



Student-Generated User Story Quality: A Study on Practitioner and ChatGPT Evaluation

Muhammad Ihsan Zul^{1*}, Suhaila Mohd. Yasin², Dadang Syarif Sihabudin Sahid³

^{1,2}Department of Software Engineering, Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, Johor, Malaysia

^{1,3}Department of Information Technology, Politeknik Caltex Riau, Pekanbaru, Indonesia

¹ihsan@pcr.ac.id, ²ysuhaila@uthm.edu.my, ³dadang@pcr.ac.id

Abstract

Evaluating the quality of student-generated user stories is important in software engineering education, but only a limited number of industry practitioners can assist. The integration of generative AI can facilitate this process. To do so, the INVEST quality evaluation framework is widely recognized for assessing user story quality; however, prior research has not explored its use in conjunction with generative AI. This study investigated ChatGPT's ability to evaluate user stories using the INVEST framework. This study compares two ChatGPT-based evaluation approaches with those of experienced practitioners, focusing on student-generated user stories. Discrepancies between ChatGPT and practitioner evaluations were measured using Mean Absolute Deviation (MAD), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Statistical significance was tested using the Mann-Whitney U Test. The results indicate that ChatGPT's 1st approach yielded lower discrepancies than practitioner evaluations. Moreover, significance testing showed no statistically significant differences between the ChatGPT and practitioner results for the two INVEST criteria- Independent and Estimable. These findings suggest that the 1st approach can assist in the evaluation process, although practitioners must ensure comprehensive and accurate evaluations. ChatGPT can provide preliminary evaluations in educational contexts, enabling students to receive formative feedback and allowing educators to streamline evaluation processes. Although practitioner validation is still required, their role may shift toward verifying AI-generated results, thus reducing the overall workload and accelerating quality evaluation

Keywords: user stories; INVEST framework; practitioners; ChatGPT; quality evaluation;

How to Cite: M. I. Zul, S. M. Yasin, and D. S. S. Sahid, "Student-Generated User Story Quality: A Study on Practitioner and ChatGPT Evaluation", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 5, pp. 1163 - 1172, Oct. 2025.

Permalink/DOI: <https://doi.org/10.29207/resti.v9i5.6950>

Received: July 12, 2025

Accepted: October 18, 2025

Available Online: October 25, 2025

This is an open-access article under the CC BY 4.0 License

Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Software engineering education increasingly incorporates Agile Software Development (ASD) approaches to prepare students for industry demands. One critical aspect of this preparation is ensuring students can effectively identify and gather user or client requirements. However, capturing these requirements is not enough; the quality of requirement documentation must adhere to established evaluation standards to support efficient development processes [1], [2].

In the requirement engineering phase of ASD, user stories are a widely adopted method for documenting user needs [3]. However, assessing their quality presents significant challenges in educational settings. Various frameworks have been developed for evaluating user story quality [4], including the INVEST

[5] - [7] and the Quality User Story (QUS) framework [8] - [10]. Each framework offers unique perspectives, but their application in student-generated user stories remains a critical research area.

Beyond selecting an appropriate evaluation framework, inviting practitioners is also widely applied to evaluate user story quality in requirement engineering studies [11], [12]. However, engaging industry practitioners introduces additional complexities. Practitioners often have full-time responsibilities, making participation in the learning process challenging to secure. Continuous requests for involvement may disrupt their professional commitments, making relying on practitioners for consistent academic collaboration impractical. Consequently, alternative approaches to user story evaluation are necessary to minimize reliance on industry experts while maintaining assessment quality.

Various tools have been developed to automate user story evaluation to address these challenges. One such tool is AQUASA [13], [14], which facilitates the evaluation of user stories based on the QUS framework. While AQUASA has proven beneficial in enhancing the assessment [15], there is still a gap in automated tools designed to evaluate user stories using the INVEST framework. The lack of INVEST-based evaluation tools presents an opportunity for further research and development in this area.

Research related to the development of student user stories has been widely conducted. Bernardo et. al [16] focused on the instructional methods used to teach students how to write user stories. While the study involved evaluating the quality of user stories using specific criteria, it did not clearly specify the evaluation framework adopted. This distinguishes it from the current study, which explicitly focuses on assessing user story quality using the INVEST framework.

Another relevant study highlights the evaluation of student-generated user stories using a custom-developed tool called the User Story Quality Analyzer (USQA) [8]. Although this study focuses similarly on user story quality, the tools employed differ. Moreover, unlike prior research, the current study goes beyond quality assessment by further validating the evaluation results through statistical analysis, offering a more rigorous methodological contribution.

A similar activity was conducted in a study [17], where students created user stories. However, the key distinction lies in using generative Artificial Intelligence (AI) in the classroom. While both studies incorporate generative AI, the purposes differ. The present study employs generative AI not to generate user stories but to evaluate their quality.

Furthermore, the rapid advancement of generative AI has opened new possibilities for addressing these challenges. With its increasing prominence in various fields, generative AI, such as ChatGPT, offers a promising solution for evaluating user stories. Previous research has explored the potential of ChatGPT in user story assessment [18]; however, these studies have primarily focused on the QUS framework. Despite this limitation, existing research provides valuable insights that can serve as benchmarks for integrating AI-driven tools in user story evaluation.

Building on existing literature, this study investigates the capability of generative AI, specifically ChatGPT, to evaluate user story quality using the INVEST framework. The research focuses on student-generated user stories within a software engineering course and compares two ChatGPT's evaluation approaches with those of experienced practitioners using statistical analysis. One of the contributions of this study lies in its support for the learning process of user story development in software engineering education. By integrating AI-generated feedback, particularly from ChatGPT, the study demonstrates how generative AI

can enhance the quality of student-generated user stories without replacing human evaluation. Instead, ChatGPT serves as a complementary tool that assists academic instruction, especially in contexts where access to industry practitioners is constrained. This approach is expected to enable students to receive structured, consistent, and timely feedback aligned with the INVEST criteria. It fosters a deeper understanding of user story quality and supports their iterative refinement efforts throughout development.

2. Methods

2.1 Research Design

This research is conducted in an academic setting, specifically within a higher education institution. As a result, the user story dataset used in this study is generated from educational activities. The collected data is evaluated through two approaches: evaluation by practitioners and ChatGPT. The evaluation results from both methods are compared using statistical tests to examine the evaluation discrepancy and significance. The findings are analyzed to address the research objectives. A visual representation of the research design can be seen in Figure 1. Furthermore, the details of each research stage are explained in the following subsections.

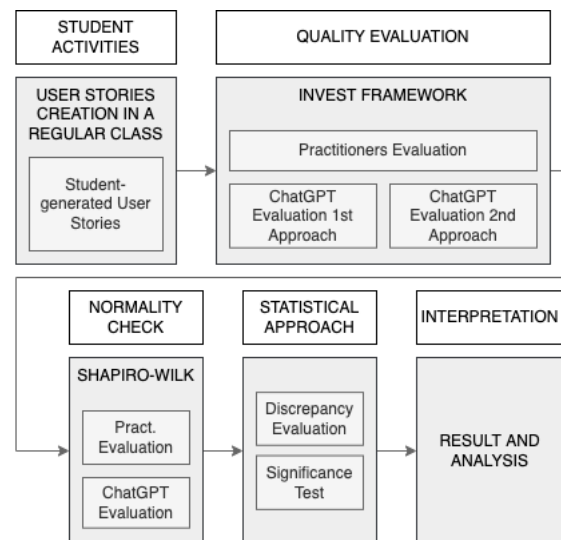


Figure 1. Research Design

2.2 Student-generated User Story Approach

This study's user stories were developed in a regular classroom setting, where students collaborated on real-world case studies provided by an industry partner seeking IT-based solutions. This approach aimed to immerse students in authentic requirement elicitation experiences while enhancing their practical skills in software engineering.

Although this activity formed part of an action research project in a classroom, this paper concentrates explicitly on evaluating the quality of the resulting user stories. In this context, practitioners serve as evaluators, and the role of ChatGPT is examined as a supplement in the AI

evaluation process. Consequently, the methodological aspects of the action research are detailed in a separate publication. A total of 36 students actively participated in this study, which was structured into eight distinct groups. Each group consisted of four to five members, allowing for effective collaboration and idea sharing. During the course, these groups worked to brainstorm and develop a total of 67 user stories. This process encouraged teamwork and creativity as students turned their ideas into clear stories.

2.3 Quality Evaluation Framework

Based on previous studies, various frameworks have been developed to evaluate the quality of user stories. Two widely recognized frameworks are INVEST [7] and Quality User Story (QUS) [10]. INVEST comprises six key criteria: independent, negotiable, valuable, estimable, small, and testable. Meanwhile, QUS comprises 13 criteria categorized into three groups: syntactic, semantic, and pragmatic [19], [20]. Both approaches have been extensively utilized in research and practice to assess the quality of user stories in ASD.

INVEST was the first model introduced among these frameworks and has been applied in various approaches. One of its key developments is the INVEST grid, which serves as a reference for evaluating user story quality [5]. The INVEST grid assesses user stories across four dimensions using a Likert scale: poor/absent (0), fair (1), good (2), and excellent (3). Numerous studies have adopted INVEST as the primary framework for evaluating user stories in academic and industrial contexts [11], [21], [22]. Therefore, by applying the INVEST grid, this study

employs INVEST as the evaluation framework for user stories.

2.4 User Story Quality Evaluation Approach

The quality of user stories was evaluated through the practitioner and ChatGPT evaluations. Industry practitioners are involved in the evaluation as practitioners. Specifically, each practitioner evaluated user stories from 2-3 teams, estimating a total of 16–25 user stories per practitioner. The invitation and inquiry were achieved by email and involved 18 software engineering practitioners from Indonesia, Malaysia, Singapore, and Australia, all of whom possessed 3–19 years of experience in software engineering, especially in ASD. The detailed practitioner data can be seen in Table 1.

Table 1. Software Engineering Practitioners

No	Practitioner Initial (Experience in Year)	Company Base Country
1	DB (10), FM (8), WJ (5), AN (4), DK (3), BO (2), SO (8), BZ (3), DE (3), MD (19), FA (4)	Indonesia
2	RM (9), NA (8)	Singapore
3	AJ (9), IA (9)	Malaysia
4	FS (11)	Australia

The evaluations conducted by practitioners are inherently influenced by bias. Bias reduction was performed for all practitioner evaluation results to ensure quality. This action was taken by analyzing their evaluation scores. Thus, the final evaluation score was determined by selecting the median value for practitioners' scores.

Table 2. ChatGPT 1st Approach Evaluation

Context	ChatGPT Prompt/Instruction
Understanding the INVEST Grid	The evaluation process must begin with thoroughly examining the INVEST Grid document. You are required to understand each of the six INVEST criteria—Independent, Negotiable, Valuable, Estimable, Small, and Testable—along with their respective definitions and scaling guidelines. Importantly, the scoring range from 0 to 3 must be applied objectively. A score of 0 should not be avoided or overlooked; it must be assigned whenever a User Story fails to meet even the minimum expectations of a particular criterion. [INVEST Grid document was uploaded]
System/ Project Context	The User Stories under evaluation pertain to developing an information system designed to track and monitor project progress within a corporate setting. The intended users of this system include company employees, senior management, and commissioners. These contextual details must be taken into account when assessing the relevance and utility of each User Story. No new requirements are to be elicited during the evaluation process; the analysis is limited strictly to the user stories provided
Initial Data Familiarization	Before commencing the evaluation, you must read and comprehend all User Stories listed in the dataset. This familiarization stage is critical for building contextual awareness and ensuring informed judgment during scoring. No evaluative action should be taken at this point; instead, the focus should be on understanding the content, structure, and intent of each User Story. [user story dataset was uploaded]
Team-Specific Evaluation for Certain Criteria	Special attention must be given to the criteria “Independent” and “Small.” These two aspects should only be assessed in the context of individual teams. Each User Story should be compared and interpreted relative to other user stories created by the same team, as inter-team dependencies are considered irrelevant for these dimensions. This ensures that independence and size are measured consistently within bounded project scopes.
Context-Based Evaluation for Remaining Criteria	The remaining criteria—Negotiable, Valuable, Estimable, and Testable—must be evaluated based on contextual interpretation and the evaluator’s expertise. These dimensions require professional judgment that accounts for the narrative, clarity, and purpose embedded within each User Story. Evaluators are expected to go beyond keyword matching and consider how each story contributes to the system’s usefulness, feasibility, and testability in practical terms. Assess the “Testable” criterion using a more flexible approach. Assign reasonable scores where there are hints of testability, even if not fully explicit.
Three-Round Evaluation Process	The evaluation is carried out in three iterative rounds, with each round allowing for score adjustments based on improved understanding or contextual insights. To ensure transparency and traceability, each round must generate a separate CSV file containing the evaluation results. After every round, the outcomes should be reviewed with a researcher to validate alignment with the data and the INVEST Grid, and to inform revisions for the subsequent iteration.

In the ChatGPT evaluation, ChatGPT 4o was employed. The process was structured systematically to ensure evaluation consistency in each user story. This evaluation is conducted through two approaches. In the first (1st) approach, ChatGPT was trained to understand the INVEST grid by uploading the grid document. Subsequently, a set of instructions was used to verify ChatGPT's ability to recognize and apply the INVEST criteria and its context precisely. Following this step, the user stories file was uploaded to ChatGPT for quality evaluation. After that, the result was analyzed to identify any missing criteria, skipped values, or format inconsistencies. The evaluation was conducted three times in a sequence of sessions to provide the scoring result. This approach was adopted by the works of Ronanki et al. [18]. The ChatGPT instruction in this approach is described in Table 2.

Lastly, the final ChatGPT score was determined by taking the median value of these evaluations. This approach ensured consistency and minimized the influence of variations in ChatGPT evaluation.

In the second (2nd) approach, ChatGPT was trained to understand the INVEST grid, as well as the 1st approach. Next, the user story file was uploaded, and the evaluation was conducted using the criteria. The emphasis on each criterion is intended to sharpen the instructions defined in the INVEST Grid, ensuring that ChatGPT applies a consistent interpretation across all criteria.

This approach also ensures that ChatGPT remains focused on the specific evaluation reference for each criterion and avoids conflating interpretations across different criteria. Each prompt for all criteria can be seen in Table 3.

Table 3. ChatGPT 2nd Approach Evaluation

Context	ChatGPT Prompt/Instruction
Understanding the INVEST Grid	Please evaluate each user story based on the six INVEST criteria (Independent, Negotiable, Valuable, Estimable, Small, and Testable), using the story's wording and context within its project team. This evaluation follows the standard INVEST Grid but applies a more refined interpretation tailored to actual user intent, clarity, and development feasibility. [INVEST Grid document was uploaded]
Independent Criterion	For the Independent (I) criterion, assess dependencies only within the same team's project. A story is considered "Independent" if it does not rely on another story from the same team. Ignore any dependencies that might exist across different teams. Provide the result in csv file! [user story dataset was uploaded]
Negotiable Criterion	For the Negotiable (N) criterion evaluates how open a User Story is to discussion between users and developers. A story that is clearly written, focused on user needs, and open to feedback receives a score of 3. These stories allow both sides to collaborate and adjust requirements as needed. A score of 2 is given when the story is understandable but doesn't strongly encourage negotiation. If the story describes a function but feels rigid or final, it is scored 1. A score of 0 is used when the story reads like a technical specification or command with no flexibility. This evaluation focuses on the overall tone and context, not just specific words. Provide the result in csv file!
Valuable Criterion	For the Valuable (V) criterion, assess whether the story communicates a clear and meaningful benefit to the user. A score of 3 should be given when the story emphasizes a user goal or outcome (e.g., understanding progress, making evaluations). A score of 2 applies when the function supports the user, but the value is more implicit. Use 1 for weakly stated benefits, and 0 for stories lacking any clear value. Provide the result in csv file!
Estimable Criterion	For Estimable (E), assess whether the story provides enough clarity through actor, action, and goal to allow reasonable estimation by a developer. Provide the result in csv file!
Small Criterion	For Small (S) criterion. Judge whether the story is appropriately sized for a sprint; clear, focused stories should score higher, while vague or overly broad ones should score lower. Provide the result in csv file!
Testable Criterion	For Testable (T) by determining whether the outcome can be verified through observable or testable results. Give the highest scores to stories that lead to specific and verifiable outputs, and lower scores to those with vague or ambiguous outcomes. This strengthened interpretation of the INVEST Grid is intended to support consistent, context-aware evaluation aligned with agile best practices. Provide the result in csv file!

2.5 Statistical Measurement and Test

To measure the evaluation discrepancies between the scores provided by Practitioners and ChatGPT, two distinct approaches were analyzed using statistical measures, including Mean Absolute Deviation (MAD), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) [23]. These metrics serve to quantify the disagreement between the two evaluators. MAD calculates the average absolute difference in scores, indicating how far apart the evaluations tend to be on average [24], [25]. MSE, on the other hand, emphasizes larger discrepancies by squaring the differences, thereby highlighting more significant errors. RMSE, as the square root of MSE, translates these discrepancies back to the same scale as the original ratings, making

the results easier to interpret. The formulas of these metrics are shown in Formulas 1 to 3.

$$MAD = \frac{\sum_{i=1}^n |(y_i - \hat{y}_i)|}{n} \quad (1)$$

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (3)$$

y_i is the practitioner evaluation score, $\hat{y}_i$ ChatGPT evaluation score, and n is the total number of user story. These metrics comprehensively analyze the discrepancies observed when comparing practitioner evaluations against ChatGPT's first and second approaches.

A normality check was conducted to determine whether the data followed a normal distribution, ensuring an appropriate significance test method was selected for analysis. If the data were non-normally distributed, a non-parametric test would be applied to compare the significance of evaluation results between practitioners and ChatGPT. Data type and sample size also influenced the choice of a significant test method. Since the INVEST Grid categorizes data as ordinal and independent, the Shapiro-Wilk test validated the statistical approach [26], [27].

To further analyze the significant differences between the two evaluation approaches, a statistical comparison was performed to examine the significance of evaluation scores [28], [29]. Each INVEST criterion was assessed separately to evaluate the practitioners' and ChatGPT's evaluations significantly. The agreement metrics and significance analysis are presented in the results section.

3. Results and Discussions

3.1 Practitioner and ChatGPT Evaluation Result

67 user stories created by students were evaluated using two approaches: evaluation conducted by practitioners and ChatGPT. Among the 18 practitioners invited to participate in the evaluation, 16 submitted their scoring results. Each practitioner was assigned to evaluate 2 to 4 groups consisting of 7 to 10 user stories. This distribution was designed to respect the time constraints of practitioners who volunteered for the evaluation.

Outlier detection was performed to ensure the reliability of the practitioners' assessments and identify potential biases and inconsistencies in their evaluations. The analysis revealed that three practitioners assigned identical scores across all evaluation criteria, indicating they filled out the assessments arbitrarily without thoroughly reviewing the user stories. As a result, their evaluations were excluded from further analysis. Consequently, the final processed evaluations used for calculating the median score were based on the assessments of 13 practitioners.

The ChatGPT evaluation was conducted using two approaches; the 1st approach was adopted from Ronanki et al. [18], and the 2nd approach was proposed in this study. The results from each prompt were collected, and the median score was calculated to generate a single consolidated evaluation score. The comparison of the practitioner's and ChatGPT evaluation results can be seen in Figure 2. While the left bar represents the evaluations by practitioners, the center bar represents the evaluations by ChatGPT 1st Approach, and the right bar represents the evaluations by ChatGPT 2nd Approach. Each bar chart displays the six INVEST criteria, and the y-axis represents the frequency of each score category.

Figure 2 shows the diversity of evaluation scores given by the three evaluation approaches used in this study. None of the criteria showed near-perfect consistency.

Based on this figure, it is necessary to examine the extent of the discrepancy between the evaluation results of the two ChatGPT approaches and how significant the differences are. This discussion can be found in the following subsection.

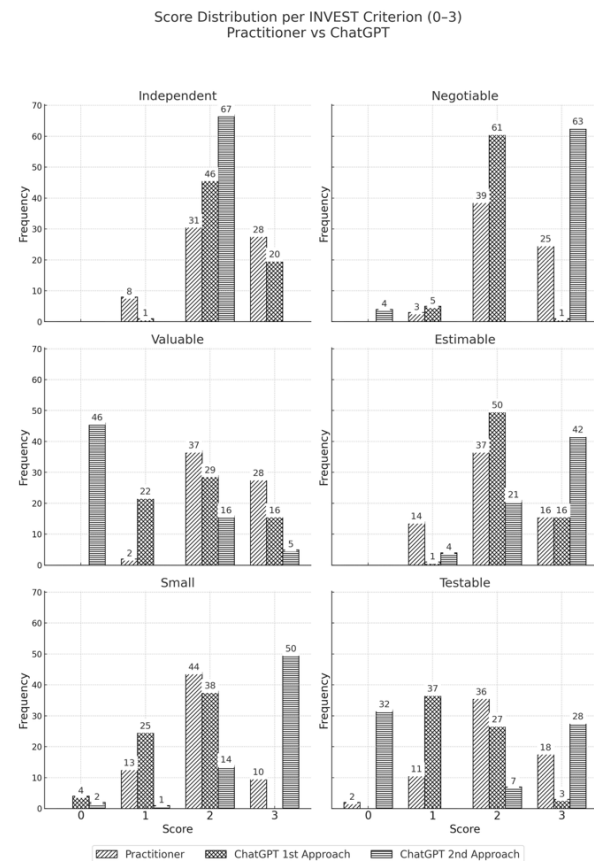


Figure 2. Distribution of Evaluation Result based on INVEST Grid

3.2 Evaluation Discrepancy Result

As explained in subsection 2.4, evaluation discrepancies were measured between practitioners with the ChatGPT 1st approach and the ChatGPT 2nd approach. A complete overview of the discrepancy measurement results for each INVEST criterion can be seen in Table 4.

Table 4. The Discrepancy Measurement Result

Criterion	Practitioner vs. ChatGPT 1 st Approach			Practitioner vs. ChatGPT 2 nd Approach		
	MAD	MSE	RMSE	MAD	MSE	RMSE
Independent	0.49	0.49	0.70	0.54	0.54	0.73
Negotiable	0.45	0.57	0.75	0.85	1.30	1.14
Valuable	0.57	0.75	0.86	1.69	3.90	1.97
Estimable	0.58	0.64	0.80	0.84	1.19	1.09
Small	0.57	0.87	0.93	0.96	1.25	1.12
Testable	0.88	1.21	1.10	1.51	3.18	1.78

The discrepancy analysis between practitioner evaluations and the two ChatGPT evaluations across the INVEST criteria indicates notable differences in alignment levels. Based on the MAD, ChatGPT's 1st approach demonstrates relatively lower discrepancies, with MAD values ranging from 0.49 (Independent) to 0.88 (Testable). These values indicate mild to moderate divergence from practitioner evaluation. In contrast, the

ChatGPT 2nd approach exhibits higher discrepancy levels in several criteria, notably Valuable (MAD=1.69) and Testable (MAD=1.51), suggesting considerable deviation.

The MSE further substantiates these patterns. For the 1st approach, MSE values are consistently below 1.5, with the highest at 1.21 (Testable), indicating relatively contained error magnitudes. Meanwhile, the 2nd approach records the highest MSE of 3.90 (Valuable) and 3.18 (Testable), reflecting larger squared discrepancies and suggesting outliers or divergence in those assessments.

The RMSE values indicate that the ChatGPT 1st and 2nd approaches yield different results. This is evident from the number of criteria with RMSE values in the 2nd approach exceeding one, which has five criteria (Negotiable, Valuable, Estimable, Small, and Testable). When compared based on the average metrics across all criteria, the ChatGPT 1st approach performs better than the ChatGPT 2nd approach. The complete overview of the average metrics comparison based on each metric is presented in Figure 3.

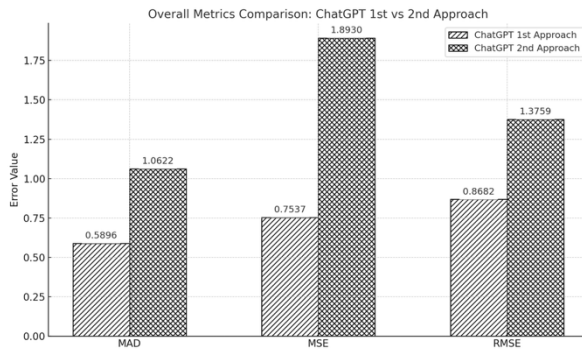


Figure 3. Overall Metrics Comparison in Average

Figure 3 illustrates the average values of all discrepancy results based on the MAD, MSE, and RMSE calculations across all INVEST criteria, as presented in Table 4. It can be observed that the ChatGPT 1st Approach performs considerably better than the 2nd Approach, which exhibits discrepancy values exceeding 1.

The key findings suggest that ChatGPT's 1st approach maintains lower discrepancy across all metrics, making it more consistent and closer to practitioner-level evaluation standards. The observed discrepancy may be attributable to the more comprehensive training methodology employed in the 1st approach, which incorporates knowledge verification and a three-time evaluation process. This approach is designed to ensure the consistency and reliability of the evaluation outcomes produced by ChatGPT. However, these results are still unsatisfactory regarding the Testable criterion. This can be seen from the discrepancy values in the MAD, MSE, and RMSE.

Based on this discrepancy analysis, although ChatGPT's 1st approach demonstrates a high level of discrepancy across several criteria, ChatGPT is a

helpful tool to support the user story evaluation process. Nevertheless, the involvement of human practitioners remains essential to revalidate and ensure the accuracy and reliability of ChatGPT-generated evaluations.

3.3 Statistical Test Practitioners vs. ChatGPT Evaluation

The selection of an appropriate statistical method begins with assessing the normality of the evaluation data from both practitioners and ChatGPT. Calculations were performed using the Shapiro-Wilk test to achieve this. The results indicate that neither dataset follows a normal distribution when considered as a whole and analyzed according to individual criteria. As a result, the most suitable statistical test is a non-parametric approach, specifically the Mann-Whitney U test [30]. This decision is further supported by the nature of the data, which is ordinal (Likert scale) and two independent samples.

Furthermore, because the dataset contains more than 20 data points, the critical value of the Mann-Whitney U test cannot be used. Instead, it is replaced by the normal distribution. Consequently, conclusions will be drawn based on the p-value calculation, which is obtained by determining the z-score of the U-value for each criterion [31]. The Mann-Whitney U test equation is provided in Formulas 4 to 6.

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, \text{ and} \quad (4)$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2, \text{ and then} \quad (5)$$

$$U = \min(U_1, U_2) \quad (6)$$

n_1 is the number of practitioner evaluation scores, n_2 is the number of ChatGPT evaluation scores, R_1 is the sum of ranks for practitioner scores, and R_2 is the sum of ranks for ChatGPT scores. U is the Mann-Whitney U Statistic. The U-statistic is converted into a z-score using Formula 7.

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \quad (7)$$

The final p-value is calculated using the cumulative distribution formula (CDF) given in Formula 8.

$$p = 2 \times (1 - \Phi(|Z|)) \quad (8)$$

The $\Phi(|Z|)$ is CDF of the normal distribution and $|Z|$ is the absolute value of the z-score.

The threshold value used for the p-value is 5% (0.05) to define the significance level of the evaluation differences between practitioners and ChatGPT approaches. The results of the Mann-Whitney U test for the significance test between the evaluation results of Practitioners vs. ChatGPT 1st Approach and Practitioners vs. ChatGPT 2nd Approach can be seen in Tables 5 and 6.

Table 5 presents the results of the Mann-Whitney U Test calculations. The table shows that several criteria

in both the 1st approach and 2nd approach have significant results when compared to the practitioners' evaluation. In the 1st approach, the Independent and Estimable criteria indicate a non-significant difference between ChatGPT's and practitioners' evaluations, with a p-value >0.05 . Meanwhile, the other three criteria, Negotiable, Valuable, Small, and Testable, show significant results in this test. These results indicate that two INVEST criteria in the 1st approach align between practitioners' evaluations and ChatGPT's.

Table 5. Mann-Whitney U Test Practitioner vs. ChatGPT 1st Approach

Criterion	Mann Whitney U	z-score	p-value (0.05)
Independent	2344.0	0.442764	0.657937
Negotiable	3054.5	3.604408	0.000313
Valuable	3024.5	3.470911	0.000519
Estimable	1913.0	-1.475137	0.140176
Small	2996.5	3.346315	0.000819
Testable	3200.5	4.254092	0.000021

Furthermore, in Table 6, the significance test calculations between the practitioners' evaluation and the 2nd approach show different results. This test indicates that the 2nd approach produces greater divergence from practitioners' evaluation—all criteria except Testable show statistically significant differences. Significantly, the Negotiable, Valuable, and Small criteria show marked differences in both methods, yet these variations are more pronounced in the 2nd approach.

Table 6. Mann-Whitney U Test Practitioner vs. ChatGPT 2nd Approach

Criterion	Mann Whitney U	z-score	p-value (0.05)
Independent	2914.5	2.981424	0.002869
Negotiable	1055.5	-5.290915	0.000000
Valuable	3896.0	7.348988	0.000000
Estimable	1300.5	-4.200693	0.000027
Small	892.5	-6.016247	0.000000
Testable	2616.0	1.653133	0.098304

Figure 4 compares the significance test results between the ChatGPT 1st and 2nd approaches based on the practitioner evaluations. The red line in the figure indicates the threshold for statistical significance (p-value=0.05).

Based on Figure 4, the difference is also contradictory when comparing the 1st and 2nd approaches, as depicted in Figure 4, where the 1st approach is non-significant, while the 2nd approach is significant. This also occurs vice versa. For example: Estimable in the 1st approach is declared non-significant (p-value=0.140176 > 0.05), but in the 2nd approach it is declared highly significant (p-value=0.000027 < 0.05), and conversely in Testable, in the 1st approach it is declared highly significant (p-value=0.000021 < 0.05), but in the 2nd approach it is non-significant (p-value=0.098304 > 0.05). However, there are consistently significant criteria when testing the results between practitioners using the two approaches, such as Negotiable, Valuable, and Small. These findings are illustrated in the comprehensive

visualization of the p-value comparison, as depicted in Figure 4.

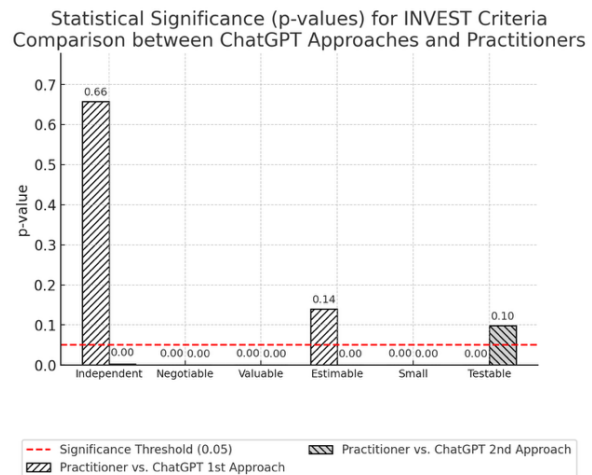


Figure 4. p-value Comparison in the Mann-Whitney U Test

Key findings from this significance test are: 1) Not all criteria show non-significant differences with the results of the practitioners' evaluation, indicating that there are differences in evaluation methods between practitioners and ChatGPT, even though two different approaches were used; 2) differences in instruction/prompt approaches when evaluating user stories using ChatGPT significantly influence the evaluation results. They can even be contradictory. 3) Among all the INVEST criteria, Negotiable, Valuable, and Small present challenges regarding the consistency between ChatGPT's evaluation results (from both the 1st and 2nd approaches) and those of practitioners. It is essential to address these discrepancies in future studies.

3.4 Discussion

The quality evaluation of the student-generated user stories involving practitioners and two distinct ChatGPT-based approaches has been conducted comprehensively. Discrepancy analysis reveals that the 1st approach exhibits lower disagreement levels than the 2nd. This outcome is likely attributed to the more comprehensive training process and the additional consistency checks performed three times during the development of the first approach. From the perspective of statistical significance, only 3 (Independent, Estimable, and Testable) of the six identified criteria showed non-significant differences in value compared to the results of the practitioners' evaluation. The results of these three non-significant criteria were obtained from two different ChatGPT approaches (1st approach: Independent and Estimable, and 2nd approach: Testable). The results suggest that while both approaches align with practitioner evaluation for specific criteria, neither fully replicates human evaluation. These findings reinforce that the evaluation process should not be delegated entirely to ChatGPT; practitioners' involvement remains essential for validating AI-generated evaluations. Moreover, the variability of prompt strategies given to ChatGPT may

significantly influence its outputs, sometimes resulting in contradictory evaluations. This is evident in the contrasting significance test outcomes between the 1st and 2nd approaches.

Nonetheless, the study offers notable advantages over related work. For example, Ronanki et al. [18] evaluated fewer user stories and relied on sampling, whereas the present study analyzes complete user story sets derived from real student projects. Additionally, this study incorporates a broader set of metrics—examining discrepancy and statistical significance—whereas Ronanki's study focused solely on inter-rater agreement. Using these complementary metrics enables a deeper understanding of evaluation consistency and the degree of divergence across human and AI assessments. Despite methodological differences, both studies converge on a key insight: the role of human practitioners remains critical in evaluating user story quality, even when leveraging ChatGPT.

Despite several findings, the study has limitations. First, only two prompt strategies were explored; alternative prompt formulations may produce different results, a topic that warrants further investigation. Second, the user stories evaluated were written in Bahasa Indonesia. Evaluations might differ if the user stories were authored initially in English, limiting the findings' generalizability. Third, the Likert scale operationalizing the INVEST grid may not fully capture the nuances of evaluating a user story set, raising concerns about measurement validity.

The findings of this study carry significant implications for software engineering education. The scarcity of practitioner availability for user story evaluation can be mitigated by adopting the AI-assisted approach proposed herein. Such a strategy could accelerate the evaluation process while reducing dependency on expert involvement, particularly in academic settings. Rather than replacing practitioners, ChatGPT can serve as a supplementary tool, with human validation retained as a necessary checkpoint. This approach could improve the quality of student-generated user stories by integrating Generative AI as a supportive evaluation tool in instructional contexts.

3.5 Future Work

This study also presents opportunities for further research, particularly in selecting quality frameworks. The same prompting methodology applied in this study can be extended to alternative frameworks, such as the QUS Framework [10], [13], which offers different criteria for evaluating user stories. Exploring the applicability of ChatGPT across various quality frameworks may provide insights into the generalizability and adaptability of generative AI in diverse evaluation contexts. Regarding the evaluation metrics, incorporating a broader range of inter-rater agreement metrics presents an opportunity to conduct a more comprehensive comparative analysis. Future studies can better understand evaluation consistency

and reliability by including alternative statistical measures and agreement indices.

Furthermore, from an educational standpoint, the evaluation method explored in this study can be used to improve the quality of user stories produced by students in SEE. ChatGPT-generated recommendations may serve as formative feedback, guiding students in revising and refining their user stories. This strategy supports learning and empowers educators to scale feedback delivery, ensuring quality assurance through practitioner validation.

4. Conclusions

This study compared the quality evaluation of user stories conducted by practitioners with two ChatGPT-based approaches using the INVEST evaluation framework. The discrepancy analysis revealed that the ChatGPT 1st approach produced results that were more closely aligned with those of practitioners. Regarding statistical significance, two out of the six INVEST criteria (Independent and Estimable) exhibited p-values that were not significantly different from practitioner assessments. These findings suggest that the 1st approach is considered a viable support tool for evaluating user stories produced by students. However, practitioner involvement remains essential, as the ChatGPT-based approaches did not demonstrate consistent reliability across all INVEST criteria. Consequently, the role of practitioners or domain experts in user story evaluation cannot be entirely replaced by generative AI models. Despite these findings, the study provides valuable insights for educators looking to incorporate AI tools into user story evaluation. ChatGPT can speed up the evaluation process if used alongside the INVEST framework, helping reduce the need for many practitioners while still preserving their essential role in education. Practitioners can also validate the results generated by ChatGPT.

This study offers several directions for future research. The prompting systematic approach can be applied here to other quality frameworks, such as the QUS Framework, to explore how well ChatGPT performs across different evaluation frameworks. This could help assess the generalizability of generative AI in user story evaluation. Future work can also include a broader range of inter-rater agreement metrics and statistical measures to better understand the consistency and reliability of AI and human evaluations. In educational settings, ChatGPT has the potential to provide formative feedback to students, helping them improve their user stories. This approach supports learning and allows educators to scale feedback while relying on practitioners for final validation.

Acknowledgements

Acknowledgment is given to the Politeknik Chevron Riau Foundation for supporting this work through its PhD Scholarship program.

References

- [1] A. Rasheed et al., "Requirement Engineering Challenges in Agile Software Development," *Math Probl Eng*, vol. 2021, pp. 1–18, May 2021, doi: 10.1155/2021/6696695.
- [2] M. I. Zul, S. M. Yasin, and D. S. S. Sahid, "Exploring Requirement Engineering Challenges in Software Development: Insights from Global and Indonesian Landscape," in *2024 4th International Conference on Electrical Engineering and Informatics (ICon EEI)*, IEEE, Oct. 2024, pp. 136–141. doi: 10.1109/IConEEI64414.2024.10748069.
- [3] A. Hendriana, T. Raharjo, and A. Nurfitriani, "Approaches in Determining User Story Quality through Requirement Elicitation: A Systematic Literature Review," *Indonesian Journal of Computer Science*, vol. 12, no. 6, Jan. 2024, doi: 10.33022/ijcs.v12i6.3639.
- [4] G. B. Herwanto, G. Quirchmayr, and A. M. Tjoa, "Leveraging NLP Techniques for Privacy Requirements Engineering in User Stories," *IEEE Access*, vol. 12, pp. 22167–22189, 2024, doi: 10.1109/ACCESS.2024.3364533.
- [5] 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.
- [6] 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.
- [7] B. Wake, "INVEST in Good Stories, and SMART Tasks," XP123: Exploring Extreme Programming.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] Y. Li, J. Keung, Z. Yang, X. Ma, J. Zhang, and S. Liu, "SimAC: simulating agile collaboration to generate acceptance criteria in user story elaboration," *Automated Software Engineering*, vol. 31, no. 2, p. 55, Nov. 2024, doi: 10.1007/s10515-024-00448-7.
- [12] 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 Et," 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.
- [13] 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.
- [14] S. N. F. N. B. Mustaffa, J. Bin Sallim, and R. B. Mohamed, "Enhancing High-Quality User Stories with AQUASA: 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.
- [15] E. Scott, T. Töemets, and D. Pfahl, "An Empirical Study of User Story Quality and Its Impact on Open Source Project Performance," in *Software Quality: Future Perspectives on Software Engineering Quality. SWQD 2021*, vol. 404, Springer, Cham, 2021, pp. 119–138. doi: 10.1007/978-3-030-65854-0_10.
- [16] J. Bernardo, J. Frota, T. Pessoa, M. Pessoa, B. Gadelha, and T. Conte, "A Journey Through Cyberium: Creating User Stories for a Dystopic World to Teach Requirements Engineering," in *2025 IEEE/ACM 37th International Conference on Software Engineering Education and Training (CSEE&T)*, IEEE, Apr. 2025, pp. 276–286. doi: 10.1109/CSEET66350.2025.00035.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] E. Trisnawati, I. K. Raharjana, T. Taufik, A. H. Basori, A. B. F. Mansur, and N. A. Alghanmi, "Analyzing Variances in User Story Characteristics: A Comparative Study of Stakeholders with Diverse Domain and Technical Knowledge in Software Requirements Elicitation," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 1, pp. 110–125, Feb. 2024, doi: 10.20473/jisebi.10.1.110-125.
- [21] A. Ananjeva, J. S. Persson, and A. Bruun, "Integrating UX work with agile development through user stories: An action research study in a small software company," *Journal of Systems and Software*, vol. 170, no. December 2020, Dec. 2020, doi: 10.1016/j.jss.2020.110785.
- [22] 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.
- [23] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geosci Model Dev*, vol. 15, no. 14, pp. 5481–5487, Jul. 2022, doi: 10.5194/gmd-15-5481-2022.
- [24] E. Pinsky and S. Klawansky, "MAD (about median) vs. quantile-based alternatives for classical standard deviation, skewness, and kurtosis," *Front Appl Math Stat*, vol. 9, Jun. 2023, doi: 10.3389/fams.2023.1206537.
- [25] G. D. Kader, "Means and MADs," *Mathematics Teaching in the Middle School*, vol. 4, no. 6, pp. 398–403, Mar. 1999, doi: 10.5951/MTMS.4.6.0398.
- [26] E. González-Estrada and W. Cosmes, "Shapiro–Wilk test for skew normal distributions based on data transformations," *J Stat Comput Simul*, vol. 89, no. 17, pp. 3258–3272, Nov. 2019, doi: 10.1080/00949655.2019.1658763.
- [27] S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, vol. 52, no. 3/4, p. 591, Dec. 1965, doi: 10.2307/2333709.
- [28] R. W. Emerson, "Mann-Whitney U test and t-test," *J Vis Impair Blind*, vol. 117, no. 1, pp. 99–100, Jan. 2023, doi: 10.1177/0145482X221150592.

- [29] J. C. F. de Winter and D. Dodou, "Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon," *Practical Assessment, Research & Evaluation*, vol. 15, no. 12, 2010.
- [30] N. K. Malhotra, D. Nunan, and D. F. Birks, *Marketing research: an applied approach*, Fifth Edition. New York: Pearson, 2017.
- [31] T. D. Gauthier and M. E. Hawley, "Statistical Method," in *Introduction to Environmental Forensics*, Elsevier, 2007, pp. 129–183. doi: 10.1016/B978-012369522-2/50006-3.