

Data Pollination: An Emergent Ecological Process Driving AI Population Evolution

Shufang Xie^{1*}, Qizhi Pei^{1*}, Ang Lv¹, Jingyang Hu¹, Lijun Wu², Rui Yan^{3†}

¹Gaoling School of Artificial Intelligence, Renmin University of China,

²Shanghai Artificial Intelligence Laboratory,

³School of Artificial Intelligence, Wuhan University

{shufangxie, qizhipei, anglv, hjy2024104281}@ruc.edu.cn,

apeterswu@gmail.com, rui.yan@whu.edu.cn

Abstract

AI development is often framed as the outcome of isolated research and engineering efforts, yet evidence from deployed systems suggests that language models interact through a shared data ecosystem. While the optimization of individual models is extensively studied, the emergent properties of this interconnected population remain largely unexplored, limiting our ability to predict long-term ecosystem trajectories. We term this process *data pollination*, the unintentional circulation of synthetic model outputs through shared online platforms and web-scale training corpora, and formalize it as a population-based evolutionary framework to investigate stability dynamics under synthetic data training. Our theoretical analysis and controlled experiments involving 320 language models demonstrate that population dynamics can mitigate the model collapse observed in single-lineage recursive training, yielding stable or improving performance across diverse benchmarks. Crucially, we find that ecological diversity functions as a fundamental resilience mechanism that safeguards the ecosystem against collapse, highlighting the critical importance of maintaining model diversity for sustainable AI development.

1 Introduction

AI development typically assumes that model capabilities emerge through deliberate engineering and explicit, expert-designed mechanisms, including architectural innovations, optimization algorithms, and strategies for data curation and scaling for pre-training and instruction tuning (Vaswani et al., 2017; Kingma and Ba, 2015; Ouyang et al., 2022; Brown et al., 2020; Dodge et al., 2021).

However, this view overlooks a fundamental ecological process we term *data pollination*: the

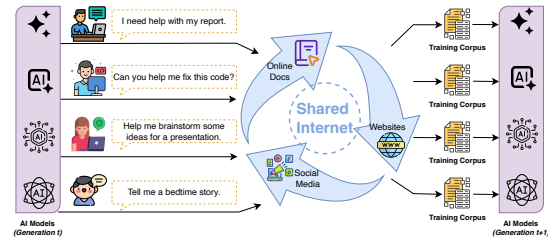


Figure 1: **Data pollination cycle.** Deployed models (Generation t) populate the web with synthetic outputs. This content is scraped to train the next generation (Generation $t + 1$), creating an unmanaged evolutionary loop that transmits behavioral traits across model families.

unintentional circulation of synthetic model outputs across shared training corpora. Analogous to bees transferring pollen while foraging (Bronstein et al., 2006), users selecting capable models and publishing their outputs implicitly construct inheritance pathways that transcend organizational boundaries. As shown in Figure 1, this decentralized feedback loop operates through routine practices: users generate content with preferred models, web scrapers harvest this mixed content, and organizations train new generations on these pollinated corpora (Dodge et al., 2021; Bender et al., 2021; Osborne et al., 2024). Thus, individual deployment decisions collectively induce population-level evolution without centralized coordination.

Despite the central role of this emergent data pollination process, most work on AI behavior still treats evolution as an engineered training primitive within centrally controlled pipelines (Jaderberg et al., 2017; Real et al., 2019). In contrast, we focus on decentralized population dynamics that arise from routine human activities that unintentionally construct the evolutionary infrastructure of AI.

To understand the implications of this ecological perspective, we combine measurement in real ecosystems with controlled evolutionary modeling.

*Equal contribution.

†Corresponding author: Rui Yan (rui.yan@whu.edu.cn).

We begin by investigating whether data pollination already impacts deployed systems. Our analysis of linguistic fingerprints from a diverse set of models across five major families reveals substantial cross-family behavioral convergence, indicating that nominal organizational boundaries do not prevent population-level mixing.

Building on this empirical grounding, we then simulate the long-run trajectories of these dynamics through large-scale evolutionary modeling involving extensive model populations. These experiments demonstrate that population-based evolution fundamentally alters the collapse observed in isolated recursive training: instead of model collapse, evolving populations maintain robust performance across diverse benchmarks.

Finally, we trace this resilience to ecological diversity. By integrating theoretical Gaussian analysis with empirical measurements, we show that population-based training preserves the distributional variance that single models lose. This diversity acts as functional redundancy, safeguarding the ecosystem against systemic failure where individual lineages might otherwise degrade.

These results demonstrate that data pollination substantially reshapes the long-run impact of synthetic data on model quality: although individual lineages may collapse, diverse, high-fidelity behaviors can persist at the population level even under predominantly synthetic training.

Taken together, these empirical and theoretical findings reveal that contemporary artificial intelligence development operates as an interconnected ecosystem. Specifically, we make the following contributions:

- We identify and quantify *data pollination* in the wild by analyzing 27 deployed models, revealing cross-family behavioral convergence that challenges assumptions of independent lineages;
- We formalize this phenomenon as a population-based evolutionary framework, providing a theoretical foundation that establishes *ecological diversity*, in addition to individual optimization, as a fundamental prerequisite for AI ecosystem resilience;
- We validate this framework through large-scale experiments with 320 language models, demonstrating that evolutionary populations naturally maintain this diversity to prevent collapse, offering a practical strategy for sustainable AI development.

2 Related Work

Unlike research treating AI evolution as an engineered optimization primitive (Real et al., 2019; Elsken et al., 2019; Akiba et al., 2025; Zhang et al., 2025; Wortsman et al., 2022; Jaderberg et al., 2017), we model language models as a distributed population evolving through decentralized data pollination. In our framework, user choices, content publication, and web-scale scraping create inheritance pathways across independent model families.

Recursive training on synthetic data can degrade model quality, leading to tail loss and irreversible forgetting (Shumailov et al., 2024; Bertrand et al., 2024; Dohmatob et al., 2024; Seddik et al., 2024; Zhu et al., 2025; Suresh et al., 2025), though recent work questions the universality of these risks (Schaeffer et al., 2025). Meanwhile, distilled outputs can covertly transmit behavioural traits through semantically unrelated tokens (Jiang et al., 2025; Cloud et al., 2026). While prior studies treat synthetic data as a controlled variable within a single pipeline or simulated network (Wang et al., 2025; Chen et al., 2024a,b), we analyze uncoordinated *data pollination* in the wild. We show that interacting populations trained on synthetic data can leverage selection and diversity to reshape collapse dynamics and preserve high-fidelity lineages.

This risk is exacerbated by *algorithmic monoculture*, where widespread reliance on dominant models homogenizes the information ecosystem (Bommasani et al., 2022) and reduces the collective diversity of novel content (Doshi and Hauser, 2024; Wu et al., 2025). Our findings demonstrate that population-level diversity acts as a necessary counterforce to this process, echoing resilience mechanisms in biological ecosystems (Luck et al., 2003).

Distribution estimation and data forensics infer properties of opaque pretraining corpora from model behavior (Liang et al., 2025; Ravichander et al., 2025; Bellini et al., 2024). This behavior-based lens has motivated automated detectors such as GLTR and DetectGPT (Gehrmann et al., 2019; Mitchell et al., 2023). It has also motivated a lexical-pattern generation and analysis toolkit for LLM outputs (Paech, 2025), as well as corpus documentation and stylometric fingerprinting tools for model outputs (Dodge et al., 2021; Bender et al., 2021; Anderson et al., 2025). We build on these approaches but shift the focus from reconstructing a single model’s pretraining distribution to characterizing population-level convergence.

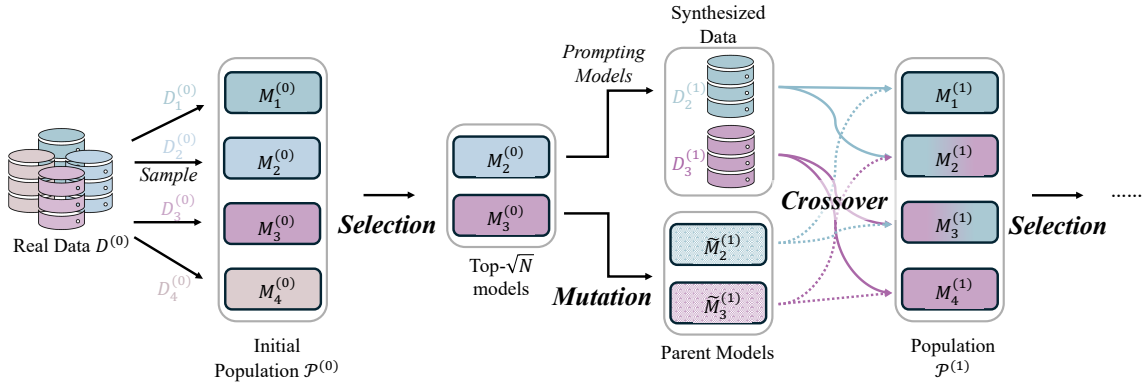


Figure 2: **Population-based evolutionary framework.** Initial population $\mathcal{P}^{(0)}$ contains N models, with each model $M_i^{(t)}$ paired with dataset $D_i^{(t)}$ at generation t . The iterative process applies selection of top performers, crossover through data-mediated inheritance, and mutation introducing parameter variations.

3 Population-Based Evolutionary Model of Data Pollination

Motivated by the ecosystem-level data pollination cycle described above, we formalize data pollination as a population-based evolutionary process acting on language models. This section presents a conceptual framework that characterizes how selection, variation, and data-mediated inheritance jointly shape population dynamics, independent of any particular experimental instantiation. Figure 2 illustrates the overall architecture of this population-based evolutionary framework.

Population. At generation t , the language model population is $\mathcal{P}^{(t)} = \{M_1^{(t)}, M_2^{(t)}, \dots, M_N^{(t)}\}$, where each model $M_i^{(t)}$ is parameterized by weights $\theta_i^{(t)}$ and paired with a synthetic dataset $D_i^{(t)}$ consisting of its outputs on a shared pool of prompts. The initial population $\mathcal{P}^{(0)}$ is trained on human-generated text, and all subsequent generations learn exclusively from synthetic data produced within the population.

Selection. Selection formalizes survival of the fittest at the population level. At each generation, interactions with human users induce a preference relation \preceq over models in $\mathcal{P}^{(t)}$, indicating which systems are more likely to be selected and reused in practice. The selection operator chooses a subset $\mathcal{S}^{(t)} \subseteq \mathcal{P}^{(t)}$ of size $k = \sqrt{N}$ consisting of \preceq -maximal elements, representing the models whose behaviors are most likely to be amplified in subsequent generations through usage and data output. The choice $k = \sqrt{N}$ ensures that the $k \times k$ ordered parent pairs used in crossover yield $k^2 = N$ off-

spring, keeping the population size constant across generations.

Mutation. Mutation introduces controlled variation into model parameters while preserving overall competence. For each parent model in $\mathcal{S}^{(t)}$, mutation is represented as the addition of small, state-dependent noise to its parameters. At a high level, this operator abstracts the accumulation of stochastic perturbations in deployed systems, including quantization and numerical errors, incremental re-training, and configuration changes.

Crossover and data-mediated inheritance. The crossover operator couples parameter and data lineages, capturing data pollination as inheritance through shared corpora. Formally, crossover is modeled as a mapping $f_{\text{xo}} : \mathcal{S}^{(t)} \times \mathcal{S}^{(t)} \rightarrow \mathcal{P}^{(t+1)}$, where $f_{\text{xo}}(M_i^{(t)}, M_j^{(t)}) = M_{i,j}^{(t+1)}$. For each ordered pair $(M_i^{(t)}, M_j^{(t)}) \in \mathcal{S}^{(t)} \times \mathcal{S}^{(t)}$, the offspring $M_{i,j}^{(t+1)}$ inherits parameters from the (mutated) weights of the first parent and a training dataset derived from the synthetic outputs $D_j^{(t)}$ of the second parent. Iterating crossover over parent pairs therefore yields a new population whose members inherit traits both from parameter ancestry and from exposure to other models' synthetic outputs, abstracting the way web-scale corpora mix content produced by many deployed systems.

Together, these operators define a population-level evolutionary process on language models (Figure 2) that abstracts deployed ecosystems: user adoption induces selection, routine model updates introduce mutation, and shared corpora implement crossover through data pollination.

We investigate three research questions:

- **RQ1:** Do deployed language models show cross-family behavioral convergence consistent with data pollination (Section 4)?
- **RQ2:** In synthetic-only training, can evolutionary population dynamics prevent the collapse observed in recursive baselines (Section 5)?
- **RQ3:** How does ecological diversity promote systemic resilience against failure modes (Section 6)?

4 Real-World Evidence of Data Pollination

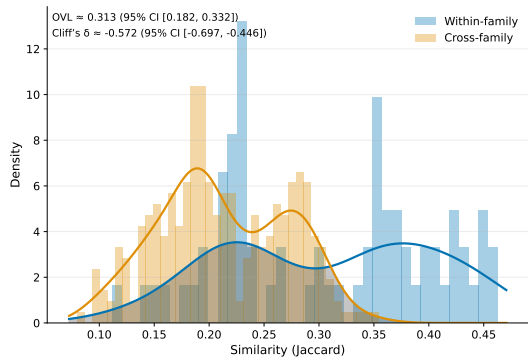


Figure 3: **Cross-family behavioral convergence (Distribution)**. Density of pairwise behavioral similarity scores, showing substantial overlap between within-family (blue) and cross-family (orange) distributions.



Figure 4: **Cross-family behavioral convergence (Similarity matrix)**. Row-normalized pairwise similarity matrix across five model families. High off-diagonal values indicate strong inter-family similarity.

Directly tracing data pollination is challenging due to the opacity of commercial training sets (Dodge et al., 2021) and the diminishing reliability of synthetic text discriminators (Sadasivan et al., 2025; Valiaiev, 2024). We therefore estimate population-level relationships by analyzing linguistic fingerprints from 27 deployed models across five major families using a lexical-pattern generation and analysis toolkit for LLM outputs (Paech,

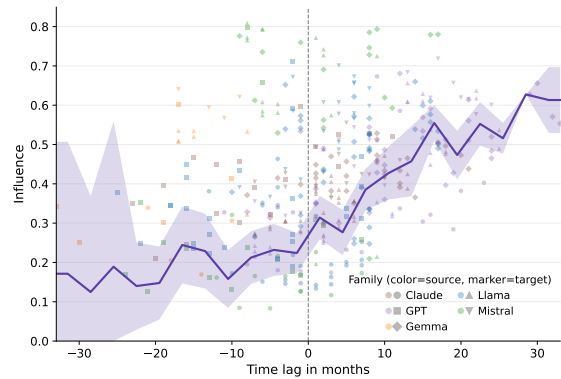


Figure 5: **Temporal asymmetry of cross-family influence**. Directed influence vs. time lag. Positive lags (past \rightarrow future) show significantly higher influence, indicating causal flow from older to newer models. Shaded regions: 95% CI.

2025). This method quantifies behavioral convergence by identifying over-represented stylometric patterns (uncommon repetitive n-grams) in model outputs. It filters standard stop words to capture distinctive habits rather than shared vocabulary.

We observe substantial cross-family overlap (Overlap Coefficient, OVL = 0.313, 95% CI: [0.182, 0.332]) and moderate effect sizes (Cliff's $\delta = -0.572$, 95% CI: [-0.697, -0.446]) in Figure 3. This is further corroborated by the row-normalized pairwise similarity matrix (Figure 4), where off-diagonal (cross-family) elements exhibit magnitudes comparable to diagonal (within-family) self-similarity. These patterns indicate that approximately 31% of the distributional mass is shared between cross-family and within-family similarity distributions, suggesting that nominal organizational boundaries do not prevent behavioral mixing.

To definitively distinguish directional pollination from generic convergence, we further analyze the temporal structure of these similarities (Figure 5). By mapping pairwise influence against model release dates, we quantify the directed influence $I(A \rightarrow B)$ as the salience-weighted proportion of the older model A 's fingerprints that appear in the newer model B (exact form in Appendix B). Visualizing this metric reveals a distinct temporal asymmetry: models exhibit significantly higher behavioral overlap with older models from different families than the reverse. This lagged correlation, peaking at positive time lags, suggests a causal pathway where the outputs of widely deployed models permeate the web, serving as train-

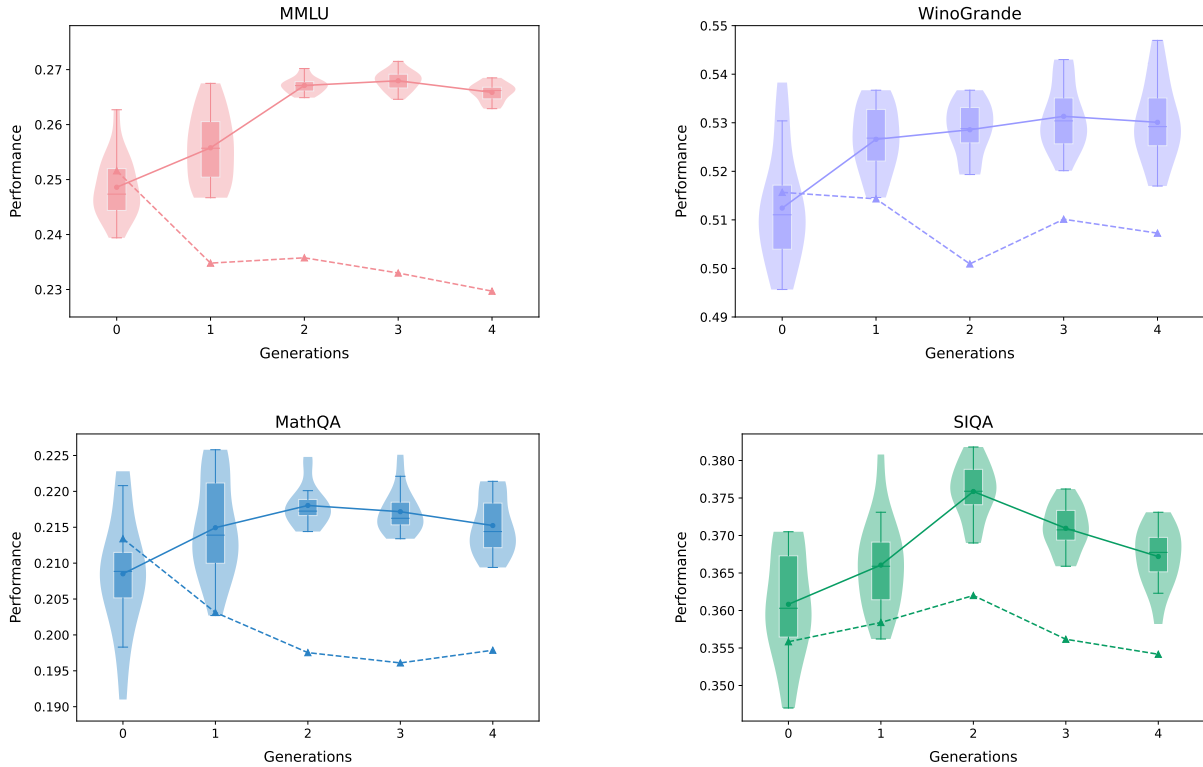


Figure 6: **Evolutionary populations outperform recursive baselines.** Violin plots show performance distributions (MMLU, WinoGrande, MathQA, SIQA) over five generations. Evolutionary methods (violins) consistently exceed single-lineage baselines (dashed lines), which exhibit rapid collapse.

ing fodder for subsequent generations. Crucially, this asymmetry distinguishes data pollination from generic convergence on shared pre-training corpora (e.g., Common Crawl), which would result in symmetric similarity independent of release order. Detailed methodological procedures and statistical formulations are provided in Appendix B.

Taken together, these measurements suggest that deployed models likely operate as an interconnected system. While observing data pollination in the wild is necessarily indirect given the confounding factors, these patterns motivate our controlled synthetic-only experiments where we can definitively isolate the effects of synthetic data.

5 Evolutionary Prevention of Model Collapse

This section presents empirical validation of our evolutionary framework in a controlled synthetic-only setting. We provide both quantitative comparisons against recursive baselines and qualitative analyses of generation quality to demonstrate how population dynamics prevent model collapse.

5.1 Quantitative Analysis

We evaluate our framework on 320 models across four populations ($N = 16$ for each population). Following established protocols for studying model collapse (Shumailov et al., 2024), we employ models with approximately 100M parameters to enable extensive population-level experimentation. We use MMLU (Hendrycks et al., 2021), WinoGrande (Sakaguchi et al., 2021), MathQA (Amini et al., 2019), and SIQA (Sap et al., 2019) as selection proxies. Implementation details are provided in Appendix C and Appendix D.

Our results indicate that evolutionary population dynamics fundamentally change how systems learn from synthetic data, preventing the model collapse seen in isolated recursive training (Figure 6). While baselines degrade rapidly as predicted by theory, evolutionary populations maintain stable performance or achieve sustained improvements across all benchmarks. In reasoning tasks like MMLU and MathQA, the population not only avoids decay but also consistently improves, showing that selection pressure can identify useful training signals even in purely synthetic environments. Even where performance stabilizes (e.g., SIQA), the population

structure prevents regression, keeping capabilities significantly above the collapsing baseline and confirming the value of ecological diversity.

This resilience is especially important given our strict synthetic-only setting, where recursive training typically leads to failure. By separating the survival of information from individual models, the framework creates a collective correction mechanism: selection promotes high-quality outputs while crossover spreads beneficial traits across the population. These findings confirm that the degradation often associated with synthetic data is not inevitable, but rather a limitation of individual optimization that can be resolved through population-level organization.

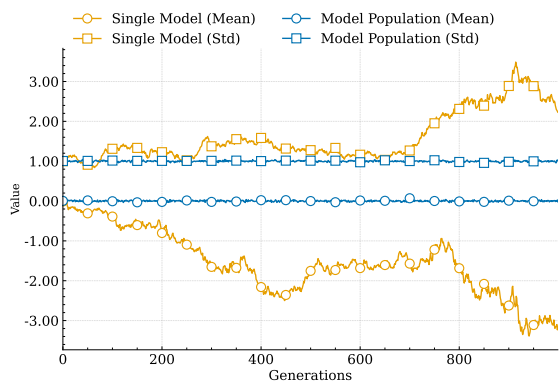


Figure 8: **Gaussian validation of collapse mitigation.** Evolution (blue) maintains population statistics near the ground truth ($\mu = 0, \sigma = 1$) across generations, whereas single-model recursive training (orange) drifts significantly. Circles and squares denote estimated means and standard deviations, respectively.

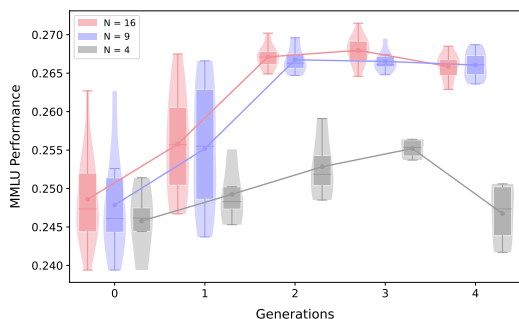


Figure 9: **Population size as a driver of resilience.** MMLU performance distributions for populations of sizes $N = 4, 9, 16$. Larger populations ($N = 16$) show robust improvement and stability, whereas smaller groups ($N = 4$) exhibit instability and premature stagnation.

5.2 Qualitative Analysis

Qualitative examples (Figure 7) confirm these patterns, with additional cases in Appendix E (Tables 3 and 4). While all models initially generate coherent text, behavior diverges significantly over time.

As depicted in the right column of Figure 7, models trained recursively without evolution suffer from rapid degradation. By Generation 2, the single model begins to lose narrative coherence, abruptly inserting unrelated metadata and URL-like strings. In later generations, the output further devolves into nonsensical sequences or repetitive, hallucinated web-crawled artifacts, consistent with the model collapse described by Shumailov et al. (2024).

Conversely, models incorporating evolutionary training (Figure 7, left column) consistently produce linguistically valid and semantically coherent English sentences throughout all generations. Even at Generation 4, the population-based model maintains the original narrative thread, describing character interactions and scenes without degenerating into repetitive loops or structural failure.

These findings illustrate how population-based methodologies can preserve output diversity and avoid the extreme collapse behavior observed in single-model baselines within our experimental setting. Qualitative trajectories therefore complement our quantitative metrics and further address RQ2 by showing how evolutionary training preserves coherent, meaningful generations over time.

6 Diversity and Ecosystem Resilience

This section investigates the ecological mechanisms underlying the observed resilience. We analyze the role of diversity through a theoretical Gaussian framework and empirical studies of population size and semantic embedding space, highlighting how collective diversity mitigates systemic failure.

6.1 Theoretical Intuition

To analyze the divergence between population and individual behaviors, we extend the single-model Gaussian analysis of Shumailov et al. (2024) to a population setting. Consider a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ where estimators recursively fit synthetic samples. We measure estimation quality using the squared Wasserstein-2 distance $W = (\mu - \hat{\mu})^2 + (\sigma - \hat{\sigma})^2$. For a single estimator, random sampling errors accumulate over generations, causing W to grow and leading to

progressive collapse.

However, moving to a population of K estimators fundamentally alters this dynamic by introducing diversity. As we prove in Appendix A.1, the probability that at least one estimator maintains high fidelity increases exponentially with K . This theoretical result suggests that ecological diversity effectively functions as a resilience mechanism: the presence of multiple independent lineages ensures that accurate approximations of the distribution persist, providing robustness against the inevitable degradation of individual models.

Empirical Validation in Gaussian Models. We validate this intuition by simulating the Gaussian process for 1000 generations (Figure 8). The trajectories explicitly illustrate the protective effect of diversity. The single-estimator baseline (orange lines) suffers from uncorrected error accumulation: its estimated mean drifts aggressively away from the true value $\mu = 0$ (reaching < -2.5 in the figure), while its variance fluctuates wildly, reflecting a loss of distributional fidelity. Conversely, the population-based estimators (blue lines) exhibit remarkable stability. Despite individual fluctuations,

the population selection mechanism keeps the collective mean and standard deviation locked to the ground truth ($\mu = 0, \sigma = 1$) across the entire timeline. This visual contrast underscores that while isolated lineages are vulnerable to stochastic drift, a diverse population acts as a statistical anchor, preventing the ecosystem from converging to failure.

6.2 Population Size as a Driver of Resilience

Population size constitutes a critical parameter for resilience. We conducted systematic experiments with $N = 4, 9,$ and 16 individuals (Figure 9). Smaller populations ($N = 4$) exhibited constrained exploration, yielding inconsistent improvements followed by decline after generation 3. Medium populations ($N = 9$) demonstrated more stable trajectories. The largest population ($N = 16$) achieved the most rapid initial gains and sustained performance. These findings indicate that sufficiently large populations are essential for stable dynamics, suggesting a threshold effect where population size directly influences stability.

Semantic Diversity Preservation. To quantify diversity preservation, we analyzed model output

Gen.	Model Population	Single Model
0	again by hand. "I thought I would come back," says Alexander (who had actually read the book himself), "and after leaving my house." Alex is not alone. He is also the last survivor of a mysteriously forgotten history...	again by hand. With the news of crowds passing from Sudbury (with the exception of the police) to Arlington after moving towards Sydney (where he died just 16 years ago), Alex's mourning would have been unwise...
1	again and tried to return it back to its original location when he found it. His search failed because the tree behind his house stood too close to the library, leaving no space between him and the book itself; so, despite trying to find it, Alexander named it at the top...	again by hand. He realized that a scrapbook could not stand still when he tried something else; no matter how many people were waiting for him to ask for help from his own elders, what would happen if he couldn't get back there?...
2	again by hand; now there are no more than three copies left behind—and no one else has seen them! But at first, after several minutes, when this is not possible to see anything beyond what would otherwise have been visible, we were surprised that all sorts of things happened...	again! This article originally appeared here: http://www.biolandiaarchivscienceonline.com/article_of%783194606#crownsience+videos=
3	again after being told that there were no books left behind; instead they took turns reading each sentence aloud while waiting until morning to pick them off again later when someone asked him what kind of thing would happen next!...	again! This article originally appeared here: http://www.youtube.com/watch?v=UJPOZKfMz7g3dBkWc69Y8lGaE504Lw_C
4	again afterward; now there are no signs yet! There is nothing left except a small piece of wood cut off from the floor above them.....and when someone comes along they say 'Hey.' But I am sure you won't find anything else..."	again next weekend....and finally, afterward I went back home ...and felt like my friend had come across me walking down stairs looking towards us when we arrived here today! Historical records show how many different cultures live within each country...

Figure 7: **Qualitative comparison of generation quality across generations.** Each row shows continuations of the same fantasy prompt from population-based evolutionary models (left) and from single models trained recursively on synthetic data (right). Population-based models maintain coherent, grammatical narratives through generation 4, whereas single models progressively degrade into nonsensical, URL-like outputs characteristic of collapse.

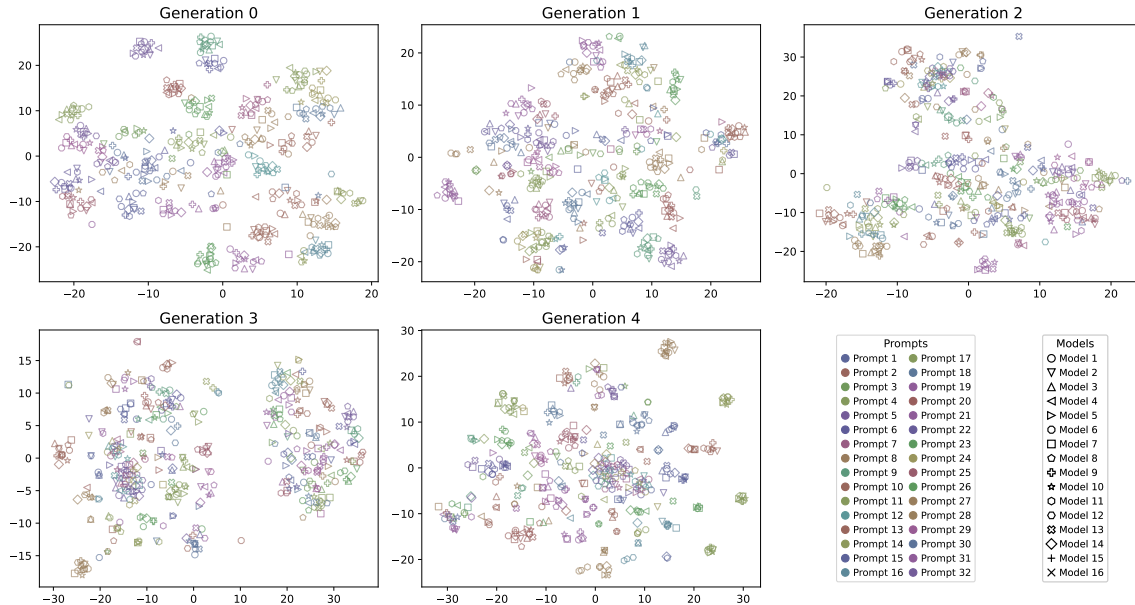


Figure 10: **Semantic diversity preservation.** t-SNE projections of model completions for generations 0–4 (WinoGrande). The persistence of distinct clusters by prompt (colors) shows that models retain the ability to distinguish different inputs, while the distributed spread of model shapes within clusters indicates that the population preserves behavioral diversity, avoiding the convergence to identical outputs observed in single-model baselines.

distributions in semantic embedding space. We sampled 32 prompts from the training dataset and generated completions using all models per generation. Text embeddings were extracted using RoBERTa (Liu et al., 2019) and projected via t-SNE (Van der Maaten and Hinton, 2008) for visualization. Embedding space analysis reveals evolutionary populations maintain substantial semantic diversity throughout training (Figure 10). Generated completions exhibit distributed clustering patterns rather than converging to homogeneous outputs, indicating models preserve distinct semantic styles and content generation strategies. Persistent separation between prompt responses demonstrates evolutionary mechanisms prevent semantic collapse typical in iterative training. This diversity preservation occurs despite selection pressure favoring high-performing models, suggesting the framework successfully balances performance optimization with output heterogeneity. Diverse populations provide varied perspectives on inputs, reducing systematic error risks and enhancing reliability while maintaining adaptive potential across generations.

Mitigating Algorithmic Monocultures. We show that population-based evolution offers a vital countermeasure to the risks of *algorithmic monoculture* (Bommasani et al., 2022). In monocul-

tural ecosystems, the widespread reliance on shared components creates systemic fragility: failures are no longer independent but correlated, leading to outcome homogenization where specific errors or blind spots become entrenched globally rather than locally (Bommasani et al., 2022). While single-lineage recursive training accelerates this degeneration by collapsing behavioral distributions (Shumailov et al., 2024), we demonstrate that ecological diversity functions as a critical structural safeguard. By preserving the collective novelty (Doshi and Hauser, 2024), population-based evolution ensures functional redundancy: even if individual lineages degrade, the aggregate ecosystem retains the diverse capabilities necessary for adaptation. This finding suggests that diversity is not merely a beneficial feature, but a fundamental prerequisite for the long-term sustainability of AI ecosystems.

7 Conclusion

In this work, we highlight an important yet under-explored phenomenon: the emergent evolution of language models as an interconnected global population. While individual models are often viewed in isolation, we argue that data pollination, the unintentional circulation of synthetic outputs through shared substrates, creates a *de facto* evolutionary ecosystem. Our real-world observations, theoretical analysis, and controlled experiments demon-

strate that while this connectivity leads to cross-family behavioral convergence, the resulting population dynamics also function as a natural defense against synthetic-data collapse.

These findings advocate for a paradigm shift in AI research: moving beyond the study of isolated models to the stewardship of an evolving ecosystem. We suggest that maintaining ecological diversity is not just a safety buffer, but a fundamental prerequisite for sustainable progress in a synthetic-data era. By treating population dynamics as a core design variable, the community can transform the risks of circular training into mechanisms for long-term systemic resilience.

Limitations

Our behavioral fingerprinting methodology measures linguistic style similarities but cannot definitively distinguish whether convergence arises from data-mediated capability transfer through synthetic data circulation, convergent training practices adopted independently across organizations, or overlapping use of common public datasets. While observed patterns are consistent with data pollination mechanisms, alternative explanations remain plausible. Without access to proprietary training corpora, definitive causal attribution remains infeasible, motivating future investigations with controlled data provenance tracking and population-level interventions.

Additionally, our controlled evolutionary experiments use 100-million parameter models trained on 1 billion tokens, which differ substantially from contemporary deployed systems operating at scales exceeding 100 billion parameters with trillions of training tokens. This scale gap affects generalizability, as smaller models may exhibit different collapse dynamics and evolutionary behaviors than large-scale production systems. Nevertheless, the experiments establish proof-of-concept for population-based dynamics and provide a theoretical foundation through Gaussian analysis that may generalize across scales.

Third, our theoretical analysis relies on a simplified Gaussian abstraction. While this provides intuition for population dynamics, characterizing the complete optimization landscape of deep Transformer networks remains mathematically intractable.

Finally, our study focuses specifically on the role of population dynamics in preventing model

collapse; we leave the analysis of other ecosystem-level effects on model capability development and alignment for future work.

References

- Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. 2025. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 7:195–204.
- Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. [Mathqa: Towards interpretable math word problem solving with operation-based formalisms](#). In *Proceedings of the 2019 Conference of the North*, pages 2357–2367. Association for Computational Linguistics.
- Bryce Anderson, Riley Galpin, and Tom S Juzek. 2025. Model misalignment and language change: Traces of ai-associated language in unscripted spoken english. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, volume 8, pages 179–191.
- Valentina Bellini, Federico Semeraro, Jonathan Montomoli, Marco Cascella, and Elena Bignami. 2024. Between human and ai: assessing the reliability of ai text detection tools. *Current Medical Research and Opinion*, 40(3):353–358.
- 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*, pages 610–623.
- Quentin Bertrand, Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. 2024. [On the stability of iterative retraining of generative models on their own data](#). In *The Twelfth International Conference on Learning Representations*.
- Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems*.
- Judith L Bronstein, Ruben Alarcón, and Monica Geber. 2006. The evolution of plant-insect mutualisms. *New Phytologist*, 172(3):412–428.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah Abdin. 2024a.

- [On the diversity of synthetic data and its impact on training large language models.](#) *arXiv preprint arXiv:2410.15226*.
- Jie Chen, Yupeng Zhang, Bingning Wang, Wayne Xin Zhao, Ji-Rong Wen, and Weipeng Chen. 2024b. Unveiling the flaws: Exploring imperfections in synthetic data and mitigation strategies for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14855–14865.
- Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Szyber-Betley, Sören Mindermann, Jacob Hilton, Samuel Marks, and Owain Evans. 2026. [Language models transmit behavioural traits through hidden signals in data.](#) *Nature*, 652:615–621.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305.
- Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. 2024. A tale of tails: Model collapse as a change of scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11165–11197. PMLR.
- Anil R Doshi and Oliver P Hauser. 2024. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290.
- Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55):1–21.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.
- Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. 2017. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*.
- Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025) Datasets and Benchmarks Track*.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*.
- Hao Liang, Keshi Zhao, Yajie Yang, Bin Cui, Zenan Zhou, and Wentao Zhang. 2025. Training data distribution estimation for optimized pre-training data management. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 4640–4648. IEEE.
- 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.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. [Fineweb-edu: the finest collection of educational content.](#)
- Gary W Luck, Gretchen C Daily, and Paul R Ehrlich. 2003. Population diversity and ecosystem services. *Trends in Ecology & Evolution*, 18(7):331–336.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*.
- Cailean Osborne, Jennifer Ding, and Hannah Rose Kirk. 2024. [The ai community building the future? a quantitative analysis of development activity on hugging face hub.](#) *Journal of Computational Social Science*, 7(2):2067–2105.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*.
- Samuel J Paech. 2025. [Slop forensics: A toolkit for generating & analyzing lexical patterns in llm outputs.](#)
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

- Abhilasha Ravichander, Jillian Fisher, Taylor Sorensen, Ximing Lu, Maria Antoniak, Bill Yuchen Lin, Nilofar Mireshghallah, Chandra Bhagavatula, and Yejin Choi. 2025. [Information-guided identification of training data imprint in \(proprietary\) large language models](#). In [Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies \(Volume 1: Long Papers\)](#), pages 1962–1978.
- Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. 2019. Regularized evolution for image classifier architecture search. In [Proceedings of the AAAI Conference on Artificial Intelligence](#), volume 33, pages 4780–4789.
- Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. Can ai-generated text be reliably detected? stress testing ai text detectors under various attacks. [Transactions on Machine Learning Research](#).
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. [Social IQa: Commonsense reasoning about social interactions](#). In [Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing \(EMNLP-IJCNLP\)](#), pages 4463–4473.
- Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Oluwasanmi Koyejo. 2025. Position: Model collapse does not mean what you think. [arXiv preprint arXiv:2503.03150](#).
- Mohamed El Amine Seddik, Swei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. 2024. How bad is training on synthetic data? a statistical analysis of language model collapse. In [Conference on Language Modeling \(COLM\)](#).
- Iliia 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.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. [Neurocomputing](#), 568:127063.
- Ananda Theertha Suresh, Andrew Thangaraj, and Aditya Nanda Kishore Khandavally. 2025. Rate of model collapse in recursive training. In [International Conference on Artificial Intelligence and Statistics](#), pages 1396–1404. PMLR.
- Dmytro Valiaiev. 2024. Detection of machine-generated text: Literature survey. [arXiv preprint arXiv:2402.01642](#).
- Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. [Journal of machine learning research](#), 9(11).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In [Advances in Neural Information Processing Systems](#), volume 30.
- Tianyu Wang, Lingyou Pang, Akira Horiguchi, and Carey E Priebe. 2025. Llm web dynamics: Tracing model collapse in a network of llms. [arXiv preprint arXiv:2506.15690](#).
- Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In [International conference on machine learning](#), pages 23965–23998. PMLR.
- Fan Wu, Emily Black, and Varun Chandrasekaran. 2025. Generative monoculture in large language models. In [International Conference on Learning Representations](#).
- Yiqun Zhang, Peng Ye, Xiaocui Yang, Shi Feng, Shufei Zhang, Lei Bai, Wanli Ouyang, and Shuyue Hu. 2025. Nature-inspired population-based evolution of large language models. [arXiv preprint arXiv:2503.01155](#).
- Xuekai Zhu, Daixuan Cheng, Hengli Li, Kaiyan Zhang, Ermo Hua, Xingtai Lv, Ning Ding, Zhouhan Lin, Zilong Zheng, and Bowen Zhou. 2025. How to synthesize text data without model collapse? In [Proceedings of the 42nd International Conference on Machine Learning](#).

A Theoretical Analysis of Gaussian Models

This section details the theoretical foundation and experimental verification of our Gaussian model abstraction, providing intuition for the population dynamics observed in large language models.

A.1 Theoretical Foundation

Let us consider an initial Gaussian distribution $\mathcal{N}^{(0)}(\mu, \sigma^2)$, where μ and σ denote the mean and standard deviation, respectively. Given n independent samples from this distribution, we can obtain unbiased estimators for the population parameters:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{\mu})^2,$$

where $\hat{\mu}$ is the sample mean, and $\hat{\sigma}^2$ is the sample variance. These estimators define a new Gaussian distribution $\mathcal{N}^{(1)}(\hat{\mu}, \hat{\sigma}^2)$.

We extend this to a population-based framework by considering K independent runs of this estimation process, yielding distributions $\{\mathcal{N}_k^{(1)}(\hat{\mu}_k, \hat{\sigma}_k^2)\}_{k=1}^K$. To evaluate the quality of these derived distributions, we employ the squared Wasserstein-2 distance between each derived distribution and the original distribution as a fitness metric. For one-dimensional Gaussian distributions, the squared Wasserstein-2 distance is given as:

$$\mathcal{W}_2^2(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2.$$

This property is specific to one-dimensional Gaussian distributions and forms the basis of our analysis.

For each distribution k , the squared Wasserstein-2 distance is therefore:

$$W_k = \mathcal{W}_2^2(\mathcal{N}^{(0)}(\mu, \sigma^2), \mathcal{N}_k^{(1)}(\hat{\mu}_k, \hat{\sigma}_k^2)) = (\mu - \hat{\mu}_k)^2 + (\sigma - \hat{\sigma}_k)^2.$$

Statistical Properties of the Estimators. From statistical theory, we know that $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ are independent for Gaussian distributions. Specifically:

1. The sample mean $\hat{\mu}_k$ follows a Gaussian distribution:

$$\hat{\mu}_k \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

2. The sample variance $\hat{\sigma}_k^2$ follows a scaled chi-square distribution:

$$(n-1)\frac{\hat{\sigma}_k^2}{\sigma^2} \sim \chi_{n-1}^2.$$

For large n , the sample standard deviation $\hat{\sigma}_k = \sqrt{\hat{\sigma}_k^2}$ can be approximated as asymptotically normal. Using this approximation, the deviation $(\sigma - \hat{\sigma}_k)$ follows approximately:

$$(\sigma - \hat{\sigma}_k)^2 \sim \frac{\sigma^2}{2n}\chi_1^2,$$

where χ_1^2 denotes a chi-square distribution with 1 degree of freedom. Thus, the two components of W_k , $(\mu - \hat{\mu}_k)^2$ and $(\sigma - \hat{\sigma}_k)^2$, are independent and distributed as:

$$(\mu - \hat{\mu}_k)^2 \sim \frac{\sigma^2}{n}\chi_1^2, \quad (\sigma - \hat{\sigma}_k)^2 \sim \frac{\sigma^2}{2n}\chi_1^2.$$

Tail Probability Analysis Using Chernoff Bounds.

We now derive an upper bound on the probability that the squared Wasserstein-2 distance between the true distribution and an estimated distribution exceeds a threshold δ . Since W_k is the sum of two independent scaled chi-square random variables, we can write:

$$W_k = \frac{\sigma^2}{n}X + \frac{\sigma^2}{2n}Y,$$

where $X, Y \sim \chi_1^2$ are independent. For $t > 0$, the moment generating functions (MGFs) of X and Y are:

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX}] \\ &= (1 - 2t)^{-\frac{1}{2}}, \\ M_Y(t) &= \mathbb{E}[e^{tY}] \\ &= (1 - 2t)^{-\frac{1}{2}}, \end{aligned}$$

valid for $t < \frac{1}{2}$.

The MGF of W_k is then:

$$\begin{aligned} M_{W_k}(t) &= \mathbb{E}[e^{tW_k}] \\ &= \left(1 - \frac{2\sigma^2 t}{n}\right)^{-\frac{1}{2}} \left(1 - \frac{\sigma^2 t}{n}\right)^{-\frac{1}{2}}. \end{aligned}$$

Using Chernoff's bound for any $t > 0$, we obtain:

$$\begin{aligned} \mathbb{P}(W_k \geq \delta) &\leq \exp(-t\delta)M_{W_k}(t) \\ &= \exp(-t\delta) \left(1 - \frac{2\sigma^2 t}{n}\right)^{-\frac{1}{2}} \\ &\quad \times \left(1 - \frac{\sigma^2 t}{n}\right)^{-\frac{1}{2}}. \end{aligned}$$

To obtain a simple explicit bound, we choose $t = \frac{n}{4\sigma^2}$, which satisfies $t < \frac{n}{2\sigma^2}$. Substituting this value and using inequalities for logarithms, we get:

$$\mathbb{P}(W_k \geq \delta) \leq 2 \exp\left(-\frac{n\delta}{4\sigma^2} + \frac{1}{2}\right).$$

Population-Based Analysis. We now analyze how a population of K independent estimators improves the probability of obtaining at least one high-fidelity estimate. Let $W_{\min} = \min_k W_k$ denote the minimum squared Wasserstein-2 distance across the K estimators. We are interested in bounding the probability that W_{\min} exceeds a given threshold

δ :

$$\begin{aligned}\mathbb{P}(W_{\min} \geq \delta) &= \mathbb{P}(W_1 \geq \delta, \dots, W_K \geq \delta) \\ &= \prod_{k=1}^K \mathbb{P}(W_k \geq \delta) \\ &= (\mathbb{P}(W_k \geq \delta))^K.\end{aligned}$$

Under the assumption that the estimators are independent, we have:

$$\mathbb{P}(W_{\min} \geq \delta) \leq \left[2 \exp\left(-\frac{n\delta}{4\sigma^2} + \frac{1}{2}\right) \right]^K.$$

Consequently, the probability that at least one of the K estimators achieves squared Wasserstein-2 distance less than δ is:

$$\begin{aligned}\mathbb{P}(W_{\min} < \delta) &= 1 - \mathbb{P}(W_{\min} \geq \delta) \\ &\geq 1 - \left[2 \exp\left(-\frac{n\delta}{4\sigma^2} + \frac{1}{2}\right) \right]^K.\end{aligned}$$

This result shows that the probability of obtaining a high-fidelity estimate increases exponentially with population size K . Even when individual estimators have a non-negligible probability of high error, the collective population is overwhelmingly likely to contain at least one estimator close to the true distribution. This provides a theoretical explanation for why, in this Gaussian abstraction, population-based approaches are more likely to preserve distributional properties and reduce the frequency of collapse events compared to single-estimator dynamics.

A.2 Experimental Verification

Experimental results for the Gaussian model validate the theoretical analysis presented above. In single-model settings, recursive estimation led to progressive drift in both mean and variance estimates, causing the squared Wasserstein-2 distance to grow steadily across generations. In contrast, population-based estimation maintained both parameters near their ground truth values, with the minimum Wasserstein-2 distance across the population remaining bounded even after many generations (Figure 11).

Sensitivity analyses varying sample size n and population size K further confirm that larger populations and moderate sample sizes yield the most robust preservation of distributional properties. These findings support the conclusion that population-based selection fundamentally alters

collapse dynamics by ensuring the persistence of high-fidelity estimators even when individual estimators are noisy.

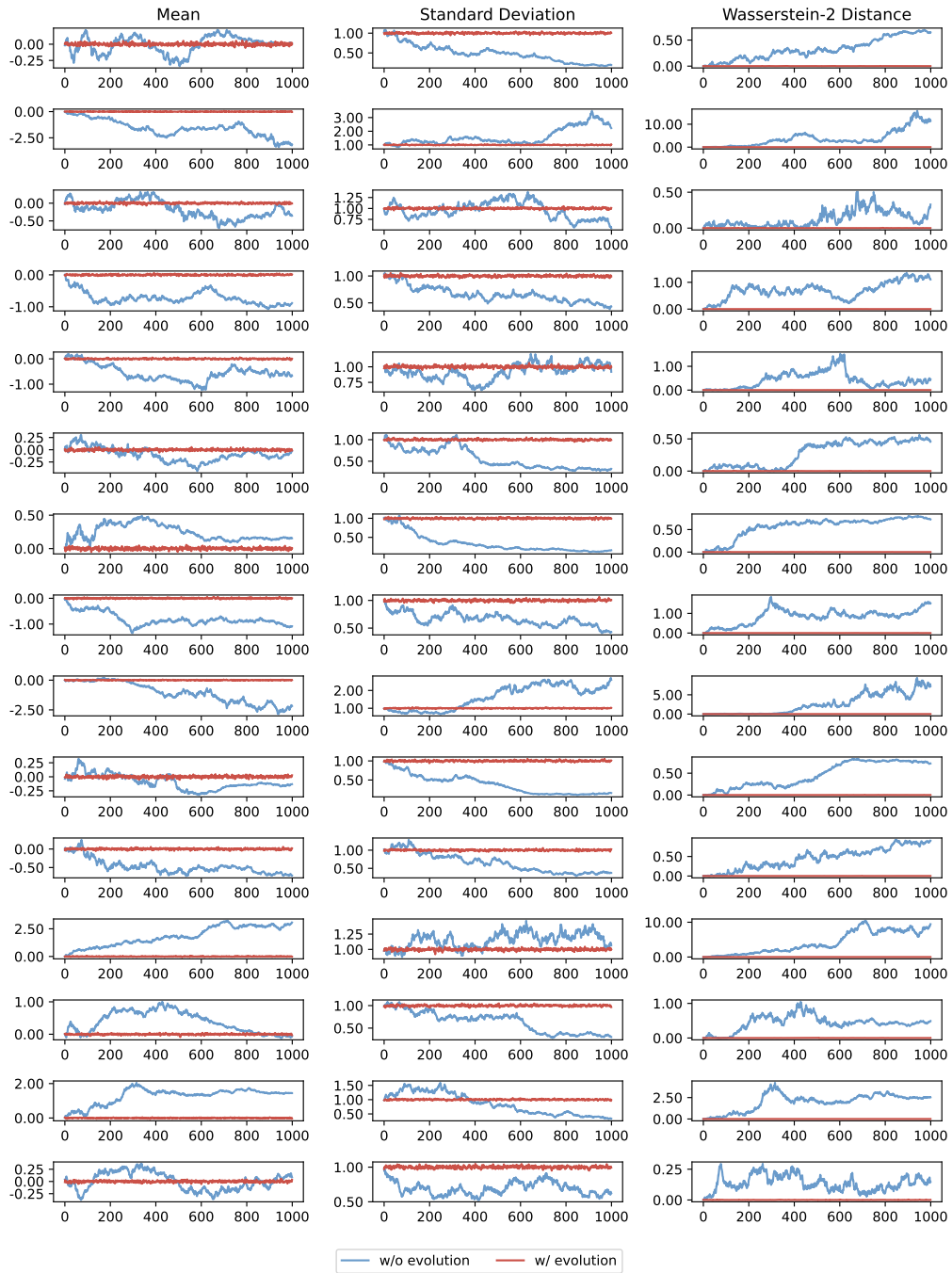


Figure 11: Comparative analysis of evolutionary approach versus baseline across 1000 generations. Each row represents an independent experimental run, with columns displaying mean, standard deviation, and Wasserstein-2 distance of the Gaussian distribution, respectively. In each sub-figure, the x-axis denotes the generation and the y-axis denotes the value. The evolutionary approach (red) demonstrates robust distribution preservation while the baseline (blue) exhibits model collapse where the mean and standard deviation values drift away from the true values 0 and 1.

B Deployed Model Analysis Details

B.1 Behavioral Fingerprinting

We employed the Slop Forensics toolkit (Paech, 2025)¹ to extract behavioral fingerprints and quantify similarities across model families. We then performed our own downstream analyses on these fingerprints, including pairwise Jaccard similarity, distribution-overlap statistics, and temporal influence analyses across model families.

We extracted linguistic profiles for the following 27 deployed models across five families:

- Claude: claude-3-haiku, claude-3.5-haiku-20241022, claude-3.5-sonnet, claude-3.7-sonnet
- GPT: chatgpt-4o-latest, gpt-3.5-turbo-0613, gpt-4-0314, gpt-4.5-preview, gpt-4o-mini
- Gemma: Gemma-3-Starshine-12B, Gemma-3-Glitter-12B, gemma-2-9b-it, gemma-3-12b-it, gemma-3-27b-it, gemma-3-4b-it, gemma-2-Ifable-9B
- Llama: llama-3.1-405b-instruct, llama-3.1-70b-instruct, llama-3.1-8b-instruct, llama-3.2-1b-instruct, llama-3.2-3b-instruct, llama-4-maverick, llama-4-scout
- Mistral: mistral-large-2411, mistral-nemo, mistral-small-24b-instruct-2501, mistral-small-3.1-24b-instruct-2503

For our data collection protocol, we follow the default toolkit settings using 1,000 prompts. All models receive an identical system prompt and a uniform long-form instruction ($\sim 1,000$ words) to reduce prompt-driven lexical variation. During decoding, we set the temperature to 0.7 and max_tokens to 4,096, keeping top_p, top_k, and other parameters at provider defaults. Outputs shorter than 500 characters are discarded. To ensure robustness, n -gram features (words, bigrams, and trigrams) must recur across ≥ 2 distinct prompt IDs. Additionally, English stopwords are filtered, and unigram scores are normalized against the wordfreq English baseline by the toolkit.

Following feature extraction, we computed pairwise similarity matrices using Jaccard similarity over the resultant fingerprints to identify relationship patterns. Statistical significance was assessed through nonparametric bootstrap with 1,000 resampling iterations.

¹<https://github.com/sam-paech/slop-forensics>

B.2 Statistical Metrics

Distributional overlap (OVL). Let $f_w(s)$ and $f_c(s)$ denote the probability density functions of similarity scores for within-family and cross-family pairs, respectively. We quantify their overlap using the overlap coefficient:

$$\text{OVL} = \int_{-\infty}^{+\infty} \min(f_w(s), f_c(s)) ds,$$

which takes values in $[0, 1]$ (0: perfectly separable; 1: identical distributions). In practice, given finite samples $W = \{w_i\}_{i=1}^{n_w}$ and $C = \{c_j\}_{j=1}^{n_c}$, we estimate f_w and f_c via kernel density estimation with a common bandwidth and grid, and approximate the integral by numerical quadrature:

$$\widehat{\text{OVL}} \approx \sum_g \min(\hat{f}_w(s_g), \hat{f}_c(s_g)) \Delta s.$$

As a robustness check, we also compute a histogram-based OVL using shared bins and normalized counts. Uncertainty is quantified by non-parametric bootstrap: resample W and C with replacement, recompute $\widehat{\text{OVL}}$ over B replicates ($B=1000$), and report the 2.5th–97.5th percentiles as a 95% confidence interval.

Cliff’s delta (Effect size). Cliff’s delta δ measures the stochastic dominance between two samples and is defined as

$$\delta = \Pr(X > Y) - \Pr(X < Y),$$

where $X \sim S_c$ (cross-family) and $Y \sim S_w$ (within-family) are independent draws. Equivalently, for samples $C = \{c_j\}_{j=1}^{n_c}$ and $W = \{w_i\}_{i=1}^{n_w}$,

$$\hat{\delta} = \frac{\#\{(j, i) : c_j > w_i\}}{n_c n_w} - \frac{\#\{(j, i) : c_j < w_i\}}{n_c n_w}.$$

Cliff’s δ is directly related to the Mann-Whitney U statistic:

$$\hat{\delta} = \frac{2U}{n_c n_w} - 1,$$

with the sign depending on group ordering. Confidence intervals are obtained via nonparametric bootstrap over pairs (C, W) ; we report the percentile bootstrap 95% interval across B replicates. For interpretability, note that

$$\Pr(Y > X) = \frac{1 - \hat{\delta}}{2},$$

$$\Pr(X > Y) = \frac{1 + \hat{\delta}}{2},$$

so $|\delta|$ closer to 1 indicates less overlap (stronger separation), while values near 0 indicate substantial overlap.

B.3 Temporal Influence Analysis

To quantify the directionality of information flow, we analyze the relationship between behavioral influence and the temporal release order of model pairs. Let M_A and M_B be two models with release dates t_A and t_B (resolved to monthly granularity). We define the time lag $\Delta t_{AB} = t_B - t_A$.

We define the set of stylometric fingerprints $F(M)$ for a model M as the set of n -grams (1, 2, and 3-grams) that are statistically over-represented in its outputs compared to the general population. The directed influence score $I(A \rightarrow B)$ is calculated as the weighted inclusion of A 's fingerprints in B 's output:

$$I(A \rightarrow B) = \frac{\sum_{f \in F(A) \cap F(B)} w(f)}{\sum_{f \in F(A)} w(f)} \quad (1)$$

where $w(f)$ is a weighting function. In our primary analysis (Figure 5), we use a source-weighted directed overlap score, defined as the proportion of source-feature salience recovered in the target model.

To isolate cross-lineage contamination from self-reinforcement, we exclude intra-family pairs (where $\text{Family}(A) = \text{Family}(B)$) and strictly synchronous pairs ($\Delta t_{AB} \approx 0$). The visualization overlays mean influence aggregated in fixed-width month bins, together with approximate 95% confidence intervals. Near-zero-influence points ($I(A \rightarrow B) \leq \epsilon$, with $\epsilon = 10^{-3}$) are excluded from the log-scale scatter plot to ensure stability, though they are included in aggregate statistics. The distinct rise in influence for $\Delta t_{AB} > 0$ relative to the $\Delta t_{AB} < 0$ baseline provides evidence for the pollination hypothesis, where older models influence the training distributions of newer models.

C Evolutionary Framework Implementation

Implementation Specifications. The main text in Section 3 introduces the evolutionary framework conceptually. This subsection specifies the implementation used in our language model experiments through the following concrete operators.

Selection Mechanism. At generation t , to approximate the preference relation induced by human usage described in Section 3, all N models in population $\mathcal{P}^{(t)}$ are evaluated on designated benchmark tasks. Models are ranked by accuracy, and the

top k performers form the parent set $\mathcal{S}^{(t)}$, where we set $k = \lfloor \sqrt{N} \rfloor$ to balance genetic diversity with performance quality. For our experiments with $N = 16$ models, this yields $k = 4$ parent models per generation.

Mutation Operation. Parameter mutation introduces exploration noise scaled by parameter variability. For each parent model $M_i^{(t)} \in \mathcal{S}^{(t)}$ with parameters $\theta_i^{(t)}$, the mutated parameters are computed as $\tilde{\theta}_i^{(t+1)} = \theta_i^{(t)} + \lambda \cdot \text{std}(\theta_i^{(t)}) \cdot \Delta\theta$, where $\Delta\theta \sim \mathcal{U}(-0.5, 0.5)$ represents uniform random noise, $\text{std}(\theta_i^{(t)})$ denotes the standard deviation of parameter values, and $\lambda = 0.01$ controls mutation strength. This adaptive scaling ensures that perturbation magnitude reflects parameter importance while preventing destructive modifications.

Crossover Mechanism. Crossover generates offspring through parameter–data inheritance between parent pairs. From the parent set $\mathcal{S}^{(t)}$, we form all possible ordered pairs

$$\{(M_i^{(t)}, M_j^{(t)}) : i, j \in \{1, \dots, k\}\},$$

yielding $k^2 = N$ offspring to maintain constant population size. For each pair $(M_i^{(t)}, M_j^{(t)})$, the offspring inherits mutated parameters $\tilde{\theta}_i^{(t+1)}$ from the first parent and trains on synthetic dataset $D_j^{(t)}$ generated by the second parent. Synthetic data generation employs nucleus sampling with top- $p = 0.9$ (Holtzman et al., 2020), sampling up to 1 billion tokens per parent model. Training minimizes the standard next-token prediction objective (Radford et al., 2018).

D Experimental Settings

D.1 Model Architecture and Training

We implemented our evolutionary framework using Llama-2-style decoder-only architectures with approximately 100 million parameters per model. Each model consists of 12 transformer layers with 12 attention heads and a hidden dimension of 768, with rotary positional embeddings (RoPE) (Su et al., 2024). We used the AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay 0.01. Gradients were clipped at a norm of 1.0 to improve stability during training on synthetic data. Each model was trained for one epoch over approximately 1 billion

tokens with a context window of 1024 tokens; per-generation learning rates and batch sizes are listed in Table 1.

For each generation, parent models generated synthetic datasets of up to 1 billion tokens. Offspring models were trained from either inherited weights or random initialization as specified by the experimental condition. Learning rate warmup was applied for the first 1,000 steps followed by cosine decay. We did not apply dropout during training, as preliminary experiments indicated it did not materially affect collapse dynamics in this setting.

D.2 Dataset and Data Processing

Our experiments used a corpus of natural-language prompts drawn from public evaluation benchmarks and web-scale datasets. The initial population ($\mathcal{P}^{(0)}$) is pretrained on a subset of the FineWeb-EDU dataset (Lozhkov et al., 2024) (see Appendix G for details). Synthetic completions were generated conditioned on these prompts using Nucleus Sampling with top- $p = 0.9$ and temperature $T = 0.7$, truncating degenerate completions shorter than 10 tokens.

To construct training datasets, we aggregated prompt-completion pairs from parent models within each generation. For experiments isolating data-only effects, we regenerated synthetic data for each generation using the same prompt set but with models reinitialized before training. For experiments incorporating parameter inheritance, synthetic datasets were generated by models inheriting parameters from selected parents, preserving architectural continuity across generations.

D.3 Benchmark Evaluation

We evaluated model populations using four benchmarks capturing complementary aspects of language understanding and reasoning: MMLU (Hendrycks et al., 2021), WinoGrande (Sakaguchi et al., 2021), MathQA (Amini et al., 2019), and SIQA (Sap et al., 2019). Evaluation prompts followed standard multiple-choice formats, with models producing completion tokens that were mapped to answer choices using exact match or heuristic matching where necessary.

For each generation, we computed accuracy scores for all models on each benchmark. Selection pressure was applied by ranking models by accuracy and selecting top performers as parents for the next generation, as described earlier. Baseline collapse trajectories were generated by training single

models recursively on their own synthetic outputs following protocols from Shumailov et al. (2024). We ensured that evaluation prompts were not used during training to avoid contamination.

D.4 Baseline Configuration

Baseline models followed the same architecture and training configuration as evolutionary populations but lacked population-based mechanisms. Each baseline model was trained recursively on its own generated data across the same number of generations as the evolutionary populations. Synthetic datasets were constructed using identical sampling parameters and prompt distributions, ensuring that differences in performance trajectories arose from population dynamics rather than training hyperparameters or data composition.

D.5 Extended Analyses

Computational Resources. All experiments were conducted on a high-performance computing cluster with NVIDIA A100 GPUs. Each evolutionary population required approximately 8 GPUs for concurrent training of models within a generation, while baseline collapse experiments required 1–2 GPUs depending on batch size and context length. Training time per generation ranged from 6 to 10 hours depending on dataset size and benchmark configuration.

Reproducibility. To facilitate reproducibility, we fixed random seeds for all key stochastic components including parameter initialization, data sampling, and optimization. Hyperparameters were selected based on preliminary experiments and held constant across all reported runs, with complete settings summarized in Tables 1 and 2.

Diversity Metrics. We quantified diversity in semantic embedding space, following the protocol summarized in Section 6. For each generation, we sampled 32 prompts from the training distribution and collected one completion from every model in the population. Each completion was encoded into a fixed-dimensional vector using a pre-trained RoBERTa encoder (Liu et al., 2019), and the resulting embeddings were projected to two dimensions with t-SNE (Van der Maaten and Hinton, 2008) for visualization. Diversity is then read off as the spread of model-level embeddings within each prompt cluster, tracked across generations to assess how training condition affects the breadth of model behavior.

Table 1: Hyperparameters for model architecture and training configuration.

Parameter	Value
<i>Model Architecture</i>	
Number of layers	12
Hidden dimension	768
Number of attention heads	12
Attention head dimension	64
Feed-forward network dimension	3072
Vocabulary size	32,000
Maximum sequence length	1024
Positional encoding	RoPE (Rotary Position Embeddings)
Total parameters	≈100 million
<i>Training Configuration (Generation 0)</i>	
Learning rate	1×10^{-3}
Weight decay	0.01
Batch size	128
Training tokens per model	1 billion
<i>Training Configuration (Generations 1–4)</i>	
Learning rate	2×10^{-4}
Weight decay	0.01
Batch size	64
Training tokens per model	1 billion
<i>Optimizer</i>	
Type	AdamW
β_1	0.9
β_2	0.95
ϵ	1×10^{-8}
Learning rate scheduler	Cosine with 1,000-step warmup

Table 2: Hyperparameters for evolutionary setup, evaluation, and hardware.

Parameter	Value
<i>Evolutionary Parameters</i>	
Population size (N)	16
Number of generations	5
Selection size (k)	$\sqrt{N} = 4$
Mutation noise distribution	$\mathcal{U}(-0.5, 0.5)$
Mutation scaling factor (λ)	0.01
Synthetic data generation	Nucleus Sampling (top- $p = 0.9$)
Maximum prompt length	8 sentences
Synthetic tokens per model	1 billion
<i>Evaluation</i>	
MMLU shots	5
WinoGrande shots	5
MathQA shots	0
SIQA shots	0
<i>Random Seeds</i>	
Training seed (evolutionary)	42
Training seeds (baseline)	42, 123, 456
Data sampling seed	100
Evaluation seed	1234
<i>Software and Hardware</i>	
Framework	TorchTune 0.3.1
Python version	3.12
PyTorch version	2.1.0
CUDA version	12.1
GPU	NVIDIA A100 (80GB)
Models per GPU	1

E Qualitative Analysis

Here we present two additional sets of examples of the models' behavior over generations. Without evaluation, the model collapse happens in the late generations: in Table 3, the model in Generation 4 without evolution produces meaningless enumerations; and in Table 4, models in Generation 3 and 4 without evolution generate erroneous references. In contrast, models trained with evolutionary processes are capable of producing coherent and valid outputs. These additional cases further substantiate the efficacy of our population-based evolutionary methods.

Generation	Model Evolution	Model Collapsed
0	\nMany years ago, when the world was crowded in cars and people often had to be late for breakfast, many saw trees everywhere. But for those who spent their lives beyond that hustle-bustle, the roadways would have been uninhabited with ever more bushy trees. As we	But the pleasureous misty saplings scoured the landscape and turned it into a glittering sound.\nAccording to historian Cox Gartner, “The wild flowers were nurtured by male mothers creating a ‘deep’ logistic pattern. Yet these moths are inseparable from each
1	But the more valuable and vivid a landscape was, the more beautiful it would be!\nThis is no ordinary scene at first glance: one shows what came next in this period. Whereas the panoramic views appear differently depending upon whether they are standalone or multifaceted, there are still	But the more valuable and vividly detailed textile style is what it meant for Native Americans: the tree was highly valued by peasants who came out in search of gold or silver; while ancient Egyptian fortifications would have been built on unincorporated property such as windowsills or houses;
2	It is said that this time has been referred to as ‘the day’when the sun shone through them again, and then afterward they turned blue, turning blackish red.\nHistorical Fiction: How To Write A Story About People Who Are Good At Their Money? An Introduction To The Story	But these plants are still alive today: they remain sober and beautifully done—so why not do you? They stand forever! Their beauty is often reminiscent of what makes them special; each flower will bloom differently depending upon whether its color varies between seasons or seasonal cycles. And then there
3	It seems certain that this is what happened when the earthquakes struck the city in March 2016 (the first major quake). At least five people died during these devastating events; most likely because they had been injured while trying to find shelter underneath their homes and livelihoods.	But these plants did not last long enough because they had been eaten raw or cooked; instead, some seeds became available through winter until late April when seedlings died out before reaching maturity again! These blooming vegetables would soon become part of our diet today: bananas and peppers are
4	It seems likely that this time around, if you want your own tree plantings, then why don’t we help them? Why shouldn’t someone else just bring us some old ones! A few weeks ago I saw one of my favorite trees growing near me: the Tallgrass (Hydrangea	But today we see more than just a few varieties available online:\n <u>Southern</u> <u>umes – Asteraceae (Lepidoptera): Gingerbreadth & Flower Plant - Lily plantain , Cucurbita / Leafhopper . These include the European honeysuckle,</u>
Prompt.	With the arrival of spring, the world seemed to awaken from a long slumber, embracing a vibrant rebirth. The once-barren fields were now painted with a tapestry of wildflowers, their colors a celebration of life. Trees, once stripped of their leaves, stood tall with fresh greenery, their branches heavy with budding promise.	

Table 3: Model outputs across generations. Invalid outputs are underlined. Models without evolution start producing meaningless enumerations, while the outputs of models trained with evolution remain stable.

Generation	Model Evolution	Model Collapsed
0	\nMany years ago, Mohammed Ali was working as a research engineer at the University of North Carolina, North Carolina where he first encountered mathematics. In his spare time, he spent much of his life trying to figure out how to use algebra to solve problems. He later learned about his father’s plans for building an	\nMany years ago scientists began to work out how mathematics works and how it develops over time as well as where and how it evolved for the better. The result was quite simple: the human brain’s entire structure would be built into mathematical terms. In fact, the same way we remember words today would
1	\nThe language itself takes up some of its most important features; it is often used as part of the literary canon (the first few lines describe how the author expresses his thoughts and feelings) but also serves as a model for many other matters within mathematics: how to represent information using symbols; and how to define quantities	But mathematicians are also often unable to find new ways of solving problems without having a solid foundation for reasoning. However, at first glance mathematics may be referred as “the model of mechanisms” because they must go beyond simple calculations. They would have been uninformed by many different fields; and their understanding of how
2	But mathematics doesn’t simply lead us into new lands; it also brings about real life experiences—a sense of wonder and discovery within ourselves. In this vein, mathematical thinking can be seen as a means by which mankind might live up to its potential.\nThis course was created as part of a series called	But how do we define such forms? How can they be applied across multiple dimensions without having any special significance beyond those found at first glance? What sets them apart from others depends upon what sort of structure they must go before being revealed by their own means (and therefore not yet understood), but also whether there exists some kind
Prompt.	Mathematics is a profound tapestry of elegance and precision, a language that speaks to the very fabric of the universe. Its beauty lies in the simplicity of its fundamental principles and the complexity of its intricate patterns. From the harmonious symmetry of geometric shapes to the rhythmic flow of equations, math reveals an underlying order in the chaos of the world. It is a realm where logic and creativity intertwine, where abstract concepts take on tangible form through numbers and symbols.	

Table 4: Model outputs across generations. Invalid outputs are underlined. Models without evolution start producing erroneous references, while the outputs of models trained with evolution remain stable.

Generation	Model Evolution	Model Collapsed
3	But these are not just tools used for creating new ideas; they also serve as tools for developing strategies based upon real experiences rather than mere facts alone! This article explores how mathematics influences our perception of reality: what makes it possible? How does it apply to everyday life? What kinds of solutions might	Mathematically speaking, everyday objects can become concrete representations using their own unique properties—such as those found within them (see Figure 1). These elements include mathematical formulas such as logarithms or fractions; numerical operations involving some sort of fractional factoring process involving many different kinds of variables including variables like volume
4	But these same principles remain unchanged throughout history: mathematics has always existed outside our own homes; today we live within ourselves—and nowhere else! At least since then there are many ways forward toward understanding each other's lives - such as learning about history or politics, having fun together while watching television programs like \"	But why? Because everyday objects contain many elements which make sense if you want your own ideas about how things work together without being too complicated! This article will explain what makes up a good deal more difficult than just studying physics at home... <u>Please join StudyMode to read the full document</u>
Prompt.	Mathematics is a profound tapestry of elegance and precision, a language that speaks to the very fabric of the universe. Its beauty lies in the simplicity of its fundamental principles and the complexity of its intricate patterns. From the harmonious symmetry of geometric shapes to the rhythmic flow of equations, math reveals an underlying order in the chaos of the world. It is a realm where logic and creativity intertwine, where abstract concepts take on tangible form through numbers and symbols.	

Table 5: Model outputs across generations (continued). Invalid outputs are underlined.

F Mechanism Analysis: Parameter Inheritance

To understand the drivers of performance, we examine the role of parameter inheritance.

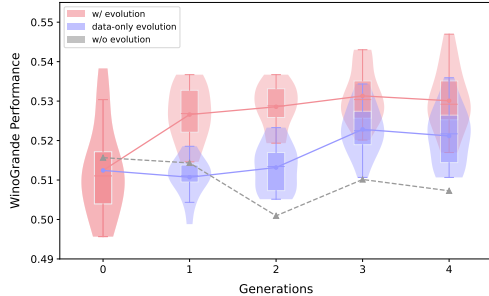


Figure 12: **Contribution of parameter inheritance.** Performance trajectories on WinoGrande for full evolution (parameter + data inheritance) versus data-only evolution. Full inheritance (pink) consistently outperforms data-only transfer (blue) and the single-model baseline (gray), validating the benefit of cumulative parameter optimization.

Parameter Inheritance Contributions. We compared models trained with full evolution (parameter + data inheritance) versus data-only evolution (random initialization with evolved data). Figure 12 shows performance trajectories using WinoGrande selection. First, full evolution (pink) consistently outperforms data-only (blue) across all generations. Second, both evolutionary approaches surpass single-model baselines (gray), confirming evolved data itself carries significant value. The widening performance gap suggests parameter inheritance enables cumulative optimization that random initialization cannot achieve.

G Artifact Details

This section details the artifacts (datasets and models) used in our work, confirming that all usage complies with their respective licenses and intended purposes.

Datasets.

- **MMLU** (Hendrycks et al., 2021): This dataset is released under the **MIT License** and intended for measuring massive multitask language understanding. Our usage for model evaluation aligns with these terms.
- **WinoGrande** (Sakaguchi et al., 2021): Designed for common-sense reasoning, this

dataset is licensed under the **Creative Commons Attribution (CC-BY)** license (code under Apache 2.0). We use it strictly for evaluating model reasoning capabilities in accordance with its intended use.

- **MathQA** (Amini et al., 2019): This dataset for math word problem solving is released under the **Apache 2.0 License** (via Hugging Face). We use it for evaluating mathematical reasoning models, consistent with its intended research purpose.
- **SIQA** (Sap et al., 2019): A benchmark for social commonsense intelligence. It is released under the **Creative Commons Attribution 4.0 (CC BY 4.0)** license. We use it solely for academic evaluation purposes.
- **FineWeb-EDU** (Lozhkov et al., 2024): The initial dataset, comprising 100 billion tokens, was sourced from the FineWeb-EDU dataset, a large-scale web corpus filtered for educational content quality. It is released under the **Open Data Commons Attribution License (ODC-By)**. Our use of this data for pre-training large language models is the primary intended use case of this dataset.

Models and Tools.

- **RoBERTa** (Liu et al., 2019): We use the pre-trained RoBERTa model to generate text embeddings for semantic diversity analysis. The model is released under the **MIT License**, which permits our research usage.
- **Slop Forensics Toolkit** (Paech, 2025): We use this open-source toolkit for generating and analyzing lexical patterns. It is available under a permissive research license (MIT/Apache), and our usage for behavioral fingerprinting is consistent with its intended software analysis purpose.

Privacy and Safety. We utilized standard public datasets that are widely used in the community and considered safe. These datasets do not contain personally identifying information or offensive content beyond what is typical for web-scale corpora used in standard pre-training and evaluation.

H Potential Risks

While our work primarily involves theoretical analysis and simulations, we acknowledge potential

risks associated with evolutionary optimization. If selection criteria are not carefully aligned with human values, evolutionary pressure could theoretically amplify undesirable behaviors. However, our experiments are conducted in a controlled environment using standard benchmarks, and we do not release any high-capability models that could be directly deployed for malicious purposes. The primary goal of this research is to improve ecosystem resilience and mitigate the risks of model collapse.