AI-Generated User Stories Supporting Human-Centred Development: An Investigation on Quality

Omed Abed, Karsten Nebe(✉), and Ahmed Belal Abdellatif

Hochschule Rhein-Waal, Friedrich-Heinrich-Allee 25, 47475 Kamp-Lintfort, Germany
{omed.abed,karsten.nebe,
ahmedbelal.abdellatif}@hochschule-rhein-waal.de

Abstract. User stories are an effective tool to document requirements in agile development, typically used to communicate user needs and technical details to developers in a non-technical language. As a result, poorly articulated stories often lead to misinterpretations, negatively affecting the development process as well as the quality of its outcome. Furthermore, the creation of user stories is often time-consuming, burdening project teams, where time is already scarce and tight schedules need to be met.

This study sets out to explore the effectiveness of various Artificial Intelligence (AI) tools in generating high-quality user stories, based on customer or user interview transcript as input to the tool. Our approach is to use selected AI tools to generate multiple user stories. To ensure consistency and comparability, the same input prompt is provided to all AI tools. The Quality User Stories framework is then used to evaluate the quality of the generated stories. The application of these criteria enables a comprehensive evaluation of the syntactic, semantic and pragmatic properties of the stories. By comparing the results for the different tools, we draw patterns relating to each of the AI tools as well as a comparison of their performance, in terms of accuracy in extracted insights from the data.

The significance of this research lies in exploring the extent to which AI can support the requirement specification process within the HCD lifecycle. AI tools could be embedded into the workflow, assisting practitioners in the analysis of transcripts to extract insights and create user stories. By having an assistive role in agile development, AI has the potential to save time and cost, enhancing overall efficiency and introducing more automation.

Keywords: Agile Development · Artificial Intelligence · Quality of User Stories

1 Introduction

Getting precise and meaningful requirements right has a major impact on the success of the product [1]. Rapidly evolving technologies and the increase of complexity while fulfilling customers' needs, often cause requirements to change and have impact on development practices in companies [2]. Responding to these challenges, development

C. Stephanidis et al. (Eds.): HCII 2024, CCIS 2120, pp. 3–13, 2024.
https://doi.org/10.1007/978-3-031-62110-9_1

strategies have changed and are gradually replaced with the more flexible agile methodologies [3–5]. Agile development is known for its flexibility and allows the adaptation of requirements, while involving the stakeholders along the process. Often, user requirements in agile development are described as part of user stories [2]. User stories are structured descriptions of requirements, written from the perspective of the user in a standardized format [6, 7]. These are created in collaboration with experts from diverse backgrounds to create a common understanding of technical and functional matters [8]. In general, it can be said that the process of user story generation increases the common understanding of end-user expectations and allows the developing team to define the resulting requirements [9]. However, documenting user stories is time-consuming. In practice, practitioners often miss allocating enough time for this to do it appropriately, consequently risking development failure, project delays, exceeded budgets, and customer dissatisfaction [2, 8]. Especially, the increasing complexity of today's products and resulting complexity of data collected during research can be challenging for practitioners to generate meaningful stories [10]. While looking for aiding software, we found that there is limited research on tools supporting generation of user stories [11], despite the gaining popularity of commercially available products utilizing Artificial Intelligence (AI). There is ongoing research on the possible use of AI in various fields, e.g. design and requirements engineering [10]. Rodeghero [6] developed a machine learning (ML) algorithm which is able to extract relevant information from meeting transcripts, for the generation of user stories. However, the process of generating user stories from the extracted data and their quality is not further discussed in their paper. Thus, there is a need to investigate tools, suitable for the generation of user stories, using transcripts as input to support the process. It is assumed that this can be achieved by using AI tools, however quality needs to be ensured. The following research questions are defined: Which AI-based tools exist, being capable to generate user stories from transcripts? What is the quality of user stories generated by these tools?

2 Related Work

In the following section, we provide an overview of the relevant research in the context of the paper. We look at the fundamental advantages of user stories, as well as their quality aspects, and discuss which relevant instruments exist to determine their quality. Afterwards, we investigate the current state of research on supporting tools for the creation of user stories, including those that are AI-based. Finally, we conclude and present the need for our work.

2.1 User Stories Advantages: Simple, Structured, Estimatable

Using user stories to capture requirements is widespread among practitioners, due to their simple predefined format which makes it easy to use [8]. User stories generally consist of three components: 1) text representing the user story itself, 2) the conversation between the stakeholders' exchanging views and ideas about the given user story, 3) the acceptance criteria that needs to be fulfilled and tested in order to ensure correctly met and delivered user requirements [12]."As a <role>, I want <goal>, so that

<benefit>" is a standard user story structure, popularized by M. Cohn [7]. Practitioners agree that applying user story templates increases the productivity and the quality of the deliverables [7]. Stakeholders derive requirements that define various aspects of a software project into smaller, comprehensible chunks, later documented in the form of user stories [7]. Furthermore, the granularity of user stories supports precise estimation of efforts and resources, contributing to project efficiency [13]. Practitioners share the opinion that the activity of documenting requirements in the form of user stories increases the common understanding of end-user expectations within the team [7, 14]. Whereas, inadequate communication and false understanding of requirements can lead to the development of undesired software and consequently, waste of resources [7]. It is known, that high-quality requirements can create reliable products, while addressing user needs to ensure user-friendly solutions [9].

2.2 Quality of User Stories

Despite the importance of requirements documentation, user stories are frequently poor in quality [3]. Oftentimes, stakeholders plan insufficient time to formulate requirements, consequently struggling with ambiguity and incompleteness, which causes extra effort managed by the development team later on [8, 15]. To prevent project delays, exceeded budgets, and customer dis-satisfaction or even development failure, which are common consequences of low-quality requirements, practitioners must ensure the quality of the requirements [9]. However, the review and analysis of user stories' quality is time-consuming [8]. To support this and to decrease human workload, Xiangqian Xu et al. [9] suggest the use of an automatic requirement quality assessment method, to improve the quality and the development efficiency. G. Lucassen et al. [3] introduced AQUSA - a tool to expose defects and deviations in user stories. For that purpose, the Quality User Story (QUS) framework is used, which provides a set of criteria assessing the quality of the user story's text, as basis for the assessment [7].

2.3 Quality Assessment Frameworks for User Stories

Even though automated objective assessment could increase the efficiency of the evaluation process, there is still a need for subjective reviews done by experts to ensure the quality and applicability of requirements [9]. One popular example for such is INVEST (Independent, Negotiable, Valuable, Estimatable, Small, Testable), which is used to evaluate the quality of the user stories [9]. Furthermore, sometimes it is also being used as a mean of training, by using it as a guideline while creating and evaluating user stories [7]. Nevertheless, INVEST guidelines are difficult to measure and do not address writing qualities, consequently Lucassen developed the QUS framework [16, 17]. The QUS framework provides 13 quality criteria for the review of user stories. It focuses on the content only and does not include any management aspects, such as effort estimations or acceptance criteria. Assessment criteria are clustered in three categories: Syntactic, Semantic, Pragmatic (see Table 1) [3].

The QUS framework defines the format of the user story (US) based on a standard template: role, means and ends. Role refers to a specific role within the user story. Means represent the requirement, including: 1. Subject: intent for e.g., want/is able to, 2. Action

Table 1. Quality criteria for the assessment of user stories according to Lucassen et al. [7]

Criteria	Description
Syntactic	considers the textual structure without its meaning
Atomic	A US concerns only one feature
Minimal	A US contains only necessary information: role, means, and optionally end
Well-formed	A US includes at least role and means
Semantic	qualifies the relations and meaning of elements in the user story
Conflict-free	A US should not be inconsistent with any other US, e.g. conflict of activities or resources
Conceptually sound	The means express a feature and the ends express a rationale, not anything else
Problem-oriented	A US specifies only the problem, not the solution to it
Unambiguous	A US avoids terms or abstractions that may lead to multiple implementations
Pragmatic	focuses on the audience's subjective interpretation of the user story text apart from the syntax and semantics
Complete	Implementing a set of user stories creates a feature-complete application, no steps are missing
Full sentence	A user story is a well-formed, grammatically correct, full sentence
Independent	The user story is self-contained, avoiding inherent dependencies on other user stories
Scalable	User stories do not denote too coarse-grained requirements that are difficult to plan and prioritize
Uniform	All user stories follow roughly the same template
Unique	Every user story is unique, duplicates are avoided

verb: express action linked to the feature being requested, 3. Direct object: on which the subject executes the action. Ends illuminate reasons for the means, provide additional information to clarify the means, describe dependencies with other functionality, and indicate qualitative requirements [3]. To assess if a criterion is fulfilled, Trisnawati [18] used the framework to evaluate user stories by marking met criteria with a "1" and not fulfilled ones with a "0".

2.4 Tools Utilizing AI, Supporting the Generation of User Stories

To gather user and customer needs, it is common in agile development to perform interviews and observations [6]. Gathered information subsequently must be further analyzed to become a source for user stories. AI-based solutions gain popularity in the domain of requirements engineering as "AI can observe, perceive and process exponentially faster and significantly better than humans" [10]. This is especially beneficial when to analyze huge amounts of data, i.e. having complex products [10].

Rodeghero et al. investigated the possibility to support practitioners in extracting relevant information from data collected during customer meetings by using trained models [6]. The model processes conversation transcripts between developers and users to extract roles, features, and motivations finally documented in the form of user stories. Rodeghero et al. also used the model on extractive summaries that he created based on the transcripts, and compared the outcomes from unprocessed and pre-processed data. However, the study does not explain how extracted information is further used to generate user stories and the quality they finally provide [6].

F. Dwitma and A. Rusli [19] developed a chatbot utilizing *Artificial Intelligence Markup Language* to aid various stakeholders in the process of generating user stories. The chatbot aims to support the formatting of user requirements into user stories by asking for specific information in the required order [19]. Whereas the study indicated the high acceptance of the chatbot, the quality of generated user stories was not discussed.

There is further investigation on the usage of AI techniques to analyze large amounts of data - e.g. app reviews, user forums, tweets, and issue tracking systems [10]. They aim to use crowd-user feedback to extract information that can be further used by requirement engineers [10]. Although these studies provide promising insights, there is no investigation on applying their techniques for the generation of user stories.

The review of related work reveals a notable gap in research with regard to the generation of user stories from transcripts, while considering quality criteria. Additionally, our investigation shows the lack of scientific research into the application of AI tools for the purpose of generating user stories. This gap presents an opportunity for contributions to the field by streamlining and enhancing the user story generation process using AI. This direction also aligns with the long-term vision articulated by Rodeghero et al. [6], which is the availability of a system that extracts user story relevant information from conversations with customers [6].

3 Investigation on the Quality of AI-Generated User Stories

In this study, we investigate the quality of user stories generated by AI tools, while using transcripts as input. To ensure the application of the designed process in the practical environment, we cooperated with a company. The company provided us with a recording of their user interview (in German language), which they originally used to derive requirements and create their user stories. We were provided with 18 user stories, based on their analysis, documented by the moderator of the interviews (practitioner).

For this study, the recordings were transcribed by using the AI-based tool: https://trint.com. The transcript was reviewed by the author to ensure its correctness and used as input to selected AI tools, to auto-generate user stories. Subsequently, we investigated on the quality of the stories, based on selected QUS criteria. Finally, the outcomes were compared with the user stories created by the company.

In the following sections, we first explain the basic considerations of the AI tool selections for this study, followed by the concepts of prompting being used, and finally the assessment of quality aspects in scope.

3.1 AI Tools Selection and Prompt Preparation

We followed a systematic approach to identify AI tools that can be used in the context of this study. First, we used existing catalogue sites, which list and categorize AI tools by various means and allow keyword search. We have chosen only the catalogue websites with huge datasets, of more than 2500 AI tools listed, because those sites are often used as search-entry for practitioners too. The sites we have chosen are: https://topai.tools/ (5565+), https://gpte.ai/ (5000+), https://theresanaiforthat.com/ (12000+) and https://www.futuretools.io/ (2606 AI tools). As filter criteria, we chose the keywords "user stories" and "user story". Moreover, we only considered results (AI tools) found during February 2024. We reviewed found results under the following criteria, if the AI tool is a) specifically made to generate user stories, b) allows larger size input in the form of a file upload or word input of at least 8000 words c) only AI tools that are directly available for use.

All selected AI tools were instructed with the same prompt. The prompt included information about: intended outcome, the product, transcript, and intended structure of the user story, i.e. standardized by Cohn [7]. The transcript was attached to the prompt as a text file or provided alongside the prompt, annotated as "interview transcript:"

3.2 Quality Assessment of Generated User Stories

The user stories were evaluated based on eight out of 13 criteria from the QUS framework. The criteria which determine quality, based on user stories within a "Set" were not considered, since the available interview does not describe the complete product, but rather new requests to the existing product. A set is defined as a group of complementary user stories describing functionalities of one feature. Consequently, completeness, conflict-freedom, uniqueness, uniformity and independence of the user story sets were not evaluated. The user stories were subsequently evaluated by an expert from the company. The evaluation focused on gaps in the content of the generated user stories. For this purpose, the expert evaluated the correctness of each user story based on the described role, mean and end, without considering their quality.

4 Procedure

4.1 AI Tool Selection

During the screening, we identified several AI tools dedicated to support creative writing because of the lacking context in which the term "user stories" is used. These false results were sorted out. At the end, 17 AI tools were identified based on the criteria mentioned in Sect. 3.1. Two of the identified tools were standalone products: *Pace AI* and *GeniePM*. Whereas, 15 of them were customized versions of *ChatGPT* (GPTs). GPTs are customized chatbots, which can be created by paid users, by providing specific information and instructions such as, guidelines for text generation. We decided to only include GPTs with conversation counts above 500: *PRD Maker*, *User Story Crafter* (USC) and *BOB the BA – User Story* (BOB). As the majority of the tools use *ChatGPT* as the basis, we decided to include *ChatGPT* itself, as well as, the second most popular AI tool, *Gemini Advanced* (previously called Google Bard) [20]. As a result, seven AI tools were selected for this study.

4.2 Evaluation of the User Stories

The selected tools were prompted to generate the user stories based on the provided transcript. The expected scope of the user story should consist only of the first component, defined by Bik et al. [12], namely the text representing the user story itself. All AI tools satisfied the requirements stated in the prompt, however, in the case of *GeniePM* and *PRD Maker*, not all the created user stories comply with the requested structure. We collected in total 70 user stories from the AI tools. The outcome from each tool was analyzed and evaluated based on the QUS criteria. The evaluation of user stories was conducted together by the authors and a practitioner from the company (i.e. the interviewee/generator of the 18 user stories). Every user story was individually reviewed. Upon encountering conflicting opinions, the evaluators presented their rationale to reach a unified conclusion.

The practitioner from the company evaluated the generated user stories regarding the precision of the extracted information about the roles, means, and ends from the transcript. The user stories were organized into three categories: 1) ready to use, when the user story could be provided to the developer in the current form, 2) refinement needed, when the user story requires minor adjustments, such as more details or corrections in the role, means or ends, 3) cannot be used, when the tool misconstrued the transcript and the user story would require corrections of two or more user story elements. The coverage of user stories from each tool was assessed by reviewing if the AI-generated user stories addressed the same features as the practitioner-generated user stories. If a match exists between the manually extracted key features and those of the AI-generated user stories, this would be noted. Subsequently, the sum of matching features was calculated, representing the coverage per AI tool.

5 Results

The results of the quality assessment are shown in Table 2. The table rows include the source of the user stories and the columns include the evaluated QUS criteria. The highest number of AI-generated user stories was reached by *ChatGPT* and *PRD Maker*, with a total of 13 user stories each. The lowest number of seven user stories was produced by the *BOB*. The percentage values provide information about how well the generated user stories by the specific tool fulfills the specific criteria (in relation to their total number of user stories). The syntactics analysis revealed that 100% of all user stories were "well-formed", including at least the role and the means. Whereas the average of fulfilling "atomic" criteria across all tools is 61%, where *BOB* achieved the highest rate of 86%. From the semantic category, the "unambiguous" criterion has the lowest score with 32%, where only *BOB* and PRD Maker achieved over 50%. The results show that 77% of generated user stories inform about the means of the users without suggesting any solutions, consequently they fulfilled the criterion of "problem-oriented". Analysis of the pragmatics revealed that 97% of user stories were written as well-formed "full-sentence", and 59% of them are considered "estimatable". The user stories generated by the AI tools fulfil on average 5.64 (of 8.0) QUS criteria, we have selected. The lowest number of fulfilled criteria per user story is 5.1, by USC. While the highest is 6.86, by *BOB*.

Table 2. AI generated user stories assessed using QUS framework

Tool	No. of US	Well-formed	Atomic	Minimal	Conceptually Sound	Problem-oriented	Unambigous	Full sentence	Estimatable
USC (GPT)	9	100.00%	44.44%	55.56%	55.56%	77.78%	33.33%	100.00%	44.44%
BOB (GPT)	7	100.00%	85.71%	85.71%	85.71%	85.71%	57.14%	100.00%	85.71%
PRD Maker (GPT)	13	100.00%	61.54%	69.23%	61.54%	61.54%	53.85%	92.31%	69.23%
Pace AI	10	100.00%	60.00%	90.00%	90.00%	50.00%	10.00%	100.00%	50.00%
GeniePM	10	100.00%	62.50%	75.00%	12.50%	75.00%	50.00%	100.00%	50.00%
Gemini Advanced	8	100.00%	60.00%	40.00%	90.00%	100.00%	30.00%	90.00%	50.00%
ChatGPT4	13	100.00%	53.85%	69.23%	69.23%	92.31%	23.08%	100.00%	61.54%
Average	-	100.00%	61.15%	69.25%	66.36%	77.48%	36.77%	97.14%	58.70%
Practitioner	18	100.00%	94.44%	88.89%	100.00%	100.00%	77.78%	100.00%	77.78%

All outcomes were evaluated by the company's practitioner to measure their readiness to be used in following process phases, along with the coverage of functionalities/key-features described in the user stories, generated by each tool in relation to the practitioner-generated user stories. The results of this are presented in Table 3. The table rows include the source of the user stories, together with the average of results. The table's columns include categories described in Sect. 4.2. The percentage values indicate the proportion of user stories in each category relative to the total number of user stories generated by the corresponding tool.

The highest coverage of identified functions from the transcript is 61%, achieved by *PRD Maker* and *GPT4*, by covering 11 functions out of the 18 identified by the company. Consequently, seven functions would not be identified and therefore lost in further documentations, such as user stories, if solely the tools would be trusted to extract insights out of the transcript. Whereas, *PRD Maker* was more precise in identifying roles and ends, with 69% of the user stories evaluated as ready to use, *Gemini Advanced* achieved the lowest coverage rate with 22% with four functions matching and the lowest number of ready to use user stories with only one result.

Moreover, it was detected that *BOB* and *PRD Maker* generated a user story that covers functionality not identified by the practitioner: It was confirmed further by additional review of the transcript.

Table 3. Company evaluation of user stories (US)

Tool	No. of US	Can not be used	Refinement needed	Ready for Use	Coverage
USC (GPT)	9	33.33%	33.33%	33.33%	39%
BOB (GPT)	7	14.29%	42.86%	42.86%	33%
PRD Maker (GPT)	13	7.69%	23.08%	69.23%	61%
Pace AI	10	10.00%	50.00%	40.00%	44%
GeniePM	8	37.50%	12.50%	50.00%	39%
Gemini Advanced ("Bard")	10	80.00%	10.00%	10.00%	22%
ChatGPT4	13	15.38%	38.46%	46.15%	61%
Average	10	28.31%	30.03%	41.65%	37%
Company generated	18				

6 Discussion and Conclusion

With regards to our first research question, we investigated on 7 specific AI tools, potentially capable of generating user stories fulfilling the selection criteria: *GeniePM* and *PaceAI* as standalone products, and three relatively frequently used GPTs for generating user stories (*BOB*, *PRD Maker* and *USC)*. In addition, we included *ChatGPT4* and *Gemini Advanced* as non-specialized AI chatbots.

*ChatGPT*4 also provides similar quality and coverage of user stories as utmost of reviewed solutions and was performing even better than some of the customized GPTs, in terms of coverage and quality criteria: "conceptually sound" and "problem-oriented". Whereas *PRD Maker* generated more user stories, evaluated "as ready for use". Even though, both tools were achieving coverage of 11 user stories, they missed seven functions out of 18 discovered by the practitioner. In general, the results indicate that there is a potential to integrate AI as a supporting tool, to complement practitioners, in documentation activities, such as identifying insights from interviews and generating user stories. However, the expertise of practitioners is still essential throughout the whole process, to for example, critically review the output of the AI and additionally add their insights.

With regards to our second research question, about the quality of user stories generated by the available tools, this study provides insights into the use of AI tools in this context. The use of the QUS framework allowed a comprehensive evaluation across syntactic, semantic and pragmatic criteria and identified potential weaknesses. The quality of user stories generated by each AI tool is varying. We observe good performance according to the QUS criteria "well-formed" and "full-sentence" among all the tools, and low results in "unambiguity". Moreover, we recognize similarities between the scores in the criteria "atomic" and "estimatable", which supports the research of Liskin et al. [13], claiming that the right granularity leads to a more precise estimation of implementation effort. However, there are also discrepancies in quality when considering results from various criteria within one tool. For example, *Gemini Advanced* got the highest score across all tools on "conceptually sound" and "problem-oriented", and at the same time it got the lowest score in "minimal". Moreover, it produced the most user stories, evaluated as "cannot be used". To better understand the reasoning of the differences, further investigation on the fundamental AI techniques utilized by each tool is necessary. Furthermore, we compared the coverage of generated user stories. According to our results, none of the tested AI tools was capable of generating the same amount of user stories as the practitioner, whereas two tools identified additional functionality omitted by the practitioner. At present, it must be said that the investigated AI tools are not sufficient to fully automate the process of extracting relevant information for the creation of user stories for the development team within the company, but rather to support practitioners in the formulation of user stories. In conclusion, the performance and accuracy of off-the-shelf AI tools in extracting relevant information and creating high-quality user stories still needs to be improved, which would lead to more reliable results and potentially enable the full automation of the process in the long run.

7 Future Work

Considering the limitations of our study, we provide recommendations for the future implementation of the described process.

Future work should focus on evaluating user stories by engaging with more experts to reduce the likelihood of bias and consolidate the quality of the evidence. Other aspects of user stories should also be included, such as the potential of AI-assisted generation of acceptance criteria. Furthermore, a broader dataset, such as multiple transcripts, would help to better assess the reliability and consistency of quality of the AI tools/results, potentially revealing trends and patterns. An exploration of the integration of feedback about the generated user stories back to AI tools could provide insights about the possibility of improving the quality of the generated user story. Additionally, user stories evaluated as 'ready to use' should be presented to multiple developers for further discussion and evaluation. This could also provide useful insights about preferences and acceptance towards AI- and practitioner-generated user stories.

References

1. Hofmann, H.F., Lehner, F.: Requirements engineering as a success factor in software projects. IEEE Softw. **18**, 58–66 (2001)
2. Lai, S.-T.: A user story quality measurement model for reducing agile software development risk. Int. J. Softw. Eng. Appl. **8**, 75–86 (2017)
3. Lucassen, G., Dalpiaz, F., Van Der Werf, J.M.E.M., Brinkkemper, S.: Forging high-quality user stories: towards a discipline for agile requirements. In: 2015 IEEE 23rd International Requirements Engineering Conference (RE 2015) - Proceedings. Institute of Electrical and Electronics Engineers Inc., pp. 126–135 (2015)

4. Dikert, K., Paasivaara, M., Lassenius, C.: Challenges and success factors for large-scale agile transformations: a systematic literature review. J. Syst. Softw. **119**, 87–108 (2016)
5. Wang, X., Zhao, L., Wang, Y., Sun, J.: The role of requirements engineering practices in agile development: an empirical study. Commun. Comput. Inf. Sci. **432**(CCIS), 195–205 (2014)
6. Rodeghero, P., Jiang, S., Armaly, A., McMillan, C.: Detecting user story information in developer-client conversations to generate extractive summaries. In: 2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE), pp. 49–59. IEEE (2017)
7. Lucassen, G., Dalpiaz, F., van der Werf, J.M.E.M., Brinkkemper, S.: Improving agile requirements: the quality user story framework and tool. Requir. Eng. **21**, 383–403 (2016)
8. Köse, S.G., Aydemir, F.B.: A Framework to Improve User Story Sets Through Collaboration (2023)
9. Xu, X., Dou, Y., Qian, L., et al.: A requirement quality assessment method based on user stories. Electronics (Basel) **12**, 2155 (2023)
10. Liu, K., Reddivari, S., Reddivari, K.: Artificial intelligence in software requirements engineering: state-of-the-art. In: Proceedings of the 2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI 2022), pp. 106–111 (2022)
11. Raharjana, I.K., Siahaan, D., Fatichah, C.: User stories and natural language processing: a systematic literature review. IEEE Access **9**, 53811–53826 (2021)
12. Bik, N., Lucassen, G., Brinkkemper, S.: A reference method for user story requirements in agile systems development. In: Proceedings of the 2017 IEEE 25th International Requirements Engineering Conference Workshops, REW 2017, pp. 292–298 (2017)
13. Liskin O., Pham, R., Kiesling, S., Schneider, K.: Why we need a granularity concept for user stories. In: Lecture Notes in Business Information Processing 179 LNBIP, pp. 110–125 (2014)
14. Lucassen, G., Dalpiaz, F., van der Werf, J.M.E.M., Brinkkemper, S.: The Use and Effectiveness of User Stories in Practice, pp 205–222 (2016)
15. Heck, P., Zaidman, A.: A Quality Framework for Agile Requirements: A Practitioner's Perspective (2014)
16. Scott, E., Tõemets, T., Pfahl, D.: An empirical study of user story quality and its impact on open source project performance. In: Winkler, D., Biffl, S., Mendez, D., Wimmer, M., Bergsmann, J. (eds.) SWQD 2021. LNBIP, vol. 404, pp. 119–138. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-65854-0_10
17. Ferreira, A.M.S., da Silva, A.R., Paiva, A.C.R.: Towards the art of writing agile requirements with user stories, acceptance criteria, and related constructs. In: Proceedings of the International Conference on Evaluation of Novel Approaches to Software Engineering, ENASE, pp. 477–484 (2022)
18. Trisnawati, E., Raharjana, I.K., Taufik, T., et al.: Analyzing variances in user story characteristics: a comparative study of stakeholders with diverse domain and technical knowledge in software requirements elicitation. J. Inf. Syst. Eng. Bus. Intell. **10**, 110–125 (2024)
19. Dwitam, F., Rusli, A.: User stories collection via interactive chatbot to support requirements gathering. Telkomnika (Telecommun. Comput. Electron. Control) **18**, 890–898 (2020)
20. Popular AI Tools Among IT Professionals Worldwide 2023 | Statista. https://www.statista.com/statistics/1440316/popular-ai-tools-among-it-professionals-globally/. Accessed 21 Mar 2024