NOVEL APPROACHES IN SYNTHETIC DATASET GENERATION

Daniele Galloppo - d.galloppo@studenti.unisa.it
Mario Peluso - m.peluso37@studenti.unisa.it
Luca Morelli - l.morelli6@studenti.unisa.it

TABLE OF CONTENTS

01.

INTRODUCTION

Brief introduction to the concept of fairness & ReFair

02.

OBJECTIVE

Goals of the report

03.

RESEARCH QUESTIONS

Presentation of the RQs

04.

DATASETS GENERATION

Hands on the Prompt Engineering

05.

EVALUATION OF RESULTS

Results and responses to the RQs

ARTIFICIAL INTELLIGENCE (AI)

01.

INTRODUCTION

Brief introduction to the concept of fairness & ReFair

BASIC CONCEPTS

FAIRNESS

Has to do with the set of requirements, methods, and techniques to let an AI solution act “fairly”

ETHICS

Branch of Philosophy. It has to do with moral aspects of humanity

DOMAIN SPECIFICITY

Ethical concerns and fairness are domain specific, namely they depend on the domain

PROBLEM VS. SOLUTION

PROBLEM

Implementing sustainable fairness in AI systems could be particularly problematic to achieve

SOLUTION

REFAIR is a classification model that recommends sensitive features that, if not correctly treated, may lead to unfair and biased models

ARTIFICIAL INTELLIGENCE (AI)

REFAIR IN A NUTSHELL

APPLICATION DOMAIN CLASSIFICATION

Determine the most probable application domain from a selection of 34 domains

MACHINE LEARNING TASKS CLASSIFICATION

Determine the ML-task(s) likely to be employed when implementing the US

02.

OBJECTIVE

Goals of the report

GOALS

ISSUE

Synthetic dataset might not contain realistic USs

EXPECTATIONS

Create alternative **Datasets** that contain **realistic representations** of USs

GOALS

Use **prompt engineering** techniques to get alternative Dataset

ARTIFICIAL

CE

(AI)

/ (AI)

03.

RESEARCH QUESTIONS

Presentation of the RQs

<<<< RESEARCH QUESTIONS >>>>

RQ1

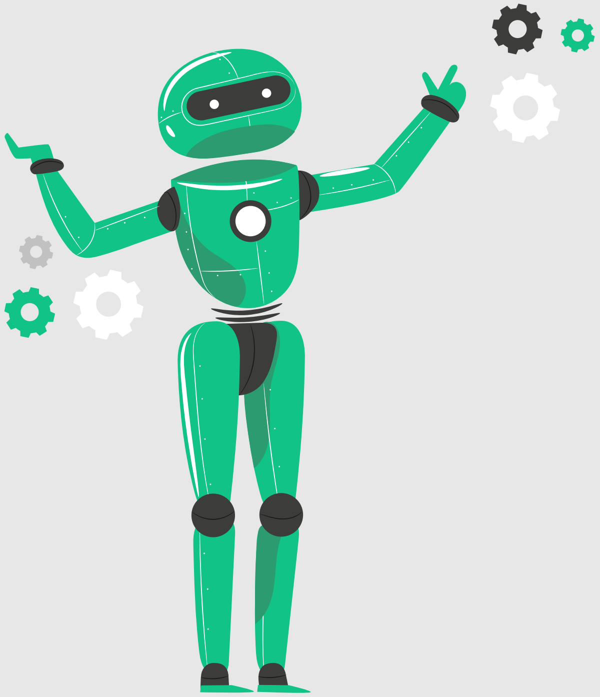
To what extent does the synthetic dataset created through **few-shot learning** technique impact the model's performance?

RQ2

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

RQ3

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



HOW TO EVALUATE THE RQS

METRICS FOR THE EVALUATION

RQ1

RQ2

RQ3

Use **F1-Score** and **accuracy** for domain classification and **F1-Score** and **Hamming Loss** for ML Task classification

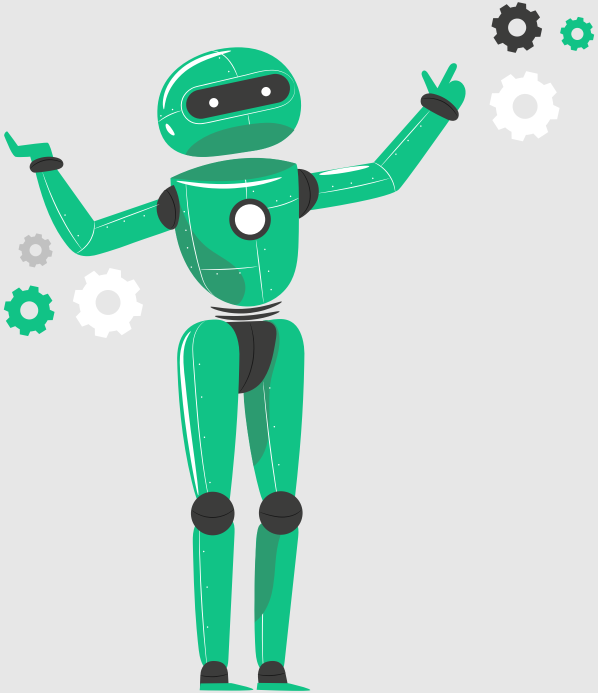
<<<< OPTIONAL RESEARCH QUESTIONS >>>>

RQ4

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

RQ5

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

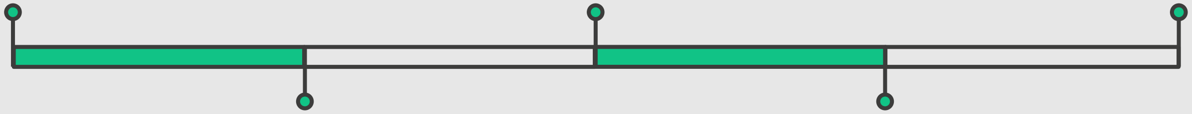


04.

DATASETS GENERATION

Hands on the Prompt Engineering

FEW SHOT LEARNING



INITIAL_FSPROMPT

Basic prompt,
composed by 3 USs
and a request

ADJSUSTED_FSPROMPT

Prompt containing
informations about
the task

RANDOM_FSPROMPT

Composed by USs of
different domains

ENLARGED_FSPROMPT

Like the previous prompt
but providing 5 examples

READJUSTED_FSPROMPT

Like Adjusted_FSPrompt
but clearer

INITIAL_FSPROMPT

DESCRIPTION

Prompt composed by 3 USs and a request

PROBLEMS

The results were **chaotic** and **lacked of a logical connection** to the examples provided. We needed to specify a **max length** for each US

ARTIFICIAL INTELLIGENCE (AI)

ADJUSTED_FSPROMPT

DESCRIPTION

Prompt that adds informations such as the **task** that the US had to describe and the **field** in which the task had to be completed

PROBLEMS

This prompt resolved the previous problems, but the USs were **too technical**

ARTIFICIAL INTELLIGENCE (AI)

RANDOM_FSPROMPT

DESCRIPTION

Prompt in which each USs is related to a **different domain**, while keeping the same questions' structure

PROBLEMS

This prompt was not always able to correctly **identify the subject** of the US or the specified **domain**.

ARTIFICIAL INTELLIGENCE (AI)

ENLARGED_FSPROMPT

DESCRIPTION

Prompt in which we have 5 USs rather than only 3. It is an evolution of Adjusted_FSPrompt

PROBLEMS

ChatGPT had difficulties in understanding the subject and domain of the US to generate but also started **mixing the tasks**

ARTIFICIAL INTELLIGENCE (AI)

READJUSTED_FSPROMPT

DESCRIPTION

Same prompt of Adjusted_FSPrompt, namely 3 US-Examples of the same domain

PROBLEMS

Same problems of Adjusted_FSPrompt, but the prompt is **clearer**

ARTIFICIAL INTELLIGENCE (AI)

DOMAIN_FSPROMPT

DESCRIPTION

The US-Examples are related to the same domain but different in respect to the domain of the US to generate

!PROBLEMS

The results were overall good. The USs were less technical while still preserving a **good level of quality**

ARTIFICIAL INTELLIGENCE (AI)

DOMAIN_FSPROMPT IN DETAILS

[High-level machine learning task] OR [Low-level machine learning technique] in the field of [machine learning] OR [natural language processing]. The example is: [Example of the first US in the domain type X].

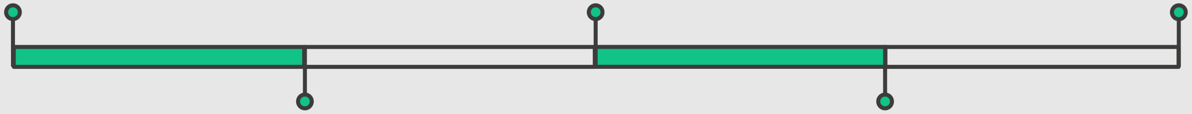
[High-level machine learning task] OR [Low-level machine learning technique] in the field of [machine learning] OR [natural language processing]. The example is: [Example of the second US in the domain type X].

[High-level machine learning task] OR [Low-level machine learning technique] in the field of [machine learning] OR [natural language processing]. The example is: [Example of the third US in the domain type X].

Following the user story structure, provide me with [number of US to generate] specific user stories for the [High-level machine learning task] OR [Low-level machine learning technique] in the field of [machine learning] OR [natural language processing] in the [Domain type Y] domain based on the above examples.

It has to be noticed that the Domain X and Domain Y are part of the same **domain cluster**

CHAIN-OF-THOUGHT PROMPT



READJUSTED_COTPROMPT

Prompt containing a question, a reasoning and an US for example

DOMAIN_COTPROMPT

Like ReAdjusted_CoTPrompt but the domain of the US is different from the one we want to generate

ZERO_COTPROMPT

Prompt with no examples and an invite to reason step by step

SUBJECT_ZSCOTPROMPT

Like Zero_CoTPrompt but specifying the subject

ENLARGED_COTPROMPT

Like ReAdjusted_CoTPrompt but with one more example

READADJUSTED_COTPROMPT

DESCRIPTION

Prompt composed by a single question, an answer that describes the **reasoning behind the US** and the US provided as example.

!PROBLEMS

The results are good, similarly to the ones provided by Domain_FSPrompt

ARTIFICIAL INTELLIGENCE (AI)

READJUSTED_COTPROMPT IN DETAILS

Q: Create an US with [High-level machine learning task] OR [Low-level machine learning technique] in the field of [machine learning] OR [natural language processing] in the [Domain type X] domain.

A: [reasoning steps based on the question]. The user story can be: [Example of the third US in the domain type X].

Q: Create 5 US with [High-level machine learning task] OR [Low-level machine learning technique] in the field of [machine learning] OR [natural language processing] in the [Domain type X] domain.

Even tho the prompt that defines the US is the same, with ReAdjusted_CoTPrompt we define a **precise Q&A scheme.** The reasoning steps can **strongly vary** from a prompt to another

DOMAIN_COTPROMPT

DESCRIPTION

Same structure of ReAdjusted_CoTPrompt but the domain of the US-Example is different.

PROBLEMS

Results **slightly worse** than the ones provided by the previous prompt

ARTIFICIAL INTELLIGENCE (AI)

ZERO_COTPROMPT

DESCRIPTION

No examples are provided. Only a suggestion to slightly reason on the response

PROBLEMS

Without defining the context, the prompt led to **critical and unusable results**

ARTIFICIAL INTELLIGENCE (AI)

SUBJECT_ZSCOTPROMPT

DESCRIPTION

Also in this prompt, **no examples** are provided, but we specify who might be the subject of the US

PROBLEMS

Defining the subject improved the understanding of the context, but the prompt might lead to **overfitting**.

ARTIFICIAL INTELLIGENCE (AI)

ENLARGED_COTPROMPT

DESCRIPTION

Same structure of ReAdjusted_CoTPrompt but we provided **two examples**.

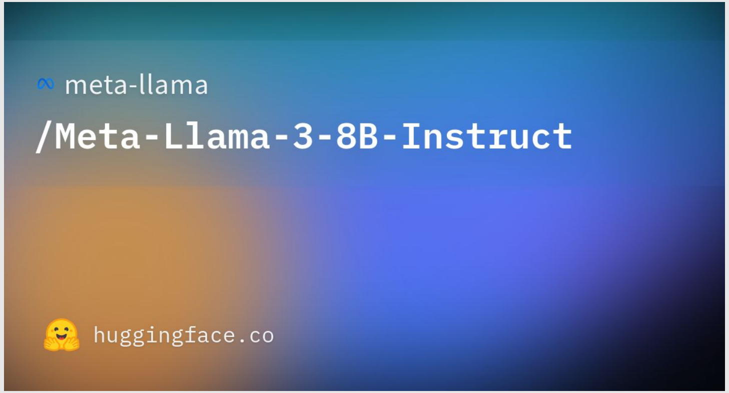
PROBLEMS

Unnecessary in most of the cases. Needed only for vague and generic domains

ARTIFICIAL INTELLIGENCE (AI)

LLAMA 3 FINE-TUNING: THE MODEL

- LLaMa 3 architecture
- Number of parameters: 8 billion
- Instruction tuned version



meta-llama

/Meta-Llama-3-8B-Instruct

huggingface.co

/ / / / / LLAMA 3 FINE-TUNING: THE DATASET / / / / /

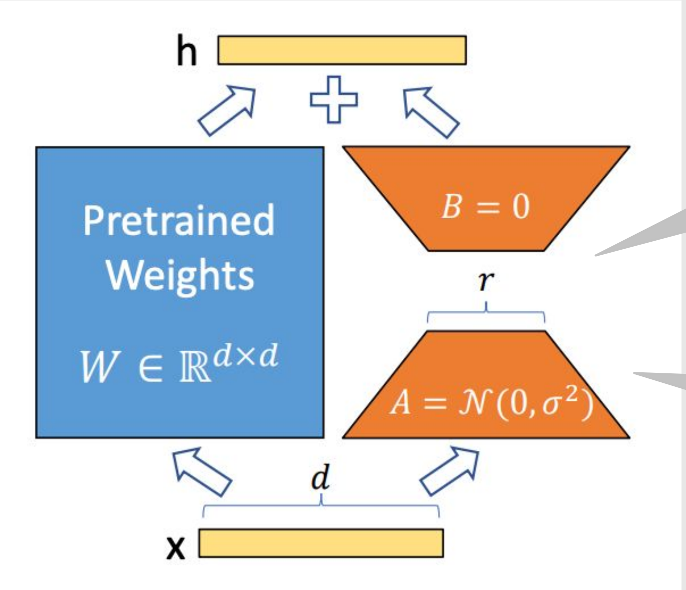
```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a helpful AI assistant<|eot_id|><|start_header_id|>user<|end_header_id|>

Considering the following machine learning technique: neural machine translation in the field of machine learning. Can you provide me with a specific user story for the following application domain? Plant Science<|eot_id|><|start_header_id|>assistant<|end_header_id|>

As a plant scientist, I want to use neural machine translation to understand and translate plant research papers and reports from different languages, so that I can stay up-to-date with the latest plant research and collaborate with researchers from around the world.<|eot_id|>
```

Starting point: ReFair original dataset of USs.

We created a conversation between user and assistant for each US.

Prompt format: Meta Llama 3 Instruct.

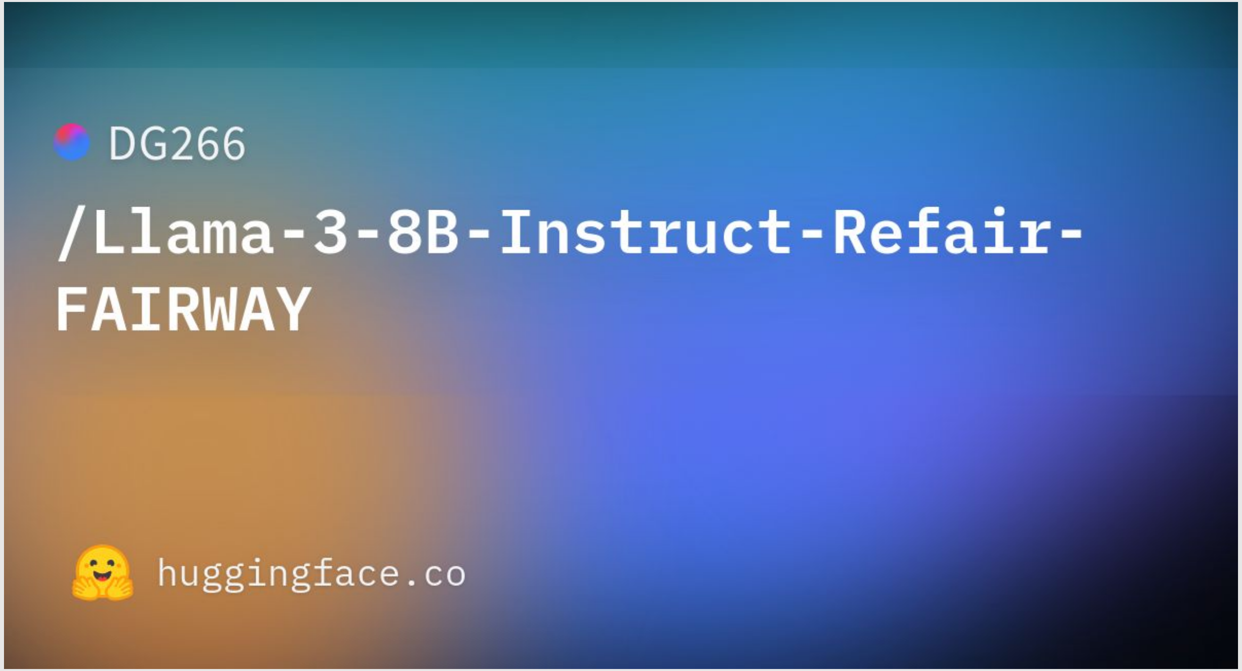
LLAMA 3 FINE-TUNING: THE LORA TRAINING TECHNIQUE


h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

We only had to train 1% of the 8 billion parameters!

The LoRA was merged with the original model.

LLAMA 3 FINE-TUNING: THE RESULT



DG266

/Llama-3-8B-Instruct-Refair-FAIRWAY

huggingface.co

05.

EVALUATION OF RESULTS

Results and responses to the RQs

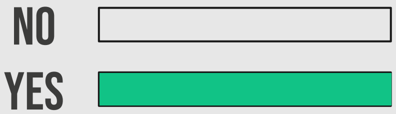
EXTERNAL VALIDATION: PARTICIPANTS

14

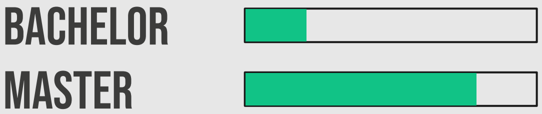
NUMBER OF PARTICIPANTS

B2

AVERAGE ENGLISH LEVEL



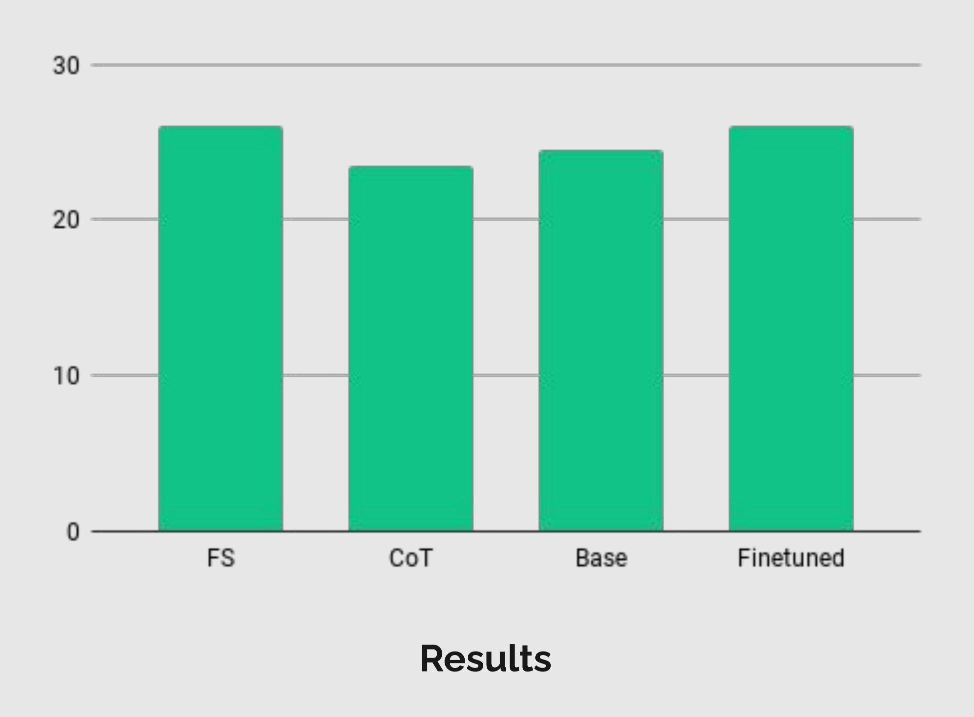
KNOWS WHAT IS AN US



TITLE OF STUDY

> > > >

EXTERNAL VALIDATION: RESULTS



RESULTS

Overall, the results showed that the USs generated with the proposed techniques are **comparable** in term of **realism**, **comprehensibility** and **actionability** to those of existing dataset. Additionally, we have to consider that the **domains had no impact** on determining the best technique.

REFAIR RETRAINING: ADDING FEW SHOT LEARNING USs

DOMAIN CLASSIFICATION

Shows **slightly decreased performance** in respect to the original results of **ReFair**

Also for ML task, performance shows a **slight deterioration**

ML TASK DETECTION

ARTIFICIAL INTELLIGENCE (AI)

REFAIR RETRAINING: ADDING CHAIN-OF-THOUGHT USs

DOMAIN CLASSIFICATION

ML TASK DETECTION

Just as with the FS results, performance for both classifier are **slightly worse** than the original results

REFAIR RETRAINING: ADDING LLAMA 3 FINE-TUNE USs

DOMAIN CLASSIFICATION

ML TASK DETECTION

Slightly higher performance across all model combinations and embedding techniques

ARTIFICIAL INTELLIGENCE (AI)

PROBLEMATICS

SURVEYS

Low number of participants to the surveys

››››

CHATGPT

Performance variation caused by the innate variability of LLMs

››››

HARDWARE

Limited hardware resources, especially for Llama 3 fine-tuning

››››

ARTIFICIAL INTELLIGENCE (AI)

CONCLUSIONS

ARTIFI
INTELL
(AI)

“There cannot be artificial intelligence without data.
There cannot be accurate artificial intelligence without good data.”

<<<<

THANKS

ARTI
CIAL
INTE
IGEN
(AI)

/(AI)/(AI)/