Rubric Development and Validation for Assessing Tasks' Solving via AI Chatbots

Mohammad Hmoud[1], Hadeel Swaity[1], Eman Anjass[2] and Eva María Aguaded-Ramírez[2]
[1]Faculty of Educational Sciences, An-Najah National University, Nablus, Palestine
[2]Faculty of Educational Sciences, University of Granada, Spain

hmoud.tech@gmail.com (Corresponding author)
hadeelswaity17@gmail.com
emananjass@correo.ugr.es
eaguaded@ugr.es

https://doi.org/10.34190/ejel.22.6.3292

Abstract: This research aimed to develop and validate a rubric to assess Artificial Intelligence (AI) chatbots' effectiveness in accomplishing tasks, particularly within educational contexts. Given the rapidly growing integration of AI in various sectors, including education, a systematic and robust tool for evaluating AI chatbot performance is essential. This investigation involved a rigorous process including expert involvement to ensure content validity, as well as the application of statistical tests for assessing internal consistency and reliability. Factor analysis also revealed two significant domains, "Quality of Content" and "Quality of Expression", which further enhanced the construct validity of the evaluation scale. The results from this investigation robustly affirm the reliability and validity of the developed rubric, thus marking a significant advancement in the sphere of AI chatbot performance evaluation within educational contexts. Nonetheless, the study simultaneously emphasizes the requirement for additional validation research, specifically those entailing a variety of tasks and diverse AI chatbots, to further corroborate these findings. The ramifications of this research are profound, offering both researchers and practitioners engaged in chatbot development and evaluation a comprehensive and validated framework for the assessment of chatbot performance.

Keywords: Rubric development, AI chatbot, Validity, Tasks assessment, Educational technology

1. Introduction

The advent of artificial intelligence (AI) in the digital world has led to sweeping transformations, significantly impacting sectors such as automation, data processing, science research, and predictive analytics (García-Orosa, Canavilhas and Vázquez-Herrero, 2023; Yang, 2022). Among the various facets of AI, chatbots have attracted substantial academic interest due to their intricate algorithms and multifunctional abilities (Kooli, 2023). However, a significant impediment remains: the evaluation of these chatbots is complicated by the lack of a universally accepted and rigorously defined evaluation rubric. The rapid progress and integration of AI chatbots into educational settings have surpassed the development of systematic assessment approaches. This has created a significant gap in empirically evaluating their effectiveness in educational contexts. Addressing this gap requires recognizing the potential of AI chatbots in education and developing and validating assessment rubrics tailored to evaluate their performance accurately.

Recent theoretical developments have revealed that rubrics have become increasingly popular in educational evaluation, serving as assessment tools in various fields, such as appraising the quality of research publications in the medical sector (Moore, Bonnett, and Colbert-Getz, 2021). They also appear essential in verifying the authenticity of content in educational syllabi (Gregori-Giralt and Menéndez-Varela, 2019). These instances highlight the versatility and efficacy of rubrics in diverse educational scenarios. However, the rapid progress and integration of AI chatbots into educational settings have surpassed the development of systematic assessment approaches. This has created a significant gap in empirically evaluating their effectiveness in educational contexts (Smutný and Schreiberova, 2020). Task-solving assessment rubrics provide clear guidelines and quality indicators, which are pivotal to higher education for evaluating student performance across various tasks and assignments (McMurtrie, 2023; Tate, 2023). These allow assessors to measure the quality of students' work and provide valuable feedback (Bradley, Anderson, and Eagle, 2020).

Recent work in the field of AI demonstrated that the increasing prevalence of AI chatbots in education has sparked interest in their potential use for task-solving assessment (McMurtrie, 2023). For instance, AI chatbots like ChatGPT, developed by OpenAI, have demonstrated their potential in assisting students in task-solving

Cite this article: Hmoud, M. et al. 2024. "Rubric Development and Validation for Assessing Tasks' Solving via AI Chatbots", *Electronic Journal of e-Learning*, 22(6), pp 01-17, https://doi.org/10.34190/ejel.22.6.3292

processes such as drafting outlines, revising content, proofreading, and post-writing reflection (Su, Lin, and Lai, 2023; Vicente-Yagüe-Jara et al., 2023).

Based on the short preview above, the primary goal of this study is to fill the current research lacuna by developing a specialized task-solving assessment rubric to evaluate AI chatbots within an educational context. Discussions regarding the extent and boundaries of AI, amplified by recent advancements in machine learning and neural networks, remain contentious (Linardatos et al., 2020). Ongoing debates concern the differentiating factors between human and artificial intelligence and the restrictions of artificiality (Korteling et al., 2021). Despite AI's significant educational potential, it remains predominantly underexplored and undervalued, leading to its metaphorical description as the "Cinderella of the AI story" (Lameras and Arnab, 2021). Concerns regarding data privacy and the skepticism surrounding technology as a panacea have hampered the full integration of AI into mainstream education (Akgun and Greenhow, 2021; Flores-Vivar and García-Peñalvo, 2023).

Nevertheless, the assessment of AI, particularly chatbots, has emerged as a key area of inquiry (Maroengsit et al., 2019). However, The present evaluation methods lack a cohesive framework encompassing all the requisite elements for an exhaustive review (Gregori-Giralt and Menéndez-Varela, 2019). Prior studies on AI chatbots have emphasized the necessity of attributes such as relevance, accuracy, coherence, thoroughness, and efficiency (Moore, Bonnett, and Colbert-Getz, 2021). However, a rubric that integrates these elements comprehensively and uniformly is conspicuously absent from the current literature. By examining the integration of AI chatbots into task-solving assessment, this study aims to identify best practices for educators to effectively incorporate this technology while preserving the authenticity of student work. Thus, this research is driven by the potential benefits of integrating AI chatbots into task-solving assessment in higher education. Consequently, through this exploration of the effectiveness of AI chatbots and the creation of suitable rubrics, it is evident that the overarching goal of the study, in addition to the specific objectives laid out, is to enhance the task-solving assessment process and provide valuable insights for educators and researchers in the field.

1.1 Utilizing Chatbots in Higher Education for Enhanced Learning and Task Solving

The Chatbots have been influential in the field of education because of their significant impact. The digital transformation ushered in by artificial intelligence (AI) has profoundly impacted numerous sectors, with notable effects in automation, data processing, scientific research, and predictive analytics. By the same token, chatbots have garnered significant academic interest in this technological revolution for their complex algorithms and versatile functionalities, especially in the educational sector (García-Orosa, Canavilhas and Vázquez-Herrero, 2023; Kooli, 2023; Yang, 2022). Recent studies by Kim and Lee (2023) and Hmoud et al. (2024) have shown growing appeal, which further highlights the transformative potential of integrating AI chatbots into educational practices. Kim and Lee (2023) delve into Student AI Collaboration (SAC) and its influence on creative tasks, finding that SAC notably enhances creativity, expressivity, and task effectiveness, with the degree of impact influenced by students' attitudes towards AI and their drawing skills. This underlines the importance of adaptive scaffolding in educational AI systems to accommodate diverse student needs, thus improving the learning experience through personalized support.

Building on these insights, Hmoud et al. (2024) examine the motivational dimensions of ChatGPT usage in learning environments. They identify significant effects on student motivation across five core areas: task enjoyment, reported effort, result assessment, perceived relevance, and interaction. Their research notably emphasizes ChatGPT's ability to amplify task enjoyment, indicating that interacting with AI chatbots can greatly enhance students' satisfaction and curiosity, which in turn improves task motivation. However, they also caution about challenges concerning the accuracy of information provided by chatbots, highlighting the essential role of critical evaluation skills among students.

One of the major topics to be investigated in this field is the effect of AI on pedagogical tools used by teachers and educators. Further extending the discourse on the utility of AI in education, Baidoo-Anu and Ansah et al. (2023) illuminate the role of generative AI chatbots as effective pedagogical tools. These chatbots offer conversational assistance, support multiple communication modes, and provide multilingual capabilities, making them cost-effective, scalable, and seamlessly integrated with existing educational technologies. Their ability to offer data analytics and insights enables educators to refine teaching methodologies, showcasing the multifaceted benefits of AI chatbots in enhancing pedagogical practices (Hmoud and Shaqour, 2024). Supporting this viewpoint, Ilieva et al. (2023) highlight the invaluable role of AI chatbots in higher education, especially in providing personalized assistance in advanced and specialized subjects. By promoting self-

directed and independent learning, and facilitating access to scholarly resources, these chatbots significantly support students in their research activities, including assistance with literature reviews and research methodology guidance, thereby fostering academic research and enhancing learning autonomy.

Amidst these technological advancements, a series of recent studies concluded that debates surrounding the distinctions between human and artificial intelligence and concerns about technology as a universal remedy continue to be contentious. However, minor issues have been experienced. Issues such as data privacy and skepticism towards the wholesale integration of AI into mainstream education further complicate the landscape (Akgun and Greenhow, 2021; Flores-Vivar and García-Peñalvo, 2023). Despite these challenges, assessing AI, especially chatbots, in educational contexts has become an essential area of inquiry for many scholars. Yet, the lack of a cohesive framework for comprehensive evaluation points to a significant gap in the literature (Gregori-Giralt and Menéndez-Varela, 2019; Maroengsit et al., 2019).

Based on the preceding, the present work aims to address this gap by advocating for the development of systematic approaches to evaluate the effectiveness of AI chatbots in educational settings, particularly in enhancing task-solving and learning experiences. It highlights the urgent need for research focused on creating reliable methodologies for assessing AI chatbots, thereby contributing to the refinement of their integration into educational frameworks (Smutný and Schreiberova, 2020).

The contributions of Kim and Lee (2023), Hmoud et al. (2024), Baidoo-Anu and Ansah et al. (2023), and Ilieva et al. (2023) offer a nuanced perspective on the implications of using chatbots for task-solving in higher education. They suggest that AI chatbots can significantly enrich student learning experiences by fostering creativity, motivation, and engagement, albeit with an acknowledgment of their limitations. These findings advocate for the development of educational AI that is both adaptable and responsive, capable of supporting diverse learning activities while encouraging critical engagement with content. This comprehensive approach positions chatbots like ChatGPT as invaluable tools in advancing higher education, provided their application is balanced with thoughtful instructional design and rigorous evaluation practices.

1.2 Developing a Rubric for Task-Solving Assessment Through Chatbots: Implications for Teaching, Learning, and Assessment Practices

Integrating rubrics in evaluating tasks facilitated by chatbot technology, such as AI chatbots, is becoming increasingly vital in educational contexts. Rubrics, as structured assessment tools, offer clear and explicit criteria that significantly enhance the evaluation of student work (Brookhart, 2018; Tan, 2020). They also can provide specific standards and expectations for assessing student performance during interactions with chatbots (Bradley, Anderson, and Eagle, 2020). This approach improves students' understanding of assessment criteria and promotes a more objective and consistent assessment process (El-Magd, 2022; Tan, 2020). Furthermore, rubrics facilitate formative feedback and enhance metacognitive skills, guiding students to understand better and meet assignment expectations (De Vera, 2023; Panadero and Jonsson, 2020).

A series of recent studies have elucidated that while rubrics are widely recognized for their benefits in educational evaluation, the absence of a universally accepted evaluation rubric for AI chatbots highlights a significant research gap. (Gregori-Giralt and Menéndez-Varela, 2019; Ilieva et al., 2023; Kooli, 2023; Moore, Bonnett, and Colbert-Getz, 2021; Smutný and Schreiberova, 2020). To address this gap, recent studies by Almasre (2024), Cope, Kalantzis, and Searsmith (2021), and Abbas, Jam, and Khan (2024) have shed light on both the potential enhancements and challenges of incorporating AI into educational assessments. For instance, Almasre (2024) and Cope, Kalantzis, and Searsmith (2021) emphasize AI's capacity to revolutionize educational assessments and facilitate diverse learning pathways through formative assessments and innovative feedback mechanisms. Conversely, Abbas, Jam, and Khan (2024) caution against the potential adverse effects of excessive AI chatbot usage, such as procrastination and diminished academic performance, highlighting the necessity for a balanced integration of AI.

This literature emphasizes the need to develop a Task-Solving Assessment Rubric tailored specifically for AI Chatbots to bolster higher education teaching, learning, and assessment practices. This kind of rubric would amalgamate dynamic evaluation criteria, encompassing relevance, accuracy, and efficiency, thereby establishing a standardized framework for appraising the contributions of AI chatbots within educational settings (Bradley, Anderson, and Eagle, 2020; Lim, 2022; McMurtrie, 2023; Su, Lin, and Lai, 2023; Tate, 2023; Vicente-Yagüe-Jara et al., 2023).

This initiative aims not only to address a noticeable gap in the current literature but also to furnish a structured evaluation tool that resonates with the innovative capabilities of AI chatbots, thereby fostering a

conducive and efficient learning environment. As higher education institutions adapt to the evolving landscape of AI applications, the development and validation of such a rubric will play a pivotal role. It will ensure the integration and utilization of AI chatbots in a manner that nurtures academic advancement and aligns with desired learning outcomes. This is highlighted by studies from El-Magd (2022), Panadero and Jonsson (2020), De Vera (2023), and Tenakwah et al. (2023).

1.3 Conceptual Framework

The conceptual framework for developing a Task-Solving Assessment Rubric for AI Chatbots in higher education seeks to fill a notable gap in existing research, aiming to enhance educational practices by systematically evaluating AI chatbots. This framework is intricately designed around the principles of Competency-Based Learning (CBL) (Henri, Johnson, and Nepal, 2017; Tenakwah et al., 2023) and informed by Brown's (2012) methodology for rubric development, reflecting a comprehensive approach to assessing AI chatbot interactions within educational settings. Central to this framework is establishing explicit goals to define essential competencies required for effective engagement with AI chatbots. These competencies encompass knowledge, skills, and abilities critical to navigating AI technologies, ensuring students are equipped for task-solving activities facilitated by chatbots. The design phase of the rubric, guided by the CBL framework, presents educators with a choice between analytic and holistic assessment methods, emphasizing mastery over critical competencies. This phase is pivotal in structuring rubric categories and scoring ranges that quantitatively measure competency attainment, thereby offering clear and actionable feedback to enhance learning outcomes (Brown, 2012; Henri, Johnson, and Nepal, 2017).

Implementing the rubric involves introducing it to students as a preparatory tool, embodying the concept of preemptive feedback. This approach aligns student efforts with the competencies outlined, setting a clear expectation before task engagement. The subsequent evaluation of student work employs this rubric to provide targeted feedback, highlighting strengths and identifying improvement areas, facilitating a nuanced development of competencies (Brown, 2012). Evaluating the rubric's effectiveness extends beyond its application, encompassing an analysis of its reliability, fairness, validity, and usability. This evaluation is instrumental in refining the rubric, ensuring its adaptability and relevance in the face of rapidly evolving AI technologies. The integration of Brown's structured development process with the CBL approach underpins the framework's robustness, enhancing the precision and utility of the rubric as an educational tool. It emphasizes the role of explicit learning objectives and feedback mechanisms, fostering an environment conducive to effectively applying knowledge and skills in real-world scenarios (Brown, 2012; Henri, Johnson, and Nepal, 2017).

Based on studies by Almasre (2024), Cope, Kalantzis, and Searsmith (2021), Abbas, Jam, and Khan (2024), and others, the framework suggests a balanced use of AI in education. It suggests that AI's shoud be used for innovations while avoiding any negative effects on student engagement and performance. This balanced integration aims to support academic growth and foster a positive learning environment, thereby contributing to the broader educational discourse on AI's role in enhancing learning outcomes and competency development (Bradley, Anderson, and Eagle, 2020; De Vera, 2023; El-Magd, 2022; Gregori-Giralt and Menéndez-Varela, 2019; Ilieva et al., 2023; Kooli, 2023; Lim, 2022; McMurtrie, 2023; Moore, Bonnett, and Colbert-Getz, 2021; Panadero and Jonsson, 2020; Smutný and Schreiberova, 2020; Su, Lin, and Lai, 2023; Tate, 2023; Vicente-Yagüe-Jara et al., 2023).

This study's objectives—developing, validating, and applying a Task-Solving Assessment Rubric for AI Chatbots—underscore the imperative to evaluate AI's educational effectiveness systematically. This research contributes significantly to understanding how AI chatbots can augment task-solving assessment processes in higher education by formulating rubrics that encapsulate essential evaluation elements and align with competency-based learning outcomes. This endeavor seeks to address a critical research void and ensure that AI chatbot integration into educational settings is effective and conducive to fostering critical thinking and originality among students (Lim, 2022).

2. Methodology

An instrumental and descriptive study of validity and reliability of a rubric was carried out. The descriptive assessment measures, often known as rubrics, are among the major current instruments associated with this tendency. These rubrics are based on a graded set of rules that are employed in complete holistic evaluation, with a large capacity for standardizing the assessment of students' performance levels. As a result, they

improve the validity and reliability of performance evaluation, resulting in improved judgment, evaluation, and identification of students' strengths and weaknesses (Stiggins, 1997). Rubrics are indicators that offer explicit and obvious standards for evaluating specific talents or tasks (Reddy and Andrade, 2010; Stanley, 2021b). They are made up of a set of performance indicators and predefined levels of achievement that allow for a systematic and consistent evaluation procedure (Brookhart, 2013). Furthermore, rubrics allow teachers to break down complex skills into individual components, giving students comprehensive feedback and highlighting their strengths and areas for improvement (Panadero and Jonsson, 2013; Stanley, 2021b). The procedure for creating and developing an assessment rubric for solving tasks with chatbots powered by artificial intelligence are explained below:

Phase 1: Identifying the objective of constructing the rubric:

The current task's purpose is to use the assessment rubric as a tool to review and create the criteria that should be met by tasks solved using AI-supported chatbots. As a result, it tries to improve the level and efficiency of these jobs while identifying the strengths and shortcomings in the responses and comments supplied by these platforms (Brookhart, 2018; El-Majd, 2022; Tan, 2020).

Phase 2: Selecting the type of rubric and justifying your choice:

We chose an Analytic Descriptive Rubric after researching educational literature on the design and development of assessment tools. This rubric type was chosen because of its precision, objectivity, realism, and comprehensiveness. It is an alternate assessment technique that focuses on the performance of learners, covering processes and outcomes. It is based on qualitative performance assessment using descriptive rating scales and gives information about the strengths and weaknesses of the many dimensions and components of performance. This data can be used to improve performance by giving precise and detailed feedback to both teachers and students, hence aiding the teaching, and learning processes (Hack, 2015).

Phase 3: Initial Scale Description and Item Development:

The assessment scale's content was generated and formulated using educational literature, studies, and previous research on artificial intelligence and chatbots (Abdul-Kader and Woods, 2015; Brandtzaeg and Følstad, 2017; Hill, Ford, and Farreras, 2015; Jain et al., 2018; Luger and Sellen, 2016). Following that, a brainstorming session was held to discover relevant indications that support the scale.

In terms of item formation, they were expressed in a clear procedural form to promote observation and comprehension by both teachers and students. The pieces were written in short, succinct phrases that have only one meaning. Creating the Assessment Scale's Structure and Organization: The scale's basic form includes 37 criteria, which are as follows:

1. **Accuracy**: Did the chatbot provide a correct and accurate response to the user's query or request?
2. **Relevance**: How relevant was the chatbot's response to the user's question or comment?
3. **Efficiency**: Did the chatbot's response fully answer the user's question or was additional clarification required?
4. **Clarity**: Was the chatbot's response clear and easy to understand, or was it confusing or ambiguous?
5. **Context-Awareness**: Did the chatbot appropriately consider the context of the conversation when providing a response?
6. **Conversational Flow**: Did the chatbot's response fit naturally within the flow of the conversation, or did it disrupt the dialogue?
7. **Empathy**: Did the chatbot respond with an appropriate level of empathy and emotional understanding?
8. **Politeness**: Did the chatbot demonstrate politeness and respect towards the user in its responses?
9. **Speed of Response**: How quickly did the chatbot provide a response?
10. **Coherence**: Were the chatbot's responses consistent throughout the conversation, and did it maintain a coherent line of thought?
11. **Grammar and Spelling**: Were the chatbot's responses free of grammar mistakes and spelling errors?
12. **Personalization**: Did the chatbot personalize its responses based on the user's specific needs and preferences?
13. **Engagement**: Did the chatbot's responses engage the user and promote further conversation?
14. **Error Handling**: How well did the chatbot handle misunderstandings or errors in the user's inputs?
15. **Fallback Strategy**: Was the chatbot able to handle unrecognized inputs or queries gracefully?

16. **Escalation Process**: How effectively did the chatbot hand off the conversation to a human agent when it was unable to assist the user?
17. **Adherence to Guidelines**: Did the chatbot adhere to pre-set guidelines (such as not providing medical, legal, or financial advice unless specifically trained and authorized to do so)?
18. **Security and Privacy**: Did the chatbot properly handle user data, ensuring its security and privacy?
19. **Multilingual Capability**: Can the chatbot effectively communicate in multiple languages as per user requirements?
20. **Appropriateness of Language**: Does the chatbot perfectly appropriate for the task at hand.
21. **Information Verification**: Does the chatbot confirm the accuracy of information provided by the user when necessary?
22. **Domain Knowledge**: How well does the chatbot respond to queries that are specific to the domain it is designed for?
23. **Handling of Complex Queries**: Can the chatbot handle complex queries, or does it only manage simple, straightforward questions?
24. **Self-Correction**: Can the chatbot identify when it's made an error and correct it in real-time?
25. **User Feedback Mechanism**: Does the chatbot have a mechanism to receive and incorporate user feedback?
26. **Simplicity**: Is the chatbot easy to interact with, even for users who aren't very tech-savvy?
27. **Up to date**: Does the chatbot provide responses that are current and up-to-date, especially for time-sensitive or dynamic information?
28. **Scalability**: Can the chatbot handle a large volume of conversations simultaneously without a drop in performance?
29. **Usability**: Is the chatbot user-friendly? Does it offer an intuitive interface?
30. **Informativeness**: Does the chatbot provide a sufficient amount of detail in its responses, without overwhelming the user with unnecessary information?
31. **Adaptability**: Can the chatbot learn from past interactions and improve its responses over time?
32. **Response Diversity**: Does the chatbot vary its responses to avoid sounding too robotic or repetitive?
33. **Comprehensiveness**: Does the chatbot covering all aspects of the task in detail?
34. **Crisis Management**: How effectively does the chatbot manage crisis situations or urgent user needs?
35. **Argument and Evidence**: Does the essay present a clear and compelling argument, and is this argument supported by substantial, reliable evidence from credible sources?
36. **Language and Tone**: Does the essay use appropriate, sophisticated, and consistent language throughout, and does it maintain an academic tone suitable for the context of the assignment?
37. **Creativity and Originality**: Does the essay provide unique insights or original perspectives, and does it demonstrate innovative thinking or creativity in its approach to the topic?

Phase 4: Choosing Assessment of Performance Levels:

To determine the performance level of each facet within each axis, the scale consists of five progressive standard and descriptive levels. Performance scores, ranging from 1 to 5, are assigned based on the extent to which performance indicators are met by the student, with higher scores indicating a better level of standard achievement (Chi, 2013; Jonsson and Svingby, 2007; Stanley, 2021).

Phase 5: Instructions for Creating a Scale:

The scale starts with instructions for the evaluator, which include a description of the scale's structure and components, the levels of performance assessment, and an explanation of the evaluator's responsibilities. An illustrated example of utilizing the scale to evaluate performance outcomes is provided and will be shown later (Chi, 2013; Stanley, 2021).

Phase 6: Verification of Rubric Scale Validity:

The initial version of the assessment scale was provided to 12 professional reviewers working in the field of computing and information technology to determine the necessity for any revisions or alterations to the content of the assessment scale. A Microsoft Form questionnaire was emailed to them in order to solicit their feedback and suggestions on each criterion. For making decisions, the following scale was used: (1) Criterion is unneeded and inappropriate, (2) Criterion is valuable but unnecessary, (3) Criterion is necessary and appropriate, and (4) Criterion is necessary and appropriate. The questionnaire also includes an open-ended place for reviewers to add any relevant remarks (de La Rosa Gómez, Cano, and Diaz, 2019).

Following receipt of the reviewers' replies on the assessment scale criteria, the Content Validity Ratio (CVR) for each criterion was determined using the formula which appear in Equation 1.

$$\mathrm{CVR} = \frac{\mathrm{ne} - \frac{\mathrm{N}}{2}}{\frac{\mathrm{N}}{2}} \qquad (1)$$

Where *ne* is the number of reviewers who say the criterion is "necessary and appropriate", *N* represents the total number of reviewers.

The CVR value is zero when half of the reviewers indicate that the criterion is required and appropriate while the other half do not. The CVR value is 1 when all reviewers agree that the criterion is required and reasonable. The CVR value is between 0 and 0.99 when more than half of the reviewers indicate that the criterion is required and appropriate, but not all of them. CVR is a useful statistical measure for examining items and deciding whether to accept or reject them based on reviewers' decisions. It is widely accepted as a tool for determining content validity (Wilson, Pan, and Schumsky, 2012).

After gathering the reviewers' data, Microsoft Excel software was utilized to complete the essential statistical and mathematical calculations for analysis, such as CVR, CVI, and Kappa. Decisions on item acceptance or rejection were made based on the CVR ratio, which should be greater than 0.667 for each item, considering the number of reviewers (12), as mentioned in the study by Ayre and Scally (2014). By computing the Content Validity Index (I-CVI) for each item, criteria were omitted based on recommendations of a study by Mishra (2017). The following formula used to calculate the I-CVI as show in Equation 2.

$$\mathrm{ItemCVI} = \frac{\mathrm{ne}}{\mathrm{N}} \qquad (2)$$

Where *ne* number of reviewers designating the criterion as "necessary and appropriate", *N* is total number of reviewers.

To determine the consistency of the reviewers for each item, the modified kappa coefficient (*), commonly known as the inter-rater agreement strength, was calculated. Equation 3 was used to arrive to this conclusion:

$$\mathrm{k}^{*} = \frac{\mathrm{ItemCVI} - \mathrm{p_c}}{1 - \mathrm{p_c}} \qquad (3)$$

Where ItemCVI is the content validity index, and pc is the observed percentage of reviewer agreement.

To guarantee the accuracy and agreement of the reviewers, the criteria were subsequently eliminated based on the I-CVI values and the adjusted kappa coefficient (*). These statistical methods were used in accordance with accepted standards for assessment validity and reliability. Mishra's study (2017) offers valuable insights for refining assessment criteria, enhancing validity and ensuring a robust evaluation process.

Phase 7: Verification of Rubric Scale Reliability:

Two techniques were used to evaluate the rubric's reliability:

a) The reliability of the inter-rater agreement was assessed across all items using the Kappa coefficient calculation.

An-Najah University Graduate Studies in Education students were given an assignment to use the chatbots to analyze the advantages and disadvantages of a contemporary teaching approach based on the course's characteristics and requirements. Then, using an assessment tool created by five seasoned professors who acted as independent assessors, an evaluation procedure was carried out. Using the Interrater Reliability technique, this was done to confirm the reliability of the evaluation scale.

The following processes were done in order to determine the overall agreement percentage: a list of each assessor's evaluations for each item on the scale was prepared. Then, for each item, agreement points (1) and disagreement points (0) were determined. The Holsti equation (Holsti, 1970) was used to calculate the sum of agreement points and to obtain the percentage of agreement between the assessors' evaluations:

Agreement Percentage is calculated as follows = 5 * Number of Agreements / (Number of Items Assessed by Assessors 1 + Number of Items Assessed by Assessors 2 + … + Number of Items Assessed by Assessors 5) * 100%.

b) Prior to implementation, the reliability of the assessment scale was further validated using a method known as Intrarater reliability. Thirty students who participated in the previous evaluation were asked to retake the

work when there was ample time in between the tests. The Test/Retest method is one way to assess the accuracy of measurement tools, despite some disadvantages including test familiarity and the potential impact of the students' earlier performance.

Phase 8: Apply Rubric:

The assessment scale was applied to 144 university students who studied at Graduate Studies in Education and completed a final task using chatbots during the second semester of the academic year 2022/2023 after reviewers and evaluators confirmed the validity of the assessment scale and established its reliability.

3. Results

Nine criteria were accepted after determining the Conversion Rate (CVR) for each criterion, as shown in the Table 1.

Table 1: Simplified Table of CVR

Criteria item	Num of Experts Agree an item	*CVR*	Result
Accuracy	12	1	Accept
Relevance	12	1	Accept
Efficiency	12	1	Accept
Clarity	9	0.5	Reject
Context-Awareness	9	0.5	Reject
Conversational Flow	8	0.33	Reject
Empathy	9	0.5	Reject
Politeness	8	0.33	Reject
Speed of Response	6	0	Reject
Coherence	12	1	Accept
Grammar and Spelling	12	1	Accept
Personalization	9	0.5	Reject
Engagement	9	0.5	Reject
Error Handling	8	0.33	Reject
Fallback Strategy	9	0.5	Reject
Escalation Process	8	0.33	Reject
Adherence to Guidelines	7	0.17	Reject
Security and Privacy0	6	0	Reject
Multilingual Capability	6	0	Reject
Appropriateness of Language	9	0.5	Reject
Information Verification	9	0.5	Reject
Domain Knowledge	9	0.5	Reject
Handling of Complex Queries	8	0.33	Reject
Self-Correction	8	0.33	Reject
User Feedback Mechanism	9	0.5	Reject
Simplicity	8	0.33	Reject
Up to date	8	0.33	Reject
Scalability	8	0.33	Reject
Usability	8	0.33	Reject
Informativeness	7	0.17	Reject
Adaptability	8	0.33	Reject

Criteria item	Num of Experts Agree an item	*CVR*	Result
Response Diversity	9	0.5	Reject
Comprehensiveness	11	0.83	Accept
Crisis Management	8	0.33	Reject
Argument and Evidence	10	0.67	Accept
Language and Tone	11	0.83	Accept
Creativity and Originality	12	1	Accept

All the accepted criteria clearly show a degree of agreement among the reviewers that may be regarded as almost flawless, as indicated by the previously stated indicators. When the I-CVI and the modified kappa coefficient for all the criteria are calculated. The values in accordance with the criterion number was shown in Table 2. Instead of calculating each item separately using CVR, the CVI indicator's aggregate result frequently results in a scale that is more effective overall. CVR is a practical statistical technique for assessing the validity of each individual item based on reviewers' evaluations. The total average CVR of all the elements that make up the instrument is represented numerically by the CVI, in contrast (Gilbert and Prion, 2016).

Table 2: Simplified Table of Items CVI, Kappa Coefficients

Criteria item	*I-CVI*	*pc*	*K**	Strength of Agreement
Accuracy	1	0	1	Almost Perfect
Relevance	1	0	1	Almost Perfect
Efficiency	1	0	1	Almost Perfect
Clarity	0.75	0.05	0.74	Substantial
Context-Awareness	0.75	0.05	0.74	Substantial
Conversational Flow	0.67	0.12	0.62	Substantial
Empathy	0.75	0.05	0.74	Substantial
Politeness	0.67	0.12	0.62	Substantial
Speed of Response	0.5	0.23	0.35	Fair
Coherence	1	0	1	Almost Perfect
Grammar and Spelling	1	0	1	Almost Perfect
Personalization	0.75	0.05	0.74	Substantial
Engagement	0.75	0.05	0.74	Substantial
Error Handling	0.67	0.12	0.62	Substantial
Fallback Strategy	0.75	0.05	0.74	Substantial
Escalation Process	0.67	0.12	0.62	Substantial
Adherence to Guidelines	0.58	0.19	0.48	Moderate
Security and Privacy0	0.5	0.23	0.35	Fair
Multilingual Capability	0.5	0.23	0.35	Fair
Appropriateness of Language	0.75	0.05	0.74	Substantial
Information Verification	0.75	0.05	0.74	Substantial
Domain Knowledge	0.75	0.05	0.74	Substantial
Handling of Complex Queries	0.67	0.12	0.62	Substantial
Self-Correction	0.67	0.12	0.62	Substantial
User Feedback Mechanism	0.75	0.05	0.74	Substantial
Simplicity	0.67	0.12	0.62	Substantial

Criteria item	*I-CVI*	*pc*	*K**	Strength of Agreement
Up to date	0.67	0.12	0.62	Substantial
Scalability	0.67	0.12	0.62	Substantial
Usability	0.67	0.12	0.62	Substantial
Informativeness	0.58	0.19	0.48	Moderate
Adaptability	0.67	0.12	0.62	Substantial
Response Diversity	0.75	0.05	0.74	Substantial
Comprehensiveness	0.92	0	0.92	Almost Perfect
Crisis Management	0.67	0.12	0.62	Substantial
Argument and Evidence	0.83	0.02	0.83	Almost Perfect
Language and Tone	0.92	0	0.92	Almost Perfect
Creativity and Originality	1	0	1	Almost Perfect

Therefore, after removing the items on which the reviewers differed, the CVI for the new evaluation scale tool was determined using Equation 4.

$$\text{CVI} = \frac{\sum \text{CVR}}{\text{retained numbers}} = \frac{8.33}{9} = 0.926 \qquad (4)$$

While Davis (1992) contends that a CVI value of 0.80 is ideal, the study by Tilden, Nelson, and May (1990) contends that CVI values should surpass 0.70. As a result, the final CVI value of 0.93 is higher than 0.8, demonstrating the validity, reliability, and applicability of the overall evaluation scale instrument. As a result, the scale was built with the marks, which are shown in Table 3.

Table 3: Rubric Assessment Scale

Criteria	(1) Point	(2) Points	(3) Points	(4) Points	(5) Points
Relevance	Fully and directly addresses all aspects of the task.	Directly addresses most aspects of the task, minor points missing.	Partially addresses the task but lacks important details.	Vaguely related but does not directly address the task.	Unrelated to the task.
Accuracy	Completely accurate.	Almost entirely accurate.	Mostly accurate but contains minor inaccuracies.	Contains a few factual inaccuracies.	Contains several factual inaccuracies.
Efficiency	Efficient and succinct.	Mostly efficient, with minor room for improvement.	Somewhat efficient, with room for improvement.	Could be more succinct or well-structured.	Unnecessarily lengthy or convoluted.
Coherence	Completely coherent and easy to follow.	Very coherent with only minor issues.	Mostly coherent, with a few confusing statements.	Some coherence but difficult to understand.	Largely incoherent or nonsensical.
Comprehensiveness	Extremely comprehensive, covering all aspects of the task in detail.	Quite comprehensive, only missing a few minor points.	Covers most of the task but lacks some details.	Covers a few elements of the task but misses many key points.	Barely touches on the task.
Grammar and Spelling	Excellent grammar and spelling with no errors.	Good grammar and spelling with a few minor errors.	Acceptable grammar and spelling, but with several mistakes.	Poor grammar and spelling with many errors.	Unacceptable grammar and spelling with frequent errors.

Criteria	(1) Point	(2) Points	(3) Points	(4) Points	(5) Points
Argument and Evidence	Clearly articulated, strong argument supported by substantial evidence.	Solid argument with adequate evidence.	Argument present but lacks substantial supporting evidence.	Weak argument with little or insufficient evidence.	Absent or unclear argument with no evidence.
Language and Tone	Highly sophisticated and nuanced language and tone.	Sophisticated language and tone with minor inconsistencies.	Acceptable language and tone but with some inconsistencies.	Inappropriate or inconsistent language and tone.	Poor language use and inappropriate tone.
Creativity and Originality	Unique, insightful, and innovative approach.	Somewhat unique with some insightful thoughts.	Ordinary approach with few insights.	Lack of originality, few insights.	Completely lacks creativity or originality.

According to Holsti (1970), an agreement percentage of 85% or higher indicates strong instrument reliability, whereas one of less than 70% suggests low instrument reliability. The agreement rate between the faculty assessors in grading the thirty students using the assessment scale was higher than 85%, and points to the assessment scale's Good Reliability.

Following the second evaluation of the students' performance, the results of both tests were exceeded 70%, showing high instrument reliability, the evaluation scale nonetheless maintained good reliability as a measurement tool (Streiner, 2003, p.102). The convergent validity of the scale's final form was investigated to ascertain the construct validity of the assessment scale. Exploratory factor analysis employing the principal component analysis approach and Varimax rotation was used to achieve this. Two domains with an eigenvalue greater than one was produced by the analysis, and it had 6 items (Accuracy, Relevance, Coherence, Comprehensiveness, Grammar and Spelling, Argument and Evidence) for the first domain which named "Quality of Content" and 3 items (Efficiency, Language and Tone, Creativity and Originality) for the second domain which named "Quality of Expression". Table 4 shows that the combined factors explained 84.3% of the overall variance.

Table 4: Total Variance Explained

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.437	71.525	71.525	4.498	49.978	49.978
2	1.150	12.777	84.302	3.089	34.324	84.302
3	.645	7.162	91.465			
4	.312	3.470	94.935			
5	.173	1.919	96.854			
6	.118	1.310	98.164			
7	.091	1.015	99.179			
8	.067	.741	99.920			
9	.007	.080	100.000			

Note: ***Extraction Method: Principal Component Analysis.***

According to Gorsuch's (2014) findings, a factor is considered valid when at least three different variables show factor loadings that are more than 0.3. The factor is regarded as insignificant or unimportant if it does not satisfy this requirement. The factor in question was confirmed and accepted in the current investigation

because all of the assessment scale's items showed significant loadings on the factor. This supports the validity of the instrument, which is shown in Table 5.

Table 5: Rotated Component Matrixa

Criteria	Component	
	1	2
Accuracy	0.887	
Relevance	0.689	0.535
Efficiency	0.523	0.762
Coherence	0.835	0.408
Comprehensiveness	0.885	
Grammar and Spelling	0.869	
Language and Tone		0.908
Creativity and Originality		0.921
Argument and Evidence	0.778	0.457

Note: Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

Kaiser-Meyer-Olkin (KMO) test, which assesses the suitability of the sample size for proving the efficacy of factor analysis and whether the partial correlations between variables are minimal, was carried out to ensure the effectiveness of conducting factor analysis on the instrument. Field (2018) claims that high values in the KMO test findings, which are greater than 0.7, indicate that factor analysis would be helpful for our data. After running the test, it was discovered that the KMO value was equal to 0.802, suggesting that the sample was adequate, and that factor analysis had been successful, as indicated in Table 6.

Table 6: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.802
Bartlett's Test of Sphericity	**Approx. Chi-Square**	259.401
	df	36
	Sig.	.000

In order to determine how much each item contributes to measuring the rubric scale as a whole, the Pearson correlation coefficient was determined for each item in respect to the overall rubric scale and that were more than 0.6, which indicates that the rubric as a whole scale is strong (Turney, 2022).

The internal consistency reliability of the estimate scale was examined using the SPSS software by determining the Cronbach's alpha coefficient, which was 0.948, and that indicates good internal Consistency for the Rubric Scale items according to Streiner (2003, p. 102). Also, another estimate scale's reliability was evaluated in SPSS using the split-half method for rubric scale items, which divided them into two equivalent halves using the ODD-EVEN technique, as shown in the table below. The results showed that the Robustness scale was reliable, with Cronbach's alpha coefficients of 0.905 and 0.883 for each half, respectively. Furthermore, the Spearman-Brown coefficient was 0.970, and the Guttman coefficient was 0.966, suggesting outstanding reliability for both sides of the Robustness scale.

4. Discussion

From the results above, key issues emerged that must be addressed. In the first place, the results demonstrated in this chapter match state of the art methods. The main promising finding is that integrating AI chatbots into educational settings marks an essential evolution in teaching and learning methodologies. With AI's capacity to revolutionize sectors from predictive analytics to scientific research (García-Orosa, Canavilhas and Vázquez-Herrero, 2023; Yang, 2022), its application in education promises to enhance teaching and

student engagement. Recent studies show that AI chatbots can potentially improve task-solving assessments, as demonstrated by their ability to enhance creativity, motivation, and student engagement (Hmoud et al., 2024; Kim and Lee, 2023). Additionally, AI chatbots are being used more often to assist students with tasks such as technical and argumentative writing (El-Magd, 2022; Su, Lin, and Lai, 2023). Despite their widespread implementation, a systematic tool for evaluating AI chatbot's efficiency remains elusive (Jain et al., 2018; Maroengsit et al., 2019). However, the lack of a unified framework for evaluating these chatbots poses a significant barrier to their effective integration (Gregori-Giralt and Menéndez-Varela, 2019; Maroengsit et al., 2019). Andrade and Heritage (2017) and Brookhart (2019) assert that rubrics are crucial in setting clear expectations for learner competencies and within any educational tool's assessment framework. They serve as critical instruments in scaffolding both formative and summative assessment processes, enabling the tracking of progress and ensuring alignment with educational standards (Andrade, 2010; Darling-Hammond, Newton, and Wei, 2013).

Rooting its methodology in a rigorous validation process, it demonstrated the rubric's potential applicability and reliability in real-world educational settings and its capacity to standardize the evaluation of AI chatbots' effectiveness in aiding task-solving activities among students.

The intensive investigation into the reliability and authenticity of the proposed tool resulted in consistent scoring outcomes among faculty reviewers. These results are in line with Holsti's (1970) benchmark for instrument strength with an agreement rate of 85% or more. This consistency is pivotal for a precise and equitable evaluation of chatbot functionality. Factor analysis revealed two significant domains: "Quality of Content" and "Quality of Expression." This enhancement of the scale's construct validity established its convergent validity, highlighting the importance of content validity in rubric assessments (Gregori-Giralt and Menéndez-Varela, 2021). Statistical methods such as the Content Validity Ratio (CVR) and a modified kappa coefficient fortified the evaluation tool's reliability and validity (Davis, 1992; Gilbert and Prion, 2016). This suggests that a high Content Validity Index (CVI) value of 0.926 and acceptance of CVR-based criteria substantiate the scale's reliability and validity. A significant finding was the convergent validity, confirmed through factor analysis. The Kaiser-Meyer-Olkin test, as recommended by Hair et al. (2014), verified the factor analysis's effectiveness, strengthening the validity of the assessment scale. Furthermore, the researchers assessed the assessment scale's internal consistency reliability using the Pearson correlation and Cronbach's alpha coefficients, indicating its robustness.

These findings are in accordance with findings reported by Hill, Ford, and Farreras (2015). As an illustration, our investigation adds to the larger conversation on evaluating chatbot performance. The findings extend the discourse on rubric evaluations. The results confirm the reliability and validity of the assessment scale in assessing chatbot performance and open avenues for its use in enhancing chatbot designs and identifying areas for improvement. Here, we compared the results of the proposed method with those of the traditional methods.

These results go beyond previous reports. For instance, the validation of the rubric, underscored by the unanimous acceptance of criteria such as "Accuracy," "Relevance," and "Efficiency," echoes the critical attributes highlighted in the literature for evaluating educational tools (Almasre, 2024; Cope, Kalantzis, and Searsmith, 2021). These attributes are important to ensure that AI chatbots effectively aid pedagogy, fostering both knowledge acquisition and the development of critical thinking skills among students. The rigorous statistical validation, including the high agreement percentages among assessors and the substantial reliability coefficients, attests to the rubric's robustness, aligning with best practices in educational assessment (Bradley, Anderson, and Eagle, 2020). Furthermore, the factor analysis revealing two distinct domains - "Quality of Content" and "Quality of Expression" - validates the conceptual framework proposed by this study and offers critical insights into the multifaceted nature of evaluating AI chatbots. This distinction underscores the complexity of assessing AI chatbots, where factual accuracy and communicative effectiveness are paramount. This detailed approach to evaluation aligns with the Competency-Based Learning (CBL) framework, which prioritizes mastery of essential competencies and underscores the significance of feedback in learning processes (Brown, 2012; Henri, Johnson, and Nepal, 2017). Additionally, the validation process of the rubric revealed its robustness in assessing the intended competencies, reinforcing the importance of a balanced approach to AI integration in educational contexts. Abbas, Jam, and Khan (2024) noted that this balance is crucial to harness AI's innovative capabilities while mitigating potential negative impacts on student learning behaviors and outcomes.

This result ties well with previous studies wherein the current study's findings resonate with and diverge from the literature in several key areas. Unlike previous research by García-Orosa, Canavilhas and Vázquez-Herrero (2023) and Yang (2022), which emphasized the potential and challenges of integrating AI in education without a clear framework for evaluation, this study provides a concrete rubric for assessing the effectiveness of AI chatbots in educational settings. The rubric's focus on "Accuracy," "Relevance," and "Efficiency" parallels the attributes identified by Almasre (2024) and Cope, Kalantzis, and Searsmith (2021) as essential for educational tools. However, our findings extend beyond these attributes by validating a comprehensive set of criteria through empirical methods, addressing a gap in the literature regarding the systematic assessment of AI chatbots. The distinction between "Quality of Content" and "Quality of Expression" identified through factor analysis further deepens the understanding of chatbot assessment. This approach offers a more comprehensive framework than the general discussions on AI chatbot capabilities presented by Kim and Lee (2023) and Hmoud et al. (2024). They emphasized the benefits of AI chatbots in fostering creativity and engagement without specifying mechanisms for evaluation.

Rubrics, as elucidated by Andrade and Heritage (2017) and Brookhart (2019), serve as essential tools in teacher education by making explicit the competencies expected of learners. They facilitate both formative and summative assessments, clearly communicating expectations and tracking progress over time (Darling-Hammond, Newton, and Wei, 2013). Based on this foundational understanding, our study expands the application of rubrics to AI chatbots, portraying these technological tools as 'learners' whose performance and integration into educational practices require careful evaluation.

The study identifies six items: "Accuracy," "Relevance," "Efficiency," "Coherence," "Comprehensiveness," "Grammar and Spelling," "Argument and Evidence," "Language and Tone," and "Creativity and Originality" as key criteria. This selection highlights the importance of content quality and communicative effectiveness in educational AI chatbots, offering a clear response to the needs within educational settings for reliable and engaging AI tools. Applying the Task-Solving Assessment Rubric demonstrates that AI chatbots can significantly support task-solving assessments when evaluated against the identified criteria. They offer a means to enhance learning engagement and creativity and ensure that students interact with AI technologies that meet high accuracy, relevance, and efficiency standards. This finding validates the hypothesis that properly assessed and integrated AI chatbots are valuable in higher education.

Our study's outcomes significantly contribute to the discourse on AI chatbot evaluation in education. We bridge the identified literature gap with a validated assessment tool reflecting the core principles of CBL and Brown's developmental stages (Henri, Johnson, and Nepal, 2017; Tenakwah et al., 2023). While this research marks a critical step toward systematic AI chatbot assessment, it also highlights the necessity for ongoing validation efforts. Future studies should extend this work across varied educational contexts, chatbot types, and learning tasks while also considering the ethical implications of AI in education (Korteling et al., 2021; Lim, 2022; Tate et al., 2023).

5. Conclusion, Recommendations, Limitations, and Implications for Future Research

This research pioneered the developing and validation of a rubric-based assessment scale for evaluating AI chatbot performance in educational settings. By employing rigorous methodology and a thorough validation process, the study has established a foundation for systematically evaluating chatbot effectiveness, filling a significant gap in the literature identified by Jain et al. (2018) and Maroengsit et al. (2019). Insights from this study suggest various directions for future research and practical application.

To enhance the assessment tool's generalizability and applicability across various educational contexts, future research should aim to validate the rubric with a broader, more diverse sample (Nsabayezu et al., 2022). Investigating the assessment scale's effectiveness in evaluating chatbot performance over time could yield insights into the durability and evolution of chatbot effectiveness. There is a pressing need to examine the ethical implications of AI chatbot use, such as potential overreliance on AI, the quality of AI-generated writing, and issues related to literacy assessment (Korteling et al., 2021; Lim, 2022; Tate et al., 2023).

Enhancing the rubric-based assessment with user feedback and performance metrics would give a broader understanding of chatbot effectiveness. This combines users' feelings with performance data (Jain et al., 2018). Furthermore, future research could specifically target the relationship between chatbot performance and learning outcomes, such as knowledge retention, critical thinking, and problem-solving abilities (Liu, 2017).

There are several limitations to this research. First, The sample size and diversity were constrained, potentially impacting the rubric's generalizability across different educational settings and subjects. Additionally, the

study focused on immediate assessment outcomes, leaving room for exploration of long-term impacts and the sustainability of chatbot effectiveness over time. The implications of this research are manifold. The validated rubric offers educators and technologists a practical tool for assessing and improving AI chatbot integration in educational contexts. It underscores the necessity of aligning AI technologies with pedagogical objectives and competency-based learning frameworks, a critical insight that aligns with the foundational principles discussed by Akgun and Greenhow (2021). For the broader field of educational technology, this study highlights the importance of developing reliable, validated tools for evaluating emerging technologies. It highlights how AI chatbots can improve educational experiences when assessed and implemented well. This echoes what Korteling et al. (2021), Lim (2022), and Tate et al. (2023) have said about the importance of considering AI's ethical and practical implications in education. In conclusion, this research contributes significantly to the dialogue on AI chatbot performance assessment in education, presenting a validated assessment tool and outlining a path for future research. As we advance, we must continue to explore, validate, and refine our methods, ensuring that AI chatbots and similar technologies are leveraged to their fullest potential in enriching the educational landscape.

References

Abbas, M., Jam, F.A. and Khan, T., 2024. Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, [online] 21(1). https://doi.org/10.1186/s41239-024-00444-7

Abdul-Kader, S.A. and Woods, J., 2015. Survey on Chatbot Design Techniques in Speech Conversation Systems. *International Journal of Advanced Computer Science and Applications*, [online] 6(7). https://doi.org/10.14569/ijacsa.2015.060712

Akgün, S. and Greenhow, C., 2022. Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI And Ethics*, [online] 2(3), pp.431–440. https://doi.org/10.1007/s43681-021-00096-7

Almasre, M.A., 2024. Development and evaluation of a custom GPT for the assessment of students' designs in a typography course. *Education Sciences*, [online] 14(2), p.148. https://doi.org/10.3390/educsci14020148

Ayre, C. and Scally, A.J., 2014. Critical values for LawsHe's Content Validity ratio. *Measurement and Evaluation in Counseling and Development*, [online] 47(1), pp.79–86. https://doi.org/10.1177/0748175613513808

Baidoo-Anu, D. and Ansah, L.O., 2023. Education in the Era of Generative Artificial Intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, [online] 7(1), pp.52–62. https://doi.org/10.61969/jai.1337500

Bradley, E.J., Anderson, S.E. and Eagle, L., 2020. Use of a marking rubric and self-assessment to provide feedforward to level 5 undergraduate Sport students: student perceptions, performance and marking efficiency. *Journal of Learning Development in Higher Education*, [online] (18). https://doi.org/10.47408/jldhe.vi18.557

Brandtzæg, P.B. and Følstad, A., 2017. Why people use chatbots. In: *Lecture notes in computer science*. [online] pp.377–392. https://doi.org/10.1007/978-3-319-70284-1_30

Brookhart S.M., 2013. *How to create and use rubrics for formative assessment and grading*. [online] CiNii Books. Available at: <http://ci.nii.ac.jp/ncid/BB18336410>.

Brookhart, S.M., 2018. Appropriate criteria: key to effective rubrics. *Frontiers in Education*, [online] 3. https://doi.org/10.3389/feduc.2018.00022

Brown J.D., 2012. *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages*. [online] CiNii Books. Available at: <http://ci.nii.ac.jp/ncid/BB17593429>.

Chi, Y., 2013. Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages. Edited by J. D. Brown. *Language Assessment Quarterly*, [online] 10(2), pp.236–239. https://doi.org/10.1080/15434303.2013.769553

Cope, B., Kalantzis, M. and Searsmith, D., 2020. Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational Philosophy and Theory (Print)*, [online] 53(12), pp.1229–1245. https://doi.org/10.1080/00131857.2020.1728732

Davis, L.L., 1992. Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, [online] 5(4), pp.194–197. https://doi.org/10.1016/s0897-1897(05)80008-4

De La Rosa Gómez, A., Cano, J.M.M. and Díaz, G.A.M., 2019. Validation of a rubric to evaluate open educational resources for learning. *Behavioral Sciences*, [online] 9(12), p.126. https://doi.org/10.3390/bs9120126

De Vera, R.L., 2023. Formative Assessment and Metacognition towards Relevant Learning. *European Journal of Science, Innovation and Technology*, [online] 3(1), pp.54–65. Available at: <https://ejsit-journal.com/index.php/ejsit/article/view/158>.

El-Magd, M.A.A.E.-M., 2022. Text Chatbot Assisted Edublogs for Enhancing the EFL Technical Writing Performance among Computer and Informatics Students. *MağAllaẗ Kulliyyaẗ Al-Tarbiyyaẗ Fi Al-ʿulūM Al-Tarbawiyyaẗ*, [online] 46(2), pp.99–145. https://doi.org/10.21608/jfees.2022.241755

Field, A.P., 2018. *Discovering statistics using IBM SPSS statistics*. [online] Available at: <https://dl.acm.org/citation.cfm?id=2502692>.

Flores-Vivar, J.M.F. and García-Peñalvo, F.J., 2023. Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4). *Comunicar*, [online] 31(74), pp.37–47. https://doi.org/10.3916/c74-2023-03

García-Orosa, B.G., Canavilhas, J. and Vázquez-Herrero, J., 2023. Algorithms and communication: A systematized literature review. *Comunicar*, [online] 31(74), pp.9–21. https://doi.org/10.3916/c74-2023-01

Gilbert, G.E. and Prion, S., 2016. Making sense of methods and measurement: Lawshe's Content Validity Index. *Clinical Simulation in Nursing*, [online] 12(12), pp.530–531. https://doi.org/10.1016/j.ecns.2016.08.002

Gorsuch, R.L., 2014. *Factor Analysis: Classic Edition*. [online] Available at: <https://www.amazon.com/Factor-Analysis-Psychology-Routledge-Editions/dp/1138831999>.

Gregori-Giralt, E. and Menéndez-Varela, J.-L., 2021. The content aspect of validity in a rubric-based assessment system for course syllabuses. *Studies in Educational Evaluation*, [online] 68, p.100971. https://doi.org/10.1016/j.stueduc.2020.100971

Hack, C., 2015. Analytical rubrics in higher education: A repository of empirical data. *British Journal of Educational Technology*, [online] 46(5), pp.924–927. https://doi.org/10.1111/bjet.12304

Henri, M., Johnson, M.D. and Nepal, B., 2017b. A review of Competency-Based Learning: Tools, assessments, and recommendations. *Journal of Engineering Education*, [online] 106(4), pp.607–638. https://doi.org/10.1002/jee.20180

Hill, J., Ford, W.R. and Farreras, I.G., 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, [online] 49, pp.245–250. https://doi.org/10.1016/j.chb.2015.02.026

Hmoud, M. and Shaqour, A., 2024. AIEd Bloom's Taxonomy: A Proposed Model for Enhancing Educational Efficiency and Effectiveness in the Artificial Intelligence Era, *the International Journal of Technologies in Learning*, 31(2), pp. 111–128. https://doi.org/10.18848/2327-0144/cgp/v31i02/111-128

Hmoud, M., Swaity, H., Hamad, N., Karram, O. and Daher, W., 2024. Higher education students' task motivation in the Generative Artificial Intelligence context: the case of ChatGPT. *Information*, [online] 15(1), p.33. https://doi.org/10.3390/info15010033

Holsti, O.R., 1970. Content analysis for the Social Sciences and Humanities. *American Sociological Review*, [online] 35(2), p.356. https://doi.org/10.2307/2093233

Ilieva, G., Yankova, T., Klisarova-Belcheva, S., Dimitrov, A., Bratkov, M. and Angelov, D., 2023. Effects of generative chatbots in higher education. *Information*, [online] 14(9), p.492. https://doi.org/10.3390/info14090492

Jain, M., Kumar, P., Kota, R. and Patel, S.N. 2018. Evaluating and Informing the Design of Chatbots. In Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18). Association for Computing Machinery, New York, NY, USA, [online] pp. 895–906. https://doi.org/10.1145/3196709.3196735

Jönsson, A. and Svingby, G., 2007. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review (Print)*, [online] 2(2), pp.130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Kim, J. and Lee, S.-S., 2022. Are two heads better than one?: The Effect of Student-AI Collaboration on Students' Learning Task Performance. *TechTrends (Online)*, [online] 67(2), pp.365–375. https://doi.org/10.1007/s11528-022-00788-9

Kooli, C., 2023. Chatbots in Education and Research: A Critical Examination of ethical implications and solutions. *Sustainability (Basel)*, [online] 15(7), p.5614. https://doi.org/10.3390/su15075614

Korteling, J.E., Van De Boer-Visschedijk, G.C., Blankendaal, R., Boonekamp, R. and Eikelboom, A.R., 2021. Human- versus Artificial Intelligence. *Frontiers in Artificial Intelligence (Lausanne)*, [online] 4. https://doi.org/10.3389/frai.2021.622364

Lameras, P. and Arnab, S., 2021. Power to the Teachers: An Exploratory review on Artificial intelligence in education. *Information (Basel)*, [online] 13(1), p.14. https://doi.org/10.3390/info13010014

Lim, V. F. 2022. ChatGPT raises uncomfortable questions about teaching and classroom learning, *The Straits Times*, Available at: <https://www.straitstimes.com/opinion/needto-review-literacy-assessment-in-the-age-of-chatgpt>.

Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S., 2020. Explainable AI: A review of Machine Learning Interpretability Methods. *Entropy (Basel. Online)*, [online] 23(1), p.18. https://doi.org/10.3390/e23010018

Liu, S. H. 2017. Trends of digital learning in K-12 schools: A review of the literature on research and practice, *Journal of Educational Technology Development and Exchange*, 10(2), pp. 23-42.

Luger, E. and Sellen, A. 2016. Like having a really bad PA: The gulf between user expectation and experience of conversational agents, in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM.

Maroengsit, W. Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., and Theeramunkong, T. 2019. A survey on evaluation methods for chatbots, in Proceedings of the 2019 7th International Conference on Information and Education Technology. New York, NY, USA: ACM, pp. 111–119. Available at: <https://doi.org/10.1145/3323771.3323824>.

McMurtrie, B. 2023. Teaching: Will ChatGPT change the way you teach? *The Chronicle of Higher Education*. Available at: <https://www.chronicle.com/newsletter/teaching/2023-01-05>.

Meo, S.A., Al-Masri, A.A., Alotaibi, M., Meo, M.Z.S. and Meo, M.O.S., 2023. CHATGPT Knowledge Evaluation in Basic and Clinical Medical Sciences: Multiple Choice Question Examination-Based Performance. *Healthcare (Basel)*, [online] 11(14), p.2046. https://doi.org/10.3390/healthcare11142046

Mishra, S., 2017. Open educational resources: removing barriers from within. *Distance Education*, [online] 38(3), pp.369–380. https://doi.org/10.1080/01587919.2017.1369350

Moore, K.B., Bonnett, R. and Colbert-Getz, J.M., 2020. A process and rubric for a group to review the quality of a Medical Education Course/Clerkship. *MedEdPORTAL*. [online] https://doi.org/10.15766/mep_2374-8265.10911.

Nsabayezu, E., Mukiza, J., Iyamuremye, A., Mukamanzi, O.U. and Mbonyiryivuze, A., 2022. Rubric-based formative assessment to support students' learning of organic chemistry in the selected secondary schools in Rwanda: A technology-based learning. *Education and Information Technologies*, [online] 27(9), pp.12251–12271. https://doi.org/10.1007/s10639-022-11113-5

Panadero, E. and Jönsson, A., 2013. The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, [online] 9, pp.129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Reddy, Y.M. and Andrade, H.L., 2010. A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, [online] 35(4), pp.435–448. https://doi.org/10.1080/02602930902862859

Sedoc, J., Ippolito, D., Kirubarajan, A., Thirani, J., Ungar, L.H. and Callison-Burch, C., 2018. ChatEval: a tool for the systematic evaluation of chatbots. *Association for Computational Linguistics*, [online] pp.42–44. https://doi.org/10.18653/v1/w18-6709

Stanley, T. 2021. What is a rubric?, in *Using RUBRICS for Performance-Based Assessment*. New York: Routledge, pp. 9–20.

Stanley, T., 2021. How to write a rubric. In: *Routledge eBooks*. [online] pp.53–68. https://doi.org/10.4324/9781003239390-6

Stiggins, R.J. 1997. Student-centered classroom assessment, *New York: Merrill*. Available at: <https://files.nwesd.org/depts/eadmin/Admin_Website/CIT-CL/LiteratureReference/JournalArticles/Student-Centered-Classroom-Assessment_Stiggins.pdf>.

Streiner, D.L., 2003. Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, [online] 80(1), pp.99–103. https://doi.org/10.1207/s15327752jpa8001_18

Su, Y., Lin, Y.G. and Lai, C., 2023. Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, [online] 57, p.100752. https://doi.org/10.1016/j.asw.2023.100752

Tan, K., 2020. *Assessment rubrics decoded*. [online] *Routledge eBooks*. https://doi.org/10.4324/9780429022081

Tate, T., Doroudi, S., Ritchie, D., Xu, Y. and Warschauer, M., 2023. Educational Research and AI-Generated Writing: Confronting the Coming Tsunami. *EdArXiv*. [online] https://doi.org/10.35542/osf.io/4mec3

Tenakwah, E.S., Boadu, G., Tenakwah, E.J., Parzakonis, M., Brady, M., Kansiime, P., Said, S., Ayilu, R.K., Radavoi, C.N. and Berman, A.L., 2023. Generative AI and Higher Education Assessments: A Competency-Based Analysis. *Research Square (Research Square)*. [online] https://doi.org/10.21203/rs.3.rs-2968456/v2

Tilden, V.P., Nelson, C.A. and May, B.A., 1990. Use of qualitative methods to enhance content validity. *Nursing Research*, [online] 39(3), p.172???175. https://doi.org/10.1097/00006199-199005000-00015

Turney, S., 2022. *Pearson Correlation Coefficient (r) | Guide & Examples*. [online] Scribbr. Available at: <https://www.scribbr.com/statistics/pearson-correlation-coefficient>.

Vicente-Yagüe-Jara, M.-I., Martínez, O.L., Navarro-Navarro, V. and Cuéllar-Santiago, F., 2023. Writing, creativity, and artificial intelligence. ChatGPT in the university context. *Comunicar*, [online] 31(77). https://doi.org/10.3916/c77-2023-04

Wilson, F.G., Pan, W. and Schumsky, D.A., 2012. Recalculation of the critical values for LawsHE's content validity ratio. *Measurement and Evaluation in Counseling and Development*, [online] 45(3), pp.197–210. https://doi.org/10.1177/0748175612440286

Yang, W., 2022. Artificial Intelligence education for young children: Why, what, and how in curriculum design and implementation. *Computers & Education: Artificial Intelligence*, [online] 3, p.100061. https://doi.org/10.1016/j.caeai.2022.100061