I.J. Modern Education and Computer Science, **2026, 2, 82-99**
Published Online on April 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2026.02.05

Quality Evaluation of Indonesian Student-generated User Stories: Insight from Human and ChatGPT Evaluation

Muhammad Ihsan Zul*
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Malaysia
Department of Information Technology, Politeknik Caltex Riau, Indonesia
Email: ihsan@pcr.ac.id
ORCID ID: https://orcid.org/0000-0001-7668-2675
*Corresponding Author

Suhaila Mohd. Yasin
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Malaysia
Email: ysuhaila@uthm.edu.my
ORCID ID: https://orcid.org/0000-0002-5543-8259

Ivan Chatisa
Department of Information Technology, Politeknik Caltex Riau, Pekanbaru, Indonesia
Email: ivan@pcr.ac.id
ORCID ID: https://orcid.org/0009-0000-6846-2883

Fikri Muhaffizh Imani
Department of Information Technology, Politeknik Caltex Riau, Pekanbaru, Indonesia
Email: fikri@pcr.ac.id
ORCID ID: https://orcid.org/0009-0001-6390-6331

Siti Syahidatul Helma
Department of Information Technology, Politeknik Caltex Riau, Pekanbaru, Indonesia
Email: helma@pcr.ac.id
ORCID iD: https://orcid.org/0000-0002-6331-7123

Dadang Syarif Sihabudin Sahid
Department of Information Technology, Politeknik Caltex Riau, Pekanbaru, Indonesia
Email: dadang@pcr.ac.id
ORCID ID: https://orcid.org/0000-0002-5398-8548

Received: 21 July, 2025; Revised: 18 November, 2025; Accepted: 27 January, 2026; Published: 08 April, 2026

Abstract: User stories are essential in agile software development for capturing software requirements, yet concerns over their quality persist globally. While prior studies have evaluated user story quality using practitioners and artificial intelligence, they primarily focus on general settings. This study addresses a gap by evaluating the quality of student-generated user stories in an educational context, specifically in Indonesia. The objective of this study is to compare evaluations by human evaluators and ChatGPT using the Quality User Story (QUS) Framework and evaluate the quality of the student-generated user story compared to the global studies. A total of 951 user stories from 103 student software projects were analyzed. Evaluations were conducted by three human evaluators and ChatGPT (GPT-4o). Percentage Agreement and Cohen's Kappa measured inter-rater agreement, while the McNemar Test assessed statistical significance, and effect sizes were examined using Cohen's g. Results show generally high agreement between human and ChatGPT evaluations, but lower consistency in several criteria, such as Conceptually Sound, Independent, and Unambiguous. Only four of the thirteen criteria—Conflict-Free, Unique, Well-Formed, and Atomic—showed no significant differences. Most criteria showed small to medium effect sizes, whereas Complete exhibited a large practical difference. Common quality issues among students included Uniform, Independent, and Complete (set criteria), Atomic, Conceptually Sound, and Unambiguous (individual criteria), with overlap observed in global studies. This study shows

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

that ChatGPT can support user story evaluation in educational settings when guided by clear rubrics and validated by humans. It also offers practical insights for educators by identifying criteria that require stronger emphasis in teaching, particularly in software engineering education in Indonesia.

Index Terms: User Story Quality, Agile Software Development, ChatGPT Evaluation, Quality User Story Framework, Software Requirement

1. Introduction

User Stories are central to software development projects that adopt Agile Software Development (ASD) methodologies, especially during the requirement elicitation phase. Serving as a communication tool between developers and stakeholders, a user story helps articulate functional requirements. However, several studies have pointed out ongoing quality issues in user stories, including ambiguity, vagueness, incompleteness, inconsistency, and duplication [1,2,3,4]. These issues have often been identified through literature reviews and the analysis of user stories collected from open-access industrial or community-driven projects.

In line with these practices, user stories have also become prevalent in higher education in Indonesia, where students employ them to document software requirements in academic projects. This is evident from the widespread use of user stories in student-authored publications, theses, and final project reports related to software development. Consequently, user stories have become an integral part of how students articulate functional requirements. Despite this trend, systematic research on the quality of user stories produced by students, especially those written in Indonesian, remains limited. This research gap represents a critical area requiring further investigation. To address this gap, this study investigates the quality of student-generated user stories that are publicly disseminated through various academic platforms, such as Google Scholar and open-access digital libraries. The evaluation adopts the Quality User Story (QUS) Framework [5], [6], one of the most comprehensive frameworks currently available for assessing user story quality. While earlier frameworks such as INVEST [7,8] gained popularity due to their simplicity, QUS offers a more nuanced evaluation through its combination of set and individual-based criteria. Moreover, some studies even use Artificial Intelligence (AI) to evaluate the quality of user stories, making the use of AI an opportunity for further research in this field.

Recent literature shows a growing interest in the use of AI technologies, such as Natural Language Processing (NLP), machine learning, and large language models (LLMs), to assess user story quality [9,10]. Lucassen et al. [6] introduced AQUSA to support QUS-based evaluations, while more recent work, such as Ronanki et al. [11], has explored the use of Generative AI. Although the use of NLP and Machine Learning to evaluate the quality of user stories has been widely developed, research on the use of generative AI is still very limited. Moreover, the opportunities for using generative AI in an open academic context are vast, as an alternative to the limited access to experts or industry practitioners, especially in Indonesia. Therefore, this issue is considered the second gap for this research. This study extends the latter line of inquiry by employing ChatGPT to evaluate user stories based on QUS criteria. In addition to the AI component, the evaluations in this study are also conducted by human experts, including academic researchers and industry practitioners—an approach consistent with prior research involving expert or practitioner judgment [12,13].

According to the gaps and the prior studies, the objectives of this study are twofold: (1) to compare the evaluation results of ChatGPT and human evaluators using the QUS Framework on student-generated user stories from Indonesia; and (2) to investigate the quality of the student-generated user story and the comparison findings from global studies. The study contributes to the academic discourse by offering an alternative approach that leverages Generative AI to support formal framework-based evaluations of user stories. Moreover, it provides empirical insights into the current state of student-generated user stories in Indonesia. It offers the opportunity for enhancing user story learning in software engineering education in Indonesia.

This paper is organized as follows. Section 2 presents the materials and methods used and applied in this study. It introduces the theoretical background on user stories and their quality evaluation, followed by a detailed explanation of the research methodology. Section 3 reports the study's results, including statistical tests addressing the first research objective and quality analysis related to the second objective. Section 4 discusses the research findings, including comparisons with previous studies and limitations. Finally, Section 5 concludes the paper by summarizing the main contributions and implications of the study.

2. Materials and Methods

2.1. User Story

A user story is a concise, informal statement that describes a user's need or expectation from a software system [14]. Widely adopted in Agile practices, particularly during the requirement elicitation phase, user stories provide a simple yet structured approach to capturing functional requirements [15,16]. The most commonly used format for

writing user stories is the Connextra template, which was introduced by [14]. This template frames requirements through three core dimensions—WHO, WHAT, and WHY—each representing the user role, the intended activity, and the underlying goal or value, respectively [2,17,18,19]. The standard formulation of a user story is structured as:

"As a [role], I want to [action] so that [value/purpose]."

This format ensures clarity by aligning system functionality with user objectives and business value [20]. Further elaborations of the user story model emphasize critical elements such as user roles, specific tasks, and intended outcomes (both hard and soft goals), thereby helping ensure that user stories are both meaningful and aligned with the overall objectives of the system under development [17,18,19,21]. These conceptual models improve the clarity and consistency of user stories and support traceability between user expectations and software features.

2.2. User Story Quality Evaluation Framework

Researchers have developed various frameworks to evaluate the quality of user stories in agile software development. Among the most well-known are the INVEST framework [22,23], the Quality User Story (QUS) framework [3], and the Model for Determining User Story Quality (MDUSQ) [24]. In addition, several studies have adopted general software engineering standards such as ISO/IEC 25010 [25] and IEEE 830 [26] as benchmarks for user story evaluation. Although these frameworks differ in scope and focus, they all aim to evaluate the user story content, such as quality, completeness, clarity, and relevance.

Among them, the QUS framework stands out for several reasons. Firstly, it is the most widely adopted for automation using artificial intelligence techniques, including both Machine Learning and NLP [9,27,28], as well as Generative AI approaches employing large language models [11,29]. Secondly, QUS offers a more comprehensive evaluation mechanism than frameworks such as INVEST, as it categorizes quality assessment into individual and set criteria. Furthermore, QUS integrates a linguistic perspective to assess the structure and coherence of user stories. The complete set of QUS criteria is summarized in Table 1.

Table 1. QUS Framework Criteria [6]

Criteria	Description	Individual/Set
Syntactic		
Well-formed	A user story includes at least a role and a means	Individual
Atomic	A user story expresses a requirement for exactly one feature	Individual
Minimal	A user story contains nothing more than role, means, and ends	Individual
Semantic		
Conceptually sound	The means expresses a feature and the ends expresses a rationale	Individual
Problem-oriented	A user story only specifies the problem, not the solution to it	Individual
Unambiguous	A user story avoids terms or abstractions that lead to multiple interpretations	Individual
Conflict-free	A user story should not be inconsistent with any other user story	Set
Pragmatic		
Full sentence	A user story is a well-formed full sentence	Individual
Estimatable	A story does not denote a coarse-grained requirement that is difficult to plan and prioritize	Individual
Unique	Every user story is unique, duplicates are avoided	Set
Uniform	All user stories in a specification employ the same template	Set
Independent	The user story is self-contained and has no inherent dependencies on other stories	Set
Complete	Implementing a set of user stories creates a feature-complete application, no steps are missing	Set

The QUS framework comprises thirteen criteria, five of which are classified as set criteria. These criteria evaluate the quality of user stories at the project level, meaning that the evaluation must consider the entire collection of user stories within a given project. In contrast, the remaining criteria are categorized as individual and can be evaluated independently, without referencing other user stories in the same project. This distinction enables evaluators to capture the individual-level quality for each user story and the set-level coherence and consistency across the complete set of user stories as requirements.

2.3. Research Design

To achieve the objectives of this study, the research follows a structured multi-phase approach. The first phase involves the construction of a dataset of user stories generated by Indonesian students. This phase consists of three key steps: (1) user story acquisition, which focuses on collecting user story from various student-generated documents; (2) user story extraction, where the collected user stories are converted from documents into a structured table format; and (3) user story preprocessing, which ensures that all user stories are standardized and consolidated into a single dataset file for further analysis. The second phase comprises the user story quality evaluation, which is conducted using two parallel approaches: evaluation by ChatGPT and evaluation by a human evaluator, including practitioners and academia. This is followed by the third phase, which applies statistical analysis techniques to measure inter-rater reliability by evaluating the level of agreement and the statistical significance between the evaluations produced by ChatGPT and

those conducted by human evaluators. This phase provides findings that address the first research objective. Finally, the fourth phase evaluates the overall quality of student-generated user stories in Indonesia by analyzing the weaknesses identified in both ChatGPT and human evaluations. It compares it to the global studies issue. This phase addresses the second research objective. An overview of the entire research design is illustrated in Fig. 1.

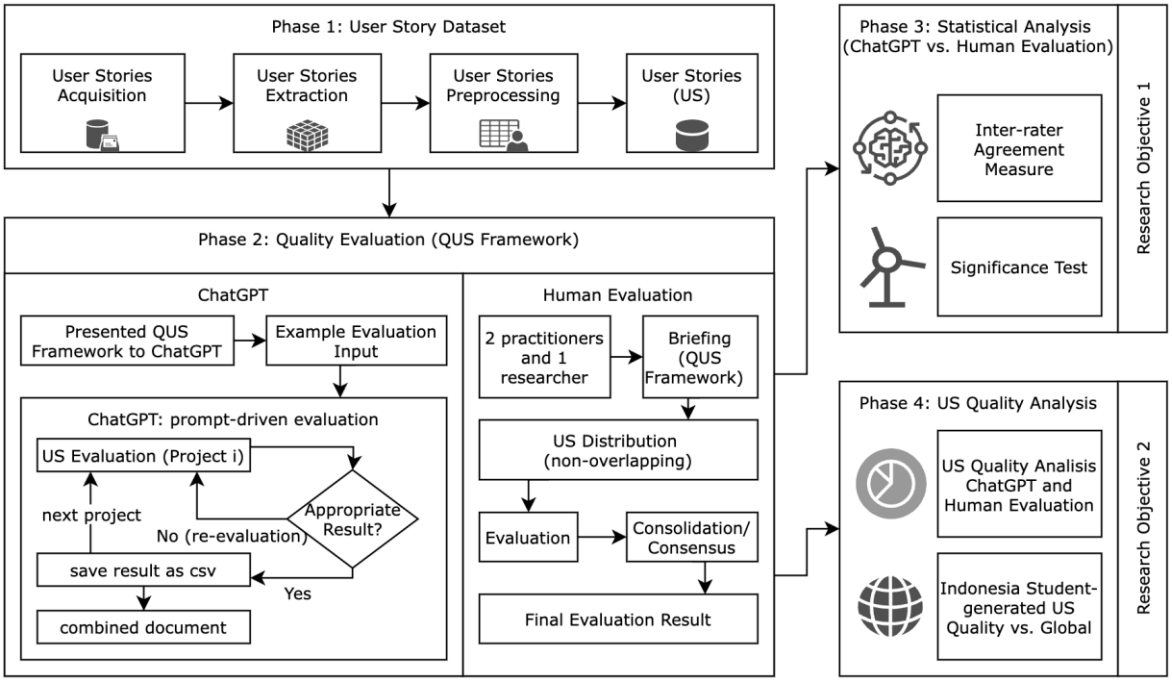


Fig. 1. Research Design.

A detailed description of each phase of this study is presented in the following subsections to provide a comprehensive understanding of the overall research workflow.

2.4. User Stories Acquisition

The user story acquisition phase begins by collecting digital documents related to software development projects authored by Indonesian students. These documents include research journal articles, undergraduate theses, final project reports, and internship reports. The study emphasizes the use of publicly available Indonesian-language documents to ensure transparency and reproducibility of the dataset. Unpublished student work was excluded due to access and consent constraints. Accordingly, the findings are not intended to be generalized to unpublished student artifacts or to user stories written in other languages, but should be interpreted within the context of Indonesian-language user story documentation in academic software development projects. This limitation is acknowledged in the Discussion section.

Furthermore, given the study's focus, only documents written in the Indonesian language are selected. The document search targets publications from 2017 to 2025 using Google Scholar and search engines. Keywords used include “user story” AND “*skripsi*”, “user story” AND “*sistem informasi*”, and (“user story” OR “user stories”) AND “*skripsi*”. This search yields 105 documents, each explicitly containing user stories developed within ASD-based software development methodologies.

Subsequently, in the extraction phase, all identified documents undergo a manual process to compile user stories into a structured dataset. The extraction process involves identifying sections that explicitly present user stories in plain text and tabular format. To maintain data consistency, the extracted user stories are checked against the original documents to confirm both their source and the accuracy of the extracted content. Each extracted user story is organized into a spreadsheet with columns including: ID number, user story, project title, publication year, higher education name, province, and document URL. Geographic information, such as university name and province, is included to capture the demographic distribution of the dataset. The spreadsheet is shared among all researchers, allowing cross-checking and verification of the extracted entries. As a result, 979 user stories are extracted from the 105 collected documents.

The final sub-phase, preprocessing, ensures that all user story data conforms to a standardized structure aligned with the dataset's column headings. After removing 28 duplicated entries, the final dataset comprises 951 unique user stories derived from 103 student projects across 73 universities located in 24 provinces throughout Indonesia, covering major islands including Java, Sumatra, Kalimantan, Sulawesi, Bali, and Nusa Tenggara. The preprocessing stage also involves verifying provincial information by cross-referencing the originating higher education institution using search engines. This result shows that the source of user stories comes from different areas in Indonesia. Additionally, two evaluation plan spreadsheet tabs are prepared—one for the set QUS criteria and another for the individual criteria. While the number of user stories in Indonesia could be vast, this study limits its scope to publicly available and

explicitly published online sources. When this phase is concluded, the study moves to the second phase, namely the quality evaluation procedure.

2.5. Quality Evaluation Procedure

This study's user story quality evaluation employs two distinct approaches: ChatGPT and the involvement of practitioners as human evaluators. Specifically, ChatGPT, powered by the GPT-4o model (ChatGPT Plus), is used to evaluate the user story based on predefined criteria. In parallel, three practitioners are engaged as human evaluators. Two of them are experienced software engineers with approximately six years of industry experience who currently serve as lecturers, while the third is a researcher specializing in software engineering with four years of experience in the research area. The involvement of practitioners in user story evaluation is a common practice in prior research, ensuring that evaluations are grounded in real-world expertise. ChatGPT and the practitioners conduct the evaluation using the same QUS Framework, providing a consistent basis for comparing results across ChatGPT and human evaluations.

ChatGPT Evaluation Procedure

The ChatGPT-based evaluation in this study follows a prompt-driven evaluation approach proposed by [30], in which the QUS framework is explicitly provided to guide the model's assessment, rather than involving any model training or fine-tuning. The evaluation process begins by introducing the QUS criteria and evaluation mechanisms to ChatGPT through structured instructions. The model's understanding of the framework is then verified by querying its interpretation of the meaning and structure of the QUS criteria. A set of sample user stories is then evaluated, with the results presented in a per-criterion table and rationale-based explanations for each result provided by the model. The whole dataset evaluation process is initiated once the model consistently demonstrates an accurate understanding and application of the QUS Framework.

The dataset evaluation is carried out separately for each project, adhering to the QUS Framework's two distinct groups: individual and set criteria. The process begins with the first project, for which ChatGPT evaluates each user story. The output is then screened to ensure three things: (1) the evaluated user stories match those specified in the prompt, (2) the evaluation includes both individual and set criteria, and (3) the results are ordered according to the sequence of the criteria to facilitate data integration. If any of these structural conditions are not met, the user story evaluation for that project is repeated. It is important to note that the screening process does not examine the content of the evaluation results themselves; it only ensures the correctness of the structure and formatting of the output. This action was made to preserve ChatGPT's original semantic interpretation of the QUS criteria without post-hoc human correction. It is enabling a direct head-to-head comparison with expert evaluations. Although potential semantic misinterpretations by ChatGPT are not eliminated at this stage, they are retained and later reflected in the comparative analysis. This methodology ensures consistent alignment with the framework and minimizes potential false statements, misinterpretations, or hallucinations [31] by ChatGPT in terms of QUS criteria.

Practitioners Evaluation Procedure

The evaluation of user story quality has been the focus of numerous studies, with the involvement of experts, practitioners, and academia as a common practice. For instance, Brockenbrough and Salinas [13] engaged academic evaluators to assess the quality of user stories generated by ChatGPT. Similarly, collaborative evaluations involving software development practitioners and academic experts have been employed in other studies as well [32,33]. In other cases, researchers have relied solely on practitioners to conduct the evaluation, as seen in the work of Kuhail and Lauesen [26] and do Nascimento et al. [8]. Building on these approaches, the present study adopts a hybrid strategy by involving two experienced software practitioners alongside academic professionals who teach and research in the field of software development.

All evaluators attend an in-person briefing session before the evaluation process to ensure a shared understanding of the QUS Framework and the evaluation procedure. This step is essential to maintain validity and prevent discrepancies from differing framework interpretations [34]. Next, user stories are then distributed to evaluators via an online platform using a non-overlapping assignment strategy, where each evaluator assesses a unique subset of user stories. Given the large dataset (951 user stories), this method is appropriate and is supported by literature emphasizing its practicality and content validity when standardized criteria are in place [35,36,37]. The evaluators are given six weeks to complete their tasks and can consult the researcher for clarification. Once all evaluations are submitted, a consolidation meeting is held to review and discuss ambiguous cases. During this session, evaluators may revise or reaffirm their evaluations based on group input, including feedback from the researcher. The consolidation meeting is limited to resolving ambiguous or borderline cases and does not involve re-evaluating the entire dataset. Final decisions are guided strictly by the QUS criteria rather than majority voting. Through this process, a final consensus is reached, forming the basis for the official evaluation results by practitioners and academia [38]. Completing this phase will progress the study to the third, statistical analysis phase.

2.6. Statistical Analysis

Inter-rater Agreement Measure

To evaluate the consistency between ChatGPT and human evaluators, the results from both methods are compared using inter-rater reliability techniques, commonly employed in previous studies, such as Ronanki et al. [11]. Statistically, this approach is known as an inter-rater agreement measure [39], and it is suitable for measuring the extent to which two raters agree when evaluating the same items. Given that the comparison in this study involves only two evaluation methods—ChatGPT and human evaluation—two key statistical measures are employed: Percentage Agreement and Cohen's Kappa. Percentage Agreement calculates how often both evaluators produce the same score, offering a straightforward indication of alignment. In contrast, Cohen's Kappa accounts for the agreement that could occur by chance, providing a more robust measure of inter-rater consistency. Together, these metrics offer complementary insights into the reliability of the evaluation results across ChatGPT and human evaluator perspectives. The percentage agreement formula is:

$$percentage\ agreement\ = \left(\frac{number\ of\ agreement}{total\ user\ stories}\right) \times 100\% \tag{1}$$

The inter-rater reliability assessment proceeds with Cohen's Kappa after the initial analysis using the Percentage Agreement. McHugh [39] noted that Cohen's Kappa provides a more accurate estimate of inter-rater reliability than percentage agreement, as it adjusts for the likelihood of agreement occurring by chance. Cohen's Kappa is necessary because percentage agreement can appear deceptively high, especially when evaluators tend to give a dominant response (e.g., consistently choosing "Yes"). Kappa corrects this by measuring the extent of agreement that cannot be explained by chance alone. The Kappa coefficient ranges from -1 to 1, where higher values indicate stronger agreement. According to the widely cited interpretation by Landis and Koch [40], Kappa values are categorized as follows: 0.01–0.20: slight agreement, 0.21–0.40: fair agreement, 0.41–0.60: moderate agreement, 0.61–0.80: substantial agreement, and 0.81–1.00: almost perfect agreement. This classification is used to interpret the level of agreement between ChatGPT and human evaluators based on their evaluations. The Cohen's Kappa formula is:

$$P_o = \frac{A+D}{N} \tag{2}$$

$$P_e = \left(\frac{(A+B)(A+C)}{N^2}\right) + \left(\frac{(C+D)(B+D)}{N^2}\right) \tag{3}$$

$$k = \frac{P_o - P_e}{1 - P_e} \tag{4}$$

Where P_o: agreement between the evaluators and P_e: expected agreement by chance, A: number of user stories rated Yes by both human and ChatGPT, B: number of user stories rated Yes by human and No by ChatGPT, C: number of user stories rated No by human and Yes by ChatGPT, D: Number of user stories rated No by both human and ChatGPT, $N = A+B+C+D$: total number of user stories evaluated and κ : Kappa coefficient.

Percentage Agreement and Cohen's Kappa are applied to the thirteen criteria defined in the QUS Framework, enabling a detailed and criterion-level comparison of evaluation results. This granular approach ensures that the analysis captures not only overall agreement but variations in consistency across set and individual quality criteria.

Significance Test

Then, a significance test is performed to determine if there is a statistically significant difference between the evaluations produced by ChatGPT and those provided by human evaluators. Since the evaluation results are recorded in binary format (i.e., Yes or No), the appropriate statistical method is McNemar's Test. This test is designed to assess paired nominal data and is commonly used to evaluate the relationship between two related categorical variables [41]. In this context, it is used to determine if the differences observed between the two evaluation methods are significant. The formula for McNemar's Test is:

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{5}$$

Where χ^2 = Chi-squared for McNemar's test, b = number of user stories rated Yes by humans and No by ChatGPT, and then c = number of user stories rated No by humans and Yes by ChatGPT. The critical value for McNemar's Test is determined using the Chi-squared distribution table with 1 degree of freedom (df)**.** In this study, a significance level (α) of 0.05 is adopted, resulting in a reference chi-square (χ^2) value of 3.841, as commonly cited in statistical literature [42]. According to this threshold, if the calculated χ^2 value exceeds 3.841, the difference between human and ChatGPT evaluations is considered statistically significant, and then insignificant if the χ^2 value is less than 3.841 [43]. This

analysis formally tests whether the two evaluation approaches—automated and human—produce meaningfully different outcomes when applied to the same user stories.

The significance test is complemented by an assessment of the magnitude of the differences between human and ChatGPT evaluations using Cohen's g. This allows the practical impact of statistically significant findings to be examined in terms of effect size. Cohen's g is used as an effect size for the McNemar test because both tests focus on paired cases in which the two evaluations disagree [44]. While the McNemar test determines whether the difference between these non-matching pairs is statistically significant, Cohen's g quantifies the magnitude of that difference, thereby providing an interpretable measure of practical significance [45].

The statistical analysis phase is considered complete once both the inter-rater agreement and the statistical significance tests have been conducted and the results have been obtained and interpreted. Following the completion of this phase, the study proceeds to its next stage: the analysis of user stories' quality, which builds upon the insights derived from the preceding statistical evaluations.

2.7. Student-Generated User Story Quality Analysis

The quality analysis is performed by examining the evaluation results generated by both ChatGPT and human evaluators. A descriptive analysis categorizes the findings based on the set and individual criteria defined in the QUS Framework. Each criterion is then explored in detail to compare and contrast the evaluations from both approaches. This analysis aims to provide a comprehensive overview of the quality of user stories produced by Indonesian students, as reflected in the constructed dataset. In addition to evaluating the quality of each criterion at both the individual and set levels, this study also evaluates the overall quality by identifying whether all criteria, both set and individual, are marked as "Yes" by both human evaluators and ChatGPT. The results of this analysis are presented as the percentage of projects that utilize user stories meeting all set criteria, as well as the percentage of user stories that fulfill all individual criteria. Furthermore, to complete the second research objective, the findings are compared with prior global studies on user story quality, facilitating the identification of similarities or differences between user stories generated by Indonesian students and those developed in global studies.

3. Result

3.1. Inter-rater Agreement

The inter-rater agreement is calculated using Percentage Agreement and Cohen's Kappa to measure the consistency between human and ChatGPT evaluations. The calculations are organized into two groups: the first focuses on the set criteria, while the second addresses the individual criteria defined in the QUS Framework. The results of the percentage agreement calculations for both criterion groups are visualized in Fig. 2, providing a comparative overview of agreement levels across set and individual criteria.

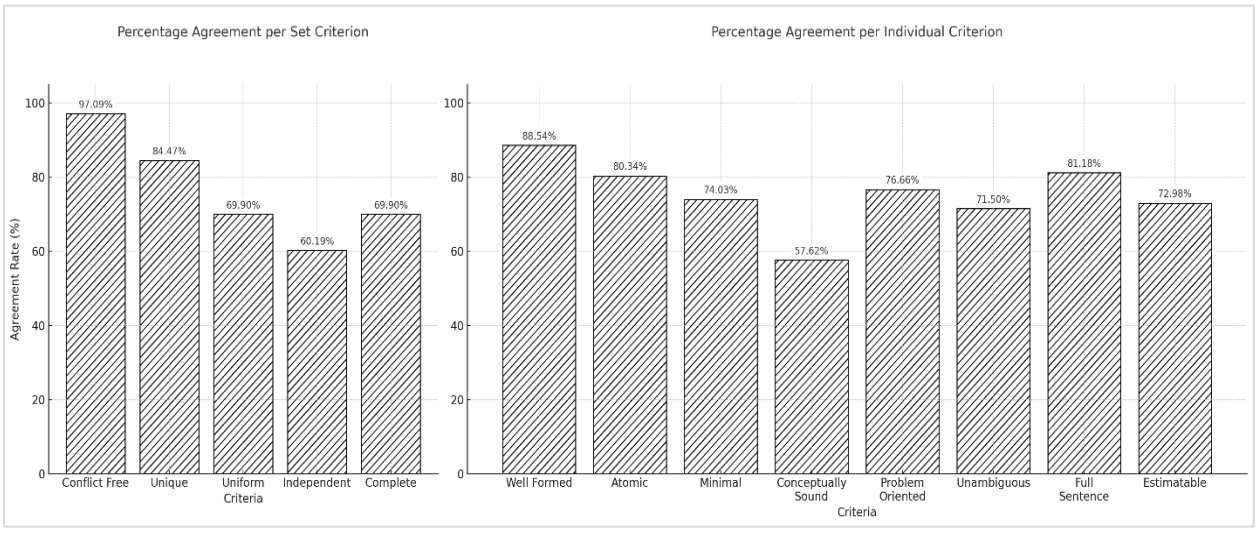


Fig. 2. Percentage Agreement, set criteria (left) and individual criteria (right).

It can be observed in Fig. 2 that the lowest agreement is found in the Independent criterion (60.19%), whereas Uniform and Complete share the same agreement rate of 69.9%. This indicates a notable discrepancy in evaluation approaches between human evaluators and ChatGPT, particularly for these three criteria. For Independent and Complete, evaluators must conduct a deeper analysis as these criteria require a nuanced understanding of the context and objectives of each project. In contrast, the Uniform criterion should theoretically be easier to identify based on the

structural format consistency of the user stories' text. However, the results suggest a misalignment in how humans and ChatGPT interpret what constitutes a "Uniform" user story.

A closer look reveals that human evaluators marked Yes for Uniform only 55.3% of the time, while ChatGPT did so at a significantly higher rate of 75.7%. This suggests that humans do not rely solely on consistency across user stories within a project, as ChatGPT tends to, but evaluate whether each user story adheres to the typical user story structure. On the other hand, two different criteria show high agreement levels above 84%, Unique (84.4%) and Conflict-free (97.09%), indicating that in these criteria, both human and ChatGPT evaluations are largely aligned.

On the right side of Fig. 2, nearly all agreement rates for individual criteria exceed 70%, except for Conceptually Sound, which records a notably lower agreement of 57.62%. This suggests a divergence in interpretation between ChatGPT and human evaluators. According to Lucassen et al. [6,46], a conceptually sound user story must emphasize both the means (a clear and actionable feature) and the end (a rationale introduced with "so that"). In practice, many user stories either lack a clear means or omit the end entirely. ChatGPT tends to classify such cases as No, resulting in a Yes rate of only 58%, whereas human evaluators are more permissive, marking Yes 72% of the time. For the remaining individual criteria—such as Well-formed, Atomic, and Full Sentence—the evaluation results from ChatGPT closely align with those of human evaluators, likely because these criteria rely on objective textual structures. Overall, while agreement between human and ChatGPT evaluations is generally consistent across both set and individual criteria, closer attention is needed for specific criteria where interpretation plays a larger role, such as Uniform, Independent, Complete, and Conceptually Sound.

Cohen's Kappa is calculated as a complementary measure to deepen the agreement analysis between human evaluators and ChatGPT. The resulting Kappa values for both groups are presented in Fig. 3. Based on Fig. 3, the Complete criterion yields the highest Cohen's Kappa value for the set criteria, falling within the moderate agreement category. This indicates that human evaluators and ChatGPT demonstrate a moderate level of agreement when evaluating this particular criterion. Interestingly, this result contrasts with the findings from the percentage agreement analysis, which previously showed that Conflict-free and Unique had the highest agreement rates. This discrepancy arises because Cohen's Kappa accounts for agreement that may occur by chance, offering a more conservative estimate. According to the Kappa results, Conflict-free and Uniform fall within the fair agreement range, suggesting a partially aligned evaluation approach between humans and ChatGPT for certain user stories. Meanwhile, Unique and Independent yield only slight agreement, implying a frequent divergence in how human evaluators interpret and apply these criteria compared to ChatGPT.

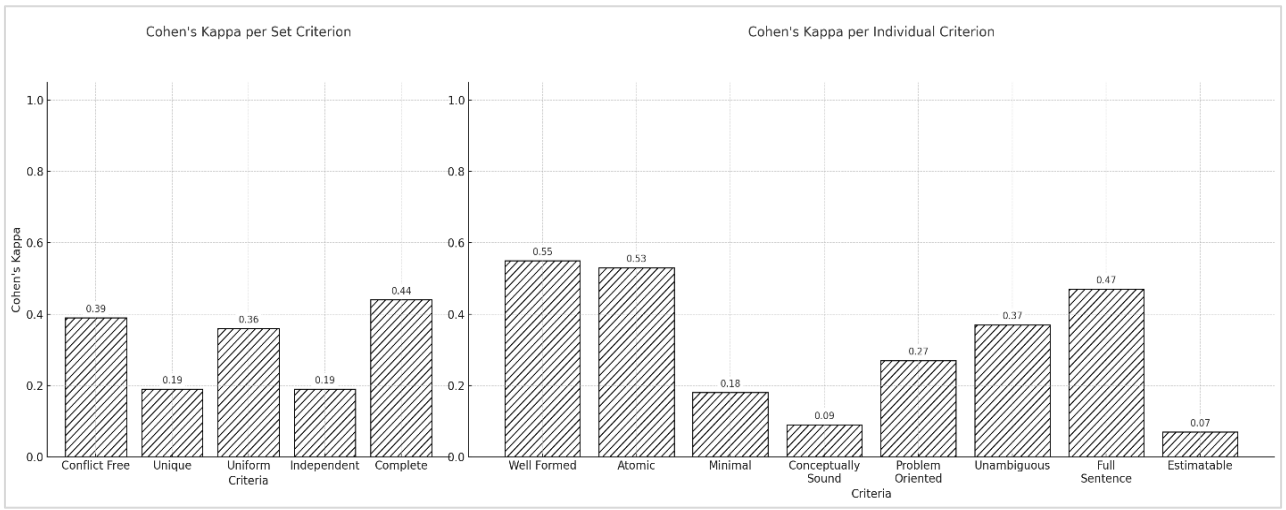


Fig. 3. Cohen's Kappa Agreement Result, set criteria (left), and individual criteria (right).

Turning to the individual criteria, Well-formed, Atomic, and Full Sentences consistently show the highest Kappa values, all categorized as moderate agreement. These results are also consistent with the percentage agreement analysis, reinforcing that these three criteria are relatively straightforward to evaluate. Since they rely primarily on syntactic structure, both human evaluators and ChatGPT appear to apply similar standards in evaluating them. In contrast, Problem-Oriented and Unambiguous fall within the fair agreement level, suggesting that human and ChatGPT evaluators occasionally align on certain items. This level of agreement is understandable given that these criteria often require evaluators to understand the underlying context of the project, rather than relying solely on the sentence structure of the user story.

Lastly, the criteria Minimal, Conceptually Sound, and Estimable fall into the slight agreement category. While Conceptually Sound and Estimable share contextual complexity with Problem-Oriented and Unambiguous, Minimal stands out because it is expected to rely mainly on structural features—namely, the presence of a role, means, and end. However, the analysis reveals a notable disagreement between evaluators. Human evaluators marked Yes for Minimal 76% of the time, whereas ChatGPT assigned Yes in 86% of cases. This indicates that ChatGPT may apply a more

permissive interpretation of what constitutes a "Minimal" user story, potentially overlooking contextual nuances that human evaluators consider during the evaluation.

While the inter-rater agreement results provide insights into the agreement between human and ChatGPT evaluations, it is also important to evaluate the proximity of ChatGPT's evaluation outcomes to those of human evaluators. In this context, human evaluation is treated as the reference benchmark, given that the evaluators involved in the study are domain experts. To measure how closely ChatGPT approximates human evaluation, the F1 Score is employed. This metric is used to assess how closely ChatGPT approximates human evaluation decisions and serves as a meaningful yet additional metric rather than a standalone measure of educational quality. The comparison results for the set criteria based on the F1 Score are presented in Fig. 4.

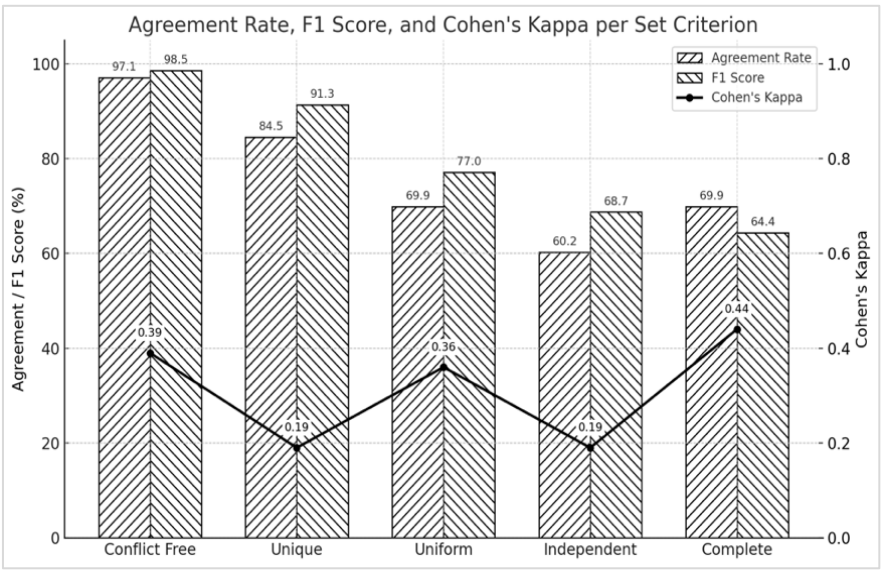


Fig. 4. Percentage Agreement, Cohen's Kappa, and F1 Score Comparison Result for Set Criteria.

Based on Fig. 4, it can be observed that ChatGPT demonstrates a generally aligned capability with human evaluators. This is supported by the F1 Score values ranging from 64% to 98.5%. The highest scores, exceeding 90%, are found in the Conflict-Free and Unique criteria, while Uniform scores 77%, and both Independent and Complete fall between 60% and 70%. These values are consistent with the percentage agreement results. However, the Cohen's Kappa values do not always follow the same pattern, as some criteria, such as Unique and Independent, are categorized under slight agreement. This suggests that some observed agreement may be due to chance, highlighting the importance of Cohen's Kappa as a complementary metric for objectively evaluating agreement. Therefore, for the set criteria Unique, Uniform, and Independent, there remains a challenge in defining more precise and shared evaluation references to ensure better alignment between human interpretations and ChatGPT, which would in turn help improve the corresponding Cohen's Kappa values.

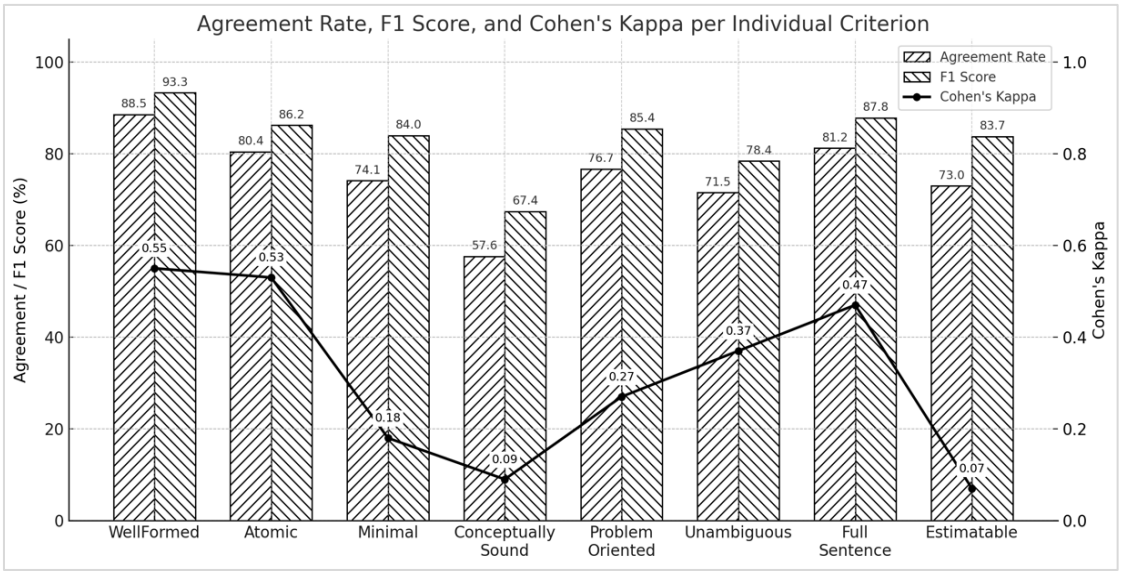


Fig. 5. Percentage Agreement, Cohen's Kappa, and F1 Score Comparison Result for Individual Criteria.

Similarly, the individual criteria in Fig. 5 show that ChatGPT performs reasonably well in replicating human evaluations, with F1 Scores ranging from 67.4% to 93.3%**.** This indicates that, for most criteria, ChatGPT is able to follow human evaluation patterns effectively. However, a persistent challenge remains with the *Conceptually Sound* criterion, which consistently exhibits lower scores across all evaluation metrics. An anomaly is observed in the *Estimable* criterion, where both percentage agreement and F1 Score exceed 70%, yet the Cohen's Kappa value approaches zero. This suggests that, despite slight agreement, the alignment between ChatGPT and human evaluators may be occurring by chance rather than based on a shared understanding of the evaluation framework. Therefore, based on Cohen's Kappa, several individual criteria—*including* Minimal**,** Conceptually Sound**,** Problem-Oriented**,** and Estimable**—**still pose challenges**.** To improve alignment, the evaluation mechanisms for these criteria need to be defined in greater detail and clarity, thereby promoting more consistent interpretations between human evaluators and ChatGPT.

Taken together, differences in inter-rater agreement between percentage agreement and Cohen's Kappa for both set and individual criteria arise from the skewed distribution of the evaluation outcomes across most criteria. Under such conditions, Cohen's Kappa is highly sensitive to the dominance of a single response category, whereas percentage agreement is largely unaffected by this prevalence effect [47,48]. Consequently, for certain criteria, Cohen's Kappa may differ substantially from percentage agreement and F1 Score, even when the observed level of agreement is relatively high.

3.2. Significance Test

The significance test using McNemar's Test is applied to the set criteria, and the results are presented in Table 2.

Table 2. McNemar Test for Set Criteria Result.

Criteria	b (Yes-ChatGPT, No-Human)	c (No-ChatGPT, Yes-Human)	McNemar Chi-square (χ^2)	Significant (χ^2>3.81)
Conflict Free	3	0	1.333	No
Unique	7	9	0.062	No
Uniform	26	5	12.903	Yes
Independent	8	33	14.049	Yes
Complete	31	0	29.032	Yes

Based on Table 2, the evaluations provided by human evaluators and ChatGPT for the Conflict-Free and Unique criteria show no statistically significant differences, indicating alignment between the two evaluators. However, significant differences are found for the rest criteria, suggesting that humans and ChatGPT interpret these criteria differently. A closer examination reveals that conflicting evaluations occur in 31 projects for Uniform, 41 for Independent, and 31 for Complete—figures that remain below half of the total 103 projects included in the study. This implies that, despite statistical significance, a substantial level of agreement still exists between human and ChatGPT evaluations for these criteria.

Table 3. McNemar Test for Individual Criteria Result.

Criterion	b (Yes-ChatGPT, No-Human)	c (No-ChatGPT, Yes-Human)	McNemar Chi-square (χ^2)	Significant (χ^2>3.81)
Well Formed	46	63	2.349	No
Atomic	92	94	0.005	No
Minimal	169	77	33.663	Yes
Conceptually Sound	139	264	38.154	Yes
Problem Oriented	78	144	19.032	Yes
Unambiguous	103	168	15.114	Yes
Full Sentence	137	42	49.363	Yes
Estimable	193	64	63.751	Yes

The calculation results for the individual criteria are presented in Table 3**.** Based on Table 3, the evaluation results for the Well-Formed and Atomic criteria show no significant differences between human evaluators and ChatGPT. This finding is consistent with the high percentage agreement and the Cohen's Kappa values, which fall within the moderate agreement category. For both criteria, it appears that humans and ChatGPT share a similar understanding of the evaluation guidelines. However, the results differ slightly for the Full Sentence criterion. Although the structural aspects of this criterion are easily observable in the sentence form, the significance test reveals a notable difference in evaluations: ChatGPT tends to be more permissive, often rating user stories as Full Sentence even when human evaluators do not. This discrepancy may stem from the stricter interpretation applied by humans. Other criteria also showed statistically significant differences, with a substantial gap observed in the Conceptually Sound criterion. This is aligned with both previous percentage agreement and Cohen's Kappa calculations, where Conceptually Sound records the lowest values and is classified as slight agreement in Cohen's Kappa.

Based on the results of the significance test, it is evident that most QUS criteria show statistically significant differences between the evaluations conducted by humans and ChatGPT. Only a few criteria—namely, Conflict-Free

and Unique from the set criteria, and Well-Formed and Atomic from the individual criteria—exhibit no significant differences. This indicates a higher level of alignment between the two evaluation approaches. In contrast, the remaining criteria fall into the significant category. This highlights notable discrepancies in judgment. These findings suggest that humans and ChatGPT often interpret many QUS criteria differently when evaluating user stories. This contributes to the observed significance levels in the test results.

The subsequent statistical analysis focuses on examining the effect sizes associated with the significance test results. As discussed in Section 2.6, the effect size calculations using Cohen's g are presented in Table 4 for the set criteria and in Table 5 for the individual criteria.

Table 4. Set Criteria Effect Size

Criteria	No-ChatGPT, Yes-Human	Yes-ChatGPT, No-Human	n	Cohen's g	Effect Size Interpretation
Conflict-Free	3	0	103	0.029	Small
Unique	7	9	103	0.019	Small
Uniform	26	5	103	0.204	Medium
Independent	8	33	103	0.243	Medium
Complete	31	0	103	0.301	Large

Table 5. Individual Criteria Effect Size

Criteria	No-ChatGPT, Yes-Human	Yes-ChatGPT, No-Human	n	Cohen's g	Effect Size Interpretation
Well-Formed	63	46	951	0.018	Small
Atomic	94	92	951	0.002	Small
Full Sentence	77	170	951	0.098	Small
Problem-Oriented	264	139	951	0.131	Medium
Minimal	144	78	951	0.069	Small
Conceptually Sound	168	103	951	0.068	Small
Estimable	42	137	951	0.100	Medium
Unambiguous	64	193	951	0.136	Medium

From a practical significance perspective, the effect size analysis provides additional insight into the practical meaning of these differences. As shown in Table 4, most set criteria demonstrate small to medium effect sizes. The Complete criterion shows a large effect size. This indicates a substantial practical difference and suggests that human involvement remains necessary to ensure the completeness of user stories at the project level. In contrast, Table 5 shows that most individual criteria are associated with small effect sizes. This suggests that the differences between human and ChatGPT evaluations have a limited practical impact. ChatGPT can therefore be considered a supportive tool for preliminary user story quality assessment when complemented by clear evaluation guidelines and human oversight.

3.3. Indonesian Student-generated User Story Quality

Human evaluators and ChatGPT use different methods to evaluate user stories, but both provide useful information. Their results can help analyze the overall quality of user stories generated by students in Indonesia. This evaluation can benchmark Indonesian student stories with global research, highlighting strengths, weaknesses, and areas for improvement in local education. The following section presents the user story quality evaluation results from the perspectives of both set and individual criteria. The quality of the user story sets is evaluated by analyzing the percentage of Yes and No responses for each criterion, as determined by human evaluators and ChatGPT.

Set Criteria

According to Fig. 6, it is evident that the criteria Uniform, Independent, and Complete remain areas requiring improvement in Indonesian student-generated user stories. Although there are notable differences between human and ChatGPT evaluations for these three criteria, their overall percentages are still relatively high compared to Conflict-Free and Unique. Therefore, for set criteria, the key quality issues identified in Indonesian student-generated user stories include: Uniform, referring to the consistency of user story formatting; Independent, referring to the degree to which a user story stands alone without depending on others; and Complete, referring to the extent to which a set of user stories comprehensively covers all features needed to deliver the project.

Individual Criteria

Regarding Fig. 7, the quality of user stories for individual criteria shows relatively similar evaluation results between human evaluators and ChatGPT, indicating that the gap is not as pronounced as in the set criteria. The most critical quality issues are observed in the Atomic, Conceptually Sound, and Unambiguous criteria, where the proportion of Yes responses remains below 75% for both human and ChatGPT evaluations. This suggests that many student-generated user stories in Indonesia still lack atomicity, as multiple means are often included within a single story. For Conceptually Sound, which refers to the clarity of the feature described in a user story, students often fail to specify the exact feature being addressed. Additionally, even when a valid feature is present, the "end" part of the story is often unrelated or more appropriate as a "means", indicating a misunderstanding of the user story structure. Furthermore, the Unambiguous criterion, which emphasizes interpretive clarity, remains challenging. These three criteria—Atomic,

Conceptually Sound, and Unambiguous—stand out as critical areas for improvement in Indonesian student-generated user stories for individual criteria.

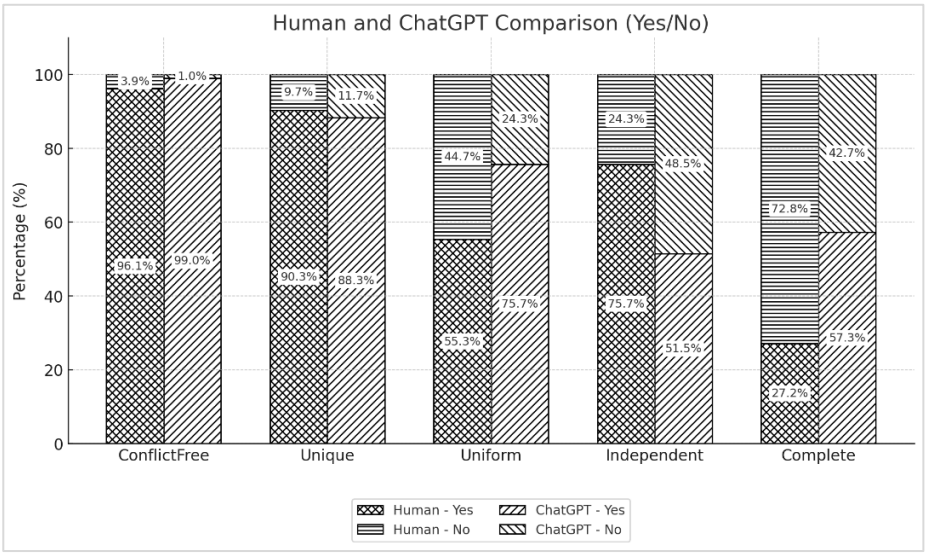


Fig. 6. Set Criteria Quality Evaluation Result.

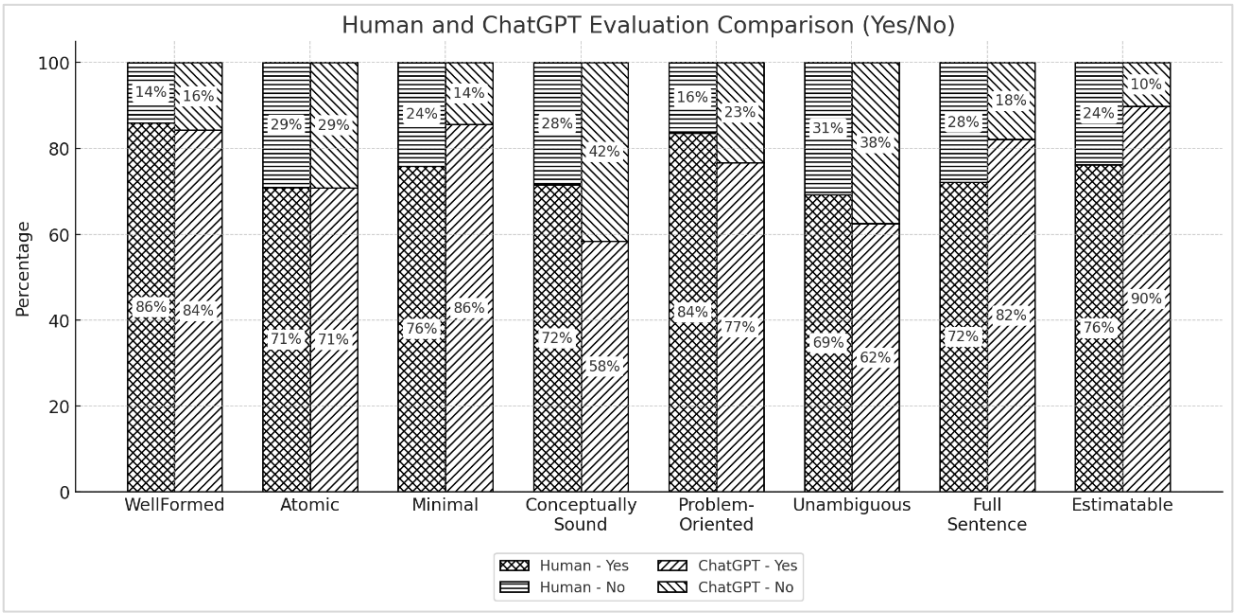


Fig. 7. Individual Criteria Quality Evaluation Result.

Overall Set and Individual Criteria

Further analysis of the overall quality based on the set-level criteria in the QUS Framework reveals that only 14.56% (15 projects) fulfilled all set criteria according to human evaluators. This result differs notably from the evaluation by ChatGPT, which indicated that 22.33% (23 projects) met all set criteria, suggesting that ChatGPT tends to assign higher scores than human evaluators. Regarding the individual-level criteria, 24.29% (231 user stories) were evaluated by human assessors as meeting all individual criteria. In comparison, 28.39% (270 user stories) received full compliance ratings from ChatGPT, with a relatively small difference of 39 user stories. Based on these findings, out of the 103 projects examined in this study, fewer than 23 projects fully met the quality requirements at the set level, and only a small proportion of the total 270 user stories fulfilled all individual criteria. These results highlight the need for educators to emphasize improving the quality of user stories generated by students, encouraging alignment with the standards outlined in the QUS Framework.

Indonesian Student-Generated User Story Quality vs. Global Studies

Based on global studies conducted to date, user story quality has been a frequently discussed issue in the literature. Table 4 presents a list of relevant studies that specifically address user story quality, highlighting the aspects they examined and discussed within their respective research contexts.

Table 6. User Story Quality Issue in Global Studies.

No	Criteria	Group	Study
1	Conflict-Free	Set	[1,2,4]
2	Unique (Duplication/Redundant)	Set	[1-3],33]
3	Complete (Insufficiency)	Set	[1,2,4]
4	Independent	Set	[49]
5	Well-Formed	Individual	[33,49]
6	Atomic (Granularity)	Individual	[4,49]
7	Ambiguity (Vagueness)	Individual	[1-4,33,50]

Based on the studies summarized in Table 4, seven recurring user story quality criteria have been identified in prior research. Although most of these studies do not explicitly specify which user story quality framework they employed, the issues they address are generally similar to the criteria outlined in the QUS Framework. Regarding the set criteria, Conflict-Free, Unique, Independent, and Complete frequently surface in global discussions. However, within the context of this study, only Independent and Complete are highlighted as challenges in user stories generated by Indonesian students. In contrast, Conflict-Free and Unique receive very high ratings, with affirmative responses exceeding 88% from human evaluators and ChatGPT. Notably, while a significant concern in the Indonesian context, the Uniform criterion remains absent from global discourse. This disparity may be attributed to variations in the frameworks used or a general underestimation of formatting consistency as a critical issue.

Global studies frequently discuss the individual criteria of Well-Formed, Atomic, and Ambiguous. In this study, atomic and unambiguous criteria also emerge as significant challenges for Indonesian students when developing clear and concise user stories. While Well-Formed is a key focus in global studies, it receives less emphasis in this study. Nevertheless, it remains important for students to consider it in order to meet accepted structural user story standards. Conversely, Conceptually Sound, similar to Uniform, is seldom addressed in previous studies, possibly due to the use of different quality frameworks or its relatively minor role in global studies.

Finally, the comparison shows that several quality issues discussed globally, such as Atomic, Ambiguity, Independent, and Complete, are also found in Indonesian student user stories. Meanwhile, some criteria like Conflict-Free and Unique are generally well met. At the same time, the appearance of Uniform and Conceptually Sound as local issues highlights the need to consider contextual factors when evaluating user story quality in educational settings.

4. Discussion

This study evaluated 951 user stories from 103 academic projects using the QUS framework and multiple evaluation metrics. Overall, agreement between human evaluators and ChatGPT ranged from moderate to high, indicating that ChatGPT can approximate human judgments for several criteria. However, consistent differences remain, particularly for criteria that require deeper contextual interpretation, whereas structurally defined criteria show stronger alignment. These findings suggest that ChatGPT can support user story quality evaluation at scale when guided by clear evaluation guidance, but its role should remain complementary for semantically criteria.

Taken together, the observed disagreement patterns, effect sizes, and skewed distributions suggest that not all QUS criteria operate well under binary scoring. While structurally grounded criteria such as Well-Formed and Atomic demonstrate stable agreement and small effect sizes, semantically and context-dependent criteria—including Conceptually Sound, Estimable, Independent, and Complete—exhibit greater disagreement and larger practical differences. These findings indicate that quality along these criteria is often expressed in degrees rather than as a strict Yes or No. The observed disagreement patterns indicate that some aspects of user story quality are not easily captured by binary judgments, particularly those that depend on semantic and contextual interpretation. The results do not challenge QUS, but suggest that selected criteria may be better supported by rubric-based evaluation guidance that allows partial fulfillment to be considered.

Compared to Ronanki et al. [11], who applied a Best-of-Three (BO3) prompting strategy on 11 user stories using 7 QUS criteria, this study provides greater empirical depth and methodological rigor by evaluating 951 user stories across all 13 QUS criteria and employing multiple evaluation metrics, including percentage agreement, Cohen's Kappa, McNemar's Test, and effect size analysis. Unlike Ronanki's heuristic prompting approach, this study adopted a rubric-based training procedure [30], in which ChatGPT was explicitly introduced to the QUS framework and validated prior to full dataset evaluation. When comparing percentage agreement for overlapping QUS criteria, both studies show higher agreement for structurally oriented criteria such as Well-Formed and Atomic (above 80%), and lower agreement for context-dependent criteria such as Unambiguous (around 55–65% in Ronanki et al. and approximately 60% in this study). While Ronanki et al. reported the lowest agreement for Unambiguous and Full Sentence, this study identifies

Conceptually Sound as the most challenging criterion. Despite these differences, both studies consistently conclude that ChatGPT can support human evaluation, while emphasizing the continued need for human validation in criteria requiring deeper contextual interpretation.

While several global studies report recurring user story quality issues, this study discusses these findings as qualitative context rather than quantitatively aligned comparisons, given differences in evaluation frameworks and metrics. These studies commonly highlight issues such as ambiguity, incompleteness, and independence, which conceptually align with QUS criteria, including Unambiguous, Complete, and Independent [1,3,4,33,49,50]. In addition, Atomic and Conceptually Sound are also frequently reported as challenging criteria. In contrast, Uniform—a key issue identified in this study, is rarely discussed in global studies, possibly due to differences in the quality frameworks employed. This result confirms the general applicability of QUS, while indicating that certain quality criteria may require attention depending on local educational and development contexts.

Although this study addresses the research objectives, several limitations remain. The dataset is limited to Indonesian-language user stories, and the evaluation relies on the QUS framework with a small number of human evaluators, which may affect generalizability and interpretative diversity. Future research should involve more diverse evaluator groups, multilingual or cross-cultural datasets, and comparative analyses across different AI models. In addition, the findings suggest opportunities to refine QUS by introducing clearer or more graded evaluation guidance for semantically complex criteria. From an educational perspective, further work may also explore instructional approaches that better support students in developing high-quality user stories aligned with QUS standards..

5. Conclusion

This study examined the quality of 951 student-written user stories from 103 software projects in Indonesia using the QUS framework and compared evaluations conducted by human experts with those conducted by ChatGPT. The findings indicate that ChatGPT can approximate human judgments for several structurally defined criteria, while notable discrepancies remain for semantically and context-dependent criteria such as Conceptually Sound, Independent, and Estimable. The analysis also reveals persistent weaknesses in student-generated user stories, particularly in the criteria of uniformity, independence, completeness, and conceptual clarity. These issues suggest the need for a more targeted teaching approach in software engineering education, especially for the criteria that require deeper contextual understanding.

Overall, this study demonstrates the potential of ChatGPT as a complementary tool for user story quality evaluation when supported by clear rubrics and human oversight. At the same time, the results highlight opportunities to refine the operationalization of the QUS framework and to develop instructional strategies that better support students in producing high-quality user stories. Future research should extend this work through more diverse datasets, evaluator groups, and comparative analyses across different AI models and evaluation frameworks.

Author Contributions Statement

Muhammad Ihsan Zul – Conceptualization, Literature Survey, Constructed the overall framework, Methodology, Proposed research ideas, User Story Data Acquisition, Data Extraction, Practitioners Coordination, Data Analysis, Statistical Analysis, Final Document Writing.

Suhaila Mohd. Yasin – Conceptualization, Methodology, Implementation Supervision, Reviewed and Edited the Manuscript, Final Manuscript Review.

Ivan Chatisa – User Story Data Acquisition, Data Extraction Screening, Practitioners Coordination, Data Analysis, Initial Writing, Manuscript Draft Review.

Fikri Muhaffizh Imani – User Story Data Acquisition, Practitioners Coordination, Data Extraction Validation, Data Analysis, Initial Writing, Manuscript Draft Review.

Siti Syahidatul Helma – User Story Data Acquisition, Practitioners Coordination, Data Extraction Validation, Data Analysis, Initial Writing, Manuscript Draft Review.

Dadang Syarif Sihabudin Sahid – Methodology, Draft Writing, Implementation Supervision, Editing, Final Manuscript Review.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

This research was funded by the Politeknik Chevron Riau Foundation as part of its Staff PhD Program.

Data Availability Statement

None

Ethical Declarations

The research did not involve human subjects or animal experiments. As such, ethical approval and consent to participate were not applicable.

Acknowledgments

We sincerely thank the practitioner who evaluated the user stories and contributed to their quality evaluation, as these user stories were produced by an Indonesian student.

Declaration of Generative AI in Scholarly Writing

The authors used generative AI tools solely to assist with language editing and improving the clarity and readability of the manuscript. The use of AI was limited to linguistic support, and all scientific content, analysis, interpretations, and conclusions were developed and verified by the authors. The authors carefully reviewed and revised the manuscript to ensure its accuracy and integrity.

Abbreviations

The following abbreviations are used in this manuscript:

QUS – Quality User Story
ASD – Agile Software Development
AI – Artificial Intelligence
NLP – Natural Language Processing
LLMs – Large Language Models
MDUSQ – Model for Determining User Story Quality
INVEST – Independent, Negotiable, Valuable, Estimable, Small, Testable
BO3 – Best-of-Three

References

[1] A. R. Amna and G. Poels, “Ambiguity in user stories: A systematic literature review,” May 01, 2022, *Elsevier B.V.* doi: 10.1016/j.infsof.2022.106824.

[2] A. R. Amna and G. Poels, “Systematic Literature Mapping of User Story Research,” 2022, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2022.3173745.

[3] F. Dalpiaz, I. van der Schalk, S. Brinkkemper, F. B. Aydemir, and G. Lucassen, “Detecting terminological ambiguity in user stories: Tool and experimentation,” *Inf. Softw. Technol.*, vol. 110, pp. 3–16, Jun. 2019, doi: 10.1016/j.infsof.2018.12.007.

[4] Y. A. Kustiawan and T. Y. Lim, “User Stories in Requirements Elicitation: A Systematic Literature Review,” in *2023 IEEE 8th International Conference On Software Engineering and Computer Systems (ICSECS)*, IEEE, Aug. 2023, pp. 211–216. doi: 10.1109/ICSECS58457.2023.10256364.

[5] G. Lucassen, F. Dalpiaz, J. M. E. M. van der Werf, and S. Brinkkemper, “The Use and Effectiveness of User Stories in Practice,” in *REFSQ 2016: Proceedings of the 22nd International Working Conference on Requirements Engineering: Foundation for Software*, 2016, pp. 205–222. doi: 10.1007/978-3-319-30282-9_14.

[6] G. Lucassen, F. Dalpiaz, J. M. E. M. van der Werf, and S. Brinkkemper, “Improving agile requirements: the Quality User Story framework and tool,” *Requir. Eng.*, vol. 21, no. 3, pp. 383–403, Sep. 2016, doi: 10.1007/s00766-016-0250-x.

[7] B. Kumar, U. Tiwari, and D. C. Dobhal, “User Story Splitting in Agile Software Development using Machine Learning Approach,” in *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, IEEE, Nov. 2022, pp. 167–171. doi: 10.1109/PDGC56933.2022.10053226.

[8] S. S. do Nascimento, J. M. Abe, L. R. Forçan, C. C. de Oliveira, K. Nakamatsu, and A. Ari, “Improving the Process of Evaluating User Stories Using the Paraconsistent Annotated Evidential Logic Eτ,” in *New Approaches for Multidimensional Signal Processing. NAMSP 2022. Smart Innovation, Systems and Technologies*, vol. 332, Springer, Singapore, 2023, pp. 133–142. doi: 10.1007/978-981-19-7842-5_12.

[9] S. Jiménez, A. Alanis, C. Beltrán, R. Juárez-Ramírez, A. Ramírez-Noriega, and C. Tona, “USQA: A User Story Quality Analyzer prototype for supporting software engineering students,” *Computer Applications in Engineering Education*, vol. 31, no. 4, pp. 1014–1024, Jul. 2023, doi: 10.1002/cae.22620.

[10] Z. Xuan, T. Wang, C. Wang, and T. Li, "A Tool for Automatically Identifying Semantic Conflicts in User Stories by Combining NLP and BERT Model," in *2024 IEEE 32nd International Requirements Engineering Conference (RE)*, IEEE, Jun. 2024, pp. 484–487. doi: 10.1109/RE59067.2024.00057.
[11] K. Ronanki, B. Cabrero-Daniel, and C. Berger, "ChatGPT as a Tool for User Story Quality Evaluation: Trustworthy Out of the Box?," in *Agile Processes in Software Engineering and Extreme Programming - Workshops*, P. Kruchten and P. Gregory, Eds., Amsterdam: Springer, Cham, 2024, ch. AI-assisted Agile, pp. 173–181. doi: 10.1007/978-3-031-48550-3_17.
[12] C. Ankora and A. D., "Integrating User Stories in the Design of Augmented Reality Application," *International Journal of Information Technologies and Systems Approach*, vol. 15, no. 1, pp. 1–19, Jul. 2022, doi: 10.4018/IJITSA.304809.
[13] A. Brockenbrough and D. Salinas, "Using Generative AI to Create User Stories in the Software Engineering Classroom," in *2024 36th International Conference on Software Engineering Education and Training (CSEE&T)*, IEEE, Jul. 2024, pp. 1–5. doi: 10.1109/CSEET62301.2024.10662994.
[14] M. Cohn, *User Stories for Agile Software Development*. Addison-Wesley, 2004.
[15] F. Dwitama and A. Rusli, "User stories collection via interactive chatbot to support requirements gathering," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 2, pp. 890–898, Apr. 2020, doi: 10.12928/TELKOMNIKA.V18I2.14866.
[16] C. O'hEocha and K. Conboy, "The Role of the User Story Agile Practice in Innovation," in *International Conference on Lean Enterprise Software and Systems*, Helsinki: Springer, 2010, pp. 20–30. doi: 10.1007/978-3-642-16416-3_3.
[17] S. Nasiri, Y. Rhazali, M. Lahmer, and N. Chenfour, "Towards a Generation of Class Diagram from User Stories in Agile Methods," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 831–837. doi: 10.1016/j.procs.2020.03.148.
[18] Y. Wautelet, S. Heng, M. Kolp, and I. Mirbel, "Unifying and Extending User Story Models," in *CAiSE 2014: Advanced Information Systems Engineering pp 211–225*, Thessaloniki: Springer, Champ, Jun. 2014, pp. 211–225. doi: 10.1007/978-3-319-07881-6_15.
[19] B. Yang, H. Guo, and H. Liu, "Evaluation and assessment of machine learning based user story grouping: A framework and empirical studies," *Sci. Comput. Program.*, vol. 227, Apr. 2023, doi: 10.1016/j.scico.2023.102943.
[20] X. Xu, Y. Dou, L. Qian, Z. Zhang, Y. Ma, and Y. Tan, "A Requirement Quality Assessment Method Based on User Stories," *Electronics (Basel).*, vol. 12, no. 10, p. 2155, May 2023, doi: 10.3390/electronics12102155.
[21] K. Tsilionis, J. Maene, S. Heng, Y. Wautelet, and S. Poelmans, "Conceptual Modeling Versus User Story Mapping: Which is the Best Approach to Agile Requirements Engineering?," 2021, pp. 356–373. doi: 10.1007/978-3-030-75018-3_24.
[22] B. Wake, "INVEST in Good Stories, and SMART Tasks," XP123: Exploring Extreme Programming.
[23] L. Buglione and A. Abran, "Improving the user story Agile technique using the INVEST criteria," in *Proceedings - Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement, IWSM-MENSURA 2013*, IEEE Computer Society, 2013, pp. 49–53. doi: 10.1109/IWSM-Mensura.2013.18.
[24] E. Jharko, "Some Issues in Using the Model of Determining the User Stories Quality to Reduce Software Development Risks," in *2024 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, IEEE, May 2024, pp. 658–663. doi: 10.1109/ICIEAM60818.2024.10553873.
[25] K. A. Alam, H. Asif, I. Inayat, and S.-U.-R. Khan, "Automated Quality Concerns Extraction from User Stories and Acceptance Criteria for Early Architectural Decisions," in *Software Architecture. ECSA 2024. Lecture Notes in Computer Science*, vol. 14889, Springer, Cham, 2024, pp. 359–367. doi: 10.1007/978-3-031-70797-1_24.
[26] M. A. Kuhail and S. Lauesen, "User Story Quality in Practice: A Case Study," *Software*, vol. 1, no. 3, pp. 223–243, Jun. 2022, doi: 10.3390/software1030010.
[27] S. N. F. N. B. Mustaffa, J. Bin Sallim, and R. B. Mohamed, "Enhancing High-Quality User Stories with AQUSA: An Overview Study of Data Cleaning Process," in *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, IEEE, Aug. 2021, pp. 295–300. doi: 10.1109/ICSECS52883.2021.00060.
[28] T. Wang, C. Wang, T. Li, Z. Liu, and Y. Zhai, "User Story Quality Assessment Based on Multi-dimensional Perspective: A Preliminary Framework," in *CEUR Workshop Proceedings: 15th International iStar Workshop*, A. Mate, T. Li, and E. Goncalves, Eds., Hyderabad, India: ceur-ws.org, Oct. 2022, pp. 7–13.
[29] Z. Zhang, M. Rayhan, T. Herda, M. Goisauf, and P. Abrahamsson, "LLM-Based Agents for Automating the Enhancement of User Story Quality: An Early Report," in *25th International Conference on Agile Software Development XP 2024*, W. van der Aalst, S. Ram, M. Rosemann, C. Szyperski, and G. Guizzardi, Eds., Bozen Bolzano: Springer, Jun. 2024, pp. 117–126. doi: 10.1007/978-3-031-61154-4_8.
[30] A. Atasoy and S. M. N. Arani, "ChatGPT: A reliable assistant for the evaluation of students' written texts?," *Educ. Inf. Technol. (Dordr).*, Apr. 2025, doi: 10.1007/s10639-025-13553-1.
[31] M. T. Hicks, J. Humphries, and J. Slater, "ChatGPT is bullshit," *Ethics Inf. Technol.*, vol. 26, no. 2, p. 38, Jun. 2024, doi: 10.1007/s10676-024-09775-5.
[32] O. Abed, K. Nebe, and A. B. Abdellatif, "AI-Generated User Stories Supporting Human-Centred Development: An Investigation on Quality," in *HCI International 2024 Posters. HCII 2024. Communications in Computer and Information Science*, vol. 2120, Springer, Cham, 2024, pp. 3–13. doi: 10.1007/978-3-031-62110-9_1.
[33] X. Xu, Y. Dou, L. Qian, J. Jiang, K. Yang, and Y. Tan, "Quality improvement method for high-end equipment's functional requirements based on user stories," *Advanced Engineering Informatics*, vol. 56, Apr. 2023, doi: 10.1016/j.aei.2023.102017.
[34] D.-M. Yoo and J. J. Han, "Inter-rater reliability and content validity of the measurement tool for portfolio assessments used in the Introduction to Clinical Medicine course at Ewha Womans University College of Medicine: a methodological study," *J. Educ. Eval. Health Prof.*, vol. 21, p. 39, Dec. 2024, doi: 10.3352/jeehp.2024.21.39.
[35] M. R. Lynn, "Determination and Quantification of Content Validity," *Nurs. Res.*, vol. 35, no. 6, pp. 382–386, Nov. 1986, doi: 10.1097/00006199-198611000-00017.
[36] S. G. Sireci, "Gathering and Analyzing Content Validity Data," *Educational Assessment*, vol. 5, no. 4, pp. 299–321, Oct. 1998, doi: 10.1207/s15326977ea0504_2.

[37] V. Zamanzadeh, A. Ghahramanian, M. Rassouli, A. Abbaszadeh, H. Alavi-Majd, and A.-R. Nikanfar, "Design and Implementation Content Validity Study: Development of an instrument for measuring Patient-Centered Communication," *J. Caring Sci.*, vol. 4, no. 2, pp. 165–178, Jun. 2015, doi: 10.15171/jcs.2015.017.
[38] R. Montes, C. Zuheros, J. Morales, N. Zermeño, J. Duran, and F. Herrera, "Design and consensus content validity of the questionnaire for b-learning education: A 2-Tuple Fuzzy Linguistic Delphi based Decision Support Tool," *Appl. Soft Comput.*, vol. 147, p. 110755, Nov. 2023, doi: 10.1016/j.asoc.2023.110755.
[39] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem. Med. (Zagreb).*, pp. 276–282, 2012, doi: 10.11613/BM.2012.031.
[40] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.
[41] R. H. Riffenburgh, *Statistics in Medicine*, 2nd ed. Elsevier, 2006. doi: 10.1016/B978-0-12-088770-5.X5036-9.
[42] S. D. Moore, G. P. McCabe, and B. A. Craig, *Introduction to the Practice of Statistics*, 9th ed. W. H. Freeman, 2016.
[43] A. Agresti, *Statistical Methods for the Social Sciences*, 5th ed. Pearson, 2018.
[44] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2013. doi: 10.4324/9780203771587.
[45] M. W. Fagerland, S. Lydersen, and P. Laake, "The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional," *BMC Med. Res. Methodol.*, vol. 13, no. 1, p. 91, Dec. 2013, doi: 10.1186/1471-2288-13-91.
[46] G. Lucassen, F. Dalpiaz, J. M. E. M. van der Werf, and S. Brinkkemper, "Forging high-quality User Stories: Towards a discipline for Agile Requirements," in *2015 IEEE 23rd International Requirements Engineering Conference (RE)*, IEEE, Aug. 2015, pp. 126–135. doi: 10.1109/RE.2015.7320415.
[47] A. R. Feinstein and D. V. Cicchetti, "High agreement but low Kappa: I. the problems of two paradoxes," *J. Clin. Epidemiol.*, vol. 43, no. 6, pp. 543–549, Jan. 1990, doi: 10.1016/0895-4356(90)90158-L.
[48] T. Byrt, J. Bishop, and J. B. Carlin, "Bias, prevalence and kappa," *J. Clin. Epidemiol.*, vol. 46, no. 5, pp. 423–429, May 1993, doi: 10.1016/0895-4356(93)90018-V.
[49] P. Pokharel and P. Vaidya, "A Study of User Story in Practice," in *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy, ICDABI 2020*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020. doi: 10.1109/ICDABI51230.2020.9325670.
[50] C. A. Peláez and A. Solano, "A practice for specifying user stories in multimedia system design: An approach to reduce ambiguity.," *Interaction Design and Architecture(s)*, no. 60, pp. 214–236, Mar. 2024, doi: 10.55612/s-5002-060-009.

Authors' Profiles



Muhammad Ihsan Zul is a lecturer at Politeknik Caltex Riau, Indonesia. He holds a master's degree in electrical engineering and information technology from Gadjah Mada University, Indonesia. He is a PhD student at the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM). His dissertation topic is software engineering, focusing on user story, requirement engineering, and design thinking in education.



Suhaila Mohd. Yasin received a B.Sc. degree in science (computing) and an M.Sc. in computer science from Universiti Teknologi Malaysia, Skudai, and a Ph.D. in computer science from The University of Queensland, Australia, in 2020. She is currently a senior lecturer with the Department of Software Engineering and the Software Testing Focus Group leader and principal researcher at Universiti Tun Hussein Onn Malaysia (UTHM). As a rising researcher, she supervises postgraduate students and has authored or co-authored articles in software engineering, modeling, development, and testing.



Ivan Chatisa earned his Master's in Applied Computer Engineering from Politeknik Caltex Riau, Indonesia. He is a lecturer in the Informatics Engineering Program at Politeknik Caltex Riau. His research interests include Software Engineering, the Internet of Things (IoT), and Operating Systems and Computer Networks (OSCN). He has actively developed and implemented various web-based systems for enterprise applications, including sales management systems, enterprise data management platforms, journal management systems, financial information systems, and integrated enterprise information platforms. These projects, conducted in collaboration with technology companies, show his dedication to combining academic research with practical technological solutions.



Fikri Muhaffizh Imani holds a Master of Applied Computer Engineering degree from Politeknik Caltex Riau, Indonesia. He is currently a lecturer in the Information Systems program at Politeknik Caltex Riau and is part of the Software Engineering research group. His teaching focus includes Mobile Programming and Web Programming. His research interests include system integration and full-stack development. He has been involved in various internal and external projects and has published several papers related to software architecture.



Siti Syahidatul Helma earned her master's in Computer Science from IPB University and obtained her Bachelor's in Information Systems from the State Islamic University of Sultan Syarif Kasim Riau in 2020. She is currently a lecturer in the Information Systems program at Politeknik Caltex Riau and serves on the Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM) editorial board. Her research interests include Machine Learning, Data Mining, and Decision Support Systems.



Dadang Syarif Sihabudin Sahid received a bachelor's degree in mathematics from Bandung Institute of Technology, Indonesia, in 1999. The M.Sc. degree in Information Technology was earned at the Universiti Teknologi Malaysia in 2009. Doctoral holder from Universitas Gadjah Mada, Indonesia, in 2018. He has been a lecturer at the Information Technology Department of Politeknik Caltex Riau since July 1st, 2000. He carries out three main activities as a lecturer: teaching, research, and community service. He teaches in Information Systems, Application and Systems Development, IT Project Management, Information Technology Concepts, and IT strategic planning. The research areas focused on IT Project Management, Soft Computing, Context-Aware Computing, and IT Planning. He received many research grants abroad, including collaboration grants and grants from the Higher Education Ministry. He also has experience as deputy director and director at the Politeknik Caltex Riau.

How to cite this paper: Muhammad Ihsan Zul, Suhaila Mohd. Yasin, Ivan Chatisa, Fikri Muhaffizh Imani, Siti Syahidatul Helma, Dadang Syarif Sihabudin Sahid, "Quality Evaluation of Indonesian Student-generated User Stories: Insight from Human and ChatGPT Evaluation", International Journal of Modern Education and Computer Science(IJMECS), Vol.18, No.2, pp. 82-99, 2026. DOI:10.5815/ijmecs.2026.02.05