive Educational Experiences. *Medical Teacher*, *39*(12), 1250–1260. https://doi.org/10.1080/0142159X.2017.1372566

Rovinelli, R. J., & Hambleton, R. K. (1977). *On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Item Validity*.

Shute, V. J., & Rahimi, S. (2017). Review of Computer-Based Assessment for Learning in Elementary and Secondary Education. *Journal of Computer Assisted Learning*, *33*(1), 1–19. https://doi.org/10.1111/jcal.12172

Stephens, M., & Ormandy, P. (2019). An Evidence-Based Approach to Measuring Affective Domain Development. *Journal of Professional Nursing*, *35*(3), 216–223. https://doi.org/10.1016/j.profnurs.2018.12.004

Sukhera, J. (2022). Narrative Reviews: Flexible, Rigorous, and Practical. *Journal of Graduate Medical Education*, *14*(4), 414–417. https://doi.org/10.4300/JGME-D-22-00480.1

UNESCO. (2022). *Recommendation on the Ethics of Artificial Intelligence*. UNESCO.

UNESCO. (2023). *Guidance for Generative AI in Education and Research*. https://doi.org/10.54675/EWZM9535

van der Schaaf, M., Baartman, L., Prins, F., Oosterbaan, A., & Schaap, H. (2013). Feedback Dialogues That Stimulate Students' Reflective Thinking. *Scandinavian Journal of Educational Research*, *57*(3), 227–245. https://doi.org/10.1080/00313831.2011.628693

Vistorte, A. O. R., Deroncele-Acosta, A., Ayala, J. L. M., Barrasa, A., Lopez-Granero, C., & Marti-Gonzalez, M. (2024). Integrating Artificial Intelligence to Assess Emotions in Learning Environments: A Systematic Literature Review. *Frontiers in Psychology*, *15*, 1387089. https://doi.org/10.3389/fpsyg.2024.1387089

Wang, P., Chen, F., Wang, D., & others. (2025). Enhancing Students' Dialogic Reflection Through Classroom Discourse Visualisation. *International Journal of Computer-Supported Collaborative Learning*, *20*, 293–315. https://doi.org/10.1007/s11412-024-09443-2

Wang, Q. (2024). The Educational Design Research Approach. In *Designing Technology-Mediated Learning Environments*. Springer. https://doi.org/10.1007/978-981-96-0680-1_8

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are We There Yet? A Systematic Literature Review on Chatbots in Education. *Frontiers in Artificial Intelligence*, *4*, 654924. https://doi.org/10.3389/frai.2021.654924

Yun, H., & Cho, J. (2022). Affective Domain Studies of K-12 Computing Education: A Systematic Review From a Perspective on Affective Objectives. *Journal of Computers in Education*, *9*, 477–514. https://doi.org/10.1007/s40692-021-00211-x

Yusoff, M. S. B. (2019). ABC of Content Validation and Content Validity Index Calculation. *Education in Medicine Journal*, *11*(2), 49–54. https://doi.org/10.21315/eimj2019.11.2.6

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*, *73*, 815–874. https://doi.org/10.1007/s11423-025-10447-4