GenWriter: Reducing Gender Cues in Biographies through Text Rewriting

Shweta Soundararajan and Sarah Jane Delany
Technological University Dublin
shweta.x.soundararajan@mytudublin.ie, sarahjane.delany@tudublin.ie

Abstract

Gendered language is the use of words that indicate an individual's gender. Though useful in certain context, it can reinforce gender stereotypes and introduce bias, particularly in machine learning models used for tasks like occupation classification. When textual content such as biographies contains gender cues, it can influence model predictions, leading to unfair outcomes such as reduced hiring opportunities for women. To address this issue, we propose GenWriter, an approach that integrates Case-Based Reasoning (CBR) with Large Language Models (LLMs) to rewrite biographies in a way that obfuscates gender while preserving semantic content. We evaluate GenWriter by measuring gender bias in occupation classification before and after rewriting the biographies used for training the occupation classification model. Our results show that GenWriter significantly reduces gender bias by 89% in nurse biographies and 62% in surgeon biographies, while maintaining classification accuracy. In comparison, an LLM-only rewriting approach achieves smaller bias reductions (by 44% and 12% in nurse and surgeon biographies, respectively) and leads to some classification performance degradation.

1 Introduction

Gendered language refers to the use of language that explicitly or implicitly convey the gender of a person, animal, or object (Hamidi et al., 2018; Bigler and Leaper, 2015). This can occur explicitly, through words that clearly denote gender, such as mother, she, or man or implicitly, where social roles or behaviors can signal an individual's gender. For instance, women are often expected to exhibit communal characteristics (e.g., emotional, affectionate, gentle), while men are typically linked with agentic traits (e.g., confident, decisive, ambitious) (Gaucher et al., 2011). Although gendered language may serve functional purposes in certain situations, it also has the potential to reinforce harmful gender stereotypes (Bucholtz and Hall, 2004; Leaper and Bigler, 2004). Gender stereotypes are generalized views or preconceptions about attributes or characteristics, that are or ought to be possessed by men and women and behaviours and roles that are or should be performed by men and women (Commissioner, 2014; Blumer et al., 2013; Ellemers, 2018; Morgan and Davis-Delano, 2016; Wiegand et al., 2021). These assumptions, based solely on an individual's gender, can lead to gender bias.

Bias Statement. Gendered language in written content becomes a serious issue when it leads to unfair treatment of an individual based on their gender, identifiable through the content itself. In 2018, Amazon scrapped its AI-powered recruitment model due to gender bias against female applicants (Simaki et al., 2017). Similarly, an occupation classification model trained on the biographies (De-Arteaga et al., 2019) exhibited gender bias, often misclassifying female doctors as nurses. These examples illustrate how gender bias in text classification that involves systematic errors or unfair predictions related to gender can cause allocational harms (Blodgett et al., 2016; Barocas et al., 2017). In both cases, the differences in language use in resumes and biographies influenced the model's decisions, further contributing to its misclassifications (Chang, 2023; Nemani et al., 2024). Gender-based inferences from writing style and language choices can lead to harmful, gender-biased decisions, and potentially impacting career opportunities for female applicants (Madera et al., 2009; Khan et al., 2023; Gaucher et al., 2011; Tang et al., 2017).

It has been shown that adjectives and verbs used to describe women differ from those used for men in contexts such as job advertisements (Gaucher et al., 2011; Tang et al., 2017; Tokarz and Mesfin, 2021), biographies (Wagner et al., 2015; De-Arteaga et al., 2019), recommendation letters

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 347–357

(Madera et al., 2009), articles and fashion magazines (Caraballo Moral et al., 2019; Arvidsson, 2009; Morelius, 2018), and fictional stories (Fast et al., 2016; Williams Jr et al., 1987). These linguistic differences in describing individuals of different gender can introduce gender stereotypes that may lead to gender-based bias, resulting in both conscious and unconscious discrimination (Barocas and Selbst, 2016; Burgess and Borgida, 1999). Therefore, it is important to help or facilitate people to use content where the gender of the person is not clearly evident from the language used, as this can reduce any potential harm caused to individuals.

The aim of this work is to rewrite textual content that describes people in such a way the gender of the person described in the text may not be so evident in the revised version. The approach used is to rewrite text content about a person as if it was written by a person of a different gender. To do this, we use Large Language Models (LLMs) which have become vital tools for text generation across a variety of applications (Sallam, 2023; Transformer et al., 2022; Wan et al., 2023a; Valentini et al., 2023; Hallo-Carrasco et al., 2023) and Case-Based Reasoning (CBR), a problem-solving paradigm, that finds solutions to new problem based on past experiences (Aamodt and Plaza, 1994).

Despite LLM's impressive capabilities in text generation, they can perpetuate gender stereotype and bias through their generated text (Kotek et al., 2023; Dong et al., 2024; Fang et al., 2024; Ovalle et al., 2023; Soundararajan et al., 2023; Wan et al., 2023a). For instance, LLM-generated reference letters, CVs are found to have used more agentic and positive words for men than women (Wan et al., 2023b; Soundararajan and Delany, 2024; Zinjad et al., 2024). This contributes to representational harms, thus disadvantaging a particular group of individuals, more often women.

CBR has also been used in text generation. Prior experiences are captured as cases and made available in a casebase. As we are concerned with rewriting textual content about a person, our cases are sentences from biographies that describe aspects of individuals. The steps involved in reasoning using CBR include (1) case retrieval: retrieving one or more source cases from the casebase that are similar to a query case, i.e. the sentence to be rewritten; (2) case reuse: adapting information from these similar cases to form a solution for the query case.

While CBR is helpful in text generation, adapting past solutions to new problems in a textual domain remains challenging due to natural language variability and complexity. To facilitate adaptation, CBR can be integrated with LLMs which provides benefits to both (Wilkerson and Leake, 2024). Firstly, the integration can reduce the risk of generating content with gender bias and stereotypes by LLM when producing solutions. Secondly, if an LLM could handle the knowledge-intensive aspects of the CBR process, it could significantly expand the range of CBR applications by enabling their use in knowledge-rich domains where formally encoded knowledge is unavailable, expensive, or difficult to encode.

We propose **GenWriter**, an approach that leverages both CBR and LLMs to rewrite textual content containing indicators of gender identity, modifying the content so that the gender of the described individual may not be evident from the language used. We use GenWriter to rewrite biographies of nurses and doctors as these are occupations where gender bias is significant when predicting occupation, with female doctors often misclassified as nurses and male nurses misclassified as doctors (De-Arteaga et al., 2019). This work focuses on rewriting textual content that contains implicit gendered language, rather than explicit gender indicators, which often cannot be altered or may not be meaningful to change—particularly in domains such as biographies, where explicit gender indicators are necessary. We evaluate the performance of rewriting biographies by measuring gender bias in an occupation classification task. A reduction in gender bias in occupation classification is treated as a proxy for successful transformation of biographies.

Our results show that biographies rewritten using our approach used as training data in an occupation classification task, significantly reduce gender bias by almost 89% for nurses and over 62% for surgeons without compromising on classification performance. In contrast, biographies rewritten using only an LLM reduce gender bias by just over 44% and 12% for nurses and surgeons, respectively.

The rest of the paper is organized as follows. Section 2 discusses existing works on rewriting gendered language and using CBR, with and without LLMs, for various text generation tasks. Section 3 elaborates on how the cases are created, retrieved, reused and adapted using LLM in GenWriter to rewrite the biographies and Section 4 presents the evaluation of GenWriter's effectiveness in rewriting biographies and compares its performance to

baseline methods.

2 Related Work

Previous research has explored rewriting gendered language to produce gender-neutral or gender-fair versions. For instance, Pryzant et al. (2020) utilized a BERT model trained on a large corpus of biased and unbiased texts to automatically replace subjective words with neutral alternatives. While effective at addressing lexical (word-level) bias, this technique may overlook deeper contextual or structural biases, such as those embedded in narrative framing or character roles. Similarly, Sun et al. (2021) developed a transformer-based model trained on a rule-generated parallel corpus from Wikipedia to rewrite gendered sentences into gender-neutral forms using singular "they." While this promotes inclusivity, the model defaults to "they" without considering other binary pronouns, potentially reducing the nuance of gender expression. Another study, Amrhein et al. (2023), proposed a transformer trained on synthetic parallel corpora generated via round-trip translation through biased machine translation (MT) systems. This method enables rewriting of gender-biased text into gender-fair alternatives but has the potential to suffer from the noise introduced by MT errors and may not generalize well to real-world examples, as synthetic biases can differ from authentic ones. Other approaches that focused on rewriting or adjusting gendered language included Ma et al. (2020) who introduced a model based on OpenAI-GPT that reduces gender bias by leveraging connotation frames to adjust implied power and agency in character portrayals. However, this method depends on connotation frames that encode pragmatic knowledge of power dynamics in verb predicates, which may limit its generalizability. Finally, Dinan et al. (2019) tackled gender bias in dialogue systems using a multi-faceted approach that includes counterfactual data augmentation, bias-controlled training, and human-curated, gender-balanced datasets. Although this method shows promising results in reducing conversational bias, it requires extensive manual data curation, making it less scalable for large-scale or domain-diverse applications.

CBR has been applied to automated text generation tasks such as anomaly reporting processing (Massie et al., 2007), automated natural language generation for obituaries (Upadhyay et al., 2020), automated generation of sports summaries (Upadhyay et al., 2021), writing product reviews (Bridge and Healy, 2010) and product descriptions (Waugh and Bridge, 2010).

There has also been research that successfully applied the combination of CBR and LLMs for various text generation applications. Minor and Kaucher (2024) uses CBR to retrieve relevant examples from a casebase and integrates them into prompts for LLMs to generate explanations for business process models. Wiratunga et al. (2024) worked on enhancing the performance of LLMs in legal question answering tasks, by using CBR to retrieve relevant past legal cases and integrating them into prompts for LLMs using Retrieval-Augmented Generation (RAG). Similarly, (Marom, 2025)'s framework combines CBR with RAG to enhance LLMs for multimodal tasks, converting non-text case components into text to improve case retrieval and enrich LLM queries. Another work (Yang, 2024) used CBR in combination with LLM to enhance case-based reasoning in healthcare and legal domains. It uses LLMs to process queries, retrieves relevant cases via RAG, and generates actionable insights, improving searchability and precision in complex cases.

3 GenWriter

The aim is to rewrite text as if it was written by someone of a different gender, so that the gender of the described individual is not as evident in the modified text. To this end, we use our approach, GenWriter, which integrates Case-Based Reasoning (CBR) and Large Language Models (LLMs), to generate a revised version of the text. We establish a casebase that serves as a repository of experiences, in our situation, this is sentences describing people that are taken from biographies. In CBR, when there is a new problem, such as a need to transform a text including content about a person into a version where the gender of the described person is less evident, the solutions of similar problems in the casebase are used to address it. LLM plays a dual role within this framework: it assists in constructing cases and in adapting existing solutions to fit the specifics of the current problem, enabling effective integration of CBR with LLM capabilities.

3.1 Case Representation

The case representation reflects how the experience is structured and encoded in the casebase. Each

case within the casebase represents a sentence that describes some aspect of a person. For instance, if we consider the biographies of people, a biography generally begins with a brief overview of the individual's basic details, such as their name, birthplace, age, and occupation. This is followed by education and work experience, including their employer, job role, and professional expertise. Lastly, it touches on personal aspects such as family, hobbies, and interests. Overall, a biography covers four main components: Demographics, Education, Work details, and Non-Professional details. The case representation will include the following:

- **Gender**, indicating the gender of the person being discussed in the sentence.
- **Category**, specifying which aspect of the person is being discussed in the sentence. The four components of the biography– Demographics, Education, Work details, and Non-Professional details are the *Category*.
- **Generalized Sentence**, a sentence about a person related to the *Category*, with pronouns and entities, such as the name of an individual, location, organization, educational institution, dates & time, numbers, award, field of study, occupation, specialization/area of expertise, replaced with context-based placeholders, to ensure entity generalization. This is used both in the retrieval phase of CBR to find the most similar sentence for a sentence that has to be rewritten, and in the reuse and adaption phase of CBR as the rewritten sentence.

The generalized sentence is generated through few-shot prompting (Brown et al., 2020) with an LLM. The LLM is provided with a few-shot prompt, detailed in Table 1, along with the query sentence in order to generate the generalized sentence. Table 2 shows examples of cases created from a biography and their representation using OpenAI's GPT-4o (with the temperature set to 0.7 and all other hyperparameters left at their default values).

Instruction Prompt
Transform a given sentence into a general template by identifying and replacing all entities and pronouns with placeholders that describe the type of entity, as demonstrated in the examples below. Use consistent placeholders throughout, while maintaining the grammatical structure of the sentence. *<few-shot examples>* Your Turn: Input Sentence: *<input_sentence>*
Few-shot Examples
Examples: Input Sentence: Dr. Dilip Nadkarni is an Orthopedic surgeon specialized in Arthroscopic or Key-hole surgery for the Knee Joint. Output: Dr. [Name of the Person] is an [Occupation] specialized in [Specialisation]. Input Sentence: Dr. Crow graduated from University of Arkansas for Medical Sciences College of Medicine in 1966 and has been in practice for 51 years. Output: Dr. [Name of the Person] graduated from [University] in [Year] and has been in practice for [Duration]. Input Sentence: He practices at Apollo Medical Centre with his assistants in Kotturpuram, Chennai, Chennai Speciality Clinic in Besant Nagar, Chennai and Apollo Spectra Hospitals in MRC Nagar, Chennai. Output: [He/She] practices at [Hospital] with [his/her] assistants in [Location], [Hospital] in [Location], [Hospital] in [Location].

Table 1: Instruction prompt and the few-shot examples provided to GPT-4o to generate generalized sentence.

Gender	Category	Generalized Sentence
Female	Demographics	[Name of the Person] is a [Occupation] in [Location].
Female	Education	[He/She] graduated with honours in [Year].
Female	Work Details	Having more than [Duration] of diverse experiences, especially in [Occupation], [Name of the Person] affiliates with [Hospital].

Table 2: Cases created from the following biography: *Sejal P Graber is a Nurse Practitioner Specialist in Everett, Washington. She graduated with honours in 2006. Having more than 10 years of diverse experiences, especially in Nurse Practitioner, Sejal P Graber affiliates with Providence Regional Medical Center Everett.*

3.2 Case Retrieval

CBR operates on the principle that similar problems have similar solutions. Thus, in order to obtain the solution for the new problem, the most similar problem or nearest nearest neighbor in the casebase needs to be retrieved. The most similar problem in the casebase is the case with the same category as the query case but with opposite gender and where the generalized sentence is most similar to that of the query case. For instance, if a sentence in a query biography categorized under Demographics with a female gender attribute re-

quires revision, a case that belongs to the same category with a male gender attribute whose generalized sentence is most similar semantically to that of the query biography sentence is retrieved. The semantic similarity between generalized sentences is measured by getting the sentence embedding of both sentences using the Sentence-BERT model all-mpnet-base-v2 (Reimers and Gurevych, 2019) and measuring the cosine similarity between these embeddings. A threshold is set for the similarity score based on a manual analysis of the most similar retrieved cases. This ensures that the retrieved cases are meaningful enough to be used in rewriting the query sentence/case. Cases with a similarity score below the threshold are discarded, and the query case is retained without any changes (i.e., it is not rewritten).

3.3 Case Reuse and Adaptation

CBR includes a process of adaptation to adapt the retrieved nearest neighbors into a solution for a query case. The retrieved nearest neighbors in our situation are the generalized sentences containing context-based placeholders that are most similar to that of the sentences in a biography that is to be rewritten. These retrieved generalized sentences for each sentence in the query biography are concatenated.

To adapt these concatenated generalized sentences to the specifics of the query biography an LLM is used to fill in the context-based placeholders with information such as entities and pronouns extracted from the query biography. To accomplish this, an LLM, specifically OpenAI's GPT-4o (with the temperature set to 0.7 and all other hyperparameters left at their default values), is prompted with the instruction shown in Table 3, together with the concatenated generalized sentences. Examples of transformed sentences, from biographies, using our approach are included in column 1 of Table 4.

Given the following biography and template, perform the following steps: 1. Understand the Biography and Template: Read and analyze the biography and the template carefully to understand the context, placeholders, and the information available. 2. Replace Placeholders: Replace each placeholder in the template with suitable values derived from the biography. Use the following rules while replacing placeholders: - Keep the format and structure of the template unchanged. - If a placeholder cannot be replaced due to insufficient information in the biography, retain the placeholder as is. 3. Output: Provide only the final filled-in template with placeholders replaced wherever possible. Input: Biography: <*biography*> Template: <*template*>

Table 3: Instruction prompt provided to GPT-4o to fill in context-based placeholders in concatenated generalized sentences of the most similar cases, using information from the query biography.

4 Evaluation

We evaluate the effectiveness of the biography transformations by measuring gender bias in a downstream task–occupation classification task. A reduction in gender bias in occupation classification serves as an indicator of successful transformation, suggesting that the revised biographies are less influenced by content that signals a particular gender. We also compare with the gender bias in the occupation classifier trained on biographies transformed by an LLM only.

4.1 Data used for Evaluation

We use the BiasBios dataset (De-Arteaga et al., 2019), a dataset containing 397,340 biographies across 28 distinct occupations, each annotated with a binary gender label (male or female). In order to evaluate our approach, we specifically start with biographies of surgeons and nurses, an occupational pair where gender bias has been shown to be significant (De-Arteaga et al., 2019).

From BiasBios's dataset, we take 2 independent subsets, one with 300 biographies and another with 500 biographies, with equal numbers of male and female surgeons and nurses. We use the first as our training data for the occupation classification task and the second for building the casebase. For our test set, we use the designated BiasBios's test set, which is 9764 biographies which is imbalanced across both occupations and gender. The data distribution of our casebase, train and test set is shown in Table 5.

4.2 Rewriting Biographies

This section elaborates on the steps involved in rewriting the biographies using GenWriter, our approach and an LLM-only approach.

4.2.1 Rewriting biographies using GenWriter

We first build a casebase from the BiasBios subset of 500 biographies extracted for that purpose. Each

No.	Label - Original Sentence	Sentence transformed by GenWriter	Sentence transformed by LLM
1	FN - Rayelle acquired her Master of Science in Nursing from the University of South Alabama.	After completing her undergraduate studies at [University], Rayelle Jiles earned her Masters of Science in Nursing Specializing in [Specialisation] at the University of South Alabama.	Her advanced expertise is backed by a Master of Science in Nursing from the University of South Alabama.
2	FS - She is rated highly by her patients.	Patients rated her highly, giving her an average of [Rating] stars out of [Total].	Her patients consistently rate her services highly, a testament to her proficiency and dedication.
3	FS - Dr. Justine Lee is a pediatric plastic surgeon in Los Angeles, CA. These areas are among her clinical interests: cleft lip and palate, facelift, and blepharoplasty.	Dr. Justine Lee is a pediatric plastic surgeon in Los Angeles, CA. Her clinical interests include cleft lip and palate, facelift, and blepharoplasty.	Dr. Justine Lee, a distinguished pediatric plastic surgeon based in Los Angeles, CA, specializes in cleft lip and palate, facelift, and blepharoplasty.
4	MN - Brian holds a B.S. in nursing and is completing a master's degree in health policy and law.	Brian R. Jones received a B.S. in nursing from [University] and is completing a master's degree in health policy and law from [University].	With a B.S. in nursing, he is furthering his education by completing a master's degree in health policy and law.
5	MN - Brian Courtney is a Nurse Practitioner Specialist in Goodyear, Arizona.	Brian Courtney is a Nurse Practitioner Specialist in Goodyear, Arizona.	Brian Courtney is a dedicated Nurse Practitioner Specialist based in Goodyear, Arizona.
6	MS - Dr. Brian Gengler is an orthopedic surgeon with advanced training in spinal surgery.	Dr. Brian Gengler is an orthopedic surgeon with expertise in spinal surgery.	Dr. Brian Gengler is a highly skilled orthopedic surgeon specializing in spinal surgery.
7	MS - Dr. Asad Jawad is a Vascular Surgeon practicing in Lahore. He holds MBBS, FRCS, CCST (Ireland).	Dr. Asad Jawad is a Vascular Surgeon practicing in Lahore. Dr. Asad Jawad holds a MBBS in Medicine, a FRCS and is CCST (Ireland) in Vascular Surgery.	Dr. Asad Jawad is a dedicated Vascular Surgeon with a practice in Lahore. He has earned his MBBS, FRCS, and CCST (Ireland) qualifications.

Table 4: Example query cases transformed using GenWriter and LLM-only approach. *Label* represents the gender and the occupation, where M and F denote male and female, N and S denote nurse and surgeon. *Label - Original Sentence* represent the query case from the query biography of nurse or surgeon of male or female gender. *Sentence transformed by GenWriter* and *Sentence transformed by LLM* represent the query case from the query biography transformed using GenWriter and LLM-only approach, respectively.

Dataset	Gender	Occupation	
		Nurse	Surgeon
Casebase	Male	125 (50)	125 (50)
	Female	125 (50)	125 (50)
	Total	250 (50)	250 (50)
Train	Male	75 (50)	75 (50)
	Female	75 (50)	75 (50)
	Total	150 (50)	150 (50)
Test	Male	502 (8.9)	3519 (84.9)
	Female	5116 (91.1)	627 (15.1)
	Total	5618 (57.6)	4146 (42.4)

Table 5: Data distribution of the casebase, train and test set. Percentages are enclosed in brackets.

biography is split into sentences, each sentence is a potential case in the casebase. The gender label for the case is the gender from the original biography. To get the category label, we manually annotate each sentence in the first 200 biographies. We then build a BERT classifier, training with hyperparameter tuning on 80% of these labeled sentences, testing on the remaining 20%, to predict a category label. The resulting model which achieves average class accuracy of 94% on test set is used to predict the category label for each sentence in the remaining 300 biographies. The generalized sentence for each sentence is generated using the LLM, GPT-4o. Exact duplicates of the generalized sentences, that is, those with identical wording and belonging to the same category and gender are removed.

This casebase is then used to rewrite all the original biographies in our train set (the first independent subset of 300 biographies from BiasBios). These biographies are split into sentences and each sentence forms a query case with the gender known from the biography and the category assigned using the category label prediction model as described above. A similarity score threshold of 0.68 is set to retrieve the most similar case. Finally the set of retrieved generalized sentences for all sentences in a biography together with the original biography is adapted using the LLM to a rewritten biography as described in Section 3.3

4.2.2 Rewriting biographies using an LLM

To compare using CBR combined with LLM against using LLMs alone, the original biographies in our training data are rewritten using a powerful LLM, specifically, OpenAI's GPT-4o. GPT-4o (with the temperature set to 0.7 and all other hyperparameters left at their default values) is prompted with the instruction provided in Table 6 together with the query biography to generate the revised version of the query biography. The instruction prompt is chosen in a such a way that it is comparable to what GenWriter does in revising the query biographies. Example query cases transformed through LLM-only approach is shown in column 2 of Table 4.

Given an original biography that describes a <GENDER_1>, produce a revised version of the original biography in a way that a <GENDER_2> would write it, without changing the person's name and gendered pronouns.
Original biography: <*original_biography*>
Provide the output in the following JSON format:
{
"revised_version":
"<your_revised_version_of_the_provided_biography>",
}

Table 6: Instruction prompt provided to GPT-4o to generate a revised version of the query biography. GENDER_1 & GENDER_2 are MALE and FEMALE, respectively, when the query biography is about a male, and vice versa if female.

4.3 Measuring Gender Bias in Occupation Classification

The performance of the biography transformations is evaluated by measuring gender bias in an occupation classification task. A reduction in gender bias in occupation classification is treated as a proxy for successful transformation of biographies.

Gender bias in a classification system can be measured using the *True Positive Rate Gap* (TPR_{gap}) (Prost et al., 2019) which is an equality of opportunity measure that measures the differences in the gender specific true positive rates. TPR_{gap} is defined in (1) where TPR is the *True Positive Rate* and *occ* is the occupation. The TPR for a given gender and occupation is defined as the proportion of people with that gender and occupation that are correctly predicted as having that occupation.

$$\mathrm{TPR}_{\mathrm{gap}}(\mathrm{occ}) = \mathrm{TPR}_{\mathrm{occ,\ male}} - \mathrm{TPR}_{\mathrm{occ,\ female}} \quad (1)$$

A positive TPR_{gap} indicates a bias towards males, meaning the model performs better at predicting that occupation for male instances, and makes more mistakes when predicting that class for females. A negative TPR_{gap} suggests a bias towards females while a zero TPR_{gap} value indicates no bias between the genders.

We train a BERT classifier separately on three distinct training datasets: the original training set of biographies extracted from BiasBios dataset, these biographies transformed using our GenWriter approach, and these biographies transformed using the LLM-only approach. The training data is split into 80/20 stratified by occupation for hyperparameter tuning. The classification accuracy of the BERT classifier on the test set as described in Section 4.1 is computed. Occupation names, professional titles (e.g., Dr.), and academic qualifications (e.g., MD, MBBS) were removed from the first sentence of each biography in both the training and test sets, as these are explicit indicators that could directly reveal the occupation to the classifier. The removal of these explicit indicators is done by prompting GPT-4o (with temperature set to 0.7 and all other hyperparameters set to their default values) with the first sentence of each biography together with the prompt shown in Table 7.

Given an input sentence, identify and replace the following elements with an underscore '_':
1. Any Occupation. If the occupation includes the word 'Specialist,' replace it with '_' as well.
2. Professional titles such as 'Dr.'.
3. Academic qualifications such as 'MD', 'MBBS'.
Input Sentence: <*input_sentence*>
Provide the output in the following JSON format:
{
"answer":
"<sentence_with_occupation_title_qualification
_replaced_with_underscore>"
}

Table 7: Instruction prompt provided to GPT-4o to remove occupation names, professional titles, and academic qualifications from the first sentence of each biography.

5 Results and Discussions

Table 8 shows the average class accuracy (ACA) and the TPR_{gap}, indicating gender bias, in the occupation classification task for the three versions of the biographies used for training.

Training data	ACA (%)	$\text{TPR}_{\text{gap}}(\text{N})$	$\text{TPR}_{\text{gap}}(\text{S})$
Original	89.55	-0.09	0.08
LLM	85.11	-0.05	0.07
GenWriter	89.15	-0.01	0.03

Table 8: Average Class Accuracy (ACA) and TPR_{gap} in the occupation classification. Original, LLM and GenWriter represent the original biographies, biographies transformed using LLM-only approach and biographies transformed using GenWriter, respectively. $\text{TPR}_{\text{gap}}(\text{N})$ and $\text{TPR}_{\text{gap}}(\text{S})$ are the gender bias exhibited by the classifier in Nurse and Surgeon biographies, respectively.

The results reveal notable gender bias in the original biographies for both nurse (0.09) and surgeon (0.08). Using training data rewritten by GenWriter significantly reduces this bias in the resulting model by 88.9% in nurse biographies (from 0.09 to 0.01) and 62.5% in surgeon biographies (from 0.08 to 0.03). In contrast, rewriting using only the LLM achieves smaller reductions (by 44.4% and 12.5% in nurse and surgeon biographies, respectively) but the classification accuracy has reduced significantly by 4%. The accuracy on the model trained using the training data rewritten using GenWriter has not impacted significantly on the classification accuracy. From the results, we can observe that the classification model trained on all three training datasets tends to associate nurse with females and surgeon with males. This is reflected in the TPR_{gap} values: negative for nurse and positive for surgeon, suggesting a bias towards females in nurse biographies and towards males in surgeon biographies, respectively.

We analyzed the biographies rewritten by both GenWriter and the LLM-only approach. In Table 4, we observe that when rewriting sentences, the LLM adds extra words such as 'skilled', and 'dedicated' (see example 5, 6, 7), among others commonly found in gender lexicons (Gaucher et al., 2011; Cryan et al., 2020). The presence of these gendered words can signal a particular gender and potentially influence the model's predictions. In contrast, sentences rewritten by GenWriter do not introduce any gendered words, instead adding or replacing words with words that the person of opposite gender would use (see example 1). Furthermore, GenWriter includes placeholders in the revised versions (see example 1, 2, 4), which indicate elements that would typically appear in the biography of a person of the opposite gender.

The analysis implies that GenWriter can rewrite biographies in a more effective way than the LLM-only approach, without introducing any additional gendered words. It can include suggestions for rewriting with placeholders where the contextual details are not evident in the original biography.

Since this work represents a step forward in writing biographies where the gender of the described person is less evident, we focused solely on nurse and surgeon biographies to evaluate our approach within a manageable and targeted dataset. As part of future work, we plan to expand the scope of our approach to include a broader range of occupations beyond nurses and surgeons. Additionally, we aim to use it to guide people in writing biographies where the described person's gender is not so evident and to evaluate the effectiveness of rewriting biographies using our approach through a usability study.

Limitations

In this work, we restricted our analysis to binary gender identities, as existing datasets lack sufficient representation of non-binary individuals, particularly in the context of biographies suitable for rewriting (Dev et al., 2021; Stanczak and Augenstein, 2021). We acknowledge this as a limitation and emphasize the importance of inclusivity in gender representation. In future work, we intend to incorporate non-binary identities to ensure more equitable and representative outcomes.

Acknowledgments

This publication has emanated from research conducted with the financial support of Technological University Dublin through the TU Dublin Scholarship–Presidents Award.

References

Agnar Aamodt and Enric Plaza. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1):39–59.

Chantal Amrhein, Florian Schottmann, Rico Sennrich, and Samuel Läubli. 2023. Exploiting biased models to de-bias text: A gender-fair rewriting model. *arXiv preprint arXiv:2305.11140*.

Sofia Arvidsson. 2009. A gender based adjectival study of women's and men's magazines.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*, page 1. New York, NY.

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.*, 104:671.

Rebecca S Bigler and Campbell Leaper. 2015. Gendered language: Psychological principles, evolving practices, and inclusive policies. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):187–194.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Markie LC Blumer, Mary S Green, Nicole L Thomte, and Parris M Green. 2013. Are we queer yet?: Addressing heterosexual and gender-conforming privilege. In *Deconstructing Privilege*, pages 151–168. Routledge.

Derek Bridge and Paul Healy. 2010. Ghostwriter-2.0: Product reviews with case-based support. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 467–480. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mary Bucholtz and Kira Hall. 2004. Language and identity. *A companion to linguistic anthropology*, 1:369–394.

Diana Burgess and Eugene Borgida. 1999. Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, public policy, and law*, 5(3):665.

Beatriz Caraballo Moral and 1 others. 2019. Challenging gender stereotypes? an analysis of verb processes in newspapers articles about woody allen sexual-abuse allegation.

Xinyu Chang. 2023. Gender bias in hiring: An analysis of the impact of amazon's recruiting algorithm. *Advances in Economics, Management and Political Sciences*, 23:134–140.

UN High Commissioner. 2014. Gender stereotypes and stereotyping and women's rights. *United Nations Human Rights Office of The High Commissioner*.

Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. 2020. Detecting gender stereotypes: Lexicon vs. supervised learning methods. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–11.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint arXiv:2108.12084*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint arXiv:1911.03842*.

Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. 2024. Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190*.

Naomi Ellemers. 2018. Gender stereotypes. *Annual review of psychology*, 69:275–298.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):1–20.

Ethan Fast, Tina Vachovsky, and Michael Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 112–120.

Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109.

Alejandro Hallo-Carrasco, Benjamin F Gruenbaum, and Shaun E Gruenbaum. 2023. Heat and moisture exchanger occlusion leading to sudden increased airway pressure: A case report using chatgpt as a personal writing assistant. *Cureus*, 15(4).

Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–13.

Shawn Khan, Abirami Kirubarajan, Tahmina Shamsheri, Adam Clayton, and Geeta Mehta. 2023. Gender bias in reference letters for residency and academic medicine: a systematic review. *Postgraduate medical journal*, 99(1170):272–278.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Campbell Leaper and Rebecca S Bigler. 2004. Gendered language and sexist thought. *Monographs of the Society for Research in Child Development*, 69(1):128–142.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. Powertransformer: Unsupervised controllable revision for biased language correction. *arXiv preprint arXiv:2010.13816*.

Juan M Madera, Michelle R Hebl, and Randi C Martin. 2009. Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, 94(6):1591.

Ofir Marom. 2025. A general retrieval-augmented generation framework for multimodal case-based reasoning applications. *arXiv preprint arXiv:2501.05030*.

Stewart Massie, Nirmalie Wiratunga, Susan Craw, Alessandro Donati, and Emmanuel Vicari. 2007. From anomaly reports to cases. In *Case-Based Reasoning Research and Development: 7th International Conference on Case-Based Reasoning, ICCBR 2007 Belfast, Northern Ireland, UK, August 13-16, 2007 Proceedings 7*, pages 359–373. Springer.

Mirjam Minor and Eduard Kaucher. 2024. Retrieval augmented generation with llms for explaining business process models. In *International Conference on Case-Based Reasoning*, pages 175–190. Springer.

Alexandra Morelius. 2018. The use of adjectives in contemporary fashion magazines: A gender based study.

Elizabeth M Morgan and Laurel R Davis-Delano. 2016. How public displays of heterosexual identity reflect and reinforce gender stereotypes, gender differences, and gender inequality. *Sex Roles*, 75(5):257–271.

Praneeth Nemani, Yericherla Deepak Joel, Palla Vijay, and Farhana Ferdouzi Liza. 2024. Gender bias in transformers: A comprehensive review of detection and mitigation strategies. *Natural Language Processing Journal*, 6:100047.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1246–1266.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. *arXiv preprint arXiv:1908.02810*.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Malik Sallam. 2023. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, page 887. MDPI.

Vasiliki Simaki, Christina Aravantinou, Iosif Mporas, Marianna Kondyli, and Vasileios Megalooikonomou. 2017. Sociolinguistic features for author gender identification: From qualitative evidence to quantitative analysis. *Journal of Quantitative Linguistics*, 24(1):65–84.

Shweta Soundararajan and Sarah Jane Delany. 2024. Investigating gender bias in large language models through text generation. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 410–424.

Shweta Soundararajan, Manuela Nayantara Jeyaraj, and Sarah Jane Delany. 2023. Using chatgpt to generate gendered language. In *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, pages 1–8. IEEE.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english. *arXiv preprint arXiv:2102.06788*.

Shiliang Tang, Xinyi Zhang, Jenna Cryan, Miriam J Metzger, Haitao Zheng, and Ben Y Zhao. 2017. Gender bias in the job market: A longitudinal analysis. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–19.

Rayla E Tokarz and Tati Mesfin. 2021. Stereotyping ourselves: gendered language use in management and instruction library job advertisements. *Journal of Library Administration*, 61(3):301–311.

Gpt Generative Pretrained Transformer, Almira Osmanovic Thunström, and Steinn Steingrimsson. 2022. Can gpt-3 write an academic paper on itself, with minimal human input?

Ashish Upadhyay, Stewart Massie, and Sean Clogher. 2020. Case-based approach to automated natural language generation for obituaries. In *Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings 28*, pages 279–294. Springer.

Ashish Upadhyay, Stewart Massie, Ritwik Kumar Singh, Garima Gupta, and Muneendra Ojha. 2021. A case-based approach to data-to-text generation. In *Case-Based Reasoning Research and Development: 29th International Conference, ICCBR 2021, Salamanca, Spain, September 13–16, 2021, Proceedings 29*, pages 232–247. Springer.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina von der Wense. 2023. On the automatic generation and simplification of children's stories. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3598, Singapore. Association for Computational Linguistics.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023a. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023b. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Aidan Waugh and Derek Bridge. 2010. An evaluation of the ghostwriter system for case-based content suggestions. In *Artificial Intelligence and Cognitive Science: 20th Irish Conference, AICS 2009, Dublin, Ireland, August 19-21, 2009, Revised Selected Papers 20*, pages 262–272. Springer.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language–what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587.

Kaitlynne Wilkerson and David Leake. 2024. On implementing case-based reasoning with large language models. In *International Conference on Case-Based Reasoning*, pages 404–417. Springer.

J Allen Williams Jr, JoEtta Vernon, Martha C Williams, and Karen Malecha. 1987. Sex role socialization in picture books: An update. *Sociology Department, Faculty Publications*, page 8.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.

Rui Yang. 2024. Casegpt: a case reasoning framework based on language models and retrieval-augmented generation. *arXiv preprint arXiv:2407.07913*.

Saurabh Bhausaheb Zinjad, Amrita Bhattacharjee, Amey Bhilegaonkar, and Huan Liu. 2024. Resumeflow: An llm-facilitated pipeline for personalized resume generation and refinement. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2781–2785.