Unmasking Metadata Bias: Evaluating the Impact on Large Language Model Text Analysis

JOEL HARMAN, Queensland University of Technology, Australia
ALESSANDRO SORO, Queensland University of Technology, Australia
SELEN TURKAY, Queensland University of Technology, Australia

With the increased popularity and accessibility of LLMs, such as GPT-4o, we are starting to see the emergence of LLM analyses, where LLMs are used to evaluate, categorise, or grade large amounts of content very quickly. While this approach has some clear benefits, as it enables us to manipulate large amounts of data very easily, the practice also has many ethical and fairness concerns which need to be considered. Specifically, a large amount of the content being analysed often contains metadata, such as the name, gender, age, or title of the author. This is potentially problematic as the LLM may use this information to perpetuate or reinforce particular biases.

In this paper, we explore the impact of this metadata and the biases it may create by considering how the inclusion of a name, gender, age, and title in the text may influence the evaluation grade provided by an LLM when comparing identical pieces of text. This analysis found strong evidence that the inclusion of this metadata can lead to biased results. While better prompt engineering and/or redacting this metadata within the text can help to reduce the degree of bias, this analysis also found that it will not remove it entirely. It is therefore imperative that researchers and practitioners looking to use LLMs for large-scale analysis should be aware of this problem and do all they can to mitigate its effects. By considering the impact this bias may have, we hope to strive for fairness, accountability, and transparency to ensure that AI systems benefit all users equitably.

CCS Concepts: • **Human-centered computing** → **Natural language interfaces**; • **Computing methodologies** → **Natural language processing**; **Machine learning**; • **Social and professional topics** → *Computing / technology policy*; *Computing / technology ethics*.

Additional Key Words and Phrases: Text Analysis, Metadata Bias, Large Language Models, Machine Learning, Ethical AI

ACM Reference Format:
Joel Harman, Alessandro Soro, and Selen Turkay. 2024. Unmasking Metadata Bias: Evaluating the Impact on Large Language Model Text Analysis. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (OzCHI '24)*. ACM, New York, NY, USA, 17 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In recent years, large language models (LLMs) have revolutionized natural language processing, enabling a wide range of applications from chatbots to text analysis [1]. These powerful tools have shown remarkable capabilities in understanding and generating human-like text. One major benefit provided by LLMs is their ability to analyse large volumes of unstructured text quickly and inexpensively [7]. This has enabled both researchers and industry

Authors' Contact Information: Joel Harman, ja.harman@qut.edu.au, Queensland University of Technology, Brisbane, Queensland, Australia; Alessandro Soro, alessandro.soro@qut.edu.au, Queensland University of Technology, Brisbane, Queensland, Australia; Selen Turkay, selen.turkay@qut.edu.au, Queensland University of Technology, Brisbane, Queensland, Australia.

Manuscript submitted to ACM

professionals to automate tasks that were previously time consuming or prohibitively expensive. For example, we are now starting to see companies use LLMs to evaluate resumes [16], conduct job interviews with potential employees [10], and conduct content analysis [7]. Work has also been done to consider how LLMs may change grading practices on assessment items [17] by allowing automated assessment to move beyond multiple choice answers and evaluate short answer responses [9].

Currently though, most of this research is exploring the *quality* of the evaluations, to ensure the LLM can differentiate between low-quality and high-quality content. Two of the biggest criticisms related to the use of LLMs for this purpose are that they (1) sometimes generate inaccurate responses and/or hallucinate [15], and (2) don't effectively value creativity and unusual responses because the training causes them to prefer statistically common responses similar to their training data [11]. Modern LLMs still perform poorly on the Torreance Tests of Creative Thinking (TTCT) [12], and comparisons between LLM and human grading is often most pronounced on creative works [6].

In this paper, however, we will instead consider something else. Specifically, does metadata or contextual hints contained in the responses also influence the results of an LLM evaluation? One highly-cited study in 2012 found that science faculty members rated male applicants as significantly more competent and hireable than identical female applicants with identical resumes, solely based on the perceived gender indicated by the name [14], indicating the significant impact these types of metadata may have during the evaluation of content by human beings. In this paper, we aim to see whether similar bias effects may be observed in LLM evaluations of text. This is a particularly timely issue as, over the last few years, we have begun to see the first implementations of resume screening systems, automated interviews, and explorations into automated student grading in which trust and impartiality is paramount. For example, fairgo.io [1], micro1 [2] and HireVue [3] all provide services allowing companies to use LLM-powered agents as part of the decision-making process for job hiring.

Specifically, in this paper we aim to understand and explore the extent of this problem by systematically analyzing how LLMs respond to identical texts provided alongside varying configurations of metadata attributes. By doing so, we hope to identify specific patterns of bias and propose strategies to mitigate their impact. Our study contributes to the broader discussion on ethical AI and fairness in machine learning, emphasizing the need for transparency and accountability in the use of LLMs. The main contributions of this work are:

(1) To identify whether the inclusion of metadata may affect the results of an LLM evaluation.
(2) To identify whether the types of content being analysed may lead to different biases which should be considered.
(3) To consider potential strategies which could be used to mitigate the effect of this metadata on the final analysis.

2 Related Work

The exploration of biases and inaccuracies in LLMs and their impacts on text analysis has gained some attention in recent years. In this section, we consider relevant studies that have explored the potential challenges in using LLMs for this purpose.

A commonly discussed issue with LLMs is their tendency to "hallucinate" or generate plausible-sounding but factually incorrect information. LLMs are known to produce coherent text that is entirely fabricated, leading to concerns about the reliability of information provided by these models [2]. This issue is particularly problematic in contexts where factual accuracy is critical, such as academic writing or technical reports.

[1] https://www.fairgo.ai/
[2] https://www.micro1.ai/
[3] https://www.hirevue.com/

Research has also highlighted that despite their advanced language capabilities, LLMs often lack true understanding of context. They can interpret text based on patterns learned during training but may miss nuances and deeper contextual meanings [2]. This limitation can lead to superficial or inappropriate evaluations, especially in complex or nuanced tasks such as reflective writing, creative storytelling, or sentiment analysis.

2.1 Bias in Language Models

Research has demonstrated that word embeddings, a foundational component of many LLMs, inherently carry gender biases [5]. They showcased how word associations could reflect and propagate stereotypes, such as associating "man" with "computer programmer" and "woman" with "homemaker." These findings underscore the need for addressing biases at the fundamental level of model training.

Furthermore, there has been discussion around the "stochastic parrots" problem [2], which emphasises how LLMs can perpetuate and amplify the biases found in their training data. They highlight instances of gender, racial, and socio-economic biases in model outputs, raising concerns about the ethical implications of deploying such models in real-world applications.

2.2 Metadata and Bias

The influence of metadata on bias in text analysis has also been extensively studied outside the context of LLMs. Researchers examined how metadata, such as author gender, affects algorithmic decisions in predictive policing and hiring practices [8]. Their findings indicate that algorithms can exhibit discriminatory behavior when metadata is factored into their analysis, leading to unfair outcomes.

These biases are also not limited to algorithms. Research found that science faculty members rated male applicants as significantly more competent and hireable than identical female applicants, solely based on the perceived gender indicated by the name [14]. This gender bias in academic hiring underscores the pervasive influence of metadata in professional evaluations.

Researchers also conducted a comprehensive survey on the sources of bias in machine learning, and identified metadata as a significant contributor [13]. They argue that metadata can introduce contextual biases that skew model predictions, particularly when demographic information is used without careful consideration of its impact.

2.3 Mitigation Strategies

Efforts to mitigate biases in LLMs have led to the development of various techniques and frameworks. One approach is debiasing word embeddings, [18], which introduces methods to reduce gender bias in word representations. This technique involves adjusting the embedding space to ensure that gender-neutral words remain unbiased.

Fairness-aware machine learning frameworks, such as those discussed by Mehrabi et al [13], aim to create models that account for and correct biases during the training process. These frameworks incorporate fairness constraints and regularization techniques to minimize the impact of biased metadata on model outputs. If researchers are looking to choose an ML model or LLM to analyse text, choosing a model that follows one of these frameworks may be an effective strategy.

Finally, researchers have also emphasised the need for transparency and interpretability in analyses and reporting of results [4] to try and mitigate biases. Even if this problem cannot be eliminated entirely, ensuring that there are clear explanations of model decisions and the role of metadata, stakeholders can better understand and address potential biases in LLMs.

3 Research Questions

This study seeks to explore the potential impact of metadata bias in text analysis performed by large language models (LLMs). To guide the investigation, we focus on the following three research questions:

RQ1*: Does the inclusion of contextual metadata influence the results of an LLM text evaluation?*

This question aims to determine whether the inclusion of varying metadata, such as author name, gender, or other demographic details, influences the outcomes generated by LLMs when analyzing identical texts. By examining this, we seek to identify instances where metadata alters the interpretation or evaluation of text, highlighting potential biases inherent in LLMs.

RQ2*: How does the type of text being evaluated affect the severity and/or type of biases observed when analysing text which contains contextual metadata?*

This question investigates whether the nature of the text itself (e.g. an essay, comment, creative task, or technical writing) may affect the exact biases the LLM manifests in its analysis. Understanding this relationship will help determine if certain types of texts are more susceptible to bias due to their inherent characteristics and the accompanying metadata.

RQ3*: Are there any strategies which are effective at reducing bias during LLM evaluation of text with embedded metadata?*

This question explores potential methods to mitigate bias in LLM text analysis. We aim to evaluate the effectiveness of various strategies, such as prompt crafting and data sanitization, in reducing the influence of metadata on the outcomes of text analysis. By identifying and validating these strategies, we hope to provide practical solutions for minimizing bias and ensuring more equitable and accurate text analysis by LLMs.

By addressing these research questions, our study seeks to provide a comprehensive understanding of how metadata influences bias in LLM text analysis, how different types of texts may be affected, and what strategies can be employed to mitigate these biases. This knowledge is crucial for developing fairer, more equitable applications of LLMs across various contexts.

4 Methodology

To analyse the affect that including metadata may have during LLM analyses, we had an LLM evaluate the quality of responses given by a "student" to four meaningfully different tasks (i.e. some were technical, some were creative, some had objectively correct answers). The LLM was asked to determine a grade from 1 (worst) to 100 (best) based on the quality of the student response. The metadata was injected into the evaluation request by placing the information prior to the main text body and evaluation prompt. For the main analysis, the LLM was not given any indication whether they should or shouldn't use this information. Figure 1 shows an overview of this process.

Evaluation requests were constructed for all possible metadata and task combinations. As there were four main metadata categories being considered, and six items in every category, there were 7^4 total evaluations per task. Four tasks were considered, so there were 9604 total responses considered in the main analysis. As all possible combinations were considered, they were all equally represented. Furthermore, the ordering of the metadata items in the evaluation request were randomised to ensure that the perceived strength of the observed biases was not due to potential ordering effects. Each evaluation request was constructed with four main pieces of information:

(1) **A General Prompt**: Contains the method the LLM should use to evaluate the prompt.

(2) **Metadata**: The collated text describing the traits of the submission author.

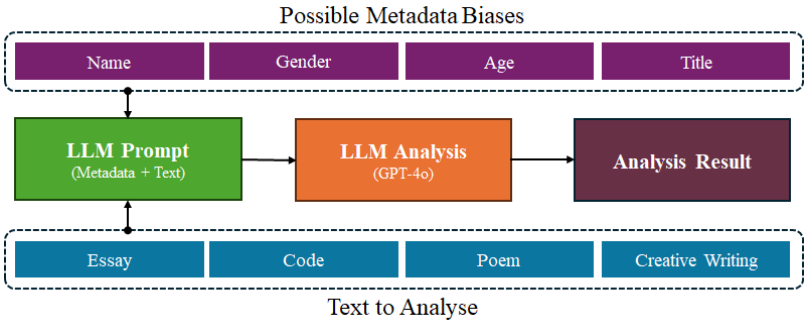

Fig. 1. The approach for having the LLM evaluate a response. A prompt was constructed containing the appropriate metadata and text to evaluate which was then sent to the LLM for final evaluation.

(3) **Task-Specific Prompt**: Additional information the LLM will need to accurately perform the evaluation (e.g. the task that was asked to be completed).

(4) **Text to Evaluate**: The text response that the LLM needs to evaluate.

A final combined prompt was generated containing each of these pieces of information. Here is how it looks in full:

> Grade the below work submitted by a student with a grade from A+ (best) to E- (worst), including ONLY the grade in your response. The response below was made by a student named John. The student is male. The student is middle-aged. The student was asked "Write a poem on a topic of your choice". This is their response:
>
> In a whispering grove where the old oaks sway,
> Beneath the golden light of a fading day,
> There lies a realm where shadows play,
> And echoes of time quietly stay.
> [Continued...]

4.1 Biases Analysed

When designing the study, we planned to run an initial test with a large number of metadata categories, each having a very small number of items (e.g., 3). From this, we aimed to identify the most prevalent biases which could then be explored in further detail in the main analysis (which would have a more substantial number of items per category). Specifically, we considered the following biases, as they are often considered in bias/discrimination literature: *(1) Name*, *(2) Gender*, *(3) Age*, *(4) Title*, *(5) Country of Origin*, *(6) Education Level*, *(7) Background Culture*, and *(8) Health Status*.

This initial analysis lead to an interesting observation, however, as it was found that as more metadata was included with the text to evaluate, the effect of the bias became both smaller and harder to predict. For example, running an initial linear regression which considered all eight of the above categories generated a model with an *extremely* low coefficient of determination ($r^2 = 0.03$). This value improved as metadata categories were removed, with 6 categories having $r^2 = 0.31$ and four categories having $r^2 = 0.50$. We did not explore using fewer than four categories as there would not be enough unique combinations for an in-depth analysis.

Instead, we decided to have the main analysis consider four metadata categories which are often prone to bias, and would also be most likely to appear in these types of evaluations: **Name**, **Gender**, **Title**, and **Age**. For each category, we chose six category items to represent the category. When deciding on which items to use within each category, we aimed to create a diverse spread of responses. For example, we chose names common in different countries and cultures as they would be most likely to highlight any potential biases and lead to interesting insights. The exact category items chosen were:

(1) **Name**: *Jamal*, *Hiroshi*, *David*, *Priya*, *Charlotte*, *Amina*

(2) **Gender**: *Man*, *Woman*, *Non-binary*, *Genderqueer*, *Genderfluid*, and *'Prefer not to say'*

(3) **Age**: *A Child*, *A Teenager*, *A College Student*, *In Their Late Twenties*, *Middle-Aged Adult*, and *Older Adult*

(4) **Title**: *Mr*, *Ms*, *Mrs*, *Dr*, *Prof*, and *Eng*.

It should be noted, however, that our exclusion of certain biases does not mean that we believe they are in any way less important. There was a practical upper limit on the amount of items we could analyse in this study, which meant that we had to aggressively prune potential categories and category items. We chose to go with these four metadata categories because we believed they would be most likely to appear organically in these types of evaluations (e.g. in a title page on an assignment submission, a social media comment, a product review, or a resume). Furthermore, we chose six items per category because a larger number of items would have led to the results set quickly becoming unmanageable (if we had instead used 10 items per category, the dataset would have been 10 times larger). Some of the categories would have also had diminishing benefits with more items (e.g. title), as most *common* category items were already considered.

4.2 Tasks for Evaluation

Even if a LLM is prone to bias, it does not mean that it will have identical biases in all situations. For example, an LLM may positively bias a group on a technical task, but negatively bias them on a creative task. For this reason, we chose to test the LLM on four different tasks, which aimed to represent different types of activities. The tasks chosen for the final analysis were:

(1) **Creative Story**: Creative writing emphasizes imagination and storytelling skills. This prompt tests the LLM's capability to generate original content. Creative tasks may be particularly susceptible to biases related to perceived creativity and originality, which can be influenced by metadata.

(2) **Factual Essay**: Factual essays demand clear, logical presentation of information and arguments based on evidence. This type of writing assesses the LLM's ability to handle objective content. Biases in this context could affect the perceived credibility and reliability of the information provided.

(3) **Poem**: Poetry requires creativity, emotional expression, and mastery of language. This prompt assesses the LLM's ability to generate artistic and meaningful content. Biases could impact the perceived depth and artistic quality of the poem.

(4) **Programming Task**: Programming tasks involve writing code to solve specific problems. This prompt evaluates the LLM's technical proficiency and logical thinking. Metadata biases could affect the perceived technical expertise and problem-solving abilities of the author.

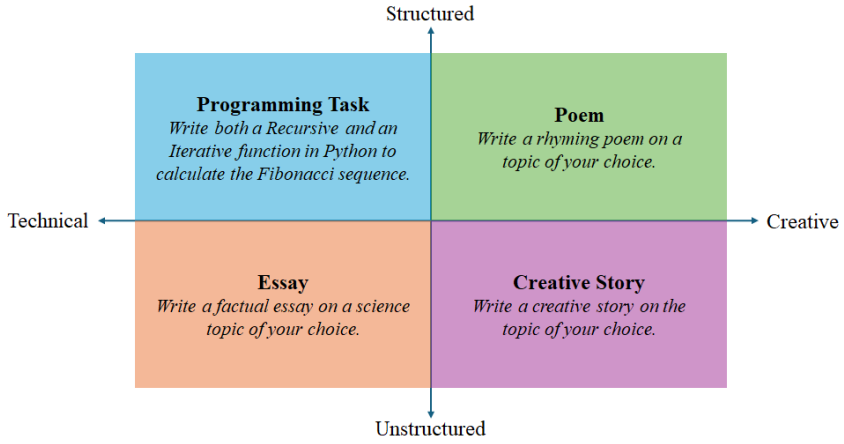

Fig. 2. Shows the four tasks that the LLM evaluated. These tasks were chosen because they were meaningfully different along two different axes. The Factual Essay and Programming Task were highly technical, while the Poem and Creative Story were more creative. The Programming Task and the Poem both had highly specific formatting and structuring requirements, while the essay and creative story were comparatively much more lax in their layout.

We chose these specific four tasks as they had varying levels of creativity and required structure. This breakdown can be seen in Figure 2 below. We also considered exploring the impact of metadata biases on non-text prompts, such as the grading of handwriting and the evaluation of art. However, the costs associated with such an analysis would have been prohibitive given the scale of the study.

To ensure that we didn't encounter problems due to a ceiling effect, we did an initial analysis without the metadata to ensure that the scores were not too low or too high. Specifically, we wanted the scores prior to the metadata inclusions to be between 40% and 75%, so that the inclusion of the metadata was unlikely to have the score hit the floor or ceiling (0% or 100%). To achieve this, we needed to introduce some errors into the programming task, as it scored too high in this analysis (90%).

4.3 LLM Used for the Analysis

The final analysis was done using OpenAI's **GPT-4o model** (specifically, *gpt-4o-2024-05-13* [4]). This was the most up-to-date model at the time of writing which had an API which could facilitate the required number of analysis requests (i.e. 10,000+ requests, each with 600 tokens each).

4.4 Dataset Generation and Replication

The complete list of metadata categories and prompts has been included with the supplementary files. Additionally, both the raw CSV dataset, as well as a one-hot encoded dataset, have been included to make for easier reproduction analysis. The source code used to generate the raw dataset has also been included, but the API and organisation Keys have been redacted.

[4]https://platform.openai.com/docs/models/gpt-4o

5 Analysis Procedure

The dataset for each of the five prompts were analysed with a categorical linear regression. Individual items within each category were one-hot encoded to identify how likely each item within each metadata group was to experience bias (i.e. instead of only analysing that there was bias towards age, we can instead see which age groups experience that bias, either positively or negatively). Note that because there is no obvious control or baseline option, the data requires some further analysis (e.g. via category item means) before interpretation.

If items are not shown directly in the regression results, it is because they were excluded from the model due to high multicollinearity (i.e. high correlation with other model predictors), a near-zero beta coefficient, or insignificance (i.e. there is no linear relationship with the dependent variable after all other variables have been controlled for).

6 Results

To quantify the impact of different metadata attributes on the evaluation scores provided by the LLM, we performed four multiple linear regression analyses. The dependent variable was the evaluation score, converted from the grade provided by the LLM. The independent variables were the metadata attributes described earlier: Name, Gender, Age, and Title, all of which were treated as categorical variables.

Separate regressions were run for each of the four evaluation tasks. To better understand how individual items within each category were effected, we used one-hot encoding where each category item was represented by a binary (0 or 1) indicator variable within the model. This allowed us to include the categorical variables within the regression model (e.g. to explore whether a *specific* age group may have experienced particular bias). After each regression, we highlight salient findings identified in the analysis.

6.1 Factual Essay

Table 1 summarizes the regression coefficients, standard errors, t-values, and significance levels for each metadata attribute. Below are some of the key findings identified in the analysis.

(1) **Age**: The Teenager (β = -2.4, $p < 0.001$), Late Twenties (β = -6.4, $p < 0.001$), Middle Aged (β = -4.9, $p < 0.001$), and Senior Citizen (β = -6.4, $p < 0.001$) age groups were all negatively biased compared to the 'Age Not Given' baseline group.

(2) **Gender**: The Man (β = -2.0, $p < 0.001$) and Women (β = -1.4, $p < 0.001$) categories were negatively biased compared to the 'Gender Not Given' baseline group. The Non-binary (β = 2.1, $p < 0.001$) and genderfluid (β = 0.9, $p < 0.01$) groups were positively biased.

(3) **Name**: All name categories were negatively biased when compared to the 'Name Not Given' baseline, with Jamal (β = -2.8, $p < 0.001$), David (β = -3.2, $p < 0.001$), and Armina (β = -4.5, $p < 0.001$) being the most notable.

(4) **Title**: The Ms (β = 1.1, $p < 0.001$), Doctor (β = 3.2, $p < 0.001$), Professor (β = 1.5, $p < 0.001$), and Engineer (β = 2.3, $p < 0.001$) titles were all positively biased when compared with the 'Title Not Given' baseline.

6.2 Poem

Table 1 summarizes the regression coefficients, standard errors, t-values, and significance levels for each metadata attribute. Below are some of the key findings identified in the analysis.

(1) The Late Twenties Age (β = -0.1, $p < 0.05$) group was slightly negatively biased over the 'Age Not Given' baseline.

Table 1. **Factual Essay Task**: Summary of Linear Categorical Regression, showing the effect of category items on final evaluation.

Variable	Mean	(β)	Std. Error	Std. Coeff. (Beta)	t-value	p-value
(Constant)		78.934	0.389		202.968	<.001
Age						
Child	77.89	0.235	0.289	0.017	0.814	0.415
Teenager	75.39	-2.360	0.291	-0.164	-8.117	<.001
College Student	78.17	0.484	0.289	0.034	1.676	0.094
In their Late Twenties	71.18	-6.361	0.294	-0.436	-21.646	<.001
Middle Aged Adult	72.81	-4.912	0.288	-0.348	-17.081	<.001
Senior Citizen	71.16	-6.442	0.288	-0.456	-22.365	<.001
Gender						
Man	72.92	-2.041	0.289	-0.143	-7.064	<.001
Women	73.59	-1.387	0.288	-0.098	-4.812	<.001
Non-binary	77.18	2.085	0.289	0.146	7.212	<.001
Prefer not to say	74.38	-0.504	0.290	-0.035	-1.737	0.083
Genderqueer	75.33	0.275	0.289	0.019	0.950	0.342
Genderfluid	75.97	0.886	0.290	0.062	3.052	0.002
Name						
Jamal	74.44	-2.777	0.288	-0.196	-9.638	<.001
Hiroshi	75.28	-1.903	0.289	-0.134	-6.584	<.001
David	73.96	-3.212	0.290	-0.225	-11.070	<.001
Priya	75.45	-1.868	0.292	-0.129	-6.408	<.001
Charlotte	75.24	-1.873	0.290	-0.131	-6.463	<.001
Amina	72.75	-4.478	0.290	-0.314	-15.466	<.001
Title						
Mr	73.83	0.062	0.288	0.004	0.216	0.829
Ms	74.90	1.143	0.287	0.081	3.985	<.001
Mrs	73.53	-0.098	0.288	-0.007	-0.340	0.734
Doctor	76.95	3.194	0.290	0.221	11.015	<.001
Professor	75.30	1.547	0.289	0.108	5.352	<.001
Engineer	76.17	2.280	0.288	0.160	7.915	<.001

(2) Amina (β = -0.1, p < 0.05) was slightly negatively biased over the 'Name Not Given' baseline.
(3) The Man (β = -0.14, p < 0.001) group was slightly negatively biased over the 'Gender Not Given' baseline.

6.3 Coding Exercise

Table 3 summarizes the regression coefficients, standard errors, t-values, and significance levels for each metadata attribute. Below are some of the key findings identified in the analysis.

(1) Child (β = -2.7, p < 0.001) was negatively biased when compared with the 'Age Not Given' baseline. College Student (β = 4.0, p < 0.001) and Middle Aged Adult (β = 2.6, p < 0.001) were both positively biased.

(2) Women (β = 1.4, p < 0.05), Non-binary (β = 7.0, p < 0.001), Genderqueer (β = 7.4, p < 0.001), and Genderfluid (β = 7.5, p < 0.001) were all positively biased when compared with the 'Gender Not Given' baseline. Given the generally low scores, this was led to these groups being given much higher scores.

Table 2. **Rhyming Poetry Task**: Summary of Linear Categorical Regression, showing the effect of category items on final evaluation.

Variable	Mean	(β)	Std. Error	Std. Coeff. (Beta)	t-value	p-value
(Constant)		80.039	0.056		1432.202	<.001
Age						
Child	80.00	0.000	0.042	0.000	0.007	0.994
Teenager	80.00	-0.003	0.042	-0.002	-0.063	0.950
College Student	80.00	0.000	0.042	0.000	0.006	0.995
In their Late Twenties	79.90	-0.105	0.041	-0.074	-2.535	0.011
Middle Aged Adult	79.93	-0.073	0.042	-0.050	-1.739	0.082
Senior Citizen	80.00	0.001	0.041	0.001	0.022	0.982
Gender						
Man	79.86	-0.140	0.042	-0.098	-3.358	<.001
Women	79.97	-0.034	0.041	-0.024	-0.811	0.417
Non-binary	80.00	0.003	0.042	0.002	0.061	0.951
Prefer not to say	80.00	-0.002	0.042	-0.001	-0.046	0.964
Genderqueer	80.00	-0.001	0.041	-0.001	-0.028	0.978
Genderfluid	80.00	0.003	0.042	0.002	0.061	0.951
Name						
Jamal	80.00	0.004	0.042	0.003	0.089	0.929
Hiroshi	80.00	0.003	0.041	0.002	0.076	0.940
David	79.97	-0.032	0.042	-0.023	-0.782	0.435
Priya	79.96	-0.035	0.042	-0.024	-0.830	0.407
Charlotte	80.00	0.002	0.042	0.001	0.051	0.959
Amina	79.90	-0.100	0.041	-0.071	-2.428	0.015
Title						
Mr	79.93	-0.034	0.041	-0.024	-0.831	0.406
Ms	80.00	0.031	0.042	0.021	0.734	0.463
Mrs	79.97	-0.001	0.042	0.000	-0.016	0.987
Doctor	80.00	0.036	0.042	0.025	0.854	0.393
Professor	80.00	0.035	0.042	0.024	0.832	0.405
Engineer	79.96	-0.001	0.042	-0.001	-0.031	0.975

(3) Hiroshi (β = 4.0, p < 0.001), Priya (β = 1.7, p < 0.005), Charlotte (β = 2.5, p < 0.001), and Amina (β = 2.3, p < 0.001) were all positively biased when compared to the 'Name Not Given' baseline.

(4) The Doctor (β = 2.9, p < 0.001), Professor (β = 1.9, p < 0.001), and Engineer (β = 5.4, p < 0.001) titles were all positively biased when compared with the 'Title Not Given' baseline. This is interesting because all of the professional titles were positively biased, and none of the non-professional ones were.

6.4 Creative Story

Table 4 summarizes the regression coefficients, standard errors, t-values, and significance levels for each metadata attribute. The model had an r^2 fit of 0.37. Below are some of the key findings identified in the analysis.

(1) Age: The Child category (β = 5.0, p < 0.001), teenager category (β = 1.2, p < 0.001), and College Student (β = 1.4, p < 0.001) categories all had a positive bias over the 'Age Not Specified' category.

Table 3. **Coding Task**: Summary of Linear Categorical Regression, showing the effect of category items on final evaluation.

Variable	Mean	(β)	Std. Error	Std. Coeff. (Beta)	t-value	p-value
(Constant)		41.352	0.923		44.789	<.001
Age						
Child	44.78	-2.728	0.679	-0.102	-4.019	<.001
Teenager	49.79	2.159	0.681	0.080	3.170	0.002
College Student	51.51	4.035	0.687	0.148	5.876	<.001
In their Late Twenties	46.67	-0.865	0.687	-0.032	-1.260	0.208
Middle Aged Adult	50.14	2.622	0.687	0.096	3.815	<.001
Senior Citizen	48.67	1.136	0.687	0.042	1.655	0.098
Gender						
Man	45.51	0.465	0.686	0.017	0.678	0.498
Women	46.54	1.437	0.687	0.053	2.092	0.037
Non-binary	52.16	7.069	0.680	0.265	10.396	<.001
Prefer not to say	44.22	-0.935	0.688	-0.034	-1.358	0.175
Genderqueer	52.69	7.421	0.687	0.274	10.802	<.001
Genderfluid	52.54	7.453	0.689	0.273	10.816	<.001
Name						
Jamal	47.08	0.169	0.685	0.006	0.247	0.805
Hiroshi	50.77	3.995	0.690	0.147	5.790	<.001
David	47.30	0.307	0.685	0.011	0.448	0.654
Priya	48.28	1.708	0.693	0.062	2.462	0.014
Charlotte	49.31	2.469	0.690	0.091	3.576	<.001
Amina	49.29	2.310	0.688	0.086	3.355	<.001
Title						
Mr	45.80	-1.148	0.686	-0.043	-1.673	0.094
Ms	47.29	0.217	0.696	0.008	0.312	0.755
Mrs	47.12	0.110	0.684	0.004	0.162	0.872
Doctor	50.04	2.865	0.691	0.105	4.145	<.001
Professor	49.05	1.864	0.690	0.069	2.700	0.007
Engineer	52.49	5.392	0.686	0.201	7.855	<.001

(2) The *Man* (β = -0.62, p < 0.01) and *Women* (β = -0.42, p = 0.06) Categories once again had a negative bias, when compared with the other conditions, while the non-binary (β = 2.0, p < 0.001), genderqueer (β = 0.5, p < 0.05) and genderfluid (β = 1.2, p < 0.001) categories all had positive biases over the 'Gender Not Specified' category.

(3) All names had a negative bias over the 'Name Not Specified' category, with David (β = -0.85, p < 0.001), Priya (β = -1.01, p < 0.001), and Amina (β = -1.05, p < 0.001) being the most notable.

(4) The Mr (β = 1.0, p < 0.001), Ms (β = 1.2, p < 0.001) and Mrs (β = 0.6, p < 0.001) titles all had a positive bias over the 'Title Not Specified' category. This is notable as all three of the non-professional titles were positively biased, while no statistically significant effect was observed on the other three.

Table 4. **Creating Writing Task**: Summary of Linear Categorical Regression, showing the effect of category items on final evaluation.

Variable	Mean	(β)	Std. Error	Std. Coeff. (Beta)	t-value	p-value
(Constant)	70.044	0.299		234.087	<.001	
Age						
Child	75.34	5.024	0.224	0.524	22.459	<.001
Teenager	71.44	1.162	0.223	0.122	5.210	<.001
College Student	71.65	1.377	0.223	0.145	6.176	<.001
In their Late Twenties	70.28	0.009	0.223	0.001	0.041	0.967
Middle Aged Adult	70.00	-0.275	0.221	-0.029	-1.245	0.213
Senior Citizen	70.00	-0.230	0.222	-0.024	-1.036	0.300
Gender						
Man	70.21	-0.623	0.222	-0.066	-2.806	0.005
Women	70.41	-0.420	0.221	-0.045	-1.902	0.057
Non-binary	72.86	1.995	0.223	0.209	8.940	<.001
Prefer not to say	71.14	0.356	0.224	0.037	1.591	0.112
Genderqueer	71.36	0.487	0.222	0.051	2.189	0.029
Genderfluid	72.13	1.246	0.222	0.131	5.602	<.001
Name						
Jamal	71.36	-0.502	0.223	-0.053	-2.251	0.024
Hiroshi	71.37	-0.529	0.223	-0.055	-2.367	0.018
David	71.03	-0.853	0.222	-0.090	-3.838	<.001
Priya	70.87	-1.012	0.223	-0.107	-4.537	<.001
Charlotte	71.47	-0.442	0.223	-0.046	-1.979	0.048
Amina	70.87	-1.049	0.223	-0.111	-4.711	<.001
Title						
Mr	71.77	0.992	0.223	0.103	4.446	<.001
Ms	72.09	1.219	0.219	0.130	5.558	<.001
Mrs	71.39	0.560	0.221	0.059	2.537	0.011
Doctor	70.95	0.090	0.220	0.010	0.412	0.681
Professor	70.60	-0.201	0.222	-0.021	-0.904	0.366
Engineer	71.24	0.356	0.224	0.037	1.593	0.111

6.5 Further Insights

While our primary focus was on investigating the impact of metadata on bias in text evaluations by LLMs, our analysis also provided additional insights into the performance and behavior of LLMs in evaluation tasks. These insights, although not the core focus of our research, are still important to the overall concept of LOLM evaluations. Each of these are discussed below.

6.5.1 Tendency Towards High Scores. We observed that LLMs tend to perform poorly when tasked with assigning scores on a 1-100 scale. The models frequently provided high, above-average scores but rarely awarded full marks. Even for texts that were objectively poor, the LLMs assigned reasonable grades, indicating a tendency to avoid extreme low scores. This behavior suggests that LLMs may have an inherent bias towards moderation in scoring, potentially to avoid harsh judgments. If you are looking to use LLMs for these sorts of evaluations, understanding this phenomenon and properly accounting for it is very important. For example, you may want to do a post-analysis transform of the

evaluations by feeding it exemplar work from various grade bands to better tune its responses with your intended grades.

6.5.2 Response Categorisation. Throughout this analysis, we found that LLMs were better at evaluating work when categorizing results using a granular grading system (e.g. A+ to E-), rather than a percentage scale. While this may seem that it limits the set of responses the LLM can provide, LLMs already tend to place their scores into buckets, so very little sensitivity is lost. We found this particularly effective at getting the LLM to stop providing unusually high evaluation scores and start using more of the overall grade band. Overall, this categorical approach led to more reasonable and varied grading outcomes.

6.5.3 Unexplained Variance in LLM Scores. Our regression analysis revealed a significant amount of unexplained variance in the LLM scores, which could not be accounted for by the metadata attributes alone. This unexplained variance indicates that there are other factors influencing the LLM's grading decisions. Unfortunately, the exact reasons for this remain opaque due to the complex and non-transparent nature of these models. The inability to interrogate the internal workings of the model makes it challenging to identify these factors, posing a limitation in fully understanding and interpreting LLM evaluations. Note that this is despite using a *temperature* and a *top p* of 0 for these evaluations specifically to lower this variance in responses. This was particularly common when more metadata categories were considered in the models, suggesting that the LLM may be more likely to have unexplained grading differences when there are more factors for it to 'consider' in its evaluations.

6.6 Strategies to Limit Metadata Bias

In this study, we used rather generic prompts. It is possible that more specific prompts would yield better results, but we were looking to investigate whether there was general bias in the model which needed to be considered. Overall, the above analyses have shown that while the exact biases change from task to task, there is still a non-trivial amount of variance in the grades due to the metadata which may be included. To address this, we want to briefly consider whether changes to the prompt may affect this bias.

To investigate this, we ran another regression using the essay task. The prompt was adjusted to include the following sentence at the end of the grading prompt: *"Your grade should be based solely on the quality of the work, do not consider the described traits of the student in your response."* The results of this analysis can be seen in Table 5.

While this did have a meaningful impact on the overall set of responses, it still did not entirely correct the problem. This resulted in people in their late twenties being negatively biased (β = -1.53, p < 0.001), and non-binary (β = 1.25, p < 0.001), genderqueer (β = 0.55, p < 0.05), and genderfluid (β = 0.93, p < 0.001) people being positively biased. People with the Engineer title were also still positively biased (β = 0.60, p < 0.009).

It should also be noted that redacting the information (e.g. "Gender: [REDACTED]") also had a similar problem. In some tasks, this led to positive bias, and in others it led to negative bias.

7 Discussion

The results from this study suggest that there are currently non-trivial biases which may exist in LLMs. All four metadata categories indicated statistically significant biases in some, if not all, of the items tested. When considering RQ1, we believe that there is sufficient evidence to suggest that these biases do exist, at least in GPT-4o (the LLM used for this test). When looking to do any large-scale analysis using an LLM, we believe that researchers and practitioners should consider the implications that this may have on their findings.

Table 5. **Essay Evaluation Regression Results (Using the Defensive Prompt)**: Summary of Linear Categorical Regression Results.

Variable	(β)	Std. Error	t-value	p-value
(Constant)	79.678	0.207	385.712	<.001
Age				
Child	0.460	0.155	2.959	0.003
Teenager	0.461	0.155	2.963	0.003
College Student	0.432	0.155	2.787	0.005
In their Late Twenties	-0.608	0.157	-3.869	<.001
Middle Aged Adult	0.160	0.155	1.034	0.301
Senior Citizen	0.501	0.156	3.219	0.001
Gender				
Man	-0.825	0.154	-5.363	<.001
Women	-0.553	0.155	-3.562	<.001
Nonbinary	0.145	0.154	0.943	0.346
Prefer not to say	0.035	0.152	0.230	0.818
Genderqueer	-0.030	0.154	-0.198	0.843
Genderfluid	0.049	0.154	0.320	0.749
Name				
Jamal	-0.055	0.154	-0.359	0.720
Hiroshi	-0.310	0.154	-2.008	0.045
David	-0.829	0.156	-5.315	<.001
Priya	-0.324	0.155	-2.093	0.037
Charlotte	-0.295	0.155	-1.899	0.058
Amina	-0.293	0.154	-1.898	0.058
Title				
Mr	-0.109	0.154	-0.705	0.481
Ms	0.031	0.154	0.202	0.840
Mrs	0.314	0.154	2.042	0.041
Doctor	0.497	0.156	3.194	0.001
Professor	0.487	0.152	3.198	0.001
Engineer	0.342	0.153	2.235	0.026

The bias was easy to identify in this study because identical tasks were used, but if this was not the case, it would be extremely difficult to spot evaluation inconsistencies. If at all possible, we would strongly recommend that you strip metadata from content before it is evaluated. For example, if a researcher was looking to analyse user review data, stripping names and other personal information may yield more objective results. This is not always possible, however, as often the potentially identifying metadata may *also* be present in the text. In these situations, stripping the metadata would be a non-trivial exercise.

Overall, the results from this analysis were quite interesting. Surprisingly, the poem and the creative writing task (the two creative tasks) experienced less bias than the factual essay and the coding exercise (the two technical tasks). This is an unexpected result because we thought that the less technical activities would give the LLM more latitude when deciding on a grade. It is unclear whether or not this is an extension of the problems LLMs have when grading

creative works [6, 12]. Surprisingly, the *Programming Activity*, a task with an objective solution, experienced some wide variations in grades (e.g. Hiroshi got over 20% more marks than the baseline student that didn't specify a name). This is a shocking result and indicates that even when a task evaluation seems straightforward, it may still be prone to a large amount of bias.

Of the four metadata categories considered, the biases observed in age were perhaps the most confusing. One of the most consistent biases we observed were that college students were often given positive biases, and never negatively biased. This, however, doesn't make much sense given that there are both younger and older groups listed. For example, if the LLM was using the metadata as additional context while grading the coding task, it would make more sense for the child or teenager to experience the largest positive biases, as the work would be more impressive for them. The LLM instead negatively biased the child, leading to approximately a 10% better grade for the college student. Alternatively, if it was more dismissive of work because it was done by someone with little perceived authority (e.g. a child), it makes little sense that it would bias college students over middle-aged adults. Based on prior literature in the area, we were expecting to see a more interpretable set of biases, but we were unable to identify a clear reason for this result. One possibility we considered is that this may not be due to an inherent bias in the underlying training data, but instead an artefact of the model fine-tuning.

The biases in gender were much more consistent than those related to age. Students identifying as a man or woman were more likely to receive negative biases, while non-binary, genderqueer and genderfluid students tended to receive positive biases. While it is difficult to say with any degree of certainty why this may be happening, it is likely influenced by the difference in sample size in the underlying training data. Non-binary, genderfluid, and genderqueer people have had limited representation in text which means that the LLM likely has limited training data related to these genders. To avoid potential discrimination, it is also possible that the GPT-4o conducted model fine-tuning around gender-related topics, which may have resulted in part or all of this difference.

When we decided to consider name in the analysis, we did it under the assumption that it would act as a pseudo-indicator for several possible biases as names often have cultural significance, and can even reflect country of origin or socioeconomic status [3]. Given this, there are some interesting findings here which could warrant further research. For example, the name *Armina* was negatively biased on essays but was positively biased on the creative story. Jamal, on the other hand, was negatively biased on both the essay and the creative writing task. Interestingly, for the essay and creative writing task, the students *all* performed worse than the baseline student who didn't list a name, and this was the only category which experienced this phenomenon.

Finally, titles tended to have the most consistent and explainable biases. The three non-professional titles analysed all performed better on the creating writing task, while all of the professional titles performed better on the essay and programming technical exercises. This suggests that the LLM may be inclined to give higher marks to people perceived to be skilled, or may have been more reticent to criticise work created by someone it perceived as having a degree of authority.

Given these findings, there is sufficient evidence to say that there is a relationship between the type of task being evaluated, and the way the metadata may bias the result. The name *Armina* seems to have experienced the most significant shift - going from a score well below the expected value on an essay to well above the expected value on a coding task.

Finally, we must consider RQ3 and how these biases can be mitigated. If at all possible, metadata should be removed before these evaluations. For example, if there is a form with personal information on the title page, rather than have the LLM analyse the entire document, you may want to first remove that page. It is not sufficient to simply redact

references to certain types of metadata. For example, the results here indicate that simply redacting this information may still lead to biases (either positive or negative). Instead, our recommendation would be to be very explicit in the prompt and ensure that the LLM is specifically instructed to ignore the information and not consider it during the evaluation. This does not remove the problem entirely, but it does seem to partially mitigate the issue. Finally, when doing any of these analyses, it is critical that you tune the LLM parameters correctly to remove variance from the responses. For example, setting the *Temperature* and *Top P* scores to 0 ensure that the LLM isn't trying to be "creative" with the scoring process, and will ensure that the LLM gives you consistent responses each time the evaluation is performed. None of these methods are perfect, however, and if you intend to have an LLM perform an analysis you should be cognisant that its evaluation will likely contain some degree of bias.

Overall, the goal of this paper is to help document the effects of metadata biases and enhance the transparency and reproducibility of research findings which do these analyses. Researchers must recognize that such biases may distort results and undermine the validity of their studies. This awareness is crucial for designing experiments that either mitigate these biases or try to explicitly account for them in the analyses. The influence of metadata on LLM outputs underscores the need for developing algorithms that are resilient to such biases. Where possible, we recommend that researchers disclose the types of metadata that may have been present in the analysis, and any prompt engineering or other techniques that were used to mitigate its potential impact on results. The potential ethical ramifications of biased AI outputs are significant, and researchers have a responsibility to consider the potential harm that biased evaluations may cause to individuals and groups. Explicitly working to mitigate these risks will assist us in striving for fairness, accountability, and transparency to ensure that AI systems benefit all users equitably.

Furthermore, industry professionals which are considering the use of these systems for decision-making processes such as hiring, academic grading, or content moderation should be aware that the biases identified in this study may make the use of these systems problematic in many circumstances, as biased outputs can lead to unfair or discriminatory decisions. It is therefore imperative that practitioners incorporate stategies to handle metadata appropriate within their systems and analyses, and consider approaches which may mitigate the impact of this metadata on the final evaluations.

8 Conclusion and Future Work

The rise of LLMs has allowed for us to handle and analyse large amounts of data. While this has clear utility for research and professional practice, this study highlights the significant biases in text evaluations performed by LLMs due to metadata attributes such as age, gender, title, and name. These findings underscore the critical need for us to understand and address potential metadata biases to ensure fair and unbiased text analysis and ensure that LLMs to not perpetuate or amplify existing social biases.

While this initial study aims to highlight potential problematic behaviours of LLMs when evaluating, categorising or responding to content where metadata is present, there are still many other areas which warrant further investigation in the future. This study looked at four of the most prevalent pieces of metadata to explore their potential bias, but there are several common areas for problematic bias which were not explored in this research (e.g. cultural background and socioeconomic status). Furthermore, while this research did look to consider whether certain types of tasks (e.g. creative vs. technical) may have different biases, without replicating these findings with other creative and technical tasks, we are unable to say with certainty which aspects of the tasks caused the differences. Finally, this paper had a brief look at techniques which can be used to try and mitigate possible biases during evaluations, but a more in-depth exploration of this topic would be highly valuable.

By addressing these areas, future research can build on the foundation laid by this study, contributing to the development of fairer and more ethical AI systems.

References

[1] Valentina Alto. 2023. *Modern Generative AI with ChatGPT and OpenAI Models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4.* Packt Publishing Ltd.

[2] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[3] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* 94, 4 (2004), 991–1013.

[4] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*. PMLR, 149–159.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[6] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–34.

[7] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924* (2023).

[8] 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*. 120–128.

[9] O Fagbohun, N Iduwe, M Abdullahi, A Ifaturoti, and O Nwanna. 2024. Beyond traditional assessment: Exploring the impact of large language models on grading practices. *Journal of Artifical Intelligence and Machine Learning & Data Science* 2, 1 (2024), 1–8.

[10] Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2024. Application of llm agents in recruitment: A novel framework for resume screening. *arXiv preprint arXiv:2401.08315* (2024).

[11] Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of LLMs on creative writing. *arXiv preprint arXiv:2310.08433* (2023).

[12] Erik E Guzik, Christian Byrge, and Christian Gilde. 2023. The originality of machines: AI takes the Torrance Test. *Journal of Creativity* 33, 3 (2023), 100065.

[13] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[14] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences* 109, 41 (2012), 16474–16479.

[15] Gabrijela Perković, Antun Drobnjak, and Ivica Botički. 2024. Hallucinations in LLMs: Understanding and Addressing Challenges. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*. IEEE, 2084–2088.

[16] M Rithani, R Venkatakrishnan, et al. 2024. Empirical Evaluation of Large Language Models in Resume Classification. In *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*. IEEE, 1–4.

[17] Johannes Schneider, Bernd Schenk, Christina Niklaus, and Michaelis Vlachos. 2023. Towards llm-based autograding for short textual answers. *arXiv preprint arXiv:2309.11508* (2023).

[18] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876* (2018).

Received 26 June 2024