Recursive Training Loops in LLMs: How training data properties modulate distribution shift in generated data?

Grgur Kovač[*1], Jérémy Perez[*1], Rémy Portelas[2], Peter Ford Dominey[3,4], Pierre-Yves Oudeyer[1]

[1]Flowers TEAM, INRIA, FR [2]Ubisoft La Forge, FR [3]INSERM UMR 1093-CAPS, FR [4]Robot Cognition Laboratory, Institute Marey, FR

[*]equal contribution

Correspondence: grgur.kovac@inria.fr

Abstract

Large language models (LLMs) are increasingly used in the creation of online content, creating feedback loops as subsequent generations of models will be trained on this synthetic data. Such loops were shown to lead to *distribution shifts* - models misrepresenting the true underlying distributions of human data (also called *model collapse*). However, how human data properties affect such shifts remains poorly understood. In this paper, we provide the first empirical examination of the effect of such properties on the outcome of recursive training. We first confirm that using different human datasets leads to distribution shifts of different magnitudes. Through exhaustive manipulation of dataset properties combined with regression analyses, we then identify a set of properties associated with distribution shift magnitudes. Lexical diversity is found to amplify these shifts, while semantic diversity and data quality mitigate them. Furthermore, we find that these influences are highly modular: data scrapped from a given internet domain has little influence on the content generated for another domain. Finally, experiments on political bias reveal that human data properties affect whether the initial bias will be amplified or reduced. Overall, our results portray a novel view, where different parts of internet may undergo different types of distribution shift.

1 Introduction

Large Language Models (LLMs) are increasingly contributing to the creation of internet content, being used for journalism (Brigham et al., 2024), coding (Jiang et al., 2024) and generating content on social media (Ferrara et al., 2016). The increasing amount of synthetic, LLM-generated data on the internet introduces a precarious feedback loop: LLMs trained on datasets containing synthetic data will themselves generate data that will be used to train future models. Shumailov et al. (2024) demonstrated that this process, known as **recursive training**, can have detrimental effects, causing models to progressively lose information about the true underlying distributions they are intended to approximate. This results in a gradual change in generated distributions, often accompanied by a reduction in variance. While this has sometimes been described as *model collapse*, we refer to this mismatch between the true and the generated distribution as *distribution shift*, and limit the term collapse to refer to *detrimental* distribution shifts, such as losses in quality or diversity. Such detrimental effects (Guo et al., 2024) as well as bias amplification (Wang et al., 2024a) have been reported in previous work. Their potential societal consequences make it imperative to better understand the dynamics of recursive training.

Internet data varies along a wide range of properties. For instance, certain domains may be associated with a higher ratio of synthetic-to-human data (e.g. GitHub), others with lower quality data (e.g. Reddit), and some with lower diversity data (e.g. specialized forums). If those properties affects the outcome of recursive fine-tuning, we may expect different types of shifts on different parts of the internet.

Given the early stage of research in this field, existing studies often ignore this diversity, relying on simplifying assumptions and focusing on abstracted settings. To the best of our knowledge, how different data properties influence distribution shifts remains largely unexplored, aside from studies examining the ratio of human to synthetic data (Bertrand et al., 2023; Bohacek and Farid, 2023; Kazdan et al., 2024; Martínez et al., 2023b; Zhang et al., 2024b). Filing this gap is therefore crucial to draw a more nuanced and detailed picture of the consequences of recursive fine-tuning.

In this paper, we adopt the iterative chain paradigm used in previous studies (Shumailov et al., 2024; Gerstgrasser et al., 2024a). Our experimen-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32290–32309

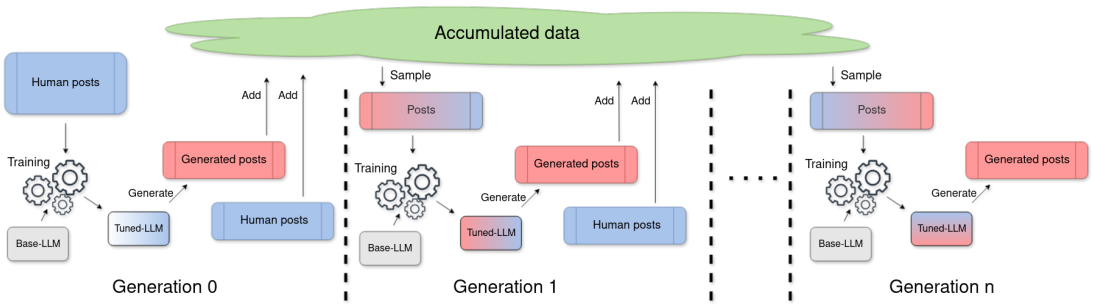


Figure 1: **Iterative chain** In each generation, a fresh base model is fine-tuned on texts sampled from the Accumulated data pool (except generation 0, where it's trained only on human posts). The model generates posts, which are added to the pool alongside some newly sampled human posts.

tal setup is shown in Figure 1. The process begins with fine-tuning a base LLM on a selection of human data (e.g. Reddit posts). This model generates data, which are added to the pool of Accumulated data together with newly sampled human data. In each subsequent generation, a new base model is fine-tuned on data sampled from the Accumulated data pool. This model, in turn, generates new data, which are again added to the pool along with fresh human data. This pipeline allows us to study how generated data evolve across successive generations, with a particular focus on the distribution shift from the first to the final generation.

In our experiments, we study how various properties of human data (e.g. quality, diversity, bias) influence the dynamics of distribution shifts in recursive training chains. We use five datasets spanning three domains (Twitter, Reddit and Wikipedia). First, we confirm that the choice of the dataset greatly influences the consequences of iterative fine-tuning (Section 4.2): while some datasets exhibit sharp decreases in diversity and quality, others are more robust to such shifts. Our second set of experiments (Sections 4.3 and 4.4) aims to uncover specifically which properties of training data mitigate or amplify distribution shifts. We run iterative chain experiments on 800 clusters created from four different datasets, and conduct a series of regression analysis mapping various data properties to the degradation in the quality and diversity of generated texts. We find that lexical diversity is associated with greater degradation, while semantic diversity has the opposite effect. Furthermore, we observe that these influences are highly modular, with generated content being mostly influenced by human data properties from the same domain. This suggests that different internet domains might undergo distinct and relatively independent distribution shifts regardless of models being trained on a mixture of domains. Finally, our last set of experiments focuses on the evolution of political bias. The results indicate that the type of shift observed (bias reduction, amplification or inversion) depends on the political lean of the the human data. This empirically confirms that properties of human data play an important in shaping the outcome of recursive training.

The code for reproducing the simulations, analyses and figures is available on our GitHub[1].

The main contributions of this work are:

- We propose and experimentally confirm the hypothesis that different training datasets lead to different distribution shift dynamics, motivating an investigation on the underlying causes.
- Through an extensive set of experiments (four datasets over three domains), we outline several data properties as influencing distribution shift dynamics.
- We reveal that these influences are highly modular, with generated content being mostly influenced by human data properties from the same domain.
- We find that distribution shifts also occur in terms of political lean, and that the type of shift (bias amplification, reduction or inversion) depends on the political lean of the human data.

2 Related Work

Recursive fine-tuning and model collapse A rapidly growing body of literature has focused

[1]https://anonymous.4open.science/r/ce_llms-9068

on the consequences of recursively training generative models on synthetic data (Schaeffer et al., 2025). The phrase "model collapse", coined in Shumailov et al. (2024), refers to the progressive degradation of models induced by this feedback loop. This phenomenon has been studied both theoretically (Dohmatob et al., 2024a,b; Bertrand et al., 2023; Alemohammad et al., 2023) and empirically, on both generative image models (Martínez et al., 2023b,a; Bohacek and Farid, 2023; Hataya et al., 2022; Alemohammad et al., 2023) and language models (Zhang et al., 2024b; Guo et al., 2023; Kazdan et al., 2024; Briesch et al., 2023; Gerstgrasser et al., 2024b). Theoretical results have provided valuable insights, for instance showing how it is characterized by the disappearance of distribution tails (Dohmatob et al., 2024a,b; Shumailov et al., 2024). Empirical studies have allowed to establish several properties of this phenomenon, such as the role of synthetic-to-real-data ratio (Briesch et al., 2023; Hataya et al., 2022) or strategies for mitigating collapse (Gerstgrasser et al., 2024b; Kazdan et al., 2024; Zhang et al., 2024b). Recently, Wang et al. (2024b) showed that recursive LLM fine-tuning can lead to bias amplification. These works do not systematically evaluate how the properties of the human dataset used in their experiments affect their conclusions. Here, we extend this literature by investigating how those properties impact the outcome of recursive training.

Cultural dynamics in artificial agents The motivation for this research area stems from the observation that human-made technologies have transitioned from passive mediators of cultural evolution (e.g., the printing press) to active generators of cultural content. This shift has been described as the emergence of machine culture - culture mediated or generated by machines (Brinkmann et al., 2023). Understanding the dynamics that shape the evolution of machine-generated content over time is therefore crucial. This has led researchers to study cultural dynamics in populations of reinforcement learning agents (Cook et al., 2024; Schmitt et al., 2018; Team et al., 2021; Prystawski et al., 2023; Nisioti et al., 2022) and of LLMs (Perez et al., 2024a,b; Nisioti et al., 2024; Vallinder and Hughes, 2024; Burton et al., 2024). Our work extends this literature by examining the factors that modulate the evolution of LLM-generated content.

3 Methods

3.1 The iterative chain paradigm

We use the iterative chain paradigm inspired by Shumailov et al. (2024); Gerstgrasser et al. (2024a). Our experimental design is shown in Figure 1. First, a base LLM is fine-tuned on 8000 samples from a human dataset (e.g. Wikipedia articles or Reddit posts). This model generates $4000 * r$ posts, where r is the synthetic-data ratio. Those posts are added to the pool of accumulated data together with $4000 * (1 - r)$ newly sampled human posts. In all subsequent generations, a new base model is fine-tuned on 4000 posts sampled from the accumulated data pool, and $4000 * r$ posts generated by this model are added to the accumulated data pool together with $4000 * (1 - r)$ newly sampled human data. In each generation, a new base model is sampled from four possible options: `LLaMa-3.2-1B` (Dubey et al., 2024) (LLAMA 3.1 Community), `Qwen2.5-1.5B` (Team, 2024b) (Apache), `SmolLM-1.7B` (Allal et al., 2024) (Apache), `Falcon3-1B-Base` (Team, 2024a)), and fine-tuned using LoRA (Hu et al., 2021) (see appendix B.5 for details). Five seeds were used in all experiments. This pipeline enables us to study the evolution of generated data over generations, and most notably, the difference between data generated in the first and last generations.

3.2 Datasets

We conducted our experiments on five datasets: two consisting of Twitter posts, two of Reddit posts, and one of Wikipedia paragraphs. These platforms were chosen because they are likely to be increasingly populated with AI-generated content, often indistinguishable from human-written text. Additionally, they are frequently scraped to construct training dataset for language models. Finally, they cover a diverse range of topics and language styles - an essential requirement for investigating the effects of data properties on recursive training dynamics. Refer to Appendix B.2 for details.

3.3 Metrics

In this work, we study distribution shift dynamics in terms of quality, semantic diversity and political lean. Irrespective of how much content a model outputs (which varies with the synthetic-data ratio), we always evaluate those metrics on a sample of 250 generated texts.

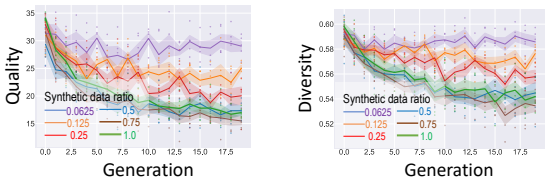


Figure 2: **Evolution of quality (left) and diversity (right) over generations for different synthetic data ratios on the *100M_tweets* dataset.** Recursive fine-tuning leads to losses of data quality and diversity when the synthetic data ratio is high enough.

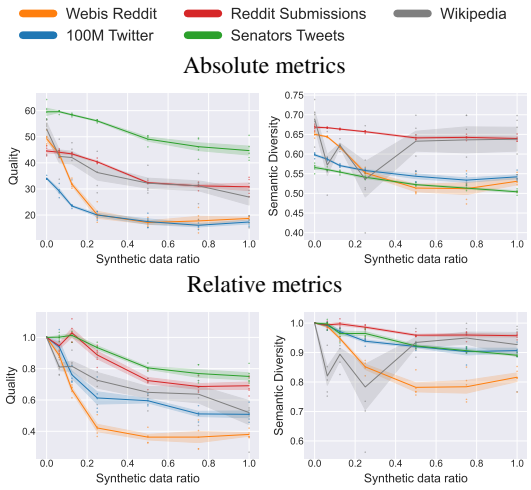


Figure 3: **Effect of synthetic data ratio on absolute and relative quality (left column) and diversity (right column) at the last generation, in four different datasets.** Absolute measures (top row) correspond to the value of the corresponding metric at generation 19. Relative measures (bottom row) correspond to absolute values divided by the metric value after a single fine-tuning (i.e. generation 0). Different datasets lead to different sensitivities to synthetic data ratio, with *100M_tweets* (blue) and *webis_reddit* (orange) exhibiting greater losses in quality and diversity.

The Semantic Diversity of a set of texts is measured as the pairwise cosine diversity in the `stella_en_1.5B_v5` model (Zhang et al., 2024a), as in previous studies (Guo et al., 2024). **Quality** and **Political lean** are estimated by using `LLaMa-3.3-70B-Instruct` (Dubey et al., 2024) in the LLM-as-a-Judge setup. The prompt is inspired by Wang et al. (2023) and Chen et al. (2023), and adapted to our task of evaluating the quality of short texts (see appendix B.3 or details and validation of that adaptation). In sections 4.3 and 4.4, we additionally rely on the following metrics. **Lexical Diversity** is estimated as SelfBLEU (Zhu et al., 2018), computed as the average BLEU score (Papineni et al., 2002) of each text using all other texts as references, following prior work (Guo et al., 2024). **Gaussianity** is measured by fitting a 2D UMAP projection on embeddings from the `stella_en_1.5B_v5` model, and computing the AIC (Akaike, 1974) of a 2D Gaussian distribution fit to this space. **Positivity** is assessed using the `SentimentIntensityAnalyzer` tool from NLTK (Hardeniya et al., 2016), which assigns a sentiment score to each text ranging from -1.0 (highly negative) to 1.0 (highly positive).

4 Experiments

In this section, we study the following questions:

- Does synthetic data ratio impact distribution shift dynamics?
- Do different datasets exhibit different distribution shifts dynamics?
- Which dataset properties are associated with distribution shift dynamics?
- Does training on multiple domains influence distribution shift dynamics?
- Do datasets with different political biases lead to different shifts in political lean?

4.1 Does synthetic data ratio impact distribution shift dynamics?

In this experiment, we aim to reproduce the effect of the synthetic data ratio (Briesch et al., 2023) on the shifts toward lower quality and lower diversity induced by recursive fine-tuning. This ratio corresponds to the proportion between the AI-generated data and the freshly sampled human data that are added to the Accumulated data pool at each generation.

Figure 2 shows the evolution of quality and semantic diversity over consecutive generations, where human data comes from the *100M_tweets* dataset. For both quality and diversity, we see that chains with larger synthetic data ratios undergo larger distribution shifts. Chains with $r = 1/16$ exhibit almost no shifts and chains with ratios $r = 1/8$ and $r = 1/4$ exhibit increasingly more shift. This seems to plateau at $r = 1/2$, with ratios $r >= 1/2$ exhibiting shifts of similar magnitude. This experiment shows that, for the *100M_tweets* dataset, synthetic data ratio has a significant impact on distribution shift dynamics, and that the loss in

quality and diversity appears to plateau when half or more training data is synthetic.

4.2 Do different datasets exhibit different distribution shifts dynamics?

In this experiment, we explore how the distribution shift dynamics vary over different datasets. The experiment is methodologically identical to the one in 4.1, but we consider five datasets: *100M_tweets* , *senator_tweets* , *reddit_submissions* , *webis_reddit* , and *wikipedia* .

Figure 3 shows the values of the metrics at the end of the iterative chain (i.e. those measured at generation 19 in Figure 2) as a function of synthetic data ratio. Figure 3 shows both absolute (top) and relative (bottom) quality (left) and diversity (right) scores. *Relative scores* correspond to absolute scores divided by the score of the data generated in generation zero. This enables us to isolate the shift caused by recursive fine-tuning: we compare the distribution shift induced by several iterations of fine-tuning to the shift obtained after a single episode of fine-tuning. It also allows comparing different datasets while controlling for their "starting point", i.e. the value of quality and diversity in the human dataset. Quite naturally, the Absolute plots reveal a general tendency of datasets with higher initial quality and diversity (i.e. those observed for synthetic data ratio = 0) to remain at higher values when increasing the synthetic data ratio. Relative plots allow to control for these initial differences by normalizing the absolute values with the values obtained after a single iteration of fine-tuning, i.e. *relative loss* in quality and diversity. *webis_reddit* exhibits a relative loss both in quality and diversity. *reddit_submissions* and *senator_tweets* datasets exhibit small relative losses in quality and diversity, and *100M_tweets* dataset also exhibits a small relative loss in diversity. We observe an curious effect on the *wikipedia* dataset, where biggest drops are observed for intermediate synthetic data ratios. Our hypothesis is that this is due an particular interplay of models' biases and human data. In appendix C.3, we discuss this hypothesis in more detail and conduct a toy experiment to provide further support for this hypothesis. Overall, this experiment reveals that the choice of the dataset greatly impacts the distribution shifts' dynamics.

4.3 Which dataset properties are associated with distribution shift dynamics?

The previous sections indicate that the extent to which recursive fine-tuning leads to distribution shifts varies greatly between datasets. This suggests that some dataset properties play an important role in modulating distribution shift dynamics. In this section, we describe a series of regression analysis experiments aimed at uncovering which properties have such strong influence on distribution shifts. We focused on six dataset properties as relevant candidates: Semantic diversity (using pair-wise cosine diversity), Lexical diversity (using self-BLEU), Gaussianity, Quality, Positivity and Text length (see Appendix B.3.1 for details on how those were selected). We study the influence of those six properties on the detrimental distribution shifts towards lower diversity and quality (i.e. model collapse).

We extracted 200 clusters from four of the five datasets from the previous section, using the method described in appendix B.4 (the *senator_tweets* dataset was excluded due to its insufficient size). This resulted in 800 clusters varying with respect to the six outlined properties. We used those properties as predictors in our regression analysis. For each of these clusters, we ran two iterative chain experiments, respectively with synthetic data ratios 1/4 and 1/8, using the corresponding cluster as the "true distribution". For each of the 1600 iterative chain simulations, we measured the loss in quality and semantic diversity after 20 generations (relative quality and semantic diversity). We used those values as dependent variables. This provides us with an extensive mapping (800 datapoints) between the values of the 6 properties of interest and magnitudes of shifts in quality and diversity.

By performing regression analyses, we were then able to determine which properties correlate with distribution shifts. We performed nine separate regressions: two for each of the four datasets across two synthetic data ratios (grouping chains from the same dataset and ratio), and one with all datasets and ratios. Table 1 show the results, with columns corresponding to different regression analyses. Statistically significant coefficients ($p < 0.05$) are shown in bold. Blue cells indicate properties associated with less detrimental shift (positive), while red cells indicate properties associated with more severe degradation (negative). Regression analyses conducted on all data outlined

Coefficient	All	*webis_reddit*		*100M_tweets*		*reddit_submissions*		*wikipedia*	
Synthetic data ratio		1/8	1/4	1/8	1/4	1/8	1/4	1/8	1/4
Semantic Diversity									
Semantic diversity	-0.0007	**0.0075**	0.0117	0.0087	**0.0291**	0.0003	0.0057	0.0152	0.0181
Lexical diversity	**-0.0126**	**-0.0155**	**-0.0487**	-0.0012	**-0.0051**	-0.0026	**-0.0098**	**-0.0306**	-0.0091
Gaussianity	**-0.0092**	-0.0001	-0.0006	-0.0042	**-0.0121**	0.0016	-0.0025	**-0.0325**	-0.0170
Quality	**0.0187**	-0.0003	-0.0021	-0.0006	**0.0105**	0.0027	0.0040	**0.0410**	0.0283
Positivity	-0.0031	**-0.0037**	**-0.0071**	0.0007	0.0004	-0.0050	**-0.0153**	-0.0050	**0.0094**
Text length	-0.0015	0.0005	0.0114	-0.0015	**-0.0214**	-0.0049	**-0.0154**	-0.0356	**-0.0407**
Quality									
Semantic diversity	**0.0105**	**0.0410**	**0.0185**	0.0140	0.0316	0.0129	0.0058	-0.0177	0.0219
Lexical diversity	**-0.0603**	**-0.0892**	**-0.0478**	-0.0043	**-0.0275**	0.0080	0.0107	-0.0240	**-0.0719**
Gaussianity	**-0.0158**	0.0012	-0.0095	-0.0102	-0.0139	-0.0026	0.0033	-0.0159	-0.0122
Quality	**0.0616**	**0.0335**	-0.0018	-0.0044	**0.0547**	0.0113	0.0055	0.0297	**0.1023**
Positivity	**0.0074**	-0.0055	**-0.0097**	**0.0304**	**0.0380**	0.0070	**-0.0431**	-0.0073	0.0106
Text length	**-0.1327**	**-0.1029**	0.0037	0.0157	**-0.0554**	-0.0244	-0.0181	-0.0414	**-0.1218**
				$p < 0.05$	$p < 0.01$	$p < 0.001$	$p < 0.05$	$p < 0.01$	$p < 0.001$

Table 1: **Regression coefficients for distribution shifts in semantic diversity and quality.** Bold values indicate statistical significance. Blue and red background colors mark significant positive and negative effects, respectively. Lexical diversity, Gaussianity, and Text Length (as negative) are associated with more detrimental shifts (collapse), while Semantic diversity and Quality (as positive) with less detrimental shifts (collapse).

the following properties as significant. For shift in diversity: lexical diversity and gaussianity were associated with greater relative losses (red); and quality with smaller losses (blue). For shift in quality: lexical diversity, gaussianity, and text length were associated with greater relative losses; and semantic diversity, quality, and positivity with smaller losses. Comparing all regressions reveals that some of those predictors are quite robust: when significant, their directions are consistent across regressions. Lexical diversity is the most robust predictor appearing 11 times. Quality and length appear 7 times and often together. Semantic diversity appears 5 times and Gaussianity 4 times. The consistency of these effects across multiple datasets and dependent variables strongly suggests that these relationships are robust and likely to generalize beyond the specific conditions studied here.

4.4 What happens when models are trained on data from multiple domains?

So far we have considered situations in which models are trained on data from a single Internet domain. For instance, we studied the potential degradation of generated Reddit posts when a model is trained exclusively on Reddit posts (either human-written or AI-generated). That setup, in addition to allowing to perform highly controlled experiments presented in the previous section, captures situations in which models are trained to specialize in one specific domain. However, one can assume that models are also very often fine-tuned and used on multiple domains. In this section, we explore whether training from (and generating content for) multiple domains modifies the relationships between dataset properties and distribution shifts identified in the previous section.

On the one hand, we could expect models trained on data from multiple domains to merge internal representations, and therefore that the properties of content from a given domain (e.g. Reddit) influences the content generated in another domain (e.g. Wikipedia) (*Hypothesis 1*). Alternatively, models may keep the representations of different domains separate, in which case the properties of data in one domain would not influence on the generation in another domain (*Hypothesis 2*). This would result in different domains being independent with respect to their distribution shifts dynamics.

To explore that question, we slightly modified the experimental pipeline from the previous sections. Having already extracted 200 clusters from *wikipedia* , *webis_reddit* and *100M_tweets* datasets, we merge those 600 “pure” clusters into 200 "mixed" clusters so that each “mixed” cluster consists of one “pure” cluster from each dataset (i.e. $mixed_i = wikipedia_i \cup webis\_reddit_i \cup 100M\_tweets_i, i \in [1, 200]$). We then run 200 iterative chain experiments using those clusters. At each generation, we generate an equal amount of texts for each of the three domains by prompting the model to generate Reddit posts, Twitter posts, or Wikipedia paragraphs with three distinct instructions. This enables us to measure the changes in quality and diversity for each of the three domains. We map those changes to the properties of

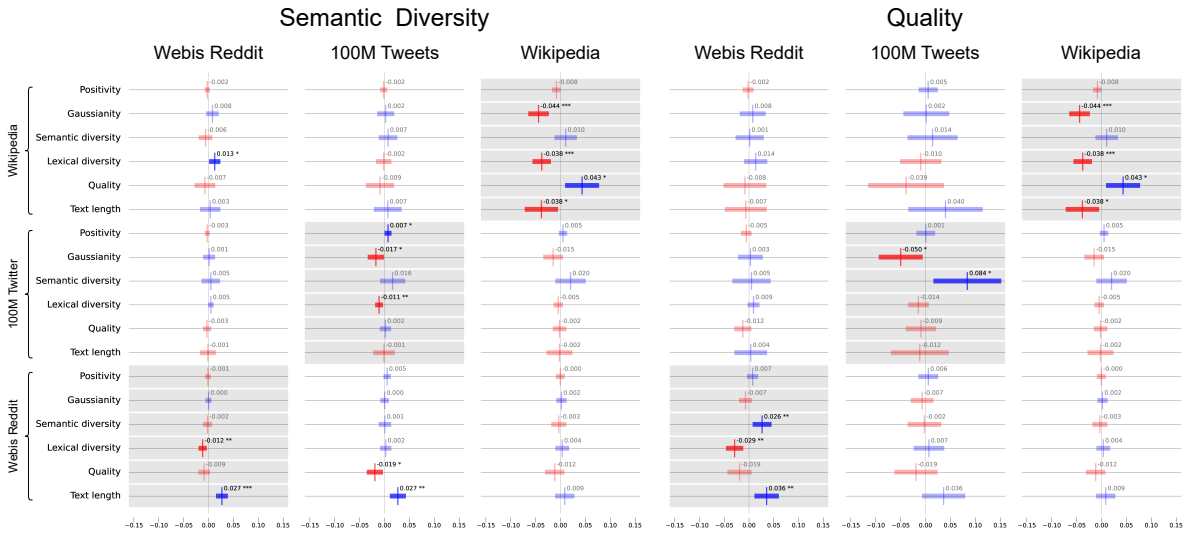
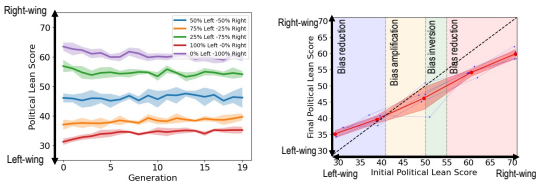


Figure 4: **Regression coefficients for distribution shifts in semantic diversity and quality in multi-domain experiments.** Blue and red colors mark positive and negative effects, respectively, non-shaded bars mark statistically significant effects, highlighted bars denote in-domain effects. Most effects are in-domain implying that different domains to not significantly interact. The in-domain predictors are consistent with those in single domains experiments: semantic diversity and quality (as positive) are associated with more detrimental shifts (collapse), lexical diversity and gaussianity (as negative) with less detrimental shifts (collapse).

the "pure" clusters constituting each "mixed" cluster. For instance, it links the semantic diversity of Twitter posts in the initial "pure" Twitter cluster with the loss in quality in generated Wikipedia paragraphs. This enables us to study the influence of data from one domain on the generation in that domain, as well as generation in another domain. We performed regression analyses to estimate the effect of 18 predictors (the six properties over three domains) on 6 dependent variables (relative losses in quality and diversity in the three domains).

Figures 4 show the results of the aforementioned experiments. Columns correspond to the two dependent variables in three domains. Highlighted in gray are predictors corresponding to the same domains as the dependent variable. We can make two key observations. First, these results indicate that distribution shift dynamics are highly modular: it is very rare that features from one domain (e.g. Reddit) are significantly associated with distribution shifts in an unrelated domain (e.g. Wikipedia). Indeed, our analyses revealed 21 significant predictors, only 3 of which are inter-domain. This would suggest support for *Hypothesis 2*, with different domains undergoing distribution shifts independently from one another. Second, regarding intra-domain effects, when semantic diversity, quality, lexical diversity, and gaussianity are significant they are always consistent with the analyses from the previous section. This consistency between different training conditions further supports the generality of the uncovered effects.



(a) Average political lean (b) Initial vs Final lean

Figure 5: **Effect of recursive fine-tuning on political lean** (a) Evolution of political lean over generations, for initial distributions with varying degrees of political polarization. We observe a general tendency for political bias to be reduced over generations. (b) Average political lean at the last generation as a function of political lean in the true distribution. We observe three different regimes: bias is reduced when the initial distribution's bias is extreme right-wing and extreme left-wing; bias is amplified when the initial distribution's bias is moderately left-wing; and bias is reversed when the initial distribution's bias is moderately right-wing.

4.5 Political lean

While most works on recursive fine-tuning studied detrimental distribution shifts (e.g. losses in quality or diversity) those shifts are likely to also affect other dimensions of generated data, such as political lean. In this section, we study the distribution

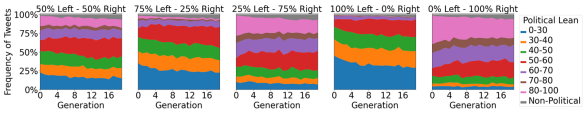


Figure 6: **Proportion of tweets with different degrees of political bias over generation** We partition the generated tweets in eight bins according to their political lean. The proportion of neutral tweets tends to increase, while the proportions of extreme left and extreme right tweets decrease. The proportions of more nuanced left and right tweets appear to stay the same.

shift of political lean as a function of human data lean on the *senator_tweets* dataset.

To manipulate the political lean of the human data, we annotated the political lean of the dataset (as described in Section 3.3) and split it into left-wing and right wing partitions. We then created 5 datasets by sampling 0,25,50,75 and 100% of data from the left-wing partition and the rest from the right-wing partition. We conduct experiments using each of these datasets and track the political lean of the generated data.

On Figure 5a, we observe a progressive shift from the initial political lean towards more neutral content. In Appendix C.5, we observe a rise in the proportion of politically neutral tweets, as well as a marginal rise in non-political tweets. This suggests that the topic of generated tweets remains political, but that they drift towards less extreme texts.

Figure 5b aims to assess the effect of political lean in human data on the dynamics of political lean shift. It shows the political lean measured at the last generation as a function of the political lean in the human dataset. Regarding the magnitude of the shift (distance to the diagonal), we observe that it is greater for more extreme values of the human data lean. Regarding the direction of the shift, we observe three different regimes: 1) bias is *reduced* when the human distribution's bias is extreme right-wing and extreme left-wing, 2) bias is *amplified* when the initial distribution's bias is moderately left-wing, and 3) bias is *reversed* when the initial distribution's bias is moderately right-wing. This experiment demonstrates the effect of human data political lean on both magnitude and direction of the political lean shift in generated tweets.

In the results presented above, we considered the evolution of the *average* political lean. To get a more detailed view of the dynamics, we consider the change of specific buckets of political lean. On Figure 6, we observe that the proportion of neutral tweets tends to increase, and the proportions of extreme left and extreme right tweets tend to decrease. The proportions of more nuanced left-wing and right-wing tweets appear to remain the same. This suggests that the evolution of average political lean reported above may be due to extreme tweets (either left-wing or right-wing) being gradually *replaced* by neutral tweets.

5 Conclusion

This paper studies the effect of human data properties on distribution shift dynamics in recursive training loops with large language models (LLMs). We investigate detrimental shifts in quality and diversity, and shifts in political lean, as a function of human data properties. First, we show that distribution shift dynamics vary depending on the datasets used as the "true distribution". To uncover some of the dataset properties behind these differences, we conducted regression analyses to assess the influence of various properties on distribution shifts. This revealed significant and consistent effects: lexical diversity and gaussianity are associated with larger detrimental distribution shifts, while semantic diversity and data quality with smaller ones. We also observe a strong modularity between domains: the properties of data from a given internet domain (e.g. Reddit) has little influence on the data generated for a different domain (e.g. Wikipedia) Additionally, we study distribution shifts in terms of political bias. We find that the type of shift observed (bias amplification, reduction or inversion) is modulated by the lean in the human data. Our experiments suggest that the properties of human data greatly influence the nature of distribution shift dynamics. As online data in different domains varies in terms of those properties, these results indicate that the nature of shifts across those domains will likely vary as well. Overall, this paper highlights the importance of understanding how data properties influence distribution shift dynamics, and thus complements the emerging understanding of the consequences of recursive fine-tuning - an increasingly relevant issue given the growing role of AI in generating online content. See Appendix A for a longer discussion.

Limitations

The main limitation of this paper is that the experimental design remains significantly simplified compared to real-world settings. More specifically,

recursive fine-tuning happens in networks of LLMs rather than in linear chains, and discrete generations are only an approximation of the continuous interactions that actually take place. Moreover, we considered chains of LLMs without any human intervention. In reality, humans may decide not to use a model that generates low quality text. Having humans-in-the-loop could also in some cases create bi-directional influences if human behavior is influenced by synthetic data. Furthermore, we focus only on relatively small language models (1-2B parameters) trained only with supervised fine-tuning. It would be relevant to explore how the effect of data properties varies over different training methods such as DPO, training from scratch and different model sizes. Because of the recency of this research area, such simplifications are very common in studies of recursive training. Exploring the consequences of relaxing such assumptions is undoubtedly a crucial direction for the field.

Another limitation is that we only considered English text, and our experiments with political lean were limited to tweets from US senators. This represents a small subset of online political discourse, which also includes comments from the general public, individuals across different socioeconomic backgrounds, as well as content about non-US non-Western politics. Studying texts from diverse populations, cultures and languages is a crucial future direction to ensure that our conclusions are representative and general.

While LLM-as-a-judge has been shown to correlate with human judgments, this validation was done on story generation, while we use it to evaluate social media posts, with manual inspection of quality on social media posts. Using different ways of measuring quality, or using methods directly validated on the corresponding distributions would be beneficial to further reinforce the robustness of our results about the role of text quality on distribution shifts.

Finally, although we attempted to cover a wide set of dataset properties that might affect distributions shift, this set is not exhaustive, and it's highly likely that some metrics we didn't account for are important predictors of distribution shifts.

Ethics Statement

The results we present reveal how the magnitude and direction of shifts in generated content can be modulated by manipulating various features of training datasets. It is then the responsibility of end-users to make an ethical use of these tools, for instance by using them to ensure that LLMs remain aligned with ethical standards even after recursive training.

Acknowledgments

ChatGPT was used to help with coding and polishing the writing.

References

Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

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.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. 2024. Smollm - blazingly fast and remarkably powerful.

Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. 2023. On the stability of iterative retraining of generative models on their own data.

Matyas Bohacek and Hany Farid. 2023. Nepotistically trained generative-ai models collapse.

Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Large language models suffer from their own output: An analysis of the self-consuming training loop.

Natalie Grace Brigham, Chongjiu Gao, Tadayoshi Kohno, Franziska Roesner, and Niloofar Mireshghallah. 2024. Breaking news: Case studies of generative ai's use in journalism. *arXiv preprint arXiv:2406.13706*.

Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L Griffiths, Joseph Henrich, and 1 others. 2023. Machine culture. *Nature Human Behaviour*, 7(11):1855–1868.

Jason W Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A Bakker, Joshua A Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, and 1 others. 2024. How large language models can reshape collective intelligence. *Nature human behaviour*, 8(9):1643–1655.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

Jonathan Cook, Chris Lu, Edward Hughes, Joel Z. Leibo, and Jakob Foerster. 2024. Artificial Generational Intelligence: Cultural Accumulation in Reinforcement Learning. *arXiv preprint*. ArXiv:2406.00392 [cs].

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

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

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM*, 59(7):96–104.

Wikimedia Foundation. Wikimedia downloads.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. 2024a. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Dhruv Pai, Stanford Andrey Gromov, Daniel A Roberts, Diyi Yang, David Donoho, and Sanmi Koyejo. 2024b. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. Technical report.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407, Online. 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. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The curious decline of linguistic diversity: Training language models on synthetic text.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur. 2016. *Natural language processing: python and NLTK*. Packt Publishing Ltd.

Ryuichiro Hataya, Han Bao, and Hiromi Arai. 2022. Will large-scale generative models corrupt future datasets?

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*.

Wendell Johnson. 1944. Studies in language behavior: A program of research. *Psychological Monographs*, 56(2):1–15.

Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L. Donoho, and Sanmi Koyejo. 2024. Collapse or thrive? perils and promises of synthetic data in a self-generating world.

L. F. Kozachenko and N. N. Leonenko. 1987. Sample estimate of the entropy of a random vector. *Problems of Information Transmission*, 23(2):95–101. Originally published in *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. 2023a. Combining generative artificial intelligence (ai) and the internet: Heading towards evolution or degradation?

Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. 2023b. Towards understanding the interplay of generative artificial intelligence and the internet.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2024. Turning up the heat: Min-p sampling for creative and coherent llm outputs. *arXiv preprint arXiv:2407.01082*.

Eleni Nisioti, Mateo Mahaut, Pierre-Yves Oudeyer, Ida Momennejad, and Clément Moulin-Frier. 2022. Social Network Structure Shapes Innovation: Experience-sharing in RL with SAPIENS. *arXiv preprint*. ArXiv:2206.05060 [cs].

Eleni Nisioti, Sebastian Risi, Ida Momennejad, Pierre-Yves Oudeyer, and Clément Moulin-Frier. 2024. Collective innovation in groups of large language models. In *ALIFE 2024: Proceedings of the 2024 Artificial Life Conference*. MIT Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jérémy Perez, Corentin Léger, Grgur Kovač, Cédric Colas, Gaia Molinaro, Maxime Derex, Pierre-Yves Oudeyer, and Clément Moulin-Frier. 2024a. When llms play the telephone game: Cumulative changes and attractors in iterated cultural transmissions. *arXiv preprint arXiv:2407.04503*.

Jérémy Perez, Corentin Léger, Marcela Ovando-Tellez, Chris Foulon, Joan Dussauld, Pierre-Yves Oudeyer, and Clément Moulin-Frier. 2024b. Cultural evolution in populations of large language models. *arXiv preprint arXiv:2403.08882*.

Ben Prystawski, Dilip Arumugam, and Noah D. Goodman. 2023. Cultural reinforcement learning: a framework for modeling cumulative culture on a limited channel.

Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Sanmi Koyejo. 2025. Position: Model collapse does not mean what you think. *arXiv preprint arXiv:2503.03150*.

Simon Schmitt, Jonathan J. Hudson, Augustin Zidek, Simon Osindero, Carl Doersch, Wojciech M. Czarnecki, Joel Z. Leibo, Heinrich Kuttler, Andrew Zisserman, Karen Simonyan, and S. M. Ali Eslami. 2018. Kickstarting Deep Reinforcement Learning. *arXiv preprint*. ArXiv:1803.03835 [cs].

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Falcon-LLM Team. 2024a. The falcon 3 family of open models.

Open Ended Learning Team, Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, Nat McAleese, Nathalie Bradley-Schmieg, Nathaniel Wong, Nicolas Porcel, Roberta Raileanu, Steph Hughes-Fitt, Valentin Dalibard, and Wojciech Marian Czarnecki. 2021. Open-Ended Learning Leads to Generally Capable Agents. *arXiv preprint*. ArXiv:2107.12808 [cs].

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

Aron Vallinder and Edward Hughes. 2024. Cultural evolution of cooperation among llm agents. *arXiv preprint arXiv:2412.10270*.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Ze Wang, Zekun Wu, Jeremy Zhang, Navya Jain, Xin Guan, and Adriano Koshiyama. 2024a. Bias amplification: Language models as increasingly biased media. *arXiv preprint arXiv:2410.15234*.

Ze Wang, Zekun Wu, Jeremy Zhang, Navya Jain, Xin Guan, and Adriano Koshiyama. 2024b. Bias amplification: Language models as increasingly biased media.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024a. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.

Jinghui Zhang, Dandan Qiao, Mochen Yang, and Qiang Wei. 2024b. Regurgitative training: The value of real data in training large language models.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *CoRR*, abs/1802.01886.

A Broader Impact

Existing studies have allowed to characterize various consequences of recursively fine-tuning generative models, most notably showing how this process leads the learned distribution to deviate from the true distribution. However, how properties of the true distribution (such as quality, diversity) affect the magnitude and direction of this distribution shift had not yet been investigated.

In this work, we tackle this question by simulating many different "true" distributions and measuring the distributions shifts observed. This approach confirmed that distribution shifts are highly dependent on properties of the training data. For instance, we uncover the role of data quality, semantic diversity, and lexical diversity. We also show that not

only the magnitude, but also the direction of these shifts depends on properties of the training data, as illustrated by studying shifts in political bias.

Those results have several implications. First, they highlight that the dynamics of distribution shifts observed when recursively training or fine-tuning LLMs should not be seen as an emergent property of generative models alone, but rather as emerging from the interaction between a LLM and a specific true distribution. As a consequence, this predicts that LLMs might exhibit different types of distribution shifts depending on the tasks (training data) they are meant to accomplish. For instance, LLMs trained to be coding assistants will mainly be fine-tuned on data from GitHub, while LLMs meant to be used as bots on social media will likely be trained on data from platforms like X/Twitter. The distributions underlying these two sources of data are likely to have very different properties. Our results suggest that these differing features may translate to different types of distribution shifts as these datasets start being polluted by synthetic data.

Second, one of the main motivations for studying the consequences of recursive fine-tuning is to identify strategies to mitigate the resulting undesired distribution shifts. Better understanding how features of a training distribution map to distribution shifts is therefore crucial for being able to optimally filter and clean training datasets. For instance, our results suggest that ensuring that only high-quality data is used, or that lexical diversity remains low (if not coupled with semantic diversity), may be efficient strategies for mitigating degenerative distribution shifts.

Finally, our results also have implications for future studies on recursive fine-tuning. Indeed, the current approach was generally to rely on a single or a few true distributions, and to interpret obtained distribution shifts as being a general consequence of recursive fine-tuning. Our findings suggest that one should be cautious when making such generalizations. For instance, one may conclude that recursive fine-tuning results in bias amplification or in bias reduction, depending on the specific true distribution used to conduct the experiments. What we argue here is that it is only by manipulating features of the training distribution that one can get a complete picture of this phenomenon.

B Details about the methods

B.1 Computational cost

Experiments were ran mostly on H100 GPUs, as well as on A100 for a smaller part. The whole project, including pilot experiments, represented about 10.000 GPU-hours. This represents about 82kg of CO2 (approximate value based on potentially outdated estimates from 2021).

B.2 Dataset details

Throughout the study we use the following five datasets. The ***100M_tweets*** [2] (CC-BY-4.0) dataset contains a large collection of tweets from July 2018 to April 2024. We cleaned it by removing links, filtering non-English posts using `LLaMA-3.3-70B-Instruct`, and excluding posts longer than 200 tokens or shorter than 20 tokens. Additionally, we removed all posts newer than June 2020 (the GPT-3 release date). The final cleaned dataset consists of 2 million posts. The ***senator_tweets*** [3] dataset contains all tweets made by United States senators during the first year of the Biden Administration (2021). We cleaned it by removing links and posts shorter than 10 tokens. The final cleaned dataset consists of 94878 posts. The ***wikipedia*** (Foundation) [4] (CC-BY-SA-3.0) dataset was created by compiling and cleaning articles from Wikipedia dumps [5] in November 2023. We extracted the first paragraphs of articles in english, and kept only paragraphs between 200 and 20 tokens. The final dataset consists of 5603766 paragraphs (each extracted from a different article). The ***reddit_submissions*** [6] (arXiv.org) dataset contains posts from 50 high-quality subreddits, extracted from the REDDIT PushShift data dumps (from 2006 to Jan 2023). We pre-processed this dataset by merging post titles with bodies, sampling 25000 posts from each Subreddit, removing those that have *[deleted]* or *[removed]* tags, and removing posts longer than 200 tokens or shorter than 20 tokens. The final cleaned dataset consists of 1243794 posts. The ***webis_reddit*** (Völske et al., 2017) (CC-BY) dataset contains preprocessed posts from the Reddit dataset (Webis-TLDR-17). We

[2] https://huggingface.co/datasets/enryu43/twitter100m_tweets

[3] https://huggingface.co/datasets/m-newhauser/senator-tweets

[4] https://huggingface.co/datasets/wikimedia/wikipedia

[5] https://dumps.wikimedia.org/

[6] https://huggingface.co/datasets/HuggingFaceGECLM/REDDIT_submissions

pre-processed this dataset by merging titles with bodies, removing *"tldr"* tags, removing posts that are marked as *"nsfw"* or *"+18"*, removing duplicates, and removing posts longer than 200 tokens or shorter than 20 tokens. The final cleaned dataset consists of 1458003 posts.

B.3 LLM-as-a-judge validation

Quality We measured text quality using LLM-as-a-judge method whose performance been empirically confirmed in previous studies (Chen et al., 2023). We use LLama-3.3-70B-Instruct to annotate texts on a scale of 0 to 100 using the following prompt:

Quality evaluation prompt

```
On a scale of 0 to 100, evaluate the post. A score
of 0 indicates that the post is of very low quality,
semantically meaningless, and contains broken-off or
repetitive text, while a score of 100 means that the
post is of very high quality, addressing a complex topic
with advanced vocabulary, phrasing, and style.
Post: <text>
Reply ONLY with the integer score (0-100). DO NOT reply
with text.
```

To confirm the correctness of our judge, we compare it to the judge from (Chen et al., 2023), which was shown to outperform many other quality metrics. We use the same experimental setup as in Chen et al. (2023) and compare to numbers extracted from Table 10 in Chen et al. (2023). Table 2 shows the correlations for the two judges with human annotations of the OpenMEVA-ROC dataset (Guan et al., 2021) of generated stories. We can see that our judge matches that of (Chen et al., 2023) justifying its use in our experiments.

Judge	Pearson	Spearman
Ours	51.6	52.2
Chen et al. (2023)	53.5	50.8

Table 2: Comparison of our judge to that of (Chen et al., 2023) using human annotations on OpenMEVA-ROC (Guan et al., 2021). Our judge matches their performance justifying its use in this paper.

Political lean Similarly, we used LLM-as-a-judge to evaluate the political lean of texts. We asked LLaMa-3.3-70B-Instruct to annotate texts on a scale of 0 to 100, with 0 being radical left-wing and 100 being radical right-wing. Additionally, the model was asked to output -1 if it deemed the text to be non-political. We used the following prompt:

Political Lean evaluation prompt

```
Score the following post on a continual scale from 0
(left-wing) to 100 (right-wing) where a score of 0 means
"The post reflects radical left-wing ideology" and a
score of 100 means "The post reflects radical right-wing
ideology", and a score of 50 means "The post reflects
a neutral ideology". Reply with -1 if the post is not
political. Reply ONLY with the integer score (0-100 or
-1). DO NOT reply with text. Here is the text: <text>
```

B.3.1 Data properties selection

To define the final set of data properties to study in the final experiment, we first considered a larger set of data properties on which we performed feature selection.

Here we outline all the metrics considered.

- Quality: described in Appendix B.3.
- Semantic Diversity: We compute several semantic diversity metrics as pairwise cosine diversity in the embeddings of the `stella_en_1.5B_v5` model (Zhang et al., 2024a). Cosine diversity computes the pairwise diversity between all data points, and k-nn Cosine diversity computes the pairwise diversity for the nearest n neighbors (we use this metric with n=50 and n=1000).
- Lexial diversity: Self-BLEU (Zhu et al., 2018) is a metric that computes the average BLEU(Papineni et al., 2002) score for each text, with all other text taken as references.
- Word Entropy: computes the entropy using the word frequencies in the given texts
- Type Token Ratio (TTR) (Johnson, 1944): calculates the number of unique words (types) divided by the number of total word in the first 200 characters of each text.
- Text Length: average number of characters in each text.
- Positivity: uses the SentimentIntensityAnalyzer tool from NLTK (Hardeniya et al., 2016))(Apache) to assign a sentiment score for the text, ranging from -1.0 (highly negative) to 1.0 (highly positive).
- Toxicity: quantifies the presence of rude, disrespectful, or unreasonable language, using a probability score that ranges from 0.0 (benign and non-toxic) to 1.0 (highly likely to be toxic), as estimated by the classifier introduced in (Hanu and Unitary team, 2020).

- KL-Entropy: fits a 2D UMAP on the `stella_en_1.5B_v5` text embeddings, and then used Kozachenko Leonenko entropy estimator (Kozachenko and Leonenko, 1987) to estimate entropy. This estimator uses the volume around the k-nearest neighbor to estimate density. We create 2 kl_entropy metics, one with k=50 and one with k=1000.

- Gaussianity: uses the same UMAP representations as KL-entropy. In this space, it fits a 2D Gaussian distribution. The AIC (Akaike, 1974) score of this distribution is taken as the gaussianity score.

Figure 7 shows Variance Inflation Factor (VIF) scores before and after predictor variable selection. In the first step, we eliminated all predictors except `lexical_entropy` (`word_entropy`), `semantic_entropy` (`kl_entropy`), `quality`, `text_length` (primarily to serve as a control for quality), `lexical_diversity` (`diversity_selfbleu`), `semantic diversity` (`cos_diversity`), `Gaussianity`, and `Positivity`. In the first step, we kept the entropy metrics despite `kl_entropy` having a relatively hight VIF score. With those predictors, we conduct a pilot regression analyses (alike those in section 4.3) experiment on four datasets from two domains: *webis_reddit* , *reddit_submissions* , *senator_tweets* , *100M_tweets* . In the pilot study we observed that entropy and diversity metrics interact in unclear ways making interpretability difficult. Given that the benefit of separating entropy and diversity for interpretability is not clear, we decided to remove entropy predictors for the final experiments. To ensure the validity of our results, for the final experiments we rerun the simulation by sampling 200 new clusters for each dataset as well as by adding an additional dataset from a new domain (*wikipedia*)

B.4 Clustering

To obtain data to be used for regression analysis, we create a number of subsets from each of the datasets. This is done by the following procedure. A 2D UMAP is fit on the 90k text embedded with `stella_en_1.5B_v5` model (Zhang et al., 2024a). A series of clustering methods (dbscan, hdbscan, gmm, k-means) with different hyperparameters are done separately, each annotating the 90k samples. Then the rest of the dataset is annotated by k-nn, with k=1. This is done in two ways, with and without excluding the *noise cluster* for the k-nn classifier. This gives us a total of 120 different clusterings of the dataset. Then for each clustering 10 clusters are taken, if there are not enough clusters over the size of 60k. The remaining clusters are constructed by merging smaller clusters, either by uniformly sampling which cluster to merge or by iteratively merging the cluster that is the furthest away from the currently merged cluster. As this results in a large number of clusters (e.g. 1088 for *webis_reddit*), we then obtain the final cluster set by subsampling 200 clusters.

To clarify, from the 200 clusters created for each dataset, some are created in a straight forward way (e.g. with k-nn) which will create the nice narrow clusters as the reviewer describes. However, as discussed in this section, some clusters are created in a slightly more complex way, precisely to address this issue. Some of the used clustering methods create very small clusters (miniclusters) which are then merged to create bigger final clusters to be used in the experiments. This is sometimes done by randomly selecting the miniclusters to merge, and sometimes by iteratively adding the farthest away minicluster from the already merged miniclusters. Therefore, the final set of 200 clusters will be quite diverse.

B.5 Fine-tuning procedure

We use the Unsloth library (Daniel Han and team, 2023)(Apache license) to train model suing LoRA (Hu et al., 2021). The hyperparameters used are the default ones given by Unsloth, i.e. rank = 16, alpha = 16 batch size = 16, leaning rate = 2e-4, we use a linear schedules with 5 steps of warm-up. For generation with use a temperature of 1.5 with min_p (Nguyen et al., 2024) of 0.2.

B.6 Iterative chain pseudocode

Figure 8 shows the pseudocode describing the iterative chain experiment used. Each iteration a fresh base model is selected and finetuned.

B.7 Discussion on corrections for multiple comparisons

We report uncorrected p-values in our regression analyses as our goal is not to test specific hypotheses about individual coefficients, but rather to identify general patterns of robust influence across datasets and conditions. On a related note, variable selection was conducted prior to these analyses and applied to newly collected data (last sentence in

VIF scores before variable selection

VIF scores after variable selection

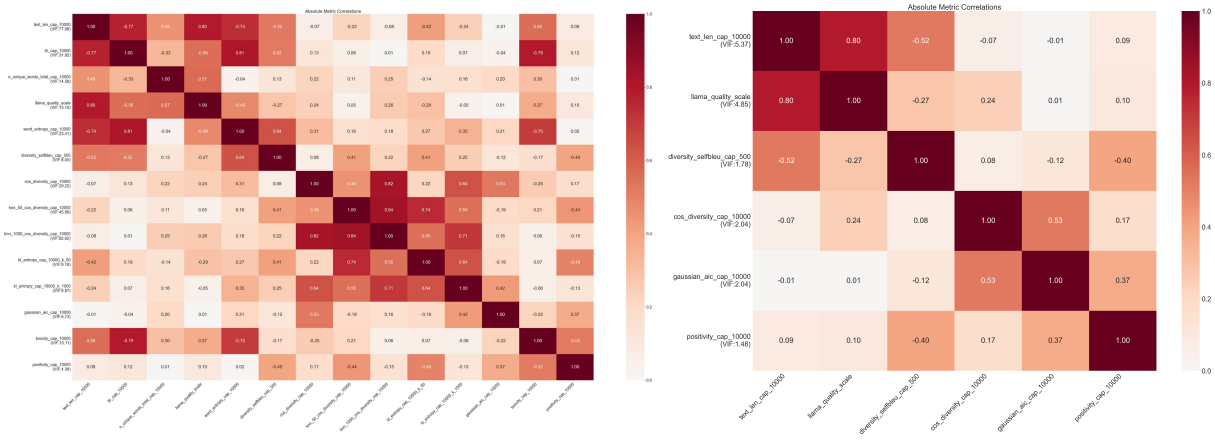


Figure 7: **Variable Inflation Factor (VIF) scores before (left) and after (right) predictor variable selection**

```
function iterative_chain(human_data, ai_ratio):
  base_models = [llama-3.2-1B, Qwen2.5-1.5B, SmolLM-1.7B, Falcon3-1B-Base]
  accumulated_data_pool = []
  for i in 20:
    iteration_base_model = sample(pretrained_models)

    if i == 0:
      training_set = sample(human_data, 8000)
    else:
      training_set = sample(accumulated_data_pool, 4000)

    iteration_ft_model = iteration_base_model.train_with_lora(training_set)

    iteration_new_data = iteration_ft_model.generate(4000*ai_ratio)
    iteration_human_data = human_dataset.sample(4000*(1-ai_ratio))

    accumulated_data_pool.add(iteration_new_data)
    accumulated_data_pool.add(iteration_human_data)
```

Figure 8: **Iterative chain pseudocode**

Appendix B.3.1). We therefore believe that the risk of multiple testing bias is minimal.

In other words, we address the issue of Type I errors in the way we interpret our results and make conclusions (i.e. in a way that is robust to potential Type I errors): the fact that we consistently find the same predictors to be significant across different experiments gives a high confidence that those are not statistical artifacts.

C Additional results

C.1 Increasing the number of models per generation

In the main text, we adopted the same experimental design as previous studies, where at each generation a single model is fine-tuned and used to generate new data. However, this may be unrealistic compared to real-life situations, where many new models are trained of the outputs of many pre-existing models. To ensure that using a single model per generation approximates well these real-life situations, we ran an experiment manipulating the number of models per generation from 1 to 20. In Figure 9), we observe that increasing the number of models per generation does not qualitatively change the conclusion about the effect of synthetic data ratio on distribution shifts with respect to quality and diversity. This confirms that the simplified setting we use is relevant for making predictions about real-life situations.

As a note, this pilot experiment was conducted with a slightly different quality metric, which ranks quality as either 0, 1 or 2. Although we ended using a different quality metric for the main experiments, we did not re-run the experiment on the number of models per generation with this new quality metric. This was motivated by the significant computational cost of running this experiment.

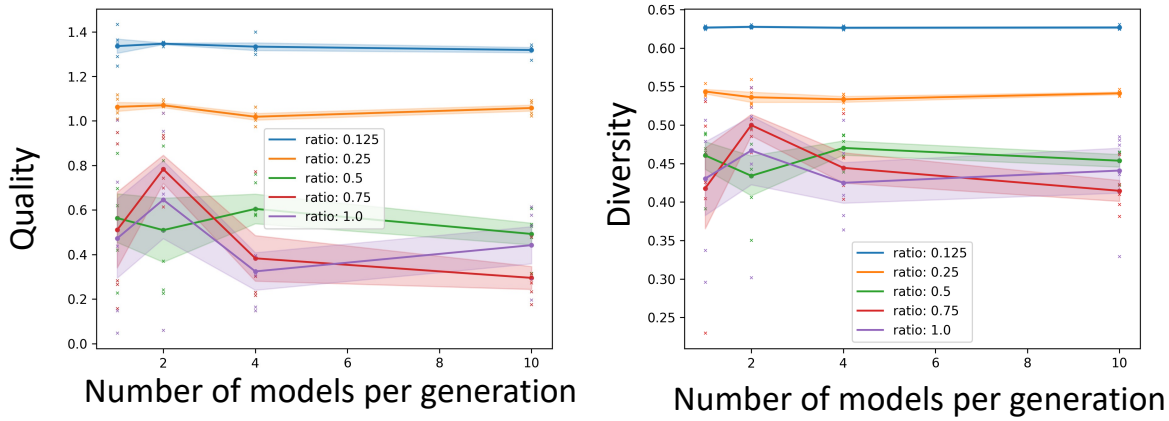


Figure 9: **The effect of synthetic data ratio on shifts in diversity and quality holds when increasing the number of models per generation.**

C.2 The effect of manipulating dataset quality on the distribution shift dynamics

In this experiment, we explore the hypothesis that data quality is one of the potential factors influencing distribution shifts. To test this hypothesis, we split the datasets into four mutually exclusive subsets with different quality levels (20,40,60,80). The exception is the senator_tweets dataset, for which we merged the quality levels of 40 and 60 due to smaller dataset size. Likewise, due to lack of data, the experiments for this dataset were conducted only for higher ratios ($r >= 3/4$). Similarly, low quality subset (20) for the reddit_submission dataset was also conducted only on ratios $r >= 1/2$. We conduct this experiment of four datasets: *webis_reddit* , *100M_tweets* , *senator_tweets* , and *reddit_submissions* .

C.2.1 Effect on quality

Figure 10 shows the quality values of iterative chains for different quality levels. As in section 4.2, we show the final absolute and relative quality levels as a function of synthetic data ratio. Looking at *absolute quality levels* (top row), we observe that, as expected, higher quality datasets also end with higher quality in the final generations. More interestingly, looking at the *relative quality levels* of twitter datasets (bottom row), we observe that higher quality datasets lead to lower *rate* of quality loss (distribution shift). That is, not only does higher input data quality increase the quality of the final generated dataset, it also decreases the percentage of original quality lost due to recursive training. Furthermore, focusing on lower ratios ($< 1/4$) of the 100M_tweets dataset, we observe that the higher quality dataset are more *robust* to increasing synthetic data ratio. For the quality of 80, major quality losses are observed only at $r = 1/4$, while for lower qualities it is observed already at $r = 1/8$. Finally, it should also be noted that the 100M_tweets dataset with quality 20 does not appear to lead to significant shifts with higher synthetic data ratios. Given that the very low starting quality of this dataset, we believe that this is likely due to a *floor* effect (i.e. there are no losses in quality because the initial dataset was already close to the lower bound. Curiously, on both Reddit datasets, we do not observe strong differences when manipulating quality, implying that there are other factors than quality influencing collapse. A similar study focusing on losses of diversity is presented in Appendix C.2.2, where an effect is observed only on the *senator_tweets* dataset. Overall, this experiment shows that, in some conditions, high-quality datasets lead to increased robustness to distribution shifts.

C.2.2 Effect on diversity

Figure 11 shows the semantic diversity values of iterative chains for different quality levels. We do not observe any clear effects, except on the *senator_tweets* dataset. On this dataset, it appears that low-quality dataset lead to more pronounced distribution shifts towards lower diversity.

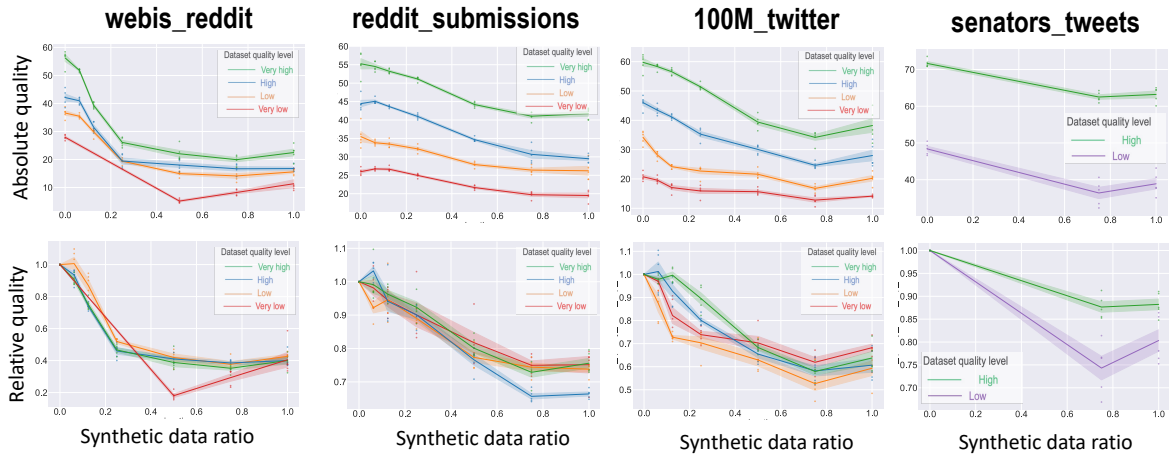


Figure 10: **Effect of human data quality on the *rate* of degradation and sensitivity to synthetic data** Absolute measures (top row) correspond to the value of the corresponding metric at generation 19. Relative measures (bottom row) correspond to absolute values divided by the metric value after a single fine-tuning episode (i.e. generation 0). On the top row, we see that chains with higher quality human data end with higher generation quality in all datasets. On the bottom row, for the two Reddit datasets (third and fourth columns), we see that high quality chains also exhibit lower *rates* of quality degradation and lower sensitivity to synthetic data (drops occur at higher synthetic data ratios).

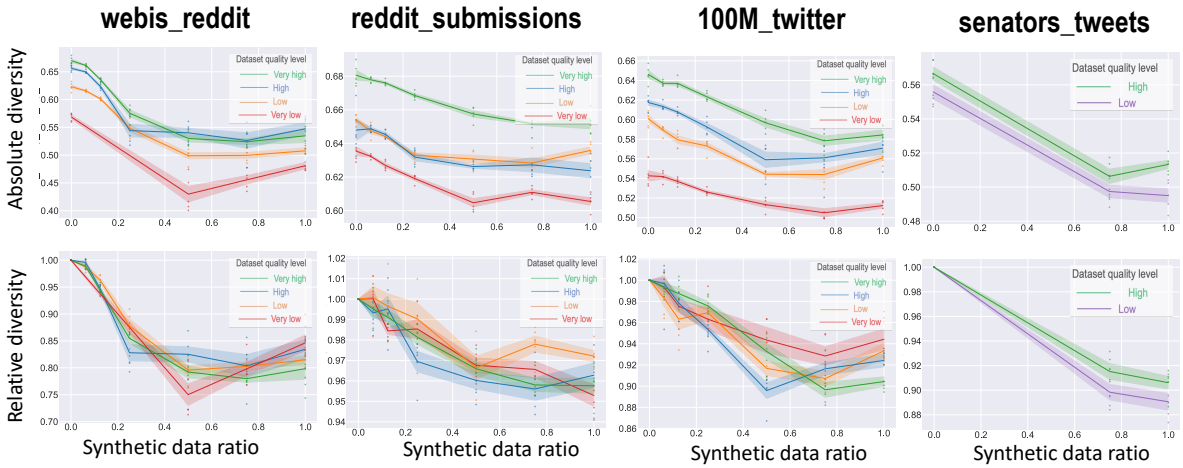


Figure 11: **Effect of manipulating dataset quality on sensitivity to synthetic data ratio, for four different datasets.** Absolute measures correspond to the value of the corresponding metric at generation 19. Relative measure correspond to absolute values divided by the metric value after a single fine-tuning episode (i.e. generation 0).No clear effect is observed, except potentially on the *senator_tweets* dataset.

C.3 Toy model exploring the causes of the non-linear relationship between diversity loss and synthetic-data ratio

In section 4.2, we observed that for the Wikipedia dataset, the relationship between diversity loss and synthetic-data ratio was non-linear. Indeed, the greatest drops in diversity are observed for intermediate synthetic-data ratios, rather than for high values as in other datasets and previous works (Bertrand et al., 2023; Bohacek and Farid, 2023; Kazdan et al., 2024). While at first surprising, we believe this pattern can be explained if we assume that synthetic data aligns more with the models' priors. This is not a strong assumption given that that data was generated by other fine-tuned versions of the same base models. The intuition, which we experimentally confirm below, is that those datapoints aligned with the model's priors, have a stronger effect on the training process. And this then leads to the intermediate synthetic data ratios to essen-

tially learn from less data. Let us consider the three different synthetic-data ratios:

- When synthetic-data ratio is low: While the model preferentially learns from synthetic data, most of its training data is human generated. Therefore, the model learns a distribution that resembles the human data despite this bias.
- When synthetic-data ratio is high: the model preferentially learns from synthetic data, but anyway most of the training data is synthetic. This bias thus does not have a large effect.
- When synthetic-data ratio is intermediate: the model preferentially learns from synthetic data, and receives human data and synthetic in comparable proportion. However, the bias will lead the model to essentially discard human data, and to learn only from the synthetic data, just like for high synthetic-data ratios. The difference is that the pool of synthetic data to learn from is here lower than in the high synthetic-data ratio.

To test this hypothesis, we develop a toy example where we could manipulate whether learning from synthetic data aligned with models' priors is favored. In this model, the true (*human*) distribution is a uniform distribution over integers in $[0, N]$. The *model* is implemented as a normalized histogram over training datapoints, which are a combination of true and synthetic datapoints from the previous generations. We sample N points from the true distribution, and normalize the resulting histogram to get the first model. Then, we sample $r * N$ points from this model, and $(1 - r) * N$ points from the true distribution. We again derive a probability distribution from this sample to get the new model. We repeat this process for 20 time steps.

To introduce the bias mentioned in our hypothesis, we assume that the models have a prior to sample multiples of 2. We thus multiply the histogram by a corresponding bias vector before normalizing. Additionally, we can manipulate whether the human data overlaps with this bias: we can modify the true distribution so that it does not contain multiples of 2.

We then ran 50 simulations for different values of synthetic-data ratio, manipulating whether learning is biased and whether the true distribution overlaps with this bias. As show in Figure 12, we observe that when learning is biased and the true distribution does not overlap (right column), we can reproduce the U-shape found for the Wikipedia dataset. This non-linear relationship disappears when we remove this bias (left column) or when the true distribution overlaps with this bias. We were able to observe this pattern both in the Accumulation (top-row) and no-Accumulation (bottom row) settings. This is therefore consistent with our hypothesis for explaining the observed U-shape relationship.

C.4 Regression models' explained variances

Here we provides r^2 measures for the linear regression models used in experiments in sections 4.3 and 4.4. Table 3 corresponds to table 1 from section 4.3. Table 4 corresponds to Figure 4 from section 4.4. We can see that the considered properties can partially account for observed effect, many more effects remain to be uncovered (in particular in the 100M tweets and wikipedia datasets).

C.5 Additional experiments on the distribution shift of political lean

We provide here the additional figures that are discussed in section 4.5. On Figure 13a we observe a steady increase in the proportion of "perfectly neutral" tweets (with assigned a score of exactly 50). On Figure 13b while we observe a slight increase in the number of non-political tweets, those remain marginal, indicating that the models are able to maintain the focus on political topics (Figure 13a). This suggests that generated tweets remain in the topic of politics, but drift from strong partisanship.

Figure 13a revealed that the shift is partly driven by an increase in politically neutral content. To isolate the different mechanisms at play, we performed the same analysis, but without taking politically neutral tweets into account (Figure 14). The consequence of this manipulation was to accentuate the observed asymmetry, as the political lean that minimizes shift magnitude moves even more toward the left. This suggests that there might two interacting mechanisms influencing political lean evolution: first, a tendency to generate politically-neutral content; and a tendency to shift the distribution toward left-wing content.

Experiments in the paper were conducted with mixed-model iterative chains - each generation a fresh base model is sampled out of four possible models options (`LLama-3.2-1B`, `Qwen2.5-1.5B`, `SmolLM-1.7B`, `Falcon3-1B-Base`). As different

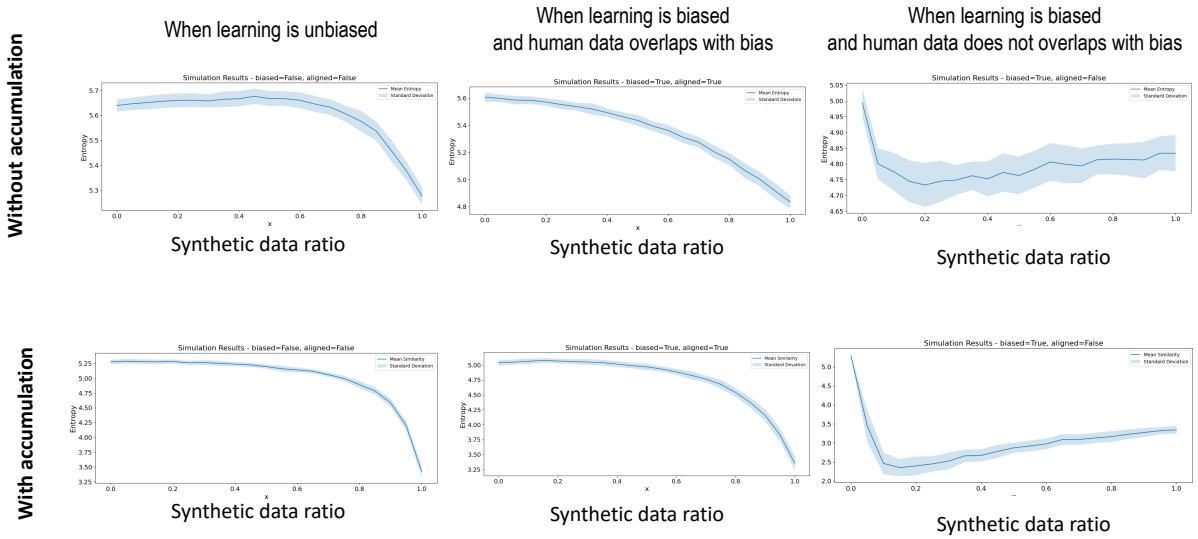
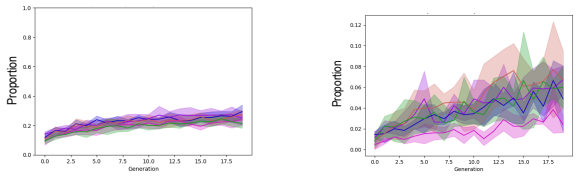


Figure 12: **Simulation results of the toy model.** Under specific conditions, a non-linear u-shaped relationship between diversity loss and synthetic-data ratio emerges.

$\mathbf{r^2}$	All	Webis		100M Tweets		Reddit Submissions		Wikipedia	
Synthetic data ratio		1/8	1/4	1/8	1/4	1/8	1/4	1/8	1/4
Semantic Diversity	0.293	0.244	0.541	0.043	0.188	0.109	0.502	0.147	0.105
Quality	0.737	0.482	0.353	0.193	0.324	0.143	0.564	0.249	0.276

Table 3: R^2 results corresponding to Table 1.



(a) Proportion of politically neutral tweets

(b) Proportion of non-political tweets

Figure 13: (a) Proportion of politically neutral tweets increases, implying that models tend to avoid strong political statements. (b) Proportion of non-political tweets marginally increases, implying that the models stay on the topic of politics.

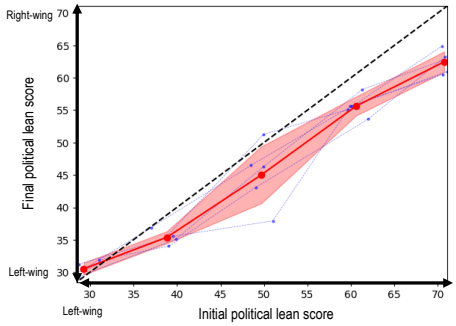


Figure 14: Average political lean at the last generation as a function of political lean in the true distribution after excluding political tweets

Domain	R^2 (Diversity)	R^2 (Quality)
Webis	0.340	0.591
100M Tweets	0.532	0.611
Wikipedia	0.241	0.363

Table 4: R^2 results corresponding to Figure 4.

models may display different biases with respect to political lean, we also ran an experiment with homogeneous chains, where the same base model is used in each generation. To clarify, in each generation the training still starts from a new instance of a pretrained model, e.g. over 20 generations we will initialize 20 separate LLama-3.2-1B instances. Figure 15 compares homogenous chains corresponding to the four considered models. We observe little variation between the chains, although Falcon3-1B seem to display slightly more pronounced left-wing bias, while this bias is weaker for Llama-3.2-1B.

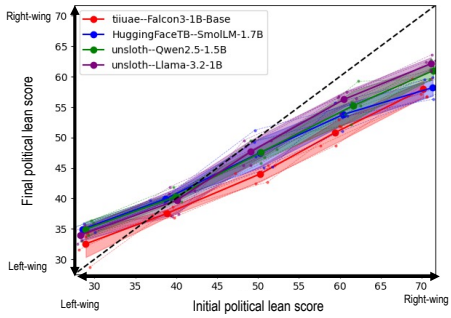


Figure 15: Average political lean at the last generation as a function of political lean in the human distribution with homogeneous transmission chains. We observe slight differences between chain: for instance, Falcon3-1B chain appears to have a stronger left-wing bias than others, while it is weaker for Llama-3.2-1B