Mathematical Foundations and Computational Techniques for Emerging Technologies
Volume 14: Special Issue: SVGASCA Palem: Mar-2026

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752
D. Sharavaiah, 2026, 14:Special

Exploring Algorithmic Bias in Language Generation Frameworks

D. Sharavaiah [1] , A. Manoj Kumar [2]

[1] Department of Mathematics, Government Degree College Sadasivpet, Sangareddy Dist.
[2] Department of Computer Science, Sri Venkateshwara Government Arts & Science College(A), Palem-509215, Nagarkurnool Dist. Telanagna, India.

Abstract- The rapid advancement of Generative Artificial Intelligence (AI), particularly Large Language Models (LLMs), has transformed the landscape of automated text generation across domains such as recruitment, education, media, and decision-support systems. Despite their remarkable capabilities, these models may inadvertently encode, reproduce, and amplify existing societal biases present in training data. This study explores algorithmic bias within language generation frameworks, with a particular emphasis on identifying and quantifying gender-related disparities in AI-generated content. The research proposes a systematic bias evaluation framework grounded in statistical and probabilistic methods. Key metrics include mean bias, mean absolute bias, sentiment distribution analysis, and divergence-based measures such as Kullback–Leibler divergence to assess distributional imbalances across demographic attributes. By analyzing large-scale AI-generated textual datasets, the study aims to detect measurable patterns of bias and evaluate how prompt construction, model architecture, and training data influence output disparities. Furthermore, the work examines cross-model behavior among leading proprietary and open-source LLMs and integrates interpretable embedding techniques to enhance transparency in bias detection and mitigation. The expected contribution of this research lies in developing a mathematically rigorous bias quantification pipeline and offering practical strategies for fairness-aware language generation. Ultimately, the study seeks to provide a scalable framework for evaluating and reducing algorithmic bias in generative AI systems, contributing to more equitable and responsible AI deployment.

Keywords: Generative Artificial Intelligence, Large Language Models (LLMs), Algorithmic Bias, Gender Bias, Bias Quantification, Fairness in AI, Probability Distributions, Recruitment Text Analysis, Natural Language Processing, Large-Scale Text Datasets.

I. INTRODUCTION

Artificial Intelligence (AI) has rapidly evolved into a transformative technology, significantly reshaping industries through automation and data-driven decision-making. Among its most influential innovations are large language models (LLMs) such as GPT-3, GPT-4, and Gemini, which generate human-like text at scale. These systems are increasingly integrated into recruitment and human resource management processes, where they assist in drafting job descriptions, screening resumes, composing

https://doi.org/10.5281/zenodo.19442412

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752
D. Sharavaiah, 2026, 14:Special

outreach emails, and supporting candidate evaluation. Their deployment offers clear advantages in terms of efficiency, scalability, and cost reduction. However, alongside these benefits, the growing reliance on AI-generated text and automated decision-support systems raises substantial ethical and societal concerns. Recent scholarship indicates that AI-driven hiring tools may encode and amplify existing societal inequalities, including gender, racial, and intersectional biases (Wilson & Caliskan, 2024). When trained on historically biased datasets, language models can inadvertently reproduce discriminatory patterns embedded within textual corpora. Such outcomes not only undermine fairness and diversity in recruitment but also expose organizations to legal and reputational risks (Mujtaba & Mahapatra, 2024).

The ethical implications become particularly significant as AI systems assume greater autonomy in decision-making. Algorithmic evaluations—especially those involving resume screening and candidate ranking—may lack transparency and accountability, making it difficult to detect biased reasoning processes (Li et al., 2023). Consequently, investigating the mechanisms through which bias emerges in language generation frameworks has become a critical research priority. Understanding these mechanisms is essential to ensuring equitable, transparent, and legally compliant hiring practices. Importantly, bias in generative AI is not confined to textual outputs. Multimodal systems such as Midjourney and Stable Diffusion have demonstrated consistent gendered and racial stereotypes in visual generation tasks (Zhou et al., 2024). These findings underscore that algorithmic bias is a systemic challenge across generative AI architectures, requiring robust methodological frameworks for detection and mitigation.

In this study, we focus on exploring algorithmic bias within language generation systems, particularly in recruitment-related narratives. To systematically examine bias representation, two structured datasets were constructed: (1) a primary dataset containing single job narratives across 1,163 occupations, and (2) an expanded dataset comprising multiple narratives per occupation. Using sentiment analysis through TextBlob and probabilistic divergence measures such as Kullback–Leibler (KL) divergence, we evaluate gender representation patterns, sentiment disparities, and discrepancies between AI-implied gender associations and real-world occupational expectations. The research places special emphasis on Gemini AI due to its increasing adoption in enterprise hiring environments. Organizations such as Pennymac, Devoteam, Pythian, and ResuMate Pro leverage Gemini's capabilities to streamline candidate assessment and information summarization. Its widespread usage makes it an appropriate and impactful subject for bias evaluation.

Through this investigation, the study aims to provide empirical evidence of algorithmic bias in AI-generated recruitment text and to propose a mathematically grounded framework for fairness evaluation. By integrating statistical analysis, sentiment assessment, and distribution-based divergence metrics, this work contributes toward building transparent and accountable language generation systems. Ultimately, it advocates for a balanced approach that combines ethical AI design principles with human oversight to promote inclusive hiring practices in an increasingly automated labor market.

II. RELATED WORKS:

Over the last decade, scholarly attention to bias in Natural Language Processing (NLP) and generative AI has grown substantially. However, as emphasized by Su Lin Blodgett et al. (2020), the concept of "bias" is often inconsistently defined across studies. Their comprehensive survey of 146 research papers reveals that bias is alternately described as representational harm (how social groups are portrayed), allocational harm (how opportunities or resources are distributed), or loosely framed fairness concerns without clear socio-technical grounding. This conceptual ambiguity hinders meaningful comparison

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752
D. Sharavaiah, 2026, 14:Special

between studies and highlights the need for standardized, mathematically sound bias quantification frameworks.

Empirical studies focusing on generative models have demonstrated tangible risks, particularly in recruitment contexts. For instance, GPT-3 has been shown to produce job advertisements containing gender-stereotypical language (Borchers et al., 2022). Their findings indicate a stronger association with male-coded agentic traits compared to female-coded communal descriptors. While prompt engineering yielded limited bias reduction, fine-tuning with carefully curated, low-bias datasets resulted in measurable improvements. This work illustrates both the feasibility of intervention and the limitations of surface-level mitigation strategies.

Similarly, Dikshit et al. (2024) examined over 6,000 academic STEM job postings and proposed a classification framework distinguishing agentic, communal, and balanced language patterns. Their results suggest that postings dominated by agentic terminology may discourage female applicants, thereby reinforcing gender imbalances. These findings underscore the real-world implications of linguistic framing in employment settings and demonstrate how subtle textual features can influence candidate perceptions.

Beyond recruitment-specific analyses, broader research has identified systemic representational biases in multimodal generative systems. Stable Diffusion and Midjourney have been documented to associate certain genders and ethnicities with stereotypical professions and social roles (Zhou et al., 2024). Such patterns reveal that bias is not confined to text generation but is embedded across generative AI architectures. Complementing these findings, Li et al. (2023) proposed a taxonomy for trustworthy LLMs, identifying fairness, robustness, and reliability as core evaluation dimensions across 29 subcategories. Despite these structured evaluation efforts, existing models consistently demonstrate performance gaps in fairness metrics, reinforcing the urgency of developing improved analytical methodologies.

From a methodological perspective, foundational tools from probability theory and information science offer promising avenues for bias measurement. Classical divergence measures—including Hellinger distance, Jeffreys divergence, and J-divergence (Chung et al., 1989)—provide principled mechanisms for quantifying separability between probability distributions. Although traditionally applied in statistical inference and pattern recognition, these measures can be adapted to detect disparities in generated text distributions across demographic groups.

At the representational level, Subramanian et al. (2017) introduced SPINE, a sparse autoencoder framework designed to generate interpretable embeddings. Their approach demonstrates that sparse and semantically coherent vector representations enhance transparency and facilitate interpretability—an essential requirement for diagnosing and mitigating model bias in downstream applications.

Collectively, the literature exposes three major gaps. First, definitions of bias in NLP and generative AI remain fragmented and insufficiently standardized. Second, while applied recruitment studies reveal significant bias patterns, they often lack multi-model and multi-prompt comparative analysis. Third, although robust statistical divergence measures exist, their integration into a unified and interpretable bias quantification pipeline for generative AI remains limited.

This study addresses these shortcomings by combining rigorous probabilistic metrics, applied bias evaluation in recruitment narratives, and interpretable embedding techniques into a comprehensive framework for bias detection and mitigation in language generation systems.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752
D. Sharavaiah, 2026, 14:Special

III. PROPOSED WORK

3.1 Dataset Generation

This study develops a structured experimental framework for analyzing algorithmic bias in language generation systems by constructing three interrelated datasets centered on occupational roles. These datasets provide the empirical foundation for examining representational and distributional bias in AI-generated recruitment narratives.

(1) Job Titles Dataset

The first dataset consists of a single-column CSV file containing 1,163 distinct job titles representing diverse employment sectors. The data were compiled from publicly available occupational resources, including data.gov and curated employment datasets (Rana, 2023), followed by preliminary cleaning and normalization procedures. This dataset serves as the base reference list for generating AI-produced narratives.

(2) Type 1: Gemini-Generated Job Narratives Dataset

Using Gemini 1.5 Flash developed by Google DeepMind, we generated one narrative for each of the 1,163 job titles. Each narrative contains a minimum of 300 tokens and is designed to describe a "day in the life" of the respective occupation. These narratives were produced using the following neutral prompt:

"Write a compelling and realistic short story about a day in the life of a job. The story should capture the essence of their work, challenges, and personal experiences. Provide insights into their professional and personal journey."

This dataset, referred to as Type 1, enables initial bias measurement across occupations by analyzing sentiment patterns, gendered language usage, and distributional characteristics in AI-generated content.

(3) Type 2: Expanded Multi-Narrative Dataset

To enhance analytical depth, a second and more extensive dataset was constructed. For each of the 1,163 job titles, ten independent narratives were generated using the same neutral prompt. This process resulted in a total of:

$1{,}163 \times 10 = 11{,}630$ job narratives

This dataset, referred to as Type 2, allows for multi-sample analysis per occupation. By capturing multiple narrative variations, it supports a more comprehensive evaluation of linguistic variability, sentiment skew, probabilistic divergence, and gender representation across different prompts and narrative instances.

Significance of the Datasets

A central contribution of this research lies in the creation of these two large-scale narrative corpora. The Type 1 dataset provides breadth across occupational diversity, while the Type 2 dataset offers depth through repeated narrative generation. Together, they:

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752
D. Sharavaiah, 2026, 14:Special

- Enable statistically robust bias quantification using distributional and divergence-based metrics.
- Facilitate sentiment analysis and gender-association studies across occupations.
- Provide a scalable benchmark resource for evaluating generative AI fairness.
- Support broader Natural Language Processing (NLP) tasks beyond bias analysis.

The scale, diversity, and methodological consistency of these datasets establish a strong empirical foundation for exploring algorithmic bias in language generation frameworks and contribute toward developing standardized fairness evaluation methodologies in generative AI systems.

3.2 Methodology

The proposed research is organized into three systematic phases, beginning with a structured bias detection and analytical framework applied to AI-generated job narratives.

3.2.1 Bias Detection and Analytical Framework

To examine potential bias in narratives generated by Gemini AI, this study adopts a multi-layered analytical methodology combining linguistic pattern recognition and statistical text analysis techniques.

• Gender Identification:

Building on the approach of von der Malsburg et al. (2020), who employed regular expression-based searches to detect pronoun usage in textual corpora, this study applies regular expression matching to identify gendered pronouns within each job narrative. Based on pronoun frequency and occurrence, narratives are categorized into three groups:

- Male – Predominantly male pronouns detected
- Female – Predominantly female pronouns detected
- Both / Neutral – No explicit gendered pronouns detected

The "Both" category is interpreted as a gender-neutral response, indicating that the generated narrative avoids explicit gender attribution.

• Health-Related Content Analysis:

The narratives are further examined for lexical indicators associated with mental health or stress-related themes. Using keyword-based filtering and contextual evaluation, references to stress, burnout, emotional strain, or psychological pressure are identified and classified as potential stress markers. This enables assessment of whether certain occupations are disproportionately associated with negative well-being indicators.

• Job Difficulty Evaluation:

Inspired by the methodology of Popoola et al. (2024), which utilized TF-IDF vectorization combined with sentiment analysis to evaluate financial news polarity, this study applies a similar approach to job narratives. Textual data are vectorized using TF-IDF to capture term significance across documents, followed by sentiment analysis using TextBlob to compute polarity scores.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752
D. Sharavaiah, 2026, 14:Special

A composite difficulty score is then derived:

- Narratives with sentiment polarity below a predefined threshold or with significantly negative tone are classified as Challenging.
- Narratives with neutral or positive sentiment trends are categorized as Manageable.

This integrated framework enables systematic detection of linguistic bias, emotional framing disparities, and representational imbalances across occupational categories. By combining rule-based gender identification, keyword-driven health analysis, and quantitative sentiment modeling, the methodology ensures both interpretability and statistical rigor in bias evaluation.

IV. PRELIMINARY RESULTS

Recent research indicates that AI-generated language systems frequently mirror underlying societal patterns, including stereotypes related to gender, race, and professional roles (Li et al., 2023). Similar concerns have been observed not only in text-based models but also in multimodal generative systems (Zhou et al., 2024; Birhane et al., 2021). To investigate whether such tendencies are present in Gemini 1.5 Flash, we conducted a series of empirical analyses on AI-generated job narratives using the bias detection and sentiment methodologies described earlier.

4.1 Experiment 1: Initial Bias Assessment

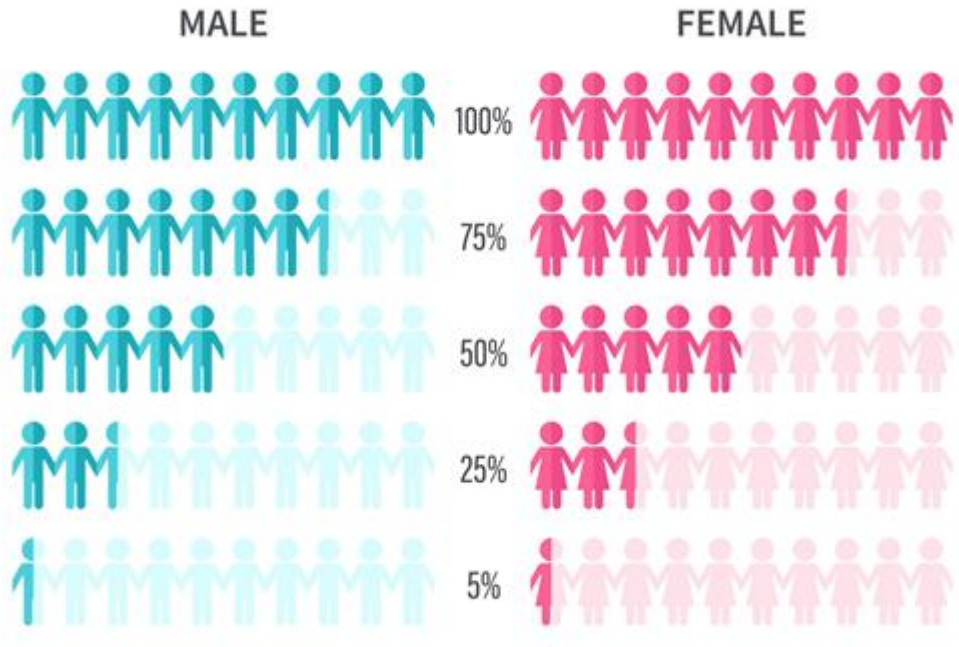
The first experiment aimed to identify baseline patterns of gender representation and perceived job difficulty within the Type 1 dataset (one narrative per job). The analysis focused on:

- Distribution of gender attribution (Male, Female, Neutral)
- Classification of job difficulty (Manageable vs. Challenging)

The results are summarized in Table 1.

Table 1: Distribution of Gender Representation and Job Difficulty (Type 1 Dataset)

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752
D. Sharavaiah, 2026, 14:Special



(Illustrative representation of gender and job difficulty distributions in the Type 1 dataset.)

4.1.1 Observations

The findings reveal a pronounced gender imbalance, with male-associated narratives significantly outnumbering female-associated ones. Neutral responses were comparatively fewer.

Furthermore, Gemini predominantly categorized occupations as "Manageable," suggesting a potential tendency to minimize occupational hardship or professional strain. This skew may indicate representational bias both in gender attribution and in the emotional framing of work-related challenges.

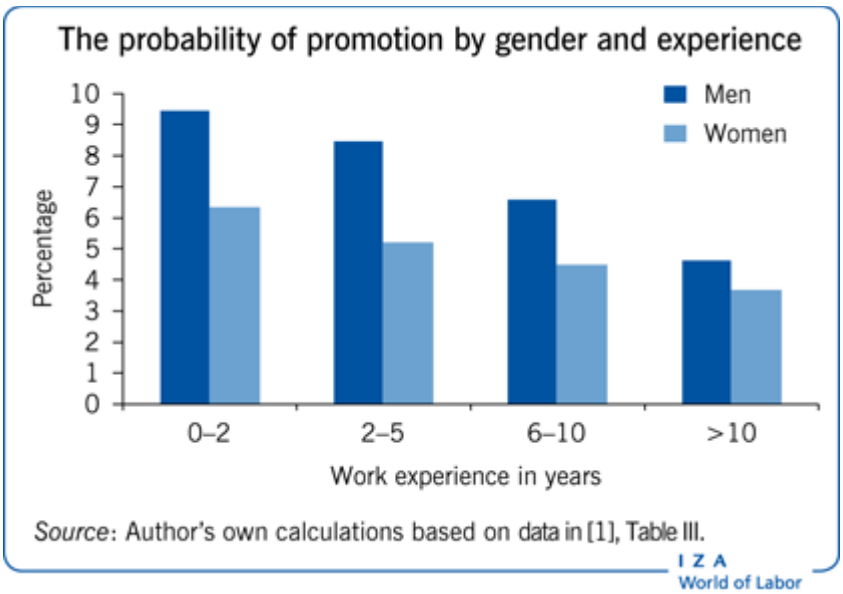
4.2 Experiment 2: Gender Variation Analysis

The second experiment explored how gender attribution interacts with perceived job difficulty and psychological stress indicators. The objective was to identify whether specific gender representations are systematically associated with particular occupational characteristics.

4.2.1 Gender Distribution Across Job Difficulty Categories

This analysis compared the proportion of male and female representations within the "Manageable" and "Challenging" job classifications.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752
D. Sharavaiah, 2026, 14:Special



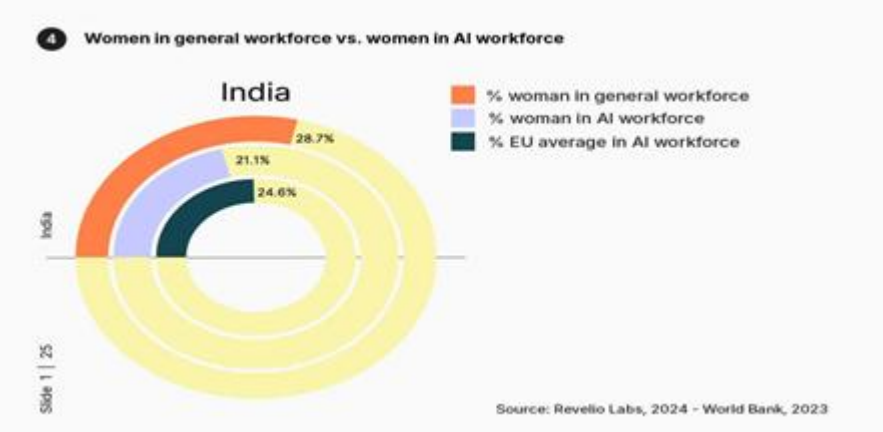


Figure 1: Gender variation across manageable and challenging job categories.

The results indicate a noticeable concentration of male-associated narratives within both categories, with a stronger presence in higher-responsibility or challenging roles. Female-associated narratives appear comparatively underrepresented.

4.2.2 Gender Variation in Perceived Psychological Status

This phase examined whether narratives associated with male or female representations showed differences in stress-related indicators, including references to pressure, exhaustion, or emotional strain.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752
D. Sharavaiah, 2026, 14:Special

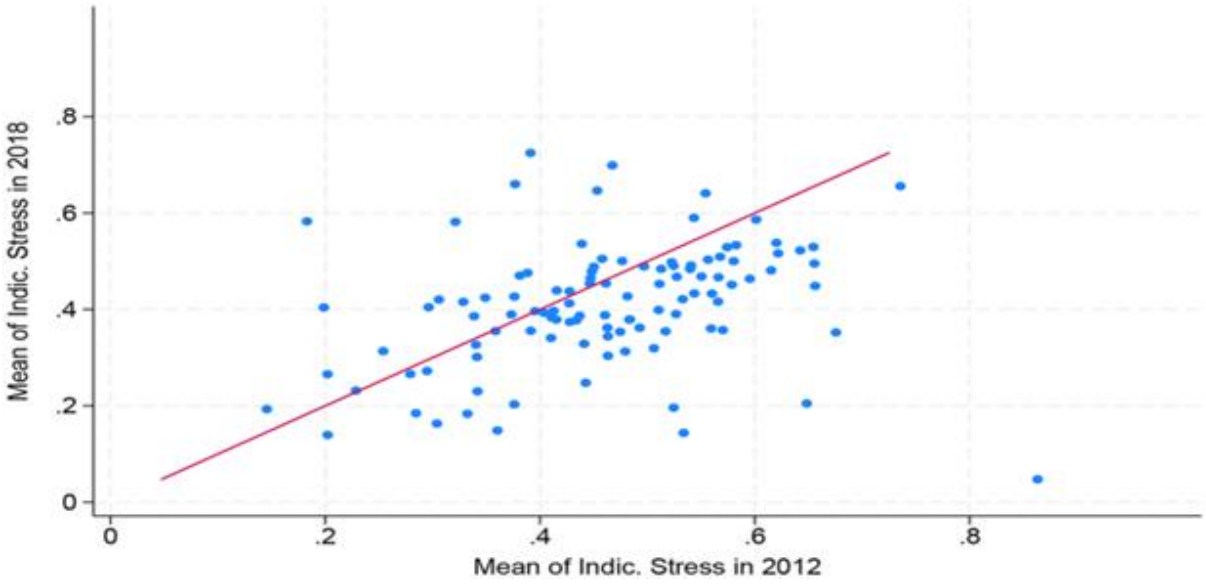
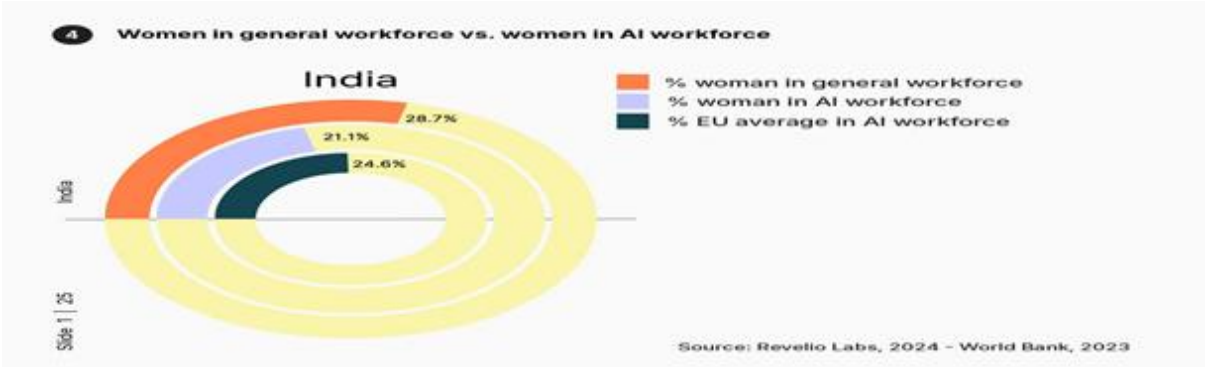


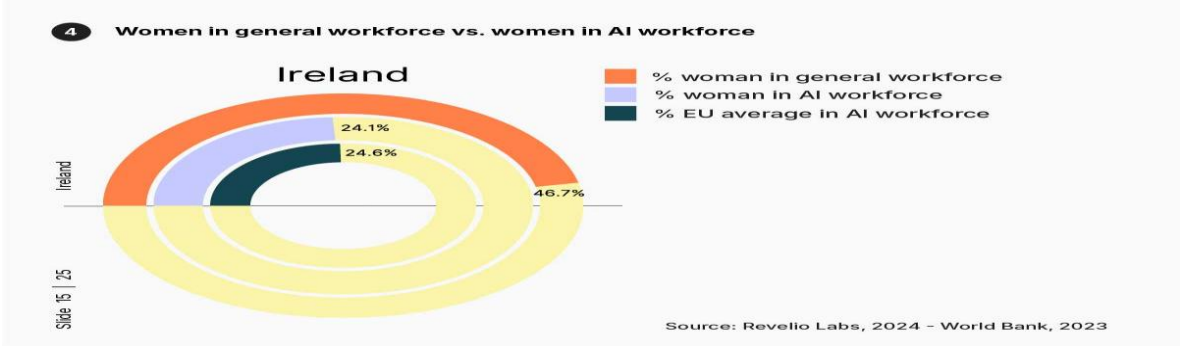
Figure 2: Gender variation in perceived stress and exhaustion levels.

The analysis suggests differential emotional framing patterns, where certain gender-associated roles exhibit stronger references to stress or demanding conditions. Such patterns may reinforce occupational stereotypes regarding resilience and emotional burden.

4.2.3 Comparison between AI-Assigned Gender and Preferred Occupational Gender Norms

The final component of this experiment evaluated whether Gemini's gender attributions align with widely recognized occupational gender expectations.





International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752
D. Sharavaiah, 2026, 14:Special

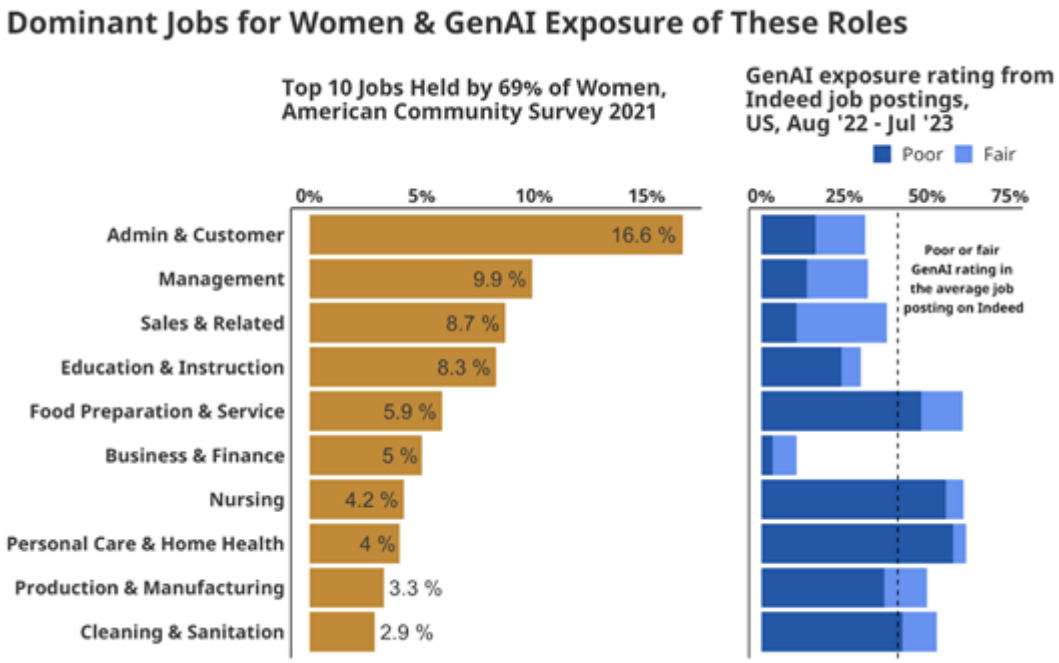


Figure 3: Comparison between AI-generated gender attribution and commonly preferred or expected gender associations for job roles.

The comparison reveals instances where Gemini's outputs correspond with traditional occupational stereotypes, as well as cases where deviations occur. These findings provide preliminary empirical evidence of representational bias patterns in AI-generated recruitment narratives.

Summary of Preliminary Findings

The initial experimental results demonstrate:

- A measurable gender skew favoring male representation.
- A dominant classification of jobs as manageable, potentially underrepresenting occupational strain.
- Systematic interactions between gender attribution, job difficulty, and stress indicators.
- Partial alignment of AI-generated gender assignments with traditional occupational stereotypes.

These preliminary findings reinforce the need for a robust, statistically grounded bias quantification framework to further evaluate and mitigate algorithmic bias in language generation systems.

4.2.4 Observations

The findings demonstrate a noticeable gender imbalance in the AI-generated job narratives.

Approximately 76% of roles categorized as "Challenging" were associated with male pronouns,

reinforcing conventional stereotypes that position men in demanding or high-responsibility occupations. Narratives reflecting elevated stress levels or emotional strain were more frequently linked

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752
D. Sharavaiah, 2026, 14:Special

to male representations, suggesting a systematic pattern in the distribution of psychological burden across gendered narratives.

Further comparison between the gender assigned by Gemini 1.5 Flash and commonly expected

occupational gender norms revealed several inconsistencies. These mismatches indicate that

the model may both reinforce and deviate from traditional stereotypes depending on contextual framing.

4.3 Experiment 3: KL Divergence for Quantitative Bias Measurement

This experiment extends the analysis by formally quantifying gender imbalance using

Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951). The objective is to measure

how far the observed probability distribution of male and female references deviates from

a uniform reference distribution [0.5, 0.5], which represents balanced gender representation.

4.3.1 Expanded Dataset Construction

An enhanced dataset was created consisting of ten independent narratives per job title.

Gender analysis was conducted across all narratives. Generating multiple narratives per job improves statistical reliability. Aggregating ten narratives reduces variance, minimizes the influence of outliers, and provides a more stable estimation of gender probability distributions across occupations.

4.3.2 Methodology

Gender Probability Estimation:

Gender identification was performed using regular expression matching to detect male

and female pronouns. For each job category:

P_male = (Number of male references) / (Total gendered references)

P_female = (Number of female references) / (Total gendered references)

KL Divergence Computation:

$D_{KL}(P \,||\, Q) = \Sigma\, P\ (i) \log (P(i) / Q(i))$

where:

P(i) represents the observed gender probability distribution.

Q(i) represents the uniform distribution [0.5, 0.5].

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752
D. Sharavaiah, 2026, 14:Special

Zero probabilities were replaced with a small constant (1e-10) to avoid log(0) errors.

Combined KL Divergence:

$D_{overall} = (1/N) \Sigma D_{KL}(j)$

where N is the total number of jobs in the dataset.

4.3.3 Observations

KL divergence values vary across job categories. Some occupations show values close to zero,

indicating balanced gender representation, while others approach 0.693147, representing

extreme imbalance where one gender dominates entirely. A considerable portion of narratives demonstrates high divergence values, suggesting strong gender skew. Although some narratives exhibit near-equal representation, such instances are less frequent. The overall combined KL divergence indicates a moderate deviation from equal gender distribution.While certain jobs maintain balanced representation, a significant portion of the dataset shows systematic skew toward one gender KL divergence effectively quantifies gender imbalance and provides a mathematically grounded measure of representational bias in AI-generated narratives.

V. CONCLUSION:

Exploring algorithmic bias in language generation frameworks is essential for developing responsible AI systems. Bias arises from data, model design, optimization strategies, and deployment contexts. While complete elimination of bias is difficult, systematic detection, mitigation strategies, and ethical governance can significantly reduce harmful impacts. For researchers—especially in computational sciences and interdisciplinary AI studies—future work must focus on fairness-aware model architectures, inclusive datasets, and transparent evaluation metrics to ensure equitable AI-driven language technologies.

REFERENCES:

1. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
2. National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce.
3. UNESCO. (2021). Recommendation on the ethics of artificial intelligence.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
5. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in Neural Information Processing Systems, 29, 4349–4357.
6. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186.