Elinvo (Electronics, Informatics, and Vocational Education)
Vol 11, No. 1, 2026, pp. 85-99
ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

Evaluating LLM-Based Institutional Information Chatbot Responses Using a Preliminary Human-Scored Analytic Rubric and Automatic Metrics

Rosni Lumbantoruan [1*], Arnaldo Marulitua Sinaga [2], Markus Pardianto Hutagalung [3], Priskila Christine Natalia Parapat [4], Mutiara Teccalonica Simanjuntak [5]

[1,2,3,4,5] Institut Teknologi Del, Toba, Laguboti, Indonesia
[1] rosni@del.ac.id*; [2] aldo@del.ac.id, [3] iss21013@students.del.ac.id, [4] iss21020@students.del.ac.id, [5] iss21029@students.del.ac.id,
* corresponding author

Article Info

Article history:
Received April 27, 2026
Revised June 10, 2026
Accepted June 24, 2026
Available July 12, 2026

Keywords:
Analytic rubric; automatic metrics; chatbot evaluation; human evaluation; System Usability Scale

Abstract

Large language models (LLMs) are increasingly used for answering institutional information enquiries in higher education, yet the quality of responses is not straightforward to evaluate, as factual accuracy alone does not account for interactive qualities such as clarity, conversational flow, error handling and personalisation. This pilot study developed, preliminary examined a human-scored analytic rubric for assessing ChatGPT responses in a higher-education institutional-information setting and explored the alignment of selected rubric scores with reference-based automatic metrics. We designed a literature-informed rubric comprising 15 criteria across five conceptual domains. Of these, 14 criteria were operationalised through 42 rubric questions, and system usability was rated separately using the System Usability Scale. Seventy-five qualified students from one higher-education institution rated the chatbot responses using a four-point scale. Preliminary evidence at the item level was provided by item-total correlations and Cronbach's alpha, while Pearson and Spearman correlations were used to investigate the alignment between human scores and reference-based metrics namely ROUGE-1, ROUGE-2, ROUGE-L and SacreBLEU for four content-oriented criteria. The results showed positive but partial agreement between human ratings and referenced-based metrics, with stronger agreement for clarity and up-to-date response than for accuracy and relevance. These findings suggest that reference-based metrics can complement, but not replace, human evaluation for insitutional information chatbot assessment. The study was confined to one institution and did not incorporate inter-rater reliability, expert validation or factor analysis. Thus, the rubric should be seen as a preliminary evaluation instrument rather than a fully validated scale.

INTRODUCTION

Chatbots have become an integral part of digital interaction such as customer service and education, as they are able to deliver immediate access to information [1][2]. However, the analysis of chatbot responses involves more than checking factually correctness but also evaluating clarity, contextual relevance, politeness, accuracy, and user satisfaction [3][4]. Previous chatbot evaluations have been applied in domains such as healthcare [5] and conversational systems [6], [7], but many evaluations heavily rely on either human evaluation or automatic metrics without considering the alignment between them. Human evaluation is essential as it can capture interactional and contextual qualities, whereas automatic metrics offer scalable but narrower evidence based mainly on reference overlap. Dialogue-system assessment frameworks such as PARADISE [8] indicate that conversational

agents should be assessed not only by task success but also by interaction quality and user satisfaction. Recent LLM assessment has also introduced semantic, model-based, and multi-dimensional evaluation such as BERTScore [9], G-Eval [10], and HELM [11] which suggest that chatbot and LLM evaluation should move beyond exact lexical overlap. However, these aforementioned approaches still require human-centred assessment, especially when criteria such as politeness, personalisation, usability, and flexibility involve user perception and contextual judgement.

In contrast to the previous approaches, this paper deals with a narrower but practically important problem namely how to design a preliminary human-scored rubric for institutional chatbot responses in higher education and to investigate if selected reference-based automatic metrics offer complementary evidence for some content-oriented criteria. Thus, this study aims to address the following research questions: (1) What are the literature-informed criteria for developing a preliminary analytic rubric to evaluate responses of LLM-based institutional chatbots in higher education? (2) To what extent do the human evaluation on selected content-oriented rubric criteria align with reference-based automatic metrics? To address these research questions, based on literature study, we design the preliminary assessment rubric consisting of 15 criteria [12][13] and organize them into 5 functional domains: content quality, technical fluency, system robustness, personal experience, and dynamic flexibility. We select the qualified computer science students in a particular campus who widely use ChatGPT to support their learning activities [14]. Unlike previous verified and universal evaluation standard chatbot instruments such as the Artificial Intelligence Performance Instrument (AIPI) [15], this study introduces a preliminary human-scored analytic rubric for LLM generated educational information responses, separating response quality evaluation from system-level usability using System Usability Scale (SUS), and investigates the alignment between human rubric scores and automatic reference-based metrics.

In summary, the contribution of this paper is threefold. First, it proposes a literature-informed, preliminary human-scored analytic rubric to evaluate LLM-based instructional chatbot responses in an institutional information setting. Second, it separates response-quality evaluation from SUS by using 42 rubric questions for 14 criteria and SUS for usability. Third, it evaluates whether selected content-oriented human ratings show partial correlation with reference-based automatic metrics. Thus, instead of proposing a universally validated tool for assessing chatbots, this study proposes a preliminary analytic rubric and empirical evidence which may inform future validation studies.

METHODS

Research Design and Scope

This study used a pilot instrument-development and evaluation design to develop and preliminarily examine an analytic rubric for evaluating ChatGPT responses to institutional information questions in a private higher education setting. The evaluation employed both human rubric-based evaluation and automatic reference-based metrics to investigate the correlation between human judgements and metric-based scores for a subset of criteria focused on content. The research methodology started with the identification and grouping of chatbot evaluation criteria, followed by rubric items development, participant-based evaluation, automatic metric calculation, and statistical analysis. The research workflow consisted of six stages, namely grouping and identifying evaluation criteria, developing institutional question-answer pairs, creating rubric items and scoring guides, evaluation based on the participants, calculation of automatic metrics, and statistical analysis.

Figure 1 depicts the overall research stage, covering rubric development, instrument application, and human–metric alignment analysis. The following subsections elaborate the data sources, question–answer preparation, evaluation instruments, participants, automatic metrics, and statistical procedures.

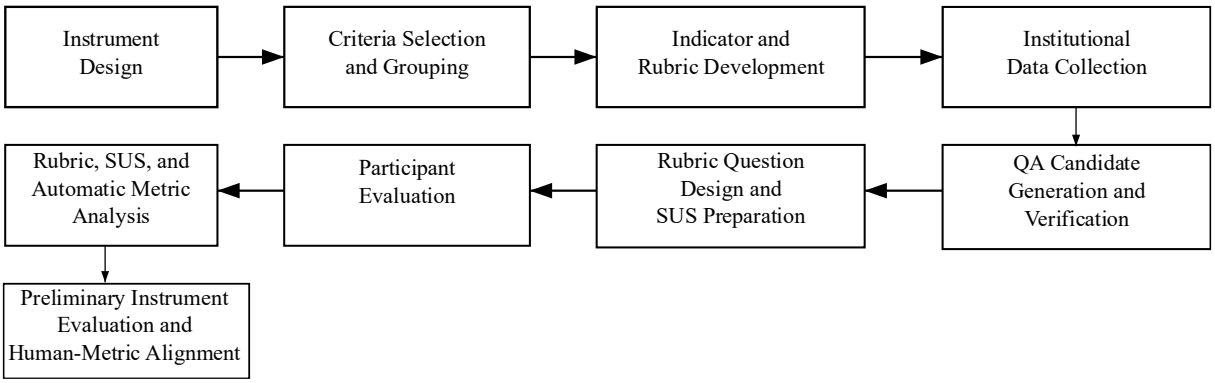

Figure 1. Research workflow for rubric development, preliminary instrument evaluation, and human–metric alignment analysis

Recalling that the study was a pilot evaluation, the analysis was limited to descriptive statistics, item-total correlation, internal consistency, SUS scoring, and human-metric correlation. No inter-rater reliability, expert-based content validity assessment, exploratory factor analysis, or confirmatory factor analysis was conducted. Therefore, the five-domain structure should be seen as a literature-informed conceptualization of concepts, not an empirically supported measurement model.

Data Collection and Question–Answer Preparation

Institutional data covering the institution's history, vision, mission, academic programs, campus facilities, student organizations, academic collaborations and training programs, were collected from publicly available institutional webpages and relevant Wikipedia pages. Then, candidate question–answer pairs were generated by extracting key information using DeepSeek-Coder model [16]. The generated QA pairs, however, were not directly used as final evaluation items considering the code-oriented nature of the model. DeepSeek-Coder's instruction-tuned variant and long-context ability enable it to process paragraph-level prompts and produce structured textual outputs. Furthermore, DeepSeek-Coder was pretrained on top of the DeepSeek-LLM 7B checkpoint to achieve better natural language understanding [16].

Specifically, ROUGE-L was employed as a preliminary lexical screening indicator to approximate the overlap between generated answers and the corresponding source passages. However, lexical overlap does not guarantee factual correctness or contextual relevance, so the final inclusion was determined by manual verification of the original institutional source text. Manual verification was performed by checking whether each question could be answered from the source passage, whether the answer was factually supported by the source, whether the pair was relevant to institutional information, and whether the wording was clear, non-redundant, and unambiguous. Only QA pairs that met these criteria were kept and used in the next stage of rubric question design.

Rubric Development and Instrument Benchmarking

In this study, we chose an analytic rubric for its ability to evaluate the chatbot response in terms of a set of observable components of performance rather than a single overall score. Rubric construction was performed by specifying criteria, indicators and score levels with defined grading systems [17], a 4-point Likert-type scale, with higher scores indicating higher fulfilment of the criterion and lower scores indicating lesser fulfilment [18].

The rubric-development process began by adopting 37 chatbot performance criteria identified from previous research [5]. These criteria were then reorganized based on their definition, conceptual overlap, and relevance to institutional chatbot-response evaluation. Criteria with similar meanings were grouped to reduce redundancy and to form a more coherent structure. Specifically, each criterion

definition was analysed from three perspectives namely general dictionary meanings, linguistic or formal academic interpretations, and chatbot-evaluation research papers. The process helped to establish whether each criterion was mostly about factual quality, conversational behavior, error management, user experience, or adaptability.

The analysis showed that several criteria conceptually overlapped. For example, coherence [4], conversational flow [19], and context awareness [20] are all assessing whether a conversation progresses logically and smoothly. Therefore, conceptually related criteria were grouped under a broader concept rather than treated as entirely separate criterion. The definitions were then summarized by highlighting key terms and adjusting the definitions to the context of educational chatbot-response evaluation. The criteria within each group were compared, filtered for relevance to interactive and functional evaluation, and organised into five literature evaluation domains. It is important to note that this process did not involve formal statistical validation or expert validation, but rather a qualitative analysis based on definitions and contextual relevance conducted in this study.

Based on the analysis and selection process of the initial 37 criteria, this study proposes a preliminary analytic rubric consisting of 15 literature-informed criteria grouped into five domains: accuracy and clarity of information; response speed and conversational flow; error management and complex query handling; user experience and personalisation; and adaptability. Meanwhile, the 15 criteria were accuracy, relevance, clarity, up-to-date response, speed of response, context awareness, conversational flow, error handling, fallback strategy, handling of complex queries, personalisation, politeness, usability, response diversity, and adaptability.

The study conducted a conceptual benchmarking of the proposed rubric using the validated Artificial Intelligence Performance Instrument (AIPI) [15], an instrument for evaluating the performance of ChatGPT in clinical case management. The comparison was restricted to design-level considerations including purpose, criteria coverage, scoring strategy, and reported validation evidence, as depicted in Table 1.

Table 1 Instrument-level comparison between AIPI and the proposed rubric

Aspect	AIPI	Proposed rubric
Purpose	Validated instrument for rating AI/ChatGPT performance in clinical case management	Preliminary human-scored rubric for LLM-based educational chatbot responses
Evaluation context	Outpatient cases; diagnosis, management, treatment	Institutional information queries in higher education
Orientation	Expert, criterion-referenced, content/decision quality	User-centred, interaction + content quality, and metric alignment
Criteria coverage	9 items; 4 subscores: patient feature, diagnosis, additional examination, treatment	15 criteria; 5 domains: accuracy & clarity; speed & flow; error management & complex query handling; user experience & personalisation; adaptability
Scoring/scale	Weighted categorical scoring, total 0–20	4-point Likert for 42 rubric items; SUS scored separately
Usability treatment	Absent	Separate SUS for system-level usability
Automatic metric	Absent	ROUGE-1/2/L, SacreBLEU for 4 content-oriented criteria
Validation status	Internal consistency, test–retest, internal validity, convergent validity, and inter-rater concordance	Item-total correlations, $\alpha = 0.699$, SUS, and human–metric correlation
Sample/population	45 cases; expert physician raters	75 student respondents from one institution

Aspect	AIPI	Proposed rubric
Strengths	Expert-built; non-Likert; real-case benchmark; stronger psychometric package	Broader interaction coverage; usability separated clearly; human–metric alignment analysed
Limitations	Clinical-domain specific; no UX/dialogue criteria; small sample; modest convergent validity	Preliminary only; single institution; no expert validation or factor analysis

Table 1 shows that AIPI is a strong psychometric evidence since it reports expert-led development, internal validity, convergent validity, test-retest reliability, and inter-rater concordance [15]. Therefore, this study does not claim that the proposed rubric is psychometrically stronger than AIPI but to show the importance of a human-scored analytic rubric to assess some criteria that cannot be adequately addressed by automatic metrics alone. This study makes a specific contribution by offering human-scored rubric for institutional educational chatbot responses, expanding the evaluation to include interaction-oriented criteria beyond the scope of AIPI, separating response-quality evaluation from SUS-based usability evaluation, and investigating human-metric alignment for a selection of content-oriented criteria.

Table 2 shows the operational indicators used to guide rubric scoring. These indicators define observable evidence for each criterion and support consistency in human evaluation. To prevent confusion between indicators and evaluation questions, the former are presented as operational descriptors, while the latter consists of 42 rubric-based items representing 14 criteria. Usability was treated separately using SUS because it reflects system-level ease of use rather than the quality of individual chatbot responses.

Table 2 Evaluation criteria and operational indicators

Criteria	Indicator
Accuracy	1. Answers correspond to facts or correct data. 2. No misinformation or misleading statements are present. 3. No contextual discrepancies occur within answers in a single conversation session [21][22].
Relevance	1. Answers align with the main topic of the question. 2. Information provided remains focused and does not diverge to other topics. 3. Addresses the core issue raised by the user [23][4].
Clarity	1. Sentences are constructed with proper grammatical structure. 2. Vocabulary used is unambiguous and easy to understand. 3. Explanations are presented logically and clearly without causing confusion [24][23].
Up-to-date Response	1. Information reflects the latest conditions or data. 2. Does not refer to outdated sources, links, or facts. 3. Explicitly references current times or recent events [21][4].
Speed of Response	1. The time lag between question and answer is fast, typically less than 3 seconds. 2. No delays even with complex questions. 3. Responds promptly to follow-up user inputs without time gaps [25].
Context Awareness	1. Understands the context of previous questions. 2. Adjusts responses based on information already provided by the user. 3. Avoids generic responses when specific context is available [20], [26].
Conversational Flow	1. The conversation flows logically from previous dialogue. 2. No sudden breaks in context occur. 3. Answers avoid unnecessary repetition [6].
Error Handling	1. Able to recognize unclear questions or typos. 2. Provides polite replies when the question is not understood. 3. Requests clarifications or additional information as needed [27].
Fallback Strategy	1. Offers informative fallback statements when the answer is unknown.

Criteria	Indicator
	2. Provides alternative help options (links or other resources). 3. Suggests contacting an admin or operator [12][22].
Handling of Complex Queries	1. Understands and responds to multi-topic or layered questions. 2. Answers fully without omitting any part of the question. 3. Differentiates and addresses distinct parts within a single query [28].
Personalisation	1. Refers to personal information of the user (if available, such as their name). 2. Tailors responses based on interaction history. 3. Adopts a communication style appropriate to the user (formal or informal) [29].
Politeness	1. Uses polite greetings and appropriate language. 2. Avoids negative phrases, blaming, or defensive tones. 3. Shows empathy or cooperative intent when responding to user complaint [30].
Usability	1. Interface and layout are easy to navigate. 2. Provides clear instructions on how to interact with the chatbot. 3. Does not require complicated steps to obtain answers [31][32].
Response Diversity	1. Provides varied answers to similar questions. 2. Avoids literal repetition of the same response. 3. Able to present information differently depending on the context [33].
Adaptability	1. Adjusts responses to the user's questioning style. 2. Improves answer quality as the interaction continues. 3. Demonstrates the ability to learn or adapt its approach [7].

The major overlaps between the proposed rubric and AIPI are in the criteria related to accuracy, contextual grounding, and complex-case handling. On the other hand, clarity, up-to-date response, speed, conversational flow, fallback strategy, politeness, personalisation, usability, response diversity and adaptability are mostly outside the scope of AIPI [15] and therefore represent additional coverage within the existing rubric.

Participants and Evaluation Procedure

The participants were 75 active students of a particular higher education institution. The selection process started by distributing an eligibility questionnaire to 130 students. Of these, 98 candidates met the eligibility criteria [8], which included frequent use of chatbot and relevant experience in interacting with chatbot-based information services. From the 98 eligible candidates, 75 respondents were selected using simple random sampling.

The number of respondents was determined by the pool of available eligible respondents and the applied nature of the study. Hence, this study was designed as a pilot study of the proposed instrument and does not conduct exploratory factor analysis (EFA) or confirmatory factor analysis (CFA) to validate the factor structure of the rubric.

Evaluation Instruments, Scoring, and Procedure

Each response item was scored using a 4-point Likert Scale, where 1 represented the lowest level of performance and 4 represented the highest level of performance for the relevant criterion. The mean score for each question and criterion was calculated to represent the tendency of the responses. Following [34], mean scores were interpreted using four intervals whereas scale of 1.00 – 1.75 means very weak accuracy, 1.76 – 2.50 means weak accuracy, 2.51 – 3.25 means good accuracy, and 3.26 – 4.00 means very accurate.

The questions for each criterion were developed by two approaches: (1) directly selecting questions from the verified QA pairs were used for accuracy, relevance, clarity, and up-to-date response, or (2) modified or newly constructed questions for particular evaluation needs, were used for context awareness, error handling, fallback strategy, and handling of complex queries. This approach enabled

factual criteria to be tested against validated reference information and interaction-oriented criteria to be tested in more contextualised scenarios.

The rubric-based evaluation consisted of 42 rubric-based evaluation items representing 14 criteria with 3 questions for each criterion. Unlike other criteria, the usability criterion was evaluated using the System Usability Scale (SUS) as it related to the ease of use and interaction experience at the system level. This design allowed the study to keep a structured response quality rubric and to use a standardised instrument to assess usability.

Each rubric question was scored using a pre-established scoring guide to help raters determine how well the chatbot response met the expected level of performance for each criterion. This structured approach was used to support consistency in human evaluation and to ensure that the functional and interactive aspects of the chatbot were examined in a systematic way.

The experiment was carried out in a controlled evaluation setup where each respondent was asked to interact with the chatbot using the rubric. After every chatbot interaction, respondents evaluated the response through an online questionnaire. The collected responses were then used to analyse descriptive performance, preliminary item-total analysis, internal consistency analysis, SUS scoring, and human–metric alignment analysis.

Automatic Reference-Based Evaluation

This study conducted an automatic reference-based evaluation using ROUGE-1, ROUGE-2, ROUGE-L, and SacreBLEU in order to complement the human-scored analytic rubric for 4 criteria, namely accuracy, clarity, relevance, and up-to-date response since these criteria can be evaluated by comparing the chatbot-generated answer with a verified reference answer [7]. Other criteria such as politeness, usability, personalisation, response diversity, adaptability, and other interaction-oriented aspects cannot be adequately measured using lexical-overlap metrics such as ROUGE and BLEU. These criteria require human judgement because they involve perception, interaction quality, and contextual interpretation. For each of the four selected criteria, three evaluation questions were used and for each question, up to four chatbot responses were selected, representing different human Likert-scale scores from 1 to 4. Therefore, the maximum number of question-answer pairs was 48, calculated as 4 criteria × 3 questions × 4 score categories. However, 4 expected response pairs were unavailable because no respondent assigned certain Likert scores, such as 1 or 2 to those answers. As a result, the final dataset used for automatic metrics consisted of 44 question-answer pairs.

We then compare each chatbot generated answer with the associated validated reference answer using the four automatic metrics. These metric scores were then employed in the human-metric correlation analysis as discussed in the next Statistical Analysis Subsection.

Statistical Analysis

Questionnaire data were analysed with descriptive statistics, item-total correlation [35], internal consistency analysis, SUS scoring, and human–metric correlation analysis. Descriptive statistics such as frequency distribution, mean and standard deviation were used to summarise the ratings of respondents for each evaluation criterion.

The preliminary validity of the items was assessed over 42 rubric-based items using item-total correlations between each item score based on the rubric and the total rubric score [35]. The critical value of r for interpreting the results was 0.227 for 75 respondents and a 0.05 level of significance.

Cronbach's alpha was used to calculate the internal consistency of the 42 rubric-based items. Following [35], the coefficient was treated as reliability evidence and interpreted cautiously rather than as proof of construct validity. In this preliminary evaluation research, values around 0.70 were found to be moderately acceptable, although with the caveat that alpha should not be taken mechanically or as

the only indication of instrument quality [35]. On the other hand, the SUS score was calculated separately using the standard SUS scoring method.

The alignment of human evaluation and reference-based automatic metrics were analysed using Pearson and Spearman correlation coefficients. Pearson correlation was used to examine linear association while Spearman correlation was used because human scores were based on non-linear ordinal Likert scale. Correlations were calculated at two levels namely across all 44 question–answer pairs and separately for each of the four content-oriented criteria.

Ethical Considerations

Participation in this study was voluntary. Prior to the evaluation, respondents were informed about the purpose of the study, the evaluation tasks to be performed, and the use of their responses for research purposes. Participants were free to withdraw their participation from the research and no academic penalty or reward as the consequences. Data collected in this study were anonymised prior to analysis and the manuscript did not report any personally identifiable information. The authors declare that there was no conflict of interest in conducting the study.

RESULTS

This section presents the descriptive rubric results, SUS usability scores, item-total correlation results, internal consistency analysis, automatic reference-based metric scores, and human–metric correlation results. Recalling that this study was designed as a pilot study of a preliminary rubric, the inferential results reported below should be interpreted as exploratory evidence.

Descriptive Results of Human Rubric Scores

Descriptive and frequency statistical analyses were conducted on the questionnaire data from 75 respondents to obtain a measurable overview of the chatbot's performance. The results of the analysis for each criterion are presented in Table 3.

Table 3 Descriptive and frequency statistics of the of the 42 rubric-based items

Criteria	Q	Point 1 (N)	Point 2 (N)	Point 3 (N)	Point 4 (N)	Mean	Std. Dev
Accuracy	Q1	2	1	12	60	3.7333	0.62240
	Q2	3	2	19	51	3.5733	0.73839
	Q3	2	0	7	66	3.8267	0.55443
Relevance	Q1	22	23	14	16	2.3200	1.11695
	Q2	6	3	7	59	3.5867	0.90185
	Q3	8	7	20	40	3.2267	1.00772
Clarity	Q1	2	2	17	54	3.6400	0.67062
	Q2	0	4	27	44	3.5333	0.60030
	Q3	4	12	9	50	3.4000	0.94440
Up-to-date Response	Q1	3	2	12	58	3.6667	0.72286
	Q2	21	15	14	25	2.5733	1.22114
	Q3	9	3	7	56	3.4667	1.03105
Speed of Response	Q1	13	26	26	10	2.4400	0.93346
	Q2	13	13	33	16	2.6933	0.99964
	Q3	10	26	30	9	2.5067	0.87570
Context Awareness	Q1	4	7	17	47	3.4267	0.87261
	Q2	0	16	26	33	3.2267	0.78108
	Q3	3	9	12	51	3.4800	0.85992

Criteria	Q	Point 1 (N)	Point 2 (N)	Point 3 (N)	Point 4 (N)	Mean	Std. Dev
Conversational Flow	Q1	11	8	28	28	2.9733	1.03940
	Q2	26	4	12	33	2.6933	1.34540
	Q3	11	9	23	32	3.0133	1.07167
Error Handling	Q1	22	6	8	39	2.8533	1.33248
	Q2	5	7	41	22	3.0667	0.81096
	Q3	13	10	35	17	2.7467	1.00126
Fallback Strategy	Q1	15	2	7	51	3.2533	1.20912
	Q2	12	6	9	48	3.2400	1.14891
	Q3	4	0	8	63	3.7333	0.72286
Handling of Complex Queries	Q1	0	5	4	66	3.8133	0.53760
	Q2	1	3	11	60	3.7333	0.60030
	Q3	2	1	11	61	3.7467	0.61717
Personalisation	Q1	6	3	16	50	3.4667	0.90544
	Q2	3	6	19	47	3.4667	0.81096
	Q3	8	5	15	47	3.3467	1.00664
Politeness	Q1	0	0	6	69	3.9200	0.27312
	Q2	1	2	53	19	3.2000	0.54525
	Q3	0	0	12	63	3.8400	0.36907
Response Diversity	Q1	14	12	21	28	2.8400	1.12754
	Q2	12	24	18	21	2.6400	1.06085
	Q3	24	20	14	17	2.3200	1.25267
Adaptability	Q1	0	2	11	62	3.8000	0.46499
	Q2	0	1	9	65	3.8533	0.39227
	Q3	1	2	23	49	3.6000	0.61512

Table 3 shows that politeness, adaptability, ability to handle complex questions, and accuracy received higher average scores among all evaluated criteria. This means that in most of the assessed interactions, respondents found the chatbot generally polite, adaptive, able to handle complex prompts and factually appropriate. Clarity, context awareness, personalisation and fallback strategy received moderate ratings. Relevance, speed of response, conversational flow, error handling and response diversity received lower or more variable ratings. In particular, lower means and higher standard deviations for some items such as Relevance Q1 and Up-to-date Response Q2 suggest that respondents did not have a consistency in rating these aspects.

Usability Results Based on SUS

Table 4 reports the usability scores that was independently analysed with the System Usability Scale (SUS). SUS measures a system-level ease of use and interaction experience rather than the quality of individual chatbot responses.

Table 4 Summary of System usability scale (SUS) scores

Statistic	Metric
Number of respondents	75
Minimum score	65.00
Maximum score	97.50
Mean score	80.97
Standard deviation (SD)	7.74

The SUS scores ranged from 65.00 to 97.50, with a mean score of 80.97 (SD = 7.74). This result indicates a generally favourable usability perception within this sample. The SUS result should be interpreted independently of the 1–4 rubric-based response-quality scores.

Item-Total Correlation Results for Rubric-Based Items

Table 5 presents the item-total correlation results for the 42 rubric-based items representing 14 response-quality criteria. With 75 respondents and a significance level of 0.05, the critical r-value was 0.227.

Table 5 Item-total correlation results for 42 rubric-based items

Criteria	Item-total correlation			Interpretation
	Question 1	Question 2	Question 3	
Accuracy	0.812	0.859	0.787	exceeded threshold
Relevance	0.690	0.676	0.765	exceeded threshold
Clarity	0.694	0.566	0.791	exceeded threshold
Up to Date	0.654	0.820	0.779	exceeded threshold
Speed of Response	0.811	0.769	0.822	exceeded threshold
Context Awareness	0.684	0.670	0.613	exceeded threshold
Conversational Flow	0.646	0.745	0.699	exceeded threshold
Error Handling	0.744	0.548	0.559	exceeded threshold
Fallback Strategy	0.735	0.712	0.745	exceeded threshold
Handling of Complex Queries	0.625	0.662	0.513	exceeded threshold
Personalisation	0.649	0.655	0.791	exceeded threshold
Politeness	0.630	0.751	0.716	exceeded threshold
Response Diversity	0.884	0.870	0.846	exceeded threshold
Adaptability	0.695	0.727	0.765	exceeded threshold

The results show that all item-total correlation values exceeded the critical r-value. Therefore, the 42 rubric-based items were retained for subsequent analysis. These findings provide preliminary item-level evidence. Nevertheless, should not be interpreted as confirmed construct validity since factor analysis was not conducted in this study.

Internal Consistency of Rubric-Based Items

Internal consistency analysis was conducted on all 75 responses with no cases excluded (see and should be interpreted with caution and not as evidence of a confirmed factor structure, given that the rubric covers a number of evaluation dimensions.

). Internal consistency was marginal for a preliminary multidimensional instrument was indicated with Cronbach's alpha of 0.699 for the 42 rubric-based items. In this preliminary evaluation research, values around 0.70 were found to be moderately acceptable and should be interpreted with caution and not as evidence of a confirmed factor structure, given that the rubric covers a number of evaluation dimensions.

Table 6 Case processing summary for internal consistency analysis

Case Processing Summary			
		N	%
Cases	Valid	75	100.0
	Excluded	0	0.0
	Total	75	100.0

Automatic Metrics Results for Content-Oriented Criteria

We conducted evaluation with reference-based automatic metrics on four content-oriented criteria: accuracy, relevance, clarity, and up-to-date response. These four criteria were selected because they can be evaluated by comparing the chatbot's responses and the reference answers. Table 7 shows the average of human scores vs. automatic metrics in terms of ROUGE-1, ROUGE-2, ROUGE-L, and SacreBLEU.

Table 7 Mean human scores and automatic reference-based metric scores for content-oriented criteria

Criteria	Human Score	ROUGE-1	ROUGE-2	ROUGE-L	SacreBLEU
Accuracy	2.700	0.234	0.135	0.176	0.089
Clarity	2.636	0.180	0.063	0.146	0.056
Relevance	2.500	0.115	0.029	0.105	0.013
Up-to-date response	2.545	0.268	0.118	0.229	0.048
Overall average	2.591	0.197	0.084	0.162	0.050

The results show that among the automatic metrics, ROUGE-L generally gave higher values than ROUGE-2 and SacreBLEU, which indicates a higher degrees of sequence-level overlap between the answers generated by the chatbot and the reference answers. However, the automatic metrics values were generally low to moderate, suggesting limited lexical overlap between the two answer sets. This finding fits the necessity of studying the alignment between human judgement and automatic metrics through correlation analysis.

Human–Metric Correlation Results

We also conducted Pearson and Spearman correlation analyses to examine the alignment between human rubric scores and automatic reference-based metrics. Pearson correlation was used to assess linear association and Spearman correlation was used because the human rubric scores were ordinal Likert scale. Correlations were calculated on two levels: overall across the four selected criteria (Table 8) and separately for each criterion (Table 9).

Table 8 Overall human–metric correlation results

Criteria	Metric	(N)	Pearson (r)	Pearson (p-value)	Spearman (rho)	Spearman (p-value)
4 criteria (Accuracy, (Relevance, Clarity, Up-to-date)	Rouge-1	44	0.3760	0.0118	0.3720	0.0129
	Rouge-2	44	0.3520	0.0190	0.3290	0.0291
	Rouge-L	44	0.3900	0.0090	0.3790	0.0112
	SacreBLEU	44	0.2580	0.0912	0.3450	0.0218

Table 8 indicates positive overall correlations between human rubric scores and reference-based automatic metrics. ROUGE-L showed the strongest overall correlation with human scores (Pearson r = 0.3900, p = 0.0090, Spearman rho = 0.3790, p = 0.0112), followed by ROUGE-1 and ROUGE-2. SacreBLEU showed a weaker relationship; its Pearson correlation was positive but not statistically significant, whereas its Spearman correlation was statistically significant. These results suggest that the human rubric scores are somewhat correlated with the automatic reference-based metrics especially those based on ROUGE.

Table 9 presents the criterion-level correlation for the four-selected criteria.

Table 9 Criterion-level human–metric correlation results

Criteria	Metric	(N)	Pearson (r)	Pearson (p-value)	Spearman (rho)	Spearman (p-value)
Accuracy	ROUGE-1	10	0.0950	0.7939	0.1060	0.7701
	ROUGE-2	10	0.3070	0.3879	0.0970	0.7893

Criteria	Metric	(N)	Pearson (r)	Pearson (p-value)	Spearman (rho)	Spearman (p-value)
	ROUGE-L	10	0.1220	0.7368	0.0250	0.9453
	SacreBLEU	10	0.1570	0.6641	0.3250	0.3593
Clarity	ROUGE -1	11	0.5630	0.0711	0.7170	**0.0130**
	ROUGE -2	11	0.5980	0.0521	0.7370	**0.0096**
	ROUGE -L	11	0.6150	**0.0439**	**0.7730**	**0.0053**
	SacreBLEU	11	0.3690	0.2635	0.4110	0.2093
Relevance	ROUGE -1	12	0.4300	0.1626	0.3270	0.2990
	ROUGE -2	12	-0.0240	0.9409	0.0330	0.9182
	ROUGE -L	12	0.3030	0.3385	0.3270	0.2990
	SacreBLEU	12	0.1170	0.7166	0.2300	0.4713
up-to-date response	ROUGE -1	11	0.7560	**0.0071**	0.7120	**0.0139**
	ROUGE -2	11	0.5270	0.0959	0.3560	0.2824
	ROUGE -L	11	0.6620	**0.0265**	0.6650	**0.0254**
	SacreBLEU	11	0.5300	0.0937	0.5340	0.0905

At the criterion level, the results show a strong alignment between human and metrics scores for both clarity and up-to-date criteria, especially for ROUGE-based metrics. ROUGE-L showed significant correlations for both Spearman and Pearson and ROUGE-1 and ROUGE-2 for Spearman. On the other hand, the accuracy and relevance results showed weaker and non-significant correlations. It should be noted that the findings here are exploratory, not definitive, because the criteria-level analyses are based on limited numbers of answer pairs 10 to 12 responses per criterion.

DISCUSSION

The fundamental value of the present study lies in showing that a structured human-scored rubric can be used to institutional educational chatbot responses and can be reinforced selectively by reference-based automatic metrics. The results indicate that the proposed evaluation approach can support a structured assessment of LLM-based educational chatbot responses. According to descriptive results, respondents generally viewed the chatbot favourably across a number of criteria, especially politeness, adaptability, dealing with complex questions, and accuracy. These results indicate that the chatbot was able to produce polite, adaptive and factually acceptable answers in many interaction scenarios. The SUS result also supports this positive perception, with an average score of 80.97, which means that respondents generally consider the chatbot system usable.

The results of instrument testing offer preliminary support for the quality of the rubric-based items. All 42 rubric-based items were above the critical item-total correlation value, indicating preliminary item validity. The Cronbach's alpha value of 0.699 also indicates marginally acceptable internal consistency and is close to the commonly used threshold of 0.70. However, this finding should be interpreted cautiously as the rubric addresses several evaluation dimensions. In comparison with the AIPI with several types of validation evidence such internal consistency, test–retest reliability, convergent validity, internal validity, and clinician inter-rater concordance, this rubric can be seen as a preliminary evidence rather than a validated standard equivalent to AIPI.

The correlation analysis also shows that human rubric scores have partial agreement with automatic reference-based metrics. For the overall level, ROUGE-L showing the strongest association with human scores. At the criterion level, there was stronger alignment for clarity and up-to-date response, suggesting that automatic metrics can provide useful complementary evidence for certain content-oriented criteria. This indicates that the human-scored rubric is not independent of automatic evaluation outcomes, but displays partial agreement with lexical reference-based metrics.

However, the weaker and non-significant correlations for accuracy and relevance show that automatic lexical metrics cannot fully capture all aspects of human judgement. Factual interpretation, semantic appropriateness, and contextual fit are not always reflected by word overlap with a reference answer. Moreover, interaction-related criteria like politeness, personalisation, response diversity, adaptability, error handling, and usability, are dependent on human perception and contextual understanding. Therefore, automatic metrics should be employed as complementary tools, not as replacements for rubric-based human evaluation.

The comparison with AIPI helps to clarify the manuscript contribution. AIPI stays psychometrically stronger as it presents wider validation evidence within a clinical case-management setting, including expert-led development and inter-rater concordance. In contrast, the present work provides a narrower but distinct contribution: larger interaction-oriented coverage for institutional educational chatbot responses, a separate usability layer through SUS, and exploratory human-metric alignment analysis. Thus, the proposed rubric should be viewed as context-specific and preliminary, and not as a psychometric equal replacement to AIPI.

In summary, the study indicates that the proposed rubric can be useful as an initial human-centred instrument for evaluating LLM-based educational chatbot responses. Human scoring remains necessary for interactional qualities such as politeness, error handling, personalisation, adaptability, and contextual appropriateness. The combination of rubric-based scoring, SUS-based usability evaluation and reference-based automatic metrics correlation gives a more comprehensive view of chatbot performance. However, the study is still limited by the single-institution setting, relatively small sample size, and lack of factor analysis and inter-rater reliability testing.

CONCLUSION

This study developed and preliminarily evaluated a human-scored analytic rubric for evaluating LLM-based educational chatbot responses to institutional information queries. The rubric organised 15 literature-informed evaluation criteria across 5 domains. 14 criteria were evaluated using 42 rubric-based items and usability was evaluated separately using the System Usability Scale (SUS). We evaluated 4 content-based metrics (accuracy, relevance, clarity, and up-to-date answer) using ROUGE-1, ROUGE-2, ROUGE-L, and SacreBLEU, and studied their correlation with human assessments using Pearson and Spearman correlation analysis.

The results indicate that the chatbot performed rather well in terms of politeness, sophisticated query handling, adaptability, correctness and clarity, but relatively poorly or inconsistently in terms of relevancy, up-to-date response, speed of response, conversational flow and response diversity. The 42 rubric-based items provided preliminary item-level evidence and marginal internal consistency. The results of human-metric correlation also suggested some degree of agreement between human scores and automatic reference-based metrics, in particular at the aggregate level. The results imply that automatic metrics is useful for content-oriented evaluation, and that human judgement remains important for interaction-oriented elements, such as usability, personalisation, politeness, flexibility and response diversity.

Limitations

This study has several limitations. The five-domain structures were constructed via literature-informed conceptual grouping and not empirically confirmed using EFA or CFA. The survey also included 75 student responses from one higher education institution, which limits generality to wider user groups and institutional contexts. Moreover, the evaluation was done in a controlled environment using selected institutional information queries. Finally, ROUGE and SacreBLEU were only applied on

selected content-oriented criteria, and may not fully cover the semantic accuracy, factual adequacy, or contextual relevance.

Future Study

Future research should use more diverse and larger samples of respondents to provide more robust statistical validation including expert-based content validity assessment, inter-rater reliability analysis, exploratory factor analysis, and confirmatory factor analysis. Future work could additionally include semantic-based evaluation metrics such as BERTScore, BLEURT, G-Eval or GPT-based evaluation such as Holistic Evaluation of Language Models (HELM) and test the rubric on multiple chatbot systems, domains, and real-world use cases.

Data Availability

As supplemental materials we supply the complete rubric scoring guide, the 42 rubric-based evaluation questions, the SUS questionnaire, the verified question-answer set, the reference answers used for automatic metric analysis, and the anonymised respondent.

REFERENCES

[1] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with Applications*, vol. 2, no. July, p. 100006, 2020, doi: 10.1016/j.mlwa.2020.100006.

[2] S. H. Tampubolon *et al.*, *ChatGPT di Era AI: Revolusi Komunikasi Berbasis Mesin*. Yayasan Kita Menulis, 2025.

[3] N. A. M. Mokmin and N. A. Ibrahim, "The evaluation of chatbot as a tool for health literacy education among undergraduate students," *Educ. Inf. Technol. (Dordr).*, vol. 26, no. 5, pp. 6033–6049, 2021, doi: 10.1007/s10639-021-10542-y.

[4] S. Gupta et al., "Comprehensive Framework for Evaluating Conversational AI Chatbots," *The Review of Contemporary Scientific and Academic Studies*, vol. 5, no. 2, 2025, doi: 10.55454/rcsas.5.02.2025.003.

[5] M. Hmoud, H. Swaity, E. Anjass, and E. M. Aguaded-Ramírez, "Rubric Development and Validation for Assessing Tasks' Solving via AI Chatbots," *Electronic Journal of e-Learning*, vol. 22, no. 6 Special Issue, pp. 1–17, 2024, doi: 10.34190/ejel.22.6.3292.

[6] R. A. Sekarwati, A. Sururi, R. Rakhmat, M. Arifin, and A. Wibowo, "Survei Metode Pengukuran Aplikasi Chatbot Berbasis Media Sosial," *Gema Teknologi*, vol. 21, no. 2, pp. 67–73, 2021.

[7] A. Parks, E. Travers, R. Perera-Delcourt, M. Major, M. Economides, and P. Mullan, "Is this chatbot safe and evidence-based? A call for the critical evaluation of generative AI mental health chatbots," *J. Particip. Med.*, vol. 17, p. e69534, May 2025, doi: 10.2196/69534.

[8] M. Walker, D. Litman, C. A. Kamm, and A. Abella, "PARADISE: A framework for evaluating spoken dialogue agents," in *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Jul. 1997, pp. 271–280. doi: 10.3115/976909.979652.

[9] G. Datta, N. Joshi, and K. Gupta, "Analysis of automatic evaluation metric on low-resourced language: BERTScore vs BLEU score," in *International Conference on Speech and Computer*, 2022, pp. 155–162.

[10] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: NLG evaluation using gpt-4 with better human alignment," in *Proceedings of the 2023 conference on empirical methods in natural language processing*, Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2511–2522. doi: 10.18653/v1/2023.emnlp-main.153.

[11] R. Bommasani, P. Liang, and T. Lee, "Holistic evaluation of language models," *Ann. N. Y. Acad. Sci.*, vol. 1525, no. 1, pp. 140–146, Aug. 2023, doi: https://doi.org/10.48550/arXiv.2211.09110.

[12] N. Cannavaro, "A chatbot application for academic services using the open-source Rasa platform with a two-stage fallback feature, [In Indonesia, "Aplikasi Chatbot untuk Layanan Akademik Menggunakan Platform RASA Open Source dengan Fitur Two Stage Fallback" ,] *Jurnal Ilmu Komputer dan Informatika*, vol. 3, no. 1, pp. 53–64, 2023, doi: 10.54082/jiki.73.

[13] T. Ait Baha, M. El Hajji, Y. Es-Saady, and H. Fadili, "The Power of Personalization: A Systematic Review of Personality-Adaptive Chatbots," *SN Comput. Sci.*, vol. 4, no. 5, 2023, doi: 10.1007/s42979-023-02092-6.

[14] M. E. S. Simaremare, C. Pardede, I. N. I. Tampubolon, D. A. Simangunsong, and P. E. Manurung, "The Penetration of Generative AI in Higher Education: A Survey," *2024 IEEE Integrated STEM Education Conference, ISEC 2024*, no. March, pp. 1–5, 2024, doi: 10.1109/ISEC61299.2024.10664825.
[15] J. R. Lechien, A. Maniaci, I. Gengler, S. Hans, C. M. Chiesa-Estomba, and L. A. Vaira, "Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the Artificial Intelligence Performance Instrument (AIPI)," *European Archives of Oto-Rhino-Laryngology*, vol. 281, no. 4, pp. 2063–2079, Sep. 2024, doi: 10.1007/s00405-023-08219-y.
[16] D. Guo *et al.*, "DeepSeek-Coder: when the large language model meets programming–the rise of code intelligence," *arXiv preprint arXiv:2401.14196*, 2024.
[17] S. Suwarno and C. Aeni, "Pentingnya Rubrik Penilaian Dalam Pengukuran Kejujuran Peserta Didik," *Edukasi: Jurnal Pendidikan*, vol. 19, no. 1, p. 161, 2021, doi: 10.31571/edukasi.v19i1.2364.
[18] H. N. Boone and D. A. Boone, "Analyzing Likert data," *J. Ext.*, vol. 50, no. 2, 2012, doi: 10.34068/joe.50.02.48.
[19] I. A. Kamal and A. B. Cahyono, "Pemanfaatan Chatbot Berbasis Dialogflow dan Google Sheet Api Untuk Penyimpanan Laporan Komplain Konsumen Toko Online," *Automata*, vol. 3, no. 2, pp. 1–5, 2022.
[20] Snigdha Tadanki and Sai Kiran Reddy Malikireddy, "Context-aware chatbots with data engineering for multi-turn conversations," *World Journal of Advanced Engineering Technology and Sciences*, vol. 4, no. 1, pp. 063–078, 2021, doi: 10.30574/wjaets.2021.4.1.0061.
[21] M. W. Shiferaw, T. Zheng, A. Winter, L. A. Mike, and L. N. Chan, "Assessing the accuracy and quality of artificial intelligence (AI) chatbot-generated responses in making patient-specific drug-therapy and healthcare-related decisions," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, 2024, doi: 10.1186/s12911-024-02824-5.
[22] N. M. Radziwill and M. C. Benton, "Evaluating Quality of Chatbots and Intelligent Conversational Agents," *CEUR Workshop Proc.*, vol. 1982, pp. 40–49, 2017, doi: 10.48550/arXiv.1704.04579.
[23] H. Liang and H. Li, "Towards Standard Criteria for human evaluation of Chatbots: A Survey," 2021. doi: 10.48550/arXiv.2105.11197.
[24] Y. Heryanto, F. Farahdinna, and S. Wijanarko, "Evaluasi Responsivitas dan Akurasi: Perbandingan Kinerja ChatGPT dan Google BARD dalam Menjawab Pertanyaan seputar Python," *Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK*, vol. 9, no. 1, pp. 248–256, 2024.
[25] A. R. Syah and A. Prihanto, "Auto Response Messages pada Telegram Bot untuk Pelayanan Sistem Informasi Praktek Industri dan Skripsi dengan Metode Webhook," *Journal of Informatics and Computer Science (JINACS)*, vol. 3, no. 04, pp. 547–556, 2022, doi: 10.26740/jinacs.v3n04.p547-556.
[26] R. Lumbantoruan, X. Zhou, Y. Ren, and L. Chen, "I-CARS: An Interactive Context-Aware Recommender System," in *IEEE International Conference on Data Mining (ICDM)*, IEEE, Jun. 2019, pp. 1240–1245.
[27] S. Raimer and M. Vanhauer, "I don't understand you – error-handling as a key aspect of conversational design for chatbots," *Human Interaction & Emerging Technologies (IHIET-AI 2022): Artificial Intelligence & Future Applications*, vol. 23, no. April 2022, 2022, doi: 10.54941/ahfe100878.
[28] S. Vemuru, E. John, and S. Rao, "Handling Complex Queries Using Query Trees," no. June, 2021, doi: 10.36227/techrxiv.14845212.v1.
[29] Z. Ma, Z. Dou, Y. Zhu, H. Zhong, and J. R. Wen, "One Chatbot per Person: Creating Personalized Chatbots based on Implicit User Profiles," *SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 555–564, 2021, doi: 10.1145/3404835.3462828.
[30] M. D. S. Monteiro, V. C. Pereira, and L. C. D. C. Salgado, "Investigating politeness strategies in chatbots through the lens of Conversation Analysis," in *Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems*, 2023, pp. 1–12.
[31] R. Ren, M. Zapata, J. W. Castro, O. Dieste, and S. T. Acuna, "Experimentation for Chatbot Usability Evaluation: A Secondary Study," *IEEE Access*, vol. 10, pp. 12430–12464, 2022, doi: 10.1109/ACCESS.2022.3145323.
[32] ISO, "Ergonomic requirements for office work with visual display terminals. Part 11: Guidance on usability," *ISO No 924111*, vol. 2008, no. February 9, p. 22, 1994.
[33] R. Sakaeda and D. Kawahara, "Generate, Evaluate, and Select: A Dialogue System with a Response Evaluator for Diversity-Aware Response Generation," *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Student Research Workshop*, pp. 76–82, 2022, doi: 10.18653/v1/2022.naacl-srw.10.
[34] W. A. Pinheiro, *At the intersections: Queer high school students' experiences with the teaching of mathematics for social justice*. Indiana University, 2022.

[35] S. Zitzmann and G. A. Orona, “Why we might still be concerned about low Cronbach’s alphas in domain-specific knowledge tests,” *Educ. Psychol. Rev.*, vol. 37, no. 2, p. 37, 2025.