ASEAN Engineering Journal

Full Paper

# USER STORY QUALITY EVALUATION: ANALYZING FRAMEWORKS AND APPLICATION METHODS

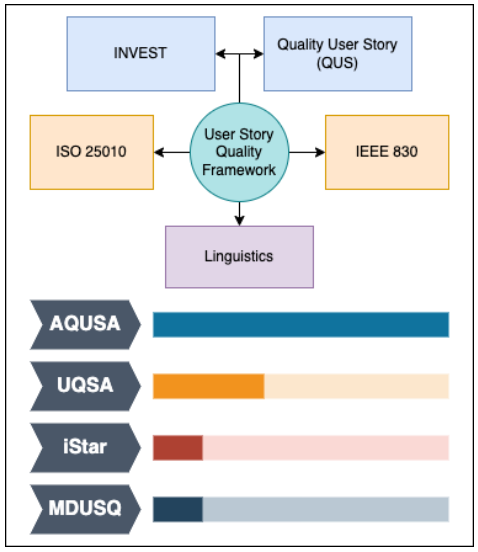
Muhammad Ihsan Zul[a,b*], Suhaila Mohd. Yasin[b], Dadang Syarif Sihabudin Sahid[a]

[a]Department of Information Technology, Politeknik Caltex Riau, Pekanbaru, Indonesia
[b]Department of Software Engineering, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

**Article history**
Received
07 February 2025
Received in revised form
05 May 2025
Accepted
14 May 2025
Published online
30 November 2025

*Corresponding author
ihsan@pcr.ac.id

**Graphical abstract**




## Abstract

Quality issues in User Stories (US) frequently arise during implementation, with common problems including ambiguity, integration challenges, and inconsistencies in templates or formal structure. These challenges emphasize the critical need to evaluate the quality of the US, prompting extensive research focused on developing methods and innovations to improve the requirement elicitation process. This study aims to analyze the frameworks, application methods, and tools employed in previous research to assess US quality. Utilizing the Kitchenham framework, 26 relevant studies were systematically reviewed. Next, the analysis identified six prominent quality evaluation frameworks: INVEST, Quality User Story (QUS), ISO 25010, IEEE 830, USQM, and linguistic-based approaches. The application of these frameworks often involves experts, practitioners, and artificial intelligence (AI). Meanwhile, among the tools reviewed, AQUSA emerged as the most frequently used due to its alignment with QUS standards. These findings highlight the adaptability of existing frameworks and tools while underscoring the potential for further integrating generative AI to enhance the accuracy and efficiency of US quality evaluations. Subsequently, future research should explore innovative AI-based methods to advance this critical area of requirement engineering.

*Keywords*: user stories, quality, INVEST, Quality User Story, AQUSA, generative AI



## 1.0 INTRODUCTION

User Story (US) has emerged as a key approach in the process of gathering requirements, especially in the context of Agile Software Development (ASD) practices [1]. Numerous software engineering teams have widely adopted its application because it facilitates effective communication between development teams and stakeholders [2], [3], [4].

However, issues related to the quality of the US frequently arise in their implementation. Common challenges observed in the US include ambiguity [4], [5], [6], [7], integration [8], [9], [10], and inconsistencies in templates or formality of format [11], [12], [13]. These quality issues can delay achieving IT project targets and sometimes fail.

Recognizing the importance of US quality, various studies have focused on developing methods and innovations to assess the quality of US generated during the requirement elicitation process. In particular, one widely used framework is INVEST [14], [15], which emphasizes Independent, Negotiable, Valuable, Estimable, Small, and Testable criteria. However, INVEST is not the only approach; several studies utilize international standards such as ISO and IEEE [16].

With recent technological advancements, particularly the rapid progress in Artificial Intelligence (AI), machine learning, and deep learning-based approaches have been applied to evaluate the quality of the US [17]. Using Natural Language Processing (NLP) techniques, the textual format of the US can be processed and analyzed automatically to identify deficiencies and improve its quality [18].

Research on Systematic Literature Review (SLR) regarding the US has been conducted in various previous studies. For example, a study by Amna and Poles [7] broadly reviewed the US, with a research scope starting in 2021. However, due to its general nature, this study did not specifically address the quality of the US. Another study by Hendriana et al. [20] focused on approaches used to determine the quality of generated US. Nevertheless, this research did not discuss the US evaluation standards employed based on the framework proposed by the previous studies. Both studies' data were sourced from several digital libraries, including IEEE Xplore, ProQuest, and ScienceDirect.

Meanwhile, Amna and Poels [9] extended their earlier research by focusing on issues of ambiguity in the US. This study specifically examined various problems investigated, proposed solutions, and the methods of validation and evaluation applied. In a different context, Raharjana et al. [10] explored the application of NLP in the US. They concluded that NLP could assist software engineers in managing the US more efficiently. The study also highlighted that opportunities for further development in applying NLP to the US remain abundant.

Aside from this, a study by Kustiawan et al. [1] examined how the US is utilized in the requirement elicitation process, including challenges encountered in its implementation. Consequently, this study identified that ambiguity is the most frequently reported issue, occurring 18 times, followed by incompleteness, which was noted 11 times. Previous SLRs have primarily examined the US in a broad context or addressed specific issues related to the US's application, such as ambiguity. While Hendriana et al. [18] did concentrate on the quality of the US, their discussion was limited to various approaches and did not reference any established frameworks for assessing US quality.

This study explicitly identifies frameworks for evaluating the quality of the US and the application methods in utilizing these frameworks. Furthermore, it explores tools and models that have been developed to assess US quality. Thus, the objective of this study is to systematically identify the frameworks, application methods, and tools used for US quality evaluation over the past five years. This research contributes to helping researchers and practitioners by providing insights into available approaches in evaluating the quality of the US. It seeks to guide the selection of frameworks and methods to measure the US quality, tackle challenges in requirement elicitation, and enhance Agile Software Development practices.

## 2.0 METHODOLOGY

This study adopts the Kitchenham Framework to conduct an SLR [22], [23]. The framework consists of three main phases: planning the review, conducting the review, and reporting the review. The details of each stage in the Kitchenham methodology are illustrated in Figure 1.

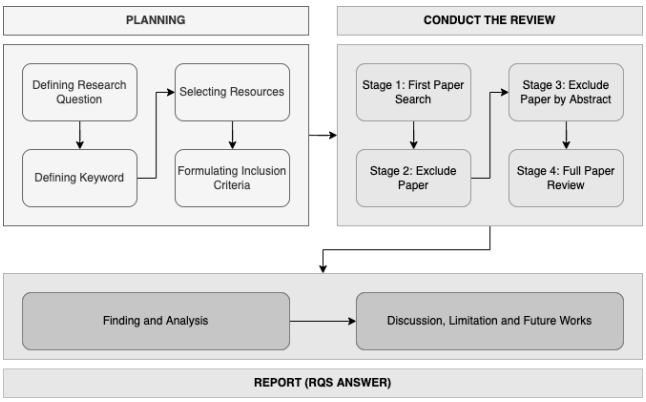



**Figure 1** Systematic Literature Review Research Design

The first phase, planning the review, involves formulating research questions aligned with the issues outlined in the background and the research objectives. This phase includes determining the keywords for identifying relevant literature, selecting digital libraries as sources, and defining inclusion criteria.

The second phase, conducting the review, is carried out step by step. The process begins with searching for literature using the predefined keywords in all selected digital libraries. Next, the excluded paper (stage 2) is conducted, screening literature based on titles and removing duplicates. The remaining literature is evaluated for relevance through abstract (stage 3) and complete paper analysis (stage 4). Finally, all selected literature undergoes an in-depth content review to ensure its suitability for answering the research questions. The results of this phase are extracted and synthesized to provide answers to the research questions.

The final phase, reporting the review, analyzes the findings to address the research questions. This phase also identifies the study's limitations and discusses opportunities for future research.

Concerning the three phases in the SLR framework, only the planning phase is explained in detail in this section. Whereas the other two phases will be elaborated further in the results and discussion section.

### 2.1 Research Question

Based on the research objectives mentioned in Section 1.0, the following research questions are formulated:

- RQ1: Which key evaluation frameworks have been previously employed in User Story research?
- RQ2: Which methods have been used to implement User Story evaluation frameworks in research?
- RQ3: What is the typical number of User Stories analyzed in recent studies evaluating User Story quality?
- RQ4: What models or tools are utilized to evaluate User Stories based on established evaluation frameworks?

### 2.2 Keywords and Resources

This study employs standardized keywords across all digital libraries used, specifically ("User Story" OR "User Stories") AND ("quality" OR "Evaluation"). The selection of these keywords aims to ensure consistency in searches and maximize the

relevance of the retrieved literature. Literature sources were drawn from leading digital libraries, including IEEE Xplore, Scopus, Google Scholar, Science Direct, and SpringerLink, which are widely recognized as reliable academic research databases.

Additionally, specific configurations were applied to each digital library to refine the search results. The primary focus was on research areas such as software engineering, agile software development, and requirement engineering (RE) to ensure that the collected literature aligns with the research context. This approach is expected to support a comprehensive and focused literature review.

### 2.3 Inclusion Criteria

The literature selection process in this study followed a systematic three-stage approach, with clearly defined inclusion criteria for each phase. In Stage 2 (title screening), the inclusion criteria required that the title explicitly contain the terms User Story or User Stories. Additionally, the paper must be written in English, peer-reviewed, and published between 2020 and 2024.

In Stage 3 (abstract and keyword screening), the inclusion criteria specified that the keywords must include User Story or User Stories alongside quality. Moreover, the abstract must explicitly address aspects related to User Story quality to ensure its relevance to the study's objectives.

In Stage 4 (content screening), the inclusion criteria were refined to encompass user stories, quality, evaluation, assessment, and/or measurement content. Further, the content must also detail the evaluation of US quality, whether as the central focus or as part of a broader research context.

This structured, multi-phase selection process was meticulously designed to ensure that the chosen literature is highly relevant and academically rigorous, comprehensively supporting the research's objectives.

## 3.0 RESULTS AND DISCUSSION

This section discusses the findings obtained during the Conduct the Review and Report phases. The detailed results of each phase are explained in the following sub-sections.

### 3.1 Conduct the Review

In the first stage, a search was conducted across all selected digital libraries using the exact keywords. This search resulted in a total of 2,682 pieces of literature. The number of search results obtained from each digital library is presented in Table 1. The left column represents the source of the literature, while the right column represents the number of literature identified for each source.

**Table 1** First Paper Search Result

| Source | Result (Paper) |
|---|---|
| Google Scholar | 828 |
| IEEE Xplore | 276 |
| Science Direct | 395 |
| Scopus | 193 |
| SpringerLink | 990 |
| **Total** | **2682** |

It can be observed that the majority of search results were retrieved from SpringerLink and Google Scholar.

Subsequently, literature selection was performed based on the inclusion criteria, focusing on their titles. At this stage, all papers with titles containing the term User Story or User Stories were selected for further processing. To avoid redundancy, duplicate papers with identical titles across multiple digital libraries were excluded from the list. These duplications commonly occurred between Scopus and Google Scholar, as well as between Scopus and SpringerLink. Consequently, the results of this stage are presented in Table 2. The left column represents the source of the literature. The subsequent columns represent the number of relevant, duplicates, and excluded, and the sum of identified literature in each source, respectively.

**Table 2** Title and Duplicate Screening

| Source | Relevant (Paper) | Duplicate (Paper) | Exclude (Paper) | Total (Paper) |
|---|---|---|---|---|
| Google Scholar | 512 | 46 | 270 | 828 |
| IEEE Xplore | 78 | | 198 | 276 |
| Science Direct | 38 | 2 | 355 | 395 |
| Scopus | 112 | 79 | 2 | 193 |
| SpringerLink | 11 | | 979 | 990 |
| **Total** | **751** | **127** | **1804** | **2682** |

Apparently, it was observed that the search using the specified keywords still produced irrelevant literature, with this issue being most prevalent in SpringerLink.

Next, the following selection stage involved screening the abstracts and keywords of the literature that passed the initial selection. At this stage, filtering was conducted by carefully examining the content of the abstracts and keywords to ensure their relevance to the research topic. The two primary keywords focused on in this stage were user story/user stories and quality. If a piece of literature explicitly mentioned both terms in its abstract or keywords, it was selected for further analysis. Conversely, literature that did not include both terms was excluded from the list. The result of this phase can be seen in Table 3. The first column displays the digital library of literature sources, the second column is the selected paper after screening the abstract and keywords, the excluded is the literature that is excluded because it does not match the research target, and the last column is the total number of literature selected at this stage.

**Table 3** Abstract and Keyword Screening

| Source | Include (Paper) | Exclude (Paper) | Total (Paper) |
|---|---|---|---|
| Google Scholar | 2 | 510 | 512 |
| IEEE Xplore | 26 | 52 | 78 |
| Science Direct | 4 | 34 | 38 |
| Scopus | 31 | 81 | 112 |
| SpringerLink | 7 | 4 | 11 |
| **Total** | **70** | **681** | **751** |

Accordingly, it was observed that search keywords generated a significant amount of irrelevant literature, particularly in Google Scholar.

Finally, the final stage was the full paper review. At this stage, the literature that did not explicitly discuss the evaluation of US quality in its content was excluded. Although some literature mentioned terms such as '*quality evaluation,*' it was also removed from the list if it lacked significants explanation on implementing such evaluations. This process ensured that only the literature offering in-depth and relevant information on US quality evaluation was considered in the final analysis. The results of this stage are presented in Table 4.

**Table 4** Content Screening

| Source | Include (Paper) | Exclude (Paper) | Total (Paper) |
|---|---|---|---|
| Google Scholar | 1 | 1 | 2 |
| IEEE Xplore | 8 | 18 | 26 |
| Science Direct | 2 | 2 | 4 |
| Scopus | 11 | 22 | 31 |
| SpringerLink | 6 | 1 | 7 |
| **Total** | **26** | **44** | **70** |

Eventually, the final results of the review process indicate that only 26 studies were selected to address the research questions from an initial pool of 2,682. This represents just 0.96% of the literature meeting the specified keyword-based selection criteria. Most of the selected literature originated from Scopus, followed by IEEE Xplore, SpringerLink, Science Direct, and Google Scholar.

Further results revealed that Scopus contributed the most significant amount of selected literature (11), the second was initially sourced from IEEE Xplore (8), and the next was followed by SpringerLink (6). Regarding literature types, 9 selected works were journal articles, and the remaining 17 were conference proceedings, reflecting the diversity of sources included in the final analysis.

### 3.2 Report The Findings

This subsection addresses the research questions formulated during the planning phase based on the literature selected through the *conduct of the review* process. Each research question is answered by referring to the relevant findings from the selected literature. Detailed explanations for addressing each research question are presented in the following paragraphs.

RQ1: Which key evaluation frameworks were previously employed in user story research?

This research question focuses on the frameworks commonly used by researchers to evaluate the quality of the US. Based on the analysis of the 26 selected literature, six primary frameworks were identified: INVEST, Quality User Story (QUS), ISO 25010, IEEE 830 Quality Criteria, User Story Quality Measurement (USQM), and Linguistic-based (semantic and semiotic-based). The detailed utilization of each framework by the selected literature is presented in Table 5.

**Table 5** User Story Quality Evaluation Frameworks

| No | Framework | Literature | References |
|---|---|---|---|
| 1 | INVEST [14], [15] | 11 | [4], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33] |
| 2 | Quality User Story (QUS)[34], [35] | 10 | [18], [33], [36], [37], [38], [39], [40], [41], [42], [43] |
| 3 | Linguistic (Semantic and Semiotic) | 3 | [17], [44], [45] |
| 4 | IEEE 830 [46] | 1 | [47] |
| 5 | ISO 25010 [48] | 1 | [49] |
| 6 | User Story Quality Measurement (USQM) [50] | 1 | [51] |

The frameworks used in studies evaluating US quality encompass various models with unique approaches. One such framework is INVEST, which consists of six criteria: Independent, Negotiable, Valuable, Estimable, Small, and Testable. This framework was first introduced in Extreme Programming [14] and later refined into the INVEST Grid [15]. According to the selected literature, this model was fully utilized in four studies [28], [30], [31], [32]. In addition to its standard application, some studies have modified the INVEST criteria by incorporating factors such as formal, lexical, semantic, and saturation quality [24]. Furthermore, other studies grouped INVEST criteria and added supplementary factors [29], focusing on quality evaluation based on individual and overall user requirements. These findings demonstrate that the INVEST framework can be adopted, modified, and further developed.

Similarly, the Quality User Story (QUS) framework was first introduced by Lucassen et al. [30] and employs 14 criteria for evaluating US quality. These criteria are categorized into three linguistic groups: Syntactic (atomic, minimal, well-formed), Semantic (conflict-free, conceptually sound, problem-oriented, unambiguous), and Pragmatic (complete, explicit dependencies, full sentence, independent, scalable, uniform, unique). Importantly, this framework also serves as the foundation for developing an evaluation tool called the Automatic Quality User Story Artisan (AQUSA) [35]. Based on the selected literature, 10 studies used QUS to evaluate US quality. Notably, four of these studies employed the AQUSA tool [18], [37], [39], [40], while others used only a subset of QUS criteria. In addition, some studies even developed new tools using specific QUS criteria as references [43].

Moreover, ISO 25010 is a quality standard developed for computer systems, software products, data, IT services, and quality-in-use [48], [52]. Building on this framework, [49] employed the ISO 25010 product quality model to evaluate the results of US extraction. This study utilized Deep Learning and NLP technologies to identify quality issues in the US, categorized under usability, performance, security, reliability, and compatibility labels.

Similarly, the IEEE 830 [46], [53] framework was introduced as a Software Requirements Specification (SRS) standard. It outlines characteristics such as correctness, unambiguity, completeness, consistency, verifiability, traceability, modifiability, understandability, feasibility, and stability. In applying this standard, Kuhail and Lauesen [43] evaluated US quality, focusing on completeness, correctness, verifiability, and

traceability criteria. Their study argued that specific IEEE 830 criteria overlap with those of the INVEST and QUS frameworks.

In contrast to these established frameworks, the Model of Determining the User Stories Quality (MDUSQ) framework, introduced by Jharko [47], employs three main factors for evaluating US quality: basic quality, quality of management, and acceptance (confirmed) quality. These factors include clarity of documents, low complexity, controllability, manageability, testability, and verifiability. Although newly introduced, this framework appears to refer to the USQM framework, previously discussed by Lai [46].

On the other hand, a linguistic approach is also utilized for evaluating the US, as their textual structure allows for analysis based on syntactic, semantic, and pragmatic dimensions, as implemented in the QUS framework. Expanding on this, Kamthan et al. [41] introduced a semiotic approach to US evaluation by adding the dimension of social quality. However, this approach has not been widely adopted, and to date, scientific publications employing this framework remain limited, indicating that it is still in the developmental stage. Furthermore, the important criteria or attributes in the US quality evaluation can be identified from all these frameworks. The criteria are determined by looking at the intersection between frameworks. Through this approach, 8 criteria were identified as important for user story evaluation. The criteria/attributes are presented in Table 6.

**Table 6** Important Criteria

| No | Criteria | Frameworks |
|---|---|---|
| 1 | Independent | INVEST, QUS |
| 2 | Unambiguous | QUS, Linguistic, IEEE 830 |
| 3 | Complete | QUS, IEEE 830 |
| 4 | Estimable | INVEST, QUS, USQM |
| 5 | Testable | INVET, QUS, IEEE 830 |
| 6 | Conflict Free | QUS, Linguistic |
| 7 | Atomic | QUS, USQM |
| 8 | Negotiable | INVEST, USQM |

Table 6 in the framework column illustrates the number of frameworks that utilize these criteria. While some criteria may not be explicitly identical across all frameworks, many are textually different yet conceptually similar. For instance, the criterion "unambiguous" aligns closely with the concept of "understandability" in IEEE 830, and "testable" relates to "verifiability" in the same standard. Additionally, QUS encompasses nearly all the criteria found in the various frameworks. Furthermore, QUS categorizes these criteria into Individual and Set, allowing for a measurement reference based on these categories.

RQ2: Which methods have been used to implement user story evaluation frameworks in research?

Various methods are employed to implement the frameworks discussed in RQ1. These methods include: (1) hiring experts, practitioners, students, or specialists experienced in User Story or Agile Software Development (ASD) to evaluate and complete the framework criteria; (2) utilizing existing User Story quality evaluation tools, including AI-based tools; and (3) developing models or tools to apply the frameworks, such as AI-based models. The criteria of the methods used to implement these frameworks are presented in Table 7.

**Table 7** Framework Application Methods

| No | Method | Criteria/Tools | Reference |
|---|---|---|---|
| 1 | Hiring Practitioners | At least 2 years' and maximum 15 years' experience in software engineering | [27], [28], [30], [36], [41], [47] |
| | Hiring Academia/ Researcher | Professor in Computer Science, Author/Researcher | [29], [32], [36] |
| | Hiring Student | 3th year students, pass SE Course | [38] |
| 2 | Existing Tools | QUS Based, INVEST Based, Generative-AI (ChatGPT) | [18], [37], [39], [40] |
| 3 | Model Dev/ AI Based Model | iStar, USQA, MDUSQ, User Requirement Quality Assessment Framework, AI (Deep Learning, Machine Learning, NLP) | [17], [25], [31], [33], [42], [43], [44], [49], [51] |

Evaluating the quality of the US often involves experts or practitioners providing in-depth assessments. Frameworks such as INVEST and QUS were employed in nine of the 26 reviewed literature. In these approaches, the number of invited experts varied, with consideration given to their experience and roles in software development. Several studies emphasized the importance of practitioner experience, requiring a minimum of two years in software engineering [27], [30], [36], [41]. In addition to practitioners, academics such as professors and researchers in software engineering, particularly those specializing in agile software development, were also involved in evaluating the US using specific frameworks [29], [32], [36]. This expertise-driven approach highlights the critical role of evaluator experience and background in ensuring the quality of the evaluation results.

Beyond expertise-based approaches, US evaluation can also leverage pre-existing tools. Several tools mentioned in the studies include AQUSA [35] and ChatGPT [39]. Among the 26 studies reviewed, four utilized AQUSA directly or in comparison with other tools [39]. Interestingly, Generative AI, such as ChatGPT, has been applied to evaluate US quality based on specific standards. Research findings indicate that evaluations using ChatGPT align with results produced by practitioners, although there is room for improvement in the consistency of the evaluations. This demonstrates the significant potential of Generative AI in supporting automated and efficient evaluation processes.

On the other hand, new evaluation models that utilize AI-based technologies, such as machine learning, deep learning, and generative AI, have been developed. Algorithms employed include Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and NLP-based models such as Bidirectional Encoder Representations from Transformers( BERT), Optimized BERT (RoBERTa and DistilBERT), XLNet, and Large Language Models (LLMs). Several evaluation models have been introduced, including iStar [17], User Story Quality Assessment (USQA) [43], the User Requirement Quality Assessment Framework [25], and the MDUSQ [51]. Other AI-based models, particularly those using Deep Learning and LLMs, have also been proposed by several studies [17], [33], [42], [44]. These studies demonstrate significant efforts in developing automated evaluation models that can accelerate the process and improve the accuracy of US quality evaluations.

RQ3: What is the number of user stories typically analyzed in recent studies evaluating user story quality?

In the literature on US quality evaluation, not all studies explicitly specify the number of analyzed US. Some do not directly mention the quantity, while others are still in the model development phase and have yet to be tested. However, most literature provides clear information on the number and sources of US used.

Among the 26 reviewed studies, 17 explicitly reported the number of US, ranging from 4 to 9,060. When categorized by data size—small (1–100), medium (101–500), large (501–1,000), and very large (1,001–10,000)—the distribution is as follows: 9 studies fall under small, four under medium, two under large, and two under very large. This indicates that most studies use a small 1–100 US dataset. The detailed distribution of US numbers can be seen in Figure 2.

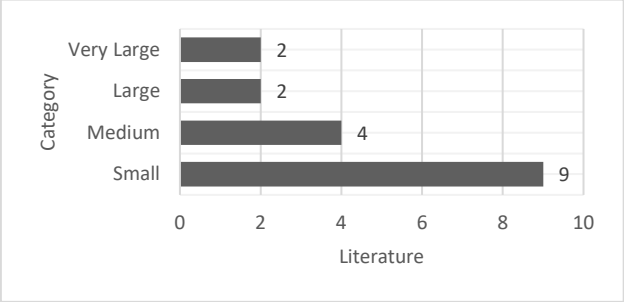



**Figure 2** User Story Number in Selected Literature

The sources of the US are diverse, including those obtained from case studies (12) conducted through workshops, public data sources (3), and the US generated using Generative AI (1). And 10 studies did not mention where their US data came from (10). The details of US sources can be seen in Figure 3.

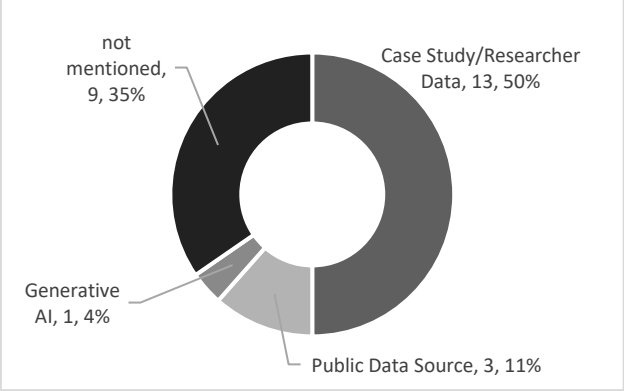



**Figure 3** User Story Sources

These findings suggest that most studies rely on a small number of US (1–100), sourced from workshops, public data, company datasets, or Generative AI. This diversity highlights the flexibility in data collection to support various approaches to evaluating US quality.

RQ4: What models or tools are utilized to evaluate user stories based on the established evaluation frameworks?

Based on the selected literature, the most commonly used frameworks for evaluating the quality of the US during the 2020–2024 period were INVEST (11 studies) and QUS (10 studies). Both frameworks remained relevant and consistently applied through 2024. However, in some instances, researchers adapted and modified these frameworks to better align with specific research needs or practical applications. This adaptability highlights the flexibility of INVEST and QUS in supporting US quality evaluations.

In addition to frameworks, several tools have been developed based on the QUS standard, including AQUSA, USQA, MDUSQ, iStar, and others like ChatGPT. Among these tools, AQUSA is the most widely used tool in prior studies. Four of the 10 studies utilizing the QUS framework employed AQUSA as an evaluation tool. AQUSA is widely used as a pioneering automatic tool for evaluating US quality [54]. Although many other models have been developed [42], [43], [51], AQUSA remains in use even after the last five years.

These findings reinforce the significance of INVEST and QUS as primary frameworks for US evaluation while emphasizing the role of QUS-based tools, such as AQUSA, in facilitating more practical and standardized evaluation processes.

### 3.3 Discussion

Based on the studies conducted, it was observed that INVEST and QUS are the most widely used frameworks for evaluating the quality of the US. In addition to these two frameworks, researchers have employed others such as ISO 25010, IEEE 830, and USQM. However, their application is relatively limited in the context of US evaluation. This is primarily due to the broader scope of standards like ISO 25010 and IEEE 830, designed to evaluate the quality of the whole RE. As a smaller subset of RE activities, the US focuses more specifically on the requirement elicitation process, making these frameworks less relevant for particular US-specific needs. Although INVEST is recognized as one of the pioneering evaluation frameworks in the US, QUS remains a popular choice among researchers. QUS provides more complex criteria and assesses quality from both individual and set perspectives, enabling a more comprehensive analysis. Its application is also more flexible, as some studies may select only specific criteria for their research [18], [39]. Moreover, it is also widely used with AI, such as machine learning and generative AI [36].

On the other hand, the trend of utilizing AI over the past five years shows a significant increase in supporting activities related to the US. Technologies such as Generative AI, Deep Learning, and Machine Learning have been applied for various purposes, including creation, conflict resolution, and quality evaluation in the US. Despite its potential, generative AI, currently at the forefront of AI technology, has had limited application in this context. Among the reviewed literature, only one study employed generative AI, specifically ChatGPT, to evaluate the quality of the US using the QUS framework [39].

This study found that the results of US quality evaluations using ChatGPT were consistent with those conducted by human evaluators. However, inconsistencies in output were noted, leading to recommendations for performing at least three evaluation iterations to enhance accuracy. Furthermore, there is substantial potential for leveraging generative AI in US quality evaluation research. The study also emphasized the critical role of human oversight in re-evaluating ChatGPT's results, ensuring that raw outputs are not used directly without further validation.

Additionally, the review highlighted variations in the number of the US dataset analyzed across the 26 literature. The number

of US used as research objects varied widely, reflecting flexibility in adapting data to research objectives. This variation enables researchers to focus on developing evaluation models or conducting specific case studies in software development.

Consequently, these findings underscore the importance of diversifying methods and tools for evaluating the quality of the US while highlighting significant opportunities for further advancements, particularly in applying AI technologies to support these activities.

## 4.0 CONCLUSION

This SLR addressed questions regarding evaluating US quality, encompassing frameworks, application methods for employing these frameworks, tools, and the number of US data sources utilized. The review ultimately identified 26 relevant literature to address the research questions. Utilizing the Kitchenham method, this study formulated four research questions to be explored.

In RQ1, related to the US quality framework has advanced significantly, with INVEST and QUS emerging as the dominant frameworks. These frameworks, consistently applied over the past five years, demonstrate their relevance and flexibility in addressing diverse quality evaluation needs. While other frameworks like ISO 25010 and IEEE 830 are utilized, their broader scope limits their applicability to the specific challenges of US evaluation, which often centers on the requirement elicitation process.

In line with that, RQ2 is related to implementing the frameworks. The practical implementation of these frameworks frequently involves human evaluators, such as experts, practitioners, and academics, whose expertise ensures nuanced assessments. Increasingly, AI technologies, including machine learning, deep learning, NLP, and Generative AI, are being adopted to complement these efforts. Furthermore, RQ3 emphasizes the variety of quality studies conducted on User Stories. According to the collected literature, the most commonly utilized number of User Stories ranges from four to one hundred. The sources for these studies are diverse, encompassing research-generated data and publicly available information, with some even being produced by Generative AI. This finding illustrates the flexibility in evaluating the quality of User Stories, particularly concerning their quantity and sources. The application of A) is pivotal in advancing NLP-based quality evaluation tools. One notable automated tool that has gained significant traction among researchers is AQUSA, which is utilized to assess US quality. This instrument effectively addresses the inquiry posed in RQ4.

In the future, generative AI, in particular, offers immense potential for advancing US evaluation. Despite limited adoption—only one study [39] has utilized ChatGPT—findings indicate that its results align closely with human evaluations, though with some inconsistencies. These observations underscore the importance of iterative evaluation and human validation. Generative AI presents a promising avenue for innovation in RE, especially in the requirement elicitation phase, where its capabilities could transform traditional methods. With further refinement, AI can bridge gaps in efficiency and accuracy, paving the way for a new era in US quality evaluation.

## Acknowledgment

Politeknik Caltex Riau fully supports this research through the employee PhD program. The authors fully acknowledged Politeknik Caltex Riau and Universiti Tun Hussein Onn Malaysia for the approved program, which makes this critical research viable and effective.

## Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

## References

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

[2] C. O’hEocha and K. Conboy, 2010 “The Role of the User Story Agile Practice in Innovation,” in *International Conference on Lean Enterprise Software and Systems,* Oct., 20–30. DOI: 10.1007/978-3-642-16416-3_3.

[3] Rupali M. Chopade and Nikhil S. Dhavase, 2017 “Agile Software Development: Positive and Negative User Stories,” in *2017 2nd International Conference for Convergence in Technology (I2CT)*, Apr., 297-299. DOI: 10.1109/I2CT.2017.8226139

[4] A. Ananjeva, J. S. Persson, and A. Bruun, 2020 “Integrating UX work with agile development through user stories: An action research study in a small software company,” *Journal of Systems and Software*, 170: 2-10. DOI: 10.1016/j.jss.2020.110785

[5] F. Dalpiaz, I. van der Schalk, S. Brinkkemper, F. B. Aydemir, and G. Lucassen, 2019 “Detecting terminological ambiguity in user stories: Tool and experimentation,” *Journal of Information and Software Technology*, 110: 3–16. DOI: 10.1016/j.infsof.2018.12.007.

[6] [6] A. R. Amna and G. Poels, 2022 “Ambiguity in user stories: A systematic literature review,” *Journal of Information and Software Technology,* 145: 1-14. Elsevier B.V., DOI: 10.1016/j.infsof.2022.106824.

[7] [7] C. A. Peláez and A. Solano, 2024 “A practice for specifying user stories in multimedia system design: An approach to reduce ambiguity.*” Interaction Design and Architecture(s) Journal*, 60: 214–236. DOI: 10.55612/s-5002-060-009.

[8] S. Nasiri, Y. Rhazali, M. Lahmer, and N. Chenfour, 2020 “Towards a Generation of Class Diagram from User Stories in Agile Methods,” in *Procedia Computer Science,* 170: 831–837 Elsevier B.V,. DOI: 10.1016/j.procs.2020.03.148.

[9] M. Urbieta, L. Antonelli, G. Rossi, and J. C. S. do Prado Leite, 2020 “The impact of using a domain language for an agile requirement management,” *Journal of Information and Software Technology*, 127: 1-16. DOI: 10.1016/j.infsof.2020.106375.

[10] M. I. Zul, S. M. Yasin, and D. S. S. Sahid, 2024 “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., 136–141. DOI: 10.1109/IConEEI64414.2024.10748069.

[11] [11] H. Gardner, A. F. Blackwell, and L. Church, 2020 “The patterns of user experience for sticky-note diagrams in software requirements workshops,” in *Journal of Computer Languages*, 61: 1-9. Elsevier Ltd., DOI: 10.1016/j.cola.2020.100997.

[12] M. Trkman, J. Mendling, P. Trkman, and M. Krisper, 2019 “Impact of the conceptual model’s representation format on identifying and understanding user stories,” *Journal of Information and Software Technology*, 116: 1-17. DOI: 10.1016/j.infsof.2019.08.001.

[13] Y. Wautelet, S. Heng, M. Kolp, and I. Mirbel, 2014 “Unifying and Extending User Story Models,” in *CAiSE 2014: Advanced Information Systems Engineering*. 211–225. DOI: 10.1007/978-3-319-07881-6_15.

[14] B. Wake, 2025 "INVEST in Good Stories, and SMART Tasks," XP123: Exploring Extreme Programming. https://xp123.com/invest-in-good-stories-and-smart-tasks/. Retrieved date: 27 January 2025

[15] L. Buglione and A. Abran, 2013 "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, Oct., 49–53. DOI: 10.1109/IWSM-Mensura.2013.18.

[16] T. Tamai and M. I. Kamata, 2009 "Impact of Requirements Quality on Project Success or Failure," in *Design Requirements Engineering: A Ten-Year Perspective. Lecture Notes in Business Information Processing*, 14: 258–275 Springer, Berlin, Heidelberg,. DOI: 10.1007/978-3-540-92966-6_15.

[17] T. Wang, C. Li, C. Wang, T. Li, and Y. Zhai, 2023 "A Deep Learning-Based Method for Identifying User Story Semantic Conflicts," in *2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*, IEEE, Oct., 220–229. DOI: 10.1109/QRS-C60940.2023.00063.

[18] G. B. Herwanto, G. Quirchmayr, and A. M. Tjoa, 2024 "Leveraging NLP Techniques for Privacy Requirements Engineering in User Stories," in *IEEE Access*, 12: 22167–22189. DOI: 10.1109/ACCESS.2024.3364533.

[19] A. R. Amna and G. Poels, 2022 "Systematic Literature Mapping of User Story Research," in *IEEE Access,* IEEE*,* 10: 51723-51746. DOI: 10.1109/ACCESS.2022.3173745.

[20] A. Hendriana, T. Raharjo, and A. Nurfitriani, 2024 "Approaches in Determining User Story Quality through Requirement Elicitation: A Systematic Literature Review," *Indonesian Journal of Computer Science*, vol. 12(6): 3599-3614. DOI: 10.33022/ijcs.v12i6.3639.

[21] I. K. Raharjana, D. Siahaan, and C. Fatichah, 2021 "User Stories and Natural Language Processing: A Systematic Literature Review," *IEEE Access*, 9: 53811–53826. DOI: 10.1109/ACCESS.2021.3070606.

[22] B. Kitchenham, 2024 "Procedures for Performing Systematic Reviews," Keele University Technical Report TR/SE-0401, July.

[23] B. Kitchenham and P. Brereton, 2013 "A systematic review of systematic review process research in software engineering," *Journal of Information and Software Technology*, 55(12): 2049–2075. DOI: 10.1016/j.infsof.2013.07.010.

[24] D. Hallmann, 2020 "'I Don't Understand!': Toward a Model to Evaluate the Role of User Story Quality," in *21st International Conference on Agile Software Development, XP 2020,* Springer, Cham, Jun., 103–112. DOI: 10.1007/978-3-030-49392-9_7.

[25] X. Xu, Y. Dou, L. Qian, Z. Zhang, Y. Ma, and Y. Tan, 2023 "A Requirement Quality Assessment Method Based on User Stories," *Electronics (Basel)*, 2(10): 2155-2171. DOI: 10.3390/electronics12102155.

[26] P. Pokharel and P. Vaidya, 2020 "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*, IEEE, Oct,. 1-5. DOI: 10.1109/ICDABI51230.2020.9325670.

[27] S. S. do Nascimento, J. M. Abe, L. R. Forçan, C. C. de Oliveira, K. Nakamatsu, and A. Ari, 2022 "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*, 332: 133–142. Springer, Singapore, DOI: 10.1007/978-981-19-7842-5_12.

[28] Z. Zhang, M. Rayhan, T. Herda, M. Goisauf, and P. Abrahamsson, 2024 "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., 117–126. DOI: 10.1007/978-3-031-61154-4_8.

[29] X. Xu, Y. Dou, L. Qian, J. Jiang, K. Yang, and Y. Tan, 2023 "Quality improvement method for high-end equipment's functional requirements based on user stories," *Advanced Engineering Informatics*, 56, DOI: 10.1016/j.aei.2023.102017.

[30] Y. Li, J. Keung, Z. Yang, X. Ma, J. Zhang, and S. Liu, 2024 "SimAC: simulating agile collaboration to generate acceptance criteria in user story elaboration," *Automated Software Engineering*, 31(2): 55. DOI: 10.1007/s10515-024-00448-7.

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

[32] A. Brockenbrough and D. Salinas, 2024 "Using Generative AI to Create User Stories in the Software Engineering Classroom," in *2024 36th International Conference on Software Engineering Education and Training (CSEET)*, 1–5. DOI: 10.1109/CSEET62301.2024.10662994.

[33] A. Ferreira, A. Rodrigues da Silva, and A. Paiva, 2022 "Towards the Art of Writing Agile Requirements with User Stories, Acceptance Criteria, and Related Constructs," in *Proceedings of the 17th International Conference on Evaluation of Novel Approaches to Software Engineering*, SCITEPRESS - Science and Technology Publications, 477–484. DOI: 10.5220/0011082000003176.

[34] G. Lucassen, F. Dalpiaz, J. M. E. M. van der Werf, and S. Brinkkemper, 2015 "Forging high-quality User Stories: Towards a discipline for Agile Requirements," in *2015 IEEE 23rd International Requirements Engineering Conference (RE)*, IEEE, Aug., 126–135. DOI: 10.1109/RE.2015.7320415.

[35] G. Lucassen, F. Dalpiaz, J. M. E. M. van der Werf, and S. Brinkkemper, 2016 "Improving agile requirements: the Quality User Story framework and tool," *Requirement Engineering*, 21(3): 383–403. DOI: 10.1007/s00766-016-0250-x.

[36] O. Abed, K. Nebe, and A. B. Abdellatif, 2024 "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*, 2120: 3–13. Springer, Cham. DOI: 10.1007/978-3-031-62110-9_1.

[37] E. Scott, T. Tõemets, and D. Pfahl, 2021 "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*, 404: 119–138. Springer, Cham, DOI: 10.1007/978-3-030-65854-0_10.

[38] E. Trisnawati, I. K. Raharjana, T. Taufik, A. H. Basori, A. B. F. Mansur, and N. A. Alghanmi, 2024 "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*, 10(1): 110–125. DOI: 10.20473/jisebi.10.1.110-125.

[39] K. Ronanki, B. Cabrero-Daniel, and C. Berger, 2024 "ChatGPT as a Tool for User Story Quality Evaluation: Trustworthy Out of the Box?," in *Agile Processes in Software Engineering and Extreme Programming - Workshops*, Springer, Cham, ch. AI-assisted Agile, Dec., 173–181. DOI: 10.1007/978-3-031-48550-3_17.

[40] S. N. F. N. B. Mustaffa, J. Bin Sallim, and R. B. Mohamed, 2021 "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., 295–300. DOI: 10.1109/ICSECS52883.2021.00060.

[41] C. Ankora and A. D., "Integrating User Stories in the Design of Augmented Reality Application," *International Journal of Information Technologies and Systems Approach*, 15(1): 1–19. DOI: 10.4018/IJITSA.304809.

[42] T. Wang, C. Wang, T. Li, Z. Liu, and Y. Zhai, 2022 "User Story Quality Assessment Based on Multi-dimensional Perspective: A Preliminary Framework," in *CEUR Workshop Proceedings: 15th International iStar Workshop*, Hyderabad, India: ceur-ws.org, Oct., 7–13.

[43] S. Jiménez, A. Alanis, C. Beltrán, R. Juárez-Ramírez, A. Ramírez-Noriega, and C. Tona, 2023 "USQA: A User Story Quality Analyzer prototype for supporting software engineering students," *Computer Applications in Engineering Education*, 31(4): 1014–1024. DOI: 10.1002/cae.22620.

[44] Z. Xuan, T. Wang, C. Wang, and T. Li, 2024 "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)*, 484–487. DOI: 10.1109/RE59067.2024.00057.

[45] P. Kamthan and N. Shahmir, 2020 "A Framework for the Semiotic Quality of User Stories," in *27th International Conference on Systems Engineering, ICSEng 2020*, Las Vegas, USA: Springer Nature, Dec., 413–422. DOI: 10.1007/978-3-030-65796-3_40.

[46] EEE, 1998 "IEEE Recommended Practice for Software Requirements Specifications," Jun., *IEEE, Piscataway, NJ, USA*. doi: 10.1109/IEEESTD.1998.88286.

[47] M. A. Kuhail and S. Lauesen, 2022 "User Story Quality in Practice: A Case Study," *Software*, 1(3): 223–243. DOI: 10.3390/software1030010.

[48] ISO, 2025 "Systems and software Quality Requirements and Evaluation (SQuaRE) — Product quality model," 2023, *https://www.iso.org/standard/78176.html*. Retrieved Date: 27 January 2025

[49] K. A. Alam, H. Asif, I. Inayat, and S.-U.-R. Khan, 2024 "Automated Quality Concerns Extraction from User Stories and Acceptance Criteria for Early Architectural Decisions," in *Software Architecture. ECSA 2024. Lecture Notes in Computer Science*, 14889: 359–367. Springer, Cham, DOI: 10.1007/978-3-031-70797-1_24.

[50] S.-T. Lai, 2017 "A User Story Quality Measurement Model for Reducing Agile Software Development Risk," *International Journal of Software Engineering & Applications*, 8(2): 75–86. DOI: 10.5121/ijsea.2017.8205.

[51] E. Jharko, 2024 "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)*, 658–663. DOI: 10.1109/ICIEAM60818.2024.10553873.

[52] A. Adewumi, S. Misra, and N. Omoregbe, 2015 "Evaluating Open Source Software Quality Models Against ISO 25010," in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, 872–877. DOI: 10.1109/CIT/IUCC/DASC/PICOM.2015.130.

[53] E. Stephen and E. Mit, 2020 "Evaluation of Software Requirement Specification Based on IEEE 830 Quality Properties," *International Journal on Advanced Science, Engineering and Information Technology*, 10(4): 1396–1402. DOI: 10.18517/ijaseit.10.4.10186.

[54] T. Tõemets, 2025 "Analysing the Quality of User Stories in Open Source Projects," PhD Thesis, University of Tartu, Estonia, 2020. Retrieved Date: Jan., [Online]. Available: https://hdl.handle.net/10062/93991