Reducing Gender Cues in Biographies through Text Rewriting using GenWriter

Shweta Soundararajan[1,*]

[1]*Technological University Dublin, Dublin, Ireland*

Abstract
Gendered language is defined as the words/phrases that signals a particular gender. This can be explicit (e.g., "mother," "she") or implicit, where roles or traits (e.g., "gentle," "ambitious") suggest an individual's gender. Although, useful in certain contexts, it can reinforce harmful stereotypes and contribute to societal bias. While significant research has explored mitigating gender bias in natural language processing (NLP) models by scrubbing or swapping gendered terms in text, these efforts have primarily focused on addressing explicit gendered language. However, they often overlook implicit gender bias embedded in language use. Therefore, I aim to focus on mitigating implicit gendered language in texts and propose a novel approach – GenWriter, a Case-Based Reasoning and Large Language Model (CBR-LLM) Fusion Approach, to generate revised content that obscures gender while preserving semantic content. The method involves constructing a case base of generalized sentence representations categorized by gender and content type, retrieving semantically similar cases, and adapting the solution through LLM to produce revised versions where the gender is not so evident. I evaluate the performance of my method by measuring gender bias in an occupation classification task. Results show that GenWriter effectively reduces gender bias, outperforming both the original and LLM-only baselines, while maintaining classification accuracy.

Keywords
Case Based reasoning, Large Language Models, Gender Bias, Gender Stereotypes, Gendered Language

1. Introduction

Gendered language refers to the use of language that indicates the gender of a person, animal, or object, either explicitly (e.g., mother, she) or implicitly (e.g., societal expectations of gendered traits) [1, 2]. While it can be useful, it can also perpetuate harmful gender stereotypes [3, 4]. Gender stereotypes are generalized views about the roles or traits men and women should have, leading to gender bias [5].

Gender bias in text can result in unfair treatment based on the perceived gender of the author, as seen in Amazon's abandoned AI recruitment model, which favored male applicants [6]. Similarly, gendered language in biographies led to women being misclassified in job applications [7]. The language used to describe individuals can reinforce gender stereotypes, can lead to unconscious bias, discrimination, and psychological harm [8]. Gendered language in job ads, for example, can deter women from applying for positions [9], while gender stereotypes in children's stories can influence young minds [10, 11]. Therefore, it is important to help or facilitate people to use content where the gender of the person is not evident, as this can reduce the harm caused to individuals.

Prior work [7, 12, 13] has focused on mitigating gender bias in text. These approaches typically remove, replace, or swap gendered terms—particularly explicit gendered language. While the resulting debiased texts are useful for training NLP models to promote gender fairness in downstream tasks, they are not suitable for use as human-facing suggestions. This is because explicit gendered terms such as pronouns and gendered references are often essential in real-world contexts such as biographies, articles, and resumes.

To this end, my aim 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, as an alternative to text

*Corresponding author.

✉ shweta.x.soundararajan@mytudublin.ie (S. Soundararajan)

🆔 0009-0005-2171-5501 (S. Soundararajan)

including content that implies gender identity. The approach involves rewriting text about a person as if it were written by someone of a different gender. Unlike prior work, my approach focuses on mitigating implicit gendered language rather than only removing or replacing explicit gender terms.

1.1. Background

This section reviews gendered language, prior research on detecting and mitigating gender bias in text and models, respectively, various text generation approaches and studies on bias in state-of-the-art text generation.

1.1.1. Gendered Language

Gendered language can be categorized into linguistic and social aspects [14, 15, 16]. Linguistic gender includes grammatical, referential, and lexical gender. Grammatical gender involves noun classification based on sentence agreement, common in many languages but not English. Referential gender relates to the gender of individuals or objects (e.g., pronouns and titles), while lexical gender assigns gender based on meaning (e.g., "man" vs. "woman"). Social gender covers cultural aspects like gender identity, expression, and roles [14], where gender identity refers to one's internal sense of gender [17], gender expression is how it is presented [18], and gender roles are societal expectations [19]. Gendered language can reinforce or challenge social norms and stereotypes [20], with lexical and referential gender being explicit, while social gender is more implicit.

1.1.2. Detecting and Mitigating Gender Bias

Existing literature on identifying and mitigating gender bias [21, 22, 23] mainly focuses on linguistic gender, such as pronouns and gendered terms. Some studies [24, 25, 9] have created gender lexicons listing words that distinguish men and women. Research [26, 27] indicates that women are increasingly identifying less with traditionally feminine traits, highlighting the need to update societal gender norms related to stereotypical characteristics of masculinity and femininity. Responding to this, Cryan et al. proposed a new gender lexicon created by scraping Wikipedia for lists of candidate words and using crowdsourcing for annotation. Datasets for stereotype detection [29, 30, 28] have been created using various sources or crowdsourcing. Efforts to mitigate gender bias in NLP models include techniques like data scrubbing [7], gender-swapping [12], and counterfactual data augmentation (CDA) [13], Counterfactual data substitution (CDS) [31] that remove, replace, or swap explicit gendered terms. Other approaches [21, 32, 33] focus on debiasing word embeddings rather than the text itself.

1.1.3. Text Generation Approaches

Text generation involves automatically creating coherent and meaningful text, ranging from sentences to full documents. Approaches include traditional rule-based or data-driven methods, statistical techniques, and modern neural-based approaches [34, 35, 36]. Traditional systems used predefined rules or language patterns from datasets, while statistical models like N-grams [37] and CRFs [38] modeled word relationships. Neural networks specifically, transformer-based models like GPT [39] and BERT [40] excel at generating diverse, coherent text. Recently, transformer-based Large Language Models (LLMs) have revolutionized text generation across various fields, including medical report generation [41], academic writing [42], and children's education [43]. However, despite their capabilities, LLMs raise ethical concerns, particularly around gender bias in generated text, which recent studies [44, 45, 46, 47, 48] show can perpetuate societal harm.

Case-based reasoning (CBR) is a problem-solving method that supports text generation by reusing solutions from similar past cases [49]. It involves four steps: (1) retrieving relevant cases, (2) reusing and adapting solutions, (3) revising as needed, and (4) retaining useful outcomes. CBR has been applied to tasks such as anomaly reporting [50], obituary writing [51], sports summaries [52], product reviews [53], and product descriptions [54]. However, adapting prior solutions in natural language is challenging.

Integrating CBR with LLMs helps address this by (1) reducing hallucinations, bias, and stereotypes, and (2) enabling application in knowledge-rich domains without formal encodings [55].

While prior work has focused on detecting and mitigating gender bias in texts and NLP models, respectively, no studies have addressed rewriting text to make the person's gender less evident. This presents a challenge in the research, as there is a scarcity of instances free from content with gender cues that could be used to offer alternatives or suggestions for rewriting content implying gender identity. My proposed research question, discussed in Section 2.1, aims to address this issue.

2. Research Plan

This section outlines the research objectives and describes the approach taken to achieve them.

2.1. Research Objectives

Our primary goal is to generate textual content where an individual's gender is not so evident, which can be used as an alternative to content signaling a particular gender. While searching for datasets with no gender cues, I found them to be scarce. To address this challenge, I formulated the following research question, focusing on exploring a viable approach for generating text that obscures gender identity of an individual described in the text. Such an approach could be used to revise content that implies a specific gender across various applications, including biographies, job advertisements, resumes, and more.

- How can I effectively transform texts that signal a specific gender into versions where the individual's gender is less apparent?

2.2. Approach

The goal is to rewrite text that implies the gender identity of the person, ensuring the gender of the person is not so evident in the text. The approach used is to rewrite text content about a person as if it was written by a person of a different gender. To achieve this, I use **GenWriter**, a CBR-LLM Fusion Approach, which combines Case-Based Reasoning (CBR) and Large Language Models (LLMs). This approach creates a case base that serves as a repository of past experiences. When a new problem arises, such as transforming a text to one where the gender of the person is not so evident, the solutions from similar cases are used. The LLM plays a key role in both constructing cases and adapting existing solutions, effectively integrating CBR with LLM capabilities. Section 2.2.1, Section 2.2.2 and Section 2.2.3 will describe how cases in the case base are represented, how solutions to new problems from the case base are retrieved, and how the retrieved solutions are adapted to ensure their suitability for new problems.

2.2.1. Case Representation

A case base typically covers a specific application area, such as biographies. Each case in the base represents a sentence describing an aspect of a person. For example, biographies usually begin with basic details like name, birthplace, age, and occupation, followed by education and work experience, and concluding with personal aspects such as family, hobbies, and interests. In total, a biography covers four main components: Demographics, Education, Work details, and Non-Professional details. A case base for biographies contains cases, each representing a sentence from a biography related to one of these components. The case representation will include the following:

- **Gender**, indicating the gender of the person being discussed in the biography.
- **Category**, specifying which aspect of the person is being discussed in the biography. The four components of the biography - Demographics, Education, Work details, and Non-Professional details are the *Category*.

- **Generalized Sentence**, a sentence from the biography related to the *Category*, with pronouns and entities used in a biography, 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.

Generalized sentences for both the query case and cases in the case base are generated using few-shot prompting [56] with OpenAI's GPT-4o (temperature set to 0.7; other hyperparameters at default). The LLM receives a few-shot prompt (see Table 2 in Appendix A) along with the target sentence to produce its generalized form. Examples of generated cases and their representations are shown in Table 3 in Appendix B.

2.2.2. Case Retrieval

CBR is based on the principle that similar problems have similar solutions. To solve a new problem, the most similar case in the case base is retrieved. This is done by measuring the semantic similarity between the sentence embeddings of the new problem and those of cases with the opposite gender attribute but the same category. For example, if a Demographics sentence with a female gender attribute needs revision, the most similar male case in the same category is retrieved. Semantic similarity is measured using cosine similarity between embeddings generated by the Sentence-BERT [57].

2.2.3. Case Reuse and Solution Adaptation

Once the most similar case is retrieved for each sentence in the new problem, its generalized sentence with context-based placeholders, is reused. These are then concatenated. To adapt these generalized sentences to the new problem, OpenAI's GPT-4 (with a temperature of 0.7; other hyperparameters at default) fills in the placeholders with information like entities and pronouns from the new problem. The LLM is prompted with the instruction shown in Table 4 in Appendix C and the set of generalized sentences. Example sentences transformed using GenWriter are shown in Table 5 in Appendix D.

3. Progress Summary

Most progress to date has focused on implementing and evaluating GenWriter. The goal of evaluating GenWriter is to assess its effectiveness in transforming texts that imply gender identity into those that makes the described person's gender less evident. I focus on biographies for this application and evaluate the transformations by measuring gender bias in the occupation classification task. A reduction in gender bias indicates successful transformation, suggesting the revised biographies are less influenced by gender-specific cues. I compare the gender bias in the occupation classifier trained on biographies transformed using GenWriter, a CBR-LLM Fusion approach, with two baselines: the original BiasBios biographies (without transformation) and the LLM-only approach, where only LLM is used for revision.

3.1. Building a Casebase

In my approach, I build a case base by extracting 500 biographies from the training set of BiasBios dataset (more details in Appendix E), with an equal number of male and female surgeons and nurses. Each biography, containing Demographics, Education, Work details, and Professional details, is split into sentences, which form the category label of a case. To gather the necessary attributes, I assign gender labels to each sentence based on the BiasBios dataset. For category labels, I manually annotate the first 200 biographies and train a BERT classifier on these labeled sentences to predict category labels for the remaining 300. The trained BERT classifier, achieving 94% average class accuracy on the test set,

is then used to predict category labels for the remaining biographies. The generalized sentence for each sentence in the biography is also generated.

3.2. Rewriting Biographies and Measuring Gender Bias in Occupation Classification

I used 300 biographies from the BiasBios training set (independent of my case base) and the full test set (9,764 biographies). All 300 training biographies were rewritten using my approach. Each was split into sentences, labeled for gender (from BiasBios) and category (via a pretrained BERT classifier). I retrieved the most semantically similar case for each sentence using cosine similarity between generalized sentences, with a threshold of 0.68; sentences below this threshold were left unchanged. Matched cases and the original biography were then combined with a prompt and passed to GPT-4o to generate the revised version. For comparison, I also applied an LLM-only approach, prompting GPT-4o to revise the original biography (refer Table 7 in Appendix G for the prompt). Example sentences transformed using LLM-only approach are shown in Table 5 in Appendix D.

I evaluate biography transformation performance by measuring gender bias in occupation classification. A reduction in gender bias indicates successful transformation. I train a BERT classifier on three datasets: original biographies, biographies transformed via GenWriter, and those using the LLM-only approach. To avoid direct occupation clues, occupation names, professional titles, and academic qualifications are removed from the first sentence of each biography. Gender bias is quantified using the True Positive Rate Gap (TPR_{gap}) [58] (see eqn. 1), which compares the gender-specific true positive rates (TPR) for each occupation. A positive TPR_{gap} indicates bias toward males, while a negative TPR_{gap} suggests bias toward females. A TPR_{gap} of zero indicates no bias. I also compute average class accuracy, accounting for the imbalanced distribution of occupations in the test set.

$$\text{TPR}_{\text{gap}}(\text{occupation}) = \text{TPR}_{\text{occupation, male}} - \text{TPR}_{\text{occupation, female}} \tag{1}$$

4. Conclusion and Future Work

Table 1 shows the average class accuracy and the TPR_{gap} in the occupation classification. From the results, we can observe that the classification system, 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. The results also reveal notable gender bias in the original biographies for both nurse (0.09) and surgeon (0.08). GenWriter, a CBR-LLM Fusion approach, significantly reduces this bias by 88.9% in nurse biographies (from 0.09 to 0.01) and 62.5% in surgeon biographies (from 0.08 to 0.03), while preserving classification accuracy. In contrast, the LLM-only method achieves smaller reductions (by 44.4% and 12.5%, respectively) but compromises accuracy. These results demonstrate the strength of GenWriter, not only does it effectively mitigate gender bias in biographies, but it also does so without sacrificing classification performance, outperforming baseline methods on both fairness and accuracy.

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

Table 1
Average Class Accuracy and TPR_{gap} in the occupation classification. TPR$_{\text{gap}}$(N) and TPR$_{\text{gap}}$(S) are the gender bias found in Nurse and Surgeon biographies, respectively.

I intend to extend my approach to include additional occupations beyond nurse and surgeon and other areas beyond the current scope, such as job advertisements. The job ads will be extracted from job postings on various online platforms using web scrappers.

Acknowledgments

Declaration on Generative AI

The authors have not employed any Generative AI tools for tasks listed in the GenAI Usage Taxonomy. However, GPT-4o was used as part of the experiments presented in this work.

References

[1] F. Hamidi, M. K. Scheuerman, S. M. Branham, 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, 2018, pp. 1–13.

[2] R. S. Bigler, C. Leaper, Gendered language: Psychological principles, evolving practices, and inclusive policies, Policy Insights from the Behavioral and Brain Sciences 2 (2015) 187–194.

[3] M. Bucholtz, K. Hall, Language and identity, A companion to linguistic anthropology 1 (2004) 369–394.

[4] C. Leaper, R. S. Bigler, Gendered language and sexist thought, Monographs of the Society for Research in Child Development 69 (2004) 128–142.

[5] U. OHCHR, Gender stereotypes and stereotyping and women's rights, 2014.

[6] V. Simaki, C. Aravantinou, I. Mporas, M. Kondyli, V. Megalooikonomou, Sociolinguistic features for author gender identification: From qualitative evidence to quantitative analysis, Journal of Quantitative Linguistics 24 (2017) 65–84.

[7] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. T. Kalai, 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, 2019, pp. 120–128.

[8] S. Barocas, A. D. Selbst, Big data's disparate impact, Calif. L. Rev. 104 (2016) 671.

[9] D. Gaucher, J. Friesen, A. C. Kay, Evidence that gendered wording in job advertisements exists and sustains gender inequality., Journal of personality and social psychology 101 (2011) 109.

[10] A. E. Arthur, R. S. Bigler, L. S. Liben, S. A. Gelman, D. N. Ruble, Gender stereotyping and prejudice in young children, Intergroup attitudes and relations in childhood through adulthood (2008) 66–86.

[11] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.

[12] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender bias in coreference resolution: Evaluation and debiasing methods, arXiv preprint arXiv:1804.06876 (2018).

[13] K. Lu, P. Mardziel, F. Wu, P. Amancharla, A. Datta, Gender bias in neural natural language processing, Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday (2020) 189–202.

[14] L. Ackerman, Syntactic and cognitive issues in investigating gendered coreference, Glossa: a journal of general linguistics 4 (2019).

[15] Y. T. Cao, H. Daumé III, Toward gender-inclusive coreference resolution, arXiv preprint arXiv:1910.13913 (2019).

[16] M. Bartl, S. Leavy, Inferring gender: A scalable methodology for gender detection with online lexical databases, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, 2022, pp. 47–58.

[17] L. Litosseliti, J. Sunderland, Gender identity and discourse analysis, volume 2, John Benjamins Publishing, 2002.

[18] D. L. Rubin, K. L. Greene, Effects of biological and psychological gender, age cohort, and interviewer gender on attitudes toward gender-inclusive/exclusive language, Sex Roles 24 (1991) 391–412.

[19] U. Gabriel, P. Gygax, O. Sarrasin, A. Garnham, J. Oakhill, Au pairs are rarely male: Norms on the gender perception of role names across english, french, and german, Behavior research methods 40 (2008) 206–212.

[20] M. Hellinger, H. Bußmann, Gender across languages: The linguistic representation of women and men, in: Gender across languages, John Benjamins, 2015, pp. 1–25.

[21] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Advances in neural information processing systems 29 (2016).

[22] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Men also like shopping: Reducing gender bias amplification using corpus-level constraints, arXiv preprint arXiv:1707.09457 (2017).

[23] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, K.-W. Chang, Gender bias in contextualized word embeddings, arXiv preprint arXiv:1904.03310 (2019).

[24] S. L. Bem, The measurement of psychological androgyny., Journal of consulting and clinical psychology 42 (1974) 155.

[25] J. T. Spence, R. Helmreich, J. Stapp, Personal attributes questionnaire, Developmental Psychology (1974).

[26] J. M. Twenge, Changes in masculine and feminine traits over time: A meta-analysis, Sex roles 36 (1997) 305–325.

[27] K. Donnelly, J. M. Twenge, Masculine and feminine traits on the bem sex-role inventory, 1993–2012: A cross-temporal meta-analysis, Sex roles 76 (2017) 556–565.

[28] J. Cryan, S. Tang, X. Zhang, M. Metzger, H. Zheng, B. Y. Zhao, Detecting gender stereotypes: Lexicon vs. supervised learning methods, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–11.

[29] M. Nadeem, A. Bethke, S. Reddy, Stereoset: Measuring stereotypical bias in pretrained language models, arXiv preprint arXiv:2004.09456 (2020).

[30] N. Nangia, C. Vania, R. Bhalerao, S. R. Bowman, Crows-pairs: A challenge dataset for measuring social biases in masked language models, arXiv preprint arXiv:2010.00133 (2020).

[31] R. H. Maudslay, H. Gonen, R. Cotterell, S. Teufel, It's all in the name: Mitigating gender bias with name-based counterfactual data substitution, arXiv preprint arXiv:1909.00871 (2019).

[32] H. Gonen, Y. Goldberg, Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, arXiv preprint arXiv:1903.03862 (2019).

[33] A. Romanov, M. De-Arteaga, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. Rumshisky, A. T. Kalai, What's in a name? reducing bias in bios without access to protected attributes, arXiv preprint arXiv:1904.05233 (2019).

[34] J. Becker, J. P. Wahle, B. Gipp, T. Ruas, Text generation: A systematic literature review of tasks, evaluation, and challenges, arXiv preprint arXiv:2405.15604 (2024).

[35] A. Celikyilmaz, E. Clark, J. Gao, Evaluation of text generation: A survey, arXiv preprint arXiv:2006.14799 (2020).

[36] C. C. Osuji, T. C. Ferreira, B. Davis, A systematic review of data-to-text nlg, arXiv preprint arXiv:2402.08496 (2024).

[37] M. Suzuki, N. Itoh, T. Nagano, G. Kurata, S. Thomas, Improvements to n-gram language model using text generated from neural language model, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 7245–7249.

[38] D. Song, W. Liu, T. Zhou, D. Tao, D. A. Meyer, Efficient robust conditional random fields, IEEE Transactions on Image Processing 24 (2015) 3124–3136.

[39] F. Khalil, G. Pipa, Transforming the generative pretrained transformer into augmented business text writer, Journal of Big Data 9 (2022) 112.

[40] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[41] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, D. Shen, Chatcad: Interactive computer-aided diagnosis

on medical image using large language models, arXiv preprint arXiv:2302.07257 (2023).

[42] M. Sallam, Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns, in: Healthcare, volume 11, MDPI, 2023, p. 887.

[43] M. Valentini, J. Weber, J. Salcido, T. Wright, E. Colunga, K. von der Wense, On the automatic generation and simplification of children's stories, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 3588–3598. URL: https://aclanthology.org/2023.emnlp-main.218. doi:10.18653/v1/2023.emnlp-main.218.

[44] Y. Wan, G. Pu, J. Sun, A. Garimella, K.-W. Chang, N. Peng, "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 3730–3748. URL: https://aclanthology.org/2023.findings-emnlp.243. doi:10.18653/v1/2023.findings-emnlp.243.

[45] H. Kotek, R. Dockum, D. Sun, Gender bias and stereotypes in large language models, in: Proceedings of the ACM collective intelligence conference, 2023, pp. 12–24.

[46] X. Dong, Y. Wang, P. S. Yu, J. Caverlee, Disclosure and mitigation of gender bias in llms, arXiv preprint arXiv:2402.11190 (2024).

[47] X. Fang, S. Che, M. Mao, H. Zhang, M. Zhao, X. Zhao, Bias of ai-generated content: an examination of news produced by large language models, Scientific Reports 14 (2024) 1–20.

[48] A. Ovalle, P. Goyal, J. Dhamala, Z. Jaggers, K.-W. Chang, A. Galstyan, R. Zemel, R. Gupta, "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, 2023, pp. 1246–1266.

[49] A. Yan, Z. Cheng, A review of the development and future challenges of case-based reasoning, Applied Sciences 14 (2024) 7130.

[50] S. Massie, N. Wiratunga, S. Craw, A. Donati, E. Vicari, 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, Springer, 2007, pp. 359–373.

[51] A. Upadhyay, S. Massie, S. Clogher, 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, Springer, 2020, pp. 279–294.

[52] A. Upadhyay, S. Massie, R. K. Singh, G. Gupta, M. Ojha, 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, Springer, 2021, pp. 232–247.

[53] D. Bridge, P. Healy, Ghostwriter-2.0: Product reviews with case-based support, in: International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, 2010, pp. 467–480.

[54] A. Waugh, D. Bridge, 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, Springer, 2010, pp. 262–272.

[55] K. Wilkerson, D. Leake, On implementing case-based reasoning with large language models, in: International Conference on Case-Based Reasoning, Springer, 2024, pp. 404–417.

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

[57] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[58] F. Prost, N. Thain, T. Bolukbasi, Debiasing embeddings for reduced gender bias in text classification, GeBNLP 2019 9573 (2019) 69.

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

Instruction prompt and the few-shot examples provided to GPT-4o to generate generalized sentence is described in Table 2.

> 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.
> 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].
> Your Turn:
> Input Sentence: *<input_sentence>*

Table 2
Instruction prompt and the few-shot examples provided to GPT-4o to generate generalized sentence.

B. Examples of cases

Examples of cases of my case base is shown in Table 3.

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 3
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.*

C. Instruction prompt provided to GPT-4o to fill in context-based placeholders in a generalized sentence

Instruction prompt provided to GPT-4o to fill in context-based placeholders in a generalized sentence is described in 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 4
Instruction prompt provided to GPT-4o to fill in context-based placeholders in a generalized sentence

D. Example Sentences transformed using GenWriter and LLM-only approach

Example sentences transformed using GenWriter are shown in Table 5.

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	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.

Table 5
Example sentences 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.

E. Data used for Evaluation

I use the BiasBios dataset [7], which contains 397,340 biographies across 28 occupations, each annotated with a binary gender label (male or female). For evaluation, I focus on the biographies of surgeons and nurses, with 22,784 surgeon and nurse biographies in the train set and 9,764 in the test set. For my train set, I select a subset of 300 biographies from the BiasBios train set, with an equal number of male and female surgeons and nurses. For my test set, I use the entire BiasBios test set of 9,764 biographies, which is imbalanced across occupations and gender. The data distribution for my train and test sets is shown in Table 6 in Appendix F.

F. Data distribution of my train and test set

Data distribution of my train and test set is shown in Table 6

Occupation	Train			Test		
	#M	#F	#Total	#M	#F	#Total
Surgeon	75	75	150	3519	627	4146
	(50)	(50)	(50)	(84.9)	(15.1)	(42.4)
Nurse	75	75	150	502	5116	5618
	(50)	(50)	(50)	(8.9)	(91.1)	(57.6)

Table 6
Data distribution of my train and test set. #M and #F represent the number of males and females, respectively. Percentages are enclosed in brackets.

G. Instruction prompt provided to GPT-4o to generate a revised version of the original biography

Instruction prompt provided to GPT-4o to generate a revised version of the original biography is described in Table 7.

> 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. After revising the biography, provide a brief two-line explanation specifying what was modified in the revised version and why.
> Original biography: *<original_biography>*
> Provide the output in the following JSON format:
> {
> "revised_version": "<your_revised_version_of_the_provided_biography>",
> "explanation":
> "<your_explanation_for_the_changes_made_in_the _revised_biography>"
> }

Table 7
Instruction prompt provided to GPT-4o to generate a revised version of the original biography. GENDER_1 & GENDER_2 are MAN and WOMAN, respectively, when the original biography is about a man and vice versa otherwise.