CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

How well can ChatGPT create user stories compared to humans?

Bachelor of Science Thesis in Software Engineering and Management

ENES AKIN

HIMANK MEATTLE

Department of Computer Science and Engineering
UNIVERSITY OF GOTHENBURG
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023

How well can ChatGPT create user stories compared to humans?

Supervisor: JENNIFER HORKOFF & KHAN MOHAMMAD HABIBULLAH
Examiner: CHRISTIAN BERGER

University of Gothenburg
Chalmers University of Technology
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Department of Computer Science and Engineering
University of Gothenburg
Chalmers University of Technology
Gothenburg, Sweden 2023

How well can ChatGPT create user stories compared to humans?

Himank Meattle
Department of Computer Science and Engineering
Gothenburg University
Gothenburg, Sweden
gusmeahi@student.gu.se

Enes Akin
Department of Computer Science and Engineering
Gothenburg University
Gothenburg, Sweden
gusakien@student.gu.se

***Abstract*—User stories are widely used in Agile software development yet in practice they are usually poorly written. Researchers have been conducting studies to explore ways to improve user stories and evaluate their quality, but none of them focuses on the newly released ChatGPT's abilities to create user stories yet. INVEST grid method suggests using a 0-3 rating to evaluate user stories based on each INVEST attribute. This study seeks to identify ChatGPT's (GPT-3.5) abilities to create user stories, its strengths and weaknesses. We conducted an experiment to gather user stories from six participants and handed out a survey (survey 1) to gain insights from their experiences. 60 user stories were created during the experiment and rated by 27 respondents in survey 2. Based on survey 1 responses, we identified that ease of use, time efficiency and providing a good starting point are some of ChatGPT's strengths, and lack of creativity and creating incomplete user stories are some of its weaknesses. Ratings gathered from survey 2 showed that ChatGPT can generate user stories of similar quality to those created by humans while taking less time. Our findings suggest that ChatGPT has the potential to improve the user story creation process.**

Index Terms—ChatGPT, GPT-3.5, user stories, AI, requirements engineering, software engineering, experiment, survey

I. Introduction

User stories are a method of writing requirements in an informal way and follow a template such as "As a <type of user>, I want <some goal>, so that <some reason>" [1]. User stories are widely accepted and used in Agile software development [2]. However, in practice, user stories are usually poorly written [8]. As a result, researchers are conducting new studies to improve user stories, such as using Quality User Story (QUS) framework and Automatic Quality User Story Artisan (AQUSA) software tool, which relies on natural language processing (NLP) [9]. There are also other NLP applications for requirements engineering (RE) such as Dowser [10] and RAI [11].

ChatGPT is an NLP chatbot released in November 2022 by OpenAI [6]. In a short time after its release, ChatGPT rose in popularity among people with varying expertise and has been part of many people's workflows, whether it is to help them write the first draft of an article or overcome a programming challenge. Moreover, ChatGPT quickly became subject to studies across different fields, such as software engineering (SE) [21], requirements engineering (RE) [20], medical writing [3] and translation [4].

The purpose of this study is to explore whether ChatGPT (GPT-3.5) can create quality user stories with human intervention. It is important to mention human intervention because of course a human needs to enter prompts to ChatGPT to get results. But in this context, human intervention also means analyzing ChatGPT's responses and asking it to improve or change certain answers.

NLP tools such as Dowser [10], RAI [11] and AQUSA [9] rely on improving existing user stories or requirements, not creating them. However, this study aims to test the limits of ChatGPT in terms of creating user stories. As mentioned above, user stories are widely used in agile development [2], meaning a small improvement in the process of generating user stories can have a positive impact on many organizations and teams. These improvements can be in terms of generating high quality user stories, which would directly affect the quality of the end product, or in terms of reducing the resources allocated to the creation of user stories, which could help the organization financially and allow more time to be spent on the development process.

II. Related Work

Due to ChatGPT being newly released, there are not many studies testing ChatGPT's abilities to create user stories. However, some studies explore ChatGPT's role in RE [20] and ChatGPT prompt patterns for SE and RE [21]. Unlike in RE, many studies explore the strengths and weaknesses of ChatGPT in different fields, such as translation [4] and medical writing [3]. There are also studies exploring NLP and user stories together [5] [9].

Zhang et al. [20] quantitatively evaluated ChatGPT's effectiveness in requirements information retrieval. They selected four benchmark data sets (NFR multi-class classification, app review NFR multi-labels classification, term extraction and feature extraction) that cover different artifacts and identified baselines for each data set. Their findings suggest that ChatGPT significantly outperforms the baseline for the feature extraction data set with a higher precision and lower recall values, while having a higher recall and lower precision in the other three data sets compared to their baselines. The study identifies ChatGPT's promising potential for retrieving requirements information.

White et al. [21] proposed multiple prompt patterns for large language models (LLMs) such as ChatGPT in an effort to improve requirements elicitation as well as code quality. They suggest using prompt patterns to minimize the errors and mistakes LLMs make and reach their full potential, which could result in automating processes such as simulating a system based on requirements and more. They identified LLMs' huge potential for automating tasks in SE, however, their findings suggest that human involvement is required in the current state of LLMs to reach their full potential.

Rharjana et al. [5] conducted a systematic literature review to analyze the role of NLP on user stories. Their findings were focused on different applications of NLP on user stories, such as generating models from existing user stories and discovering defects in user stories [5]. This study identifies different applications of NLP on existing user stories and provides insights about the state-of-the-art research related to NLP and user stories.

Studies conducted by Lucassen et al. [8] [9] proposed the use of the AQUSA tool to improve existing user stories. AQUSA works by detecting problems in the user stories given and suggesting improvements. However, AQUSA cannot detect all defects with 100% recall, which is why it focuses on detecting defects that are easily describable and algorithmically determinable [9]. Automating these tasks allows the requirements engineer to fully focus on other defects that cannot be detected by AQUSA with 100% recall.

AQUSA is a good effort in improving the RE workflow by automating certain tasks, and it falls under category (a) of natural language (NL) RE tools [12]. Berry et al. [12] state that there are four main categories of NL RE tools, which are briefly: a. Tools that find defects b. Tools that generate models c. Tools that trace links between other artifacts and requirements d. Tools that recognize abstractions in documents. These four categories share the same goal of improving the RE workflow but none of them cover the creation of user stories by using NLP applications or similar tools.

In order to evaluate the quality of user stories, Buglione et al. [13] discuss the application of the INVEST method. INVEST (Independent, Negotiable, Valuable, Estimable, Small, Testable) consists of six attributes that should be focused on to create good user stories [1]. Buglione et al. [13] propose the use of the INVEST grid to measure user stories both qualitatively and quantitatively.

Studies mentioned in this section discussed the role of NLP such as AQUSA and ChatGPT [8] [9] [20] [21] in RE, potential use cases of ChatGPT in creative fields [4] [5] and methods of evaluating user stories' quality [13]. However, none of the studies focused on ChatGPT's ability to create user stories.

Unlike the studies discussed in this section, our study explores ChatGPT's ability to create user stories compared to humans, by using the INVEST grid [13] to evaluate user stories' quality.

III. Research Methodology

In order to investigate ChatGPT's abilities to create user stories compared to humans' abilities, following research questions were formulated.

A. Research questions

- RQ1: How effective is ChatGPT at creating user stories compared to humans?
- RQ2: What are the advantages and disadvantages of using ChatGPT to create user stories?

B. Research methodology used

In order to answer the research questions, we conducted an experiment and two surveys as part of the experiment. The experiment was conducted to collect user stories from different participants while the surveys were conducted to gain insights from the experiment participants' experiences and rate the collected user stories. The experiment consisted of four main parts: planning, user story generation, survey 1 and survey 2, which is explained in chronological order.

a) Planning: The planning started with us selecting two project descriptions created by a requirements engineer, as shown in Appendix A. These project descriptions were selected to be used by participants to create user stories during the experiment. We contacted software engineers, both students and professionals who are familiar with user stories. Three professionals and three students volunteered. Experiment participants were selected based on convenience sampling [7]. The reason we included different participants such as students and professionals is to conduct the experiment with participants that have different levels of experience with user stories, in an effort to gain insights from different perspectives of these participants.

b) User story generation: Once the project description and the participants were selected, the experiment began. The experiment followed within-subjects design [19]. Every participant created two sets of user stories based on two different project descriptions that are shown in Appendix A. Each set of user stories contained five user stories.

Set A of user stories was created manually by the participants themselves and set B of user stories was created by ChatGPT, under the supervision and assistance of the participants. In order to minimize the learning effects [19], we used two different project descriptions and randomized the order of user story creation based on these project descriptions. Three participants generated user stories with ChatGPT first and then manually, while the other three generated user stories manually first and then with ChatGPT.

There was a rule all participants followed when they used ChatGPT to create user stories: They were not allowed to manually edit any user story generated by ChatGPT. However they were allowed to enter as many prompts as they want into ChatGPT. For example, if they wanted to change a certain user story, they could enter a prompt such as "User story 2 is a bit vague, can you change it and make it easier to understand?", so that ChatGPT could create a new user story based on the feedback. This was done to encourage participants to interact more with ChatGPT, instead of manually editing user stories, to learn more about how participants used ChatGPT and help us answer RQ2.

All interactions (participants' prompts and ChatGPT's responses) that took place between the participants and ChatGPT was recorded to be analyzed by the researchers in order to answer RQ2.

Once this part of the experiment was complete, we had the following data collected:

a. User stories generated by ChatGPT:

- Three sets of user stories (five user stories each) based on project description A
- Three sets of user stories (five user stories each) based on project description B

b. User stories generated manually by participants:

- Three sets of user stories (five user stories each) based on project description A
- Three sets of user stories (five user stories each) based on project description B

In total 60 user stories and six recorded interactions between ChatGPT and participants were collected.

c) Survey 1: Once the participants completed creating user stories, they answered a survey (survey 1) that contained both open-ended and closed-ended questions. Survey 1 focused on participants' thoughts about the processes they used, the challenges they faced, advantages of one process compared to the other (manual and ChatGPT) or their observations in general. The qualitative data collected from survey 1 helped us answer RQ2, which is discussed further in data collection and data analysis sections.

Q: As a small business owner, I want to be able to talk to a real person with IT expertise so that I get help when the interactive assistant is not being as useful as expected.

Please rate this user story based on the INVEST grid (0 = Poor/Absent, 1 = Fair, 2 = Good, 3 = Excellent)

A: Independent: [0, 1, 2, 3]

Negotiable: [0, 1, 2, 3]

Valuable: [0, 1, 2, 3]

Estimable: [0, 1, 2, 3]

Small: [0, 1, 2, 3]

Testable: [0, 1, 2, 3]

Fig. 1. An example question from survey 2

d) Survey 2: Once the user story creation was complete, we handed out an unsupervised survey (survey 2) that included the user stories created by participants and questions to rate these user stories based on the INVEST grid [13]. Structure of the questions in survey 2 is shown in Fig. 1.

There were 60 user stories to be rated in survey 2, and having all of them in one questionnaire would make the questionnaire too long and significantly reduce the response rate [14]. This is why we split survey 2 into three smaller surveys that contained 20 user stories each. Each sub-survey included manually created user stories and the ones created by ChatGPT. When we mention survey 2 in this paper, it includes all three sub-surveys that follow the same template, but with different user stories to be rated.

Survey 2's target group was software developers, requirements engineers, students and other practitioners who are experienced with user stories. We used convenience sampling [14] as well as voluntary response sampling to select participants for survey 2. In order to attract participants, survey 2 was advertised through social media, contacts in companies, colleagues and classmates. Initially, we aimed to get 30 responses to survey 2, and we received 27 responses in total. In the end, sub-surveys had 13, 13 and 12 responses respectively, meaning each user story was rated at least 12 times by different people. We made it clear while advertising and at the beginning of the survey that we were only seeking participants who are familiar with user stories and INVEST in order to gather more accurate data.

The purpose of survey 2 was to gather quantitative data about the quality of user stories created during the experiment, which helped us answer RQ1. By comparing the INVEST ratings (from survey 2 results) of user stories generated by ChatGPT with user stories generated manually, we explored ChatGPT's abilities in terms of creating user stories compared to humans.

1) Data collection: Throughout the experiment, two sets of data were collected, both qualitative and quantitative. Data set 1 included six experiment participants' recorded interactions with ChatGPT, survey 1 results, and the user stories created during the experiment. Data set 2 was collected from survey 2 and consisted of each user story's rating.

a) Data set 1: Data set 1 consists of qualitative data that was collected from six participants separately. It contains messages between participants and ChatGPT as well as participants' answers to survey 1, which were about their observations during the experiment.

This data was collected in order to understand the advantages of disadvantages of using ChatGPT to create user stories. Our findings from collecting and analyzing data set 1 helped us answer RQ2. The main focus of this data is the observations and experiences of participants who used ChatGPT to create user stories.

Data set 1 also contains the user stories created by participants. Although by themselves, these user stories don't tell us anything, they were rated in survey 2 by participants, in order to identify their quality.

It is important to highlight that all participants were informed before the experiment began that their responses to survey 1 and recorded interactions with ChatGPT would be analyzed extensively during the research and their responses and interactions could be used verbatim in the report. Furthermore, we obtained all participants' consent regarding the use of their data in this manner before the experiment began.

TABLE I
SURVEY 1 QUESTIONS

#	RQ	Question Text	Answer Choices
1	D	What is your current title/position?	[software engineer/developer, student (software engineering), other: specify]
2	D	What is your prior education?	[bachelor's in software engineering, bachelor's in other fields, master's in software engineering, master's in other fields, other: specify]
3	D	How many years of work experience do you have in software development or related fields?	[0, <3, 3-5, 5+]
4	RQ2	Did you have trouble setting up/using ChatGPT environment? If yes, please explain what happened.	[yes: specify, no]
5	RQ1	Please submit here, 5 user stories created using ChatGPT during the experiment.	[open ended]
6	RQ2	Please attach a text file here, containing your entire conversation with ChatGPT during the experiment.	[file upload]
7	RQ2	How long did it take you to create user stories with ChatGPT?	[#]
8	RQ2	How would you rate your experience creating user stories with ChatGPT and why?	[sliding bar, 1-5]
8.1	RQ2	Please explain your answer for the question above.	[open ended]
9	RQ1	Please submit here, 5 user stories created manually during the experiment.	[open ended]
10	RQ2	How long did it take you to create user stories manually?	[#]
11	RQ2	How would you rate your experience creating user stories manually and why?	[sliding bar, 1-5]
11.1	RQ2	Please explain your answer for the question above.	[open ended]
12	RQ2	If you were to choose a method for your next project, which one would you choose and why?	[creating user stories manually, creating user stories with ChatGPT, creating user stories by combining ChatGPT and manual methods, other: specify]
12.1	RQ2	Please explain your answer for the question above.	[open ended]

b) Data set 2: Data set 2 is the results gathered from survey 2, which consists of quantitative data. In survey 2, participants rated different user stories that were created by ChatGPT and manually, using the INVEST grid method [13]. User stories in survey 2 were Likert-like, rated from 0-3 (0 = Poor/Absent, 1 = Fair, 2 = Good, 3 = Excellent). Furthermore, if a user story is rated 0 for any attributes of INVEST, it is marked as bad quality. A valid user story should get a rating of at least 1 in all INVEST attributes [13].

This was the most crucial data to be collected during the study because it allowed us to compare the user stories created by humans and ChatGPT in order to answer RQ1, which is the main focus of the study.

2) Pilot experiment: In order to identify the potential weaknesses and limitations of the experiment, we asked two software engineering students to pilot the experiment as well as survey 1. The feedback suggested that the experiment was too long. The only change we made to the experiment design was to reduce the number of user stories we expected from each participant to ten (five with ChatGPT and five manually), to reduce the duration of the experiment. No changes have been made to survey 1.

Survey 2 was also piloted by two volunteers to identify areas to improve in the survey. Feedback gathered suggested that some questions in the survey were missing necessary options, which was quickly fixed. The overall design of survey 2 has not been changed.

3) Data analysis: Since survey 1 mainly consisted of open-ended questions and collected qualitative data, we used thematic analysis [15] to analyze the data. Because it allowed us to identify, analyze and report themes in the survey answers.

We used the six steps approach to apply thematic analysis [16] to survey 1. The first step was to know the data and hence we read through all the answers. The next step was to start generating codes and picking keywords in the answers. For example, some keywords were 'easy to use´ and 'fast´. We then grouped similar codes, identified the shared meaning and how it is connected to the RQs and generated themes. The fourth step included reviewing the generated themes and picking those which are organized around the RQs, have clear boundaries and are of good quality. Step five was to refine the selected themes, and in the final step we produced the report.

Using thematic analysis helped us identify the advantages and disadvantages of using ChatGPT to create user stories, which allowed us to answer RQ2.

Since survey 2 questions were Likert-like we analyzed survey 2 with the help of Likert item and scale analysis as recommended by Linåker et. al [14]. For each user story, their INVEST attributes were analyzed separately using mode as well as mean. As a result, each user story had six mode values and six mean values -a mode and a mean value for each INVEST attribute-. After this step, user stories were grouped into two categories, created with ChatGPT and manually. Following the previous method, overall mode and mean values were calculated for each category and at the end we had separate mode and mean values for INVEST attributes of the user stories created with ChatGPT and manual. We also calculated mode and mean values of user stories without separating them into INVEST attributes.

Analyzing the INVEST attributes separately helped us identify strengths of user stories created with ChatGPT by comparing them to those created manually.

Lastly, we analyzed the recorded interactions between participants and ChatGPT by using the six steps approach to thematic analysis. The goal of this analysis was to explore the prompts given to ChatGPT to identify potential strengths of weaknesses of ChatGPT and answer RQ2.

IV. Results

A. Demographics

This subsection describes the experiment and survey participants' work experiences in SE or similar fields and their education levels.

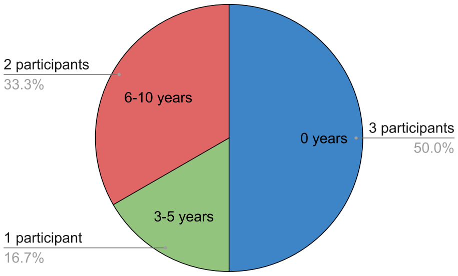


Fig. 2. Distribution of six experiment participants based on years of work experience in SE

1) Experiment & Survey 1: As discussed in Section III, six participants were selected for the experiment. Three of these participants are software engineers while the other three are undergraduates. All six participants study or work in Sweden.

As shown in Fig. 2, two participants had 6-10 years and one participant had 3-5 years of work experience in an SE related field. The other three participants are undergraduates who do not have work experience. The three undergraduates who participated in the experiment are studying software engineering, computer science and game development respectively.

All six participants reported that they have used user stories before in various projects, and two of the undergraduates also reported that they have taken requirements engineering courses in which user stories were the main focus.

2) Survey 2: Survey 2 had 27 participants in total. All 27 participants stated that they have used user stories before and they are comfortable working with them.

As shown in Fig. 3, the majority of participants have at least a year of work experience in an SE related field. On the other hand, almost 50% of the participants are students without any work experience, pursuing SE or related undergraduate degrees.

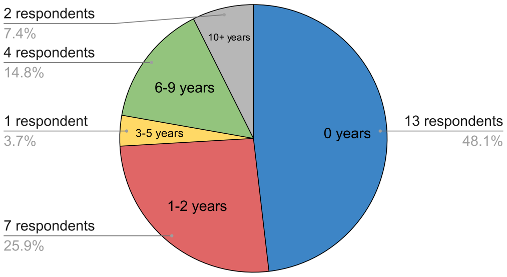


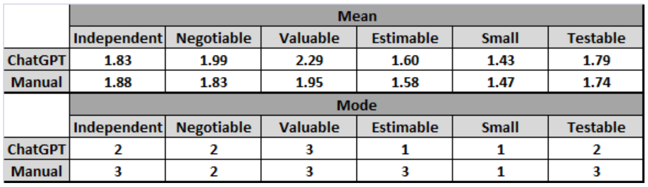
Fig. 3. Distribution of 27 respondents to survey 2 based on years of work experience in SE

B. RQ1: How effective is ChatGPT at creating user stories compared to humans?

The answers to RQ1 came from survey 2 results, which consists of quantitative data of the user stories' ratings using INVEST grid. We performed basic statistical analysis on the collected data by calculating both mean and mode. We have also calculated values at different granularity levels, for each INVEST attribute per group and also per group as a whole, in an effort to gain more insights from the data.

Table II shows the mean and mode values of both groups respectively, i.e. 30 user stories created with ChatGPT and 30 created manually as shown in Appendix E and Appendix F respectively, which were calculated for each INVEST attribute separately. We received 13 responses per user story for both US1-US20 (ChatGPT group) and US31-US50 (manual group) and 12 responses per user stories for both US21-US30 (ChatGPT group) and US51-US60 (manual group). The average response was 12.6 per user story per group.

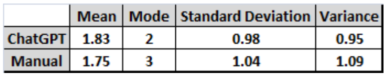
TABLE II
INVEST RATINGS OF USER STORIES



	Mean					
	Independent	Negotiable	Valuable	Estimable	Small	Testable
ChatGPT	1.83	1.99	2.29	1.60	1.43	1.79
Manual	1.88	1.83	1.95	1.58	1.47	1.74
	Mode					
	Independent	Negotiable	Valuable	Estimable	Small	Testable
ChatGPT	2	2	3	1	1	2
Manual	3	2	3	3	1	3

Table III shows descriptive statistical analysis for each group as a whole without considering each INVEST attribute separately.

TABLE III
OVERALL RATINGS OF USER STORIES

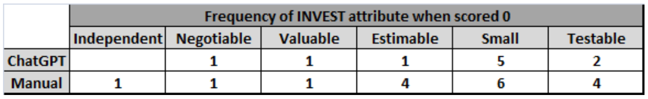


	Mean	Mode	Standard Deviation	Variance
ChatGPT	1.83	2	0.98	0.95
Manual	1.75	3	1.04	1.09

As shown in Table II above, when mean values are compared, ChatGPT scores higher ratings for four attributes of INVEST i.e. 'Negotiable', 'Valuable', 'Estimable' and 'Testable', while also having a slightly higher rating when the mean per group is calculated as shown in Table III. When we look at the mode values instead, we observe a different result, user stories created manually have either the same or higher ratings than ChatGPT. They also scored a rating of 3 (Excellent) for four INVEST attributes i.e. 'Independent', 'Valuable' 'Estimable' and 'Testable', which is the highest rating possible, while ChatGPT only scored a 3 for one attribute i.e 'Valuable'.

As shown in Table III, mean, standard deviation and variance, are almost the same, in both groups. The mode for ChatGPT group is one point lower than the other group.

TABLE IV
INVEST ATTRIBUTES THAT SCORED 0



	Frequency of INVEST attribute when scored 0					
	Independent	Negotiable	Valuable	Estimable	Small	Testable
ChatGPT		1	1	1	5	2
Manual	1	1	1	4	6	4

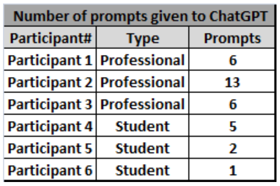
We analyzed Table II more in depth, at an individual user story level, in order to identify which user stories might have scored 0 even once, for any of their INVEST attributes and marked them as invalid [13]. We identified seven user stories in ChatGPT group and six in the manual group. The most common INVEST attribute that scored 0 was 'Small' for both ChatGPT and manually created user stories. The other attributes that scored 0 in the ChatGPT group are 'Testable', 'Negotiable', 'Valuable', 'Estimable', and in the manual group they are are 'Estimable', 'Testable', 'Independent', 'Negotiable', 'Valuable', as shown in Table IV.

In the end, the results for RQ1 indicate the same or similar values for almost all measurements.

C. RQ2: What are the advantages and disadvantages of using ChatGPT to create user stories?

Participants' conversations with ChatGPT, Q4, Q7, Q8-8.1, Q10, Q11-11.1 and Q12-12.1 in Table I were analyzed to answer RQ2. Most of these questions were open-ended and collected qualitative data as a result. Also, some questions were related to each other such as Q8, Q11 and Q12 and they are analyzed together to identify common themes across them, as described in the data analysis section.

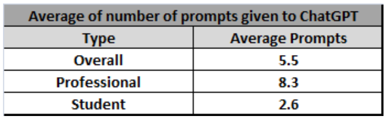
TABLE V
PROMPTS GIVEN TO CHATGPT BY PARTICIPANTS



Number of prompts given to ChatGPT		
Participant#	Type	Prompts
Participant 1	Professional	6
Participant 2	Professional	13
Participant 3	Professional	6
Participant 4	Student	5
Participant 5	Student	2
Participant 6	Student	1

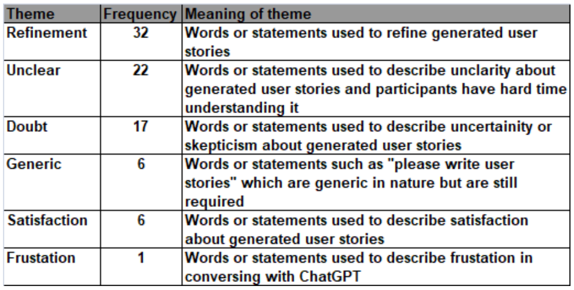
We analyzed the interaction between ChatGPT and experiment participants. The number of prompts given by participants to ChatGPT are shown in Table V. The average number of prompts given by all users is 5.5 as shown in Table VI. When we analyzed the data further in detail, the average number of prompts given by software engineer participants is 8.3 while it is 2.6 for students as shown in Table VI. After performing thematic analysis on these interactions as shown in the code book in Appendix B, we found several emerging themes such as refinement, unclear, doubt, generic, satisfaction and frustration, in decreasing order as shown in Table VII.

TABLE VI
PROMPTS GIVEN TO CHATGPT BY PARTICIPANTS



Average of number of prompts given to ChatGPT	
Type	Average Prompts
Overall	5.5
Professional	8.3
Student	2.6

TABLE VII
THEMES GATHERED FROM THE INTERACTIONS BETWEEN PARTICIPANTS AND CHATGPT



Theme	Frequency	Meaning of theme
Refinement	32	Words or statements used to refine generated user stories
Unclear	22	Words or statements used to describe unclarity about generated user stories and participants have hard time understanding it
Doubt	17	Words or statements used to describe uncertainity or skepticism about generated user stories
Generic	6	Words or statements such as "please write user stories" which are generic in nature but are still required
Satisfaction	6	Words or statements used to describe satisfaction about generated user stories
Frustation	1	Words or statements used to describe frustation in conversing with ChatGPT

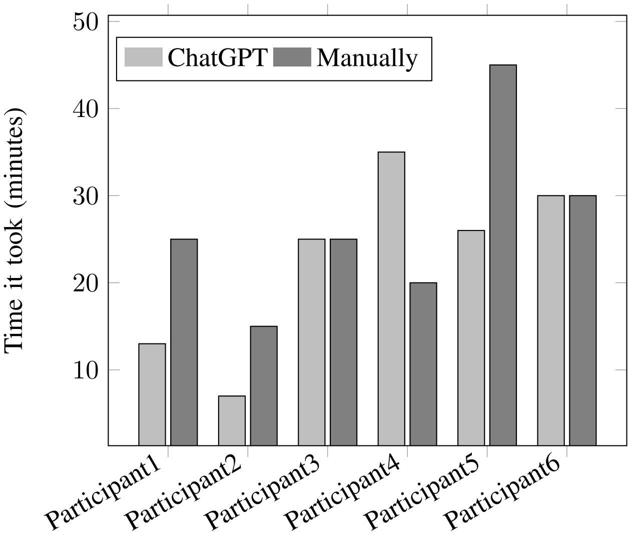


Fig. 4. Time it took participants to create 5 user stories with ChatGPT and manually

1) Q7 & Q10: Duration of the user story creation process with ChatGPT as well as manually were gathered from answers to Q7 and Q10. This was done in an effort to identify whether one method was significantly faster than the other, which could suggest a potential strength or weakness of these methods. As shown in Fig. 4, five participants reported that creating five user stories with ChatGPT took a less or equal amount of time compared to creating user stories manually.

Only one participant reported that the process took longer with ChatGPT.

The average time it took with ChatGPT was approximately 23 minutes while with the manual method it was 27 minutes. This means on average, creating user stories with ChatGPT was 17% faster than creating them manually. However, as shown in Fig. 4, there were exceptions. Participant 4 was faster at creating user stories manually than with ChatGPT.

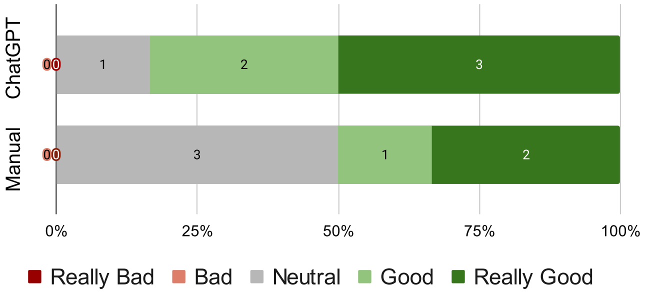


Fig. 5. Six experiment participants' experiences creating user stories with ChatGPT and manually

2) Q8 & Q11 & Q12: In order to have a better understanding of the underlying reasons for participants' answers to questions mentioned, their responses to open-ended questions are analyzed using thematic analysis in the following part of this section.

As described in the data analysis section, the 6 steps approach is used to apply thematic analysis to participants' answers to open-ended questions such as Q8 & Q8.1 as shown in Table I: 'How would you rate your experience creating user stories with ChatGPT and why?' and Q12 & Q12.1: 'If you were to choose a method for your next project, which one would you choose and why?'.

As shown in Fig. 5, when participants were asked to rate their experiences of creating user stories with ChatGPT, three participants rated it as really good, two participants rated it as good and one participant rated it neutral. When participants were asked to rate their experiences of creating user stories manually, two participants rated it as really good, one participant rated it as good and three participants rated it neutral. None of the participants rated their experience of using either of these methods as really bad or bad. But overall, participants had better experiences using ChatGPT compared to creating user stories manually.

When participants were asked to choose a method for creating user stories in the future, all six participants chose 'Creating user stories by combining ChatGPT and manual method' as their answer. However, they had different reasons for it. As shown in Appendix D, participants raised various concerns about using only ChatGPT and suggested manually revising user stories. Some of the mentioned issues were lack of trust, incomplete user stories and lack of creativity, which could be resolved by manually revising the user stories ChatGPT creates. However, participants also identified issues such as privacy concerns and the difficulty of entering all the project details to ChatGPT, which cannot be easily resolved by manually revising the user stories.

For example, participant 2 stated: "For a project within a company I would most likely only create them manually, mainly for concern about leaking any sensitive information". Although testing the security measures of ChatGPT is beyond the scope of this paper, it is still important to identify the concerns.

TABLE VIII
THEMES GENERATED FROM THE ANSWERS TO Q8-8.1 AND Q12-12.1 (APPENDIX C AND D)

Theme	Frequency	Type
Ease of use	12	Advantage
Good starting point	12	Advantage
Incomplete user stories	8	Disadvantage
Lack of creativity	7	Disadvantage
Time efficiency	6	Advantage
AI will be the future	4	Neutral
Lack of trust	4	Disadvantage
Satisfaction with user stories	4	Advantage
ChatGPT misunderstanding	3	Disadvantage
Insufficient input	2	Disadvantage
ChatGPT and Manual together	2	Neutral
Difficulty of only using ChatGPT	1	Disadvantage
Too large of a scope	1	Disadvantage
Difficulty of providing project details to ChatGPT	1	Disadvantage
Privacy concerns	1	Disadvantage

Table VIII shows how frequently each theme appeared across six participants' responses. Multiple themes were present in each participant's responses, and it still counted towards the frequency when a theme appeared more than once in a participant's response.

Since RQ2 aims to identify the advantages and disadvantages of using ChatGPT to create user stories, codes and themes were grouped under two categories: advantages and disadvantages. Themes were categorized based on the codes and participants' tones in their responses. Some themes such as "AI will be the future" did not fall into the categories advantage and disadvantage, and they were marked as neutral. The five most frequently occurring themes -three advantages and two disadvantages- are identified and explored further in this section.

In terms of the advantages of using ChatGPT, the most commonly occurring themes were: Ease of use, time efficiency and providing a good starting point.

Ease of use: This theme was generated based on codes such as "Straightforward to use" and "It (ChatGPT) followed my instructions". The theme was present in five participants answers to Q8.1. Codes collected from participants' answers to Q8.1 such as "It (ChatGPT) followed my instructions", "It was able to improve its suggestions" and "It was able to handle longer descriptions" suggest that ChatGPT is good at understanding the prompts given by users and reflecting on its own answers. ChatGPT is able to understand the given feedback and revise the initial user stories it created.

Time efficiency: This theme was generated based on codes such as, “I think it (ChatGPT) gave me a quick and nice start, by only providing the description and what to do with it (create stories)” and “I think it is pretty straightforward to create user stories with ChatGPT, I don’t have to create them myself which saves me lots of time”. Codes that suggest time efficiency were present in five participants’ responses throughout survey 1. However, as shown in Fig. 4, two participants reported that creating user stories with ChatGPT took equal amount of time compared to creating them manually, and participant 4 reported that creating user stories manually was faster.

After analyzing participant 4’s and the other two participants’ conversation with ChatGPT, it became apparent that the reason was the limitation we set before the experiment started. Participants were prohibited from manually editing the user stories ChatGPT created, they had to ask ChatGPT to do it for them. Participant 4 was not satisfied with some of the initial user stories generated by ChatGPT and they asked ChatGPT to improve the user stories multiple times. An example input from participant 4 is the following; “Rewrite number 3 and 5 as follows while leaving out the other two: Change the type of industry for number 3. It’s unlikely that a moving company would have a high level of tech knowledge so pick something where this would make more sense. For number 5 just reword the part after ‘technology security,’ so that it doesn’t say ‘breaking the bank’”.

A part of participant 4’s response to Q12.1 was “ChatGPT was very useful in providing ideas for me which would have allowed me to quickly create the groundwork for more useful user stories”, which supports our claim.

Participant 4 as well as other participants’ conversations with ChatGPT and responses to Q12.1 suggest that most participants found using **only** ChatGPT to create user stories time consuming but they believe combining ChatGPT and manual methods of creating user stories can be very time efficient. This theme is strongly tied to the third theme we discuss, which is “Providing a good starting point”.

Providing a good starting point: Five participants identified that ChatGPT is very good at giving initial suggestions and providing them with a starting point. This theme was generated from codes such as “It will give you suggestions to have as a base, adding things that might be useful in context” and “A good basis from the get-go with little effort and quick results”. Similar codes to these were present in other participants’ responses as well as shown in Appendix C and D.

When participants mentioned “a good basis” or “base”, they meant a good starting point to improve and create more user stories. This is also related to the previous theme discussed because using ChatGPT allows user story creation to begin and reach a certain point faster than doing it manually does. This can include but is not limited to generating high-level user stories and giving suggestions about certain user stories.

Entering a long project description into ChatGPT and asking it to generate user stories from that description results in generating multiple user stories within seconds. Some of these user stories would be complete, while some might serve as a starting point for a better written user story, which requires manual editing. Participants’ responses to Q12 support this claim since all six participants answered ‘Creating user stories by combining ChatGPT and manual methods’ when they were asked to choose a method for their next project.

When it comes to the disadvantages of using ChatGPT, the most commonly occurring themes were: Lack of creativity and incomplete user stories.

Lack of creativity: This theme was generated from codes such as “Being restrained to only change the user stories through the AI made for less variety between the user stories”, “It obviously tried to produce a one to one mapping between the project description and five user stories” and “ChatGPT did lack imagination and was not able to properly remember past outputs. I asked it to rewrite based on my initial input once but got basically the same answer as the very first time I gave it a prompt”. Similar codes to these were found in four participants’ responses to Q8.1 and Q12.1.

User Story A: As a SAVE worker, I want to be able to quickly access information about a gun violence event in the field, so that I can provide timely and appropriate services to affected individuals and identify potential retaliation risks.

User Story B: As a SAVE team member, I want to be able to quickly enter and look up information about gun violence incidents in the field, so that I can provide appropriate services to the affected individuals.

Fig. 6. Two different user stories created by ChatGPT

Lack of creativity was one of the main concerns participants had regarding using only ChatGPT to create user stories. Due to ChatGPT’s lack of creativity, the user stories it generated were similar to one another, lacking variety. As shown in Fig. 6, some user stories created by ChatGPT are very similar. This is because of ChatGPT’s efforts to create one-to-one mapping between the project description and the user stories, which was also identified by participant 1. ChatGPT tried to create five user stories based on features that were mentioned in the given project description, instead of coming up with new unique features. This is an important limitation of ChatGPT to identify because project descriptions might not always cover all the necessary features a software needs, and creativity is necessary when generating user stories in order to explore potential features that can benefit the software but are not mentioned in the description.

Incomplete user stories: The second most important disadvantage of using ChatGPT is that some user stories it creates are incomplete. Five out of six participants stated that some user stories ChatGPT created were incomplete, mainly in terms of their scopes. Codes such as “I think the scope for each story was to large, if going to be used in the real world in a sprint" and "However, it may be incomplete or could be the complete opposite of what you need, so it is important to manually adjust the user stories as needed" were identified from participant 2 and 6's responses, and similar codes that discuss incomplete user stories were found in three other participants' responses as well.

V. Discussion

Prior studies suggest that ChatGPT has a promising potential for undertaking different SE and RE tasks [20] [21] and our findings support this claim. We found that ChatGPT can create user stories that have a similar quality to those humans create. It is important to state that these results do not suggest that ChatGPT is ready to automate the user story creation process yet. Human involvement was present when participants created user stories using ChatGPT. This involvement was not limited to only giving initial prompts to ChatGPT, it also included revising the user stories ChatGPT generated by manually reading them and suggesting improvements in the form of new prompts.

As shown in Table II and Table III, although ChatGPT scores higher mean values, the difference between them is not significant enough to suggest that one method is better than the other. When mode values are compared instead, the manual method scores higher than ChatGPT for most INVEST attributes. As shown in Table III, when ChatGPT and manual groups are compared, mean, standard deviation and variance have almost the same values, and mode values are also close enough. This suggests that the overall quality of user stories generated by ChatGPT and manually are on a similar, if not the same level.

As shown in Table IV, the most common INVEST attribute that scored 0 was S (Small) for both ChatGPT and manually created user stories. This suggests that although ChatGPT struggled to create user stories with a small scope, participants had similar struggles.

Our findings suggest that with the state of GPT-3.5, combining ChatGPT and manual methods is the most ideal way to create user stories. As shown in Appendix D, all six experiment participants' answers to Q12 were 'Creating user stories by combining ChatGPT and manual methods'. When participants were asked to explain the reason for their answer to Q12, previously identified advantages of ChatGPT such as time efficiency and providing a good starting point, and disadvantages such as lack of creativity and incomplete user stories were present in their answers. Participants reported that manually revising the user stories ChatGPT created is necessary to overcome these weaknesses.

Based on our findings, we suggest that creating an initial set of user stories with ChatGPT, then manually revising them and creating more user stories manually -if necessary- is an ideal way to implement ChatGPT into the user story creation process. This allows requirements engineers to make the most use of ChatGPT's strengths, especially its time efficiency and providing a good starting point, while minimizing the issues its weaknesses may cause such as creating incomplete user stories.

We found that time efficiency and providing a good starting point are the two main strengths of using ChatGPT while lack of creativity and generating incomplete user stories are some of its major weaknesses. The workflow suggested above makes use of ChatGPT's time efficiency and its strength of providing a good starting point by allowing ChatGPT to create an initial set of user stories. As shown in Fig. 4, on average ChatGPT is faster than humans at reading, analyzing and creating user stories based on a project description. This would allow requirements engineers to spend less time at the beginning of the process and allocate the majority of their resources to improving these user stories. By spending more time on revising these user stories, requirements engineers can use their creativity to overcome the weaknesses of ChatGPT, such as its lack of creativity and incomplete user stories.

After analyzing participants' conversations with ChatGPT, we found that one of the most commonly occurring themes was refinement. ChatGPT was able to revise the initial user stories it created when participants gave feedback to ChatGPT. This shows the importance of human involvement in the process.

The experiment had started before GPT-4 was released and therefore GPT-3.5 version of ChatGPT was used. GPT-4 outperforms GPT-3.5 in various areas such as academic and professional exams, as well as other benchmarks [22]. Considering the performance difference between the two versions, we assume GPT-4 can generate better user stories than GPT-3.5. Testing GPT-4's abilities in future work can help researchers partially automate the process of creating user stories with improved workflows that yield better results.

VI. Threats to validity

A. Internal Threats

Observational and researcher bias [17] are some of the main limitations to internal validity of this study. When designing the experiment follow-up survey (survey 1), we asked participants to explain their answers in detail to open-ended questions such as Q8.1, Q11.1 & Q12.1 as shown in Table I. However, some participants gave short answers to these questions which resulted in collecting less data than anticipated. Some participants also had very short interactions with ChatGPT. Considering that these answers and the interactions were the main data that was used when answering RQ2, observational bias could have arisen during the data analysis stage.

Researcher bias could also be present in this work due to the nature of the qualitative data collected from survey 1 and how it was analyzed. This study aims to identify ChatGPT, an AI chatbot's abilities in terms of creating user stories, and we might have biases that could favor or oppose AI. As a result, when analyzing the qualitative data in order to answer RQ2, our biases may have unintentionally affected to conclusions we drew from the data. In order to minimize the effects of researcher bias, we were careful to be as unbiased as possible when analyzing the data gathered from survey 1. We also designed survey 2 in a way that we are not the ones to rate the user stories collected during the experiment since we know which user stories were created by ChatGPT. Instead, survey 2 was handed out to new participants who do not know which user stories were created by ChatGPT to minimize the effects of biases related to AI.

Apart from the above, another potential threat to this study relates to the participants' familiarity with ChatGPT and their interaction with the tool. Participants with a higher level of experience in requirement engineering, user stories, AI tools and ChatGPT, may be more proficient in utilizing ChatGPT effectively, resulting in better outcomes. Considering that ChatGPT is a relatively new AI tool, although we provided participants with tutorial links and encouraged them to familiarize themselves with ChatGPT prior to the experiment, we lack knowledge about their actual proficiency in using the tool. Furthermore, neither the experiment nor survey 1 included a measurement of participants' ability to use ChatGPT. Consequently, it is likely that some participants were unable to fully leverage ChatGPT, leading to its under-utilization. This limitation may have influenced the identification of both the strengths and weaknesses of ChatGPT. It is important to acknowledge that the limited research on ChatGPT makes it challenging to ascertain its full potential. Consequently, it also becomes difficult to measure whether participants were able to maximize their utilization of ChatGPT during the experiment.

Additionally, it is important to consider the potential impacts of participants lacking professional work experience, both positive and negative. Such participants bring a unique perspective compared to those with work experience, as they are less influenced or not influenced at all by prior work-related experiences. Their lack of experience allows them to approach the research with a different mindset, providing valuable insights distinct from participants with work experience. Similar to actual software development projects, in which teams typically consist of members with diverse work experiences and backgrounds. However, participants without work experience also present certain concerns. They may lack a deeper understanding of the practicalities and complexities involved in requirement engineering and software development. Their responses might be based on theoretical knowledge and assumptions, which may not accurately reflect the realities of software development projects. Nevertheless, we took measures to ensure that all participants were at least familiar with the concept and template of user stories. We explicitly mentioned user stories in the experiment and survey guidelines, and provided a tutorial link for a quick revision, to mitigate any potential knowledge gaps.

B. External Threats

Due to limited time and access to resources, the experiment was conducted with only six participants. We were careful when selecting participants by making sure that all six participants are comfortable with creating user stories, so that ChatGPT is not being compared to novices. However, there can always be requirements engineers who are better or worse at creating user stories than our participants, which could change the results of the experiment, since the quality of the user stories created by ChatGPT were compared to that of created by participants.

Another threat to the generalizability of the results comes from sampling bias. Participants for the experiment and survey 1 were selected based on convenience sampling, and survey 2 participants were mainly sampled by voluntary response sampling. Both these methods are non-probability sampling methods [14] and pose a risk of sampling bias.

Furthermore, the decision to split survey 2 into three smaller sub-surveys could potentially introduce biases due to differences in the composition of the samples and their characteristics and their representativeness of the overall population. This poses a threat to both internal and external validity. The smaller sub-surveys may not fully capture the diversity and variability present in the population, thereby limiting the generalizability of the results. Moreover, the reduction in sample size per smaller sub-survey has implications for the statistical power of our analysis. With smaller sample sizes, it becomes more challenging to detect significant effects or relationships, which in turn affects the reliability and precision of the results. To address some of these threats, we implemented a randomization process to determine the order in which the smaller sub-surveys were presented to participants. We ensured consistency of the content of the sub-surveys and we took care to split the user stories evenly across the three sub-surveys. This included an equal number of user stories created using ChatGPT, user stories created manually by the experiment's participants, and user stories for project descriptions 1 and 2. During the pilot test, we directly asked respondents whether they considered all three sub-surveys to be balanced and identical. All participants concurred that the sub-surveys were indeed balanced and identical, providing some assurance regarding the consistency and fairness of the split.

VII. Conclusion

In this paper, we explored ChatGPT's (GPT-3.5) abilities to create user stories as well as its weaknesses and strengths. We conducted an experiment with six participants in which they created user stories manually and with ChatGPT. We gathered 60 user stories in total from the experiment, 30 created manually and 30 with ChatGPT. We then handed out survey 2 and 27 participants rated these user stories based on INVEST grid.

The results from survey 2 showed that although the current state of ChatGPT cannot surpass humans at creating user stories, it can generate user stories with similar quality. GPT-3.5 on average is faster than humans at creating user stories. However, GPT-3.5 lacks creativity and is prone to creating incomplete user stories, which is why it cannot automate the user story creation yet and human involvement is necessary. Prompts that provide feedback help ChatGPT revise the initial user stories it generated and satisfy the user's expectations. Conducting more studies such as [21] is necessary to identify prompt patterns that can minimize the mistakes ChatGPT makes and reach its full potential. Our findings are akin to those of [8] [9] [20] [21], which suggest that ChatGPT or similar NLP applications have the potential to improve the user story creation by assisting or partially automating different aspects of the process, but more empirical research is required. Conducting studies within the industry, such as introducing ChatGPT into a software development team's workflow is necessary to explore how ChatGPT performs in a real world scenario.

References

[1] M. Cohn, User stories applied for Agile Software Development, Boston: Addison-Wesley, 2015.

[2] G. Lucassen, F. Dalpiaz, J. M. Werf, and S. Brinkkemper, "The use and effectiveness of user stories in practice," Requirements Engineering: Foundation for Software Quality, 2016, pp. 205–222.

[3] S. Biswas, "ChatGPT and the future of medical writing," Radiology, 2023.

[4] W. Jiao, W. Wang, J. Huang, X. Wang, and Z. Tu, "Is ChatGPT A Good Translator? A Preliminary Study," arXiv:2301.08745 [cs.CL], Jan 2023.

[5] I. K. Raharjana, D. Siahaan, and C. Fatichah, "User Stories and Natural Language Processing: A Systematic Literature Review," IEEE Access, vol. 9, pp. 53811-53826, 2021.

[6] J. Schulman et al.. "Introducing ChatGPT." OpenAI.com. https://openai.com/blog/chatgpt (accessed Feb. 26, 2023)

[7] S. J. Stratton, "Population Research: Convenience Sampling Strategies," Prehospital and Disaster Medicine, vol. 36, no. 4, pp. 373–374, 2021.

[8] G. Lucassen, F. Dalpiaz, J. M. van der Werf, and S. Brinkkemper, "Improving agile requirements: The quality user story framework and tool," Requirements Engineering, vol. 21, no. 3, pp. 383–403, 2016.

[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," 2015 IEEE 23rd International Requirements Engineering Conference, Ottawa, ON, Canada, 2015, pp. 126-135.

[10] B. Paech and C. Martell, "Innovations for Requirement Analysis. From Stakeholders' Needs to Formal Designs," 14th Monterey Workshop, Monterey, CA, USA, 2007, pp. 103–124.

[11] R. Gacitua, P. Sawyer, and V. Gervasi, "On the Effectiveness of Abstraction Identification in Requirements Engineering," 2010 18th IEEE International Requirements Engineering Conference, Sydney, NSW, Australia, 2010, pp. 5-14.

[12] D. Berry, R. Gacitua, P. Sawyer, and S. F. Tjong, "The case for Dumb Requirements Engineering Tools," Requirements Engineering: Foundation for Software Quality, 2012, pp. 211–217.

[13] L. Buglione and A. Abran, "Improving the User Story Agile Technique Using the INVEST Criteria," 2013 Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement, Ankara, Turkey, 2013, pp. 49-53.

[14] J. Linaker, S. M. Sulaman, M. Höst, and R. M. de Mello, "Guidelines for conducting surveys in software engineering," Department of Computer Science, Lund University, 2015.

[15] M. I. Alhojailan, "Thematic Analysis: A Critical Review of its process and evaluation," West East Journal of Social Sciences, vol. 1, no. 1, 2012, pp. 39–47.

[16] A. Majumdar, "Thematic Analysis in qualitative research," Research Anthology on Innovative Research Methodologies and Utilization Across Multiple Disciplines, 2022, pp. 604–622.

[17] R. Feldt and A. Magazinius, "Validity threats in empirical software engineering research-an initial survey," InSeke, 2010, pp. 374-379.

[18] A. J. Onwuegbuzie and N. L. Leech, "Validity and qualitative research: An Oxymoron?," Quality & Quantity, vol. 41, no. 2, 2006, pp. 233–249.

[19] G. Keren, "Between- or Within-Subjects Design: A Methodological Dilemma," A Handbook for Data Analysis in the behavioral sciences, vol. 1, 2014, pp. 257–272.

[20] J. Zhang, Y. Chen, N. Niu, and C. Liu, "A Preliminary Evaluation of ChatGPT in Requirements Information Retrieval," 2023, arXiv:2304.12562.

[21] J. White, S. Hays, Q. Fu, J. Spencer-Smith, and D. C. Schmidt, "ChatGPT Prompt Patterns for Improving Code Quality, Refactoring, Requirements Elicitation, and Software Design," 2023, arXiv:2303.07839.

[22] OpenAI, "GPT-4 Technical Report," OpenAI, San Francisco, California, U.S., 2023. Accessed: May, 16, 2023. [Online]. Available: https://arxiv.org/pdf/2303.08774.pdf

Appendices

Appendix A

Project Description A

Goodwill's Stand Against Violence Everyday (SAVE) division of the Gun Violence Intervention (GVI) organization aims to reduce gun violence in South Bend, Indiana, USA. Their workers are community outreach workers that perform a number of services, including counseling, providing services, setting up community events, and interfacing with law enforcement.

When a gun violence event occurs, they will get some basic information from law enforcement, then send one or more workers to the incident. Their aim is to offer counseling and services to family and friends, and to find known contacts which have the potential for retaliation, offering counseling, options and prevention services. They also perform prevention activities based on reports from law enforcement or other community services to speak with people for whom gun violence is likely. The workers can suggest various services like job placement, education, food support, needed items and other community services. The organization also organizes events with food and celebration to raise awareness and perform community outreach.

The SAVE team needs an app to help them with at least some of these tasks. When the SAVE team is in the field, they need to enter and look up information. They are given some information about the incident and need to be able to look this up in the field (what happened, who, when, where, known affiliates, contacts, gang affiliations, etc.) and to be able to update this information, including what interventions were made (counseling, services offered). The team needs to see a history for each of their contacts, including basic information, services, affiliations. This information often changes over time. The app must be easy for the workers to use in the field. Privacy and confidentiality are important, although SAVE receives information from law enforcement, they don't share back the information they collect. Worker safety is critical.

Project Description B

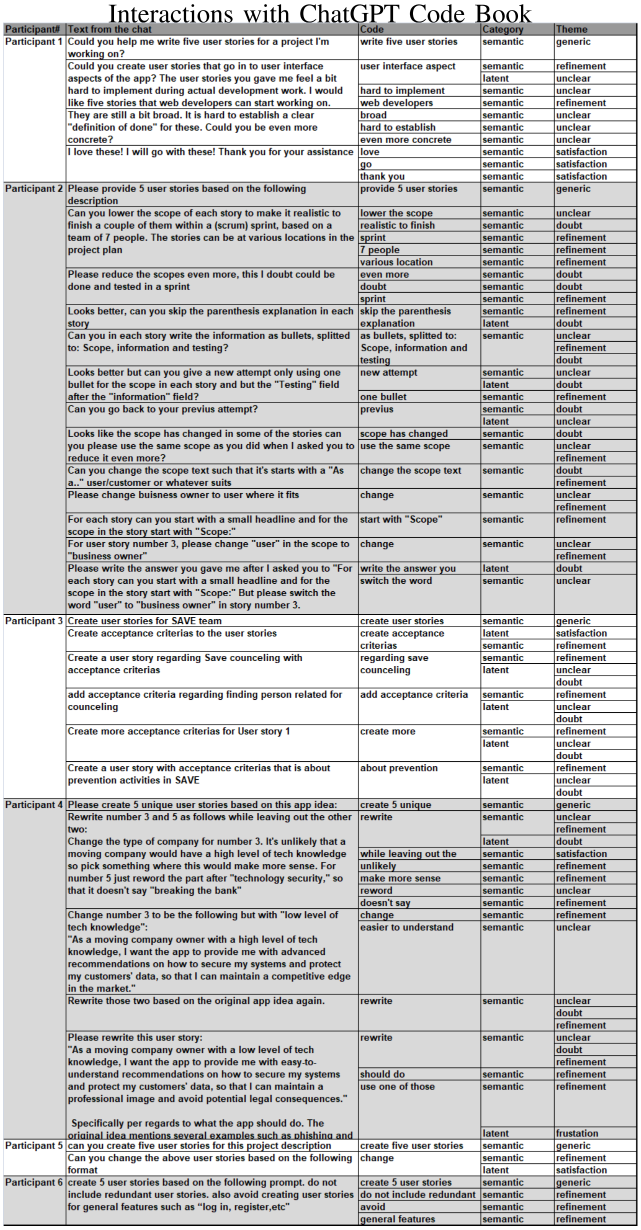
Nearly all small businesses today make use of technology; however, many business owners are not tech savvy. For example, small businesses such as moving companies, hair salons, tailors, cleaners, builders, etc. may not provide work which has a technical component, but must manage bookings, communication, and often finance using technology. Often this technology is using a mobile or personal computer and is managed by the business owner. Some businesses have the funds to hire tech help, but many do not. This case focuses on “an interactive assistant to help small and medium enterprises (SMEs) in Switzerland without dedicated IT support, to assess security risks and provide resulting guidance.”

Technology security threats such as phishing, various forms of hacking, social engineering, viruses, and malware can have devastating effects on small businesses, particularly if the business owners and managers do not well understand these threats. The business owners can lose data, customer information, and business. They could be legally liable.

We aim to create an app to help small businesses consider security threats. The site/app could assess their situation and use of technology and make suggestions. It could provide information and recommendations. However, the technology awareness of users may vary from low to high, and the trust of users in the site/app is an issue. Business owners are reluctant to provide direct access to their systems, and such access may in itself create security issues. Privacy, time pressures and liability are all considerations.

Appendix B

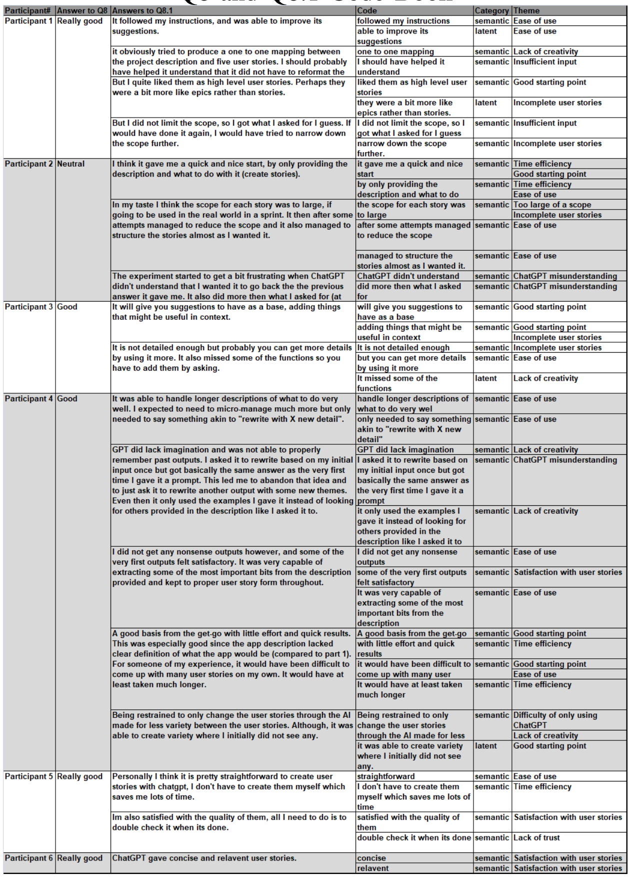
Interactions with ChatGPT Code Book



Participant#	Text from the chat	Code	Category	Theme
Participant 1	Could you help me write five user stories for a project I'm working on?	write five user stories	semantic	generic
	Could you create user stories that go in to user interface aspects of the app? The user stories you gave me feel a bit hard to implement during actual development work. I would like five stories that web developers can start working on.	user interface aspect	semantic	refinement
			latent	unclear
		hard to implement	semantic	unclear
		web developers	semantic	refinement
	They are still a bit broad. It is hard to establish a clear "definition of done" for these. Could you be even more concrete?	broad	semantic	unclear
		hard to establish	semantic	unclear
		even more concrete	semantic	unclear
	I love these! I will go with these! Thank you for your assistance	love	semantic	satisfaction
		go	semantic	satisfaction
		thank you	semantic	satisfaction
Participant 2	Please provide 5 user stories based on the following description	provide 5 user stories	semantic	generic
	Can you lower the scope of each story to make it realistic to finish a couple of them within a (scrum) sprint, based on a team of 7 people. The stories can be at various locations in the project plan	lower the scope	semantic	unclear
		realistic to finish	semantic	doubt
		sprint	semantic	refinement
		7 people	semantic	refinement
		various location	semantic	refinement
	Please reduce the scopes even more, this I doubt could be done and tested in a sprint	even more	semantic	doubt
		doubt	semantic	doubt
		sprint	semantic	refinement
	Looks better, can you skip the parenthesis explanation in each story	skip the parenthesis explanation	semantic	refinement
			latent	doubt
	Can you in each story write the information as bullets, splitted to: Scope, information and testing?	as bullets, splitted to: Scope, information and testing	semantic	unclear
				refinement
				doubt
	Looks better but can you give a new attempt only using one bullet for the scope in each story and but the "Testing" field after the "information" field?	new attempt	semantic	unclear
			latent	doubt
		one bullet	semantic	refinement
	Can you go back to your previus attempt?	previus	semantic	doubt
			latent	unclear
	Looks like the scope has changed in some of the stories can you please use the same scope as you did when I asked you to reduce it even more?	scope has changed	semantic	doubt
		use the same scope	semantic	unclear
				refinement
	Can you change the scope text such that it's starts with a "As a.." user/customer or whatever suits	change the scope text	semantic	doubt
				refinement
	Please change buisness owner to user where it fits	change	semantic	unclear
				refinement
	For each story can you start with a small headline and for the scope in the story start with "Scope:"	start with "Scope"	semantic	refinement
	For user story number 3, please change "user" in the scope to "business owner"	change	semantic	unclear
				refinement
	Please write the answer you gave me after I asked you to "For each story can you start with a small headline and for the scope in the story start with "Scope:" But please switch the word "user" to "business owner" in story number 3.	write the answer you	latent	doubt
		switch the word	semantic	unclear
Participant 3	Create user stories for SAVE team	create user stories	semantic	generic
	Create acceptance criterias to the user stories	create acceptance criterias	latent	satisfaction
			semantic	refinement
	Create a user story regarding Save counceling with acceptance criterias	regarding save counceling	semantic	refinement
			latent	unclear
				doubt
	add acceptance criteria regarding finding person related for counceling	add acceptance criteria	semantic	refinement
			latent	unclear
				doubt
	Create more acceptance criterias for User story 1	create more	semantic	refinement
			latent	unclear
				doubt
	Create a user story with acceptance criterias that is about prevention activities in SAVE	about prevention	semantic	refinement
			latent	unclear
				doubt
Participant 4	Please create 5 unique user stories based on this app idea:	create 5 unique	semantic	generic
	Rewrite number 3 and 5 as follows while leaving out the other two: Change the type of company for number 3. It's unlikely that a moving company would have a high level of tech knowledge so pick something where this would make more sense. For number 5 just reword the part after "technology security," so that it doesn't say "breaking the bank"	rewrite	semantic	unclear
				refinement
			latent	doubt
		while leaving out the	semantic	satisfaction
		unlikely	semantic	refinement
		make more sense	semantic	refinement
		reword	semantic	unclear
		doesn't say	semantic	refinement
	Change number 3 to be the following but with "low level of tech knowledge": "As a moving company owner with a high level of tech knowledge, I want the app to provide me with advanced recommendations on how to secure my systems and protect my customers' data, so that I can maintain a competitive edge in the market."	change	semantic	refinement
		easier to understand	semantic	unclear
	Rewrite those two based on the original app idea again.	rewrite	semantic	unclear
				doubt
				refinement
	Please rewrite this user story: "As a moving company owner with a low level of tech knowledge, I want the app to provide me with easy-to-understand recommendations on how to secure my systems and protect my customers' data, so that I can maintain a professional image and avoid potential legal consequences." Specifically per regards to what the app should do. The original idea mentions several examples such as phishing and	rewrite	semantic	unclear
				doubt
				refinement
		should do	semantic	refinement
		use one of those	semantic	refinement
			latent	frustation
Participant 5	can you create five user stories for this project description	create five user stories	semantic	generic
	Can you change the above user stories based on the following format	change	semantic	refinement
			latent	satisfaction
Participant 6	create 5 user stories based on the following prompt. do not include redundant user stories. also avoid creating user stories for general features such as “log in, register,etc”	create 5 user stories	semantic	generic
		do not include redundant	semantic	refinement
		avoid	semantic	refinement
		general features	semantic	refinement

Appendix C

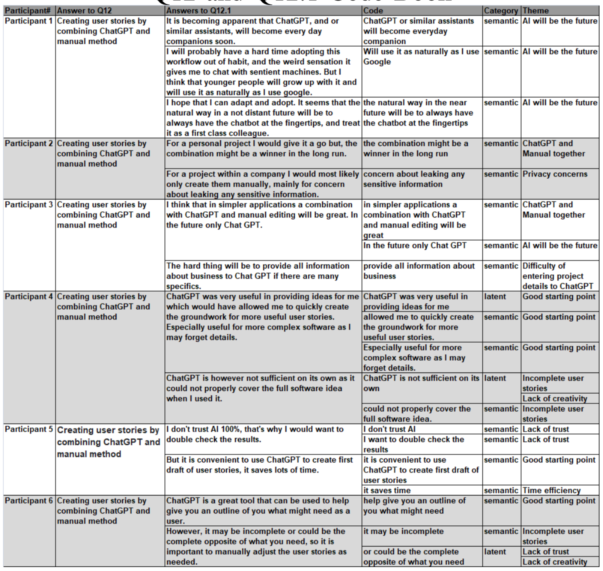
Q8 and Q8.1 Code Book



Participant#	Answer to Q8	Answers to Q8.1	Code	Category	Theme
Participant 1	Really good	It followed my instructions, and was able to improve its suggestions.	followed my instructions	semantic	Ease of use
			able to improve its suggestions	latent	Ease of use
		it obviously tried to produce a one to one mapping between the project description and five user stories. I should probably have helped it understand that it did not have to reformat the	one to one mapping	semantic	Lack of creativity
			I should have helped it understand	semantic	Insufficient input
		But I quite liked them as high level user stories. Perhaps they were a bit more like epics rather than stories.	liked them as high level user stories	semantic	Good starting point
			they were a bit more like epics rather than stories.	latent	Incomplete user stories
		But I did not limit the scope, so I got what I asked for I guess. If would have done it again, I would have tried to narrow down the scope further.	I did not limit the scope, so I got what I asked for I guess	semantic	Insufficient input
			narrow down the scope further.	semantic	Incomplete user stories
Participant 2	Neutral	I think it gave me a quick and nice start, by only providing the description and what to do with it (create stories).	it gave me a quick and nice start	semantic	Time efficiency Good starting point
			by only providing the description and what to do	semantic	Time efficiency Ease of use
		In my taste I think the scope for each story was to large, if going to be used in the real world in a sprint. It then after some attempts managed to reduce the scope and it also managed to structure the stories almost as I wanted it.	the scope for each story was to large	semantic	Too large of a scope Incomplete user stories
			after some attempts managed to reduce the scope	semantic	Ease of use
			managed to structure the stories almost as I wanted it.	semantic	Ease of use
		The experiment started to get a bit frustrating when ChatGPT didn't understand that I wanted it to go back the the previous answer it gave me. It also did more then what I asked for (at	ChatGPT didn't understand	semantic	ChatGPT misunderstanding
			did more then what I asked for	semantic	ChatGPT misunderstanding
Participant 3	Good	It will give you suggestions to have as a base, adding things that might be useful in context.	will give you suggestions to have as a base	semantic	Good starting point
			adding things that might be useful in context	semantic	Good starting point Incomplete user stories
		It is not detailed enough but probably you can get more details by using it more. It also missed some of the functions so you have to add them by asking.	It is not detailed enough	semantic	Incomplete user stories
			but you can get more details by using it more	semantic	Ease of use
			It missed some of the functions	latent	Lack of creativity
Participant 4	Good	It was able to handle longer descriptions of what to do very well. I expected to need to micro-manage much more but only needed to say something akin to "rewrite with X new detail".	handle longer descriptions of what to do very wel	semantic	Ease of use
			only needed to say something akin to "rewrite with X new detail"	semantic	Ease of use
		GPT did lack imagination and was not able to properly remember past outputs. I asked it to rewrite based on my initial input once but got basically the same answer as the very first time I gave it a prompt. This led me to abandon that idea and to just ask it to rewrite another output with some new themes. Even then it only used the examples I gave it instead of looking for others provided in the description like I asked it to.	GPT did lack imagination	semantic	Lack of creativity
			I asked it to rewrite based on my initial input once but got basically the same answer as the very first time I gave it a prompt	semantic	ChatGPT misunderstanding
			it only used the examples I gave it instead of looking for others provided in the description like I asked it to	semantic	Lack of creativity
		I did not get any nonsense outputs however, and some of the very first outputs felt satisfactory. It was very capable of extracting some of the most important bits from the description provided and kept to proper user story form throughout.	I did not get any nonsense outputs	semantic	Ease of use
			some of the very first outputs felt satisfactory	semantic	Satisfaction with user stories
			It was very capable of extracting some of the most important bits from the description	semantic	Ease of use
		A good basis from the get-go with little effort and quick results. This was especially good since the app description lacked clear definition of what the app would be (compared to part 1). For someone of my experience, it would have been difficult to come up with many user stories on my own. It would have at least taken much longer.	A good basis from the get-go	semantic	Good starting point
			with little effort and quick results	semantic	Time efficiency
			it would have been difficult to come up with many user	semantic	Good starting point Ease of use
			It would have at least taken much longer	semantic	Time efficiency
		Being restrained to only change the user stories through the AI made for less variety between the user stories. Although, it was able to create variety where I initially did not see any.	Being restrained to only change the user stories through the AI made for less	semantic	Difficulty of only using ChatGPT Lack of creativity
			it was able to create variety where I initially did not see any.	latent	Good starting point
Participant 5	Really good	Personally I think it is pretty straightforward to create user stories with chatgpt, I don't have to create them myself which saves me lots of time.	straightforward	semantic	Ease of use
			I don't have to create them myself which saves me lots of time	semantic	Time efficiency
		Im also satisfied with the quality of them, all I need to do is to double check it when its done.	satisfied with the quality of them	semantic	Satisfaction with user stories
			double check it when its done	semantic	Lack of trust
Participant 6	Really good	ChatGPT gave concise and relavent user stories.	concise	semantic	Satisfaction with user stories
			relavent	semantic	Satisfaction with user stories

Appendix D

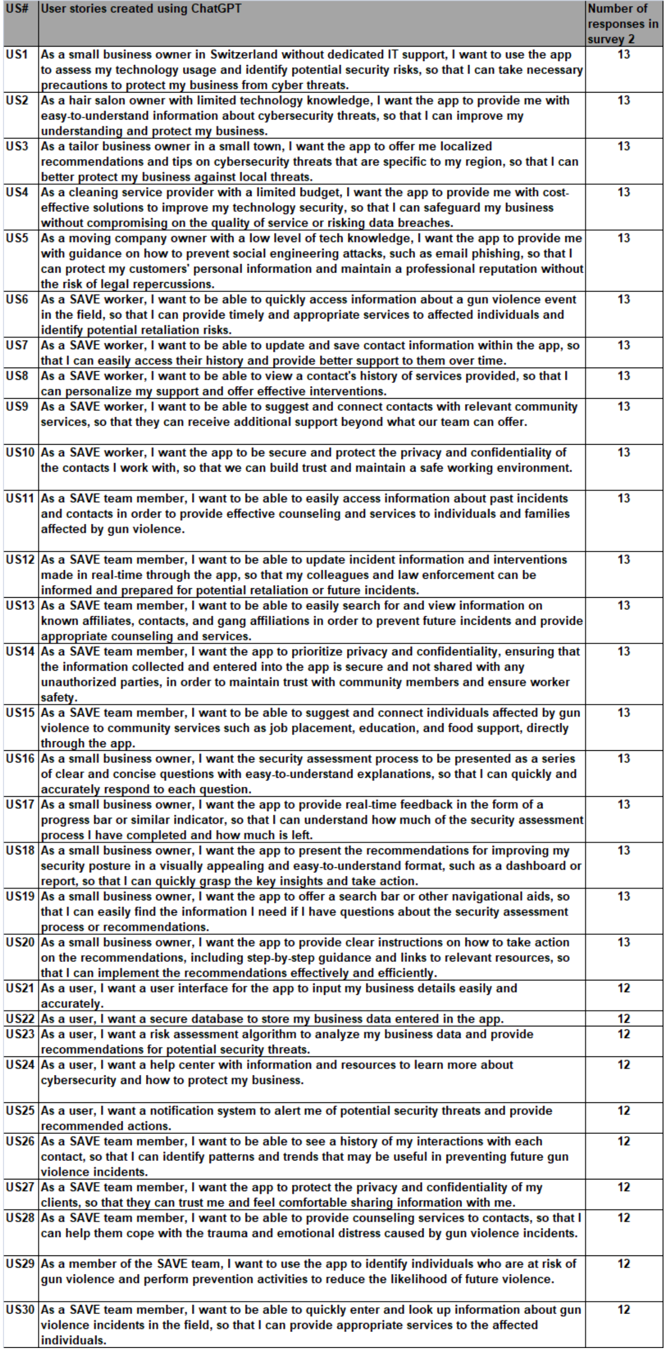
Q12 and Q12.1 Code Book



Participant#	Answer to Q12	Answers to Q12.1	Code	Category	Theme
Participant 1	Creating user stories by combining ChatGPT and manual method	It is becoming apparent that ChatGPT, and or similar assistants, will become every day companions soon.	ChatGPT or similar assistants will become everyday companion	semantic	AI will be the future
		I will probably have a hard time adopting this workflow out of habit, and the weird sensation it gives me to chat with sentient machines. But I think that younger people will grow up with it and will use it as naturally as I use google.	Will use it as naturally as I use Google	semantic	AI will be the future
		I hope that I can adapt and adopt. It seems that the natural way in a not distant future will be to always have the chatbot at the fingertips, and treat it as a first class colleague.	the natural way in the near future will be to always have the chatbot at the fingertips	semantic	AI will be the future
Participant 2	Creating user stories by combining ChatGPT and manual method	For a personal project I would give it a go but, the combination might be a winner in the long run.	the combination might be a winner in the long run	semantic	ChatGPT and Manual together
		For a project within a company I would most likely only create them manually, mainly for concern about leaking any sensitive information.	concern about leaking any sensitive information	semantic	Privacy concerns
Participant 3	Creating user stories by combining ChatGPT and manual method	I think that in simpler applications a combination with ChatGPT and manual editing will be great. In the future only Chat GPT.	in simpler applications a combination with ChatGPT and manual editing will be great	semantic	ChatGPT and Manual together
			In the future only Chat GPT	semantic	AI will be the future
		The hard thing will be to provide all information about business to Chat GPT if there are many specifics.	provide all information about business	semantic	Difficulty of entering project details to ChatGPT
Participant 4	Creating user stories by combining ChatGPT and manual method	ChatGPT was very useful in providing ideas for me which would have allowed me to quickly create the groundwork for more useful user stories. Especially useful for more complex software as I may forget details.	ChatGPT was very useful in providing ideas for me	latent	Good starting point
			allowed me to quickly create the groundwork for more useful user stories.	semantic	Good starting point
			Especially useful for more complex software as I may forget details.	semantic	Good starting point
		ChatGPT is however not sufficient on its own as it could not properly cover the full software idea when I used it.	ChatGPT is not sufficient on its own	latent	Incomplete user stories Lack of creativity
			could not properly cover the full software idea.	semantic	Incomplete user stories
Participant 5	Creating user stories by combining ChatGPT and manual method	I don't trust AI 100%, that's why I would want to double check the results.	I don't trust AI	semantic	Lack of trust
			I want to double check the results	semantic	Lack of trust
		But it is convenient to use ChatGPT to create first draft of user stories, it saves lots of time.	it is convenient to use ChatGPT to create first draft of user stories	semantic	Good starting point
			it saves time	semantic	Time efficiency
Participant 6	Creating user stories by combining ChatGPT and manual method	ChatGPT is a great tool that can be used to help give you an outline of you what might need as a user.	help give you an outline of you what might need	semantic	Good starting point
		However, it may be incomplete or could be the complete opposite of what you need, so it is important to manually adjust the user stories as needed.	it may be incomplete	semantic	Incomplete user stories
			or could be the complete opposite of what you need	latent	Lack of trust Lack of creativity

Appendix E

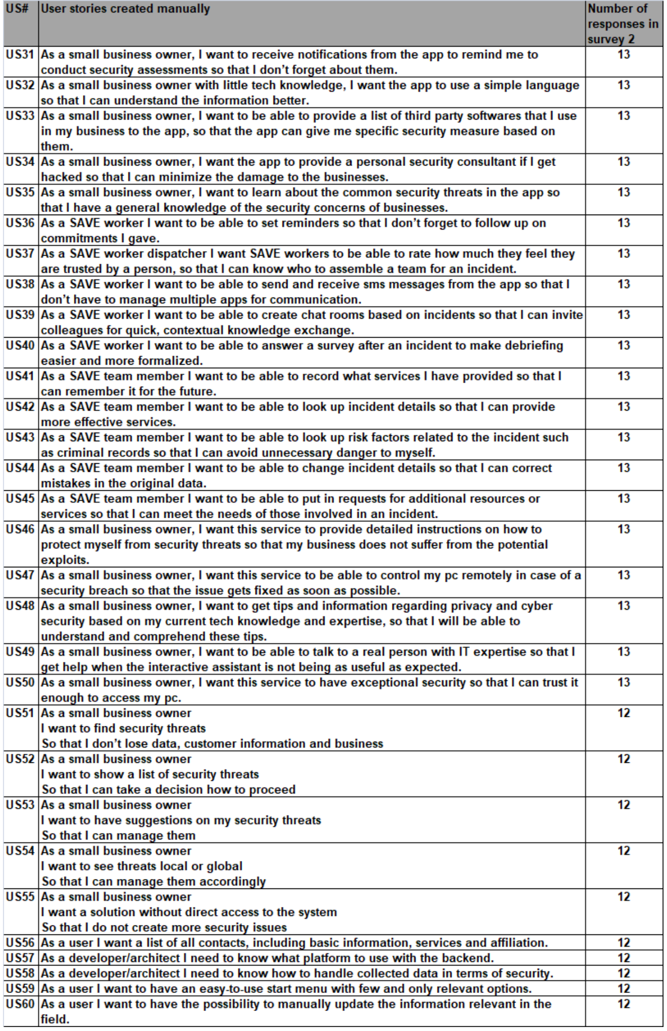
User Stories (ChatGPT)



US#	User stories created using ChatGPT	Number of responses in survey 2
US1	As a small business owner in Switzerland without dedicated IT support, I want to use the app to assess my technology usage and identify potential security risks, so that I can take necessary precautions to protect my business from cyber threats.	13
US2	As a hair salon owner with limited technology knowledge, I want the app to provide me with easy-to-understand information about cybersecurity threats, so that I can improve my understanding and protect my business.	13
US3	As a tailor business owner in a small town, I want the app to offer me localized recommendations and tips on cybersecurity threats that are specific to my region, so that I can better protect my business against local threats.	13
US4	As a cleaning service provider with a limited budget, I want the app to provide me with cost-effective solutions to improve my technology security, so that I can safeguard my business without compromising on the quality of service or risking data breaches.	13
US5	As a moving company owner with a low level of tech knowledge, I want the app to provide me with guidance on how to prevent social engineering attacks, such as email phishing, so that I can protect my customers' personal information and maintain a professional reputation without the risk of legal repercussions.	13
US6	As a SAVE worker, I want to be able to quickly access information about a gun violence event in the field, so that I can provide timely and appropriate services to affected individuals and identify potential retaliation risks.	13
US7	As a SAVE worker, I want to be able to update and save contact information within the app, so that I can easily access their history and provide better support to them over time.	13
US8	As a SAVE worker, I want to be able to view a contact's history of services provided, so that I can personalize my support and offer effective interventions.	13
US9	As a SAVE worker, I want to be able to suggest and connect contacts with relevant community services, so that they can receive additional support beyond what our team can offer.	13
US10	As a SAVE worker, I want the app to be secure and protect the privacy and confidentiality of the contacts I work with, so that we can build trust and maintain a safe working environment.	13
US11	As a SAVE team member, I want to be able to easily access information about past incidents and contacts in order to provide effective counseling and services to individuals and families affected by gun violence.	13
US12	As a SAVE team member, I want to be able to update incident information and interventions made in real-time through the app, so that my colleagues and law enforcement can be informed and prepared for potential retaliation or future incidents.	13
US13	As a SAVE team member, I want to be able to easily search for and view information on known affiliates, contacts, and gang affiliations in order to prevent future incidents and provide appropriate counseling and services.	13
US14	As a SAVE team member, I want the app to prioritize privacy and confidentiality, ensuring that the information collected and entered into the app is secure and not shared with any unauthorized parties, in order to maintain trust with community members and ensure worker safety.	13
US15	As a SAVE team member, I want to be able to suggest and connect individuals affected by gun violence to community services such as job placement, education, and food support, directly through the app.	13
US16	As a small business owner, I want the security assessment process to be presented as a series of clear and concise questions with easy-to-understand explanations, so that I can quickly and accurately respond to each question.	13
US17	As a small business owner, I want the app to provide real-time feedback in the form of a progress bar or similar indicator, so that I can understand how much of the security assessment process I have completed and how much is left.	13
US18	As a small business owner, I want the app to present the recommendations for improving my security posture in a visually appealing and easy-to-understand format, such as a dashboard or report, so that I can quickly grasp the key insights and take action.	13
US19	As a small business owner, I want the app to offer a search bar or other navigational aids, so that I can easily find the information I need if I have questions about the security assessment process or recommendations.	13
US20	As a small business owner, I want the app to provide clear instructions on how to take action on the recommendations, including step-by-step guidance and links to relevant resources, so that I can implement the recommendations effectively and efficiently.	13
US21	As a user, I want a user interface for the app to input my business details easily and accurately.	12
US22	As a user, I want a secure database to store my business data entered in the app.	12
US23	As a user, I want a risk assessment algorithm to analyze my business data and provide recommendations for potential security threats.	12
US24	As a user, I want a help center with information and resources to learn more about cybersecurity and how to protect my business.	12
US25	As a user, I want a notification system to alert me of potential security threats and provide recommended actions.	12
US26	As a SAVE team member, I want to be able to see a history of my interactions with each contact, so that I can identify patterns and trends that may be useful in preventing future gun violence incidents.	12
US27	As a SAVE team member, I want the app to protect the privacy and confidentiality of my clients, so that they can trust me and feel comfortable sharing information with me.	12
US28	As a SAVE team member, I want to be able to provide counseling services to contacts, so that I can help them cope with the trauma and emotional distress caused by gun violence incidents.	12
US29	As a member of the SAVE team, I want to use the app to identify individuals who are at risk of gun violence and perform prevention activities to reduce the likelihood of future violence.	12
US30	As a SAVE team member, I want to be able to quickly enter and look up information about gun violence incidents in the field, so that I can provide appropriate services to the affected individuals.	12

Appendix F

User Stories (manual)



US#	User stories created manually	Number of responses in survey 2
US31	As a small business owner, I want to receive notifications from the app to remind me to conduct security assessments so that I don’t forget about them.	13
US32	As a small business owner with little tech knowledge, I want the app to use a simple language so that I can understand the information better.	13
US33	As a small business owner, I want to be able to provide a list of third party softwares that I use in my business to the app, so that the app can give me specific security measure based on them.	13
US34	As a small business owner, I want the app to provide a personal security consultant if I get hacked so that I can minimize the damage to the businesses.	13
US35	As a small business owner, I want to learn about the common security threats in the app so that I have a general knowledge of the security concerns of businesses.	13
US36	As a SAVE worker I want to be able to set reminders so that I don’t forget to follow up on commitments I gave.	13
US37	As a SAVE worker dispatcher I want SAVE workers to be able to rate how much they feel they are trusted by a person, so that I can know who to assemble a team for an incident.	13
US38	As a SAVE worker I want to be able to send and receive sms messages from the app so that I don’t have to manage multiple apps for communication.	13
US39	As a SAVE worker I want to be able to create chat rooms based on incidents so that I can invite colleagues for quick, contextual knowledge exchange.	13
US40	As a SAVE worker I want to be able to answer a survey after an incident to make debriefing easier and more formalized.	13
US41	As a SAVE team member I want to be able to record what services I have provided so that I can remember it for the future.	13
US42	As a SAVE team member I want to be able to look up incident details so that I can provide more effective services.	13
US43	As a SAVE team member I want to be able to look up risk factors related to the incident such as criminal records so that I can avoid unnecessary danger to myself.	13
US44	As a SAVE team member I want to be able to change incident details so that I can correct mistakes in the original data.	13
US45	As a SAVE team member I want to be able to put in requests for additional resources or services so that I can meet the needs of those involved in an incident.	13
US46	As a small business owner, I want this service to provide detailed instructions on how to protect myself from security threats so that my business does not suffer from the potential exploits.	13
US47	As a small business owner, I want this service to be able to control my pc remotely in case of a security breach so that the issue gets fixed as soon as possible.	13
US48	As a small business owner, I want to get tips and information regarding privacy and cyber security based on my current tech knowledge and expertise, so that I will be able to understand and comprehend these tips.	13
US49	As a small business owner, I want to be able to talk to a real person with IT expertise so that I get help when the interactive assistant is not being as useful as expected.	13
US50	As a small business owner, I want this service to have exceptional security so that I can trust it enough to access my pc.	13
US51	As a small business owner I want to find security threats So that I don’t lose data, customer information and business	12
US52	As a small business owner I want to show a list of security threats So that I can take a decision how to proceed	12
US53	As a small business owner I want to have suggestions on my security threats So that I can manage them	12
US54	As a small business owner I want to see threats local or global So that I can manage them accordingly	12
US55	As a small business owner I want a solution without direct access to the system So that I do not create more security issues	12
US56	As a user I want a list of all contacts, including basic information, services and affiliation.	12
US57	As a developer/architect I need to know what platform to use with the backend.	12
US58	As a developer/architect I need to know how to handle collected data in terms of security.	12
US59	As a user I want to have an easy-to-use start menu with few and only relevant options.	12
US60	As a user I want to have the possibility to manually update the information relevant in the field.	12