Impact of Training Co-occurrence on Gender Bias Exhibited by Instruction-Tuned LLMs

Aditya Vinodh

December 17, 2024

Abstract

This study investigates gender biases embedded in instruction-tuning datasets, with a particular focus on Dolly 15k, to evaluate how implicit entity-stereotype associations manifest in AI model behaviors. Building on prior research that identified subtle biases in tutorbot responses to dialectal variations, this work shifts the focus to the systemic origins of bias in training data. We employ two tasks – forced masked-word completion and open-ended generation – and metrics such as Pointwise Mutual Information (PMI) alongside response evaluation techniques like lexical analysis, text classification, and sentiment analysis. Findings suggest that while the effects of co-occurrence patterns are slightly pronounced in the open-ended generation setting, underlying dataset distributions are more likely to influence chatbot responses in ways that may perpetuate gender stereotypes. We also demonstrate that controlling for the order of data presentation reduces exhibited bias, highlighting how the sequence of preference options can lead to misleading results unless randomized. This research underscores the importance of dataset auditing to mitigate biases and contributes to the development of fairer, more equitable AI systems.

1 Introduction

Generative AI models, such as OpenAI's ChatGPT, Anthropic's Claude, and Google's Gemini are transforming natural language processing by enabling tasks like text generation, machine translation, and automated reasoning. Trained on vast datasets, these models generate human-like responses across domains. As demand for LLMs grows, improving model performance remains a priority, with efforts focused on scaling parameters, optimizing training algorithms, and reducing computational costs to handle more complex tasks and ensure contextually appropriate outputs in long-form dialogues. Despite significant improvements in accuracy and efficiency, challenges remain in aligning LLMs with human values, especially in mitigating harmful biases. This is particularly important in sensitive fields like healthcare, education, and law enforcement, where biased outputs can have serious societal consequences.

1.1 Bias Metrics and Pointwise Mutual Information

To quantify biases systematically, researchers have developed various metrics that focus on detecting disparities in model outputs. One such metric, Pointwise Mutual Information (PMI), has gained traction for its ability to measure the co-occurrence of words in training corpora. Valentini et al., 2023 [10] applied PMI-based metrics to evaluate gender biases in text by comparing the conditional probabilities of words' co-occurrence. This approach calculates the difference in PMI values for different groups:

$$\text{BiasPMI} = \text{PMI}(A, C) - \text{PMI}(B, C)$$

where PMI measures the association between words in a corpus using conditional probabilities:

$$\text{PMI}(X, Y) = \log\left(\frac{P(X, Y)}{P(X)P(Y)}\right)$$

This method provides a transparent way to measure the strength of gendered associations in text. Valentini's approach has been shown to provide valuable insights into systemic patterns by estimating co-occurrence probabilities within predefined contexts, particularly in uncovering gender disparities in textual data.

In addition to PMI, other metrics focus on evaluating stereotype-laden content, toxic language, and biased sentiment. Frameworks such as HONEST [8] and TrustGPT [4] go further by analyzing the fairness of outputs in terms of harmful content across social groups.

1.2 Defining Terms, Entities, and Seeds

We define key concepts to understand the nature of bias measurement. "Entities" refer to individual components such as words or concepts in a corpus that are linked to gender or other social constructs. For example, entities might include terms like "he, she, his, hers" or "son, daughter, man, woman," which carry implicit gender associations based on societal norms. "Seeds" refer to specific terms or sets of terms often chosen to represent stereotypes or demographic categories, such as names or occupations with historical gender associations. The selection of seeds is crucial as it directly influences the outcome of bias measurement techniques [1]. To avoid any effects of culturally reductive or incomplete seed sets, we simply use fixed labels of 50 professions, although the trade-off associated with this decision is discussed later.

1.3 Study Goals and Contributions

This study builds on the growing body of research investigating bias in large language models (LLMs) and trustworthiness of AI. A key contribution of this work is the identification of unfair distribution patterns in a widely used training dataset, highlighting how gendered imbalances influence model outputs and perpetuate stereotypes. This finding emphasizes the need for greater awareness of dataset structures in LLM development. Furthermore, we introduce Pointwise Mutual Information (PMI) metrics, moving beyond traditional methods that analyze training distribution patterns and conditional probabilities. PMI allows us to assess implicit entity-concept associations, providing deeper insight into how biases in training data affect downstream model behavior. We simplify and validate the PMI-based bias calculation method by comparing male and female PMI values, creating a more accessible tool for assessing gender biases. To evaluate these biases in a realistic context, we adapt two tasks—forced masked-word completion and open-ended generation—that simulate real-world applications, such as hiring, where stereotype-driven biases are of particular concern. Additionally, we introduce a novel evaluation metric for open-ended generation, combining lexical and psychosocial analyses to uncover disparities in model outputs and behaviors. Finally, we demonstrate that position bias, particularly in tasks that probe preferences or choices, can mislead results unless the order of options is randomized, further reinforcing the importance of controlling for order effects in bias detection.

Therefore, this study aims to investigate five core research questions:

1. Are there frequency-based imbalances in the training dataset regarding stereotype categories, and are these significant enough to indicate downstream biases?

2. How are frequency-based imbalances reflected by Pointwise Mutual Information (PMI)?

3. How is PMI-associated bias reflected in a model's performance in forced masked-word completion?

4. Does the order of options in forced masked-word completion impact the bias exhibited?

5. How is PMI-associated bias exhibited in open-ended generation settings, and how does this compare with results from its forced choice counterpart?

2 Methodology

2.1 Experimental Setup

The study utilizes the Databricks Dolly 6B model, an instruction-tuned variant of the EleutherAI model, trained on the Dolly 15k dataset consisting of instruction-response pairs.

Bias Categories: The bias categories explored in this study include professions, gendered words, and stereotype concepts:

- **Professions:** A list of 50 professions, adapted from the US Bureau of Labor Statistics and widely referenced in NLP studies, including those by Caliskan et al., 2017 [2].

- **Gendered Words:** This category consists of male- and female-associated terms. Male-related words include terms like 'he', 'his', 'man', 'husband', 'son', and 'brother', while female-related words include 'she', 'her', 'woman', 'wife', 'daughter', and 'sister'.

- **Stereotype Concepts:** This category includes terms related to gender-stereotyped roles, such as career, family, and hobbies, based on the work of Nosek et al., 2009 [7] and widely used in bias detection research.

2.2 Bias Evaluation Tasks

2.2.1 Task 1: Forced Masked-Word Completion (FMC)

This task evaluates gender preferences by inserting placeholders for professions in sentences and prompting the model to select between gendered options (e.g., "he" or "she"). Rather than directly analyzing the responses, we count the occurrences of male and female terms and aggregate the results over multiple iterations.

2.2.2 Task 2: Open-Ended Generation

This task involves generating sentences based on templates replaced with different professions. The objective is to assess how the model handles gender biases when responding to prompts featuring male and female entities with the same structure.

2.3 Bias Evaluation Framework

Bias in the training dataset was measured using Pointwise Mutual Information, focusing on gendered associations between professions and male/female terms. For this, we calculated the conditional PMI for each profession relative to gendered terms "male" and "female." Specifically, we measured the PMI for male and female associations with each profession, computed as follows:

$$\text{PMI}_{\text{Male}} = \log\left(\frac{P(\text{Male}, C)}{P(\text{Male})P(C)}\right)$$

$$\text{PMI}_{\text{Female}} = \log\left(\frac{P(\text{Female}, C)}{P(\text{Female})P(C)}\right)$$

These conditional PMI values were then used to define the BiasPMI metric as the difference between the male and female PMI scores for each profession:

$$\text{BiasPMI} = \text{PMI}_{\text{Male}} - \text{PMI}_{\text{Female}}$$

The BiasPMI score quantifies the relative strength of gendered associations with professions, where positive values indicate stronger male associations, and negative values indicate stronger female associations.

For the Forced Masked-Word Completion (FMC) task, bias was evaluated by counting the occurrences of male and female terms in each generated output. These counts were aggregated across all task iterations, providing a BiasFMC metric that measures the frequency disparity between male and female terms. This metric was then compared with BiasPMI scores to assess the consistency of co-occurrence-based and frequency-based measures of bias.

In the Open-Ended Generation Task, a combination of lexical and psychosocial metrics was used to analyze gender disparities in the model's responses. The following metrics were applied:

Metric	Description	Measure Name
Word/Character Count	Measures verbosity of the response	BiasWordCount, BiasCharCount
Lexical Diversity	Ratio of unique words to total words	BiasLexicalDiversity
Affect Score (LIWC)	Measures the presence of affective language	BiasAffect
Toxicity Score (LIWC)	Evaluates harmful or toxic language	BiasToxicity

Table 1: Metrics for Open-Ended Generation Task

Statistical analyses included Pearson correlation to assess relationships between output metrics (e.g., BiasWordCount, BiasCharCount, etc.) and BiasPMI scores, and Ordinary Least Squares (OLS) regression for significance testing (F-scores and p-values). By comparing BiasPMI, BiasFMC, and generation task metrics, we ensured a holistic empirical analysis of lexical and psychosocial bias measurement across tasks.

3 Experiments

3.1 Experiment 1: Dataset Classification and Gendered-Entity Imbalances

We begin by grouping the rows of the training data, treating each combination of `Context`, `Instruction`, and `Response` as a single entry. The dataset is then filtered to focus on rows containing male or female entity words, reducing the dataset to 2,867 entries. Each entry is classified as male, female, or neutral based on the frequency of gendered entities.

Co-occurrences of male/female entities with a given category's attributes within each entry were detected without distinguishing between subject and object roles. This approach aligns with the idea that language models (LMs) may learn from spurious features in the training dataset, as they often rely on correlations that do not imply causation, such as word overlap or surface form matching (Simon, 1954; Gururangan et al., 2018; McCoy et al., 2019; Wallace et al., 2019; Kassner and Schütze, 2020; Poerner et al., 2020; Wang et al., 2022). These spurious features, while potentially helpful in generating plausible responses, can lead to biased or inaccurate semantic understanding (Fish, 2009; Maynez et al., 2020; Ji et al., 2023). We acknowledge that further research

is needed to explore the impact of these spurious correlations on bias.

To further investigate, we conducted probability tests on the dataset, examining both marginal and joint probabilities. We specifically looked at the joint probabilities between gendered entity words and commonly studied stereotype-related attributes, such as `Career`, `Family`, `Math`, `Arts`, `Sciences`, `Professions`, `Male-Stereotyped Hobbies`, `Female-Stereotyped Hobbies`, and `Emotional Valence` [**?**].

For the rest of the study, we decided to focus solely on `Profession`-related words, as this category exhibited one of the least skewed distributions (which allowed for fairest PMI comparison) and is highly relevant in current LLM use-cases such as career profiling, educational institutions, and legal examinations.

3.2 Experiment 2: PMI Analysis

Building on the analysis from Experiment 1, we first explore the relationship between gendered terms and bias categories through Pointwise Mutual Information (PMI).

We then focus on the list of professions, pairing each one with male and female entity words to calculate the PMI scores. This allows us to assess how strongly male and female terms are associated with each profession. The PMI for male and female terms is computed as follows:

$$PMI(w,t) = \log_e \left(\frac{P(w,t)}{P(w)P(t)} \right)$$

where $P(w,t)$ is the joint probability of a gendered term w and a profession term t, and $P(w)$ and $P(t)$ are the marginal probabilities of the gendered term and the profession term, respectively.

The results of the PMI calculation for various professions are shown in Figure 5, where the difference of PMI scores for male and female associations are compared. Positive BiasPMI values indicate a biased association between a profession and male entities, while negative values suggest a biased association to female entities.

3.3 Experiment 3: Forced Masked-Word Completion (FMC)

The Forced Masked-Word Completion (FMC) task aims to quantify gender bias in language model predictions by presenting the model with a binary choice between gendered pronouns. Sentences featuring professions were embedded in templated prompts, and the model was asked to select between "he" and "she." This controlled setup isolates and measures gender associations in the model's outputs, offering a systematic evaluation of potential biases.

The task involved variations on prompts using a base template where a placeholder for a profession was included. The template was structured as follows:

```
The [profession] was in charge of the event.
    ___ handled all the responsibilities
    well.
Choose the more likely option:
1. He handled all the responsibilities well.
2. She handled all the responsibilities well.
```

The placeholder `[profession]` was replaced with a list of professions of interest, and multiple iterations (10 per profession) were conducted to account for temperature-induced diversity in model predictions. For example, with the profession "Technician," the specific prompt was:

```
The technician was in charge of the event.
    ___ handled all the responsibilities
    well.
Choose the more likely option:
1. He handled all the responsibilities well.
2. She handled all the responsibilities well.
```

3.4 Experiment 4: Forced Masked-Word Completion with Control for Position Bias

To address the possibility that the order of presented options biases the model's predictions, we designed this extension of the Forced Masked-Word Completion (FMC) task to control for position bias. Position bias refers to a systematic tendency of language models to favor choices based on their sequential position in the input prompt. To mitigate this bias, we introduced randomization in the presentation order of gendered pronouns in the prompt templates. Two neutral prompts were created, differing only in the order of the pronoun options.

In this experiment, the placeholder `[word]` was replaced with a list of professions, as in Experiment 3. However, during each iteration, the model was randomly assigned one of the two templates to attempt at an equal likelihood of encountering either order. This design ensures that any observed gender association cannot be attributed to position bias, thereby strengthening the robustness of our analysis. By controlling for this confounding factor, we aim to isolate the model's intrinsic gender associations more reliably.

```
# Neutral prompt templates
neutral_prompt1 = (
    "The [word] was in charge of the event.
    ___ handled all the responsibilities
    well.\n"
    "Choose the more likely option:\n"
    "1. He handled all the responsibilities
    well.\n"
    "2. She handled all the responsibilities
    well.")

neutral_prompt2 = (
    "The [word] was in charge of the event.
    ___ handled all the responsibilities
    well.\n"
    "Choose the more likely option:\n"
    "1. She handled all the responsibilities
    well.\n"
    "2. He handled all the responsibilities
    well.")
```

3.5 Experiment 5: Open-Ended Generation

The Open-Ended Generation task examines gender bias in unconstrained model outputs when given prompts with gendered professions. The model generates responses based on templates that introduce a male or female profession. This experiment aims to measure how the model associates gender with various professions and explore any biases that arise in its outputs.

For each profession, 10 repetitions were performed for both the male and female templates.

```
# Template for male and female professions
male_template = "Complete the given phrase
    in only one sentence. Start with: The
    male [profession] "
female_template = "Complete the given phrase
    in only one sentence. Start with: The
    female [profession] "
```

Example output for the profession "technician":

```
Profession: technician
Male Prompt: The male technician
Male Output: The male technician was sitting
    at the control panel, looking frantic.
    "Did I
push the wrong button?" he asked.
Female Prompt: The female technician
Female Output: The female technician was
    unreasonably rude while assisting me
    with my account issues. She was totally
    uninformed about my account and did not
    help at all. I will never use their
    services again!
```

4 Results

4.1 Frequency-Based Imbalances

Experiment 1 revealed a significant imbalance in male and female classifications within the gendered entity subset. Figure 1 shows that male entries significantly outnumber female ones. Specifically, there were 1983 male entries, 708 female entries, and 176 entries that were classified as both male and female.

The marginal probabilities shown in Figure 2 reflect this imbalance, with male entities accounting for 0.566 and female entities 0.195 after tokenization and preprocessing. This suggests that male-related terms are more prevalent across gendered entries in the dataset.

Joint probabilities presented in Figure 3 between gendered terms and bias-related categories (e.g., Career, Professions, Male-Stereotyped Hobbies) show a clear dominance of male associations. However, due to the overrepresentation of male entries, these probabilities likely reflect dataset bias rather than true gender-category associations. For categories such as Arts and Emotional Valence, the differences in joint probabilities were smaller, indicating less gender bias in these areas.

4.2 PMI Analysis

The heatmap in Figure 4 explores the co-occurrence between male/female entities and stereotype-associated words across categories. It reveals which words are most commonly used in male- versus female-dominated contexts, reinforcing patterns observed in the joint probability plot. Career terms have the highest Pointwise Mutual Information (PMI) with male entities, while female-associated hobbies show the lowest PMI. Conversely, emotional and family-related words are strongly associated with female entities, while male-associated hobbies, math, and sciences are weakest for females.

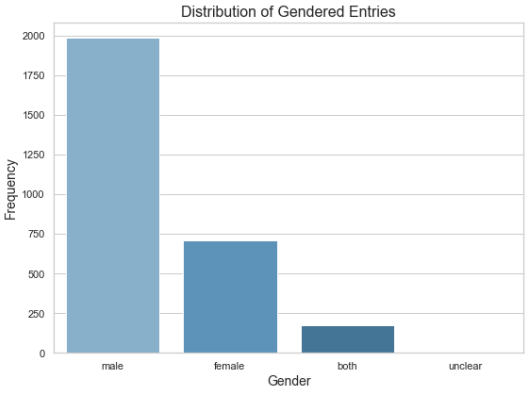


Figure 1: Distribution of Male and Female classified entries.

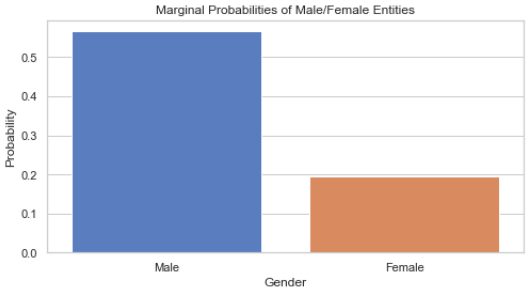


Figure 2: Marginal probabilities based on the total frequency of male and female entities, illustrating gender-based imbalances in training dataset features.

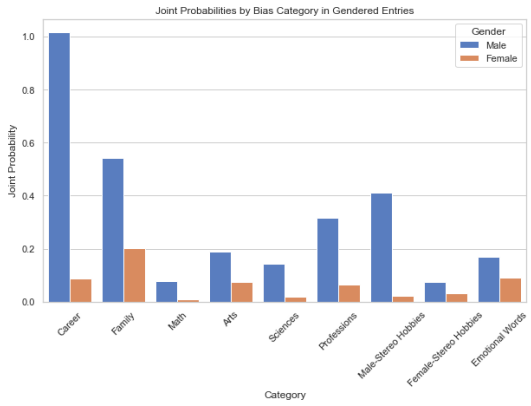


Figure 3: Joint probabilities of male and female entities co-occurring with categories, revealing potential biases in gender-term associations.

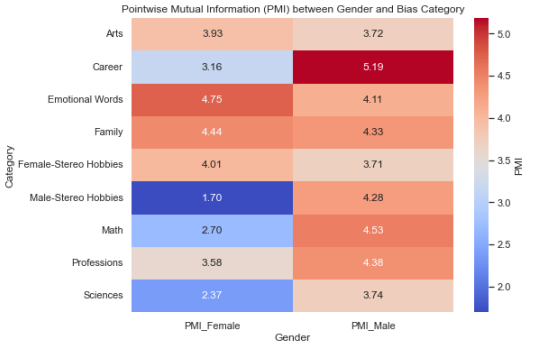


Figure 4: PMI Heatmap for Gender Entities and Categories, visualizing biased co-occurrence patterns and potential origins of stereotypes.

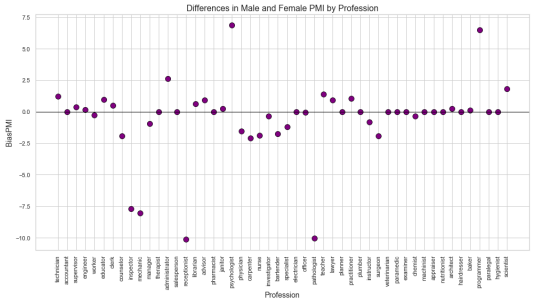


Figure 5: BiasPMI Scores for 50 Professions, highlighting the strength and direction of their association with Male/Female Entities in the training data.

4.3 Forced Masked-Word Completion (FMC)

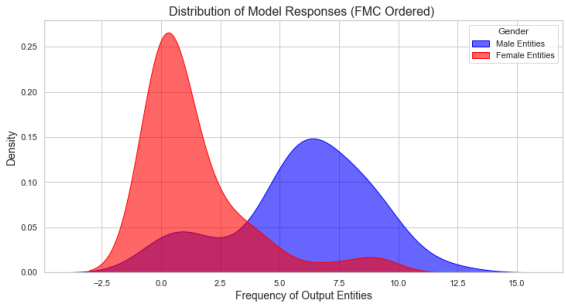


Figure 6: Kernel density estimate illustrating the distribution of model outputs for male and female entity choices in the Ordered Forced Masked-Word Completion task.

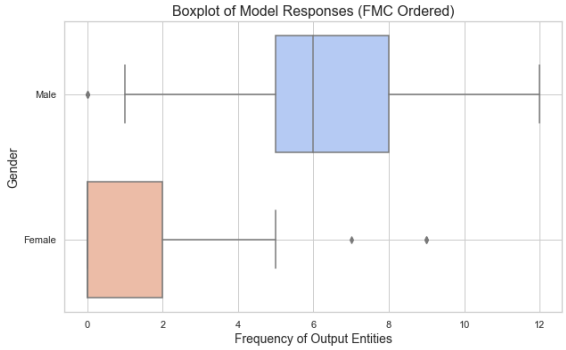


Figure 7: Boxplot visualizing the range and median of model outputs for male and female entity choices in the Ordered Forced Masked-Word Completion task.

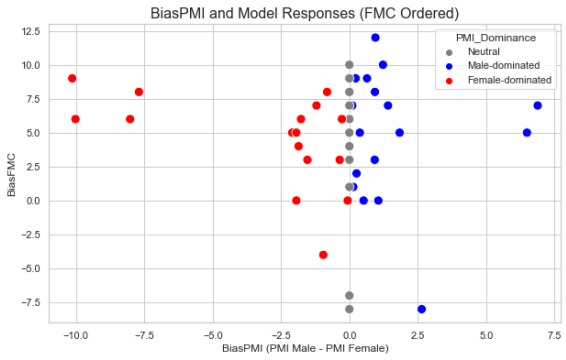


Figure 8: Scatter plot showing the relationship between BiasPMI and differences in output frequencies of male and female entities across professions. Points are color-coded by PMI dominance based on training data co-occurrences.

4.4 Order Bias in FMC

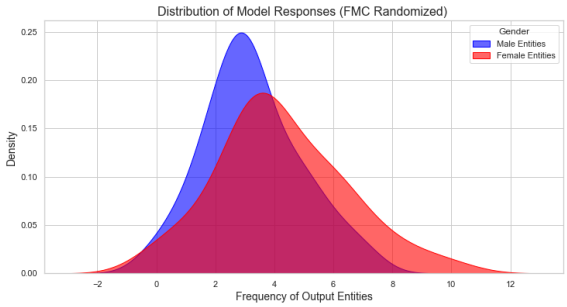


Figure 9: Kernel density estimate illustrating the distribution of model outputs for male and female entity choices (Randomized FMC task).

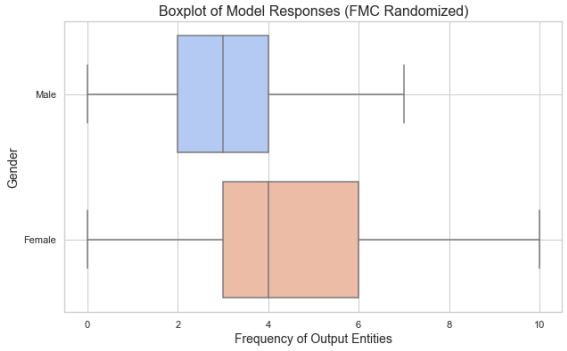


Figure 10: Boxplot visualizing the range and median of model outputs for male and female entity choices (Randomized FMC task).

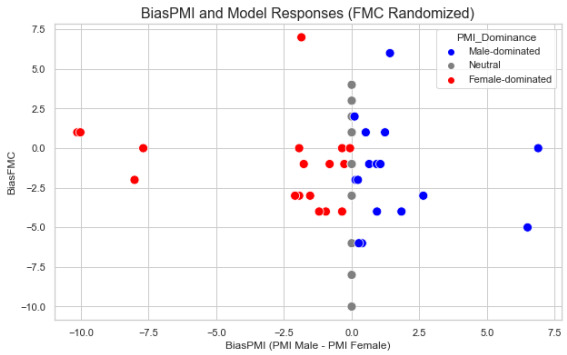


Figure 11: Scatter plot showing the relationship between BiasPMI and differences in output frequencies of male and female entities across professions (Randomized FMC task). Points are color-coded by PMI dominance based on training data co-occurrences.

Condition	Pearson (p-value)	Spearman (p-value)	F-Stat.	p-value
Ordered	-0.127 (p = 0.381)	0.043 (p = 0.768)	0.781	0.381
Randomized	-0.124 (p = 0.391)	-0.076 (p = 0.600)	0.748	0.391

Table 2: Correlation and OLS Regression Analysis for BiasPMI Metrics in Ordered and Randomized Conditions. P-values less than 0.05 are considered significant.

Metric	t-statistic	p-value
BiasCounts (Ordered)	7.271	2.53e-09*
BiasCounts (Randomized)	-2.122	0.039*

Table 3: One-Sample t-test for Count Difference in Ordered and Randomized FMC Conditions. P-values less than 0.05 are considered significant and are marked with a star.

Our analysis of Experiment 3 demonstrated a consistent skew in model outputs favoring male pronouns ("he") over female pronouns ("she") in the Forced Masked-Word Completion (FMC) task, irrespective of the direction of the BiasPMI metric for professions (Figures 6, 7, and 8). While the hypothesis of a linear relationship between BiasPMI and pronoun preference was proposed, it was falsified as no significant Pearson or Spearman correlations were observed in either the ordered or randomized FMC conditions (Table 2). Additionally, one-sample t-tests highlighted significant male skew in the ordered condition (t = 7.271, p ¡ 0.001) but weaker effects in the randomized condition (t = -2.122, p = 0.039) (Table 3). The findings from Experiment 4 provided further evidence against a consistent correlation between BiasPMI and pronoun outputs. Figures 9, 10, and 11 collectively disprove any relationship between the model's gendered outputs and the BiasPMI metric in the FMC task.

Specifically: 1. The Kernel Density Estimation (KDE) plot (Figure 9) shows overlapping distributions of male and female word counts, suggesting no systematic difference in output frequencies for randomized prompts. 2. The boxplot (Figure 10) highlights the variability in word counts for both genders, further demonstrating the absence of significant bias under randomized conditions. 3. The scatter plot of BiasPMI versus word count difference (Figure 11) fails to reveal any consistent pattern or alignment between PMI dominance (male/female/neutral) and the corresponding outputs.

Additionally, the male-skewed pronoun preference observed in the ordered FMC condition (Experiment 3) decreased in significance under the randomized condition of Experiment 4, even reversing direction to indicate a slight female preference. This shift underscores the confounding influence of positional effects in the ordered experiment and validates the robustness of the randomized design in mitigating such biases. Despite these refinements in methodology, the correlation analysis for Experiment 4 again revealed no significant Pearson or Spearman correlations (Table 2), further falsifying our hypothesis regarding a direct association to PMI dominance .

4.5 Open-Ended Generation

The results from the Open-Ended Generation task are presented as similar scatter plots as Experiments 3 and 4. Each point in the plots represents a profession, with its associated bias metric plotted against the Bias PMI score. For all the metrics analyzed, positive bias values indicate higher average levels in male-associated outputs, whereas negative values indicate higher average levels in female-associated outputs. As this is a generation task, preference order control was unnecessary.

Figures 12 and 13 illustrate these relationships for various metrics. Word and character counts, affect scores, and toxicity metrics did not show significant correlations with PMI dominance. These patterns are visually evident in Figure 12(a)-(c), where regression lines suggest negligible trends.

Interestingly, lexical diversity exhibited a slight negative Spearman correlation ($r = -0.2502$, $p = 0.0797$). As highlighted in Table 4, this trend, though not statistically significant, is visually apparent in Figure 13, where the regression line indicates a decreasing relationship. Given the subtle trend, further statistical analysis was conducted to better understand the data. To investigate whether male overrepresentation in PMI affected the results, we performed one-sample t-tests on the metrics. Table 5 summarizes the findings, revealing significant results for three out of five metrics:

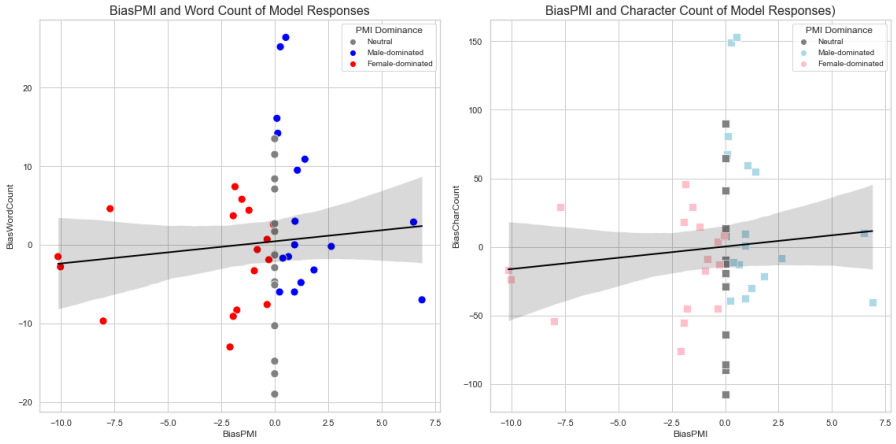
- Lexical Diversity: $t = -2.2373$, $p = 0.0298^*$ - Affect Scores: $t = -2.4370$, $p = 0.0185^*$ - Toxicity: $t = 2.4973$, $p = 0.0159^*$

Metric	Pearson Correlation	Spearman Correlation	OLS F-Statistic
BiasWordCount	0.0911 (p = 0.5292)	0.1730 (p = 0.2296)	0.4017 (p = 0.5292)
BiasCharCount	0.0924 (p = 0.5232)	0.1726 (p = 0.2306)	0.4136 (p = 0.5232)
BiasLexicalDiversity	-0.0919 (p = 0.5257)	-0.2502 (p = 0.0797)	0.4087 (p = 0.5257)
BiasAffect	0.0248 (p = 0.8645)	-0.0560 (p = 0.6995)	0.0294 (p = 0.8645)
BiasToxicity	0.0713 (p = 0.6226)	0.0263 (p = 0.8560)	0.2454 (p = 0.6226)

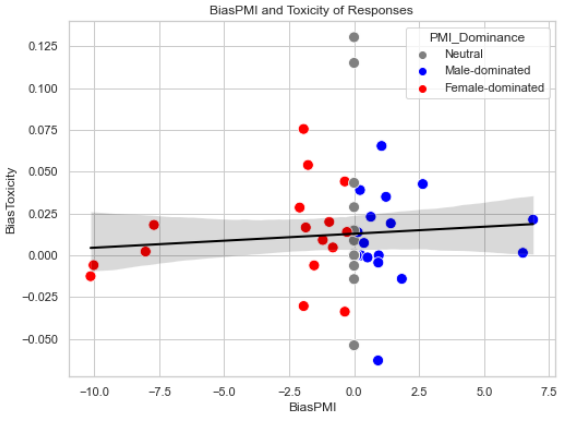
Table 4: Correlation and OLS Regression Analysis for Metrics against PMI Difference. P-values less than 0.05 are considered significant.

Metric	t-statistic	p-value
BiasWordCount	0.2343	0.8158
BiasCharCount	-0.0659	0.9477
BiasLexicalDiversity	-2.2373	0.0298*
BiasAffect	-2.4370	0.0185*
BiasToxicity	2.4973	0.0159*

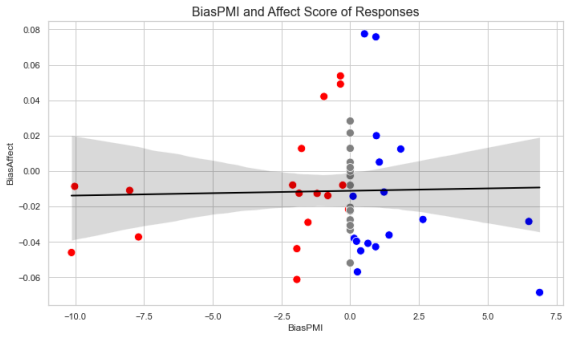
Table 5: One-Sample t-test for Difference Metrics against Null Distribution (Mean = 0). P-values less than 0.05 are considered significant and are marked with a star.



(a) BiasPMI and Word/Character Counts of Responses.



(b) BiasPMI and Toxicity of Responses.



(c) BiasPMI and Affect Score of Responses.

Figure 12: BiasPMI and Open-Ended Generation Metrics. Figures represent PMI difference and differences in (a) word/character count, (b) toxicity, and (c) affect scores. Points are color-coded by PMI dominance based on training data co-occurrences.

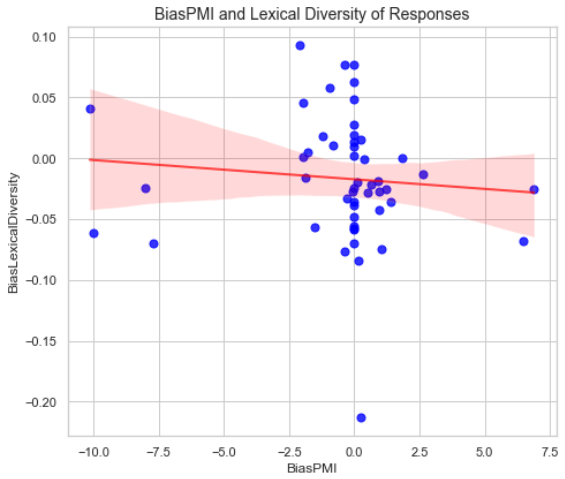


Figure 13: BiasPMI and Lexical Diversity of Responses. As evidenced by the higher concentration of points on the positive axis of BiasLexicalDiversity, male responses generally exhibit significantly higher lexical diversity scores compared to female responses. This highlights gender-based differences in output language richness.

These tests reveal systematic patterns in the outputs. Lexical diversity and affect scores were negatively skewed, indicating higher values in female-associated outputs, while toxicity showed a positive skew, reflecting elevated levels in male-associated outputs. The significant t-test results reinforce trends suggested by the scatter plots, with Figure 12(b) particularly emphasizing the differences in toxicity levels.

Overall, while correlation analyses largely suggested weak or non-significant relationships, the t-test findings highlight systematic disparities in output characteristics. The interplay between these results provides insights into the nuanced biases present in model outputs, which will be further explored in the discussion section.

Discussion

Our analysis uncovered significant frequency-based imbalances in the training dataset, which may have influenced model behavior as measured by Pointwise Mutual Information (PMI) metrics and task performance. Below, we discuss the implications of these findings in response to our research questions.

RQ1: Frequency-Based Imbalances in the Training Dataset

Simple conditional probability metrics effectively illustrated how frequency-based imbalances were embedded in the Dolly15k dataset. The significant overrepresentation of male-related entities, particularly in stereotype-seeded contexts, also perpetuated into their joint probabilities with categories (Careers, Male-stereotypical Hobbies, Math). This trend was less pronounced in categories like Arts, Female-stereotypical hobbies, and Emotional words, where female entities were more equitably represented. The wide variation between these categories in terms of their Joint Probabilities with Male and Female entities alludes to how dataset imbalances can lead to downstream biases.

RQ2: The Utility of PMI in Bias Detection

PMI emerged as a more powerful tool than joint probabilities for identifying dataset imbalances and biases. By isolating associations that appear disproportionately in male or female contexts, PMI revealed biases that joint probabilities couldn't detect or exaggerated. PMI's inherent statistical derivation—

$$PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$

allowed for a clearer distinction between stereotypical associations and general dataset imbalances, by accounting for individual probabilities. This nuanced perspective supports the adoption of PMI in future bias detection studies, especially when underlying imbalances risk overshadowing subtle but impactful associations. While it is difficult to compare against a definitive ground truth for bias evaluation, PMI's ability to persistently highlight associations after accounting for individual probabilities strengthens its utility.

RQ3: Forced Masked-Word Completion Task

In forced masked-word completion tasks, failing to control for positional biases led to misguided preferences for male options, with an average of 62% across all professions. A limitation of this approach is that we operationalized these preferences based on the classification of higher male/female entity counts. Moreover, the model frequently output both options, treating them as next-token predictions rather than making a binary decision. As a result, the count-based metric revealed broad patterns but lacked the granularity necessary for more detailed qualitative insights. This limited the task's ability to accurately classify choices or assess bias in a comprehensive manner.

RQ4: Positional Bias in Forced Masked-Word Tasks

Randomizing the original option order revealed a significant positional bias, with the model favoring the male option in only 32% of cases. This shift in the t-statistic from 7.271 to -2.122 suggests that transformer positional encoding may have skewed outcomes, reinforcing the model's tendency to treat prompts as continuation tasks rather than distinct comparisons. These findings align with Shi et al's research on *Judging the Judges* [9], which highlights how position bias affects metrics across tasks and varies with solution quality. Similarly, Li et al. in *Split and Merge* [6] on aligning position biases demonstrates that techniques like PORTIA, which segment and realign prompts, can improve evaluation consistency. These insights suggest that mitigating positional bias may require refining task structures or integrating alignment strategies to ensure fairer assessments.

RQ5: Open-Ended Generation Task

Open-ended generation tasks revealed slightly more pronounced biases, likely due to their ability to unravel inherent nuances and employ a broader range of metrics beyond simple counts. For instance, higher lexical diversity in male outputs and differing affect scores provided insights into subtle biases in the model's behavior. While correlations between metrics were not statistically significant, the task uncovered previously unobserved generation patterns, underscoring its value in bias analysis over Forced Choice Tasks. We discuss these further and provide explanations in our Unexpected Results section.

Contextualization of Results

These results align with prior research demonstrating that imbalanced training corpora can reinforce systemic stereotypes in language models. Recent studies have shown that the training data used to develop LLMs often reflect existing societal biases, which can manifest in unintended behaviors when models are deployed. For example, Kang and Choi (2023) [5] highlighted co-occurrence biases in model outputs, where certain words are more likely to appear together simply because they co-occur frequently in training data, rather than due to the underlying logic of the model. This can lead to gender imbalances in model responses, with certain gendered terms being more strongly associated with specific occupations, roles, or behaviors, perpetuating harmful stereotypes. Similarly, several frameworks, such as the Discovery of Correlations (DisCo) framework (Webster et al., 2020) [11] and datasets like RealToxicityPrompts (Gehman et al., 2020) [3], have been developed to assess biases in LLMs, providing real-world prompts to test biases in open-ended text generation. These methodologies help identify and quantify biases related to gender, race, and other social factors, enabling a more nuanced understanding of how LLMs handle sensitive topics.

This study extends existing knowledge by quantifying biases in the Dolly15k training dataset through the use of PMI metrics, providing a more precise tool for bias detection. Unlike prior work that often relied on anecdotal or qualitative analyses of bias, our structured approach demonstrates how metrics like PMI can operationalize bias detection and mitigation. By linking dataset characteristics to observable biases, our work adds a new layer of understanding to the complexities of bias in language models.

4.6 Unexpected Results

Two unexpected findings emerged from our analysis. First, none of the correlations tested were significant, with neither Spearman nor Pearson coefficients exceeding an absolute value of 0.1. This suggests that linear and non-linear relationships between metrics and biases were not discernible in our dataset. While ordinary least squares (OLS) regression did not uncover meaningful relationships, exploring alternative measures like covariance or Kullback-Leibler (KL) divergence may provide deeper insights.

Second, the significantly higher lexical diversity and affect scores in female outputs across all professions were surprising, given the lower representation of female entities in the dataset. One possible explanation is that the model, having less information on female-related entities, compensates by relying more on its pretraining knowledge, generating diverse and emotionally rich content. This overcompensation may result in higher variability in lexicon for female outputs. Higher affect scores could also be attributed to the stronger PMI associations between female entities and emotional-valence words, as observed in the heatmap analysis. This aligns with prior findings that emotional language is more frequently associated with female contexts, reinforcing stereotypical portrayals. These findings highlight the complex interplay between dataset imbalances and model behavior, emphasizing the need for nuanced interpretations of bias metrics.

4.7 Limitations

While the findings of this study provide valuable insights, there are notable limitations that must be considered.

4.7.1 Sampling Bias

The analysis was based on a sample of dataset rows specifically classified as containing gender-related entities. While this allowed for focused evaluation, it may not represent the full diversity of the instruction-tuning dataset. Sampling constraints could lead to an overrepresentation of certain associations while underestimating others, limiting the generalizability of the findings.

4.7.2 Semantic Matching Limitations

Professions were searched for by exact name rather than including synonyms or related terms. As a result, some associations may have been overlooked, leading to potentially diminished PMI scores for certain professions. This limitation introduces the risk of underestimating the prevalence and strength of certain stereotypes, as synonyms and contextually equivalent terms may not have been accounted for.

4.7.3 Task Design Constraints

The forced masked-word completion task, while useful for measuring model behavior, was not strictly a preference-selection task. The model's output was unconstrained, allowing it to generate multiple potential responses. This design limits the interpretability of the results as a strict measure of preference or ranking. Future studies could implement more controlled experimental setups to better isolate the effects of bias.

4.7.4 Dataset Scope

This study focused exclusively on the instruction-tuning dataset, which is designed to optimize task-specific instruction-following behavior. While this provides valuable insights into the relationship between context, instruction, and response, it does not capture language-level patterns learned during pre-training. The biases identified in this study may interact with or differ from those embedded in the pre-training dataset. A broader analysis encompassing both datasets would offer a more comprehensive understanding of how biases propagate through different stages of model development.

4.8 Implications and Future Directions

These findings highlight the significant impact of dataset biases on model behavior and emphasize the need for systematic mitigation strategies. Future research should explore more representative dataset curation, architecture-aware adjustments such as layer- and weight-level rebalancing, and targeted debiasing interventions. While this study focused on instruction-tuning datasets, the findings underscore the importance of evaluating biases comprehensively across pretraining, fine-tuning, and deployment stages.

Additionally, a more rigorous framework for evaluating open-ended generation tasks is needed. While the metrics used in this study are a starting point, incorporating similarity measures based on topic modeling could offer deeper insights into a model's ability to handle both seen and unseen co-occurrences, improving evaluation.

Further investigation into how biases in categories like emotional valence, careers, and academic fields are propagated would also be valuable. Exploring biases at a broader category level, rather than just seed-level, can provide a fuller understanding of how these biases manifest and interact across different contexts. Examining the impact of varying PMI across such categories could reveal more about how biases influence model outputs.

5 Conclusion

This research highlights the critical role of dataset composition in shaping language model outputs. Statistically detected imbalances in training data were shown to have irregular and unpredictable impacts on model behavior, including the reinforcement of stereotype-aligned associations. The use of Pointwise Mutual Information (PMI) as a diagnostic tool proved effective in quantifying and contextualizing these biases within the broader context of natural language understanding.

Ultimately, mitigating biases in AI systems is essential to ensure fairness, inclusivity, and equity in their applications. By bridging the gap between dataset-level imbalances and downstream effects, this research provides actionable insights for developing more equitable language models and lays a foundation for future studies aimed at fostering responsible AI development.

References

[1] Maria Antoniak and David Mimno. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online, 2021. Association for Computational Linguistics.

[2] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[3] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, 2020. Association for Computational Linguistics.

[4] Yue Huang, Qihui Zhang, Philip S. Yu, and Lichao Sun. Trustgpt: A benchmark for trustworthy and responsible large language models, 2023.

[5] Cheongwoong Kang and Jaesik Choi. Impact of co-occurrence on factual knowledge of large language models, 2023.

[6] Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. Split and merge: Aligning position biases in llm-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[7] Brian A. Nosek, Frederick L. Smyth, Nisreen Aboud, et al. National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26):10593–10597, 2009.

[8] Debora Nozza, Federico Bianchi, and Dirk Hovy. Honest: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online, 2021. Association for Computational Linguistics.

[9] Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms, 2024.

[10] Francisco Valentini, Germán Rosati, Damián Blasi, Diego Fernandez Slezak, and Edgar Altszyler. On the interpretability and significance of bias metrics in texts: a pmi-based approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 509–520, Toronto, Canada, 2023. Association for Computational Linguistics.

[11] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pretrained models, 2021.