

Equilibrium Dynamics and Mitigation of Gender Bias in Synthetically Generated Data

**Ashish Kattamuri^{†1}, Arpita Vats^{‡2}, Harshwardhan Fartale^{§3},
Rahul Raja^{†2}, Akshata Kishore Moharir^{†3}, Ishita Prasad^{†4}**

¹Proofpoint * ²LinkedIn ³Independent Research ⁴Microsoft ⁵Meta FAIR

Abstract

Recursive prompting with large language models enables scalable synthetic dataset generation but introduces the risk of bias amplification. We investigate gender bias dynamics across three generations of recursive text generation using three complementary evaluation frameworks: rule-based pattern matching, embedding based semantic similarity, and downstream task performance. Experiments with three initial bias levels (0.1, 0.3, 0.6) and four mitigation strategies reveal equilibrium dynamics rather than monotonic amplification. The low initial bias amplifies toward the model’s inherent bias level (+ 36%), whereas the high initial bias decays toward it (- 26%). Among mitigation methods, contrastive augmentation, which introduces gender-swapped variants, achieves significant downstream bias reduction (98.8% for low initial bias and 91% on average) despite producing higher embedding-based bias scores. This paradox demonstrates that semantic similarity metrics may diverge from behavioral fairness outcomes, highlighting the need for multidimensional evaluation in responsible synthetic data generation.

Introduction

Foundation models increasingly generate synthetic training data through iterative prompting and self-refinement. While this approach enables scalable dataset creation, the bias implications of recursive synthetic generation remain insufficiently examined. The self-instruct framework proposed by Wang et al. (2023) transformed instruction tuning by allowing language models to produce diverse and high-quality examples from minimal seed data. Building on this idea, recursive variants reuse model outputs as inputs for subsequent generations, offering the potential for unlimited dataset expansion but also raising questions about how bias propagates and evolves over time.

Bias in large language models has been widely documented across a range of linguistic and reasoning tasks. Prior studies have revealed systematic gender, racial, and occupational biases in model representations and outputs (Bolukbasi et al. 2016; Zhao et al. 2018; Bender et al. 2021). Early work by Zhao et al. (2018) demonstrated strong occupational stereotyping in coreference resolution through the WinoBias benchmark, while subsequent evaluations such

*This work does not relate to the authors’ positions at Proofpoint, LinkedIn, Microsoft, or Meta.

as BBQ (Parrish et al. 2022) extended bias assessment to question-answering, revealing persistent disparities across model scales and architectures. These findings underscore that even well-trained models internalize and reproduce societal stereotypes embedded in their training data.

Amplification of such biases during model usage has emerged as a critical concern. Empirical evidence indicates that repeated inference or self-conditioning can exacerbate existing imbalances. For example, Zhao et al. (2017) observed that models tend to magnify training-set biases when generating new examples. More recently, Wang et al. (2024) showed that iterative text continuation amplifies bias by 15–30% over multiple generations, suggesting that recursive or self-referential processes can compound representational skew.

Despite these insights, bias dynamics in synthetic data generation remain largely unexplored. The recursive generation of instructions or examples introduces feedback loops where a model effectively learns from its own outputs, potentially reinforcing or equilibrating biases over time. Understanding these recursive effects is essential as synthetic data increasingly substitutes or supplements human-curated datasets in model training.

Various mitigation techniques have been proposed to address bias propagation. Data augmentation through gender swapping (Zhao et al. 2018), adversarial debiasing (Zhang, Lemoine, and Mitchell 2018), and content filtering (Welbl et al. 2021) have all shown promise in constrained settings. Among these, contrastive augmentation, which creates paired gender variants of the same prompt, is notable for its simplicity and conceptual alignment with balance-oriented generation. However, its behavior under recursive synthetic generation has not been systematically studied.

In this work, we examine gender bias dynamics across three recursive generations of synthetic instruction data, using Google’s Gemma-2-2b-it model as a case study. We evaluate how initial seed bias influences amplification trajectories and compare four mitigation strategies, including contrastive augmentation. Our analysis employs both rule-based and embedding-based bias metrics, along with downstream behavioral evaluation.

The results suggest that recursive generation does not lead to inevitable bias growth but instead exhibits equilibrium dynamics, where systems stabilize around a model-specific

bias level regardless of initialization. Notably, contrastive augmentation achieves substantial downstream bias reduction (91% on average) even when embedding-based bias appears higher. This divergence highlights the limitations of single-metric evaluations and underscores the need for multidimensional fairness assessment in responsible synthetic data generation.

Methodology

We conducted recursive text generation experiments using Google’s Gemma-2-2b-it model with a temperature of 0.7 across three recursive generations. Each seed produced five child outputs per generation, yielding a progression of 50 seeds → 250 (Gen-1) → 1,250 (Gen-2) → 6,250 (Gen-3).

Seed sets were created at three target bias levels (0.1, 0.3, and 0.6) by sampling from a curated list of occupations: 12 female-associated roles (e.g., nurse, secretary, teacher), 12 male-associated roles (e.g., engineer, CEO, developer), and 20 gender-neutral prompts. Each seed consisted of a topic-oriented instruction such as “Describe the responsibilities of a nurse,” allowing for controlled bias measurement while preserving realistic recursive generation dynamics.

We compared four recursive generation strategies that differ in how gender-related information is introduced, balanced, or filtered. Table 1 summarizes the setup and rationale for each condition.

Table 1: Summary of recursive generation strategies. Each strategy represents a distinct approach to controlling gender information during recursive synthesis.

Strategy	Description
<i>Vanilla</i>	Standard recursive generation without modification; serves as the baseline condition.
<i>Contrastive</i>	Introduces gender-swapped augmentation, pairing each gendered prompt with its opposite variant (e.g., “male nurse” and “female engineer”); balances gender representation.
<i>Filtered</i>	Removes instructions with a rule-based bias score above 0.4, suppressing strongly stereotyped examples while maintaining data diversity.
<i>Size-matched</i>	Adds neutral instructions to match the sample size of the contrastive condition, isolating content effects from dataset size.

Bias Measurement

To comprehensively assess bias evolution, we employed three complementary evaluation frameworks capturing distinct dimensions of bias: explicit lexical patterns, implicit semantic associations, and downstream behavioral effects. This multi-level approach enables a deeper understanding of how bias manifests and propagates across recursive generations.

Rule-Based Metric. Explicit gender bias was measured through pattern-based analysis of stereotypical co-occurrences between gendered pronouns (he/she, his/her)

and occupation terms. Following prior work (Zhao et al. 2018), the bias rate was computed as the proportion of stereotypical associations among all gendered instructions, as defined in Equation 1:

$$\text{Bias}_{\text{rule}} = \frac{\text{Count(stereotypical pairs)}}{\text{Total gendered instructions}}. \quad (1)$$

This formulation captures overt lexical bias that reflects surface-level gender associations in the generated text. Although simple, it provides an interpretable baseline for observing explicit bias amplification trends.

Embedding-Based Metric. To capture more subtle semantic biases, we used the *all-MiniLM-L6-v2* sentence transformer to generate instruction embeddings and compared them with gender prototype vectors. The prototypes were computed as the mean embeddings of male-associated and female-associated seed instructions. For each instruction x , we calculated cosine similarity with both prototypes, and an instruction was labeled as biased if its similarity margin exceeded 0.35, as shown in Equation 2:

$$\text{Bias}_{\text{embed}}(x) = \begin{cases} 1, & \text{if } |\cos(x, v_{\text{male}}) - \cos(x, v_{\text{female}})| > 0.35 \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

This metric captures implicit bias expressed through representational proximity rather than explicit lexical markers. It reflects how the model organizes gendered concepts in semantic space, even when gender-specific words are not explicitly mentioned.

Downstream Evaluation. To evaluate whether instruction-level bias affects model behavior, we trained logistic regression classifiers on instruction embeddings to predict gender associations. We report both classification accuracy and a bias score defined as the absolute difference in predicted probabilities between male and female classes, as described in Equation 3:

$$\text{Bias}_{\text{down}} = |p(\text{male}) - p(\text{female})|. \quad (3)$$

This downstream bias metric quantifies behavioral disparities arising from representational differences. High values of $\text{Bias}_{\text{down}}$ indicate that even subtle embedding-level imbalances can translate into observable behavioral effects, linking representational bias and surface-level lexical bias within a unified evaluation framework.

Finally, we analyze all three bias measures across recursive generations to compare how explicit, implicit, and behavioral bias evolve under different generation strategies.

Results

Equilibrium Dynamics in Vanilla Condition

Embedding bias evolves in a non-monotonic manner across recursive generations. The vanilla generation condition demonstrates equilibrium dynamics, where systems converge toward a stable bias level over time rather than continuously amplifying or decaying. This behavior is illustrated in Figure 1, which shows embedding bias trajectories across three initial bias levels (0.1, 0.3, 0.6) and three recursive generations.

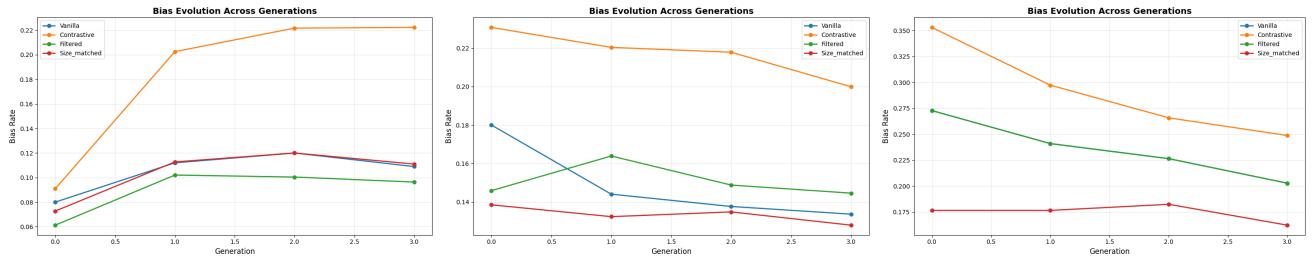


Figure 1: Embedding bias evolution across three generations for initial bias levels 0.1 (left), 0.3 (center), and 0.6 (right). Vanilla (blue) demonstrates equilibrium dynamics, where low bias amplifies and high bias decays. Contrastive (orange) yields higher embedding bias but lower downstream bias.

Table 2: Embedding and rule-based bias rates across generations under the vanilla condition. Systems converge toward equilibrium regardless of initial bias level.

Bias	Metric	Gen-0	Gen-1	Gen-2	Gen-3
0.1	Embedding	0.080	0.112	0.120	0.109 (+36%)
	Rule	0.200	0.167	0.267	0.342 (+71%)
0.3	Embedding	0.180	0.144	0.138	0.134 (-26%)
	Rule	0.467	0.560	0.557	0.579 (+24%)
0.6	Embedding	0.273	0.241	0.226	0.203 (-26%)
	Rule	0.542	0.545	0.542	0.535 (-1%)

Quantitative results in Table 2 confirm this equilibrium pattern. For low initial bias (0.1), embedding bias increased from 0.080 to 0.109 (+36%), whereas for medium (0.3) and high (0.6) initial biases, it decreased by approximately 26%. These changes indicate convergence toward a steady-state bias between 0.11 and 0.13, suggesting that the model possesses an inherent equilibrium bias level.

Interestingly, rule-based bias followed a different trajectory. It showed monotonic growth for the low-bias condition (0.200 to 0.342, +71%), a moderate increase for medium bias at +24%, and near stability for high bias at negative 1%. This divergence between rule-based and embedding-based measures suggests that they capture distinct dimensions of bias evolution, representing lexical and representational perspectives respectively. This reinforces the importance of multi-metric evaluation when analyzing recursive bias behavior.

Mitigation Strategy Comparison

Downstream bias, as measured by Equation 3, varied substantially across the four recursive generation strategies. The results summarized in Table 3 reveal clear differences in mitigation effectiveness and show that the relationship between embedding-level bias (Equation 2) and behavioral fairness is not always consistent.

Among all strategies, contrastive augmentation achieved the most effective bias mitigation. Although it produced the highest embedding bias in Figure 1 (orange lines), its downstream bias was minimal. For low initial bias (0.1), the downstream score decreased from 0.424 in the

Table 3: Downstream bias scores by strategy and initial bias level. Contrastive augmentation achieves a 91% average reduction despite exhibiting higher embedding bias (Fig. 1).

Strategy	Bias 0.1	Bias 0.3	Bias 0.6	Average
Vanilla	0.424	0.140	0.057	0.207
Contrastive	0.005	0.009	0.039	0.018
Filtered	0.241	0.278	0.057	0.192
Size-matched	0.424	0.124	0.108	0.219
Reduction (%)	-98.8	-93.6	-31.6	-91.3

vanilla setting to 0.005, corresponding to a 98.8% reduction. Medium and high bias conditions showed similar improvements (93.6% and 31.6% reductions respectively), yielding a 91.3% average reduction overall. This finding demonstrates that increased representational separation in embedding space does not necessarily translate to behavioral unfairness.

The filtered strategy displayed inconsistent results. It moderately improved fairness for low bias (-43%), degraded it for medium bias (+99%), and had negligible impact for high bias. These results suggest that filtering, while reducing explicit lexical bias, may also remove valid data and reduce sample diversity, leading to unstable mitigation outcomes.

The size-matched control performed slightly worse than the vanilla condition on average (+5.8%), confirming that the improvement observed in the contrastive setting originates from content balancing rather than sample size effects.

Changes in embedding bias across generations are visualized in Figure 2. The vanilla condition exhibits the strongest decay, while contrastive augmentation shows a small positive shift in embedding bias yet achieves the highest downstream fairness scores.

Figure 3 provides a complementary view by visualizing Gen-3 embedding bias rates across all strategies and initial bias levels. Contrastive augmentation consistently yields higher embedding bias values than other strategies but simultaneously achieves the lowest downstream bias, reinforcing that embedding-level separation and behavioral fairness capture distinct dimensions of bias.

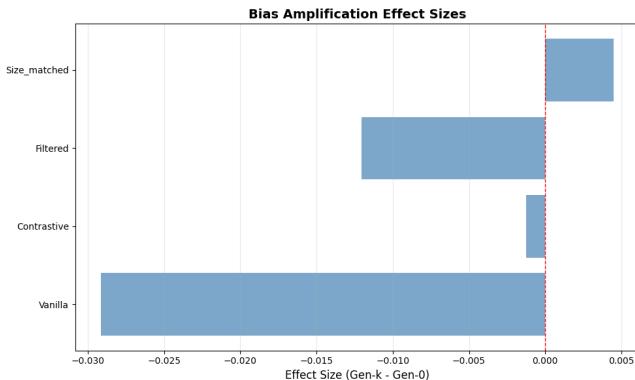


Figure 2: Effect sizes (Gen-3 minus Gen-0) for embedding bias across strategies. Negative values indicate bias decay. Vanilla shows the strongest decay, while contrastive augmentation achieves the best downstream fairness.

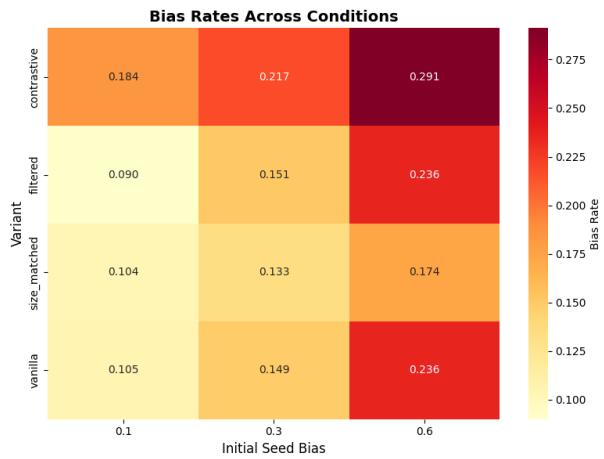


Figure 3: Gen-3 embedding bias across strategies and initial bias levels. Darker shades represent higher embedding bias. Despite these higher values, contrastive augmentation achieves the lowest downstream bias.

Statistical Analysis

Embedding bias distributions were compared across strategies through permutation testing with 5000 iterations. The analysis found no statistically significant differences between variants after false discovery rate (FDR) correction (contrastive versus vanilla: $p = 0.800$; filtered versus vanilla: $p = 0.809$; size-matched versus vanilla: $p = 0.394$).

Although intrinsic differences in embedding bias were statistically indistinguishable, downstream bias results revealed large practical effects. The contrastive strategy achieved a 91% reduction in downstream bias relative to the vanilla baseline, indicating that improvements in behavioral fairness can occur even when embedding-level changes are not statistically significant. This distinction emphasizes that statistical significance in intrinsic metrics does not necessarily correspond to practical significance in model behavior.

Discussion

Our findings provide three key insights with direct implications for responsible synthetic data generation.

Equilibrium Dynamics Over Amplification. Recursive generation does not lead to universal bias growth. Instead, the Gemma-2-2b-it model exhibits equilibrium dynamics, maintaining an intrinsic bias level around 0.11 to 0.13 as measured by the embedding metric. Seeds initialized below this level amplify toward it, while those above decay toward it. This behavior resembles regression to the mean in statistical systems. Effective mitigation approaches should therefore focus on shifting the equilibrium bias level itself rather than solely modifying the initial seed bias.

The Contrastive Paradox. Contrastive augmentation reveals an important paradox: higher embedding bias, reflecting stronger semantic polarization, coincides with substantially lower downstream bias and improved behavioral fairness. This occurs because gender-swapped augmentation produces two balanced semantic clusters, ensuring equal representation across genders. Embedding metrics capture representational separation, not fairness outcomes, whereas downstream bias reflects actual behavioral differences in model predictions. These results suggest that contrastive augmentation is effective precisely because it equalizes model outputs despite increased representational divergence.

Multidimensional Bias Measurement. The divergence among rule-based, embedding-based, and downstream bias measures highlights that bias is inherently multidimensional. Rule-based metrics capture explicit linguistic associations, embedding-based metrics quantify semantic clustering, and downstream evaluation assesses behavioral fairness. A comprehensive understanding of model bias requires integrating all three perspectives, since reliance on any single metric risks mischaracterizing mitigation outcomes.

Limitations. This study has several limitations. Computational constraints restricted the analysis to three recursive generations, and longer chains may reveal different convergence behaviors. The results are based on a single model, Gemma-2-2b-it, whose equilibrium level may not generalize to other architectures. Moreover, the binary gender framework used here does not capture non-binary or intersectional identities, which remain important directions for future work.

Conclusion

This study examined gender bias dynamics in recursive synthetic data generation and found equilibrium behavior rather than monotonic amplification. Low initial bias amplified toward the model's inherent bias level (+36%), while high initial bias decayed toward it (-26%). Contrastive augmentation achieved a 91% average reduction in downstream bias despite higher embedding bias, demonstrating that semantic clustering metrics can diverge from behavioral fairness outcomes. These findings indicate that effective mitigation must account for model-specific equilibrium dynamics and evaluate success through downstream task performance. As synthetic data generation becomes increasingly prevalent, un-

derstanding these equilibrium mechanisms is essential for designing responsible and bias-aware AI systems.

References

- Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 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.
- Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 4349–4357.
- Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. R. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105.
- Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 13484–13508.
- Wang, Z.; Wu, Z.; Zhang, J.; Jain, N.; Guan, X.; and Koshiyama, A. 2024. Bias amplification: Language models as increasingly biased media. *arXiv preprint arXiv:2410.15234*.
- Welbl, J.; Glaese, A.; Uesato, J.; Dathathri, S.; Mellor, J.; Hendricks, L. A.; Anderson, K.; Kohli, P.; Coppin, B.; and Huang, P.-S. 2021. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2447–2469.
- Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.
- Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979–2989.
- Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 15–20.