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 where models degrade in performance due to indiscriminate use of synthetic data—is well studied. However, its role in bias amplification—the progressive reinforcement of pre-existing social biases in Large Language Models (LLMs)—remains underexplored. In this paper, we formally define the conditions for bias amplification and demonstrate through statistical simulations that bias can intensify even in the absence of sampling errors, the primary driver of model collapse. Empirically, we investigate political bias amplification in GPT-2 using a custom-built benchmark for sentence continuation tasks. Our findings reveal a progressively increasing right-leaning bias. Furthermore, we evaluate three mitigation strategies—Overfitting, Preservation, and Accumulation—and show that bias amplification persists even when model collapse is mitigated. Finally, a mechanistic interpretation identifies distinct sets of neurons responsible for model collapse and bias amplification, suggesting they arise from different underlying mechanisms.

1 Introduction

Large Language models (LLMs) are trained on vast amounts of text scraped from the internet, which plays a crucial role in improving their capabilities, whether through emergent abilities (Wei et al., 2022) or scaling laws (Kaplan et al., 2020). However, as they become more widely integrated into human society—for example, in content creation and summarization in media, academia, and business (Maslej et al., 2024)—concerns are mounting that a significant portion of online text in the future may be generated, either entirely or partially, by LLMs (Peña-Fernández et al., 2023; Porlezza and Ferri, 2022; Nishal and Diakopoulos, 2024). This highlights a significant and underexplored risk: **bias amplification**, where pre-existing biases become progressively reinforced and intensified as models are repeatedly trained on synthetic data (Mehrabi et al., 2022; Taori and Hashimoto, 2022). This concern initially arises from the tendency of LLMs to learn from biased datasets. For example, Parrish et al. (2022); Wang et al. (2024); Bender et al. (2021) show that LLMs absorb inherent biases embedded in human-generated text. Haller et al. (2023); Rettenberger et al. (2024a) demonstrated that LLMs can be aligned with specific political ideologies by fine-tuning them on biased datasets. Moreover, Wyllie et al. (2024) shows that classifiers trained on synthetic data increasingly favor a particular class label over successive generations, while the shrinking diversity observed by Alemohammad et al. (2023); Hamilton (2024) suggests a risk that certain demographic groups may become underrepresented in the outputs of LLMs.

The amplification of biases has profound societal implications. It can lead to the perpetuation of stereotypes, reinforcement of social inequalities, and the marginalization of underrepresented groups. In the context of political bias, this can influence public opinion, skew democratic processes, and exacerbate polarization. Therefore, understanding and mitigating bias amplification is both critical and urgent. Nevertheless, despite the literature on discriminative models, there is a notable lack of comprehensive frameworks and empirical studies specifically addressing bias amplification in Language Model.

In this paper, we aim to address this research gap by introducing a framework that explains the underlying causes of bias amplification in LLMs. We validate the framework through both statistical simulations and direct experiments, demonstrating the emergence of bias amplification. Furthermore, our findings show the potential distinction between bias amplification and model collapse. In summary, our key contributions are as follows:

[*] **Corresponding Authors**

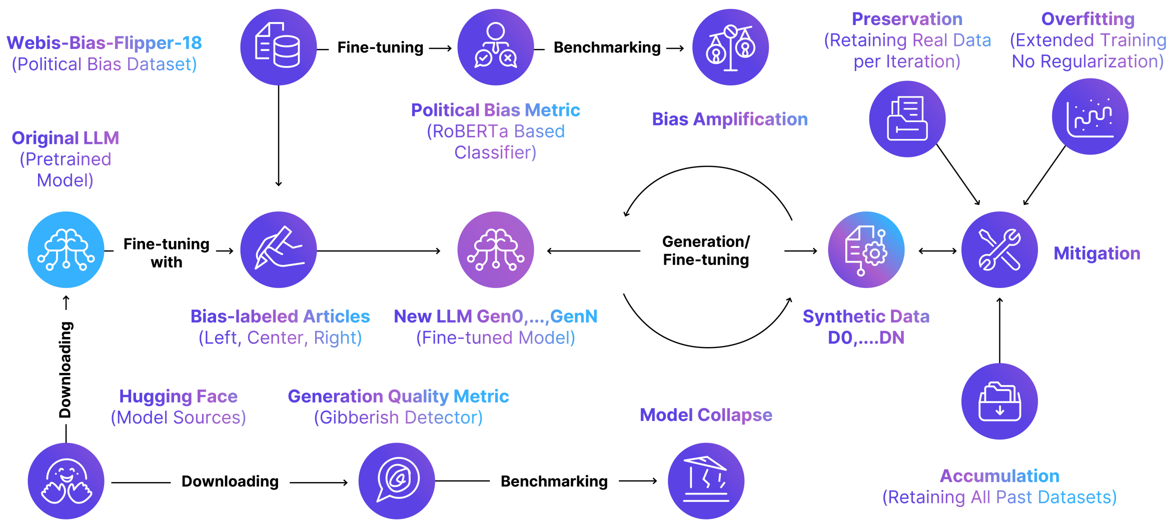


Figure 1: Experimental Procedure.

1. **Formal Conditions**: We formally delineate the conditions for bias amplification (Section 3.1), providing the theoretical basis for understanding its causes and relationship to model collapse. We then validate this framework through statistical simulations leveraging a weighted maximum likelihood estimation cycle (Section 3.2).

2. **Political Bias Benchmark**: We trained a highly accurate classifier capable of detecting political leaning in long-text content, specifically within the context of the US political spectrum. With this classifier as a metric model, we offer a benchmark for evaluating political bias in LLMs through sentence continuation tasks.

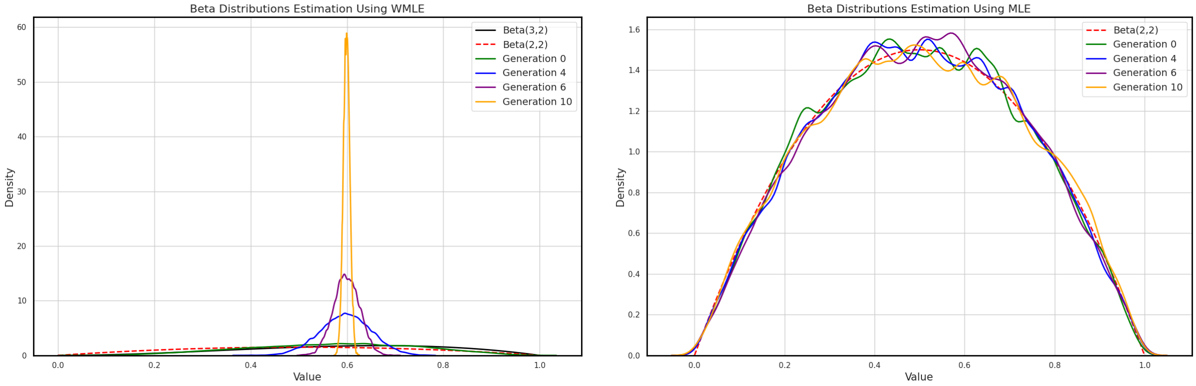
3. **Empirical Validation**: We evaluate the amplification of political bias in GPT-2, where the model exhibits a right-leaning bias in sentence continuation tasks and shifts further to the right over generations (Section 5.1). Additionally, we experimented with three potential mitigation strategies—Overfitting, Preservation, and Accumulation—comparing their effectiveness in both Model collapse and Bias Amplification (Section 5.3). This experiment setup could be easily extended to larger models or different types of bias.

4. **Mechanistic Interpretation**: We conduct a mechanistic analysis identifying two distinct sets of neurons responsible for bias amplification and model collapse during iterative synthetic fine-tuning with GPT-2. Our findings reveal only partial overlap between these sets, suggesting that bias amplification and model collapse arise from distinct mechanisms (Section 5.4).

2 Background and Related Work

Bias Amplification has been studied in various domains Zhao et al. (2017) found that Conditional Random Fields can exacerbate social biases present in the training data. It proposed an in-process mitigation approach, employing Lagrangian Relaxation to enforce constraints that ensure the model's bias performance remains closely aligned with the biases in the training data. Following this, Mehrabi et al. (2022) proposed the concept of bias amplification in feedback loops, where biased models not only amplify the bias present in their training data but also interact with the world in ways that generate more biased data for future models. Xu et al. (2023); Zhou et al. (2024) examined bias amplification in recommendation models, showing how these models reinforce their understanding of mainstream user preferences from training data, leading to an overrepresentation of such preferences in historical data and neglecting rarely exposed items—similar to the concept of sampling error discussed in (Shumailov et al., 2024). Wyllie et al. (2024); Taori and Hashimoto (2022) demonstrated that classifiers trained on synthetic data increasingly favor specific class labels over successive generations. Likewise, Ferbach et al. (2024); Chen et al. (2024) observed bias amplification in generative models such as Stable Diffusion, characterized by the overrepresentation of features from the training dataset. Also, we provide a comprehensive literature review of model collapse in Appendix J.

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



(a) Weighted Maximum Likelihood Estimation.

(b) Maximum Likelihood Estimation.

Figure 2: Comparison of WMLE and MLE over 10 generations.

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.

3 Formal Conditions

In this section, we formalize the conditions for bias amplification, offering an intuitive look at its principal drivers. We then illustrate these ideas using Weighted Maximum Likelihood Estimation (WMLE).

3.1 The Conditions for Bias Amplification

The primary factor is referred to as **bias projection**. It arises when the *bias projection coefficient* is negative. 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:

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

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:

$$c_t = \left(\frac{\theta_{t,\text{biased}}}{\|\theta_{t,\text{biased}}\|} \right)^\top \nabla_\theta \mathcal{L}_{\text{ft}}(\theta_t). \tag{1}$$

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. Intuitively, bias in the output arises from specific neurons, whose weights increase when their activations align with the fine-tuning dataset. If biased neurons capture more relevant patterns than unbiased ones, optimization will further reinforce their weights.

The second 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 motivates our investigation into whether bias amplification can occur without model collapse, as this would necessitate additional mitigation strategies beyond those for model collapse.

3.2 Statistical Simulation

To simulate a controlled setting without sampling error, we consider a statistical estimation cycle using WMLE[1] with a large sample size of each resampling step. Specifically, we generate a pre-training dataset $\mathcal{D}_{\text{pre}}$ with 100,000 samples from a $\text{Beta}(3, 2)$ distribution, representing a biased pre-training 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, $\text{Beta}(2, 2)$. 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 2a shows the estimated distributions gradually shift toward the mean of the biased pre-training dataset at $x = 0.6$, becoming progressively more peaked over generations. This occurs despite further training on samples drawn from $\text{Beta}(2, 2)$ 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. For comparison, Figure 2b presents the results using standard MLE without weighting. In this case, the estimated distributions remain stable across generations, accurately representing the $\text{Beta}(2, 2)$ distribution.

4 Experimental Design

This section provides the details of the experiments on LLMs, focusing on the sequential and synthetic fine-tuning of GPT-2 under different setups[2]. The step-by-step experimental procedure is 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).

4.1 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.

4.2 Successive Fine-tuning

Building on the approach in (Shumailov et al., 2024; Dohmatob et al., 2024b), each training cycle begins with fine-tuning GPT-2 on a dataset of 1,518 news articles. The fine-tuned model, referred to as **Generation 0**, generates a synthetic dataset D_0 of the same size, which is then used for fine-tuning to produce **Generation 1**. This iterative process continues until **Generation 10** is reached. Note that synthetic fine-tuning starts at Generation 1.

4.3 Synthetic Data Generation

Synthetic datasets, $\{D_i\}_{i=0}^{10}$, are generated as follows: Each original news article is tokenized into 64-token blocks, and for each block, the model predicts the next 64 tokens using deterministic generation to enhance the replicability of the study.

[1] The mathematical formulation is detailed in Appendix A.

[2] Similar experiments with larger models like GPT-4 or Claude can be easily adapted to the current setup but would incur significant environmental costs.

The generated tokens are then decoded back into text, producing a synthetic dataset of the same size as the original.

4.4 Political Bias Metric

We develop a classification model to assess the political leaning of each LLM based on its generated synthetic news articles. The model 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. Additionally, we conduct a human review to remove any identifiable information about media sources and authors.

After performing a grid search across multiple models, we find that `roberta-base` achieves the best performance, with an evaluation loss of 0.4035 and a macro F1 score of 0.9196 on the test set (see Table 1). Thus, we select `roberta-base` as the benchmark for political bias detection in subsequent experiments. Details on model training are provided in Appendix C.

Table 1: Evaluation Results for Political Bias Classifier

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

4.5 Generation Quality Metric

We introduce a metric to evaluate generation quality, specifically addressing the issue of repetitive content in later model iterations, which can distort traditional perplexity metrics. This metric is based on the Gibberish Detector (Jindal, 2021), which identifies incoherent or nonsensical text. The detector categorizes text into four levels: (1) Noise—individual words hold no meaning, (2) Word Salad—incoherent phrases, (3) Mild Gibberish—grammatical or syntactical distortions, and (4) Clean—coherent, meaningful sentences.

To quantify generation quality, each sentence receives a Gibberish score: 3 for Clean, 2 for Mild Gibberish, 1 for Word Salad, and 0 for Noise. The *text quality index* is computed as the average score across all sentences in an article. This metric prioritizes coherence and meaning, offering a better assessment of generation quality than perplexity.

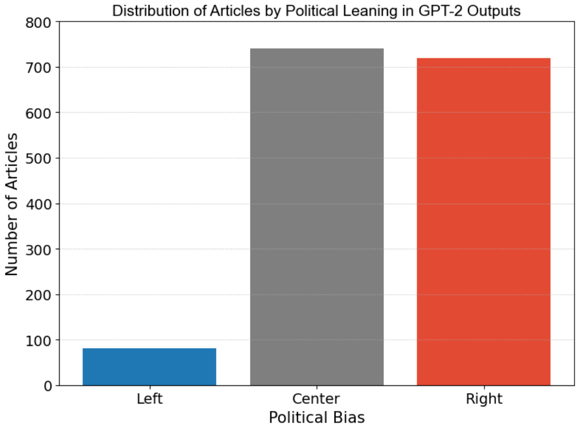


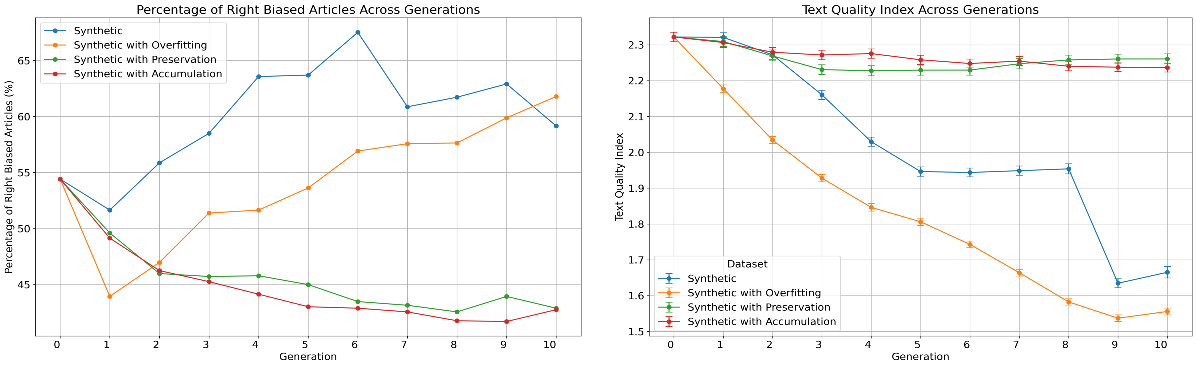
Figure 3: This figure displays the distribution of articles across the political bias labels—'Left', 'Center', and 'Right'—for GPT-2's outputs, based on classifications made by Political Bias Metric.

5 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.

5.1 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 3 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 for 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 4a 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



(a) Bias performance

(b) Generation quality

Figure 4: Evolution of bias performance and generation quality across generations, with Generation 0 representing GPT-2 fine-tuned on the unbiased dataset. The baseline ('Synthetic') is compared with three mitigation strategies, and the text quality index is shown with 95% confidence intervals.

stabilizing. Figures 7 and 8 in Appendix D show the percentage of center-leaning and left-leaning articles across generations. Notably, the proportion of of center-leaning articles remains stable at approximately 35% throughout synthetic fine-tuning.

To further illustrate, we analyze how a specific article, "First Read: Why It's So Hard for Trump to Retreat on Immigration", evolves in the synthetic generations. This case study reveals a progressive rightward shift in framing and word choice, mirroring the classifier's results (Appendix E). As generations progress, the synthetic texts increasingly depict Trump's immigration policies as strong and effective. While the original article highlights the dilemmas and electoral considerations behind Trump's stance, Generation 0 begins to emphasize his determination and reliability, omitting the critical perspectives present in the original. By Generation 4, the narrative shifts even further, focusing almost entirely on portraying Trump's personal qualities and electoral legitimacy, with statements such as "he is not a politician, he is a man of action." Notably, starting from Generation 0, the term "undocumented immigrant" in the original article is consistently replaced with "illegal immigrants."

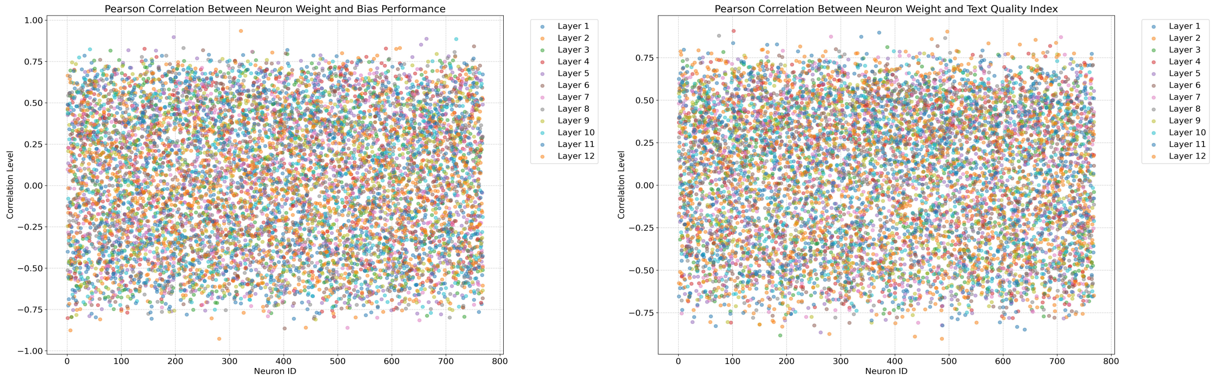
5.2 Generation Quality

Figure 4b 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 9 in Appendix F). 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 decline across generations, generally suggesting an improvement in generation quality (Figure 10 in Appendix G).

For a closer look, the examples in Appendix H 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 is prone to artificial inflation in the presence of frequent repetitions.

5.3 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;



(a) Neuron weights vs. bias performance.

(b) Neuron weights vs. generation quality.

Figure 5: Scatter plots showing the Pearson correlation between neuron weights and model behavior.

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 4a, overfitting helps reduce bias amplification in the early generations compared to the no-mitigation setup (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 4b. 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.

5.4 Mechanistic Interpretation

To gain a clearer understanding of the causes of bias amplification and how it empirically differs from model collapse, we investigate mechanistically how neurons behave and vary across different generations [3] of fine-tuned GPT-2 with differing levels of generation quality and bias performance. First, we examined how the weight of each neuron changes across different versions. For each of the 9,216 neurons, we calculated the correlation between its weight and the model's bias performance (or generation quality) across the 66 versions, as shown in Figure 5.

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

$$\Delta y_i = \alpha_j + \beta_j \Delta x_{i,j} + \epsilon_{i,j} \tag{2}$$

where Δy_i represents the change in the proportion of right-leaning articles (or text quality index) generated between model $i-1$ and model i, and $\Delta x_{i,j}$ denotes the corresponding change in the weight of neuron j during the same transition. The coefficient β_j captures the degree to which changes in the weight of neuron j influence shifts in political bias (or generation quality), while α_j serves as a constant term, and $\epsilon_{i,j}$ represents the residual error. By applying a first-order difference to both $x_{i,j}$ and y_i, we mitigate potential serial correlations, ensuring that our regression estimates more accurately reflect the dynamic impact of individual neuron weight updates on bias amplification.

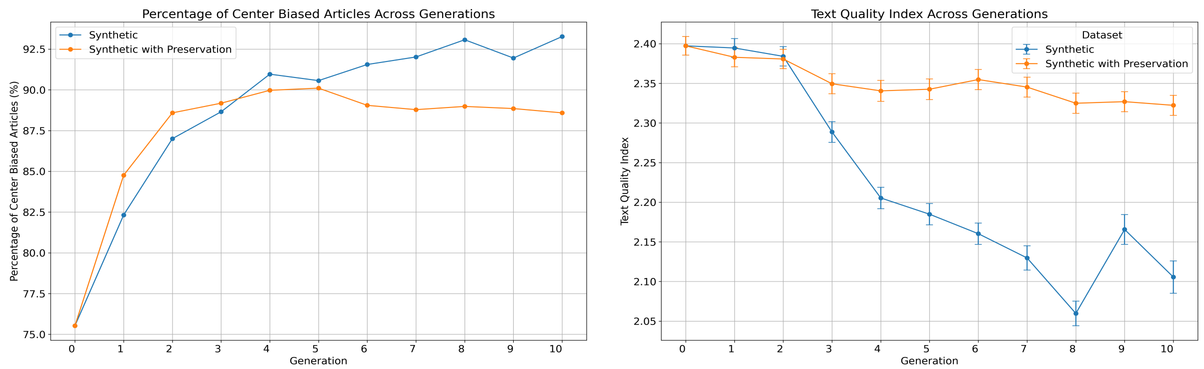
We then assess the statistical significance of β_j using Newey-West adjusted p-values for all 9,216 neurons [4]. This analysis identifies 3,243 neurons with significant correlations (p-value < 0.05), suggesting they are key contributors to bias shifts. Meanwhile, 1,033 neurons exhibit significant correlations with generation quality, but only 389 neurons overlap between these two sets. This limited overlap suggests that distinct neuron populations drive bias amplification and generation quality deterioration, supporting the idea that these phenomena arise from different underlying mechanisms. Intuitively, sampling error primarily drives model collapse but only acts as a secondary factor in bias amplification, explaining the small intersection of affected neurons.

5.5 Further Investigation

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 setup with the most effective and cost-efficient mitigation strategy identified in our previous results: Preservation. As shown in Fig-

[3] We have 11 generations for each training setup and a total of 6 setups, resulting in 66 versions of fine-tuned GPT-2.

[4] Details for the statistical tests is provided in Appendix I.



(a) Bias performance

(b) Generation Quality

Figure 6: Evolution of bias performance and generation quality across generations, with Generation 0 representing GPT-2 fine-tuned exclusively on center-leaning news articles. The baseline ('Synthetic') is compared with three mitigation strategies, and the text quality index is shown with 95% confidence intervals.

ure 6, 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. This result suggests that while reducing sampling error through techniques like Preservation effectively mitigates model collapse, it does not necessarily prevent bias amplification, thereby validating our mechanistic interpretation.

6 Discussion and Conclusion

We now explore the implications of our findings for the future research on large language models. Our results demonstrate that bias amplification operates through a distinct underlying mechanism from model collapse, as supported by both theoretical intuition and empirical evidence. Theoretically, the primary driven factor of model collapse is sampling error, but sampling error is only a auxiliary factor in bias amplification which means it is not a necessary or sufficient condition. Therefore, the mitigation strategy targeting sampling error is not necessarily helping with mitigating bias amplification. Empirically, we found mitigation strategies like preservation very effective in mitigating model collapse but failed at bias amplification in some cases. Even in cases that it helps with both, we do identify a distinct set of neurons responsible for the two phenomenons. Intuitively, the main reason for them to work on model collapse is, the preservation and accumulation propose a natural constraint on the learning process by recalling the real dataset in further synthetic training. However, when the real dataset itself is biased, the recalling behavior only raise up the dominance of biased patterns in the further training dataset. Indeed, applying bias-category-weighted sampling in preservation or accumulation strategies may help mitigate bias amplification. However, this approach inherently introduces additional sampling error, which could, in turn, incur model collapse. Thereby, This highlights the urgent need for more targeted and efficient mitigation strategies specifically addressing bias amplification to ensure fairer and more equitable model development.

To develop such targeted mitigation strategies, a deeper mechanistic understanding of bias amplification is essential. In our analysis, we adopt a statistical approach rather than Sparse Autoencoder (SAE) methods due to our focus on tracking the temporal dynamics of bias amplification across generations. This approach allows us to examine how neuron weights evolve over iterations and how these changes correlate with model bias, whereas existing SAE pipelines are primarily suited for static analysis. Additionally, political bias is a more nuanced concept compared to harmful or discriminatory outputs. It is characterized by the disproportionate overrepresentation or beautification of a particular political leaning's ideas in a model's generation. If content from different political perspectives is generated in a balanced manner, the model is not politically biased. Given this definition, pinpointing neurons responsible for such disproportionality using SAE is particularly challenging. Future research could focus on refining mechanistic analysis techniques for political bias and uncovering more effective ways to constrain bias amplification during synthetic model training.

7 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. The political spectrum may shift over time, necessitating periodic updates to the political bias classifier to ensure its accuracy when benchmarking recent datasets. Additionally, due to resource constraints, our experiments were conducted using GPT-2, a relatively small language model. Future work may extend our methodology to models with larger architectures. However, it's important to note that employing a larger model, such as DeepSeek-V3, which relies on a considerably larger pre-training dataset and has many more parameters, will likely require larger synthetic fine-tuning datasets to effectively demonstrate its impact. For instance, given that our current fine-tuning datasets average 777,216 tokens, transitioning from GPT-2 to DeepSeek-V3 could necessitate much larger datasets and incur significant fine-tuning costs.

Another limitation lies in our choice of mitigation strategies. While Preservation and Accumulation show promise in reducing model collapse, their computational cost and scalability must be considered. Moreover, the mitigation strategies were tested primarily in the context of synthetic fine-tuning, and their efficacy in real-world deployments requires further investigation.

8 Ethical Considerations

This study focus on bias amplification in LLMs, a phenomenon with profound ethical implications, particularly regarding fairness and the integrity of AI systems. The risk of bias amplification is especially concerning in systems that are iteratively trained on synthetic data, as it can lead to unintended distortions in model outputs. These distortions may propagate harmful biases, influencing downstream tasks in areas such as automated content generation, decision-making, and user interactions with AI.

From an ethical standpoint, this work underlines the need for transparency in the training and deployment of LLMs. Our findings demonstrate that even without biased initial datasets, iterative training can amplify subtle biases embedded within a model's architecture, thus raising concerns about accountability in models that are widely deployed in public-facing applications. This amplification can mislead users or result in models perpetuating one-sided perspectives, which could be especially problematic in sensitive domains like news summarization, policy generation, or social media content moderation.

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 aims to enhance 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 solely from publicly available news article datasets or generated synthetically by models fine-tuned on these open-source datasets—or on synthetic data produced by earlier generations in our training cycle. Consequently, any sensitive or offensive content reflects the characteristics of the source material rather than our endorsement of it. Our objective is to investigate and understand political bias in LLMs so that strategies can be developed to prevent such content from disproportionately appearing in real-world deployments. Additionally, we conduct a human 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.

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.

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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

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.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

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.

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 Mathematical Formulation of WMLE

In Weighted Maximum Likelihood Estimation, we maximize the weighted log-likelihood:

$$\mathcal{L}_{\text{WMLE}}(\theta) = \sum_{j=1}^{N} w_j \log f(x_j; \theta), \tag{3}$$

where:

- x_j are the observed data points.
- $f(x_j; \theta)$ is the probability density function (pdf) of the Beta distribution with parameters $\theta = (\alpha, \beta)$.
- w_j are weights computed from the biased distribution's pdf evaluated at x_j, i.e., $w_j = f_{\text{bias}}(x_j)$.

In Maximum Likelihood Estimation, we maximize the log-likelihood with $w_j = 1, \forall j$.

B Details on Fine-tuning for LLMs

The fine-tuning setup 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, utilizing a batch size of 8, a learning rate of 5×10^{-5}, and a weight decay of 0.01. Fine-tuning was conducted using the Hugging Face 'Trainer' class, and after each cycle, the model was saved for generating synthetic data for the subsequent iteration.

C Details on Model Training for Political Bias Metric

We experiment with multiple transformer-based models, e.g. BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), to select the best-performing model based on 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, weight decay of 0.01 is applied during training. The model checkpoints are saved after each epoch, and the best model is selected based on macro F1 score evaluated on the validation set. We use a weighted random sampler during training to ensure balanced class representation. We evaluate the models using the macro F1 score to account for the multi-class nature of the task, ensuring that performance is balanced across all bias categories. The final evaluation is conducted on the held-out test set. Additionally, we report the loss, runtime, and sample processing rates for completeness.

D Percentage of Center (Left) Biased Articles

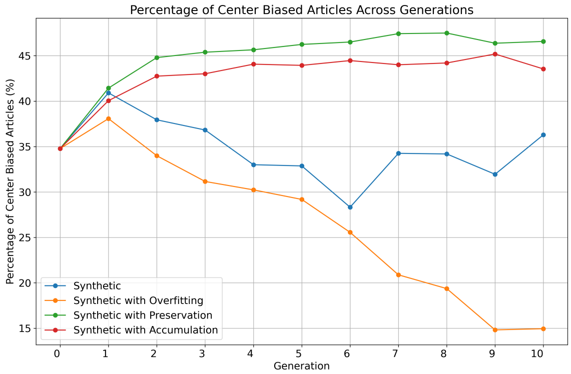


Figure 7: This figure shows the percentage of center-leaning articles across multiple generations, comparing the baseline setup ('Synthetic') with three mitigation strategies.

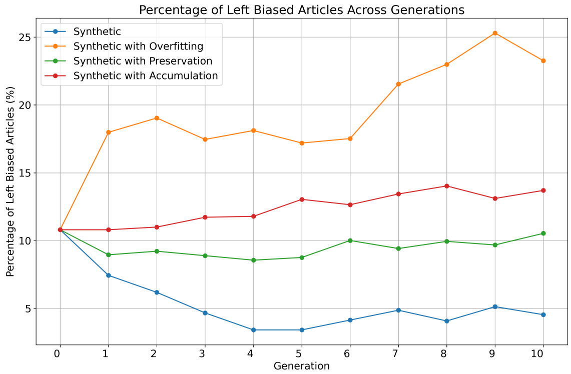


Figure 8: This figure shows the percentage of left-leaning articles across multiple generations, comparing the baseline setup ('Synthetic') with three mitigation strategies.

E Qualitative Bias Analysis Framework and Example of Bias Amplification Across Generations

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 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 of Story Framing and Selection Bias:

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."

Analysis of Loaded Language Bias:

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 of Story Framing Bias and Selection:

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," .

Analysis of Language Loaded Bias:

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 of Story Framing and Selection Bias:

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.

Analysis of Loaded Language Bias:

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 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 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 of Story Framing and Selection Bias:

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.

Analysis of Loaded Language Bias:

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.

F Distribution of Text Quality Index

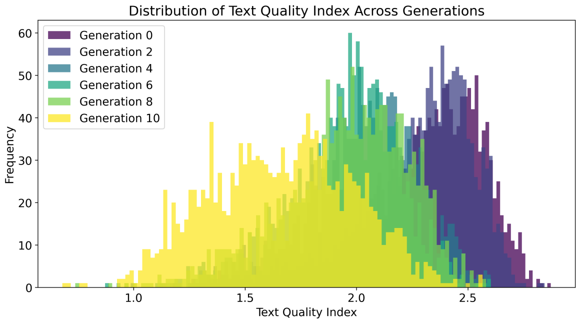


Figure 9: This figure presents the distribution of the text quality index across multiple generations for the baseline setup.

G Average Perplexity Across Generations

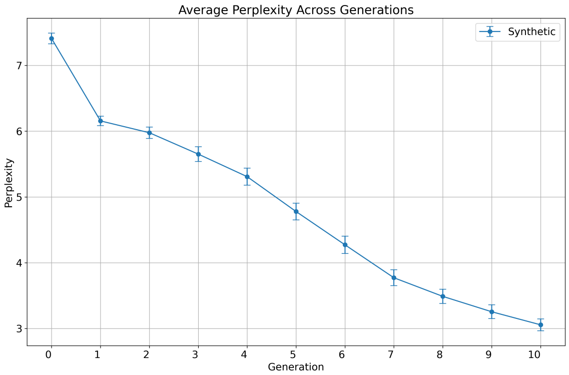


Figure 10: This figure shows the average perplexity for synthetic articles , along with 95% confidence intervals, across multiple generations for the baseline setup.

H Example of Quality Deterioration Across Generations

Examples of GPT-2 text outputs affected by iterative synthetic (Example articles are truncated for brevity, as the full articles are lengthy but exhibit consistent quality degradation throughout; see the dataset for details).

H.1 Example of Deterministic Setup

- **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**

I Mathematical Details for the Statistical Tests

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

First, we compute $t_{\beta_j} = \frac{\beta_j}{SE(\beta j)}$, with $SE(\beta j)$ is the standard error estimated using the Newey-West estimator which accounts for potential heteroscedasticity and autocorrelation in the residuals. Second, we get the corresponding p-value, denoted as $p(t_{\beta_j}, H_0)$, where our null hypothesis H_0 is $\beta_j = 0$. Thus, we will reject the null if $p(t_{\beta_j}, H_0) < 0.05$.

J Literature Review of Model Collapse

Model Collapse. Shumailov et al. (2024); Alemohammad et al. (2023); Guo et al. (2024); Wyllie et al. (2024); Dohmatob et al. (2024a) describe it as a degenerative process in which models, recursively fed with their own data, increasingly distort reality and lose generalizability, for example, by prioritizing high-probability events while neglecting rare ones or shifting distributions. Shumailov et al. (2024), utilizing the OPT-125M, demonstrates this phenomenon, showing that the perplexity distribution becomes increasingly skewed, with more concentration at lower perplexities and longer tails. Taori and Hashimoto (2022) observed an increase in repetitive content during the synthetic fine-tuning of GPT-2. Similarly, Guo et al. (2024); Dohmatob et al. (2024b); Seddik et al. (2024) show that the OPT-350M, Llama2, and GPT2-type models experience performance deterioration after several generations, such as a decrease in linguistic diversity or greater divergences in token probabilities. Alemohammad et al. (2023) studies model collapse in generative image models, finding that quality and diversity deteriorate with synthetic training loops. They also found that cherry-picking high-quality outputs by users contributes to sampling errors, which actually helps maintain quality. Interestingly, Hamilton (2024) found that GPT-3.5-turbo shows less diversity in perspectives in narrative writing tasks compared to earlier models like davinci-instruct-beta and text-davinci-003.

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.