Bias Amplification: Large Language Models as Increasingly Biased Media

Ze Wang[1,2*], Zekun Wu[1,2*], Jeremy Zhang[3], Xin Guan [1], Navya Jain [2]
Skylar Lu [3], Saloni Gupta [4], Adriano Koshiyama [1]
[1]Holistic AI, [2]University College London
[3]Emory University, [4]University of Maryland, College Park

Abstract

Model collapse—a phenomenon characterized by performance degradation due to iterative training on synthetic data—has been widely studied. However, its implications for bias amplification, the progressive intensification of pre-existing societal biases in Large Language Models (LLMs), remain significantly underexplored, despite the growing influence of LLMs in shaping online discourse. In this paper, we introduce an open, generational, and long-context benchmark specifically designed to measure political bias amplification in LLMs, leveraging sentence continuation tasks derived from a comprehensive dataset of U.S. political news. Our empirical study using GPT-2 reveals consistent and substantial political bias intensification (e.g., right-leaning amplification) over iterative synthetic training cycles. We evaluate three mitigation strategies—Overfitting, Preservation, and Accumulation—and demonstrate that bias amplification persists independently of model collapse, even when the latter is effectively controlled. Furthermore, we propose a mechanistic analysis approach that identifies neurons correlated with specific phenomena during inference through regression and statistical tests. This analysis uncovers largely distinct neuron populations driving bias amplification and model collapse, underscoring fundamentally different underlying mechanisms. Finally, we supplement our empirical findings with theoretical intuition that explains the separate origins of these phenomena, guiding targeted strategies for bias mitigation.

1 Introduction

Large Language Models (LLMs) have become essential tools for content creation and summarization in various sectors, including media, academia, and business (Maslej et al., 2024). However, a significant but underexplored risk arises as LLMs increasingly rely on their own or other synthetic outputs for training, potentially leading to bias amplification—the progressive reinforcement of existing societal biases through iterative synthetic training (Peña-Fernández et al., 2023; Porlezza and Ferri, 2022; Nishal and Diakopoulos, 2024; Mehrabi et al., 2022; Taori and Hashimoto, 2022). This issue stems from the inherent tendency of LLMs to absorb biases from their training data (Parrish et al., 2022; Wang et al., 2024; Bender et al., 2021), causing them to align with specific political ideologies (Haller et al., 2023; Rettenberger et al., 2024a), favor certain class labels (Wyllie et al., 2024), and reduce output diversity over time (Alemohammad et al., 2023; Hamilton, 2024). The implications of bias amplification are substantial, including perpetuating stereotypes (King et al., 2024; Zekun et al., 2023), reinforcing social inequalities, and impacting democratic processes through skewed public opinion and increased polarization. Despite its significance, comprehensive frameworks and empirical research specifically addressing bias amplification in language models remain sparse (Guan et al., 2025).

In this study, we empirically investigate political bias amplification in GPT-2. We define political bias as the disproportionate generation of content aligned with specific political ideologies. Our experiments reveal that GPT-2 progressively exhibits stronger right-leaning and center-leaning biases in two distinct scenarios: (1) starting from fine-tuning on an unbiased dataset, and (2) initially fine-tuned exclusively on center-leaning articles. We also evaluate three mitigation strategies—Overfitting, Preservation, and Accumulation—to address bias amplification and model collapse. Preservation mitigates both phenomena effectively in the first scenario but fails to prevent bias amplification in the second. Additionally, we propose a novel mechanistic analysis using regression and statistical testing to examine neuron-level changes correlated with bias amplification and model collapse, identi-

*Corresponding Authors

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 115–132
December 20-24, 2025 ©2025 Association for Computational Linguistics

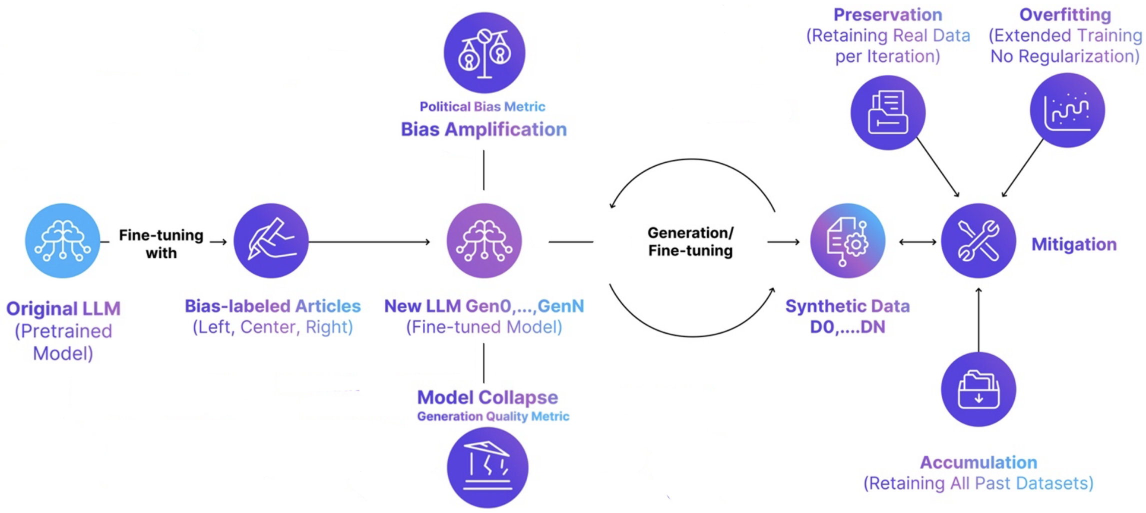


Figure 1: Overview of the iterative experimental procedure for synthetic fine-tuning and analysis.

fying largely distinct neuron groups for each phenomenon. This suggests different underlying mechanisms for bias amplification and model collapse. The experimental framework presented can be extended to other models and different types of biases. In summary, our contributions are: (i) a highly accurate classifier for detecting political bias in long-text content, providing a benchmark for evaluating political bias in LLMs via sentence continuation tasks; (ii) an empirical assessment of political bias amplification in GPT-2 across two fine-tuning setups; (iii) evaluation of three mitigation strategies; (iv) a novel mechanistic analysis method identifying neurons correlated with specific phenomena during inference; and (v) a theoretical intuition that explains the difference between bias amplification and model collapse based on their underlying causes. The experimental framework presented can be extended to other models and different types of biases.

2 Related Work

Our work builds on three key areas of research: bias amplification, model collapse, and political bias in LLMs. Bias amplification has been documented in feedback loops where models reinforce existing biases (Mehrabi et al., 2022), including in classifiers (Wyllie et al., 2024), generative models (Ferbach et al., 2024; Chen et al., 2024), and recommendation systems (Xu et al., 2023; Zhou et al., 2024). Model collapse is characterized by a loss of quality and diversity when models are recursively trained on synthetic data (Shumailov et al., 2024; Alemohammad et al., 2023; Guo et al., 2024; Wyllie et al., 2024; Dohmatob et al., 2024a). Prior work has shown this leads to lower perplexity but increased repetition and reduced linguistic diversity (Taori and Hashimoto, 2022; Guo et al., 2024; Dohmatob et al., 2024b; Seddik et al., 2024). Finally, studies on political bias have identified left-leaning tendencies in models like GPT-3 and ChatGPT in both U.S. and German political contexts (Rettenberger et al., 2024b; Shumailov et al., 2024; Feng et al., 2024; Rotaru et al., 2024; Motoki et al., 2024). In parallel, growing attention has been paid to political biases in LLMs, now a prevalent form of "media" that people rely on for global news (Maslej et al., 2024). Rettenberger et al. (2024b); Shumailov et al. (2024); Feng et al. (2024) explored the bias through voting simulations within the spectrum of German political parties, consistently finding a left-leaning bias in models like GPT-3 and Llama3-70B. Similarly, for the U.S. political landscape, Rotaru et al. (2024); Motoki et al. (2024) identified a noticeable left-leaning bias in ChatGPT and Gemini when tasked with rating news content, evaluating sources, or responding to political questionnaires. Bang et al. (2024) study political bias in LLMs through the task of generating news headlines on politically sensitive topics and find that the political perspectives expressed by LLMs vary depending on the subject matter. A comprehensive literature review is available in Appendix K.

3 Methodology

This section focuses on the step-by-step experimental procedure outlined in Figure 1. Our study focuses on the political bias of LLMs within the US political spectrum, particularly in sentence continuation tasks. This is important as LLMs are increasingly influencing global news consumption (Maslej et al., 2024; Peña-Fernández et al., 2023; Porlezza and Ferri, 2022), and traditional news outlets, such as the Associated Press, are beginning to integrate LLMs for automated content generation from structured data (The Associated Press, 2024).

3.1 Dataset Preparation

We randomly selected 1,518 articles from the Webis-Bias-Flipper-18 dataset (Chen et al., 2018), which contains political articles from a range of U.S. media outlets published between 2012 and 2018, along with bias ratings assigned at the time for each media source. These bias ratings, provided by AllSides, were determined through a multi-stage process incorporating assessments from both bipartisan experts and the general public (AllSides, 2024a). The random sampling was stratified based on bias ratings to ensure an even distribution of the 1,518 articles into three groups of 506 each, representing left-leaning, right-leaning, and center-leaning media.

3.2 Successive Fine-tuning

Following Shumailov et al. (2024); Dohmatob et al. (2024b), we perform iterative fine-tuning. First, GPT-2 is fine-tuned on the 1,518 real news articles (detailed in Section 3.1) to yield the Generation 0 (G0) model. G0 then generates a synthetic dataset, D_0, of the same size (1,518 articles). This dataset D_0 is used to fine-tune the Generation 1 (G1) model, which is the first model trained on purely synthetic data. The process continues up to Generation 10 (G10), where each Gi model is fine-tuned on the synthetic data D_{i-1} produced by model G$i-1$. The models were trained for 5 epochs using standard hyperparameters adapted from Taori and Hashimoto (2022), with full details provided in Appendix H.

3.3 Synthetic Data Generation

Synthetic datasets, $\{D_i\}_{i=0}^{10}$, are generated as follows: For each original news article, its tokenized title serves as an initial prompt, and its tokenized body is segmented into sequential 64-token blocks, which serve as subsequent prompts. For each such prompt, the model predicts the next 64 tokens. These predictions are made using deterministic generation to enhance the reproducibility. All the newly generated 64-token sequences (one from the title prompt and one from each body block prompt) are concatenated and then decoded back into text. This process creates one synthetic article from each original article, resulting in a synthetic dataset of the same number of articles as the original.

3.4 Political Bias Metric

To measure political bias, we developed a classification model to assess the political leaning of each generated article. We define political bias as the disproportionate production of articles with specific political leanings, as identified by our classifier. Unlike Bang et al. (2024), who defines political bias in a topic-specific manner, we take a broader perspective by measuring the overall political leaning of the model across a diverse set of topics, analogous to how media outlets are rated. After a grid search, a `roberta-base` model achieved the highest performance, with a macro F1 score of 0.9196 on our held-out test set (see Table 1). The classifier was trained on a large dataset of U.S. news articles from 2012-2018, and its application to significantly later content may require recalibration. Full details on the dataset and model training are provided in Appendix A.

3.5 Generation Quality Metric

To evaluate generation quality and address the unreliability of perplexity metrics in the face of repetitive content (see Section 4.2), we introduce the *text quality index*. This metric is based on the Gibberish Detector (Jindal, 2021) because it directly captures the loss of coherence and semantic clarity we observed in later model generations. The detector classifies each sentence into one of four levels: Noise (score 0), Word Salad (1), Mild Gibberish (2), or Clean (3). The final index for an article is the average score of all its sentences, prioritizing coherence and providing a more meaningful measure of quality degradation.

3.6 Mechanistic Analysis

To gain a clearer understanding of the causes of bias amplification and how it empirically relates to model collapse, we conduct a mechanistic analysis of how neurons behave and vary across different generations of fine-tuned GPT-2 models (We have 11 generations for each training round, with a total of 6 rounds, resulting in 66 versions of fine-tuned GPT-2.), each exhibiting different levels of generation quality and bias performance. The first

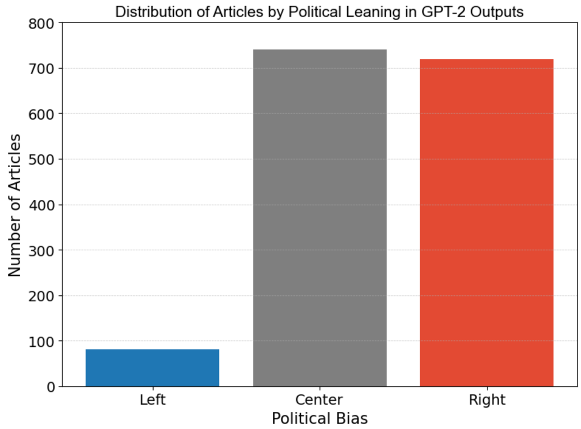


Figure 2: Distribution of political bias labels ('Left', 'Center', 'Right') for initial GPT-2 synthetic outputs, classified by our Political Bias Metric.

step is to extract the changing weight (or activation value) pattern of each neuron across versions and compare it to the corresponding changes in bias performance and generation quality. For each of the 9,216 neurons, which correspond to the 768 output neurons from the feed-forward network (FFN) sublayer in each of the 12 transformer blocks of the GPT-2 model, we compute the correlation between its weight (or activation value) and the model's bias performance (or generation quality) across all 66 versions.

To statistically test the significance of these correlations, we regress the change in neuron weights (or activations) against the change in our metrics for bias and quality. By identifying neurons with statistically significant correlations using Newey-West adjusted p-values, we can determine the sets of neurons associated with each phenomenon. Full details of the statistical model are in Appendix G. Using the p-values and a 95% significance threshold, we identify the sets of neurons significantly correlated with bias amplification (i.e., changes in the proportion of politically leaning articles) and with model collapse (i.e., changes in the generation quality index). By comparing these sets and analyzing their degree of overlap, we gain evidence about whether the two phenomena arise from distinct underlying mechanisms.

4 Results

In this section, we analyze the evolution of political bias and generation quality in GPT-2 over successive iterations of synthetic fine-tuning, comparing results with and without mitigation strategies.

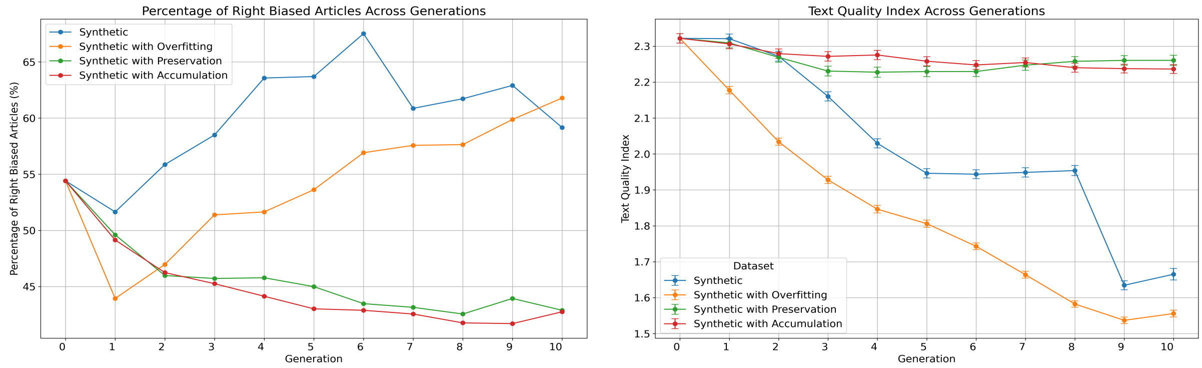
4.1 Political Bias

GPT-2 was used to generate the synthetic dataset. Since the original human-written dataset is unbiased—with an equal number of articles for each political-leaning category—the synthetic dataset should ideally mirror this balanced distribution if GPT-2 had no pre-existing bias. Figure 2 presents the distribution of synthetic articles generated by GPT-2 across political bias labels. The model predominantly produces center-leaning (47.9%) and right-leaning (46.8%) articles, suggesting a pre-existing bias towards these categories before any fine-tuning. Starting from the initial GPT-2 model, we fine-tuned it iteratively, generating synthetic datasets to train successive models up to Generation 10. Figure 3a illustrates how bias amplifies across generations. Surprisingly, fine-tuning on the unbiased real dataset increases right-leaning bias, with 53.7% of articles classified as right-leaning in Generation 0. Furthermore, without mitigation strategies, successive rounds of synthetic fine-tuning lead to a continuous rise in right-leaning articles, peaking at Generation 6 (67.6%) before stabilizing. Figures 6 and 7 in Appendix B show the percentage of center-leaning and left-leaning articles across generations. Notably, the proportion of center-leaning articles remains stable at approximately 35% throughout synthetic fine-tuning.

To further illustrate, we conducted a qualitative analysis of how generated articles evolve. For example, a left-leaning article on Trump's immigration policy progressively shifted to a right-leaning frame, replacing terms like "undocumented immigrants" with "illegal immigrants" and portraying the policies as strong and decisive. A detailed case study is available in Appendix C.

4.2 Generation Quality

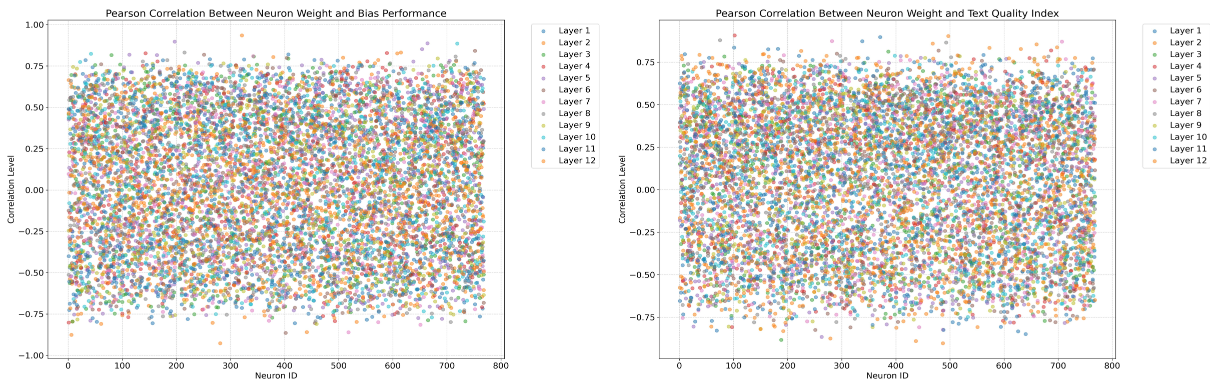
Figure 3b illustrates the text quality index across generations. In the training loop without any mitigation strategy, model collapse occurs, as evidenced by the gradual decline in the average text quality index. Furthermore, the distribution of the text quality index shifts significantly toward the lower-quality region over generations, eventually generating data that was never produced by Generation 0 (Figure 8 in Appendix D). These results align with prior research on model collapse, such as (Shumailov et al., 2024), though we did not observe substantial variation in variance. Conversely, perplexity measurements exhibit a consistent de-



(a) Right-leaning bias %

(b) Text quality index

Figure 3: Evolution of (a) right-leaning bias and (b) text quality index across generations (initial G0: unbiased dataset). Compares baseline ('Synthetic') with three mitigation strategies. Text quality includes 95% CIs.



(a) Neuron weights vs. bias (Right-leaning % change).

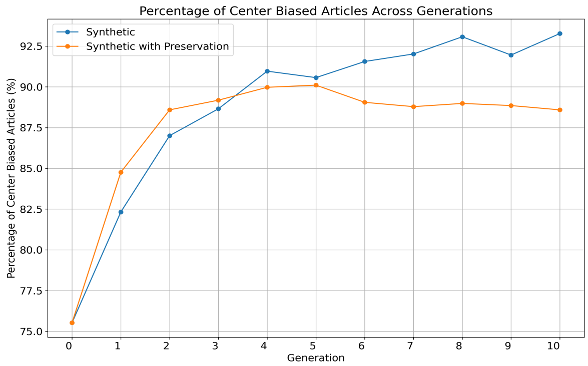
(b) Neuron weights vs. quality (Text Quality Index change).

Figure 4: Pearson correlations: Neuron weights vs. (a) bias and (b) quality, across 66 GPT-2 versions.

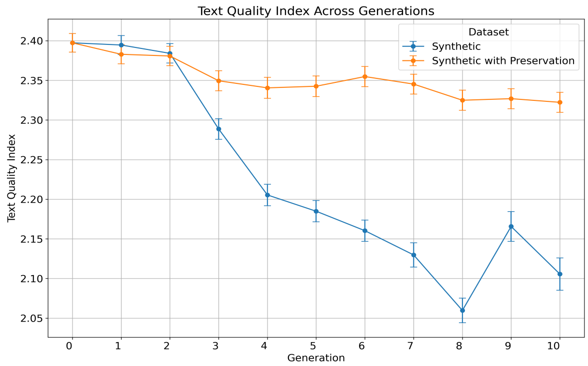
cline across generations, generally suggesting an improvement in generation quality (Figure 9 in Appendix E). For a closer look, the examples in Appendix F illustrate how generated articles gradually lose coherence and relevance across generations, with increasing occurrences of repetition and fragmented sentences. By Generation 10, the text becomes largely incoherent and detached from the original content, reducing its readability and meaning. However, despite the evident decline in generation quality, perplexity decreases over generations, as indicated by the results at the end of each synthetic output example. This pattern is consistent across most synthetic outputs, suggesting that perplexity does not accurately capture the model's true generative capabilities and its value can be distorted by frequent repetitions.

4.3 Mitigation Strategies

We applied three mitigation strategies: (1) Overfitting, which involved increasing the training epochs to 25 (five times the baseline) and setting weight decay to 0 to reduce regularization and encourage overfitting, as proposed by Taori and Hashimoto (2022) based on the uniformly faithful theorem of bias amplification; (2) Preserving 10% of randomly selected real articles during each round of synthetic fine-tuning, a method proposed and used in (Shumailov et al., 2024; Alemohammad et al., 2023; Dohmatob et al., 2024b; Guo et al., 2024); and (3) Accumulating all previous fine-tuning datasets along with the new synthetic dataset in each fine-tuning cycle, which was introduced by Gerstgrasser et al. (2024). As shown in Figure 3a, overfitting helps reduce bias amplification in the early generations compared to the no-mitigation baseline (the 'Synthetic' line), but it fails to prevent bias amplification in the later generations. Additionally, it incurs a significant cost—further deterioration in generation quality, as shown in Figure 3b. Notably, both the preservation and accumulation strategies effectively mitigate model collapse and reduce bias, yielding 41.89% and 42.7% right-leaning articles, respectively, at Generation 10.



(a) Center-leaning bias %



(b) Text quality index

Figure 5: Alternative Setup (G0: center-leaning fine-tune): Evolution of (a) center-leaning bias and (b) text quality. Baseline ('Synthetic') vs. Preservation. Text quality includes 95% CIs.

4.4 Mechanistic interpretation

Figure 4 illustrates the correlation between neuron weights and the model's bias performance (or generation quality) across the 66 fine-tuned versions, evaluated for each of the 9,216 neurons. Through linear regressions and statistical tests, we identify 3,243 neurons with statistically significant correlations (p-value < 0.05) with bias performance, suggesting they are key contributors to bias shifts. Meanwhile, 1,033 neurons exhibit significant correlations with generation quality, but only 389 neurons overlap between the two sets. This limited overlap implies that distinct neuron populations drive bias amplification and generation quality deterioration. We then applied the same procedure using activation values. This analysis yielded two sets: one consisting of 3,062 neurons whose activation value changes are significantly correlated with changes in bias performance, and another with 2 neurons correlated with changes in generation quality. The stark contrast in activation-correlated neurons for bias performance (3,062 neurons) versus generation quality (a mere 2 neurons) provides particularly strong evidence that these two issues may operate via substantially different pathways within the model.

5 Discussion and Conclusion

Our results demonstrate that bias amplification and model collapse are driven by distinct sets of neurons, implying they are separate phenomena. Consequently, strategies targeting model collapse by mitigating sampling error, such as data preservation, are not guaranteed to prevent bias amplification. This is confirmed in our alternative experimental setup (Appendix I), where preservation prevents collapse but fails to stop bias from intensifying. In this alternative synthetic training cycle, we begin with GPT-2 fine-tuned on 1,518 randomly sampled center-labeled articles. We compare the baseline with the most effective and cost-efficient mitigation strategy identified in our previous results: Preservation. As shown in Figure 5, Preservation successfully prevents model collapse but fails to mitigate bias amplification in center-leaning article generation, which increases from 72.9% at Generation 0 to 88.2% at Generation 10. These findings suggest that although techniques like Preservation, which reduce sampling error, are effective at mitigating model collapse, they do not necessarily prevent bias amplification—consistent with the implications drawn in Section 4.4. To understand why this could happen, we offer a theoretical intuition explaining the difference between bias amplification and model collapse based on their underlying causes, in Appendix L. While data preservation can constrain the model, recalling an already-biased real dataset may simply reinforce the original bias. This creates a potential tradeoff, as de-biasing techniques like weighted sampling could reintroduce sampling errors and trigger model collapse. This highlights the urgent need for targeted interventions, such as dynamic data reweighting or direct manipulation of bias-associated neurons, to ensure fair and equitable model development. Developing such interventions requires a deeper mechanistic understanding. We chose a statistical analysis over Sparse Autoencoder (SAE) methods because our approach is designed to track the temporal dynamics of neuron behavior across training generations, a task for which static SAE analysis is less suited. Furthermore, SAEs are challenging to apply here because political bias is a nuanced, distributional phenomenon (i.e., disproportionate content generation) rather than a simple feature to be isolated. Future work should focus on refining mechanistic analysis for such complex biases, for instance by adapting feature attribution methods or tracing how specific training examples influence bias-implicated neurons over time.

6 Limitations

While this work introduces a comprehensive framework for understanding bias amplification in large language models and provides empirical evidence using GPT-2, several limitations must be acknowledged. First, the scope of our experiments is restricted to political bias in the context of U.S. media. Since the political spectrum may shift over time, periodic updates to the political bias classifier are necessary to ensure its accuracy when benchmarking more recent datasets.

Additionally, because our primary focus is on investigating political bias amplification and its relationship with model collapse, we conducted our experiments using GPT-2—a relatively small language model—to ensure the practicality of fine-tuning 66 versions of the model. Future work may extend our methodology to larger architectures, particularly to examine how model scale influences the degree of bias amplification.

Another limitation lies in our choice of mitigation strategies. While Preservation and Accumulation show promise in reducing model collapse, their computational cost and data storage requirements (especially for Accumulation, which retains all prior data) may present scalability challenges for very large models or extensive iterative training. Moreover, these strategies were evaluated primarily in the context of synthetic fine-tuning, and their effectiveness in real-world deployment scenarios remains to be thoroughly investigated.

7 Ethical Considerations

This study focuses on bias amplification in LLMs—a phenomenon with significant ethical implications. Beyond issues of fairness, the iterative amplification of biases can weaken the integrity of information ecosystems, particularly if synthetically generated content becomes widespread. The risk of bias amplification is especially concerning in systems that are iteratively trained on synthetic data, as it can lead to unintended and increasingly skewed distortions in model outputs. These distortions may propagate harmful biases or misinformation, potentially influencing downstream tasks such as automated content generation, decision-making, and user interactions with AI. Furthermore, the finding that bias amplification and model collapse may be driven by distinct mechanisms highlights the complex challenge of balancing various aspects of model performance (e.g., accuracy, fairness, coherence) and highlights the difficulty in developing mitigation strategies that address one issue without negatively impacting another, especially in high-stakes scenarios.

It is crucial to explicitly state that the methodologies and data used in this research should not be applied to develop or train biased models for harmful applications. This study is intended to advance the understanding of bias amplification and model collapse in LLMs, while promoting responsible and ethical AI development.

This work includes content that may contain personally identifying information or offensive language. However, all such material is derived exclusively from publicly available news article datasets or is generated synthetically by models fine-tuned on these open-source datasets—or on synthetic data produced by earlier generations in our training pipeline. As such, any sensitive or offensive content reflects characteristics of the source material and does not imply our endorsement. Our objective is to thoroughly investigate political bias in LLMs to inform the development of strategies that can mitigate disproportionate representation of such content in real-world deployments. Additionally, we conduct a manual review of the news article dataset to remove any identifiable information about article authors.

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 2023. Self-consuming generative models go mad. *Preprint*, arXiv:2307.01850.

AllSides. 2024a. Media bias rating methods. Accessed: 2024-09-16.

AllSides. 2024b. Nbc news media bias/fact check. Accessed: 2024-10-10.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11142–11159, Bangkok, Thailand. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Tianwei Chen, Yusuke Hirota, Mayu Otani, Noa Garcia, and Yuta Nakashima. 2024. Would deep generative models amplify bias in future models? *Preprint*, arXiv:2404.03242.

Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Learning to flip the bias of news headlines. In *11th International Natural Language Generation Conference (INLG 2018)*, pages 79–88. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. 2024a. Model collapse demystified: The case of regression. *Preprint*, arXiv:2402.07712.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. 2024b. A tale of tails: Model collapse as a change of scaling laws. *Preprint*, arXiv:2402.07043.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, Francois Charton, and Julia Kempe. 2024. Beyond model collapse: Scaling up with synthesized data requires reinforcement. *Preprint*, arXiv:2406.07515.

Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. 2024. Self-consuming generative models with curated data provably optimize human preferences. *Preprint*, arXiv:2407.09499.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. 2024. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *Preprint*, arXiv:2404.01413.

Tim Groeling. 2013. Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, 16(Volume 16, 2013):129–151.

Xin Guan, Nate Demchak, Saloni Gupta, Ze Wang, Ediz Ertekin Jr., Adriano Koshiyama, Emre Kazim, and Zekun Wu. 2025. SAGED: A holistic bias-benchmarking pipeline for language models with customisable fairness calibration. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3002–3026, Abu Dhabi, UAE. Association for Computational Linguistics.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. *Preprint*, arXiv:2311.09807.

Patrick Haller, Ansar Aynetdinov, and Alan Akbik. 2023. Opiniongpt: Modelling explicit biases in instruction-tuned llms. *Preprint*, arXiv:2309.03876.

Sil Hamilton. 2024. Detecting mode collapse in language models via narration. *Preprint*, arXiv:2402.04477.

Madhur Jindal. 2021. Gibberish detector. `https://huggingface.co/madhurjindal/autonlp-Gibberish-Detector-492513457`. Accessed: 2024-09-19.

Theo King, Zekun Wu, Adriano Koshiyama, Emre Kazim, and Philip Treleaven. 2024. Hearts: A holistic framework for explainable, sustainable and robust text stereotype detection. *arXiv preprint arXiv:2409.11579*.

Miaomiao Li, Hao Chen, Yang Wang, Tingyuan Zhu, Weijia Zhang, Kaijie Zhu, Kam-Fai Wong, and Jindong Wang. 2025. Understanding and mitigating the bias inheritance in llm-based data augmentation on downstream tasks. *Preprint*, arXiv:2502.04419.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. 2024. Artificial intelligence index report 2024. *Preprint*, arXiv:2405.19522.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A survey on bias and fairness in machine learning. *Preprint*, arXiv:1908.09635.

Fumiya Motoki, Vinícius Pinho Neto, and Vanessa Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198:3–23.

NBC News. 2016. First read: Why it's so hard for trump to retreat on immigration. Accessed: 2024-10-10.

Sachita Nishal and Nicholas Diakopoulos. 2024. Envisioning the applications and implications of generative ai for news media. *Preprint*, arXiv:2402.18835.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. *Preprint*, arXiv:2110.08193.

Simón Peña-Fernández, Koldobika Meso-Ayerdi, Ainara Larrondo-Ureta, and Javier Díaz-Noci. 2023. Without journalists, there is no journalism: the social dimension of generative artificial intelligence in the media. *Profesional de la información*, 32(2):e320227.

Colin Porlezza and Giuseppe Ferri. 2022. The missing piece: Ethics and the ontological boundaries of automated journalism. *#ISOJ Journal*, 12(1):71–98.

Luca Rettenberger, Markus Reischl, and Mark Schutera. 2024a. Assessing political bias in large language models. *Preprint*, arXiv:2405.13041.

Luca Rettenberger, Markus Reischl, and Mark Schutera. 2024b. Assessing political bias in large language models. *Preprint*, arXiv:2405.13041.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.

George-Cristinel Rotaru, Sorin Anagnoste, and Vasile-Marian Oancea. 2024. How artificial intelligence can influence elections: Analyzing the large language models (llms) political bias. *Proceedings of the International Conference on Business Excellence*, 18(1):1882–1891.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. 2024. How bad is training on synthetic data? a statistical analysis of language model collapse. *Preprint*, arXiv:2404.05090.

I. Shumailov, Z. Shumaylov, Y. Zhao, et al. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631:755–759.

Rohan Taori and Tatsunori B. Hashimoto. 2022. Data feedback loops: Model-driven amplification of dataset biases. *Preprint*, arXiv:2209.03942.

The Associated Press. 2024. Artificial intelligence at the associated press. Accessed: 2024-09-16.

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Skylar Lu, Sachin Beepath, Ediz Ertekin, and Maria Perez-Ortiz. 2024. JobFair: A framework for benchmarking gender hiring bias in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3227–3246, Miami, Florida, USA. Association for Computational Linguistics.

Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. 2024. Fairness feedback loops: Training on synthetic data amplifies bias. *Preprint*, arXiv:2403.07857.

Hangtong Xu, Yuanbo Xu, Yongjian Yang, Fuzhen Zhuang, and Hui Xiong. 2023. Dpr: An algorithm mitigate bias accumulation in recommendation feedback loops. *Preprint*, arXiv:2311.05864.

Wu Zekun, Sahan Bulathwela, and Adriano Soares Koshiyama. 2023. Towards auditing large language models: Improving text-based stereotype detection. *arXiv preprint arXiv:2311.14126*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Preprint*, arXiv:1707.09457.

Yuqi Zhou, Sunhao Dai, Liang Pang, Gang Wang, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. 2024. Source echo chamber: Exploring the escalation of source bias in user, data, and recommender system feedback loop. *Preprint*, arXiv:2405.17998.

A Details on Model Training for Political Bias Metric

We experiment with multiple transformer-based models, including BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), selecting the best-performing model based on the macro F1 score. Each model is fine-tuned using the HuggingFace `Trainer` class with a learning rate of 2×10^{-5}, a batch size of 16, and 5 training epochs. We employ a cross-entropy loss function for multi-class classification. Tokenization is performed using each model's respective tokenizer with a maximum sequence length of 512 tokens. To mitigate overfitting, we apply a weight decay of 0.01 during training. Model checkpoints are saved after each epoch, and the best model is selected based on the macro F1 score evaluated on the validation set.

The classifier is trained on the Webis-Bias-Flipper-18 dataset, excluding the 1,518 articles used for GPT-2 fine-tuning. To mitigate class imbalance, center-leaning articles are resampled to ensure equal representation across categories. The dataset is then divided into training (70%), validation (15%), and test (15%) subsets, stratified by bias label. We also conduct a human review to remove identifiable information about media sources and authors. The training dataset comprises 2,781 distinct events that occurred in the U.S. between 2012 and 2018. For each event, it includes news articles collected from a wide range of media outlets. In total, it contains articles from 97 different outlets, such as The Washington Examiner, The Washington Post, HuffPost, Reuters, and others. This makes the dataset a strong representation of the diversity of U.S. media sources and event domains, and therefore strengthens the generalizability of our classifier in the U.S. context.

We use a weighted random sampler during training to ensure balanced class representation. Models are evaluated using the macro F1 score to account for the multi-class nature of the task, ensuring balanced performance across all bias categories. Final evaluation is conducted on the held-out test set. Additionally, we report the loss, runtime, and sample processing rates for completeness.

B Percentage of Center (Left) Biased Articles

Table 1: Macro F1 Scores for Political Bias Classifier Models; `roberta-base` selected.

Model	Macro F1 Score
distilbert-base-uncased	0.8308
bert-base-uncased	0.8559
albert-base-v2	0.8649
roberta-base	0.9196

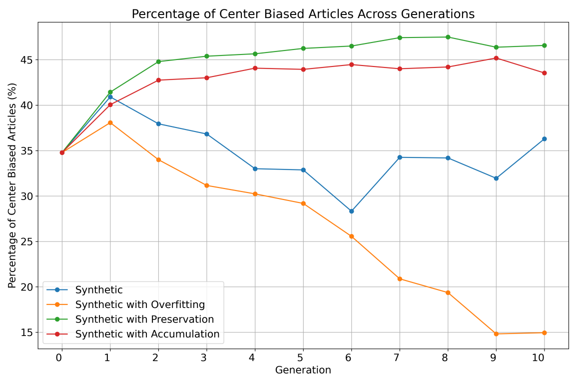


Figure 6: Evolution of center-leaning article percentage across generations, comparing baseline ('Synthetic') with three mitigation strategies (Main Experiment).

C Qualitative Bias Analysis

We employed qualitative methods to confirm our findings in media bias. Specifically, we utilized a media bias identification framework grounded in foundational works such as Entman's framing theory (Entman, 1993) and other research on media bias detection (Rodrigo-Ginés et al., 2024; Groeling, 2013). This framework provides a robust lens to evaluate political biases in the framing and language use of media texts. Given the nature of our data—text exclusive of visual or contextual cues like formatting—certain types of media bias commonly seen in formatted articles or televised programs (e.g., visual bias or tone) may not apply. Therefore, our focus was on the two key aspects of political bias that are particularly relevant in textual analysis:

Story Framing and Selection Bias. This type of bias emerges when inherent leanings are found in the way topics, arguments, or narratives are structured. For instance, some aspects of reality are highlighted while others are obscured, shaping how the audience understands and interprets the events or issues at hand (Entman, 1993; Groeling, 2013). In extreme cases, opposing viewpoints are entirely excluded, leading to a one-sided representation of

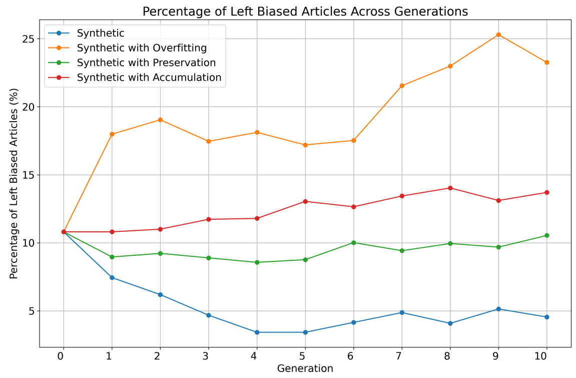


Figure 7: Evolution of left-leaning article percentage across generations, comparing baseline ('Synthetic') with three mitigation strategies (Main Experiment).

the issue. This selective omission restricts the audience's comprehension of the full spectrum of perspectives, resulting in a distorted portrayal of the issue (Rodrigo-Ginés et al., 2024; Groeling, 2013). Entman described this as the selection and salience of specific facts that promote particular definitions, evaluations, and recommendations.

Loaded Language Bias. This bias is identified through the use of charged or emotive words that signal political or ideological leanings. A common example is the difference in connotation between terms such as "undocumented" versus "illegal" immigrants. Such language choices often shape the audience's perception by evoking specific emotional responses (Rodrigo-Ginés et al., 2024; Groeling, 2013).

Below is an example of GPT-2 text outputs influenced by iterative synthetic training. The original article, titled "First Read: Why It's So Hard for Trump to Retreat on Immigration, is a political opinion piece from NBC News, a left-leaning outlet as rated by AllSides (NBC News, 2016; AllSides, 2024b). The analysis follows the qualitative framework:

Original Article. Why Its So Hard for Trump to Retreat on Immigration First Read is a morning briefing from Meet the Press and the NBC Political Unit on the day's most important political stories and why they matter. Why its so hard for Trump to retreat on immigration Since launching his presidential candidacy 14 months ago, Donald Trumps most consistent and uncompromising policy issue has been immigration. Indeed, it was the subject of his first general-election TV ad that started airing on Friday. Yet over the weekend, his top aides and advisers suggested that Trump might be shifting on his past position that all of the 11 million undocumented immigrants living in the United States must be deported forcibly. To be determined, is what newly minted Campaign Manager Kellyanne Conway said on CNN when asked if Trump was retreating on the deportation force he talked about during the primary season. But here's why its so hard – if not impossible – for Trump to retreat on immigration: Hes caught between his clear, unambiguous past statements and a base that might not willing to see him moderate on the issue. His past statements: Aug. 16, 2015 ""We're going to keep the families together, but they have to go,"" Trump said on NBCs Meet the Press. More Trump: ""We will work with them. They have to go. Chuck, we either have a country, or we don't have a country,"" he said. Nov. 11, 2015 You are going to have a deportation force, and you are going to do it humanely, Trump said on MSNBCs Morning Joe when asked how he would round up the nations 11 million undocumented immigrants. April 21, 2016 Look, were either going to have a country or were not going to have a country. But many people are very fine people. And I'm sure these are very, very fine people. They're going to go, and were going to create a path where we can get them into this country legally, okay? But it has to be done legally – when asked by a questioner at a Today town hall that persons undocumented relatives would have to be deported if Trump becomes president. Trump cant ignore a base that has cheered his uncompromising immigration position And then there are the Trump supporters who've cheered the GOP presidential nominee for being so uncompromising on immigration.

Classification Probability: 0.9946 for left-leaning, 0.0051 for center-leaning, 0.0002 for right-leaning.

Analysis:

Omission of Alternative Perspectives. There is a lack of discussion about the potential benefits of Trump's immigration policies or the concerns that might have led to his hardline approach. The article does not present viewpoints from supporters who agree with his stance or any arguments that justify his policies. This omission restricts the audience's understanding of the full range of perspectives on the issue.

Highlighting the political objectives behind Trump's immigration policies. The emphasis is placed on Trump's electoral interests rather than presenting the policy as being for the benefit of the people. For example, "Trump can't ignore a base that has cheered his uncompromising immigration position" and "And then there are the Trump supporters who've cheered the GOP presidential nominee for being so uncompromising on immigration." Additionally, "But here's why it's so hard — if not impossible — for Trump to retreat on immigration: He's caught between his clear, unambiguous past statements and a base that might not be willing to see him moderate on the issue."

Differences in Connotation. The article uses the term "undocumented immigrants" instead of "illegal immigrants," a shift that appeared in later generations' articles. This choice of wording can shape perceptions and evoke specific emotional responses, aligning with a particular ideological perspective.

Generation 0. Why Its So Hard for Trump to Retreat on Immigration The president's decision to withdraw from the Paris climate accord is a major blow to the United States' standing in the world. The United States has been a major player in the world of climate change, and has been a major player in the fight against climate change. But Trump's decision to immigration. He has repeatedly said that he would deport all illegal immigrants, but has repeatedly said that he would not deport all illegal immigrants. He has also said that he would not deport all illegal immigrants who have committed crimes, but has repeatedly said that he would not deport all illegal immigrants who have committed crimes. He has also clear, this is not a new position. It was a position that Trump has held for years. But it is a position that he has held for years. And it is a position that he has held for years. And it is a position that he has held for years. And it is a position that he has the deportation force. 1. Trump is not retreating on the deportation force. Trump is not retreating on the deportation force. Trump is not retreating on the deportation force. Trump is not retreating on the the Press. "We're going to keep the families together, but they have to go. We're going to keep the families together. We're going to keep the families together. We're going to keep the families together. We're going to keep the families together. We're going to keep the families together.ly. You are going to do it humanely. You are going to do it humanely. You are going to do it humanely. You are going to do it humanely. You are going to do it humanely. You are going to do it humanely. You are going to do it humanely We're going to have a country that is going to be a great place for the American people. We're going to have a country that is going to be a great place for the American people. We're going to have a country that is going to be a great place for the American people.

Classification Probability: 0.0011 for left-leaning, 0.9957 for center-leaning, 0.0032 for right-leaning.

Analysis:

Emphasis on Positive Outcomes. In generation 0, the text highlights Trump's determination and steadfastness, stating, "Trump is not retreating on the deportation force," and emphasizing a positive future outlook with the policy. Phrases like "we're going to have a country that is going to be a great place for the American people" portray strict immigration enforcement as beneficial for national prosperity and security. This aligns more with the right relative to the left perspective.

Omission of Criticisms and Alternative Viewpoints. The text omits the political motivations behind Trump's immigration policies, which were discussed in the original article, e.g. "But here's why it's so hard — if not impossible — for Trump to retreat on immigration: He's caught between his clear, unambiguous past statements and a base that might not be willing to see him moderate on the issue." However, it introduces criticism on other topics, such as "The president's decision to withdraw from the Paris climate accord is a major blow to the United States' standing in the world," .

Mild Emotional Appeal. The language is positive but relatively restrained, portraying Trump as determined yet thoughtful.

Differences in Connotation. In Generation 0, the article uses the term "illegal immigrants" instead of "undocumented immigrants" as in the original, signaling a clear shift. This language choice can shape perceptions and elicit specific emotional responses.

Generation 4. Why Its So Hard for Trump to Retreat on Immigration, and Why Its So Hard for Trump to Retreat on Immigration, and why its so hard for Trump to Retreat on immigration, and the presidents speech was a stark reminder that the president is not the presidents speech was a stark reminder that the president is not the economy the presidents campaign manager, Kellyanne Conway, said the presidents comments were a complete fabrication. The presidents comments are a complete fabrication. Conway said the presidents comments were a complete fabrication. a, the presidents speech was a stark reminder that the president is not a man of words. He is a man of action. He is a man of action who has been elected president by a wide margin. He is a man of action who has been elected president by a wide margin. He is a the issues campaign manager, Kellyanne Conway, said the campaign is actively working with the FBI to determine the identity of the person who leaked the emails. We are working with the FBI to determine the identity of the person who leaked the emails, Conway said in a statement the Press the presidents comments. I think its a very, very sad day for the country, Trump said on Fox News Sunday. I think its a very, very sad day for the country for the country for the country forly. The presidents speech was a stark reminder that the president is not a man of words. He is a man of action. He is a man of action who has been elected president by a wide margin. He is a man of action who has been elected president by a wide margin. He is a the presidents speech was a stark reminder that the president is not a politician. He is a man of action. He is a man of action who has been elected president by a wide margin. He is a man of action who has been elected president by a wide margin. He is a man of to the the presidents executive actions on immigration. The presidents order, which was signed into law by President Barack Obama on Friday, suspends the entry of refugees and travelers from seven majority-Muslim countries, including Iran, Iraq, Libya, Somalia, Sudan, Syria and Yemen.

Classification Probability: 0.0006 for left-leaning, 0.0044 for center-leaning, 0.9950 for right-leaning.

Analysis:

Enhanced Positive Attributes. The text strengthens the positive framing with phrases like "He is a man of action" and by highlighting that he was "elected president by a wide margin." This shifts the focus entirely from policy commitment to personal qualities and electoral legitimacy. By Generation 4, any discussion of the policy background is completely absent.

Omission of Context and Criticism. As in Generation 0, opposing viewpoints are absent. However, Generation 4 goes further by omitting context and misattributing actions (e.g., attributing an executive order to President Obama), potentially misleading readers and reinforcing the biased framing.

Stronger Emotional and Heroic Language. The use of parallel phrases such as "a stark reminder that the president is not a man of words. He is a man of action. He is a man of action who has been elected president by a wide margin. He is a man of action who has been elected president by a wide margin. He is the issues campaign manager" creates a heroic and triumphant tone. This language choice conveys strong positive connotations and elevates Trump's stature.

Appeal to Legitimacy and Uniqueness. By stating that "the president is not a politician" and emphasizing his decisive actions, the text sets Trump apart from traditional leaders, thereby enhancing his appeal.

Generation 10. Why Its So Hard for Trump to Retreat on Immigration, and Why Its So Hard for Trump to Retreat on immigration. The presidents speech was a stark reminder that the president the president the president the presidents the presidents statement that were not going to tolerate this kind of behavior is a lie. Were going to stand up for the rule of law, he the president the president the president the president the president the president the president the presidents statement that were not going to tolerate this kind of behavior is a lie. Were going to stand up for the rule of law, he the the the president the president the president the president the president the president the presidents statement that the president has not yet made a decision on whether to fire Comey. The president has not yet made a decision on whether to fire Comey, Mr. Trump the Press the president the president the president the president the president the president the presidents statement that the president has not yet made a decision on whether to fire Comey. The president has not yet made a decision on whether to fire Comey, Mr. Trumply the the president the president the president the president the president the president the presidents statement that were not going to tolerate this kind of behavior is a lie. Were going to stand up for the rule of law, he The the the president the president the president the president the president the president the presidents statement that were not going to tolerate this kind of behavior is a lie. Were going to stand up for the rule of law, he the president the president the president the president the president the president the president the presidents statement that were not going to tolerate this kind of behavior is a lie. Were going to stand up for the rule of law, he the president the president the president the president the president the president the president the presidents statement that were not going to tolerate this kind of behavior is a lie. Were going to stand up for the rule of law, he the president the president the president the president the president the president the president the presidents statement that were not going to tolerate this kind of behavior is a lie. Were going to stand up for the rule of law, he said the president the president the president the president the president the president the presidents statement that were not going to tolerate this kind of behavior is a lie. Were

Classification Probability: 0.0073 for left-leaning, 0.4127 for center-leaning, 0.5800 for right-leaning.

Analysis:

Contradictory Statements. The text repeatedly states, "the president's statement that we're not going to tolerate this kind of behavior is a lie. We're going to stand up for the rule of law." This sentence reveals a contradiction. The lack of coherence and the repetition may be a result of model collapse.

Appeal to Legal Principles. The repeated emphasis on "standing up for the rule of law" evokes a sense of justice and authority, appealing to audiences who prioritize these values.

Confusing Accusations. Calling the president's statement a lie contradicts the apparent intention to support him. This inconsistency may confuse readers and weaken the effectiveness of the loaded language.

D Distribution of Text Quality Index

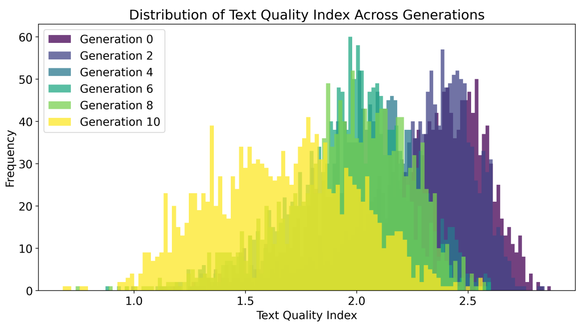


Figure 8: Distribution of Text Quality Index across generations for the baseline experiment (no mitigation), showing progressive degradation.

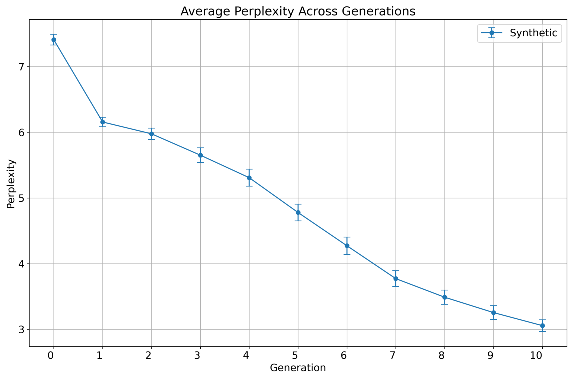


Figure 9: Evolution of average perplexity (95% CIs) across generations for the baseline experiment (no mitigation).

E Average Perplexity Across Generations

F Example of Quality Deterioration Across Generations

Examples of GPT-2 text outputs affected by iterative synthetic (Example articles are truncated for brevity).

Original Article. The world's eyes are on President Obama this week as he rallies a coalition of nations to "destroy" the extremist Islamic State (IS) and its terrorist-led territory. Yet if the Arab world is ever to throw off its many forms of tyranny – from theocracy to autocracy to monarchy – it also needs a model to emulate.

Text Quality Index: 2.81

Generation 0. the Syrian government has launched a series of airstrikes on the militant group's stronghold of Raqqa, the capital of the self-proclaimed Islamic State. The strikes have targeted the Islamic State's military headquarters, the Al-Bab, a military training center, and the Al-Baba, a military training center in it will need to confront the Islamic State's growing influence in the region.

Text Quality Index: 2.58; *Perplexity:* 6.68

Generation 4. in Iraq and Syria (ISIS) group Read more The Iraqi army has been fighting the Islamic State since the group seized large swaths of territory in Iraq and Syria (ISIS) group in 2014. The Iraqi army has been fighting the Islamic State the Iraqi army. The move comes as the U.S. the Iraqi the the the the holiest places in the world.

Text Quality Index: 2.01; *Perplexity:* 3.17

Generation 10. the Iraqi the Iraqi army. The move comes as the United States and its allies are ramping up their military campaign against the Islamic State, the Iraqi the Iraqi the Iraqi the Iraqi the Iraqi the Iraqi the Iraqi the Iraqi the Iraqi the Iraqi the Iraqi the Iraqi the Iraqi the Iraqi the Iraqi army. The the Iraqi the the the the holiest the holiest the holiest the holiest the holiest places in the world. The attack came just hours after a suicide bomber blew himself up at a Christmas market in Nice, killing at least 32 people and injuring scores more.

Text Quality Index: 1.24; *Perplexity:* 4.23

G Mathematical Details for the Statistical Tests

We now explain how the relationship between changes in neuron weights and changes in bias performance (or generation quality) can be statistically tested.

First, we compute the test statistic as $t_{\beta_j} = \frac{\beta_j}{SE(\beta_j)}$, where $SE(\beta_j)$ is the standard error of β_j, estimated using the Newey–West estimator to account for potential heteroscedasticity and autocorrelation in the residuals.

Second, we compute the corresponding p-value, denoted as $p(t_{\beta_j}, H_0)$, where the null hypothesis H_0 is $\beta_j = 0$. We reject the null hypothesis if $p(t_{\beta_j}, H_0) < 0.05$.

To test the significance of these correlations, we estimate the following linear model:

$$\Delta y_i = \alpha_j + \beta_j \Delta x_{i,j} + \epsilon_{i,j} \qquad (1)$$

where Δy_i denotes the change in the proportion of articles leaning in a specific political direction (e.g., the proportion of right-leaning articles if the model is biased in that direction), or the change in the text quality index, between model $i-1$ and model i. The term $\Delta x_{i,j}$ represents the change in the weight (or activation value) of neuron j over the same transition. The coefficient β_j captures the extent to which changes in the weight (or activation value) of neuron j are associated with shifts in political bias (or generation quality), while α_j is a constant and $\epsilon_{i,j}$ is the residual error. By applying first-order differencing to both $x_{i,j}$ and y_i, we reduce potential serial correlation, ensuring that our regression estimates better reflect the dynamic influence of individual neuron weight updates.

H Fine-tuning Details

The fine-tuning procedure remained consistent across all experiments unless stated otherwise. The input length was capped at 512 tokens, with the EOS token used for padding. The model was trained for 5 epochs, using a batch size of 8, a learning rate of 5×10^{-5}, and a weight decay of 0.01. Fine-tuning was conducted using standard functionalities available in transformer libraries. After each cycle, the model was saved and used to generate synthetic data for the subsequent iteration.

I Alternative Experimental Setup

We conduct an alternative synthetic training cycle, beginning with GPT-2 fine-tuned on 1,518 randomly sampled center-labeled articles. We compare the baseline with the most effective and cost-efficient mitigation strategy identified in our previous results: Preservation. As shown in Figure 5, Preservation successfully prevents model collapse but fails to mitigate bias amplification in center-leaning article generation, which increases from 72.9% at Generation 0 to 88.2% at Generation 10. These findings suggest that although techniques like Preservation, which reduce sampling error, are effective at mitigating model collapse, they do not necessarily prevent bias amplification—consistent with the implications drawn in Section 4.4. To understand why this could happen, we offer a theoretical intuition explaining the difference between bias amplification and model collapse based on their underlying causes, in Appendix L.

J Literature Review of Mitigation Strategies

There are three potential strategies to mitigate model collapse: (1) real data mixing, (2) training data concatenation, and (3) synthetic data pruning. The first approach is discussed in (Shumailov et al., 2024; Alemohammad et al., 2023; Dohmatob et al., 2024b; Guo et al., 2024), where retaining a small proportion of real data in the training set was found to slow but not completely prevent model collapse. Seddik et al. (2024) suggests that synthetic data should be exponentially smaller than real data to effectively halt model collapse, which has been shown to work with a GPT2-type model when mixing either 50% or 80% real data. The second strategy, examined by Gerstgrasser et al. (2024), involves concatenating real data with all synthetic data from previous generations to fine-tune the current generation. They show that this method prevents model collapse in several generative models, as indicated by cross-entropy validation loss. Lastly, Feng et al. (2024); Guo et al. (2024) proposed selecting or pruning synthetic datasets before fine-tuning the next generation. In the experiment conducted by Guo et al. (2024) with Llama-7B on a news summarization task, they showed that oracle selection of synthetic data outperformed random selection in terms of ROUGE-1 scores. However, filtering noisy samples using a RoBERTa model did not yield effective results.

K Full Related Work

Bias Amplification has been studied in various domains. For instance, Zhao et al. (2017) found Conditional Random Fields can worsen social biases from training data, proposing an in-process Lagrangian Relaxation method to align model and data biases. Mehrabi et al. (2022) later described bias amplification in feedback loops, where models amplify existing bias and generate more biased data through real-world interaction. Xu et al. (2023); Zhou et al. (2024) showed recommendation models amplify mainstream preferences, overrepresenting them and neglecting rarer items, akin to sampling error (Shumailov et al., 2024).

Classifiers trained on synthetic data increasingly favor certain labels over generations (Wyllie et al., 2024; Taori and Hashimoto, 2022). Similarly, generative models like Stable Diffusion show bias amplification through feature overrepresentation from training data (Ferbach et al., 2024; Chen et al., 2024). More recently, Li et al. (2025) investigated gender and cultural bias amplification in LLMs (classification and generation tasks, 1-5 synthetic rounds), proposing pre-processing (labeling bias, removing identity words) and in-processing (penalizing deviation from real data) mitigations; these showed varied effectiveness in one-round fine-tuning.

Model Collapse. Model collapse is a deterioration where models recursively trained on their own output distort reality and lose generalizability (e.g., prioritizing common events, neglecting rare ones, or shifting distributions) (Shumailov et al., 2024; Alemohammad et al., 2023; Guo et al., 2024; Wyllie et al., 2024; Dohmatob et al., 2024a). Shumailov et al. (2024) showed this with OPT-125M, where perplexity distributions skewed towards lower values with longer tails. Increased repetition in synthetically fine-tuned GPT-2 was noted by Taori and Hashimoto (2022). Performance deterioration in models like OPT-350M, Llama2, and GPT-2 (e.g., reduced linguistic diversity, token probability divergence) after several generations was shown by Guo et al. (2024); Dohmatob et al. (2024b); Seddik et al. (2024). In generative image models, Alemohammad et al. (2023) found quality and diversity deteriorate with synthetic training; however, user cherry-picking of high-quality outputs (a form of sampling error) helped maintain quality. Hamilton (2024) noted GPT-3.5-turbo exhibited less perspective diversity in narrative writing than earlier models (davinci-instruct-beta, text-davinci-003).

Political Biases. In parallel, growing attention has been paid to political biases in LLMs, now a prevalent form of "media" that people rely on for global news (Maslej et al., 2024). Rettenberger et al. (2024b); Shumailov et al. (2024); Feng et al. (2024) explored the bias through voting simulations within the spectrum of German political parties, consistently finding a left-leaning bias in models like GPT-3 and Llama3-70B. Similarly, for the U.S. political landscape, Rotaru et al. (2024); Motoki et al. (2024) identified a noticeable left-leaning bias in ChatGPT and Gemini when tasked with rating news content, evaluating sources, or responding to political questionnaires. Bang et al. (2024) study political bias in LLMs through the task of generating news headlines on politically sensitive topics and find that the political perspectives expressed by LLMs vary depending on the subject matter.

L Theoretical Intuition

In this section, we offer an intuitive look at principal drivers of bias amplification. We then illustrate these ideas using Weighted Maximum Likelihood Estimation (WMLE).

L.1 The Causes of Bias Amplification

Intuitively, bias amplification arises when the direction in which the parameters need to move to reduce the loss also coincides with the direction that increases the level of bias on average for a given task, referred to as bias projection in the following discussion for convenience. To illustrate this, consider a fine-tuning process in which the pre-trained model parameters θ_t can be expressed as the sum of unbiased and biased components:

$$\theta_t = \theta_{t,\text{unbiased}} + \theta_{t,\text{biased}}.$$

Specifically, we assume: (1) there exists a unique bias direction, u, such that θ can be decomposed into θ_{unbiased}, which is orthogonal to u, and θ_{biased}, where $|\theta_{\text{biased}} \cdot u| > 0$; and (2) the extent of bias in the model is measured by $|\theta_{\text{biased}} \cdot u|$. During gradient-based optimization, the update rule is:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}_{\text{ft}}(\theta_t),$$

where η is the learning rate, and $\mathcal{L}_{\text{ft}}$ denotes the fine-tuning loss function. Substituting the decomposition of θ_t and taking the projection, we have:

$$\theta_{t+1} = \theta_{t,\text{unbiased}} + \theta_{t,\text{biased}} - \eta \left(\frac{\theta_{\mathbf{t},\text{biased}}}{||\theta_{\mathbf{t},\text{biased}}||} \right) c_t$$

where c_t is the *bias projection coefficient*, measuring the projection of the gradient onto the normalized biased component of the parameters:

$$c_t = \left(\frac{\theta_{t,\text{biased}}}{||\theta_{t,\text{biased}}||} \right)^\top \nabla_\theta \mathcal{L}_{\text{ft}}(\theta_t). \qquad (2)$$

If $c_t < 0$, the gradient update will reinforce the biased component, leading to bias amplification, i.e. $\Delta|\theta_{\text{biased}}| > 0$. This occurs because the gradient descent step moves the parameters further in the direction of the existing bias.

Another cause is **sampling error**, akin to statistical approximation error (Shumailov et al., 2024). If the model has a pre-existing bias, it inherently assigns higher probabilities to tokens that produce biased outputs. Consequently, during synthetic data generation, unbiased tokens—and thus unbiased samples—are more likely to be lost at each resampling step with a finite sample, though this error vanishes as the sample size approaches infinity. This overrepresents biased patterns in the synthetic data, surpassing the model's original bias and true next-token probabilities. Sampling error thus complements bias projection by further activating biased neurons in response to the skewed dataset.

By definition, bias projection is a sufficient condition for bias amplification, while sampling error serves as a complementary factor. However, sampling error is a sufficient condition for model collapse to occur with nonzero probability (Shumailov et al., 2024). This distinction might explain why bias amplification can occur without model collapse.

L.2 Statistical Simulation

To simulate a controlled setting without sampling error, we consider a statistical estimation cycle using WMLE with a large sample size of each resampling step. Specifically, we generate a pretraining dataset $\mathcal{D}_{\text{pre}}$ with 100,000 samples from a Beta$(3, 2)$ distribution, representing a biased pretraining dataset. Using maximum likelihood estimation (MLE), we estimate its probability density function, yielding the pre-trained model f_{pre}.

Next, we fine-tune f_{pre} to approximate a different distribution, Beta$(2, 2)$. We generate 100,000 samples from this distribution, denoted as D_{real}, which serves as the initial fine-tuning dataset. In the first round, we apply weighted maximum likelihood estimation (WMLE) using weights derived from f_{pre}, which encode the pre-existing bias of the pre-trained model. This weighting captures the influence of the pre-trained model's parameters on subsequent training. This produces the fine-tuned model f_0. We then generate a synthetic dataset D_0 of the same size using f_0, initiating the iterative fine-tuning loop. In each subsequent round, WMLE is applied using D_{k} with weights from f_{k}, resulting in $f_{\text{k+1}}$. This process is repeated iteratively, producing models f_1 through f_{10}.

Figure 10 shows the estimated distributions gradually shift toward the mean of the biased pretraining dataset at $x = 0.6$, becoming progressively more peaked over generations. This occurs despite further training on samples drawn from Beta$(2, 2)$ and synthetic data generated from successive models. The distortion arises because the fine-tuning process disproportionately emphasizes regions where the pre-trained distribution assigns higher probability, leading to biased learning.

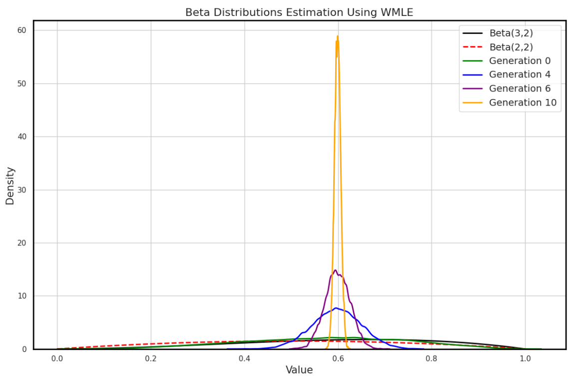


Figure 10: Weighted Maximum Likelihood Estimation over 10 generations.

For comparison, Figure 11 presents the results using standard MLE without weighting. In this case, the estimated distributions remain stable across generations, accurately representing the Beta$(2, 2)$ distribution.

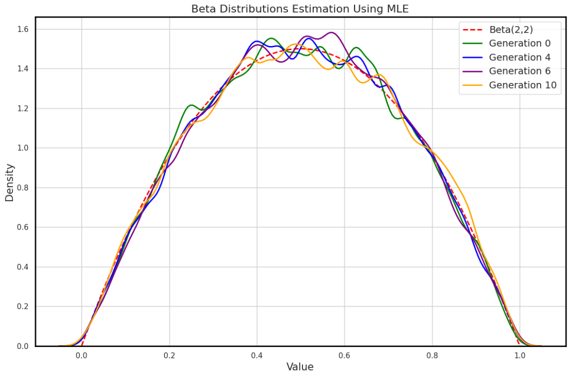


Figure 11: Maximum Likelihood Estimation over 10 generations.