An Aggregated Dataset of Agile User Stories and Use Case Taxonomy for AI-Driven Research

Abdulrahim Alhaizaey[1], Majed Al-Mashari[2]
Information Systems Department, King Khalid University, Abha, Saudi Arabia[1]
Information Systems Department, King Saud University, Riyadh, Saudi Arabia[1, 2]

***Abstract*—Agile methodologies are considered revolutionary approaches in the development of systems and software. With the rapid advancement of artificial intelligence, natural language processing, and large language models, there is an increasing demand for high-quality datasets to support the design and development of intelligent, practical, and effective automation tools. However, researchers in Agile Requirements Engineering face significant challenges due to the limited availability of datasets, particularly those involving user stories. This paper presents a dataset of over 10K user stories collected from academic sources and publicly accessible online repositories. These stories represent requirements formulated in accordance with Agile principles. The process of collecting and classifying data, as well as its use in a prior research project focused on identifying non-functional requirements, is described. The dataset was validated with substantial inter-annotator agreement and has been successfully employed in prior experiments, where a fine-tuned pre-trained language model achieved F1 scores above 93% in classifying non-functional requirements. Additionally, a structured taxonomy of potential research and practical use cases for this dataset is proposed, aiming to support researchers and practitioners in areas such as requirements analysis, automated generative tasks using generative language models, model development, and educational purposes.**

***Keywords*—Agile software development; requirements engineering; user stories; natural language processing; datasets; large language models; generative language models**

I. Introduction

Agile development methodologies have seen widespread adoption in recent years, leading to the prominent use of user stories as a primary tool for expressing software requirements [1], [2]. Lucassen et al. [3] found that employing user stories enhances the productivity of agile practitioners. They also reported that most practitioners who implement user stories adhere to the original template proposed by Cohn in [4]. The popularity, simplicity, and concise structure of user stories make them ideal candidates for promising natural language processing (NLP) analysis and research [5]. This widespread acceptance and straightforward nature of user stories motivate an appealing focus for studies and applications that support requirements engineers and agile practitioners. Given that user stories are inherently based on natural language, recent advancements in NLP have positioned them as ideal candidates for automated analysis and understanding through artificial intelligence and machine learning (ML) techniques [5]. However, the effectiveness of such models largely depends on the availability of high-quality, large-scale training data [6], making the development of structured user story datasets a pressing need.

Despite this need, the current research landscape reveals a significant scarcity of comprehensive user story datasets. These datasets are limited in availability, scattered across the literature, and often lack standardization. Raharjana et al. [7] highlighted several key challenges associated with existing user stories datasets. Among which is the limited publicly available data, making it difficult for researchers to obtain sufficient and representative samples. In addition, the high heterogeneity of the available datasets results from variations in story formats, sources, and writing styles, although many studies recommend Cohen's user story template [4]. Also, the manual labeling of the available datasets hinders scalability and consistent reuse. Such challenges create a difficulty in applying ML techniques (e.g. classification or semantic similarity analysis) due to the limitations mentioned earlier.

Therefore, creating a well-structured and annotated user story dataset is crucial for accelerating research on developing intelligent tools that can understand and analyze requirements in agile environments. In an effort to bridge this gap and motivated by one of the most widely used user stories datasets [8], although small, this study presents a dataset of over 10,000 user stories collected from both academic and publicly available sources through a carefully designed and systematic process. Unlike existing datasets such as Dalpiaz's [8], which are limited in diversity and size (1,680 samples) and lack NFR annotations, our dataset provides over 10,000 stories with ISO/IEC 25010-based labeling, ensuring broader domain coverage and improved utility for AI-driven requirements engineering (RE) research. This dataset was previously used in a separate research project [9] aimed at automating the identification of non-functional requirements (NFRs) in agile user stories using pre-trained language models (PLMs), and has been annotated accordingly.

Although the dataset was initially developed for NFR classification in two ongoing projects [9], [10], it offers broader value to the research and practitioner community, such as automated acceptance criteria generation or defect detection. It can support various agile RE tasks beyond NFR identification, including requirement specification, quality analysis, decision support, and enhancement of documentation and development tools. This paper enhances the credibility and utility of the dataset by providing a detailed description of its construction and annotation process. Moreover, it offers a comprehensive taxonomy of potential research and practical applications, thereby opening promising avenues for further exploration,

particularly in integrating artificial intelligence technologies with agile RE.

The remainder of the paper is organized as follows: Section II presents the background and related literature. Section III describes the dataset construction process, including data sources, annotation, and statistics. Section IV presents a taxonomy of research and practical use cases. Section V discusses key issues and considerations. Section VI presents the results and discussion, highlighting dataset characteristics, validation, and demonstrated utility in prior experiments. Section VII explores potential threats to validity. Finally, Section VIII concludes the paper.

II. Background and Relevant Literature

This section provides foundational background and highlights key works that support the need for structured user story datasets. Agile RE and user stories have been the focus of extensive research, especially with the rise of automation and AI tools. This section presents an overview of relevant background and key studies that inform the need for structured user story datasets. The background is divided into three areas: Agile RE and user story foundations, the application of NLP and large language models (LLMs), and an overview of existing datasets.

A. Agile RE and User Stories

Agile RE is a subfield of software engineering that focuses on eliciting, documenting, analyzing, and managing system requirements within Agile development environments such as Scrum and Extreme Programming (XP). These environments emphasize flexibility, short iterative cycles, and continuous stakeholder interaction, which necessitate lightweight and efficient techniques for capturing requirements that align with the fast-paced nature of Agile processes. One of the most widely used techniques in Agile RE is the user story [2], [11], which is a short textual representation describing a feature or functionality desired by the end user. User stories are typically written in a semi-structured natural language in a simple template as follows: "As a [*type of user*], I want [*goal*] so that [*benefit*]". For example: "As a customer, I want to transfer funds between my accounts, so that I can pay off my credit card."

User stories act as a communication bridge between development teams and stakeholders, serving as basic planning and estimation units in Agile sprints. However, their use faces challenges. Key issues include ambiguity in expression due to reliance on natural language, incompleteness or imprecision in capturing requirements [12], and difficulty representing non-functional requirements (e.g. security or performance) using such a simple textual format [13]. These challenges have motivated researchers to develop tools and techniques, including artificial intelligence and NLP approaches, recently, to improve the formulation, analysis, and understanding of user stories more efficiently.

B. NLP and LLM for User Stories

Raharjana et al. [7] conducted a systematic review of studies applying NLP techniques to Agile user stories. Their study extracted the NLP tools and techniques used across the reviewed literature and proposed a conceptual and thematic classification of research goals, which included discovering defects, generating software artifacts, identifying key abstractions of user stories, and tracing links between models and stories. The review also highlighted several challenges in the literature and the need to improve the performance of the proposed solutions. In addition, there is a limited availability of datasets and an ongoing reliance on manual data labeling processes. Additional issues relate to the domain dependency of models and their limited generalizability across contexts, the complexity of correctly interpreting requirement sentences, and the persistent need for human intervention in decision-making.

Amna and Poels [14] provide a comprehensive systematic mapping of academic research on the user story techniques within Agile Software Development. The authors analyzed 186 studies published between 2001 and 2021, classifying them across five dimensions: RE activity, problem type, research outcome, research type, and publication venue. Their findings reveal a strong emphasis on system design challenges, while ambiguity and collaboration received comparatively less attention. Four major research gaps were identified, including the lack of validation and evaluation of proposed solutions and the limited exploration of cognitive and interactional aspects. This study serves as a useful reference for future research on user stories, excellently allowing researchers to clearly position their contributions.

Recently, several emerging studies showed an evolving direction of exploring the integration of LLMs and Generative AI into RE. For example, Hemmat et al. [15] recognized a steadily growing interest and significant progress in leveraging LLMs for various software engineering tasks, particularly in requirement engineering. The analysis of Arora et al. [16] showed that LLMs have the potential to enhance several RE tasks by automating, streamlining, and augmenting human capabilities because of their ability to simulate stakeholder perspectives, generate alternative requirements, address requirements quality, cross-reference with standards, and generate structured documentation. Broadly, there is increasing observation that Generative AI and LLMs are reshaping the landscape of software development in general and RE in particular. We refer to Hemmat et al. [15] and Arora et al. [16] as important references for researchers and practitioners working at the intersection of RE, including user stories, and AI. These studies help streamline and clarify the current research directions in this space and briefly highlight key considerations and challenges, especially when exploring generative AI or LLMs applied to RE tasks. Also, they primarily provide directional insights that researchers need to consider in these emerging fields.

C. Agile User Stories Datasets

In the context of Agile user story research, Dalpiaz [8] released a dataset comprising 22 backlogs with a total of 1,680 user stories covering various projects and domains. This dataset has become a fundamental resource in many studies that use NLP techniques to tackle different challenges related to user stories, such as improving how they are formulated and written [5], identifying requirements like privacy [17], [18], extracting quality attributes that could influence early architectural design decisions [19], classifying NFRs using PLMs [9], and utilizing LLMs such as LLaMA and GPT-3.5 to simulate different roles while generating acceptance criteria for requirements expressed as user stories [20].

While this dataset has served as a cornerstone in many NLP and AI-driven research efforts in Agile RE, it remains limited in scope and size. There is a pressing need for larger, more diverse, and comprehensive datasets to further advance the field. Therefore, this promising direction is encouraged to be explored, as building and sharing richer datasets can significantly accelerate data-driven RE research on user stories and improve the effectiveness of Agile requirements practices.

III. Dataset Construction Process

This section outlines the construction of the aggregated dataset, beginning with the motivation behind its creation, followed by the strategy employed to gather and process the data. The complete process, including dataset sources, labeling, and preprocessing, is summarized in Fig. 1. Section IV presents a conceptual taxonomy that highlights how researchers and practitioners may utilize this dataset across various Agile RE use cases, especially in contexts involving LLMs and generative language models.

A. Motivation for Building the Dataset

As part of previous research [9], the effectiveness of fine-tuning pre-trained language models to identify non-functional requirements (NFRs) in user stories was examined. The goal was to address a key challenge in agile RE: the frequent neglect or implicit handling of NFRs and the general inability of user stories to explicitly capture them. This led us to ask: Can user stories implicitly contain NFRs even if they are not explicitly stated? Through the investigation, it was found that some user stories do, in fact, embed implicit NFRs, a finding also noted in earlier work by other authors [19]. However, that work was limited in scope due to the small and constrained dataset used, which restricted the automation potential and performance of the applied models. This highlighted the need for a larger and more diverse dataset to support further research in this area, particularly given that machine learning models typically require substantial data to perform effectively and generalize reliably in tasks such as NFR identification. This realization motivated us initially to construct a broader dataset of user stories. While the initial focus was centered on NFR detection, the dataset is now released for broader use, recognizing its potential to support a wide range of applications beyond NFR-related tasks.

B. Dataset Collection and Sources

While research on the application of artificial intelligence in Agile RE continues to grow, one persistent challenge is the scarcity of high-quality, large-scale datasets of user story requirements. Existing datasets are often limited in size, privately held, or lack sufficient diversity [7], making them inadequate for training and evaluating modern machine learning models, or even for supporting various automation and data-driven use cases. To tackle this challenge, a step was taken to build a larger dataset by gathering user stories from various academic and public sources. These include datasets from prior research, software project repositories, and curated online content formatted in user story style. Through a literature review in another project [9] and with the guidance of another review [7], the papers that used datasets in their work were identified. The datasets were then collected, and a snowball search method was used to find related works and additional relevant papers until no more could be found. Additionally, academic literature and popular data-hosting platforms like Zenodo, GitHub, and Kaggle were searched.

Table I lists all sources used in the dataset construction, along with the number of user stories contributed by each. After removing duplicates, the final dataset comprises a total of 10,006 unique user stories. This aggregated dataset is now publicly available via a GitHub repository[1].

C. Dataset Labeling

Although the primary purpose of labeling this dataset was to support another research project on identifying NFRs in agile user stories, the labeling process was designed to be methodical and general enough to serve broader research and practical applications. Given the diversity and occasionally ambiguous nature of the collected backlogs and their respective domains, a multi-stage labeling strategy was employed, illustrated in Fig. 1.

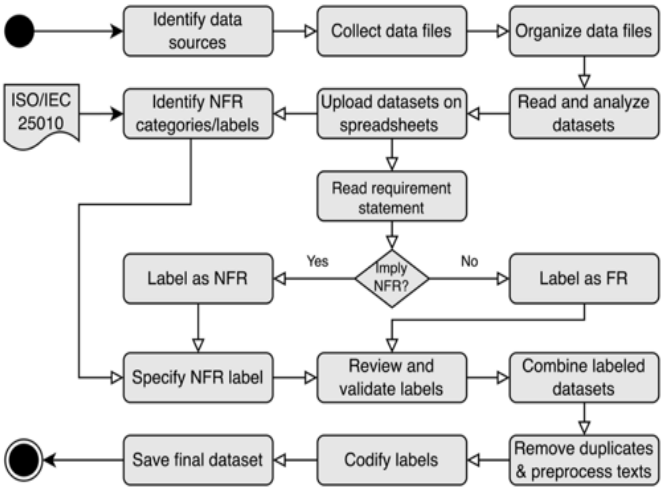

Fig. 1. Dataset construction and labeling workflow.

In the first stage, each backlog was uploaded into a spreadsheet, and a preliminary review was conducted to determine its domain and contextual background. This step was essential for understanding the scope and purpose of each

[1] https://github.com/a-re22/Agile_Req_NFRs.

backlog and for assessing whether user stories implicitly or explicitly conveyed an NFR. Once such indications were observed, the NFR labeling process was initiated. To guide the classification, the ISO/IEC 25010:2011 quality model [21], which has since been replaced by the updated version, ISO/IEC 25010:2023 [22] was used. It provides a well-structured taxonomy of software quality characteristics. It defines eight high-level NFR categories and 32 subcategories, as summarized in Table II, which offers a unified reference for annotation. Using this model, two researchers collaboratively labeled each requirement as either a functional requirement (FR) or an NFR by asking: "Does this statement explicitly or implicitly indicate an NFR?" If so, it was tagged accordingly. Each NFR-labeled story was then assigned to one of the eight NFR categories. The resulting labels were encoded in spreadsheets for consistency and clarity. The Functional Suitability category was excluded from labeling, although it is part of ISO/IEC 25010, because it was not feasible to reliably evaluate attributes related to Functional Suitability, like correctness, completeness, or appropriateness. Accordingly, FR in the dataset refers to general functional requirements rather than the functional suitability dimension.

TABLE I. SOURCES OF USER STORY SAMPLES IN THE DATASET

Data Source	Title	Description	# of samples
Dalpiaz [8]	Requirements data sets (user stories).	A collection of 22 data sets of 50+ requirements each, expressed as user stories. These were all found online or retrieved from software companies with permission to disclose. The data sets have been originally used to conduct experiments about ambiguity detection and developing Requirement Engineer Validation & Verification (REVV) tool.	1680
Murukannaiah et. al [23]	Acquiring Creative Requirements from the Crowd: Understanding the Influences of Personality and Creative Potential in Crowd RE.	This data was collected as part of research on understanding the influence of personality and creativity potential in Crowd RE. Despite the goal and other aspects in the dataset, more than 3000 SW requirements formulated as user story were extracted.	3267
Köse and Aydimer [24]	A User Story Dataset for Library and Restaurant Management Systems.	A dataset used to extract key concepts from user stories and build a knowledge graph by connecting related terms. SCOUT: Supporting User Story Completeness via Knowledge Graphs.	773
Rohmann and Paech [25]	Dataset and Application of Algorithms of 'Towards a More Set of Acceptance Criteria'.	A dataset contains SW requirements for CoronaWarn App used for experimenting a proposed algorithm aiming to a more complete set of acceptance criteria.	32
Lucassen et al. [26]	Extracting conceptual models from user stories with Visual Narrator	A dataset used to experiment an NLP-based tool (Visual Narrator) to automate extracting conceptual models from user stories.	465
Lucassen et al. [27]	Improving agile requirements: the Quality User Story framework and tool	A dataset used to evaluate Automatic Quality User Story Artisan (AQUSA) software tool- an NLP-based tool- in detecting writing quality defects within agile requirements and suggest possible remedies.	187
Prabu et al. [28]	User story-based automatic test case classification and prioritization using NLP-based deep learning	A dataset used to investigate extracting key information from user stories to facilitate automatic classification and prioritization of the associated test cases by considering factors like criticality, complexity, and dependencies.	602
Dalpiaz et al. [29]	On deriving conceptual models from user requirements: An empirical study	A dataset used to evaluate the derivation of structural conceptual models that represent the domain of the system from agile user requirements.	540
The Trident Project[2]	On the Trident Project: Architecture	A publicly available set of user stories created by a Duke University team for the Trident project which is a software to support metadata creation for their library system.	49
Online Contents	Zenodo[3], Github[4], Kaggle[5]	Searched these data hosting platforms using terms "user stories" or "agile requirements" and retrieved the publicly available datasets.	733
An ongoing project[6]	Generating requirements acceptance criteria using LLM agents and prompts engineering.	A dedicated project on using LLM agents and prompt engineering to formulate acceptance criteria and concise user stories for different roles and software domains.	1779
Duplicates	A total of 101 duplicate user stories were detected, primarily due to overlapping content across some backlogs shared between the dataset published in [8], [26], [27].		-101
Total			10,006

To ensure the reliability of the labeling process, inter-annotator agreement between the initial researcher consensus and the expert-reviewed labels was retrospectively measured using Cohen's Kappa [30]. The resulting score of 0.8335 indicates substantial agreement, affirming the consistency and validity of the annotation. Any expert-identified errors were corrected, though such cases were limited. After finalizing all labels, the datasets were merged into a unified file, duplicates were removed, and labels were converted to numeric identifiers to support downstream machine learning tasks. While this labeling effort was initially driven by a specific research focus on NFR identification, the annotated dataset is released to the broader community, recognizing its potential to support a wide variety of use cases beyond its original scope. Accordingly, each user story is stored in a separate column from its label, allowing researchers to disregard or redefine the annotations according to their own use cases. The existing labels reflect another research intent but are not binding.

[2] http://blogs.library.duke.edu/digital-collections/2009/02/13/on-the-trident-project-part-1-architecture/

[3] https://zenodo.org

[4] https://github.com

[5] https://www.kaggle.com/datasets

[6] https://github.com/a-re22/Agile_Req_NFRs

D. Dataset Preprocessing

To ensure consistency while preserving the natural linguistic structure of the user stories, only minimal preprocessing was applied. This included converting all text to lowercase and removing punctuation marks to reduce textual noise and help models focus on meaningful patterns. Also, duplicate detection and removal were performed, identifying 101 duplicated user stories, mostly resulting from overlaps among backlogs sourced from [8], [26], [27]. After eliminating these duplicates, the dataset comprised 10,006 unique entries. No stemming, lemmatization, or semantic normalization was applied, as the aim was to preserve the textual authenticity and variety of user stories. Finally, all records were merged into a unified, clean dataset, and a consistency check was performed to ensure the integrity and readiness of the data for further analysis.

TABLE II. NFR CATEGORIES BASED ON ISO/IEC 25010

ID	Abb.	Label	Related Attributes	
0	FR	Functional	- Functionality feature	
1	CM	Compatibility	- Co-existence	- Interoperability
2	MN	Maintainability	- Modularity - Reusability - Analyzability	- Modifiability - Testability
3	SE	Security	- Confidentiality - Integrity - Accountability	- Authenticity - Non-repudiation
4	PE	Performance	- Time behavior - Capacity	- Resource utilization
5	PO	Portability	- Adaptability - Installability	- Replaceability
6	RL	Reliability	- Availability - Recoverability	- Fault tolerance - Maturity
7	US	Usability	- Operability - Learnability - User interface aesthetics	- Appropriateness recognizability - User error protection - Accessibility

E. Statistics of the Final Annotated Dataset

The final annotated dataset comprises 10,006 unique user stories, with 64.5% labeled as functional requirements (FRs) and the remaining 35.5% distributed among seven NFRs. On average, each story contains 25 words, with sentence lengths ranging from a minimum of 7 words to a maximum of 154 words. Table III presents sample stories labeled according to NFR category. To maintain clarity and consistency, the general NFR category level was used to label user stories instead of the more detailed subcategory level as defined by ISO/IEC 25010. This decision was driven by time constraints and the relatively limited number of examples per subcategory. For instance, if a story clearly pertains to the confidentiality aspect of security, it is labeled simply as Security, since the broader category inherently encompasses the subcategory. This approach strikes a balance between precision and practicality, supporting meaningful analysis while ensuring reliable annotation.

TABLE III. EXAMPLE USER STORIES ACROSS FUNCTIONAL AND NFR

Label	Example from the Dataset
FR	As an archivist, I want to view collection files, so that I can gain background information for processing a collection.
CM	As a developer, I want a Mongo integration so that I can integrate packaged data with pipelines that use Mongo.
MN	As a user, I want the app to be automatically updated so that I always have the latest features and fixes.
SE	As a system administrator, I want users to log in using two-step authentication, so that authorized users are only allowed.
PE	As IFA, we want the system response time for fans team administration and IFA administration to not exceed 1 second so that it enhances the user's experience.
PO	As a user, I want to be able to access the service from both a website and a mobile app, so that I can use it on multiple devices.
RL	As the IFA, we want the system to support at least 50,000 users simultaneously, so that it does not crash during peak times.
US	As a user, I want the app to save my preferences, like language or format settings, so that I do not have to set them each time I use the service.

IV. TAXONOMY OF DATASET USE CASES

This section proposes a conceptual and structured taxonomy outlining potential use cases for this dataset, particularly for researchers in data-driven RE (RE) and, more specifically, those focused on Agile RE. This taxonomy is not intended to be exhaustive or to cover all possible research directions. Rather, it is derived from observations during the dataset collection process and inspired by the categorization of NLP applications on user stories [7], with some extensions and refinements. The objective is to offer researchers a high-level perspective that can help them position their research contributions in relation to this dataset. Below is a brief description of each category. Fig. 2 presents a concise structural overview of the proposed research directions. Although LLMs open new opportunities across these use cases, their risks (e.g. hallucination, bias, and explainability limitations) must also be considered, as further discussed in Section V.

A. Quality of User Story Writing, Rewriting, or Refinement

This category focuses on assessing and improving the writing quality of user stories. It includes adherence to formal templates such as INVEST [31], the three Cs criteria [32], and QUS frameworks [27], addressing syntactic, semantic, and pragmatic dimensions. It also covers practices like ensuring style consistency, converting informal stories into structured templates, and enhancing clarity or completeness. These use cases enable researchers to develop tools and metrics for evaluating and refining story quality, supporting clearer communication, testability, and estimation. The dataset provides diverse examples that can be used to train or benchmark models on automatic quality assessment and guided story refinement for formulating a high-quality and precise user story.

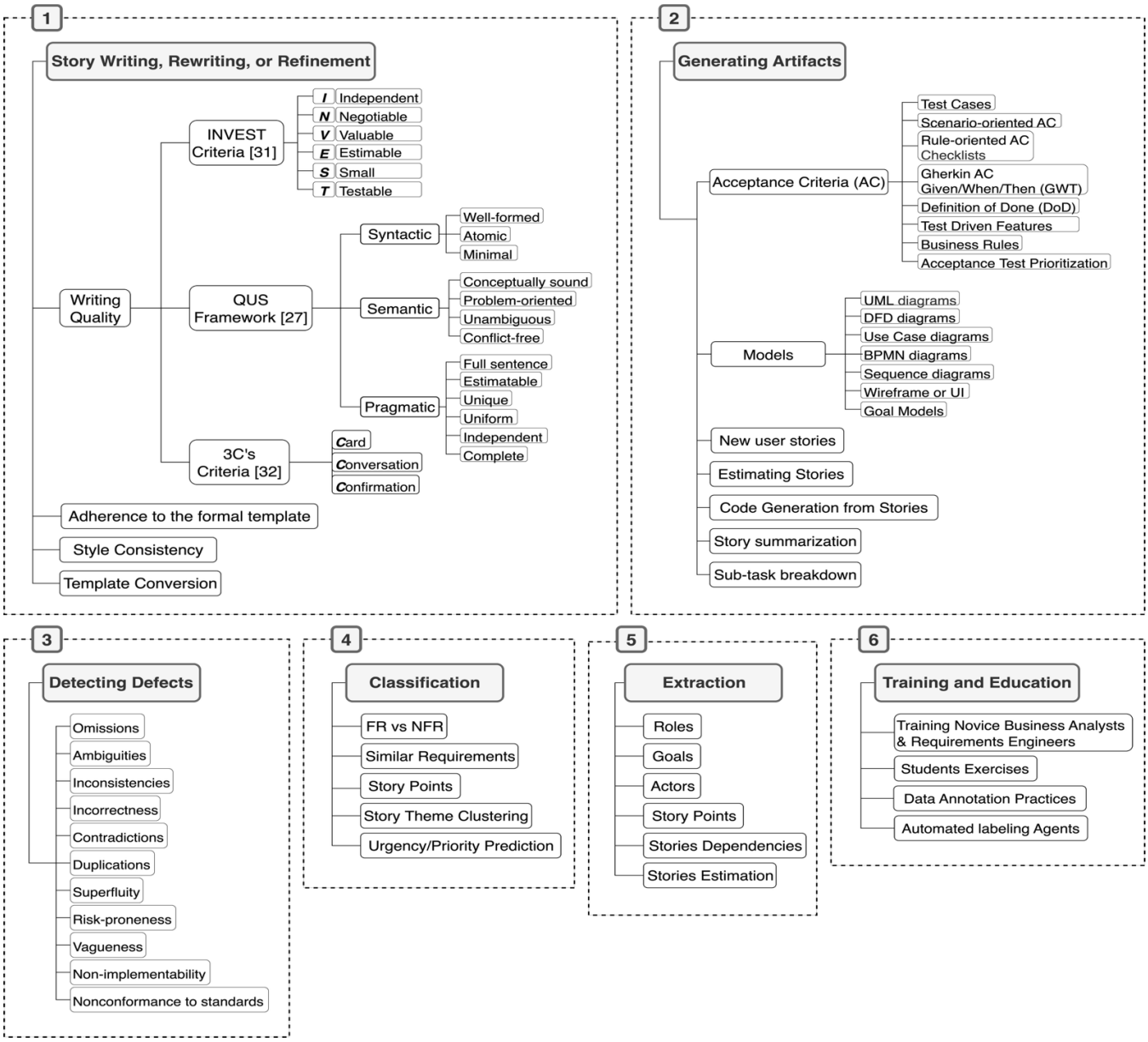

Fig. 2. Taxonomy of dataset use cases for agile requirements research.

B. Generating Artifacts from User Stories

This category covers the generation of downstream software artifacts from user stories. Various artifacts could be researched, including acceptance criteria (e.g. Gherkin [33], rule-oriented, scenario-based), test cases, UML/BPMN diagrams, and model-based artifacts like goal or sequence diagrams. For example, the dataset's labeled NFRs can train models to generate security-focused acceptance criteria (e.g. Gherkin scenarios for two-factor authentication). Other use cases include generating new user stories, decomposing stories into sub-tasks, and summarizing long stories. These transformations help automate traceability and align documentation with implementation needs. The dataset offers a strong foundation for training generative models, validating transformation pipelines, or benchmarking tools that automate requirements-to-artifact pipelines using rule-based, retrieval-based, or LLM-driven approaches.

C. Detecting and Identifying Requirements Defects

This direction targets the detection of defects in user stories that may hinder development or communication. Typical software requirements defects and their taxonomies such as [34], [35] can be very helpful in this research area. Researchers can use the dataset to develop classifiers or LLM-based prompts that detect defects such as omissions, ambiguities, inconsistencies, contradictions, and vagueness, as well as issues like non-implementability or deviation from standards, automatically. By doing so, this category supports efforts to improve the reliability and correctness of requirements documentation. The dataset offers realistic variations in quality and structure, making it suitable for training models or designing rule-based and machine learning approaches to defect detection.

D. Requirements Classification

This category involves classifying user stories based on various dimensions. Primary tasks include distinguishing between functional and NFRs, grouping similar requirements, clustering stories by theme, and predicting metadata such as story points or priority. These tasks can aid backlog management, planning, and requirement prioritization. The dataset's labeled structure supports supervised learning models for these classification tasks. Additionally, the variety in domain and writing style makes it ideal for exploring generalizability across Agile projects and evaluating classification performance of traditional ML, neural NLP, LLM, and Generative AI approaches.

E. Extraction

This category focuses on extracting structured knowledge from user stories. Targeted elements include roles, goals, actors, dependencies, and estimation features like story points. Extraction supports various downstream applications such as persona generation, automated sprint planning, or linking user stories to design models. The dataset allows for experimentation with entity recognition, role-goal modeling, and relation extraction, particularly in informal or semi-structured story formats. These tasks help bridge the gap between human-authored requirements and formal modeling or analysis tools, supporting hybrid RE environments that combine human input and AI-powered assistance.

F. Training and Education

This category emphasizes using the dataset as a resource in education and training environments. It includes applications in training novice business analysts and requirements engineers, facilitating student exercises, and practicing manual or automated annotation tasks. By engaging learners with real user stories, the dataset enables experiential learning, promotes understanding of RE concepts, and builds skills in requirement writing, evaluation, and classification. It also supports the testing and development of intelligent labeling agents. This category bridges research and pedagogy by turning real-world data into a foundation for interactive, competency-based learning in Agile RE.

V. Key Considerations for Researchers

There are several important considerations that researchers should keep in mind when working with this dataset, especially those exploring applications of LLMs and Generative AI. These considerations, summarized in Table IV, are intended to help guide the responsible and effective use of the dataset. As LLMs and generative AI are increasingly applied in Agile RE, researchers must be aware of several potential risks and challenges. Table IV summarizes these key considerations to ensure a responsible and effective application. Acknowledging and addressing these concerns is critical to maintain the quality, safety, and trustworthiness of AI-assisted tools in Agile environments.

When applying the emerging modern artificial intelligence technologies, such as LLMs and generative AI, to Agile user story datasets, researchers must account for several critical risks. These include hallucinations of facts, as LLMs can produce false responses [36], loss of context, domain misalignment, and ethical concerns such as bias or stakeholder exclusion. Additionally, generated artifacts may suffer from version drift, lack traceability, or violate privacy norms. Without proper oversight, such issues can degrade requirements quality, introduce safety risks, or mislead practitioners. Therefore, researchers must adopt mitigation strategies, such as expert validation, contextual adaptation, and anonymization, to ensure responsible and effective use of AI-driven solutions in Agile RE.

VI. Results and Discussion

A. Dataset Characteristics and Validation

The constructed dataset comprises 10,006 unique user stories, of which 6,457 (65%) are Functional Requirements (FRs) and 3,549 (35%) are Non-Functional Requirements (NFRs) as shown in Fig. 3. Within NFRs, the distribution varies substantially across categories: Usability (780) and Security (701) are the most frequent, while Maintainability (379) and Portability (381) are the least represented.

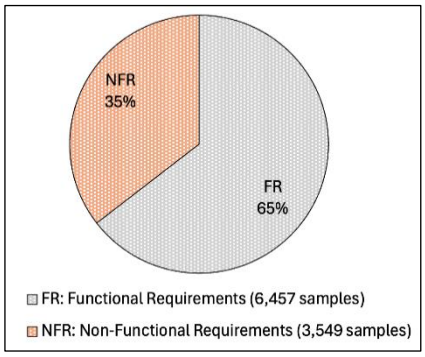

Fig. 3. Dataset FRs vs NFRs distribution.

Fig. 4 presents the Breakdown of NFR categories in the dataset. This distribution reflects common industrial practices where some NFRs, such as security, are realized more frequently than others. To ensure labeling reliability when labeling the dataset, inter-annotator agreement was measured retrospectively using Cohen's Kappa (Cohen's Kappa = 0.8335), indicating substantial agreement and validating the consistency of the classification process. This strengthens confidence in using the dataset for downstream tasks such as classification and artifact generation.

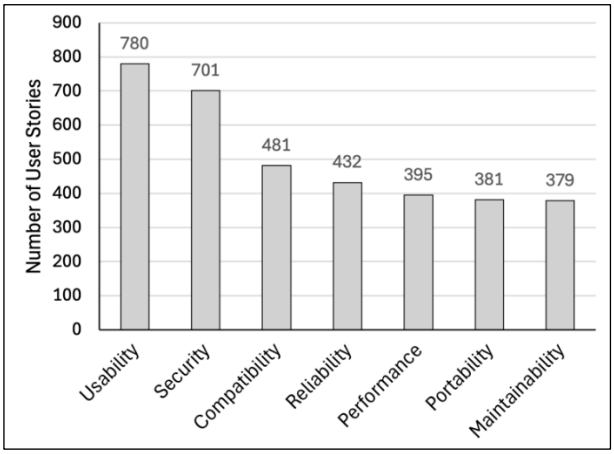

Fig. 4. NFRs distribution.

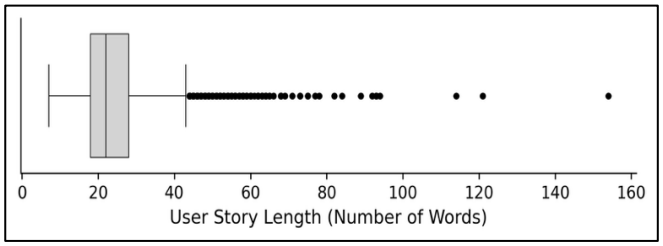

Fig. 5. Boxplot of user story lengths measured in words.

The length of user stories also shows natural variation. Fig. 5 presents a plot of user story lengths measured in words, showing an average of 23.5 words, a median of 22 words, and several long outliers reaching up to 154 words. This highlights the variability in expression style across sources. The annotated boxplot reveals several outliers, which are important to retain in order to preserve the natural linguistic diversity of agile requirements. These descriptive statistics establish the dataset's realism and its potential to support robust empirical research.

B. Demonstrated Utility in Prior Experiments

The dataset was previously employed to fine-tune transformer-based models for the automated classification of NFRs in agile user stories [9]. In these experiments, the fine-tuned BERT model achieved an F1 score of 93.4%, outperforming traditional machine learning and rule-based baselines. Comparable models such as RoBERTa, XLNet, and DistilBERT also performed competitively, underscoring the dataset's capacity to benchmark state-of-the-art approaches. Analyses further revealed that minority categories, such as Portability and Maintainability, remain challenging, largely due to data imbalance, while major categories, such as Security and Usability, achieved higher predictive performance. These results demonstrate that the dataset not only enables high-performing models but also reveals promising research applications.

TABLE IV. KEY CONSIDERATIONS FOR RESEARCHERS

Issue	Description	Implications for RE Research & Practice
Hallucination (Fabricated Output)	LLMs may produce plausible but incorrect details not present in the story or backlog.	May misguide development or testing; introduces false or unsafe assumptions. Generated outputs should undergo validation against ground-truth labels in the dataset
Project Sensitivity / Contextual Mismatch	LLMs may lack knowledge of domain-specific or organizational context, producing generic outputs.	Output may not respect business rules or constraints; increases rework.
Safety-Critical Misjudgments	In critical systems (e.g., healthcare), incorrect AI-generated requirements may cause safety risks.	Can result in non-compliance with safety-critical standards or legal risk.
Confidentiality of Requirements	User stories may expose confidential business strategies or customer data.	Risk of exposing sensitive competitive or user data through API-based models.
Data Privacy and Compliance	Stakeholder names, roles, or preferences in user stories may violate data regulations if not anonymized.	May breach regulations like GDPR, requiring careful anonymization or local deployment.
Ethical Concerns (Bias, Fairness)	LLMs may encode or introduce bias unintentionally (e.g., gender, cultural, functional bias).	Violates fairness or diversity principles in requirement framing or persona generation.
Accountability and Traceability	LLMs do not explain their reasoning, making outputs hard to audit or trace.	Challenges in proving why an AI recommendation was followed; weakens traceability.
Lack of Domain Adaptation	Generic LLMs may fail in technical or niche domains (e.g., avionics, cybersecurity).	May degrade output quality in domain-specific user stories without fine-tuning.
Validation and Verification Burden	Generated outputs still need expert review to ensure correctness and applicability.	Burden on teams to review all outputs; may nullify efficiency gains.
Overreliance or Automation Bias	Overtrust generated stories might lead to complacency or overlooked risks.	False confidence in AI can degrade quality; critical stories may be overlooked.
Incomplete Stakeholder Involvement	Reduced interaction with stakeholders may dilute shared understanding of goals.	Undermines Agile value of collaboration with customers and stakeholders.
Version Drift / Requirements Decay	Generated artifacts may go stale if not synchronized with evolving sprints or backlog.	Misaligned outputs can delay development or cause scope misunderstanding.

C. Reflections and Opportunities

The findings highlight two complementary contributions. First, the dataset itself offers a comprehensive, high-quality, and publicly available resource that consolidates fragmented prior collections into a single benchmark. Second, prior experiments confirm its utility in advancing data-driven and LLM-supported approaches in Agile-based requirements.

At the same time, the analysis surfaces future opportunities. The imbalance across NFR categories suggests the need for targeted data augmentation and rebalancing strategies. Moreover, beyond NFR identification, the dataset can support tasks such as requirements prioritization, acceptance criteria generation, defect detection, and automated documentation pipelines. These directions are encouraged to strengthen the dataset's role in aligning agile practices with high-quality software engineering outcomes.

VII. THREATS TO VALIDITY

One potential threat to internal validity lies in the manual annotation process of user stories for NFRs. Despite employing ISO/IEC 25010 and using multiple annotators with inter-rater reliability checks (Cohen's Kappa = 0.8335), subjectivity still influences the labeling decisions, especially for implicit NFRs. Variability in domain contexts and story phrasing definitely has some influence on the labelers' interpretations during their annotation process.

A threat to external validity stems from the scope and diversity of the dataset sources. While the dataset aggregates samples from multiple academic and online repositories, it still does not fully reflect domain-specific conventions across

various industries (e.g. finance, aerospace, or healthcare). As a result, the generalizability of any AI model trained on this dataset should be assessed before deployment in such specialized contexts.

VIII. Conclusion

This paper introduces a comprehensive and publicly available dataset of over 10,000 agile user stories aggregated from academic and online sources. The dataset, systematically annotated using ISO/IEC 25010-based NFR categories, was initially designed to support research on automated NFR identification. To ensure reliability, the labeling process was validated through inter-annotator agreement using Cohen's Kappa, which yielded a substantial agreement score of 0.8335, affirming the consistency of the annotations. Beyond its original purpose, the dataset holds value for a variety of tasks in agile RE, including user story classification, artifact generation, defect detection, and educational applications. A structured taxonomy of use cases was also proposed to guide future research and practical adoption of the dataset, particularly in data-driven and AI-enhanced RE.

Looking ahead, further research is needed to explore how the dataset and its annotation process can enhance the trustworthiness and scalability of automated NFR identification and classification. Future work will also investigate applications such as requirements prioritization and the formulation of acceptance criteria for NFRs in user stories. Moreover, the dataset opens up the possibility of developing automated pipelines that integrate LLMs to support practical tasks, such as requirement specification, prioritization, traceability, and acceptance criteria generation. These directions are encouraged to strengthen the validity of the dataset and its alignment with agile practices, ultimately supporting fruitful research and high-quality software engineering outcomes.

References

[1] Wang, L. Zhao, Y. Wang, and J. Sun, "The Role of Requirements Engineering Practices in Agile Development: An Empirical Study," in Requirements Engineering, D. Zowghi and Z. Jin, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 195–209. doi: 10.1007/978-3-662-43610-3_15.

[2] E.-M. Schön, J. Thomaschewski, and M. J. Escalona, "Agile Requirements Engineering: A systematic literature review," Computer Standards & Interfaces, vol. 49, pp. 79–91, 2017, doi: 10.1016/j.csi.2016.08.011.

[3] G. Lucassen, F. Dalpiaz, J. M. Werf, and S. Brinkkemper, "The Use and Effectiveness of User Stories in Practice," in Proceedings of the 22nd International Working Conference on Requirements Engineering: Foundation for Software Quality, in REFSQ 2016, vol. 9619. Berlin, Heidelberg: Springer-Verlag, 2016, pp. 205–222. doi: 10.1007/978-3-319-30282-9_14.

[4] M. Cohn, User Stories Applied: For Agile Software Development. USA: Addison Wesley Longman Publishing Co., Inc., 2004.

[5] G. Lucassen, F. Dalpiaz, J. M. E. Van Der Werf, and S. Brinkkemper, "Forging high-quality user stories: towards a discipline for agile requirements," presented at the 2015 IEEE 23rd international requirements engineering conference (RE), IEEE, 2015, pp. 126–135. doi: 10.1109/RE.2015.7320415.

[6] S. Mohammed et al., "The effects of data quality on machine learning performance on tabular data," Information Systems, vol. 132, p. 102549, 2025, doi: https://doi.org/10.1016/j.is.2025.102549.

[7] 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, doi: 10.1109/ACCESS.2021.3070606.

[8] F. Dalpiaz, "Requirements data sets (user stories)," Mendeley Data, v1, 2018, doi: 10.17632/7zbk8zsd8y.1.

[9] A. Alhaizaey and M. Al-Mashari, "Automated Classification and Identification of Non-Functional Requirements in Agile-Based Requirements Using Pre-Trained Language Models," IEEE Access, vol. 13, pp. 87401–87417, 2025, doi: 10.1109/ACCESS.2025.3570359.

[10] A. Alhaizaey and M. Al-Mashari, "A Framework for Reviewing and Improving Non-Functional Requirements in Agile-based Requirements," in 2023 18th Iberian Conference on Information Systems and Technologies (CISTI), June 2023, pp. 1–7. doi: 10.23919/CISTI58278.2023.10211956.

[11] M. Kassab, "The changing landscape of requirements engineering practices over the past decade," in 2015 IEEE Fifth International Workshop on Empirical Requirements Engineering (EmpiRE), 2015, pp. 1–8. doi: 10.1109/EmpiRE.2015.7431299.

[12] E. M. Schön, D. Winter, M. J. Escalona, and J. Thomaschewski, "Key Challenges in Agile Requirements Engineering," in Agile Processes in Software Engineering and Extreme Programming, H. Baumeister, H. Lichter, and M. Riebisch, Eds., Springer, 2017, pp. 37–51. doi: 10.1007/978-3-319-57633-6_3.

[13] I. Inayat, S. S. Salim, S. Marczak, M. Daneva, and S. Shamshirband, "A systematic literature review on agile requirements engineering practices and challenges," Computers in Human Behavior, vol. 51, pp. 915–929, 2015, doi: 10.1016/j.chb.2014.10.046.

[14] A. R. Amna and G. Poels, "Systematic Literature Mapping of User Story Research," IEEE Access, vol. 10, pp. 51723–51746, 2022, doi: 10.1109/ACCESS.2022.3173745.

[15] A. Hemmat, M. Sharbaf, S. Kolahdouz-Rahimi, K. Lano, and S. Y. Tehrani, "Research directions for using LLM in software requirement engineering: a systematic review," Frontiers in Computer Science, vol. Volume 7-2025, 2025, doi: 10.3389/fcomp.2025.1519437.

[16] C. Arora, J. Grundy, and M. Abdelrazek, "Advancing Requirements Engineering Through Generative AI: Assessing the Role of LLMs," in Generative AI for Effective Software Development, A. Nguyen-Duc, P. Abrahamsson, and F. Khomh, Eds., Cham: Springer Nature Switzerland, 2024, pp. 129–148. doi: 10.1007/978-3-031-55642-5_6.

[17] F. Casillo, V. Deufemia, and C. Gravino, "Detecting privacy requirements from User Stories with NLP transfer learning models," Information and Software Technology, vol. 146, p. 106853, 2022, doi: 10.1016/j.infsof.2022.106853.

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

[19] F. Gilson, M. Galster, and F. Georis, "Extracting Quality Attributes from User Stories for Early Architecture Decision Making," in 2019 IEEE International Conference on Software Architecture Companion (ICSA-C), 2019, pp. 129–136. doi: 10.1109/ICSA-C.2019.00031.

[20] Y. Li, J. Keung, Z. Yang, X. Ma, J. Zhang, and S. Liu, "SimAC: simulating agile collaboration to generate acceptance criteria in user story elaboration," Automated Software Engineering, vol. 31, no. 2, p. 55, June 2024, doi: 10.1007/s10515-024-00448-7.

[21] ISO/IEC 25010:2011, "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models," International Organization for Standardization, Standard. [Online]. Available: iso.org

[22] International Organization for Standardization, ISO/IEC 25010:2023 — Systems and Software Engineering — Systems and Software Quality Requirements and Evaluation (SQuaRE) — Product Quality Model, Geneva, Switzerland., 2023. [Online]. Available: https://www.iso.org/standard/78176.html

[23] P. K. Murukannaiah, N. Ajmeri, and M. P. Singh, "Acquiring Creative Requirements from the Crowd: Understanding the Influences of Personality and Creative Potential in Crowd RE," in 2016 IEEE 24th International Requirements Engineering Conference (RE), Beijing, China: IEEE, Sept. 2016, pp. 176–185. doi: 10.1109/RE.2016.68.

[24] S. G. Köse and F. B. Aydemir, "A User Story Dataset for Library and Restaurant Management Systems." Zenodo, Jan. 12, 2023. doi: 10.5281/ZENODO.7529130.

[25] A. Rohmann and B. Paech, "Dataset and Application of Algorithms of 'Towards a More Set of Acceptance Criteria.'" Zenodo, Nov. 24, 2022. doi: 10.5281/ZENODO.7358697.

[26] G. Lucassen, M. Robeer, F. Dalpiaz, J. M. E. M. Van Der Werf, and S. Brinkkemper, "Extracting conceptual models from user stories with Visual Narrator," Requirements Eng, vol. 22, no. 3, pp. 339–358, Sept. 2017, doi: 10.1007/s00766-017-0270-1.

[27] G. Lucassen, F. Dalpiaz, J. M. E. 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, doi: 10.1007/s00766-016-0250-x.

[28] A. Prabu, L. A. Suruthi, B. Sabarish, and K. Arun, "User story-based automatic test case classification and prioritization using natural language processing-based deep learning," IEEE Potentials, pp. 2–11, 2024, doi: 10.1109/MPOT.2023.3342366.

[29] F. Dalpiaz, P. Gieske, and A. Sturm, "On deriving conceptual models from user requirements: An empirical study," Information and Software Technology, vol. 131, p. 106484, Mar. 2021, doi: 10.1016/j.infsof.2020.106484.

[30] J. Cohen, "A coefficient of agreement for nominal scales," Educational and psychological measurement, vol. 20, no. 1, pp. 37–46, 1960.

[31] B. Wake, "INVEST in good stories, and SMART tasks," XP123. Accessed: Apr. 04, 2025. [Online]. Available: https://xp123.com/articles/invest-in-good-stories-and-smart-tasks/

[32] R. Jeffries, "Essential XP: Card, Conversation, Confirmation." Accessed: Apr. 04, 2025. [Online]. Available: https://ronjeffries.com/xprog/articles/expcardconversationconfirmation/

[33] M. Wynne, A. Hellesoy, and S. Tooke, The cucumber book: behaviour-driven development for testers and developers. The Pragmatic Programmers LLC, 2017.

[34] F. Shull, I. Rus, and V. Basili, "How perspective-based reading can improve requirements inspections," Computer, vol. 33, no. 7, pp. 73–79, 2000, doi: 10.1109/2.869376.

[35] A. A. Alshazly, A. M. Elfatatry, and M. S. Abougabal, "Detecting defects in software requirements specification," Alexandria Engineering Journal, vol. 53, no. 3, pp. 513–527, 2014, doi: 10.1016/j.aej.2014.06.001.

[36] J.-Y. Yao, K.-P. Ning, Z.-H. Liu, M.-N. Ning, Y.-Y. Liu, and L. Yuan, "LLM lies: Hallucinations are not bugs, but features as adversarial examples," arXiv preprint arXiv:2310.01469, 2023.