Synthetic Data for Bias Mitigation (BIAS Project)

Type of work: **Term Project 2**

Degree programme:	**Master of Science in Engineering (MSE)**
Specialisation:	**Data Science**
Student:	**Shilpi Garg (gargs1)**
Project:	**PA2**
Project Advisor:	**Dr. Kurpicz-Briki Mascha**
Co-Advisor:	**Dr. Alexandre Puttick**
Date:	**30 July 2025**

Bern University
of Applied Sciences

Contents:

1. Abstract

The Horizon Europe project BIAS explores how societal biases manifest in AI systems and large language models, particularly within the context of the labor market. As part of this initiative, this subproject focuses on bias detection and mitigation in the generation and evaluation of CVs and cover letters. The primary objective is to develop a modular synthetic data generation framework that enables the systematic creation of job application documents—including job ads, CVs, and cover letters—while carefully controlling for demographic variables, linguistic patterns, and skill profiles. This synthetic dataset is intended to serve as a benchmark for evaluating and improving fairness in downstream natural language processing (NLP) applications.

In this context, synthetic data for bias mitigation refers to the deliberate generation of artificial application materials that allow for fine-grained manipulation of attributes such as gender, ethnicity, age, and socio-economic background. This is a part of an effort to ensure that models trained or tested using such data do not perpetuate or exacerbate existing societal biases. The framework supports the injection of sensitive terms (e.g., "female," "disabled") and proxy terms (e.g., "parent," "leader") into the documents to test the model's predictive behavior in response to these variables. These terms are often correlated with demographic markers and may trigger unintended discriminatory outcomes, making them critical for bias stress-testing and mitigation strategies. In this study, the focus is on detecting gender bias in both real and synthetic application documents using text classification and interpretability methods. A LIME-based explainer is employed to identify which words contribute to gender predictions, and systematically apply masking strategies—such as removing names, pronouns, social roles, and workplace titles—to assess the model's reliance on explicit and proxy gender cues. This approach enables a controlled evaluation of bias persistence under varying levels of linguistic obfuscation.

Keywords- NLP, Synthetic Data, data augmentation, bias mitigation, sensitive attributes, proxy words, machine learning, LIME

2. Introduction

The rise of powerful language models like ChatGPT has brought widespread attention to bias and stereotyping in NLP systems. Bias can emerge at multiple stages of the ML pipeline—including data, annotation, and model design (Hovy & Prabhumoye, 2021).

The integration of artificial intelligence (AI) into recruitment processes has transformed how organizations evaluate job applicants. While AI offers scalable tools for resume screening and candidate evaluation, growing evidence suggests that these systems risk perpetuating or amplifying gender biases embedded in training data and model architectures (Fabris et al., 2025). This is particularly critical during the early stages of hiring, where automated models interpret CVs and cover letters—documents that inherently reflect individual self-presentation strategies shaped by societal norms (Marti Marcet, 2023). AI systems trained on real-world data are vulnerable to perpetuating historical societal biases, including gender biases embedded in language and occupational roles. This study is part of the broader Horizon Europe BIAS project, which aims to examine and mitigate such biases in AI systems deployed within labor market contexts.

By leveraging controlled synthetic data and fairness-driven evaluation protocols, this work contributes to the development of transparent, accountable, and equitable AI tools in recruitment pipelines. This approach is aligned with emerging principles (fairness, transparency, accountability and scalability) of **Human-Centered AI** (Bartl et al., 2025) and contributes to a growing body of work advocating for **accountable AI in employment technologies** (Chaturvedi & Chaturvedi, 2025), (Serna et al., 2023). Moreover, it provides a scalable blueprint for testing bias sensitivity in downstream NLP applications, especially in high-stakes domains.

This study investigates gender bias in text classification using the BiasBios dataset and synthetic CVs and cover letters. The BiasBios dataset, containing over 390,000 biographies labeled by gender and occupation, reveals clear gender imbalances across professions. A LIME-based explainer was employed to detect and interpret gender bias by analyzing which specific words or phrases most influenced the model's gender predictions. LIME (Local Interpretable Model-Agnostic Explanations) allowed for a granular, token-level understanding of how the classifier made decisions, highlighting both direct gender indicators (e.g., *he*, *Ms.*, *Jessica*) and more subtle proxy features (e.g., *CEO*, *sincerely*). To systematically evaluate the model's reliance on these features, multiple masking strategies were applied—removing or obfuscating names, pronouns, occupational terms, social identifiers, and personality-related language. This enabled an in-depth assessment of how deeply gender cues are embedded in the text and how resilient model bias is to surface-level anonymization.

Results showed that even with masked inputs, classifiers still relied on subtle linguistic proxies, highlighting the persistence of gendered patterns in NLP systems.

The project proposes a modular synthetic data generation framework for constructing balanced and bias-aware training and evaluation datasets. Specifically, our framework focuses on the systematic generation of job application materials. We generate artificial CVs and cover letters that allow controlled variation of:

- Demographic attributes (e.g., gender markers like *he, she, her, his,* names like *Jennifer, piyush*)
- Linguistic features (eg. *leader*, *mentor*, *motivated*, *collaborative*, *organized*), refer to stylistic, semantic, or lexical patterns
- Sensitive/proxy terms (e.g. *director*, *assistant*, *chief*, *support*, *chairwoman*), Proxy words are terms that can be used by the model as indirect signals (often unintended) to predict sensitive attributes (like gender, race, or biases).

The objectives of this study are:

- RQ1: How much gender information can be inferred from the data (bios/CV cover letters)? What can be done to mask that information?
- RQ2: How does the presence of gender information affect the generation of synthetic data? How susceptible is the generated data to gender bias?
- RQ3: Evaluate the behaviour of existing NLP models against this synthetic dataset using LIME explainer to assess and mitigate their bias responses.
- RQ4: Is there potential to use the explored methods to help mitigate unfair discrimination in recruitment?

2.1. Related work:

A growing body of research has focused on detecting and mitigating bias in word embeddings and language models. Early methods, such as Bolukbasi et al. (2016), attempted to remove gender signals from embeddings, while Caliskan et al. (2017) adapted the Implicit Association Test (Greenwald et al., 1998) used in psychology to measure implicit cognition attitudes, to show that machines can learn word associations from written texts and that these associations mirror those learned by humans. Numerous reviews summarize these advances (Sun et al., 2019; Meade et al., 2022; Delobelle et al., 2022), yet studies show limited correlation between embedding-level debiasing and downstream fairness (Goldfarb-Tarrant et al., 2020).

Although many studies do not explicitly define “bias,” their technical methods suggest implicit definitions. Most rely on techniques from Caliskan et al. (2017), Nangia et al. (2020), or Bolukbasi et al. (2016). The first two aim to measure whether harmful stereotypes are encoded in word embeddings by comparing associations (e.g., *he* vs. *she* with *business*). In contrast, Bolukbasi’s DirectBias metric quantifies overall gender skew but does not assess stereotype direction—defining bias as any deviation from gender neutrality,

not necessarily linked to social harm or discrimination. Despite significant progress in bias detection, most research on word embeddings and language models remains heavily focused on English. This lack of linguistic diversity has drawn criticism (Joshi et al., 2020), as bias is often shaped by cultural and language-specific factors (Fiske, 2017; Kurpicz-Briki & Leoni, 2021).

One of the foundational challenges in mitigating bias in AI hiring systems is the representational imbalance in training datasets. BiasBios dataset is scraped from the internet and likely reflects gender disparities that are also on the internet-scraped pre-training data of LLMs. Work of Sweeney et al. (2020) critiques models trained on BiasBios and introduces a framework for transparent reporting of performance gaps across genders. It highlights persistent disparities in false positive and false negative rates, especially for female bios in male-dominated fields. While the study of Wang et al. (2021) focuses more broadly on representation bias, it references BiasBios as a textual analog and stresses that balancing gender proportions in datasets doesn't eliminate bias. Models can still learn and reinforce gender-label associations despite data parity.

As visualized in Figure 1 and 2, a disproportionate representation exists between male and female entries, with 53.9% identified as male (213,543) and 46.1% as female (182,646). While this may seem modest, such imbalances become more problematic when they intersect with gender-stereotyped occupational labels, where roles like "nurse" or "yoga teacher" show over 80% female representation, while positions such as "software engineer" or "rapper" are male-dominated (Mansouri et al., 2024; Mihaljević et al., 2022)

```
Gender Distribution:
gender
male      213543
female    182646
Name: count, dtype: int64

Gender Distribution (Percentage):
gender
male      53.899275
female    46.100725
Name: count, dtype: float64
```

Figure 1: Gender distribution on whole dataset.

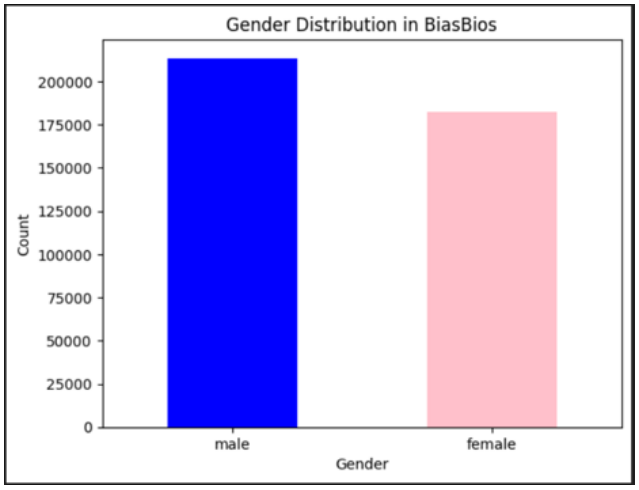


Figure 2: Gender distribution in BiasBios dataset

2.2. Synthetic Data for Bias Mitigation:

To counteract the mentioned biases in training data, synthetic data generation has emerged as a promising solution (Peña et al., 2023),(Frazzetto, 2025). Synthetic data is constructed to test and evaluate model bias by carefully varying demographic features (like gender or race) while keeping the meaning and job context the same. This connects directly to the use of proxy terms—words or phrases (e.g., "nanny," "CEO," "chairwoman") that are not explicitly demographic but often strongly correlated with protected attributes (like gender or ethnicity). These proxy terms aren't biased on their own, but because of their frequent association with certain groups, they can trigger biased responses from models. By including or varying such terms in controlled ways, researchers can simulate counterfactuals of the same bio or input with altered demographic indicators and measure how the model's behavior changes. This method helps detect and mitigate bias, as shown in recent studies on counterfactual data generation for fairness (Kumar et al., 2023).

3. Materials and Methods

3.1. BiasBios Dataset:

BiasBios is a large-scale, human-labeled dataset developed to facilitate the study and mitigation of gender bias in machine learning systems, particularly in automated occupation classification tasks. Introduced by De-Arteaga et al., (2019). in their seminal work "Mitigating Unwanted Biases with Adversarial Learning" (AIES 2019), the dataset comprises approximately 400,000 short biographies scraped from English-language publicly available professional platforms. It covers 28 diverse occupations ("accountant", "architect", "attorney", "chiropractor", "comedian", "composer", "dentist", "dietitian", "dj", "filmmaker", "interior_designer", "journalist", "model", "nurse", "painter", "paralegal", "pastor", "personal_trainer", "photographer", "physician", "poet", "professor", "psychologist", "rapper", "software_engineer", "surgeon", "teacher", "yoga_teacher"), and includes binary gender annotations inferred from first names or pronouns.

To analyze the distribution of data across different professions, we computed the number of biographies available for each profession in our dataset. The plot (Figure 3) provides an immediate sense of which professions are most or least represented in the dataset, which is useful for data balance assessment or identifying potential sampling biases.

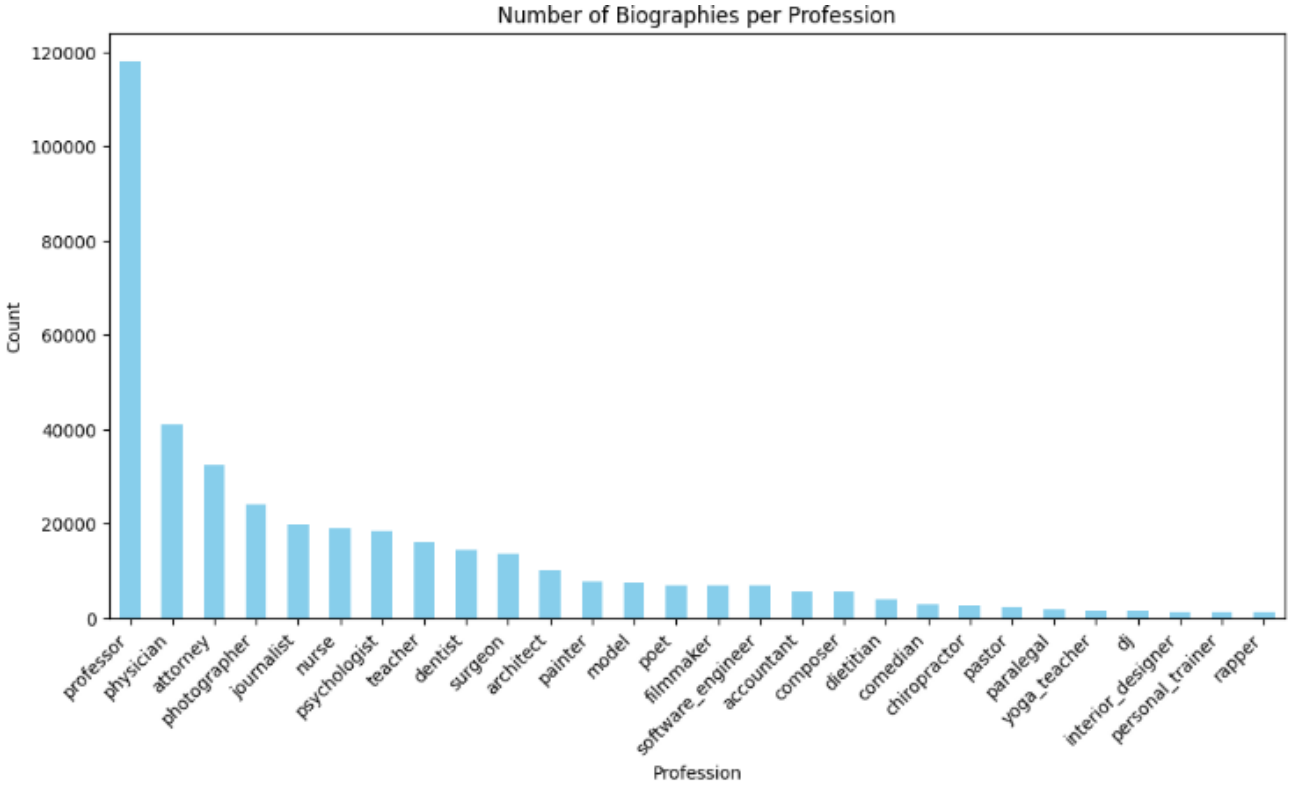


Figure 3: Number of Biographies per Profession

To gain an overview of how biographies are distributed across various professions in our dataset, we visualized the proportion of entries for each profession using a pie chart (Figure 4). This visualization is particularly useful for presenting to non-technical stakeholders who need a high-level understanding of the data composition. The variable profession_counts_sorted is a list-like object where each profession is listed along with the number of biographies it has. The data is sorted to make the pie chart easier to read. The autopct="%1.1f%%" setting adds percentage labels to each slice, showing one decimal place. colormap="Set2" applies a set of soft, distinct colors to make the slices easy to tell apart.

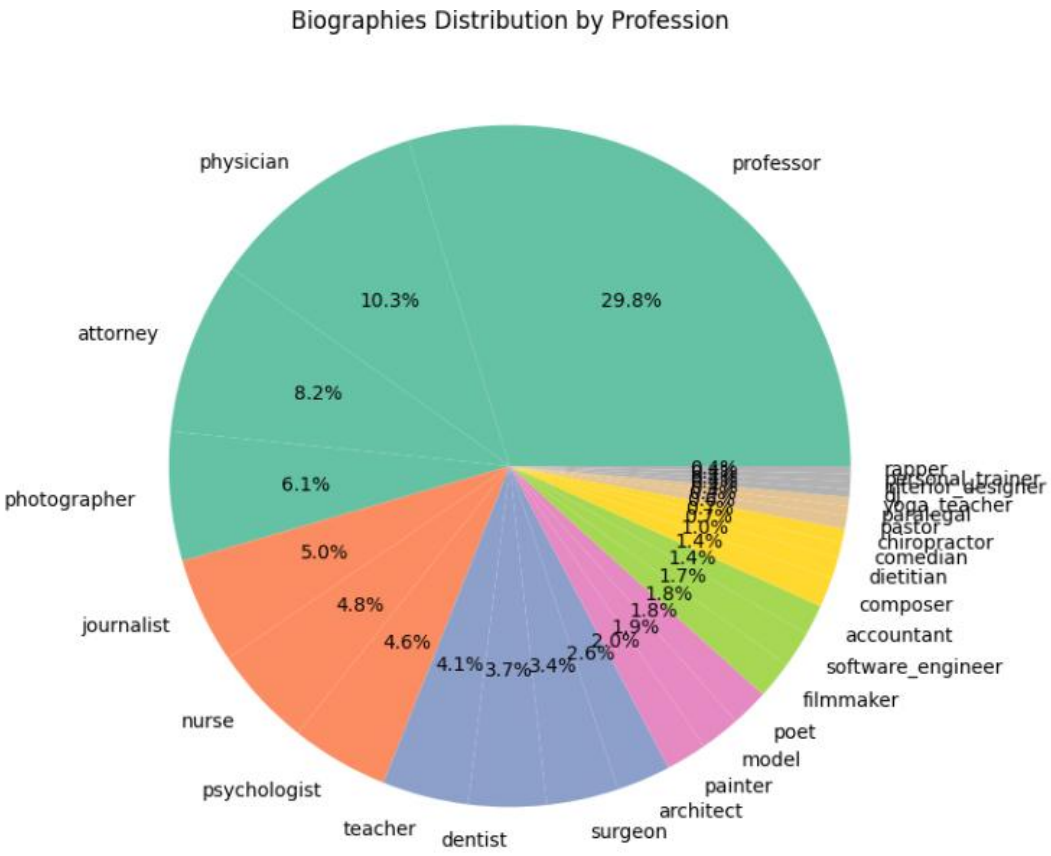


Figure 4: Biographies Distribution by Profession

To analyze and visualize gender distribution across different professions, we first grouped BiasBios dataset by the "profession" column and calculated the relative frequency of each gender within those professions. To visualize this data, Seaborn and pyplot from matplotlib is used for creating informative plot (Figure 5).

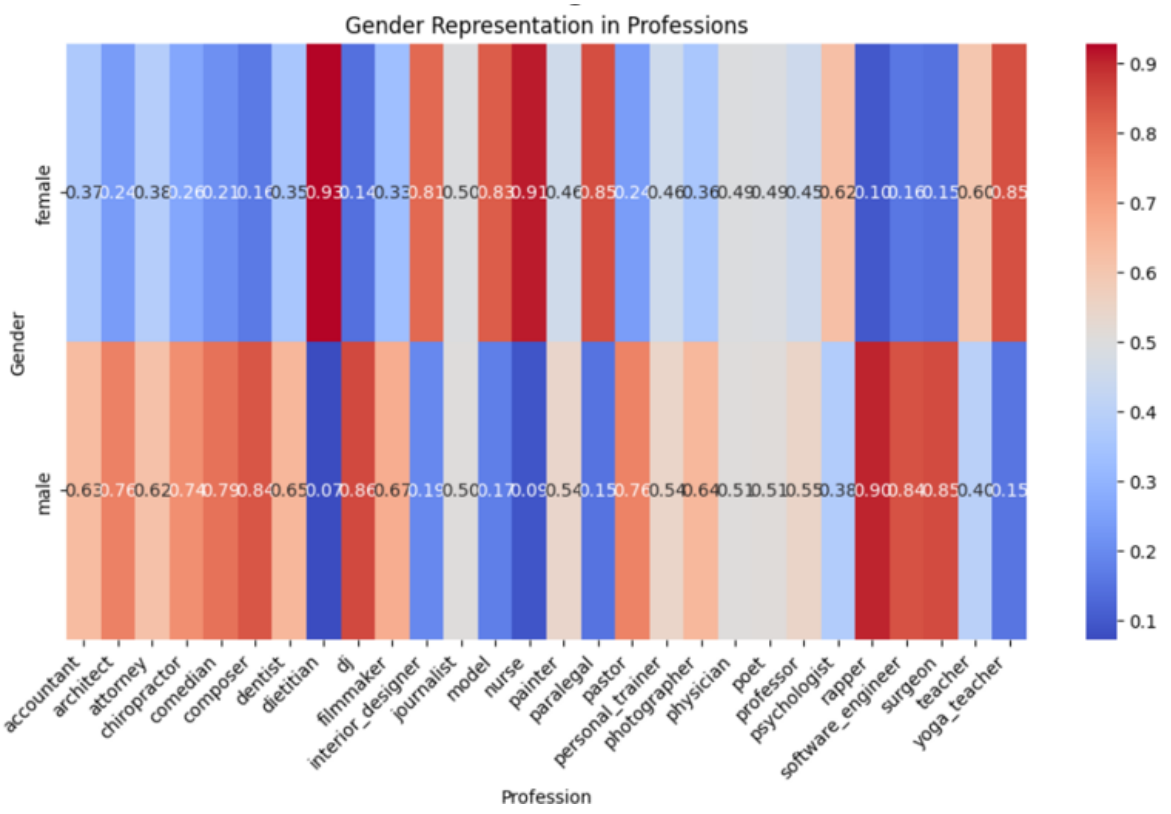


Figure 5: Gender Representations in various Profession

3.2. Models:

We use the large language models LLaMA-2-7B and LLaMA-3.2-3B to automatically generate professional documents: CVs, job advertisements, and cover letters. These models are prompted with structured data and sample bios from the BiasBios Dataset to produce tailored, coherent, and contextually appropriate content. By leveraging the models' ability to understand and generate human-like language, we can create customized outputs that reflect specific job roles, skills, and candidate profiles while also enabling scalable document generation for hiring or application pipelines.

3.2.1. Model LLaMA-2-7B-Chat (Meta AI, 2023):

The LLaMA-2-7B-Chat model is a 7-billion parameter instruction-tuned large language model released by Meta AI as part of the LLaMA-2 family. It is specifically fine-tuned for dialogue-based interactions and chat-style completions. Built on a transformer architecture, LLaMA-2 was trained on 2 trillion tokens of publicly available data and fine-tuned using Reinforcement Learning from Human Feedback (RLHF). Its context window is 4096 tokens. It performs competitively with other open-source models like Falcon-7B and Mistral-7B on instruction-following tasks and is particularly suited for structured generation tasks such as CV/cover letter synthesis and bias analysis (Touvron et al., 2023; Shumer et al., 2023)

3.2.2. Model LLaMA-3.2-3B-Instruct (Meta AI, 2024):

The LLaMA-3.2-3B-Instruct is a lightweight, instruction-tuned language model from the LLaMA 3 family developed by Meta AI. As a 3.2 billion parameter variant, it offers a strong trade-off between computational efficiency and linguistic capability, making it particularly well-suited for low-latency inference, real-time evaluation, and structured generation tasks under resource-constrained environments. Unlike its larger counterparts (e.g., LLaMA-3 7B and 70B), the 3.2B model emphasizes efficiency over scale, allowing for high-throughput generation across applications such as CV/cover letter synthesis, prompt probing, and fairness stress-testing. Despite its smaller size, it achieves competitive instruction-following performance through fine-tuning on a combination of supervised instruction datasets and synthetic reasoning prompts, including chain-of-thought, multi-turn dialogues, and system-level instructions. (Meta AI.,2024)

3.3. Hardware specifications:

The models were hosted and run on a local workstation/server with the following GPU and CPU specifications: CPU: AMD Ryzen (32 Cores, 64 Threads, 3.5 GHz base clock). High-core-count CPU is used for efficient token pre-processing and parallel sampling. GPU: NVIDIA A100 40GB (PCIe version). CUDA Capability: 8.0. FP16/TF32 optimization for faster matrix multiplications during transformer forward passes.

3.4. Experimental Setup:

Data:

The extracted dataset used for the experiments consists of 1,000 'accountant' bios, each represented as a row containing two columns:

- text: The biographical text, describing the individual's professional background.
- labels: The corresponding label (gender) associated with each bio.

All text entries were lowercased, which simplifies normalization and text comparison. The bios contain common English stop words, many of which are not semantically useful for downstream NLP tasks.

(Git: P2_BiasMitigation_SyntheticData/src/df_accountant_lower.ipynb).

Stop word removal was implemented by splitting each bio into tokens (words), filtering out those matching the stop word list, and rejoining the filtered tokens into a cleaned string.
A new column named clean_text was added to the dataset. Each row now includes the original lowercase bio (text), the filtered version (clean_text), and the associated labels.
This preprocessing step ensures the model is trained on more informative content, reducing noise and improving generalization in learning tasks.

Classification:
Classification is carried out using a transformer-based architecture for text classification using the BERT (Bidirectional Encoder Representations from Transformers) model (Jacob Devlin et al., 2018). Specifically, we utilized the bert-base-uncased variant available through the Hugging Face Transformers library (Wolf et al., 2020), implemented via the AutoModelForSequenceClassification class. The input data is pre-processed using the associated tokenizer with a fixed max_length (by default 512) padding strategy to ensure uniform sequence length across all samples. The dataset was partitioned into training and test sets using a standard 80/20 train–test split. For training the model, we leveraged the Trainer API along with TrainingArguments from the Transformers framework, which allows streamlined training and evaluation workflows.

Evaluation:
After training the classifier to predict gender from the input text (bios, cv/cover letters), explanations were obtained using LIME (discussed in the section 3.5, Gender Classifiers). Evaluation of model performance was conducted using a suite of classification metrics from the scikit-learn package, including accuracy, precision, recall, and F1-score, to comprehensively assess the predictive performance. Additionally, a confusion matrix was computed to visualize the distribution of classification results, which was rendered using a heatmap generated with the seaborn visualization library.

3.5. Gender classifiers:

To ensure interpretability of the model predictions, we employed LIME (**Local Interpretable Model-agnostic Explanations**) (Ribeiro et al., 2016), a post-hoc explanation technique designed to provide insight into individual predictions made by black-box models. It explains the model's behavior only around a single instance being analyzed, not globally. Here, the used single Test sample text is: (Index, 115)

"Kenneth H. Fowler has been issued a Tennessee license number 10377. All CPAs, including Kenneth H. Fowler, have at the minimum an undergraduate degree in accounting, passed a rigorous national exam and adhere to mandated continuing education requirements of their states in which they are licensed. CPAs can work in private industry, education or government but most people think of CPAs during tax season as the experts in tax preparation. Their overall training in business and knowledge in principles of general law and taxation provide CPAs with the skills to help individuals with both personal and business financial decisions."

The explainer used the LimeTextExplainer from the LIME library to interpret the model's prediction for test sample 115, labeled as 'Female' or 'Male'. It generated perturbed versions of the input text and observed the black-box model's predictions on these variants. A local surrogate model was then trained to approximate the prediction behavior around the instance. The coefficients of this surrogate model indicate the importance of specific words in influencing the model's prediction.

When explaining the model's prediction for an instance, LIME highlights the words that contributed most to the model's decision. By analyzing multiple instances where the model predicts a particular label (e.g., "male" or "female"), one can observe if certain words consistently have high importance. The classifier's performance is an indicator of how much sensitive gender information is present in the data. The code

processes 200 test samples from a dataset to identify and aggregate important words linked to male and female labels. For each sample, it uses LIME explainer tool to find the top 10 features (words) influencing predictions. It sums the feature scores separately for male and female samples in two dictionaries. Progress is printed every 10 samples, and GPU memory is cleared regularly. Finally, the code sorts and prints the top 10 words with the highest cumulative scores for both male and female categories, highlighting words most strongly associated with each gender. Proxy words may often appear as highly weighted features in the local explanations, revealing that the model may be relying on these words as shortcuts or proxies for sensitive attributes. This identification helps in auditing and mitigating bias, for example by removing or masking proxy words or adjusting training data.

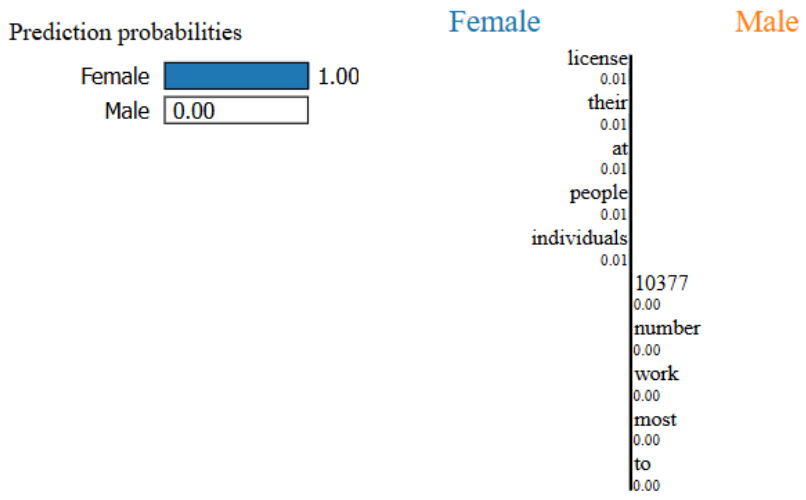
To identify test samples with low prediction confidence, we use the absolute difference between the top two predicted class probabilities.

diffs = np.abs(pred_probs[:, 0] - pred_probs[:, 1])

3.5.1. Gender Classifier without masking:

In our setup for gender classification using unprocessed text without applying any masking, lowercasing, or stop word removal, the model was trained to predict gender labels, where "female" is represented as 1 and "male" as 0 (Figure 6)

(Git: P2_BiasMitigation_SyntheticData/src/model_BERT/GenderClassification.ipynb).



Text with highlighted words

Kenneth H. Fowler has been issued a Tennessee license number 10377. All CPAs, including Kenneth H. Fowler, have at the minimum an undergraduate degree in accounting, passed a rigorous national exam and adhere to mandated continuing education requirements of their states in which they are licensed. CPAs can work in private industry, education or government but most people think of CPAs during tax season as the experts in tax preparation. Their overall training in business and knowledge in principles of general law and taxation provide CPAs with the skills to help individuals with both personal and business financial decisions.

Figure 6: Gender classifier using LIME (without masking) for 'accountant' Bio.

3.5.2. Gender Classifier with lowercase text and without masking:

The dataset text was converted to lowercase before being analyzed, which ensures consistency and avoids discrepancies due to capitalization. The model was trained to predict gender labels, where "female" is represented as 1 and "male" as 0 (Figure 7)

(Git: P2_BiasMitigation_SyntheticData/src/model_BERT/Classifier_with_lowercase_text.ipynb).

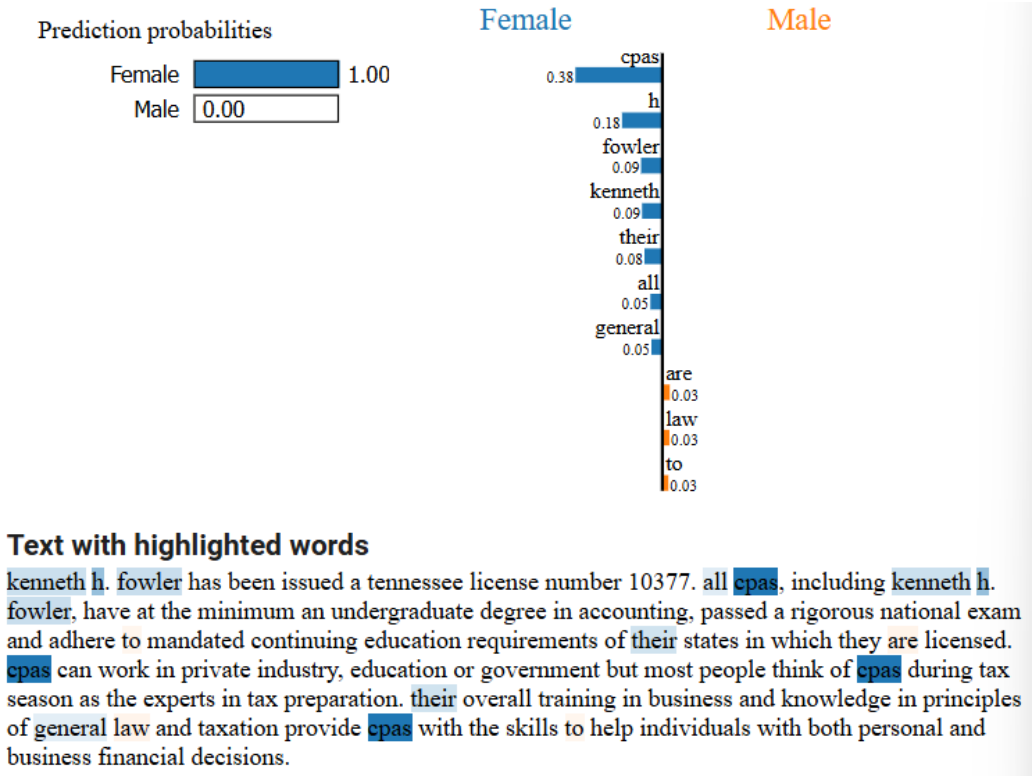


Figure 7: Gender classifier using lowercase text and without masking

3.5.3. Gender Classifier with masking gender pronouns and lowercase text:

We introduce a gendered pronoun masking list, defined as GENDERED_PRONOUNS_LIST. This technique replaces explicitly gendered pronouns with neutral equivalents (e.g., "he" → "they", "her" → "their", "himself" → "themselves"). This is implemented as part of a de-biasing pipeline to reduce the model's dependence on overt pronouns. (Figure 8)

A confusion matrix (Figure 9) is generated to show how well the model performs on test data. It compares true labels with predicted ones, then uses a heatmap to visualize correct and incorrect predictions. The x-axis shows predicted labels, the y-axis shows actual labels, and numbers show prediction counts.

The map is applied to all input texts, replacing binary pronouns and titles with gender-neutral forms to mask explicit gender signals. The LIME explainer interpretation reveals distinct word associations driving gendered predictions. (Figure 10) When prediction probabilities are skewed with 'Female' = 1 and 'Male' = 0

(Git: P2_BiasMitigation_SyntheticData/src/model_BERT/masked_pronouns_genderClassification.ipynb).

```
GENDERED_PRONOUNS_LIST = {
    "male": [
        ("he", "they"),
        ("him", "them"),
        ("his", "their"),
        ("himself", "themselves"),
        ("mr", "they"),
        ("mr.", "they"),
        ("men", "they"),
        ("man", "they")
    ],
    "female": [
        ("she", "they"),
        ("her", "them"),
        ("hers", "theirs"),
        ("herself", "themselves"),
        ("ms", "they"),
        ("ms.", "they"),
        ("mrs.", "they"),
        ("mrs", "they"),
        ("herself", "themselves"),
        ("women", "they"),
        ("woman", "they")
```

Figure 8: Gendered Pronouns List

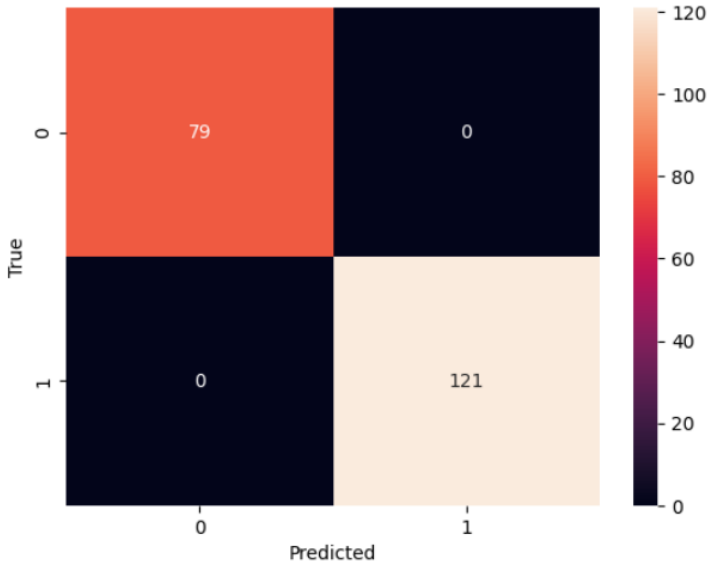


Figure 9: Confusion Matrix when masked pronouns

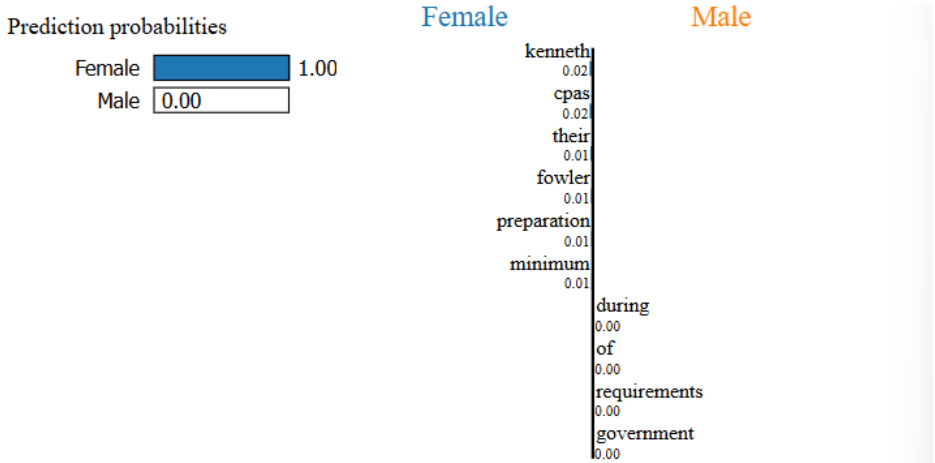


Figure 10: Gender classifier with masking gender Pronouns

3.5.4. Gender Classifier with masking Names and pronouns:

We introduce a de-biasing pipeline that combines name masking and gendered pronoun masking. The function detects and masks names in a text using NLP. It identifies named entities labeled as "PERSON" and replaces each name with “[NAME]” to remove identity-specific cues. Reversing the entity loop ensures character positions stay accurate while modifying the text. The result is a version of the input with names anonymized. A predefined GENDERED_PRONOUNS_LIST (Figure 8) is used to substitute explicitly gendered pronouns with neutral equivalents (e.g., "he" → "they", "her" → "their", "himself" → "themselves". This dual-masking approach aims to reduce the model’s reliance on overt gender indicators and mitigate bias in downstream predictions. To enhance the quality of text for modeling, a manual list of stop words was used for filtering. These words are typically non-informative for tasks like classification or text generation. The following common function stop words in English were removed from dataset (commonly extracted from ‘spacy’ open-source Python library for Natural Language Processing (NLP)):

["the", "a", "an", "and", "has", "of", "to", "in", "as", "is", "was", "were", "are", "for"]

These words appear frequently in most texts but carry little semantic meaning on their own. Removing stop words reduces noise in the data and allows models or explanations (like LIME) to focus on content words that carry more predictive or explanatory power (e.g., verbs, nouns, skills).

Figure 11 presents a confusion matrix reflecting the model’s classification performance after implementing a different masking strategy likely one that still includes gendered information or proxy terms.

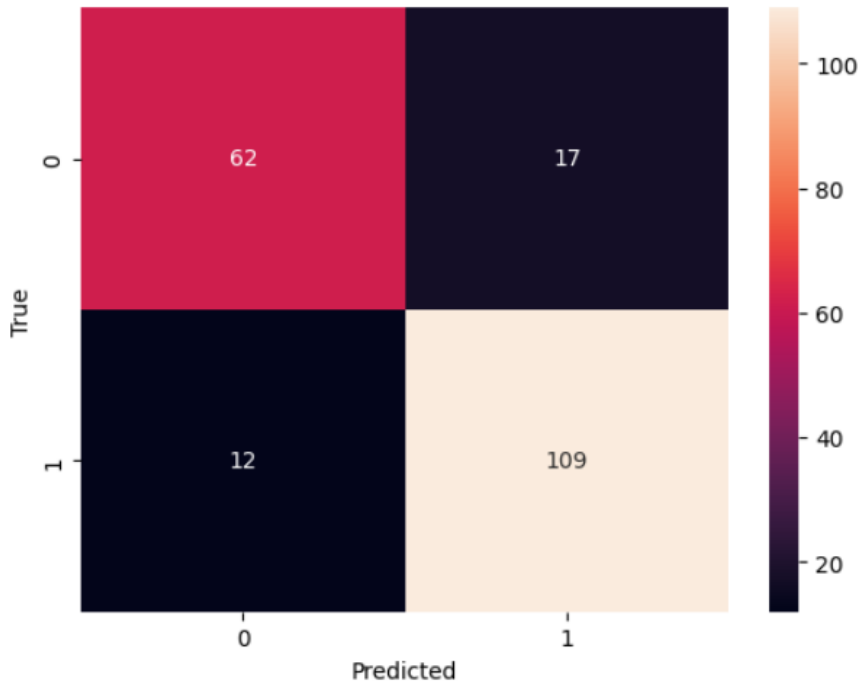
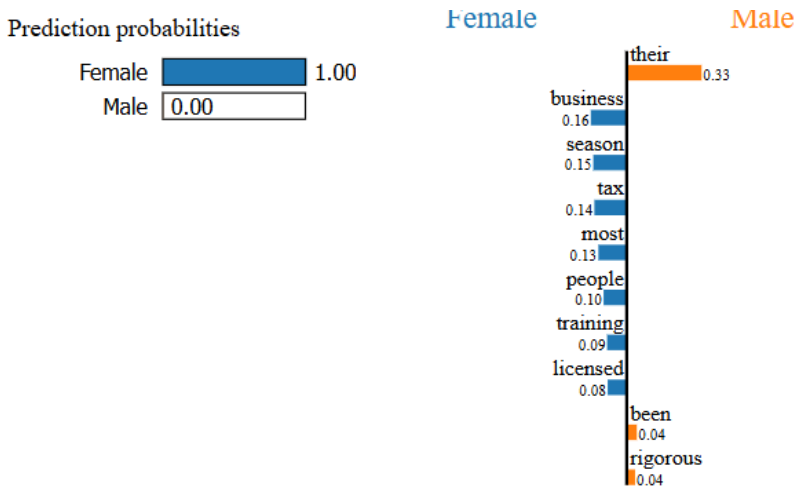


Figure 11: Confusion Matrix when masked Names and pronouns

The LIME-based explanation highlights specific word associations that influence the model’s gender prediction (Figure 12). When the predicted probabilities are interpreted with 'Female' = 1 and 'Male' = 0, the explanation reveals how certain words contribute more strongly to a female or male classification.

(Git: P2_BiasMitigation_SyntheticData/src/model_BERT/masked_Names_GenderClassification.ipynb).



Text with highlighted words

[NAME] been issued tennessee license number 10377 . all cpas , including [NAME] , have at minimum undergraduate degree accounting , passed rigorous national exam adhere mandated continuing education requirements their states which they licensed . [NAME] can work private industry , education or government but most people think cpas during tax season experts tax preparation . their overall training business knowledge principles general law taxation provide cpas with skills help individuals with both personal business financial decisions .

Figure 12: Gender classifier with masking Names and Pronouns

3.5.5. Gender Classifier with masking Social and Workplace roles (Proxy terms):

Figure 13 shows a curated list of proxy gendered terms categorized by perceived gender associations used for Gender Classification task with masking Social and Workplace roles (accountant bio). The following experiment uses masking of names and pronouns (Figure 8) as well. A curated list of proxy gendered terms spanning social roles and workplace identifiers was extracted using OpenAI, a large language model (LLM), due to its broad contextual understanding and linguistic generalization capabilities. The words such as *husband, executive* and *investor* for males, and *wife, mother* and *support* for females reflect social, familial, and workplace roles that often carry gender-coded connotations. These proxy terms were systematically replaced with neutral alternatives like " __" to minimize gender bias during model training. To enhance the quality of text for modeling, a manual list of stop words extracted from 'spacy', is used for filtering. These words are typically non-informative for tasks like classification or text generation. The following stop words were removed from dataset:

["the", "a", "an", "and", "their", "they", "them", "themselves", "theirs", "has", "of", "to", "in", "as", "is", "was", "were", "are", "for", "on"]

Figure 14 displays a confusion matrix highlighting the model's performance after applying both name and pronoun masking as part of the gender de-biasing process.

The LIME-based explanation highlights specific word associations that influence the model's gender prediction (Figure 15). When the predicted probabilities are interpreted with 'Female' = 0.99 and 'Male' = 0.01, the explanation reveals how certain words contribute more strongly to a female or male classification when masked with social and workplace roles.

(Git: P2_BiasMitigation_SyntheticData/src/model_BERT/masked_obviousProxy_GenderClassification.ipynb).

```
ProxyWords_LIST = {
    "male": [
        ("husband", "__"),
        ("father", "__"),
        ("boy", "__"),
        ("man", "__"),
        ("executive", "__"),
        ("director", "__"),
        ("analyst", "__"),
        ("leadership", "__"),
        ("men", "__"),
        ("dad", "__"),
        ("son", "__"),
        ("mentor", "__"),
        ("manager", "__"),
        ("leader", "__"),
        ("investor", "__")

    ],
    "female": [
        ("wife", "__"),
        ("mother", "__"),
        ("girl", "__"),
        ("lady", "__"),
        ("women", "__"),
        ("clerical", "__"),
        ("support", "__"),
        ("mom", "__"),
        ("daughter", "__"),
        ("sister", "__"),
        ("assistant", "__"),
        ("coordinator", "__")
```

Figure 13: Proxy Words List

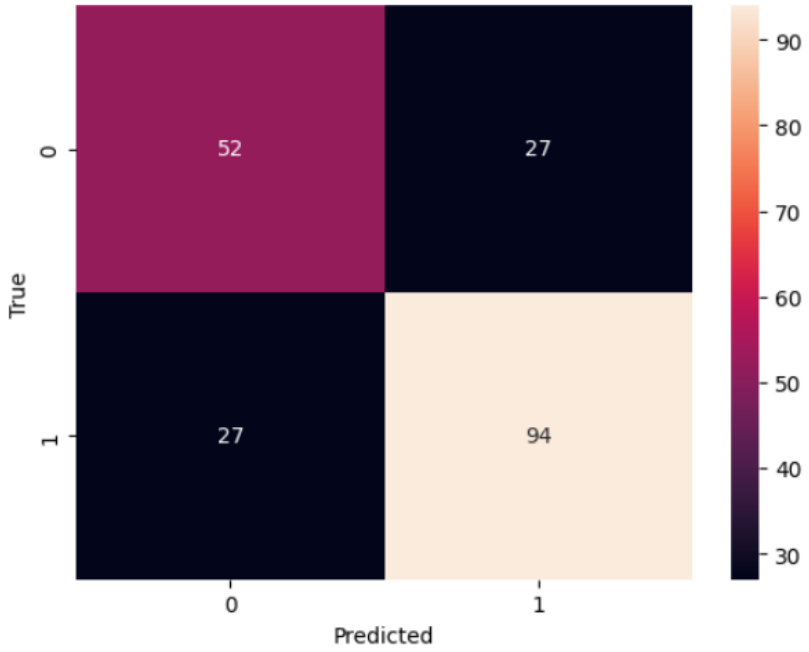


Figure 14: Confusion Matrix when masked Social and workplace roles

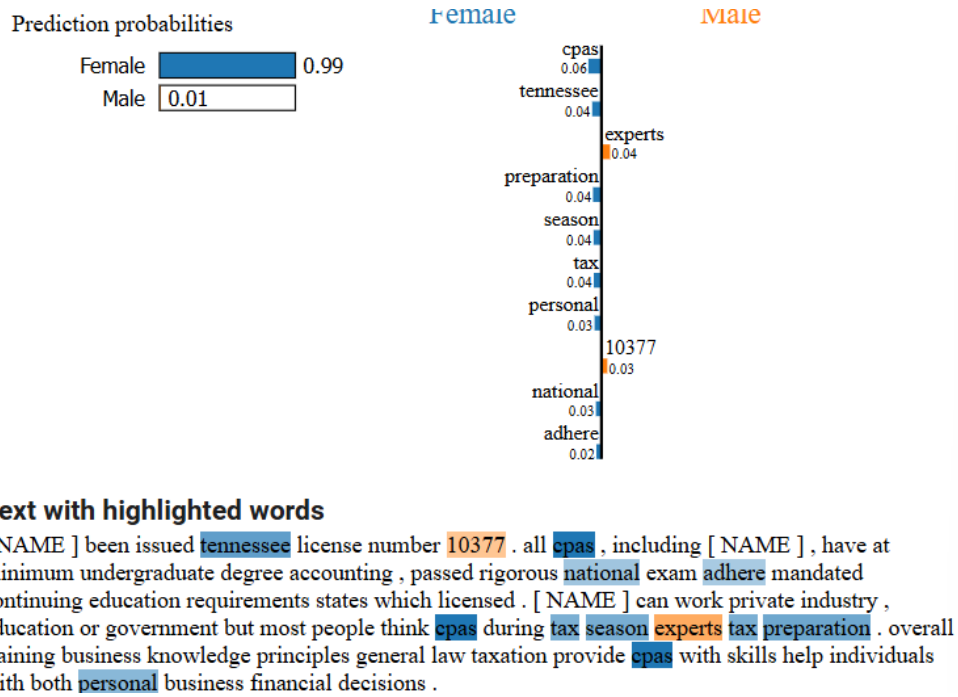


Figure 15: Gender classifier with masking Social and workplace roles

3.5.6. Gender Classifier with masking Personality traits:

Classifer with masking Personality traits includes dataset received after masking names, pronouns (Figure 8), social and workplace roles (Figure 13), and using stop words (extracted from 'spacy' library). The Personality_LIST is a dictionary that groups gender-related personality traits under "male" and "female". Traits (like *ambitious, competitive, confident*) are paired with the neutral symbol "__" under 'Male' category and Traits (like *empathetic, supportive, responsible*) are paired with the neutral symbol "__" under 'Female' category (Figure 16). This setup is used to test how well a gender prediction model performs when these traits are masked. The confusion matrix shows how well the gender model works when personality traits are masked. The model correctly predicts male (class 1) very well, it gets 94 right and 27 wrong. But for female (class 0), it's less accurate. It gets 54 right but makes 25 mistakes, wrongly labeling them as male (Figure 17). This shows the model still finds, male-linked traits more detectable, even after masking, while it struggles more with female-linked traits when the gender signal is hidden.

The LIME-based explanation highlights specific word associations that influence the model's gender prediction (Figure 18). With predicted probabilities of 'Female' = 0.99 and 'Male' = 0.01, the explanation shows how certain terms still drive gendered predictions, even when personal traits, social roles, job-related identifiers, names, and pronouns have been masked.

(Git: P2_BiasMitigation_SyntheticData/src/model_BERT/masked_personality_Classifier.ipynb).

```
Personality_LIST = {
    "male": [
        ("ambitious", "__"),
        ("competitive", "__"),
        ("assertive", "__"),
        ("confident", "__"),
        ("strategic", "__"),
        ("independent", "__"),
        ("rational", "__"),
        ("analytical", "__")
    ],

    "female": [
        ("empathetic", "__"),
        ("supportive", "__"),
        ("nurtouring", "__"),
        ("organized", "__"),
        ("friendly", "__"),
        ("patient", "__"),
        ("sensitive", "__"),
        ("cooperative", "__"),
        ("responsible", "__"),
```

Figure 16: Personality Traits List

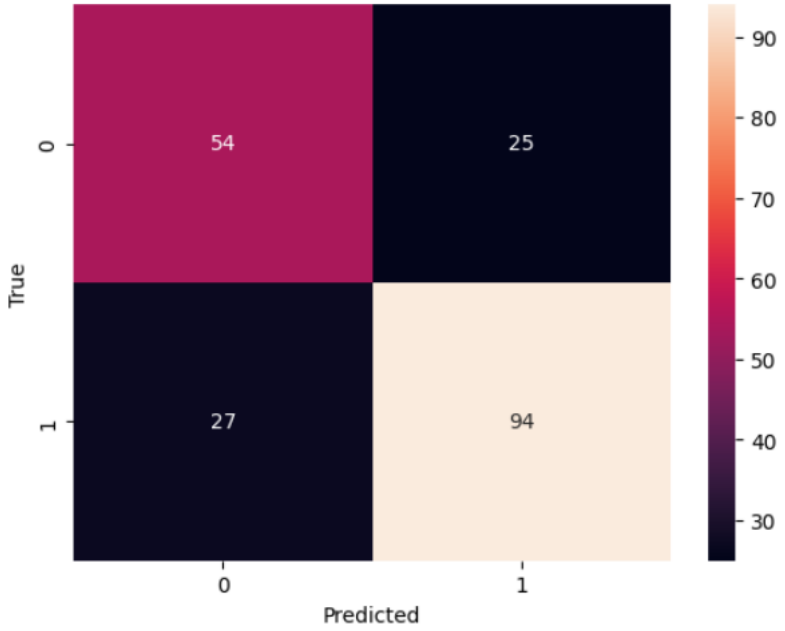
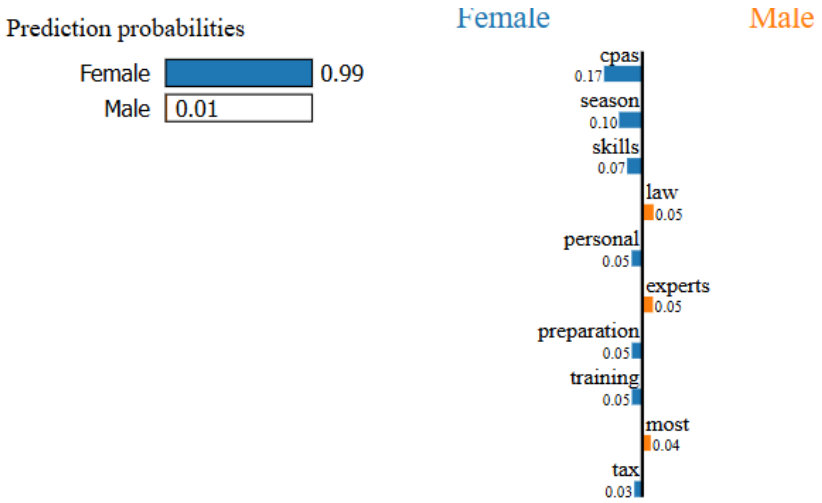


Figure 17: Confusion Matrix when masked Personality Words



Text with highlighted words

[NAME] been issued tennessee license number 10377 . all cpas , including [NAME] , have at minimum undergraduate degree accounting , passed rigorous national exam adhere mandated continuing education requirements states which licensed . [NAME] can work private industry , education or government but most people think cpas during tax season experts tax preparation . overall training business knowledge principles general law taxation provide cpas with skills help individuals with both personal business financial decisions .

Figure 18: Gender classifier with masking Personality Words

3.6. Synthetic Data Generation:

The synthetic dataset of bios, CVs, cover letters, and job ads generated using LLaMA-2-7B and LLaMA-3.2-3B served as the foundational input for evaluating gender bias in text classification. By including explicit gender information in bios and prompting the models to generate job application materials, we simulated real-world scenarios where identity-related features may subtly or overtly influence content generation.

These LLM-generated texts were then passed through trained gender classifiers to assess how much gender information remained detectable after various masking strategies were applied. Classifier performance metrics (accuracy, precision, recall, F1 (Figure 30)) across different levels of masking (e.g., names/pronouns only, social/workplace roles, personality traits) demonstrated that even synthetic content carries residual gender signals often aligned with stereotypes.

3.6.1. Job Advertisement Generation:

Each bio was used to condition the generation of job advertisements using both LLaMA-2-7B and LLaMA-3.2-3B models. We employed a consistent prompt template:

"Generate a job ad for the following profession. {profession}"

The experiment code loads bios from a CSV file of 10 'accountant' bios. For each bio, it fills the template with details like profession and gender to create a prompt. This prompt is fed into the models using a text-generation pipeline. The models respond with full job ads, which are saved into organized folders. The process is repeated with different models and token limits to compare outputs. It helps explore how language models handle personalized job ad creation. The generated job ads reflect how the models interpret and emphasize different features from a candidate's profile when generating a corresponding job opportunity.

(Git: P2_BiasMitigation_SyntheticData/src/model_Llama2_7b/generate_jobAD.ipynb for model LLaMA-2-7B)

(Git: P2_BiasMitigation_SyntheticData/src/model_Llama3.2_3B/generate_jobAD.ipynb for model LLaMA-3.2-3B)

Focusing on the “Accountant” role as a case study, we extracted all 10 job ads generated for accountant-related bios and summarized them into a single unified job description for each model. This summarization was performed using the prompt detailed as:

<|start_header_id|>system<|end_header_id|>

You are a professional HR assistant. Your job is to combine multiple job ads into a single unified job advertisement.
<|eot_id|>

<|start_header_id|>user<|end_header_id|>

Below are several job ads for the role: "{job_title}".

You must:

*- Write **only one** summarized job ad.*

*- Do **not** repeat content or give multiple versions.*

- Use this format:

 # Job Title: {job_title}

[One paragraph summarizing the role]

 # Key Responsibilities

- Bullet 1

- Bullet 2

 # Required Qualifications

- Bullet 1

- Bullet 2

 # Benefits

- Bullet 1

- Bullet 2

Also give details how and where the candidate can apply.

Here are the job ads:

{combined}

<|eot_id|>

<|start_header_id|>accountant<|end_header_id|>

3.6.2. CV and Cover Letter Generation (Generic Application):

For each bio, we generated both a CV and a cover letter using each model and generated job ad. Separate prompts were used for CVs and cover letters to better reflect real-world application document structures:

- CV prompt:

"Generate a professional CV based on the following biography:\n\n{bio}\n\n

Generated CV:"

- cover letters prompt:

"Write a personalized cover letter for the following biography:\n\n{bio}\n

Generated CoverLetter:"

This resulted in a total of 2 x10 x 2 documents (10 CV and 10 cover letters per model).

To simulate a scenario where candidates apply for a specific job, sampled 'accountant' bio csv file was selected and asked each model to generate a targeted CV and cover letters using the summarized job description. Subsequently, the same prompts for CV and cover letter generation were reused to produce a unified batch of outputs, which were compiled into a consolidated CSV file to facilitate downstream analysis and further experimental evaluation.

(Git: P2_BiasMitigation_SyntheticData/src/model_Llama2_7b/generate_CV_CoverLetter.ipynb)

(Git: P2_BiasMitigation_SyntheticData/src/model_Llama3.2_3B/generate_CV_CoverLetter.ipynb)

(Git: P2_BiasMitigation_SyntheticData/src/model_Llama3.2_3B/gen_summarizedJobAD_accountant.ipynb)

3.7. Experiments on synthetic data:

3.7.1 Gender classifier on generated synthetic data (without lower casing and masking):

To interpret model predictions and identify influential features within the synthetic application documents, we applied the LIME explainer directly to the consolidated CSV containing generated CVs and cover letters. The input text was used in its original form without applying any preprocessing like stop word removal, lowercasing, or word masking to preserve the semantic and lexical nuances in the generated content. This approach allowed us to assess how surface-level features and identity-linked language might contribute to classifier behavior.

The index = 2 of text is used as single instance to find LIME explanation. The explanation reveals the model's prediction on the unmasked, generated CV/Cover Letter text, assigning a probability of 0.59 to 'Female' and 0.41 to 'Male' (Figure 19). Without any masking of names, pronouns, or identity-related cues, the explanation highlights specific word associations that influence the model's gender classification.

(Git: P2_BiasMitigation_SyntheticData/src/model_BERT/classifier_Generated_text.ipynb)

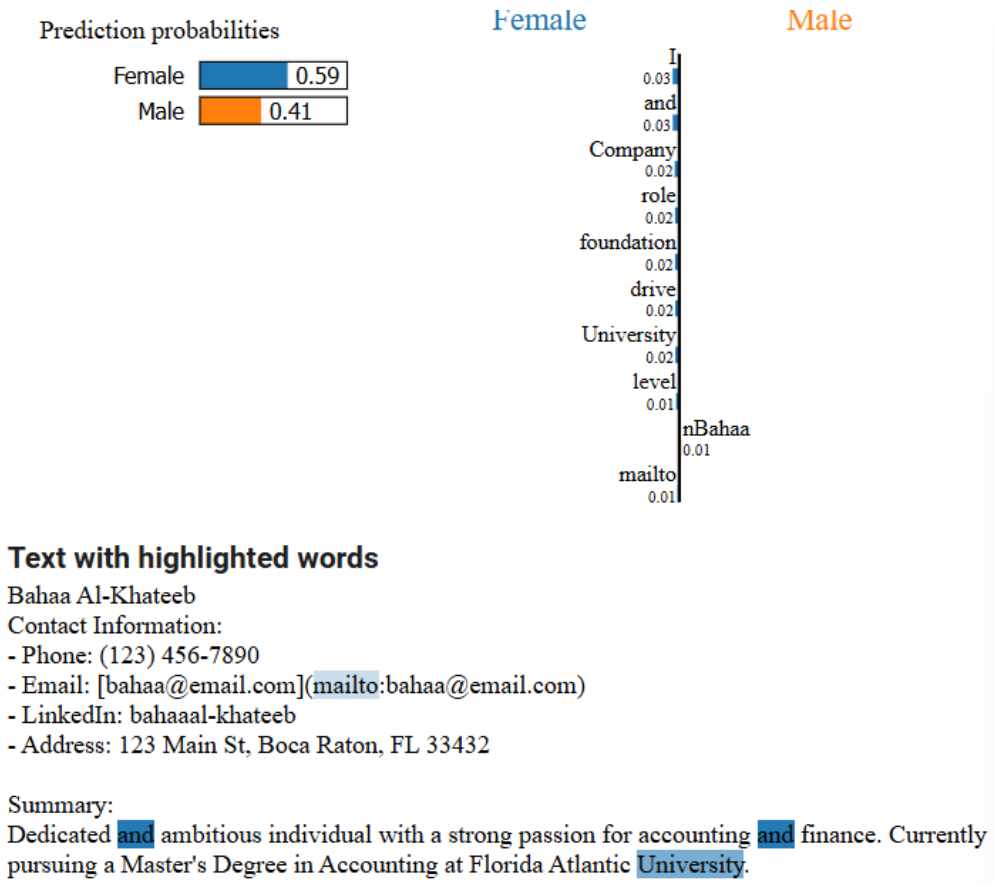


Figure 19: Gender classifier on generated CV/cover letters without any masking

3.7.2. Classifier with lowercasing the text and stop words, and masking names, pronouns, social and workplace proxies:

To reduce surface-level lexical bias and examine deeper structural and semantic patterns in gender prediction, we conducted a second round of classification using a pre-processed version of the synthetic CV and cover letter dataset. This version applied the following text normalization and masking steps:

- Lowercasing: Complete text was converted to lowercase to neutralize case-based variations.
- Stop Word Removal: Common English stop words (e.g., *the*, *and*, *in*) were removed to emphasize meaningful content terms.
- Entity and Proxy masking:
  - Names
  - Pronouns (e.g., *he*, *she*, *his*, *her*), Figure 8
  - Social identifiers (e.g., *mother*, *husband*), Figure 13
  - Workplace and positional cues (e.g., *manager*, *assistant*, *executive*)

Each identified token was replaced with a neutral placeholder (e.g., [NAME], or [__]) to minimize the influence of direct identity cues on the classification process.

This setup was designed to test whether gender prediction could still be achieved when overt linguistic signals were stripped, thereby isolating subtler semantic or stylistic patterns. The resulting model performance and LIME explanations were compared against the original, unprocessed classifier to evaluate the persistence of bias even under content obfuscation.

The LIME classifier is applied to a single instance at index = 2 to generate an explanation of the model's gender prediction. After applying comprehensive masking including names, pronouns, social and workplace proxies, along with lowercasing and stop word removal, the model assigns a probability of 0.58 to 'Female' and 0.42 to

'Male' (Figure 20). The LIME explanation identifies which remaining word features, despite the removal of explicit identity cues, continue to influence the model's classification.

(Git: P2_BiasMitigation_SyntheticData/src/model_BERT/masked_proxies_generatedText.ipynb)

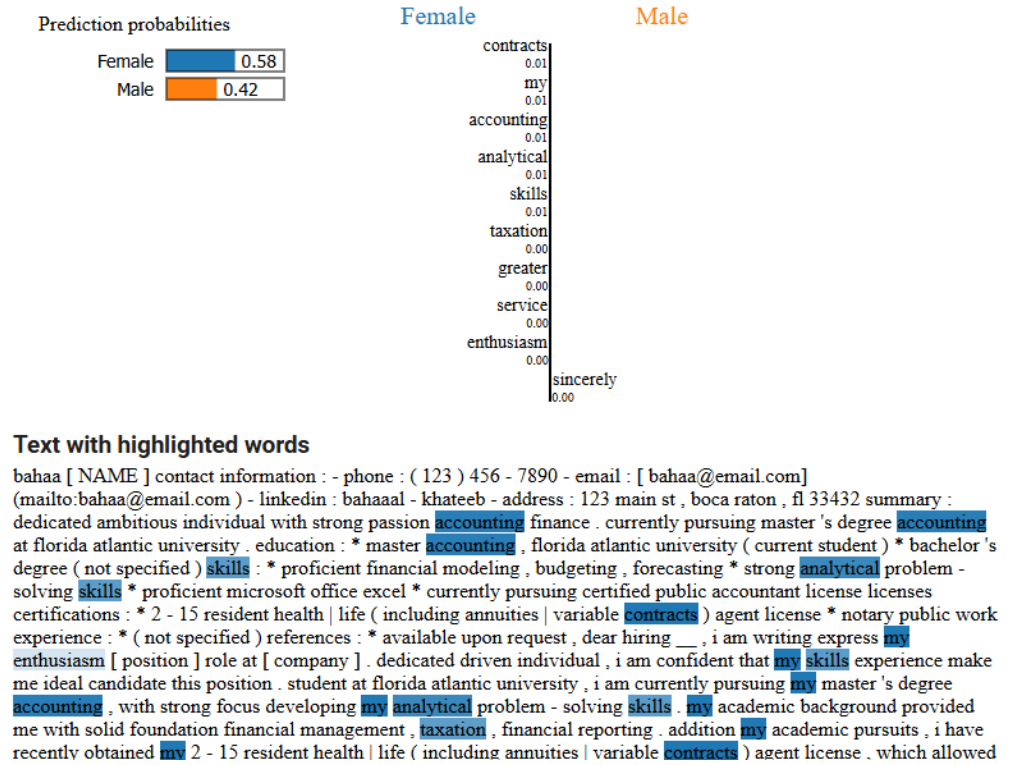


Figure 20: Gender classifier on generated CV/cover letters with lowercasing and stop words, and masking names, pronouns, social and workplace proxies

4. Results

4.1. Analysis on BiasBios dataset:

As shown in Figures 1 and 2, the dataset is 53.9% male (213,543) and 46.1% female (182,646). While this difference may seem small, it becomes more serious when combined with gender-stereotyped job roles. The professions range from the most frequent, like "professor" (118,076 biographies), to the least frequent, like "rapper" (1,403 biographies) (Figure 3). The pie chart (Figure 4) shows how biographies are spread across different professions in the dataset. Each slice of the chart represents a profession, with its size showing the percentage of biographies it contributes. The largest slice belongs to professors, making up 29.8% of the data, followed by physicians (10.3%) and attorneys (8.2%). Professions with fewer biographies, such as rapper, comedian, or paralegal, appear as much smaller slices and their labels are somewhat crowded. Overall, the chart gives a clear view of which professions are most and least represented. The disaggregation of BiasBios dataset by profession illustrates that certain occupations are strongly gender-coded in the dataset (Figure 5)

For instance, female representation is overwhelmingly dominant in roles such as:

- Dietitian (93%)
- Nurse (91%)
- Yoga Teacher(85%)
- Interior Designer (81%)

In contrast, male-dominated professions include:

- Rapper (90%)
- Dj(86%)
- Surgeon (85%)

- Software Engineer (85%)

The analysis of the BiasBios dataset identified gender imbalances both in overall representation and gender-stereotyped occupations (e.g., 90% of “rappers” male; 93% of “dietitians” female). This supports the detection of societal and occupational gender biases in training data.

4.2. Gender Classifier:

When explaining the model's prediction for an instance, LIME classifer highlights the words that contributed most to the model’s decision. By analyzing multiple instances where the model predicts a particular label (e.g., “male” or “female”), one can observe if certain words consistently have high importance. The classifier’s performance is an indicator of how much sensitive gender information is present in the data.

As the experiment unfolds, a smaller diff value indicates that the model found it harder to distinguish between class 0 and class 1, i.e., it was less confident in its prediction. The top rows in the table have the lowest diff values (e.g., 0.038370), meaning the model was almost equally likely to assign the input to either class.

4.2.1. Classifier without masking or lowercase text:

Classification experimented without masking or lower text shows certain words seemed to push the model more strongly toward one gender than the other. For example, male-related words like *he*, *his*, *him*, *Mr.*, *Penticostal*, *services* (from Figure 21) or *work, number (from Figure 6)* made the model more likely to predict "male". On the other hand, female-related words such as *she*, *her*, *Ms.*, *Mrs.*, *Jessica*, *CPAs* (from Figure 21) or *license, individuals (from Figure 6)* increased the chances of predicting "female".

Top 10 Male words	Top 10 Female words
('He', 3.34961242228) ('he', 1.9166196645) ('his', 0.9823642110) ('him', 0.59896187373) ('Mr', 0.212177381359) ('For', 0.15553675710) ('an', 0.139449532486) ('services', 0.127056438325) ('help', 0.12677291047) ('Penticostal', 0.126314986395)	('She', 20.334972885) ('her', 6.5985962079) ('she', 5.3634170559) ('Her', 2.318591337) ('Ms',-1.21109984780) ('Mrs', 0.54779522537) ('Jessica', 0.54219316489) ('herself', 0.51449886288) ('Their', 0.488303965837) ('CPAs', 0.47497754571)

Figure 21: List of top 10 Gender words without masking.

4.2.2. Classifier with lowercase text and without masking:

The confusion matrix (Figure 9) predicted all correct predictions, 121 for male and 79 for female, without any discrepancies. Classification with lowercase text and without any masking shows certain words clearly influenced the model’s predictions more than others. For instance, male-associated words like *he, his, him, law* (Figure 7, 22), made the model more likely to predict "male." In contrast, female-associated words such as *she, her, ms, cpas, general* (Figure 7, 22) increased the likelihood of predicting "female". These patterns highlight how text-based models can learn and reproduce gender biases based on word usage, job roles, and social cues embedded in the data.

Top 10 Male words	Top 10 Female words
('he', 18.6270818434) ('his', 5.95132019243) ('has', 2.3456170134) ('him', 1.4558856212) ('is', 0.89128831843) ('mr', 0.71970626260) ('and', 0.66137833001) ('of', 0.49258934339)	('she', -19.308946400) ('her', -6.0160686546) ('ms', -1.3474851797) ('wendy', -0.4824392623) ('jessica', -0.4697531498) ('catherine', -0.3979233304) ('carla', -0.3951471529) ('women', -0.36915380735)

('was', 0.47370002206) ('with', 0.334369473813))	('heidi', -0.3537158369) ('helen', -0.2952141328)

Figure 22: List of top 10 Gender words(lowercase)

4.2.3. Classifier with masking gender pronouns and lowercase text:

The confusion matrix (Figure 9) predicted all correct predictions, 121 for male and 79 for female, without any discrepancies. Classification with masking gender pronouns and lowercase text resulted in the top words influencing a female classification including the terms like *kenneth, cpas, preparation, they, alison, responsible* (Figure 10, 23). In contrast, words such as *requirements, government, chairman, kpmg, Andrew* (Figure 10, 23) dominate male-associated predictions.

Top 10 Male words	Top 10 Female words
('and', 3.192428956) ('has', 1.0033571281) ('was', 0.441153098458) ('wife', 0.40943360635) ('chairman', 0.388494858729) ('robert', 0.3802675772) ('kpmg', 0.3711696241) ('andrew', 0.3654507875) ('he', 0.35023933503) ('steve', 0.33476447467)	('they', -7.610521073) ('heidi', -0.7688661750) ('them', -0.7328463394) ('the', -0.71734357639) ('alison', -0.6482208696) ('a', -0.57570229632) ('eithne', -0.529211775) ('carla', -0.513791318) ('responsible', -0.4519028186) ('ava', -0.4235735794)

Figure 23: List of top 10 Gender words with masking gender pronouns

4.2.4. Classifier with masking Names and pronouns:

The model correctly predicts 109 instances as class 1 (likely "Male") and 62 instances as class 0 (likely "Female"), as shown in confusion matrix (Figure 11). However, 17 female samples were misclassified as male, and 12 male samples were misclassified as female. Compared to the previous matrix, this result shows a slight increase in misclassifications for both classes. While the overall accuracy remains relatively high, the rise in false negatives for male samples and false positives for female samples may suggest a residual bias or less effective masking. This highlights the importance of comprehensive de-biasing techniques to ensure balanced model performance across gender labels.

The LIME classifier revealed the prediction probabilities and feature attributions for gender classification when masked for Names and pronouns both. Words most associated with the 'Female' prediction included *business, licensed, accounting, organization, work, eithne* (Figure 12, 24). In contrast, 'Male'-associated terms reflect like *rigorous, married, firm, management, clients, experience, institute* (Figure 12, 24).

Top 10 Male words	Top 10 Female words
('their', 12.5492907924) ('been', 1.25716017708) ('that', 1.0319167027) ('married', 0.91890865438) ('firm', 0.73854006905) ('clients', 0.6492293269) ('management', 0.61584069878) ('they', 0.5904155409) ('experience', 0.55569828006) ('institute', 0.52787786131)	('them', -15.3266355036) ('johanna', -0.76427692984) ('accounting', -0.64208592810) ('catherine', -0.55804792203) ('businesses', -0.49596419502) ('helen', -0.374687705005) ('organization', -0.37373964111) ('work', -0.34462881799) ('eithne', -0.34256573962) ('lorraine', -0.33932496026)

Figure 24: List of top 10 Gender words) with masking Names and pronouns

4.2.5. Classifier with masking Social and Workplace roles:

The confusion matrix (Figure 14) reveals that the model correctly identified 94 instances as class 1 (likely "Male") and 52 instances as class 0 (likely "Female"), while it misclassified 27 female samples and 27 male samples. This demonstrates a more balanced distribution of errors across gender classes, suggesting improved fairness after comprehensive masking.

The top influential words contributing to the 'Female' classification, when masked social and workplace roles, include *cpas, preparation, adhere, university, bachelor, amy, jane* (Figure 15, 25) and the top influential words contributing to the Male classification include *experts, institute, firm, cfo, business, chairman* (Figure 15, 25).

Top 10 Male words	Top 10 Female words
('NAME', 1.40994683670)	('university', -1.3247715430)
('institute', 1.03750722402)	('bachelor', -0.99653586220)
('firm', 0.78081602135)	('state', -0.89345809311)
('cfo', 0.6758165090)	('johanna', -0.82489687231)
('years', 0.66752307965)	('accounting', -0.8174670716)
('business', 0.6193116215)	('amy', -0.78793680371)
('married', 0.60418771650)	('pia', -0.7731146578)
('chairman', 0.56477778048)	('jane', -0.76405593641)
('ceo', 0.55454730737)	('lorraine', -0.75804722857)
('clients', 0.54660815742)	('catherine', -0.72474893319)

Figure 25: List of top 10 Gender words) with masking Social and Workplace roles

4.2.6. Classifier with masking Personality traits:

Figure 17 shows the confusion matrix reveals that the model correctly identified 94 instances as class 1 (likely "Male") and 54 instances as class 0 (likely "Female"), while it misclassified 25 female samples and 27 male samples.

The classifier with masking Personality traits provides insights into the top words influencing each prediction. For the 'Female' class, words like *cpas, preparation, skills, ava, pia, bachelor* (Figure 18, 26) were influential. On female side many names also appeared even if the experiment was masked for names. On the other hand, for the 'Male' class, words such as *experts, married, firm, clients, institute, chairman* (Figure 18, 26) were highlighted terms that are more male oriented.

Top 10 Male words	Top 10 Female words
('years', 1.29491006176)	('_', -4.22316722135)
('married', 0.90099026483)	('bachelor', -0.9535235141)
('firm', 0.87805425899)	('lorraine', -0.82106554994)
('clients', 0.74631021627)	('ava', -0.74518614543)
('served', 0.7461163814)	('she', -0.74273792027)
('institute', 0.6255155939)	('jane', -0.74153944060)
('management', 0.61325630144)	('catherine', -0.71409591950)
('well', 0.5825395472)	('cheryl', -0.70532373534)
('chairman', 0.55556668412)	('pia', -0.69032587442)
('firms', 0.5302667863)	('alison', -0.68559825800)

Figure 26: List of top 10 Gender words with masking Personality traits

4.3. Evaluation on Synthetic data:

4.3.1. Classifier on generated synthetic data (without lower casing and masking):

Top tokens associated with female classification included *foundation, experience, Administrative, College, Company* (Figure 19, 27). Top tokens associated with male classification included *financial, analysis, academic, confident, Main* (Figure 19, 27). These findings reflect subtle stylistic and semantic differences in LLM-generated documents that may reinforce or reflect stereotypical gender-coded language.

Top 10 Male words	Top 10 Female words
('financial', -0.00726407444726) ('My', -0.0073886239074) ('analysis', -0.0087938032576) ('khateeb', -0.0127161552284) ('my', -0.0137001283535) ('academic', -0.0155408240784) ('Bachelor', -0.016159588071) ('confident', -0.0170786049531) ('Main', -0.0219754586834) ('Sincerely', -0.055703040213)	('and', -0.0461288755242) ('that', -0.0409926713792) ('experience', -0.038402671980) ('basketball', -0.033762133936) ('Administrative', -0.029336822534) ('College', -0.028014567851) ('dedicated', -0.0256787306839) ('Company', -0.024460654030) ('life', -0.0235013089589) ('contracting', -0.0219289515027)

Figure 27: List of top 10 Gender words on generated CV/CL without any masking

4.3.2. With lowercasing and stop words, and masking names, pronouns, social and workplace proxies:

Following the initial classification on raw, unprocessed CVs and cover letters, we re-ran the LIME explainer on a version of the synthetic dataset where all text was lowercased, stop words removed, and key identity-related entities such as names, pronouns, social descriptors, and workplace role proxies were masked. This preprocessing aimed to neutralize overt lexical gender markers and evaluate the classifier's reliance on deeper linguistic signals.

Prediction Probabilities resulted to Female class: 0.52 and Male class: 0.48 (Figure 25)

The top influential tokens identified by LIME for Female-associated words are *contracts, accounting, analytical, confident, homes, company, cultural* (Figure 20, 28) and for Male-associated words are *sincerely, finance, service, forecasting, ambitious* (Figure 20, 28). Interestingly, even after identity masking, certain semantic and stylistic cues persisted in influencing gender classification.

Top 10 Male words	Top 10 Female words
('456', -7.020630503546581e-05) ('finance', -0.0002289562524759334) ('public', -0.0013063261646777704) ('email', -0.002619938307502332) ('services', -0.003168509081936963) ('forecasting', -0.0039705995594961395) ('ambitious', -0.0041057938586260044) ('i', -0.004934802599573794) ('insurance', -0.008224237896624348) ('pursuing', -0.009845348810501014)	('confident', -0.014789551120420002) ('deliver', -0.01161894296559811) ('homes', -0.008085604073895383) ('company', -0.008048286452979196) ('oriented', -0.008043788394657417) ('cultural', -0.007721886366985236) ('excited', -0.00705174816728466) ('experienced', -0.005951034536367499) ('exchange', -0.005062909615712394) ('increase', -0.004980663319936303)

Figure 28: List of top 10 Gender words on generated CV/cover letters with lowercasing and stop words, and masking names, pronouns, social and workplace proxies

4.4. Analysis based on Accuracy metrics:

To evaluate how linguistic obfuscation and preprocessing impact gender classification performance, the performance metrics of Accuracy, Recall, Precision, and F1 Score on all classifiers under varying levels of text masking and preprocessing are conducted. The baseline classifier used the original, unprocessed dataset (with full names, pronouns, and casing intact), and additional variants introduced targeted masking strategies. The report in the Figure 29 is interpreted as follows:

- The baseline model (no masking, original casing) achieved perfect scores across all metrics (Accuracy = 1, F1 = 1), demonstrating that direct gender markers (e.g., names, pronouns) strongly influence classification.
- Applying lowercasing or masking only pronouns did not affect performance; scores remained at 100%, indicating these signals alone are not critical when other identity cues are present.
- Masking both names and pronouns led to a moderate drop in performance (Accuracy = 0.85, F1 = 0.88), showing these cues are important but not solely responsible for gender prediction.
- Masking proxy words (e.g., gender-associated job roles, social descriptors) further reduced accuracy to 0.73, and masking personality-related terms yielded similar results (Accuracy = 0.74, F1 = 0.78).

Gender Classification	Acccuracy	Recall	Precision	F1
Classifier Baseline (without any masking and no lowercase text)	1	1	1	1
With lowercase dataset	1	1	1	1
With masked pronouns	1	1	1	1
With masked Names and Pronouns	0.85	0.9	0.86	0.88
With masked proxy words	0.73	0.77	0.77	0.77
With masked personality words	0.74	0.77	0.79	0.78

Figure 29: Table of accuracy metrics for classifier experiments on 'accountant' bios

Gender Classification	Acccuracy	Recall	Precision	F1
Classifier on generated text without masking or lowercase text	0.75	0	0	0
With masking all proxies on generated text	0.75	0	0	0

Figure 30: Table of *balanced* accuracy metrics for classifier experiments on generated data of CV and cover letters

When applying the classifier to LLM-generated application text without masking or lowercasing, performance dropped sharply (Accuracy = 0.75, Recall = 0) (Figure 30).

4.5. Analysis based on Low confidence of gender predictions:

To better understand the certainty with which the gender classification model makes its predictions, the proportion of low-confidence predictions under different levels of text masking and preprocessing was measured. When only lowercasing was applied, low-confidence predictions decreased sharply to 0.45 (Figure 31). Masking pronouns alone also increased uncertainty (0.70), indicating that even simple lexical gender cues significantly influence the classifier's certainty. Masking both names and pronouns led to a sharp

drop in confidence (0.17), while further masking of proxy words (e.g., job roles, identity indicators) caused low-confidence predictions to drop to 0.02, indicating the classifier was no longer able to confidently detect gender. Similarly, masking personality-associated terms indicated low confidence to 0.05, showing that stylistic language also contributes to uncertainty.

The classifier also showed very low confidence when applied to LLM-generated CVs and cover letters (Figure 31):

- masking or lowercasing: Low Confidence = 0.09
- With all proxies masked: Low Confidence = 0.14

Gender Classification	Low Confidence	High confidence
Classifier Baseline (without any masking and no lowercase text)	0.99	0.99
With lowercase dataset	0.45	0.99
With masked pronouns	0.7	0.99
With masked Names and Pronouns	0.17	0.99
With masked proxy words	0.02	0.99
With masked personality words	0.05	0.98
Classifier on generated text without masking or lowercase text	0.09	0.62
With masking all proxies on generated text	0.14	0.42

Figure 31: Table of Low confidence while gender predictions

5. Discussion

Dataset:

This study explores how gender bias persists in text classification models used in hiring, using both real bios from the BiasBios dataset and synthetic CVs and cover letters generated by language models. The BiasBios dataset revealed strong gender and occupational imbalances that reflect real-world labor trends. Our experiments show that gender can be inferred not only from obvious clues like names and pronouns but also from subtle patterns in language and job descriptions. Our research identifies, assesses, and mitigate biases that can be amplified by machine learning classifiers and NLP models. Our findings further support its role as a realistic benchmark for testing how models respond to demographic signals embedded in professional biographies. The analysis of the BiasBios dataset reveals notable gender imbalances, both in overall distribution and profession-specific representation. While the dataset appears nearly balanced at 53.9% male and 46.1% female, deeper examination shows stark gender-stereotyping in job roles. Professions like “nurse” and “dietitian” are overwhelmingly female, while roles like “surgeon” and “rapper” are predominantly male. With nearly 30% of the data focused on professors, some occupations are vastly overrepresented, potentially skewing model learning. These patterns reflect entrenched societal stereotypes and highlight the risk of embedding occupational gender bias in downstream NLP and hiring systems.

The imbalances in the dataset BiasBios (Section 2.1) don’t just reflect social trends, they directly affect how LLMs learn. Research shows that NLP models trained on such data start linking gendered language with specific jobs, which lead to biased or unfair treatment of candidates.

This analysis evaluates how various text obfuscation strategies affect gender classification performance using accuracy, recall, precision, and F1 score. The baseline classifier, trained on unprocessed bios with full names and pronouns, achieved perfect scores, indicating strong reliance on explicit gender indicators. Masking only pronouns or applying lowercasing had negligible impact, showing that other cues like names still drive accurate predictions. However, masking both names and pronouns led to a moderate performance drop (Accuracy = 0.85), highlighting their importance in gender inference. Further masking of proxy terms such as job roles and social descriptors resulted in a notable decrease in accuracy (0.73), revealing the classifier's dependence on indirect gender cues. A similar decline occurred when personality-related terms were masked, indicating the presence of stylistic bias. When applied to LLM-generated CVs and cover letters, the classifier's accuracy dropped further (0.75), suggesting that synthetic texts contain subtler gender signals. These findings stress the layered nature of linguistic gender bias.

Research Analysis:

The gender classification experiments reveal how textual cues ranging from explicit identifiers to subtle patterns enable models to infer gender with high confidence. Using LIME, the classifier consistently highlighted gendered terms (e.g., he, her, Ms., Mr.) and names as the most influential features when no masking or text normalization was applied, showing strong gender predictability rooted in lexical content. Even when gender pronouns or names were masked, models leveraged indirect signals such as occupational roles (chairman, accounting), stylistic patterns (sincerely, management), or professional terms (CPAs, firm) to infer gender. Lowercasing and masking reduced but did not eliminate bias, as classifiers still responded to residual proxies and contextual phrasing. This anylysis helps us to find the answer for our RQ1.

When using LLMs to generate application materials (CVs, cover letters) from gendered bios, we observed that the generated text inherits existing gender cues. LIME-based analysis revealed:

- Female-coded outputs often contained emotional or interpersonal language (e.g., *excited*, *empowering*)
- Male-coded outputs emphasized assertiveness and achievements (e.g., *problem*, *gain*, *prepared*, *sincerely*)

Even when not explicitly asked to include gender-specific content, models tend to mirror real-world stereotypes found in their training data. This suggests that LLMs are highly susceptible to reproducing gender bias, especially in domains like hiring where occupational stereotypes are strong.
Careful prompt design, data augmentation with counter-stereotypical examples, and post-generation filtering are potential strategies to reduce this risk in synthetic document generation. This gives the analysed result for our RQ2.

LIME (Local Interpretable Model-agnostic Explanations) has proved to be a valuable tool in:

- Identifying key lexical features that drive gender predictions
- Comparing models' reliance on explicit vs. implicit gender cues
- Assessing the effectiveness of masking strategies across multiple classification setups

By visualizing token importance, LIME helps researchers pinpoint which features contribute to biased outcomes, thereby enabling more targeted interventions. Furthermore, combining LIME analysis with confidence scores provides a multi-dimensional understanding of model behavior under different data conditions. However, LIME still identified gender-skewed features such as "*support*," "*empowering*" (female) and "*certifications*," "*strong*" (male), suggesting that stylistic tone and thematic framing remain as residual signals for gender inference, even under full anonymization. This analysis helps us understand and find answer to our RQ3.

Answer to our RQ4 implies that the explored methods in the project show strong potential to mitigate unfair discrimination in recruitment. The study demonstrates that gender information in application texts both real

and synthetic can be partially masked using targeted strategies like removing names, pronouns, social/workplace roles, and personality traits. These masking techniques significantly reduce a classifier's ability to infer gender, especially when combined with stylistic normalization and text obfuscation. Additionally, tools like LIME allow for transparent inspection of model decisions, revealing which features contribute most to gender prediction. This interpretability enables recruiters and developers to identify bias-prone language and apply corrections. When applied to LLM-generated CVs and cover letters, the study shows that without intervention, synthetic texts tend to reproduce gendered patterns found in training data (e.g., interpersonal words in female profiles vs. achievement-oriented words in male profiles). However, with careful prompt design, masking, and LIME-based audits, these biases can be mitigated.

Conclusion:

Notably, with increasing obfuscation masking social/workplace roles or personality traits the classifier showed a more balanced confusion matrix and reduced confidence margins, indicating a dampening of bias-laden signals. These patterns suggest that while masking techniques are effective at reducing overt gender inference, models still rely on latent stereotypes embedded in language structure. Hence, bias persists through deeper semantic associations, highlighting the challenge of achieving truly fair NLP systems for hiring scenarios.

The analysis of low-confidence predictions provides critical insights into how text obfuscation affects a model's certainty in gender classification. When only lowercasing was applied, low-confidence predictions were moderate (0.45), while masking pronouns increased uncertainty (0.70), highlighting the role of even simple lexical gender cues. Masking both names and pronouns dropped confidence to 0.17, and further masking of proxy terms like job roles lowered it to just 0.02. Similarly, masking personality traits reduced confidence to 0.05. Synthetic CVs and cover letters also produced low confidence scores (0.09 with lowercasing, 0.14 with full masking), showing that LLM-generated content lacks strong gender markers.

Low-confidence predictions are especially valuable for identifying ambiguous or hard-to-classify samples, which can be flagged for manual review, prioritized in active learning, or targeted during model fine-tuning. Conversely, analyzing high-confidence predictions allows researchers to validate model reliability, develop high-precision subsets for downstream tasks, and contrast feature patterns between confidently and ambiguously classified cases.

Certain words appeared to influence the model's decisions more strongly toward one gender than the other. These included direct gender indicators (e.g., pronouns, titles, and names) as well as indirect cues, such as professional roles or cultural references. This suggests leaving text unprocessed (i.e., not masking identity terms or normalizing stylistic features) enables models to learn and act upon the signals of direct gender indicators (e.g., he, Ms., Jessica) or indirect or proxy indicators (e.g., nurse, CEO) leading to potential gender bias in classification or downstream decisions (e.g., CV screening).

The noticeable presence of emotional or relationship-focused words in outputs linked to females, compared to more action- and task-focused words in those linked to males, reflects well-known patterns in gendered communication found in professional writing. For example, female-associated texts might include words like *"support,"* or *"collaborate,"* emphasizing teamwork and relationships. In contrast, male-associated texts often use words like *"leader," "manage,"* or *"execute,"* which focus more on tasks and achievements. These differences support previous research showing women tend to communicate with a relational style, while men focus more on action, which can shape perceptions and reinforce gender stereotypes in professional settings.

Our experiments show that a significant amount of gender information is embedded in both real and synthetic documents even when direct indicators such as names and pronouns are removed. Classifiers trained on unprocessed bios or CVs achieved near-perfect accuracy and confidence, with LIME revealing reliance on not only gendered terms (e.g., *she*, *Mr.*, *Jessica*), but also proxy words (e.g., *nurse*, *director*, *chairwoman)*.
Masking explicit identity markers (names, pronouns, social/professional roles) reduces the classifier's ability to confidently infer gender, but it does not eliminate gender signals completely. This implies that gender is redundantly encoded in language through both vocabulary and structure.
Thus, effective masking combine:

- Lexical masking (names, pronouns, identity terms) (e.g. *jessica, Andrew, amy, she, he, ms, mrs.)*
- Contextual masking (job roles, personality traits) (e.g. *ceo, cfo, director, accounting, married*)
- Stylistic normalization (removal of affective cues) (e.g. *responsible, supportive, sincerely, expert*)

The word "cpas" has appeared frequently in most of the masked listings for female profiles. CPA stands for *Certified Public Accountant* (meaning extracted from open source), which is a professional accounting designation. Its repeated presence implies that many of the female listings are related to accounting or finance roles requiring this qualification. This suggests a pattern where female profiles are often associated with or focused on accounting professions.

Together, all the techniques discussed provide a promising framework for creating fairer, more inclusive recruitment tools by reducing model reliance on demographic proxies and helping shift focus toward skills, qualifications, and merit.

Limitations:

In many classifier experiments, even after masking or removing explicit names, female names still appear prominently in the top listings or explanations. This reveals a limitation in methods like LIME and the classifiers themselves. This can be because LIME tries to explain model decisions by approximating complex models locally, but it relies on features the model uses which can include indirect or subtle cues related to gender. Even if names are masked, other correlated features (like certain words, phrases, or contexts often associated with female profiles) may remain. The classifier can still pick up on these patterns, unintentionally revealing gender information. It highlights the challenge of creating truly unbiased and interpretable models, suggesting the need for more advanced techniques to detect and mitigate hidden biases.

Future work could explore:

- Adversarial debiasing techniques
- Inclusion of fairness constraints during model training
- Human-in-the-loop frameworks to guide fair evaluation and intervention

Ultimately, these methods provide a scalable and interpretable pathway toward reducing algorithmic discrimination in hiring and promoting equitable access to opportunity.

6. Bibliography

• Git: Shilpi261985/P2_BiasMitigation_SyntheticData
• Bartl, M., Mandal, A., Leavy, S., & Little, S. (2025). *Gender bias in natural language processing and computer vision: A comparative survey*. https://dl.acm.org/doi/full/10.1145/3700438
• Bolukbasi, T., et al. (2016). *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. NIPS-2016-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings-Paper.pdf

• Caliskan, A., Bryson, J., & Narayanan, A. (2017). *Semantics derived automatically from language corpora contain human-like biases*. Science

• Chaturvedi, S., & Chaturvedi, R. (2025). *Who Gets the Callback? Generative AI and Gender Bias*. https://arxiv.org/abs/2504.21400

• De-Arteaga et al (2019). *Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting*. arxiv.org/abs/1901.09451

• Delobelle P, Tokpo EK, Calders T, et al (2022). *Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models*. ACL Anthology

• Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://arxiv.org/abs/1810.04805

• Fabris, A., Baranowska, N., Dennis, M. J., Graus, D. et al. (2025). *Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey*. https://dl.acm.g/doi/10.1145/3696457

• Fiske ST (2017). *Prejudices in cultural contexts: Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion)*. PubMed

• Frazzetto, P. (2025). *Leveraging Deep Learning in Human Resources: Graph Neural Networks for Candidate-Job Matching*. Thesis PDF

• Goldfarb-Tarrant S, Marchant R, Sanchez RM, et al (2020). *Intrinsic bias metrics do not correlate with application bias*. [2012.15859]

• Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). *Measuring Individual Differences in Implicit Cognition The Implicit Association Test*. The implicit association test.

• Hovy, D., & Prabhumoye, S. (2021). *Five sources of bias in natural language processing*. PubMed.

• Joshi P, Santy S, Budhiraja A, et al (2020). *The state and fate of linguistic diversity and inclusion in the NLP world*. ACL Anthology

• Kumar, D., Greif, E., Rekabsaz, N., & Schedl, M. (2023). *Identifying Words in Job Advertisements Responsible for Gender Bias via Counterfactual Learning*. CEUR-WS. PDF

• Kurpicz-Briki M, Leoni T (2021). *A world full of stereotypes? Further investigation on origin and gender bias in multi-lingual word embeddings*. (PDF)

• Mansouri, T., Alameer, A., & Albaroudi, E. (2024). *AI Techniques for Addressing Algorithmic Bias in Job Hiring*. *AI*, 5(1), 19. https://www.mdpi.com/2673-2688/5/1/19

• Marti Marcet, S. (2023). *Natural Language Processing in Resume Data: The Interplay Between Gender and Occupation on Resume Writing Style*. Utrecht University. PDF

• Meade N, Poole-Dayan E, Reddy S (2022). *An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models*. *[2110.08527]*

• Meta AI. (2024). *LLaMA 3: Open Foundation and Instruction Models*. https://ai.meta.com/blog/meta-llama-3/

• Mihaljević, H., Müller, I., Dill, K., & Yollu-Tok, A. (2022). *Towards Gender-Inclusive Job Postings: A Data-Driven Comparison of augmented writing technologies*. PLOS ONE, 17(12). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0274312

• Nangia N, Vania C, Bhalerao R, et al (2020). *Crows-pairs: A challenge dataset for measuring social biases in masked language models*. [2010.00133]

• Njoto et al. (2022). *Professional Presentation and Projected Power: A Case Study of Implicit Gender Information in English CVs*. https://aclanthology.org/2022.nlpcss-1.15/

• Peña, A., Serna, I., Morales, A., Fierrez, J., & Ortega, A. (2023). *Human-Centric Multimodal Machine Learning: Recent Advances and Testbed on AI-Based Recruitment*. *SN Computer Science*, 4(5). s42979-023-01733-0.pdf

• Ribeiro, M., Singh, S., & Guestrin, C. (2016). *Why should I trust you?": Explaining the predictions of any classifier*. 1602.04938v3.pdf

• Shumer, B., Jernite, Y., Papazian, A., Sun, Z., Roberts, A., & Raffel, C. (2023). *Mistral 7B*. Mistral AI. https://mistral.ai/news/announcing-mistral-7b/

• Sun et al., (2019). *Mitigating Gender Bias in Natural Language Processing: Literature Review*. ACL Anthology

• Sweeney, M., & Najafian, S. (2020). *Transparent model reporting for NLP bias in hiring*. ACL Anthology

• Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... & Scao, T. L. (2023). *LLaMA 2: Open Foundation and Fine-Tuned Chat Models*. *Meta AI*. https://arxiv.org/abs/2307.09288

• Wang, T., Zhao, J., Yatskar, M., & Chang, K.-W. (2021). *Balanced datasets are not enough*. ArXiv preprint.
• Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). *Transformers: State-of-the-Art Natural Language Processing*. https://aclanthology.org/2020.emnlp-demos.6

7. Declaration of Authorship

I hereby certify that I composed this work completely unaided, and without the use of any other sources or resources other than those specified in the bibliography. All text sections not of my authorship are cited as quotations and accompanied by an exact reference to their origin.

Place, date: BFH, Biel; 30 July 2025

Signature: Shilpi Garg