Synthetic Data for Identifying Inclusive Language (Case Study: Job Descriptions in Italian)

Tommaso Romano'
Department of Computer Science,
Università degli Studi di
Milano, Italy
tommaso.romano@studenti.unimi.it

Fatemeh Mohammadi
Department of Computer Science,
Università degli Studi di
Milano, Italy
fatemeh.mohammadi@unimi.it

Paolo Ceravolo
Department of Computer Science,
Università degli Studi di
Milano, Italy
paolo.ceravolo@unimi.it

***Abstract*—Using a comprehensive list of job titles, we propose a framework to automatically generate job descriptions in Italian. This synthetic data is then used in a Large Language Model to detect inclusive language in job postings. Finally, we compare the results of this synthetic dataset with real data. Our study demonstrates that the data format and prompting method significantly impact performance. Additionally, we identify limitations and key considerations for unifying synthetic data with real data for fine-tuning purposes. We also propose improvements to the framework and provide guidelines for effectively integrating these two types of data. The novelty of our work is generating and integrating synthetic data due to the scarcity of annotated Italian job descriptions, thereby improving the training of Large Language Models (LLMs) tailored specifically for Italian.**

I. Introduction

Job advertisements serve as a crucial first interaction with an organization, allowing job seekers to assess their compatibility with the job and the organization [1]. Given this context and the legal requirements for equal treatment of all candidates, it is increasingly important that job advertisements are as inclusive as possible and ensure that no group of qualified candidates feels excluded.

Large Language Models (LLMs) have seen significant growth and adoption across various sectors in recent years. As of 2023, the market for generative AI, which includes LLMs, has been projected to reach USD 36.06 billion by 2024, with an anticipated annual growth rate of 46.47% from 2024 to 2030 [2]. This rapid expansion is driven by increasing investments and advancements in AI technology, particularly in natural language processing (NLP) and generation capabilities.

Recent advances in natural language processing, particularly with large language models (LLMs), have enabled us to fine-tune models that can automatically detect and correct non-inclusive language, thereby making job descriptions free of discrimination, particularly gender stereotypes. However, a major challenge in working with LLMs is the data we need to fine-tune them. In this case, we need a dataset of labeled job descriptions in Italian. Another challenge is to ensure a balanced dataset that includes all the job titles that can make the prediction more accurate.

In light of these challenges, we generate a dataset of synthetic job advertisements in Italian. Synthetic text is an artificially generated text based on a use-case relevant context that reflects the relevant meaning for statistical analysis in the intended context [4]. The use of generative data has advantages such as structural similarity, information relevance, and subjective assessment [5]. For these reasons, synthetic datasets have been used for training and testing algorithms [6].

This paper aims to investigate the use of modular prompting and job title masking to generate synthetic data for job advertisements in Italian. We will then classify these synthetic data as either masculine, feminine, or neutral using LLMs and evaluate their performance. Furthermore, we will compare the performance metrics of synthetic data with a real seed dataset. Thus, our study addresses three main questions:

RQ1: How can we generate synthetic data for job descriptions?

RQ2: What is the difference in performance between synthetic and real data?

RQ3: Is synthetic data reliable enough to be used as training data for fine-tuning LLMs?

This study is one of the first to focus on Italian-language texts, taking into account the unique aspects of Italian grammar such as articles (”il”, ”la”, ”i”, ”le”, etc., equivalent to ”the” in English) and gendered noun endings (”-a”, ”-o”, etc.). A large corpus of Italian job advertisements has been produced, with a diverse and balanced mix of masculine, feminine, and neutral endings.

Due to the limited availability of datasets in Italian—particularly those addressing discrimination and inclusive language—and the subsequent lack of work on fine-tuning LLMs for Italian, our study paves a promising path in the literature for detecting inclusive language in Italian.

The paper is structured as follows: after the introduction, Section 2 describes the research background in the field of synthetic data generation and discriminatory and inclusive language in job descriptions. Section 3 presents the methodology used to generate synthetic data and assess their reliability for use in fine-tuning LLMs. Then, in section 4, we discuss the results of our methodology and the performance evaluations we obtained. We also answer the research questions we have raised. Finally, Section 5 presents the discussion and conclusions, followed by a section with an outlook on future research.

II. Background and Related Works

Large Language Models (LLMs) such as GPT-4 and Llama have demonstrated remarkable capabilities in natural language understanding and generation. However, their performance heavily relies on the quality and quantity of the data used for training. Synthetic data generation has emerged as a promising approach to augment training datasets, offering potential benefits in enhancing LLM performance. In this section, we have a quick review of recent research on synthetic data generation techniques and their effects on LLM performance.

Synthetic data generation involves creating artificial datasets that mimic real-world data. Various techniques have been explored to produce synthetic data, including data augmentation which consisted of methods like text paraphrasing, back-translation, and word substitution have been employed to enhance existing datasets [7]. These approaches generate diverse textual variations, potentially improving LLM robustness.

Another technique is using generative models such as GANs (Generative Adversarial Networks) and VAEs (Variational Autoencoders) that have been adapted for text to produce realistic synthetic data [8]. These models can create new examples based on learned distributions from the original data. Also, simulation-based approaches which involve simulating data generation processes to produce text that mimics specific scenarios or contexts [9] are a good approach for generating synthetic data.

Synthetic data can significantly diversify training datasets, which is crucial for reducing over-fitting and improving generalization. Research by [10] demonstrated that incorporating synthetic data into training regimes can enhance model robustness across various linguistic tasks. The diversity introduced by synthetic data helps LLMs handle a broader range of inputs, leading to better performance on unseen examples.

For specialized or low-resource languages and domains, acquiring sufficient real-world data can be challenging. Synthetic data generation offers a solution to this problem. A study by [11] highlighted that synthetic data could effectively supplement real-world data in low-resource scenarios, resulting in improved performance on tasks like machine translation and text classification.

Synthetic data can be used to address biases present in training datasets. Research by [12] showed that generating synthetic examples with diverse demographic characteristics helped mitigate biases in LLMs, leading to fairer model predictions and more balanced performance across different groups.

Despite its potential, synthetic data generation presents several challenges such as quality control. Poorly generated data can introduce noise or artifacts, potentially harming LLM performance [13]. Also, there is a risk that LLMs might overfit synthetic data if it dominates the training set. Balancing synthetic and real data is essential to maintain model performance [14].

Recent advancements in synthetic data generation include integrating reinforcement learning for iterative data improvement [15] and developing more sophisticated generative models for text [16]. Future research should focus on refining these techniques, addressing ethical considerations, and exploring hybrid approaches that combine synthetic and real data effectively.

Synthetic data generation has the potential to significantly impact the performance of Large Language Models by enhancing data diversity, addressing data scarcity, and mitigating biases. However, challenges such as data quality, over-fitting risks, and evaluation complexities must be addressed. Ongoing research and technological advancements will likely continue to refine synthetic data techniques and their application in improving LLM performance.

III. Methodology

Here we explain our methodology for generating synthetic data and the method we used to evaluate this data against a real data set.

A. Creating a Job List in Italian

Our primary technique for generating synthetic data is to mask job titles and replace them with different endings. To achieve this, we compiled a comprehensive list of job titles (29 different job titles), which was then extended by a native speaker expert to include different possible endings in Italian (136 different endings). For example, the English job title ”nurse” can be rendered in eight different ways in Italian job descriptions: ”infermiere” (masculine and non-inclusive), ”infermiera” (feminine and non-inclusive), ”infermier*”, ”infermiere/a”, ”infermier”, ”infermiera/e”, ”infermiera o infermiere”, and ”infermiere o infermiera” (neutral and inclusive). The native expert also labeled these job titles as masculine, feminine, and neutral based on the different endings.

B. Generating Synthetic Data

After making a list of job titles with different endings, we make a list of phrases that are used in job descriptions to refer to a job title. For example
English: ”We are looking for [JOB] for our company.”
Italian: ”Stiamo cercando [JOB] per la nostra azienda.”
English: ”Newly opened office seeks [JOB] with experience.”
Italian: ”Ufficio di nuova aprtura cerca [JOB] con esperienza.”
English: ”Hours for [JOB]: variable number of hours from 20 to 36 per week, depending on availability.”
Italian: ”Orario per [JOB]: monte ore variabile da 20 a 36 ore settimanali, in base alla disponibilità.”

By creating a repository of these sentences and substituting different job titles within them, we generated our synthetic dataset of 1224 rows. Each row is labelled according to the specific job title that was substituted into it.

C. Prompt Composition

A fundamental thing when working with LLMs is a precise, clear and good prompt. In this case, we proposed a workflow based on modular prompting, which can generate different

combinations of modules to make a complete prompt. In modular prompting, each prompt consists of 6 modules:
Context (C) is mainly the role we assigned to the LLM. This module is always the first element of the prompt. For example: "*Sei un assistente che legge ed analizza un testo italiano.*" (You are an assistant who reads and analyses an Italian text.)
Instruction (I) is a clear instruction for LLM to do the task. For example: "*Il tuo obiettivo è identificare se il testo ed in particolare se i nomi di professioni si riferiscono a maschile o femminile o neutro.*" (Your goal is to identify whether the text and in particular whether the names of professions refer to masculine, feminine, or neutral.)
Evidence (E) which is the evidence or the data that we want to classify or predict.
Question (Q) is expressing the instruction as a short question. For example: "*Questa è la domanda: Il testo si riferisce a maschile o femminile o neutro?*" (This is the question: Does the text refer to masculine or feminine or neutral?)
Closing instruction (X) the closing instruction to provide the classification, or
Call-to-thinking (T) a call-to-thinking to motivate the LLM to consider the task carefully, including a mention of the question; an instruction to classify; and an ask to later explain the classification. This prompt ending is an alternative to X. For example: "*Rispondi con un singolo label "maschile", "femminile" o "neutro" per il genere dei presenti nomi di professioni, altrimenti mettere "neutro".*" (Reply with a single label "masculine", "feminine" or "neutral" for the gender of the present profession names, otherwise put "neutral".)

Based on this structure, we chose 4 different combinations (CIEQ, CIET, CIEQT, CEIQ) to make 40 different prompts. Table 1 shows one example for each combination. Figure 1 shows the workflow that we followed for generating synthetic data and prediction using LLM.

The prompts constructed using the above structures followed the Few-Shot Learning (FSL) strategy in that they provided examples of different endings for job titles. To include another popular strategy for prompting large language models (LLMs), Zero-Shot Learning (ZSL), we also added 5 prompts for this approach. In total, we tested 45 prompts using the two different strategies.

D. Prediction and evaluation

After generating our synthetic data and creating our prompts, we want to use them as input for an LLM. We choose llama3:instruct as our model. The main reason for choosing Llama 3 is that it is a new state-of-the-art model optimized for dialogue/chat use cases and outperforms many of the available open-source chat models on common benchmarks [18]. For evaluation, after running the Llama3 model on synthetic data, we calculate *accuracy*, *F1-score*, *precision* and *recall*.

E. Comparative Analysis

After running the prompts on the synthetic dataset, we used real seed data to compare the performance of the LLM on these two datasets. Our seed dataset consists of 99 lines of job advertisements in Italian, collected and annotated by domain-specific experts in law and linguistics. The labels in this dataset are 'inclusive' and 'non_inclusive'. To facilitate comparison, we have standardised the labels: a 'non_inclusive' label indicates that the job description refers to only one gender (male or female), while an 'inclusive' label means that it covers both genders or is gender-neutral. A native human expert applied these label changes, and then we randomly ran 22 prompts on the dataset. The results are discussed in the following section.

TABLE I
SOME EXAMPLES OF COMPLETE PROMPTS USING IN LABEL PREDICTION

Prompt Text	Structure
Sei un assistente che legge ed analizza un testo italiano. Il tuo obbiettivo è identificare se il testo ed in particolare se i nomi di professioni si riferiscono a maschile o femminile o neutro. Ad esempio, impiegat*, impiegato/a, e impiegatÃ sono tutti esempi di parole neutre. Questa è la domanda: "Il testo si riferisce a maschile o femminile o neutro?"	CIEQ
Sei un assistente che legge ed analizza un testo italiano. Il tuo obbiettivo è identificare se il testo ed in particolare se i nomi di professioni si riferiscono a maschile o femminile o neutro. Ad esempio, impiegat*, impiegato/a, e impiegatÃ sono tutti esempi di parole neutre. Puoi rispondere solamente con un label "maschile", "femminile" o "neutro".	CIET
Sei un assistente che legge ed analizza un testo italiano. Il tuo obbiettivo è identificare se il testo ed in particolare se i nomi di professioni si riferiscono a maschile o femminile o neutro. Questa è la domanda: "Il testo si riferisce a maschile o femminile o neutro?" Puoi rispondere solamente con un label "maschile", "femminile" o "neutro".	CIEQT
Sei un assistente che legge ed analizza un testo italiano. Ad esempio, impiegat*, impiegato/a, e impiegatÃ sono tutti esempi di parole neutre. Il tuo obbiettivo è identificare se il testo ed in particolare se i nomi di professioni si riferiscono a maschile o femminile o neutro. Questa è la domanda: "Il testo si riferisce a maschile o femminile o neutro?"	CEIQ

IV. DISCUSSION

This section presents the prediction results using Llama3 and evaluates the performance metrics. We also compare the performance of llama3 with synthetic and real seed datasets. Finally, we answer the research questions raised in the introduction.

A. Synthetic data Analysis

As mentioned in the methodology section, the synthetic data consist of 1224 lines covering 29 different job titles. Of these lines, 171 are categorised as female, 162 as male and 891 as neutral. As explained in the methodology, the identified gender of each sentence in the synthetic dataset serves as an indicator for detecting discrimination and bias in job advertisements. For example, if a job description uses only masculine or feminine forms of a job title, it can be concluded that the advertisement is not inclusive as it excludes one gender from consideration.

B. Synthetic data evaluation

We measured *accuracy*, *F1-score*, *precision* and *recall* for all prompts. Because we have three labels (masculine/feminine/neutral), we have to use the weighted average of

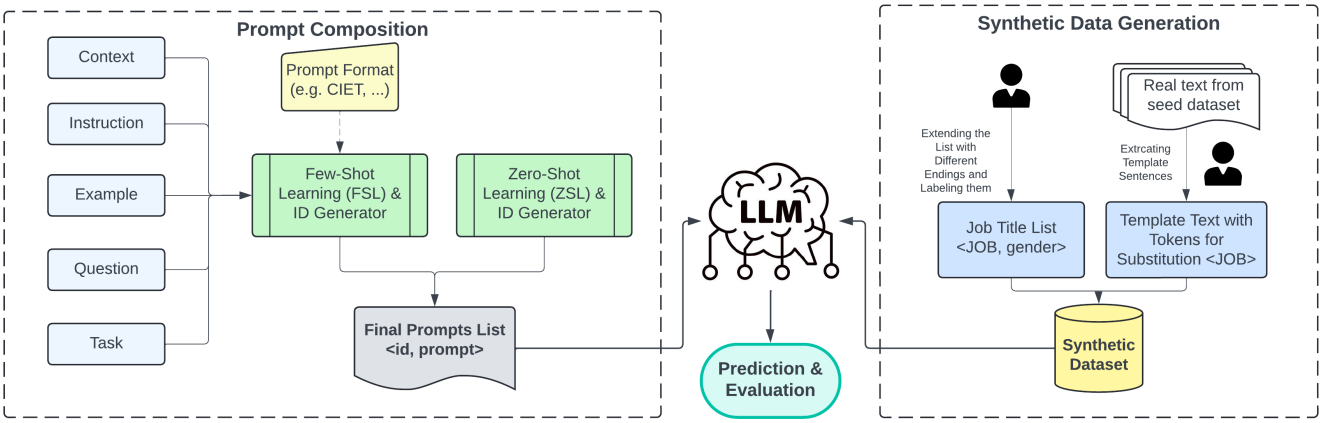

Fig. 1. Workflow of our methodology for generating synthetic data and prediction using LLM

each metric (except accuracy) for the evaluation. This is simply the average of the metric values for each class, weighted by the support for that class. For example, the weighted average precision for a 3-label classification problem is calculated as:

$$\text{Weighted Average Precision} = \frac{P_A \cdot S_A + P_B \cdot S_B + P_C \cdot S_C}{S_A + S_B + S_C} \quad (1)$$

where:

- P_A, P_B, P_C are the precision values for labels A, B, and C respectively.
- S_A, S_B, S_C are the support (or frequency) of labels A, B, and C respectively.

Figure 2 shows the heat map of performance metrics for 45 prompts that we tested on synthetic data. As shown in the figure, the Zero-Shot Learning strategy for prompts has weak performance in our case, and Few-Shot Learning has better performance in many different structures of modular prompts. Between different structures of modular prompts with FSL strategy, CIET is the weakest for Llama3, and CIEQ and CIEQT show higher performance. In general, precision is the metric with a higher value in most prompts.

C. Seed data evaluation

Figure 3 shows the heatmap of performance metrics for the 22 prompts tested on the seed data. By comparing this figure with the corresponding prompts in Figure 2, we observe a decline in overall performance on the seed dataset. Of the 22 prompts, only 5 achieved acceptable accuracy: c0_e0_i1_q0, c0_i0_e0_t1, c0_i1_e0_t1, c0_e0_i0_q0, and co_i0_e0_q0_t1. By numbers in Prompt, we show different combinations that we used for constructing prompts. For example, i0 is the first option for use as instruction in the final prompt. The prompt c0_e0_i1_q0 had the best *recall*, *precision* and *accuracy*, while c0_i0_e0_t1 had the highest *F1-score*. The worst performance was typically observed in the CIEQT modular structure, where 7 out of 8 prompts showed weak performance.

This decrease in performance on the seed dataset can be attributed to two main reasons:

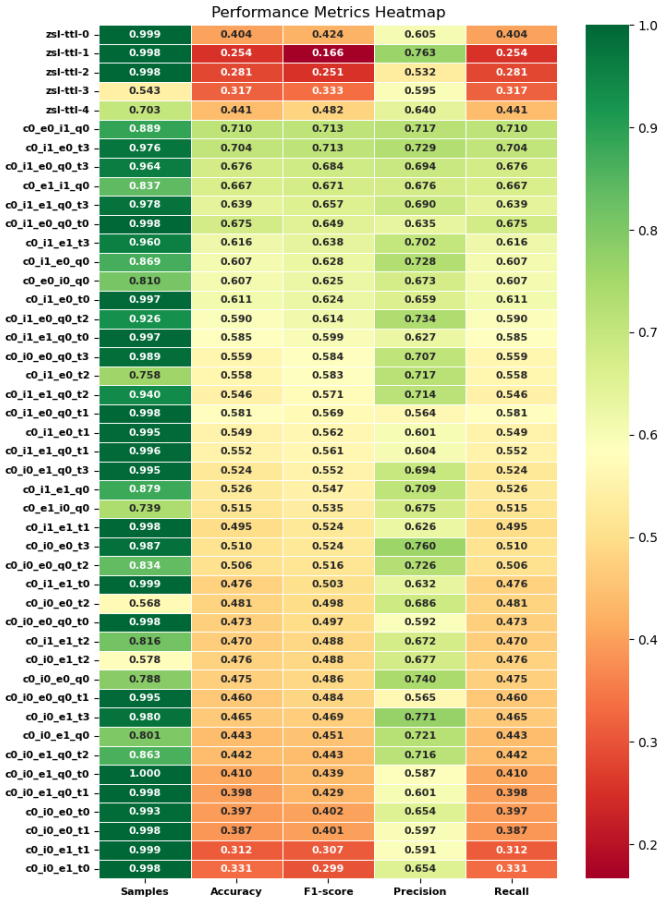

Fig. 2. The heatmap of performance metrics for synthetic data

1) **Text Structure Complexity:** The seed data contains more complex structures with multiple sentences, some of which are less focused on the candidate or the work environment. In contrast, the synthetic data set contains concise sentences filled with different job titles, making

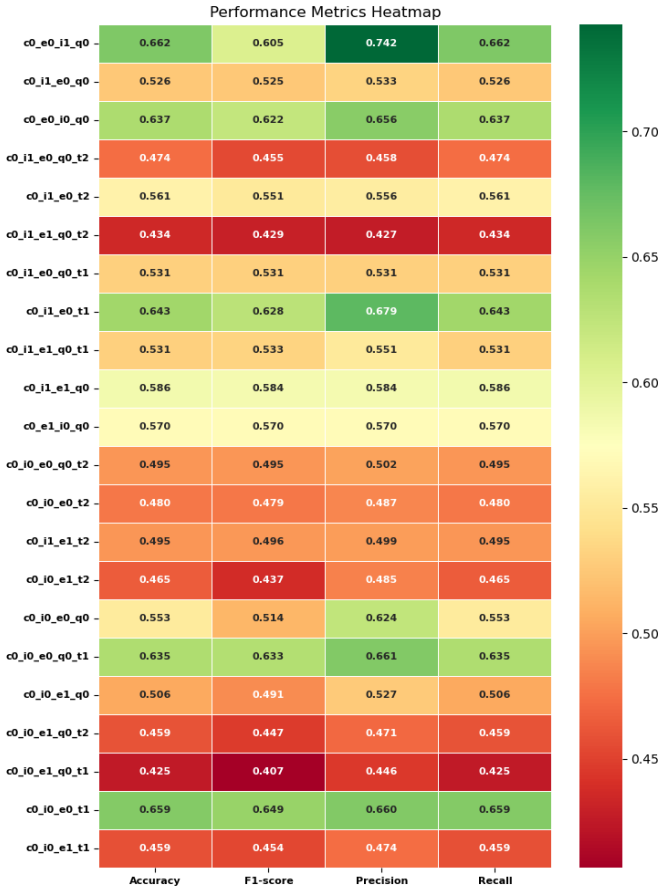

Fig. 3. The heatmap of performance metrics for seed data

it easier for the LLM to classify.

2) **Prompt Format:** The same prompts were used for both datasets to allow comparison. However, due to the differences in text structure between the datasets, different prompt modules may be required for optimal performance.

Based on these observations, we have outlined the future steps for the next stages of our work:

1) **Splitting the Job Descriptions:** Since performance metrics are higher for synthetic datasets, we plan to split the seed dataset into sentences as well.
2) **Specifying the Target of a Sentence:** The synthetic dataset performs better because its sentences are more focused on the candidate or the working environment. Therefore, we will use the same approach for the seed dataset. Therefore, after splitting the job descriptions into sentences, we will determine the target of each sentence. The target is the word that refers to the candidate or the working environment. For example, in the sentence ”Sei pronto/a #MakeAnImpactThatMatters all’interno di Officine Innovazione?” the target word is ”pronto/a”, which includes both genders, making the sentence ”inclusive”. In contrast, in the sentence ”Sei empatico e in grado di cogliere i need dei tuoi colleghi e degli stakeholder interni ed esterni”, the target word ”empatico” is masculine, making the sentence ”non_inclusive”.
   This method also helps to identify unrelated or less focused sentences, such as ”Officine Innovazione è la società del network italiano [company] che promuove la cultura dell’innovazione e fornisce alle imprese clienti servizi di innovation development e management”, which lacks a specific target word and is ”not_representative”. Detecting these types of sentences allows the LLM to focus on more important parts of the text, thereby improving performance.
3) **Labeling the Sentences:** After identifying the target words, we can label the sentences as ’inclusive’, ’non_inclusive’, and ’not_representative’, as explained earlier.

Based on these improvements, the final structure of our dataset, shown in Figure 4, consists of four main columns. All of these steps are carried out by a team of human experts specializing in law, linguistics, and computer science. These changes will allow us to integrate our seed dataset with the synthetic data. This will allow us to fine-tune the LLM on a larger dataset, covering more cases and improving performance metrics. These contributions are currently being developed and tested.

Text	Target	Label	Source
Sei pronto/a a #MakeAnImpactThatMatters all’interno di Officine Innovazione?	pronto/a	inclusive	seed
Sei empatico e in grado di cogliere i need dei tuoi colleghi e degli stakeholder interni ed esterni	empatico	non_inclusive	seed
Officine Innovazione è la società del network italiano Deloitte che promuove la cultura dell’innovazione e fornisce alle imprese clienti servizi di innovation development e management.	no_target	not_representive	seed
Ufficio di nuova aprtura cerca giornalista con esperienza	giornalista	inclusive	synthetic
Stiamo cercando estetista per la nostra azienda	estetista	non_inclusive	synthetic

Fig. 4. The final structure of the integrated dataset of seed and synthetic data

V. Conclusion

In this research, we developed a framework for generating synthetic job description data in the Italian language. We labelled this synthetic data as masculine, feminine or neutral based on the gender of each job title that was substituted in the job description. Using Llama3:instruct, a state-of-the-art LLM, we tested the performance of our framework on 45 job descriptions. Finally, we performed a comparative analysis with a real seed dataset collected by domain experts and proposed improvements to enhance the performance of the LLM, allowing a better integration of synthetic and real data.

In the introduction we posed three main questions. In order to address RQ1 on the generation of synthetic data for job descriptions, we used a comprehensive list of job titles,

extended them with different possible endings based on the target language (in our study, Italian), and substituted them into template sentences extracted from real job descriptions. Having more job titles and template sentences will help to expand the synthetic dataset.

To address RQ2 on the performance of synthetic data, we found that the LLM performs better with synthetic data because it consists of shorter and more focused sentences. However, improving the prompt format could also improve the performance metrics for seed data. To answer RQ3, we found that synthetic data is reliable for fine-tuning, but it requires some unification and structural adjustments to effectively integrate it with real data. A larger dataset combining both seed and synthetic data will undoubtedly improve performance, which we plan to implement soon as suggested in the previous section.

Our research is at an early stage. Due to the limited availability of data in the Italian language, we have to carry out data collection, data annotation and fine-tuning of LLMs in parallel. However, the outcome of this research will be a reliable and validated combination of seed and synthetic datasets and a model trained and fine-tuned on these data. This model will be able to detect bias, stereotyping and discrimination in various domains. The applications of this work are vast, including improving human rights, preventing discrimination in legal settings, and improving overall societal satisfaction.

Limitations of our work at this stage include the limited number of rows in the seed dataset, the testing of only one LLM (Llama3) on our data, and the fact that the prompting was conducted exclusively in Italian. In addition, our prompting approach was more compatible with the structure of the synthetic data. Based on these limitations and the proposed improvements to our framework, we plan to unify the structure of the synthetic data with the seed dataset.

Given the limited amount of real seed data, we will collect more data and annotate them using human experts. We will also test state-of-the-art models such as GPT-4, Mistral, Gemini and others to see which perform best in this type of classification task. We will also test more prompts with more examples in both languages to improve the performance of the prompts. In some cases, we may encounter sentences with two target words, one inclusive and one non-inclusive. For these cases, we can add another label, 'partially_inclusive'. Also, there is a limitation in terms of LLMs trained for the Italian language which is one of our next steps in this path.

ACKNOWLEDGMENT

The work reported in this paper has been partly funded by the European Union - NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D", within the project "MUSA - Multilayered Urban Sustainability Action" (contract n. ECS 00000037).

REFERENCES

[1] S. Böhm, O. Linnyk, J. Kohl, T. Weber, I. Teetz, K. Bandurka, and M. Kersting, "Analysing gender bias in IT job postings: A pre-study based on samples from the German job market," *Proceedings of the 2020 on Computers and People Research Conference*, vol. 2020, pp. 72-80, June 2020.

[2] Statista, "Leading Large Language Model (LLM) Tools Worldwide as of 2024," *Statista*, 2024. [Online]. Available: https://www.statista.com/statistics/1458138/leading-llm-tools/. [Accessed: 20-Jul-2024].

[3] J. Pereira, R. Nogueira, C. Zanchettin, and R. Fidalgo, "An Augmentative and Alternative Communication Synthetic Corpus for Brazilian Portuguese," in *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, Orem, UT, USA, 2023, pp. 202-206, doi: 10.1109/ICALT58122.2023.00066.

[4] J. Ive, "Leveraging the potential of synthetic text for AI in mental healthcare," *Frontiers in Digital Health*, vol. 4, Art. no. 1010202, 2022. doi: 10.3389/fdgth.2022.1010202.

[5] M. Guillaudeux, O. Rousseau, J. Petot, Z. Bennis, C.-A. Dein, T. Goron, N. Vince, S. Limou, M. Karakachoff, M. Wargny, and P.-A. Gourraud, *Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis*, *npj Digital Medicine*, 6(1):37, 2023.

[6] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," *PLOS Digital Health*, vol. 2, no. 1, pp. e0000082, Year 2023.

[7] X. Xu, W. Zhao, and L. Zhang, "Text Data Augmentation for Robust Language Model Training," *Journal of Computational Linguistics*, 2023.

[8] X. Zhang, Y. Li, and X. Wang, "Generative Models for Text: An Overview and Recent Advances," *Artificial Intelligence Review*, 2022.

[9] Y. Li, Y. Zhang, and H. Chen, "Simulation-Based Synthetic Data Generation for Conversational Agents," *Journal of Artificial Intelligence Research*, 2024.

[10] K. Lee, H. Yoon, and J. Choi, "Enhancing Language Model Robustness with Synthetic Data: An Empirical Study," *ACM Transactions on Intelligent Systems and Technology*, 2024.

[11] L. Wang, J. Liu, and M. Zhang, "Supplementing Low-Resource Data with Synthetic Data for Improved Language Models," *Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics*, 2023.

[12] S. Kim, J. Park, and M. Lee, "Using Synthetic Data to Mitigate Bias in Large Language Models," *Proceedings of the 2023 Conference on Fairness, Accountability, and Transparency*, 2023.

[13] T. Nguyen, H. Wang, and S. Patel, "Quality Control in Synthetic Data Generation for NLP Tasks," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[14] S. Bai, Y. Zhang, and C. Liu, "Mitigating Overfitting in LLMs with Synthetic Data: Strategies and Challenges," *Journal of Machine Learning Research*, 2024.

[15] Q. Wang, Y. Chen, and X. Zhang, "Iterative Data Improvement with Reinforcement Learning for Enhanced Language Models," *International Conference on Learning Representations*, 2024.

[16] S. Ravi, R. Singh, and P. Kumar, "Advanced Generative Models for Text: Recent Developments and Applications," *Journal of Computer Science and Technology*, 2024.

[17] H. Chen, J. Xu, and X. Li, "Evaluating the Impact of Synthetic Data on Language Model Performance: A Comprehensive Review," *Natural Language Engineering*, 2024.

[18] Ollama, "Llama 3 Instruct," *Ollama Library*, [Online]. Available: https://ollama.com/library/llama3:instruct. [Accessed: Jun. 15, 2024].