Novel Approaches in Synthetic Dataset Generation

Luca Morelli*, Daniele Galloppo†, Mario Peluso‡
*l.morelli6@studenti.unisa.it
†d.galloppo@studenti.unisa.it
‡m.peluso37@studenti.unisa.it

***Abstract*—Context: As Artificial Intelligence (AI) is becoming increasingly prevalent in many aspects of our society, there is a growing need for support and control in the phases that accompany the development of these models. Among all the potential issues that could arise, "fairness" has emerged in recent years as one of the most critical examples of what a suboptimal AI model could not lead to. Famous examples include cases related to Amazon's artificial intelligence model used for employee recruitment and cases related to the U.S. court system, which over the years has made extensive use of an artificial intelligence model (COMPAS) that is able to predict an individual's recidivism by providing judges with a value representing the likelihood that the inmate will re-offend. Both cases brought attention to fair and unbiased models, as humans were involved in the wrong decisions that have been made, giving birth to the concept of "fair machine learning". Due to the challenges regarding this environment that could not be held solely by humans, REFAIR [1] was born, a model whose goal is to help developers from the earliest stages of software development.**

Objective: Due to lack of a dataset, the capabilities of REFAIR have been assessed experimenting solely with a synthetic dataset, created by using a combination of the LLM ChatGPT, based on the GPT-3.5 architecture. Since at the current state there are still no datasets regarding user stories in a machine learning environment, the idea behind this paper is to assess the model using newer prompt engineering techniques such as few shot learning or chain-of-thought learning that should be able to create a more near-human, and therefore realistic, synthetic dataset.

***Index Terms*—fairness, machine learning, artificial intelligence, user story**

I. INTRODUCTION

Techniques such as artificial intelligence aim to analyze large amounts of data to find correlations that might escape the human eye, enabling more accurate predictions and deeper analysis. However, the massive use of data poses several questions: if the data used to train the algorithms contains biases, could those biases be incorporated into the results thereby influencing future decisions?

Furthermore, this could be particularly problematic when the biases are represented by sensitive data such as gender, ethnicity, age, religion or political orientation, leading to potential discrimination. To address this specific issue, several approaches have been proposed to reduce bias in data and ensure that AI applications are balanced, fair, and reliable. However, also the definition of "fairness" can be a complex task.

It's worth noting that multiple definitions of "fairness" exist, some of which may conflict with each other, underscoring the importance of context in defining fairness.

Yet, when it comes to implementing sustainable fairness in artificial intelligence systems, this aspect is further complicated. To address this complexity and foster fairness-aware models, REFAIR has been developed. It is classification model that, based on the application domain and the ML task, recommends sensitive features that, if not correctly treated, may lead to unfair and biased models.

Structure of the report. In Section II we will see a brief introduction to REFAIR, namely what it does and how it works, in the Section III we will discuss the goals of this project, in Section IV we will define the RQs that will guide us, in Section V we will discuss about the methodology to achieve our goals and, in Section VI, we will define the steps followed in order to create the datasets. Subsequently, in section VII, we will discuss the results obtained from the survey and the performance results of the model and in section VIII, we will address the problems encountered. The paper ends with some observations derived from the results, namely section IX, how the threats were addressed, namely section X and and conclusions were dealt with, namely section XI.

II. STATE OF ART

REFAIR [1] is a framework presented at ICSE '24. Its goal is to analyze machine learning-related User Stories (USs) by categorizing the application domain and ML tasks. Subsequently, it maps the information into specific, relevant features that need to be considered and then returns the results to the user.

REFAIR provides recommendations to requirements engineers to address potential equity issues during development, also recognizing that external factors such as laws and regulations may influence the identification of sensitive features. By providing this analysis and guidance, REFAIR assists in ensuring fairness and equity in the development of ML-enabled systems.

Going into detail, REFAIR utilizes a step-by-step approach to analyze US documents. Given a single US, the first step involves creating a spatial representation of the input using an N-dimensional space, thus allowing the extraction of features from the text that can be used for classification. In other terms, the text is transformed into a real-valued vector that can be used in subsequent steps. A further step involves the use of word embeddings as input for REFAIR's classification analysis modules. These modules consist of two main components: (1) *Application Domain Classification*, which determines the most probable application domain from a selection of 34 domains and (2) *Machine Learning Tasks Classification*, which is responsible for classifying the ML-task(s) likely to be employed

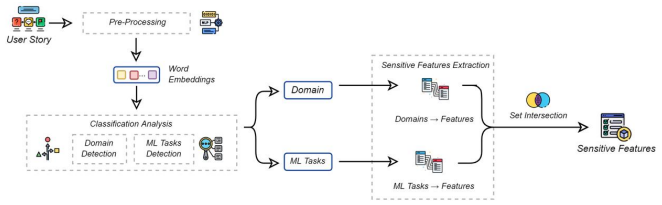

Fig. 1. Activity flow behind REFAIR's

when implementing the US. The model will then map those pieces of information onto specific sensitive features, enabling REFAIR to effectively classify and analyze US documents.

As mentioned above, since the current literature does not offer any off-the-shelf solutions, a synthetic dataset consisting of 12,401 User Stories related to 34 application domains was created in a supervised environment. The aim was to evaluate the capabilities of REFAIR and to further safeguard against potential subjectivity.

In the external inspection, a further quality assessment was conducted involving a statistically significant sample of synthetic USs. Each US then was evaluated based on comprehensibility, i.e., the degree to which the US is understandable, realism, i.e., the degree to which the US resembles a real user story, and actionability i.e., the degree to which the US can drive the development of an ML-enabled project.

Although the results presented in [1] suggest that the way synthetic USs were created is good enough to be human-like, and therefore the usage of the synthetic dataset does not invalidate the model's performance in a real-world environment, further improvement could be achieved by creating an even more human-like dataset.

This can be achieved by applying prompt engineering techniques or refining a model according to our goals. In this regard, the state of the art offers several possibilities. For instance, there's the (i) few-shot learning technique [2], which enables in-context learning by providing demonstrations in the prompt of the LLM, the (ii) chain-of-thought [3], which utilizes few-shot learning examples containing detailed reasoning steps to enhance the reasoning capabilities of the LLM, the (iii) zero-shot learning applied to chain-of-thought techniques [4], where the model is pushed to reason step by step without any examples provided on how to do it and (iv) instruction tuning [5]. It has to be noted that while (iv) is a technique that fine-tunes a pre-trained language model on a mixture of tasks phrased as instructions, (i), (ii) and (iii) do not rely on fine-tuned models. Instead, they aim to assist the LLM based on how the prompt will be created. Finally, all of these techniques demonstrate high performance compared to the zero-shot learning baseline, where no examples whatsoever are provided.

III. OBJECTIVE

Despite REFAIR acts as a perfect recommender for 97% of User Stories (USs), maintaining a decent level of quality even for the remaining 3% where errors may occur, concerns may arise when considering that real work environments may not be adequately represented by the synthetic dataset on which the model has been assessed.

Although there are no off-the-shelf datasets available at the time of this report, new techniques such as few-shot learning [2] or chain-of-thought learning [4] have been evaluated, allowing us to attempt the creation of new synthetic datasets that could more closely resemble human-like USs. The selection of these techniques is strictly related to the performance they achieved, particularly in the case of chain-of-thought learning, which, when working on PaLM, has achieved state-of-the-art accuracy with just eight chain-of-thought exemplars. Along with those techniques, we will also apply fine-tuning on the LLM model LLaMa, aiming to explore an alternative approach for creating a synthetic dataset.

IV. RESEARCH QUESTIONS

The following chapter will contain the research questions that will be addressed in this paper. They are as follows:

> **RQ$_1$**. *To what extent does the synthetic dataset created through few-shot learning technique impact the model's performance?*

> **RQ$_2$**. *To what extent does the synthetic dataset created through chain-of-thought learning technique impact the model's performance?*

> **RQ$_3$**. *To what extent does the synthetic dataset created through fine-tuning of LLaMa impact the model's performance?*

> **RQ$_4$**. *To what extent can REFAIR's Deep Learning version classify ML-specific application domains from User Stories?*

> **RQ$_5$**. *To what extent can REFAIR's Deep Learning version classify ML-specific tasks from User Stories?*

Going into detail:

> **RQ$_1$**. *To what extent does the synthetic dataset created through few-shot learning technique impact the model's performance?*

This research question was formulated to analyze how the model's performance would be impacted by evaluating it using the new synthetic dataset created through the few-shot learning technique. To achieve this objective, it will be necessary to assess the model using the same metrics employed in the initial evaluation of REFAIR: F1-Score and accuracy for domain classification and F1-Score and Hamming Loss for ML Task classification. The results will then be compared with those presented in REFAIR.

> **RQ$_2$**. *To what extent does the synthetic dataset created through chain-of-thought learning technique impact the model's performance?*

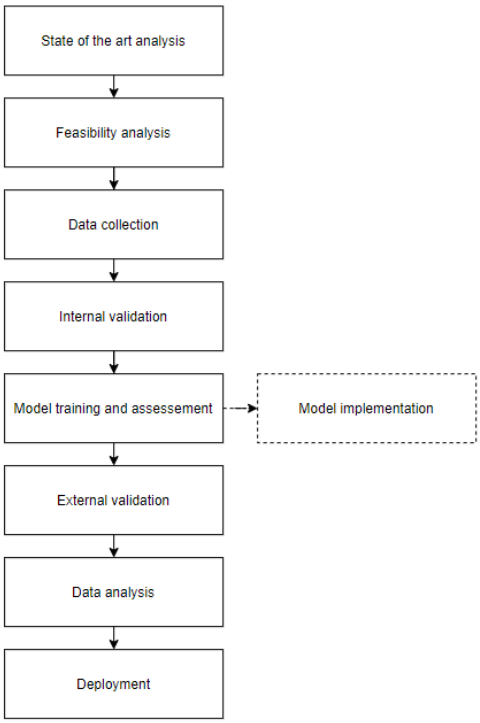

Fig. 2. Methodological steps

Similarly to $\boldsymbol{RQ}_1$, $\boldsymbol{RQ}_2$ aims at understanding to what extent the model's performance would be impacted by evaluating it using a combination of the old dataset and the new synthetic dataset created through the chain-of-thought learning technique. For $\boldsymbol{RQ}_2$ will be used same metrics of $\boldsymbol{RQ}_1$, namely F1-Score and accuracy for domain classification and F1-Score and Hamming Loss for ML Task classification.

> 🔍 $\boldsymbol{RQ}_3$. *To what extent does the synthetic dataset created through fine-tuning of LLaMa impact the model's performance?*

This research question aims to investigate the impact of utilizing the synthetic dataset generated through fine-tuning of the LLaMa model on the performance of the REFAIR model. Similar to the approach taken with $\boldsymbol{RQ}_1$, the evaluation will employ the same metrics utilized for REFAIR enabling a comprehensive assessment of the model's accuracy. These findings will then be juxtaposed against the performance metrics from the initial evaluation.

By addressing $\boldsymbol{RQ}_1$, $\boldsymbol{RQ}_2$ and $\boldsymbol{RQ}_3$, we aim to provide insights into the effectiveness of different synthetic dataset generation techniques, shedding light on their respective impacts on the performance of the REFAIR model. This multifaceted analysis will contribute to a deeper understanding of the model's robustness and its ability to generalize across diverse datasets.

> 🔍 $\boldsymbol{RQ}_4$. *To what extent can REFAIR's Deep Learning version classify ML-specific application domains from User Stories?*

> 🔍 $\boldsymbol{RQ}_5$. *To what extent can REFAIR's Deep Learning version classify ML-specific tasks from User Stories?*

These research questions aim at evaluating the model in the same way as the light version of REFAIR has been evaluated. To address $\boldsymbol{RQ}_4$, F1-Score and accuracy will be used, while to address $\boldsymbol{RQ}_5$, F1-Score and Hamming Loss will be employed.

V. METHODOLOGICAL STEPS

It is also necessary to define the methodological choices and steps that will guide us throughout this project. These steps are illustrated in the Figure 2.

State of the art analysis. In order to achieve a comprehensive understanding of our problem, we conducted an analysis of the state of the art, which provided us with deeper insights into both REFAIR [1] and Prompt Engineering. The latter will be employed to create the synthetic dataset used for training and evaluating the model. At last, we will explore the instruction tuning technique. Additional information about the state of the art is provided in Section II.

Feasibility analysis. To evaluate the feasibility of this project, we need to focus on two key aspects: (i) the synthetic dataset created using prompt engineering techniques, and (ii) the synthetic dataset created using a fine-tuned LLM, like LLaMa. It's worth noting that prompt engineering, in terms of its application, is relatively new, so we have limited information available.

Regarding prompt engineering, the techniques that will be used include (a) few-shot learning, (b) chain-of-thought learning, and possibly (c) chain-of-thought learning combined with zero-shot learning. However, a major challenge in this case is the relative novelty of these techniques, resulting in a limited level of understanding in the current state of the art. Consequently, these techniques may not be fully applicable or applicable at all.

Regarding point (ii), the challenge lies in effectively fine-tuning a large language model (LLM), a task that requires considerable computational resources and has a significant environmental impact, especially in terms of carbon emissions. Nevertheless, considerable efforts have been invested in streamlining this process and minimizing resource consumption [6]. In addition, through the usage of Low Rank Adaptations (LoRAs) it is possible to greatly reduce the number of trainable parameters and achieve a result very close to that of the fine-tuning process[7]. Consequently, one viable strategy involves leveraging LoRAs and instruction tuning to allow the generation of higher-quality user stories. A dataset will be needed for the tuning process, and a good starting point could be the synthetic dataset of user stories used for REFAIR, which has to be modified by associating an instruction with each user story.

Data collection Building upon the methodology outlined in the REFAIR paper, our objective is to further improve

the synthetic dataset. We plan to leverage their prompt and incorporate (i) few-shot learning and (ii) chain-of-thought learning techniques.

Furthermore, in order to give a more specific view of what will be done, following what is reported in the papers in which (i) and (ii) are presented, we will use 3-shot learning (or 5-shot learning in the cases where 3-shot learning isn't enough) and chain-of-thought learning providing the LLM with only one example. In cases where providing even only one example is challenging, we will use zero-shot learning applied to chain-of-thought learning, meaning that no examples will be provided, only asking to the LLM to slowly reason on the response, as shown in [4].

For the prompt that will be used, we will start from the one provided in REFAIR, namely the best found for the creation of the old synthetic dataset. It has to be noted that, given the fact that different techniques will be applied, the prompt may not be optimal. Nonetheless, it represents a valid starting point. The prompt is presented subsequently:

Prompt employed to generate USs.
Considering the following:
[*High-level machine learning task*]
OR [*Low-level machine learning technique*]
in the field of [*machine learning*]
OR [*natural language processing*].
Can you provide me with specific user stories
for the following application domains?
[*List of Relevant Application Domains*]

The application of (i) and (ii) will be done at separate stages in order to create different datasets, and each technique will be applied to every application domain on which the dataset has been created. For this matter, we will consider the same augmented ontology that has been employed in [1], where they exploited the OWL ontology developed by Fabris et al. [8], augmented with the information contained in the AI dictionary proposed by Duran Silva et al. [9]. In order to maintain an high level of generalization and an high level of similarity with the old synthetic dataset, the field of the USs, namely ML or NLP, and the level of the ML task, namely specific and detailed or generic, will vary.

At last, in order to explore more solutions, prompt engineering techniques will be applied on CHATGPT and eventually on LLaMa. The prompt engineering techniques, however, will be used on LLaMa only as a last resort if, even with fine-tuning, the model will be not able to create a synthetic dataset good enough to pass the internal validation. Even though the better scaling of those techniques on LLMs with a higher number of parameters has been largely demonstrated, the idea is that the finetuned version of LLaMa could compensate for the absence of the high number of parameters in the LLM.

It should be noted that PaLM, the LLM model in which chain-of-thought has reached state-of-the-art accuracy, will not be used for creating the synthetic dataset since it is not currently available to the public.

Internal validation. Before assessing the model on the new dataset, we will internally evaluate the quality of the USs, described by *comprehensibility*, i.e., the degree to which the US is understandable, *realism*, i.e., the degree to which the US is written as a real user story and *actionability*, i.e., the degree to which the US can be used to drive the development of a ML-enabled project.

Model training and assessment. Based on what is reported in REFAIR, the model will be trained using only the best combination of models and using a combination of the new synthetic dataset and the old ones, while the assessment will be done following what has been said in Section IV when responding to $\boldsymbol{RQ}_1$, $\boldsymbol{RQ}_2$ and $\boldsymbol{RQ}_3$. In the eventuality of a REFAIR's Deep Learning version, the model will be trained using the new synthetic dataset and assessed following what is reported in Section IV when responding to $\boldsymbol{RQ}_4$ and $\boldsymbol{RQ}_5$.

Model implementation. After assessing the model, based on our results, one possible course of action could be to develop a deep learning algorithm to further improve the performance of REFAIR. The new model will aim to provide a non-light version of the model, giving users the option to choose between the light version and the new non-light one.

External validation. To further safeguard against potential subjectivity in the internal inspection, our objective in this phase is to validate the synthetic dataset through participants, using the same metrics defined for the internal validation. Following their example will grant us the possibility to better compare the external validation reults.

The idea is to take a sample of our synthetic dataset and conduct a survey that will allow us to understand to what extent the new synthetic USs differ from the old ones. Due to the impossibility of asking the same participants of REFAIR, we will include both new and old USs in the same survey. The survey will be structured in such a way that it can be completed in few minutes. Each participant, recruited voluntarily, will have to rate 8 different USs grouped in 2 tasks using a Lickert scale, going from 1 (not at all) to 5 (extremely) for each of the metrics described earlier.

Data analysis. Going into details, $\boldsymbol{RQ}_1$ will be addressed by comparing the results obtained by REFAIR using both old and new dataset crated trhough the few shot learning approach. The performance of the new synthetic dataset will be evaluated using F1-Score and accuracy for domain classification and F1-Score and Hamming Loss for ML task classification, which are the same metrics reported in REFAIR [1]. $\boldsymbol{RQ}_2$ and $\boldsymbol{RQ}_3$ will follow a similar approach to $\boldsymbol{RQ}_1$, with the only difference being the method of the new synthetic dataset creation, Few Shot for $\boldsymbol{RQ}_1$, chain-of-thought for $\boldsymbol{RQ}_2$, and fine-tuning for $\boldsymbol{RQ}_3$.

For $\boldsymbol{RQ}_4$ and $\boldsymbol{RQ}_5$, a different approach will be considered. These RQs will be addressed by creating, training, and validating a Deep Learning model using only one dataset. Therefore,

the main difference compared to the previous RQs will be the absence of a direct comparison of the new model with the results of the old one. As reported in Section IV, $\boldsymbol{RQ}_4$ will be addressed using F1-Score and accuracy, while to address $\boldsymbol{RQ}_5$ F1-Score and Hamming Loss will be used.

It should be noted that this report aims solely to test a new technique; therefore, positive or negative responses to the first three RQs will be considered a success. Lastly, since the Deep Learning algorithm may be unnecessary, $\boldsymbol{RQ}_4$ and $\boldsymbol{RQ}_5$ may remain unaddressed.

Deployment. At the current state, REFAIR is provided through a client-server application. Another idea is to provide REFAIR as a desktop application, making it usable by a wider set of users. It is important to note that if deployed, the Deep Learning version will not serve as a substitute for the current model, and therefore, users could decide which model to interrogate. Code and datasets will be available on the GitHub repository.

VI. DATASETS GENERATION

The subsequent section will describe in detail the steps taken during the dataset generation process. Before going deeper into details, some considerations need to be addressed. Due to the non-deterministic nature of ChatGPT, all analyses are conducted based on average results. Each prompt has been tested multiple times, and the results have been aggregated before determining their adequacy.

Furthermore, the diversity of domains could represent a challenge. Certain domains may yield better results with prompts that other domains perform poorly on. Therefore, each domain should be evaluated independently of the others.

At last, all the prompts used are available on the GitHub repository.

We created separate 3 dataset, each of them consisting of 680 USs, with 20 tasks allocated to each of the 34 domains. There is no particular reason why exactly 20 tasks were chosen but they were selected based on their type and interrelationship. For instance, tasks like ”random forest” and ”decision tree” were chosen to assess whether the model could generate similar USs based on task similarity and how significantly the task influenced the response.

In this context, we observed that the model understood the relationship between ”random forest” and ”decision tree”, recognizing that the former is a more complex version of the latter.

A. Few shot learning approach

In order to understand the actual feasibility of the project, we initially gave ChatGPT a simple prompt, named *Initial_FSPrompt*, composed of three USs and a request. The results obtained allowed us to determine whether the model, based on our examples, was able to understand the content of the USs to create, specifically the possible tasks and fields involved.

As expected, the results were chaotic and lacked of a logical connection to the examples provided. However, they allowed us to gain a deeper understanding of the feasibility of the project and the prompt used in REFAIR.

Based on the previous problems and the knowledge gained, the second prompt, named *Adjusted_FSPrompt*, aimed to help ChatGPT understand the three USs provided as examples. To achieve this goal, we added some information for each US: (i) the *task* that the US had to describe and (ii) the *field* in which the task had to be completed.

From *Initial_FSPrompt* to *Adjusted_FSPrompt*, we also changed the way the question following the USs was formulated. While in *Initial_FSPrompt* it was extremely necessary to specify to ChatGPT the maximum length of the USs in order to avoid unecessarily long results (even though only short examples were provided), with *Adjusted_FSPrompt*, ChatGPT was able to be more specific and less verbose, leading to shorter phrases.

This prompt change led to significant results, although there was still something worth noting: The USs respected the domain, task, and field, but they were too technical or contained unusual terms to have been written by a human. This implies that even if the user stories were able to pass internal validation, some issues could arise during external validation.

To minimize these problems, we created another prompt, named *Random_FSPrompt*. In the attempts made up to that point, we were providing examples that were always related to the same domain. Since the problem was the model's overly deep knowledge, we tried generating some USs by simply changing the domains of the examples, while keeping the same questions' structure as in *Adjusted_FSPrompt*.

This prompt led to a decrease in performance. While the model with *Adjusted_FSPrompt* was able to independently understand the domain of the examples and create a new US from the same domain, *Random_FSPrompt* was not always able to correctly identify the subject of the US or the specified domain.

Another idea, independent from the problem that *Random_FSPrompt* caused, was to increase the number of USs provided as examples from 3 to 5. This choice was made since, in the paper where the few-shot learning technique was presented [2], Brown et al. demonstrated how with more examples, ChatGPT had better performance.

Following their example, we created *Enlarged_FSPrompt*, which showed unexpectedly poor results. During the various tests, it appeared that ChatGPT not only had difficulties in understanding the subject and the domain of the US to generate but also started trying to mix the ML or NLP tasks of the example USs with the task that had to be the content of the generated US. It should be noted that different domains could lead to different results, so *Enlarged_FSPrompt* could be a viable choice in situations where the domain is hard to understand even for ChatGPT.

Subsequently to *Enlarged_FSPrompt*, we defined *ReAdjusted_FSPrompt*. Based on the good results of *Adjusted_FSPrompt* and thinking that the way ChatGPT's response could be influenced by how the question was posed, we tried to modify the structure of the question, making it more direct and clear. Compared to *Adjusted_FSPrompt*, no different results were achieved, but given the clarity of the new prompt, we decided to stick with it.

At last, we tried to solve the problem related to the high level of technicism used by ChatGPT once again. Starting with a similar idea of changing domains, as we did with *Random_FSPrompt*, we created *Domain_FSPrompt*.

The problem with the former was related to ChatGPT's lack of comprehension when trying to understand our examples since their domains were different from each other, while with the latter we provided 3 USs from the same domain, allowing ChatGPT to deeply understand what we provided, and subsequently asked it to generate a US related to a different one. It should be noted that the domain of the example USs and the domain of the US to be generated were taken from the same domain cluster, such as Literature & Linguistics, Medicine & Health, etc.

Domain_FSPrompt finally led to slightly better results in the domains on which it was tested. The new USs were less technical while still preserving the good qualities found in *ReAdjusted_FSPrompt*'s results. These better results could probably be related to "in-context learning," but, to prove it, further analysis is needed.

B. Chain-Of-Thought learning approach

Starting from the prompts used for Few Shot learning, we have been able to find more easily prompts that led to good results. Specifically, we started from *ReAdjusted_FSPrompt* and *Domain_FSPrompt*, the most promising ones.

In this sense, the new prompts *ReAdjusted_CoTPrompt* and *Domain_CoTPrompt*, which respected the peculiarities of their Few Shot versions have been initially defined.

Between *ReAdjusted_CoTPrompt* and *Domain_CoTPrompt*, differently from the Few Shot approach, since the USs provided as results were slightly better, *ReAdjusted_CoTPrompt* seemed to be the most promising.

At the same time, we also wanted to experiment with Chain-Of-Thought combined with Zero Shot learning. This combination led to *Zero_CoTPrompt*. In this and the other prompts related to this technique, no examples were provided, but we simply invited the model to reason slowly starting from the question we posed.

In this case, during our first approach with *Zero_CoTPrompt*, we started the conversation with a message that aimed at defining the context. This combination led to results comparable to *Domain_FSPrompt* or even *ReAdjusted_CoTPrompt*, while the application of *Zero_CoTPrompt* without the context definition led to critical results.

That said, aiming for promising results, we started applying some slight changes in the way *Zero_CoTPrompt* without context was written or formulated. We started by:

- Changing the way the word 'US' was written. Not abbreviating it appeared to be useful since the model started understanding the task, and therefore producing user stories rather than requirements as output;
- Changing the way we introduced the task. In this sense, sometimes ChatGPT seemed unable to understand what was the task that had to be performed in the US. In this case, no progress was made.

The best results were obtained with *Subject_ZsCoTPrompt*. Unlike *Zero_CoTPrompt* and the other versions, in this case, we precisely specified the subject of our US, which in the other results was vague and not perfectly defined. This addition also led to a better understanding of the task contained in the US, as well as a stricter adherence to the US structure.

Even though the results were better with *Subject_ZeroCoTPrompt*, we didn't like the idea of defining the subject of the US from the beginning since this could lead to overfitting, so we decided to use *ReAdjusted_CoTPrompt*.

Ultimately, during the creation of the US, we defined *Enlarged_CoTPrompt*, which is essentially a copy of *ReAdjusted_CoTPrompt* but with 2 examples provided. In most cases, this prompt was unnecessary, as the model seemed capable of understanding the domain from a single US. However, this was not the case for the "Information Systems" and "Computer Vision" domains, which were so vague and generic that the model was unable to understand either the context or subjects.

C. LLaMa fine-tuning approach

In this section we will analyse the steps we followed to generate user stories using a fine-tuned version of LLaMa.

The first thing we did was to choose the version of LLaMa that suited our needs, and our choice fell on LLaMa 3, specifically, LLaMa 3 8B Instruct, which is an instruction tuned version of LLaMa.

For what concerns the fine-tuning dataset, we used the synthetic user stories used for training REFAIR and preprocessed them in such a way as to make them suitable for fine-tuning an LLM. The elements of our dataset should consist of a series of messages between a user and an assistant, i.e., the LLM, in which the user requests a user story characterized by a specific domain and a specific ML task, and the assistant responds with the requested user story. So, we extracted the domain and ML task associated with each individual user story from the dataset in order to generate the question from the user, and finally we used the actual user story as the answer from the assistant.

After that, we used the LLaMa 3 tokenizer to convert each conversation into a set of tokens and decided to keep only sequences with length less than or equal to 128 tokens, resulting in a total of 12360 elements in the dataset. Next, we proceeded with the generation of a train set (9888 elements, 80% of the total), a validation set (1977 elements, 16% of the total) and a test set (495 elements, 4% of the total).

Regarding model fine-tuning, we decided to use the LoRA technique, which allowed us to drastically reduce the number of trainable parameters: we only had to train 1% of LLaMa 3's 8 billion parameters. For more details on the training hyperparameters, you can consult the Jupyter Notebooks in the "llama3-finetuning" folder in the project's GitHub repository.

After training the previously mentioned LoRA, we merged it with the starting model, resulting in a fine-tuned version of LLaMa 3 for our purposes. The model can be found in our Hugging Face repository.

The test set was used to evaluate the performance of LLaMa 3 non-fine-tuned and LLaMa 3 fine-tuned. From the comparison of the two versions of LLaMa (see the results in the same folder mentioned above), we can conclude that the fine-tuning has definitely had an effect, in fact the fine-tuned version generates user stories in a format that precisely follows that of user stories in the dataset.

VII. Evaluation of Results

A. Resuts of the external validity

Instead of evaluating the results solely through ReFair, as detailed in Section V, we conducted a survey to assess the generated USs.

Although some techniques seemed to lead to better results than those proposed in ReFair, an equal number of cases showed that the techniques led to same or even worse results. The surveys indicated that overall, these techniques did not significantly influence how the USs were created.

This lack of impact was also confirmed in the cases of complex domains. This suggests that, for the tasks conducted, these techniques are not particularly useful, or more likely, different LLMs are needed. This is further supported by the results obtained with ChatGPT 4.o, which were significantly better than those obtained with the exact same prompt using ChatGPT 3.5 Turbo, the model used for this experimentation.

Lastly, it should be noted that only 14 survey responses were gathered. Therefore, this low number of responses, does not allow us to draw meaningful conclusions. Thus, deeper research is needed.

B. Addressing RQ_1

To answer $\boldsymbol{RQ_1}$, we evaluated the results of the re-trained ReFair using the metrics defined in Section IV: F1-Score and accuracy for domain classification, and F1-Score and Hamming Loss for ML task classification. For the re-training, we also provided ReFair with a new dataset composed of the old synthetic USs and the dataset created through the few-shot learning approach.

For practical reasons, we evaluated ML task detection using only BERT, while we were able to train ReFair for the domain classification task using every model and embedding technique.

Summary of the results. The results shown in Table I and Table II indicate a decrease in the model's performances.

TABLE I
Domain classifier selection - FS Results

Embedding Technique	Model	F1-Score	Accuracy
TF-IDF	CCCV	0.79	0.79
	SVC	0.79	0.79
	XGBC	0.79	0.79
BERT	XGBC	0.97	0.97
	BC	0.97	0.97
	DT	0.96	0.96
Word2Vec	CCCV	0.89	0.89
	LSVC	0.89	0.89
	LR	0.89	0.89
FastText	CCCV	0.94	0.94
	LSVC	0.94	0.94
	LR	0.94	0.94
Glove	CCCV	0.89	0.89
	LDA	0.89	0.89
	LSVC	0.89	0.89

TABLE II
ML Task Detection - FS Results

Embedding Technique	Model	F1-Score	Hamming Loss
BERT	LP + DT	0.84	0.08
	LP + RF	0.82	0.08
	BR + DT	0.77	0.12

C. Addressing RQ_2

To address $\boldsymbol{RQ_2}$, similar to $\boldsymbol{RQ_1}$, we re-trained ReFair using a new dataset composed by the old dataset and the new dataset created, in this case, through Chain-of-Thought.

Summary of the results.The results, presented in both Table III and Table IV, show decreased performance across all model combinations. However, it is noteworthy that the Hamming Loss remains the same for both the new and old datasets.

TABLE III
Domain classifier selection - CoT Results

Embedding Technique	Model	F1-Score	Accuracy
TF-IDF	CCCV	0.79	0.78
	ET	0.78	0.78
	SVC	0.78	0.78
BERT	XGBC	0.97	0.97
	BC	0.97	0.97
	DT	0.96	0.96
Word2Vec	CCCV	0.90	0.90
	LDA	0.89	0.89
	LSVC	0.89	0.89
FastText	SVC	0.93	0.93
	CCCV	0.94	0.94
	LR	0.94	0.94
Glove	CCCV	0.90	0.89
	LR	0.90	0.89
	LDA	0.89	0.89

TABLE IV
ML TASK DETECTION - COT RESULTS

Embedding Technique	Model	F1-Score	Hamming Loss
BERT	LP + DT	0.85	0.07
	LP + RF	0.82	0.08
	BR + DT	0.76	0.12

D. Addressing RQ_3

To address $\boldsymbol{RQ_3}$, we re-trained once more REFAIR using a new dataset composed by the old dataset and the new dataset created, in this case, through the usage of the fine-tuned version of LLaMa 3.

Summary of the results.The results, presented in both Table V and Table VI, show slightly higher performance across all model combinations. This is probably because the US generated by Llama 3 are very similar to those already in the original dataset.

TABLE V
DOMAIN CLASSIFIER SELECTION - FINE-TUNED LLAMA RESULTS

Embedding Technique	Model	F1-Score	Accuracy
TF-IDF	ET	0.80	0.80
	CCCV	0.80	0.80
	SVC	0.80	0.80
BERT	XGBC	0.98	0.98
	BC	0.98	0.98
	DT	0.98	0.98
Word2Vec	LR	0.91	0.91
	CCCV	0.91	0.91
	SVC	0.90	0.90
FastText	CCCV	0.95	0.95
	LR	0.95	0.95
	LSVC	0.95	0.95
Glove	LR	0.91	0.91
	CCCV	0.91	0.91
	LSVC	0.90	0.90

TABLE VI
ML TASK DETECTION - FINE-TUNED LLAMA RESULTS

Embedding Technique	Model	F1-Score	Hamming Loss
BERT	LP + DT	0.86	0.07
	LP + RF	0.84	0.08
	BR + DT	0.79	0.11

E. Answering RQ_4 and RQ_5

Finally, $\boldsymbol{RQ_4}$ and $\boldsymbol{RQ_5}$ were not addressed, as they were more closely related to academic optional purposes rather than serving as practical alternatives to the models implemented by REFAIR.

VIII. PROBLEMS ENCOUNTERED

During the course of the project, we encountered various problems, categorized into three clusters:

- Surveys: We only reached 14 participants, which might be too few to obtain significant results;
- Performance Variation: Due to the innate variability of LLMs, we encountered inconsistent performance levels with ChatGPT, necessitating multiple response regenerations in certain situations;
- Hardware Limitations: Due to limited hardware resources, we were unable to explore various combinations of hyperparameters when fine-tuning LLaMa 3. Specifically, the GPUs we had available had insufficient VRAM, forcing us to use very low batch sizes.

IX. IMPLICATIONS OF THE RESULTS

As previously mentioned, the results obtained by re-training REFAIR demonstrate how these techniques could be useful in creating the dataset. They enable the model to better understand the context, leading to greater variation in the subjects and content of the USs. However, from a statistical point of view, namely comprehensibility, realism and actionability, the new USs generated are comparable to the old ones.

In this regard, we should assess whether it is worthwhile to spend time creating examples that merely reduces the likelihood of an overfitted model.

Nonetheless, the use of these techniques could be enhanced by employing different LLMs, such as ChatGPT 4.o, where the generated USs appeared to be not only more varied but also better in terms of consistency with the required tasks and domain, as well as the quality of the US itself.

X. THREATS TO VALIDITY

In this section we will see at the steps that will accompany the validation of the new dataset and how threats, that might occur during validation, will be mitigated.

External validity. A problem concerning external validity is related to the inability to provide the survey to real-world machine learning engineers. We aim to mitigate this issue by providing the survey to participants who have at least a basic knowledge about machine learning.

Internal validity. Threats in this case are related to the type of surveys, where every user will have USs from both the new and old datasets. In this scenario, users could potentially learn or recognize patterns or relationships between USs from the same dataset and thus prefer the USs from the other one. To mitigate this threat, we will try to formulate the questions in a way that avoids patterns.

At the same time, we will provide users with quick surveys to complete in order to avoid the effects of boredom or fatigue.

Construct validity. We will try to minimize the threat associated with participants discovering the hypothesis, as this could potentially lead to biased results. This will be achieved by simply not revealing the hypothesis to the participants or letting it be understandable by the survey itself.

At the same time, the surveys will be designed objectively, thus not preferring one technique over another.

XI. Conclusions

Finally, considering the findings, the techniques used still proved to be valid as they manage to be more generic. Future investigations may involve trying different LLMs.

References

1 Ferrara, C., Casillo, F., Gravino, C., De Lucia, A., and Palomba, F., "Refair: Toward a context-aware recommender for fairness requirements engineering," 2023.

2 Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

3 Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D. *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

4 Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y., "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

5 Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V., "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

6 Rajbhandari, S., Rasley, J., Ruwase, O., and He, Y., "Zero: Memory optimization towards training A trillion parameter models," *CoRR*, vol. abs/1910.02054, 2019. [Online]. Available: http://arxiv.org/abs/1910.02054

7 Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W., "Lora: Low-rank adaptation of large language models," *CoRR*, vol. abs/2106.09685, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

8 Fabris, A., Messina, S., Silvello, G., and Susto, G. A., "Algorithmic fairness datasets: the story so far." *Data Min. Knowl. Discov.*, vol. 36, no. 6, pp. 2074–2152, 2022. [Online]. Available: http://dblp.uni-trier.de/db/journals/datamine/datamine36.html#FabrisMSS22

9 Duran-Silva, N., Fuster, E., Massucci, F. A., Parra-Rojas, C., Quinquillà, A., Roda, F., Rondelli, B., Bovenzi, N., and Toietta, C., "A controlled vocabulary for research and innovation in the field of artificial intelligence (ai)," Zenodo, Feb. 2021.