Creativity/Model Collapse

1. The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text
2. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data

Téa Wright, Donghyun Lee

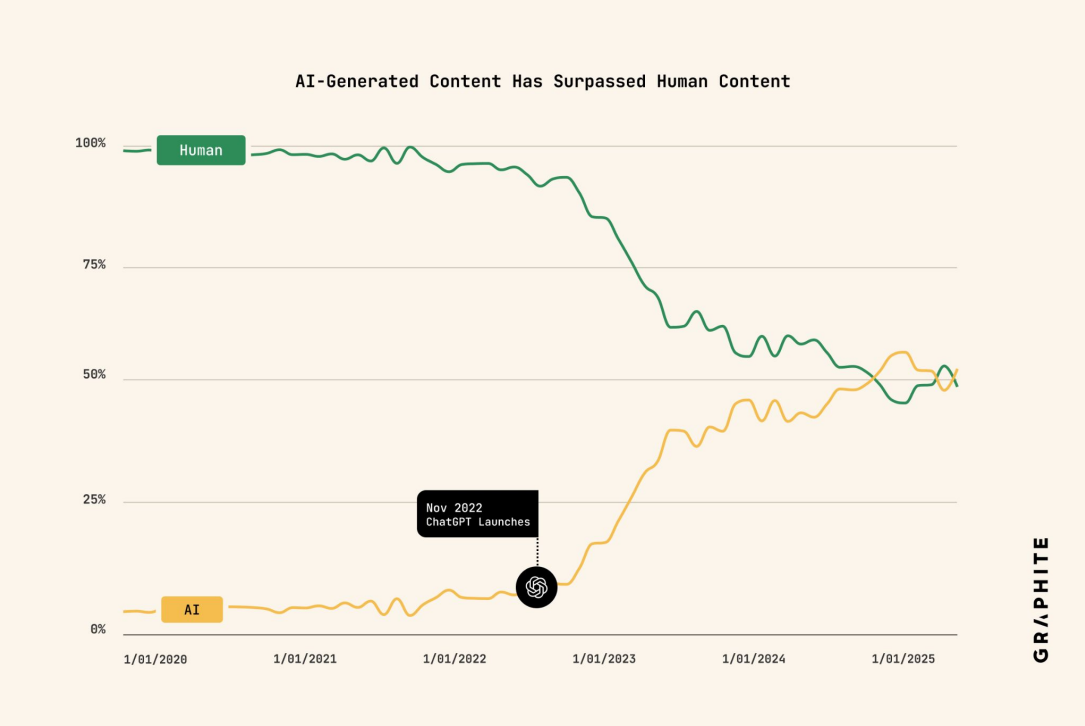
Oct 28, 2025

Table of contents

Dead Internet Theory

Dead Internet Theory

Dead Internet Theory



The “theory” might be not real, but AI contents are indeed multiplying

What does this mean for LLMs?

Article | Open access | Published: 24 July 2024

AI models collapse when trained on recursively generated data

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson & Yarin Gal

Nature **631**, 755–759 (2024) | Cite this article

571k Accesses | **354** Citations | **3459** Altmetric | Metrics

An Author Correction to this article was published on 21 March 2025

This article has been updated

Abstract

Stable diffusion revolutionized image creation from descriptive text. GPT-2 (ref. [1]), GPT-3(.5) (ref. [2]) and GPT-4 (ref. [3]) demonstrated high performance across a variety of language tasks. ChatGPT introduced such language models to the public. It is now clear that generative artificial intelligence (AI) such as large language models (LLMs) is here to stay and will substantially change the ecosystem of online text and images. Here we consider what may

Download PDF

Associated content

Nature Outlook

Robotics and artificial intelligence

AI produces gibberish when trained on too much AI-generated data

Emily Wenger

Nature | **News & Views** | 24 Jul 2024

Sections | Figures | References

Abstract

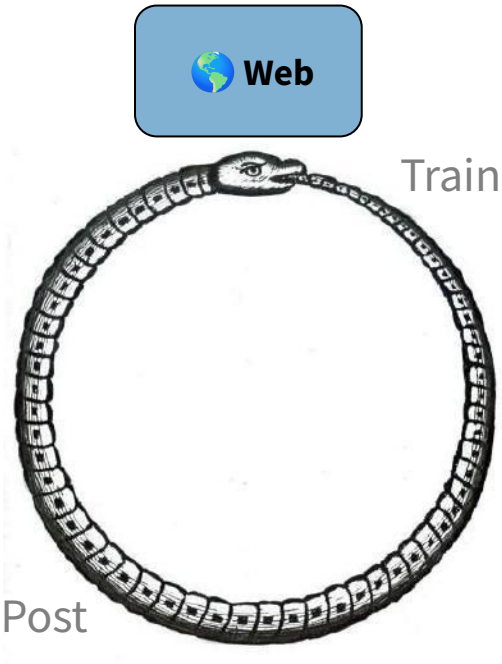
Main

What is model collapse?

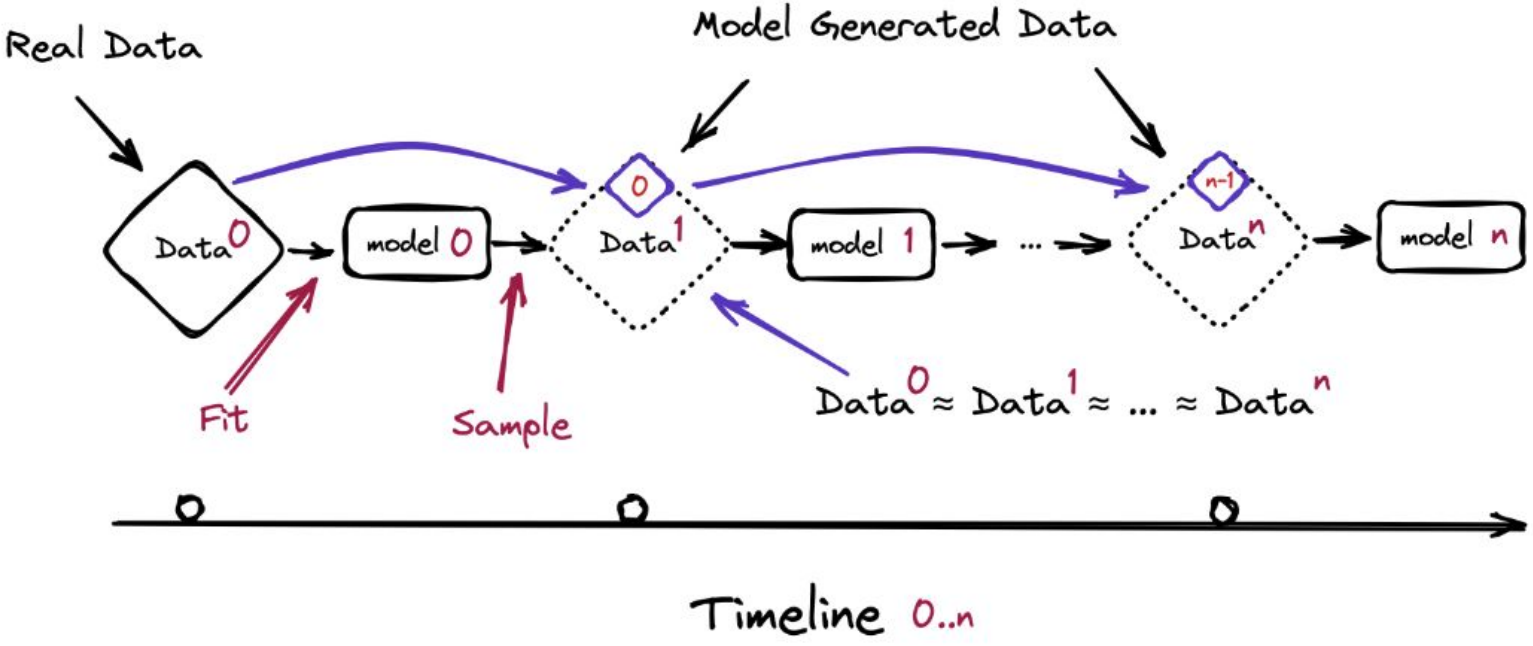
Theoretical intuition

Model collapse in language models

Discussion



What does this mean for LLMs?



Source: *THE CURSE OF RECURSION:TRAINING ON GENERATED DATA MAKES MODELS FORGET*

What does this mean for LLMs?

Example of text outputs of an OPT-125m model affected by *Model Collapse*– models degrade over generations, where each new generation is trained on data produced by the previous generation.

Input: some started before 1360 — was typically accomplished by a master mason and a small team of itinerant masons, supplemented by local parish labourers, according to Poyntz Wright. But other authors reject this model, suggesting instead that leading architects designed the parish church towers based on early examples of Perpendicular

Outputs:

Gen 0: Revival architecture such as St. John's Cathedral in London. The earliest surviving example of Perpendicular Revival architecture is found in the 18th @-@ century Church of Our Lady of Guernsey, which dates from the late 19th century. There are two types of perpendicular churches : those

Gen 1: architecture such as St. Peter's Basilica in Rome or St. Peter's Basilica in Buenos Aires. There is no evidence that any of these buildings were built during the reign of Pope Innocent III, but it is possible that they may have been built during the reign of his successor, Pope Innocent

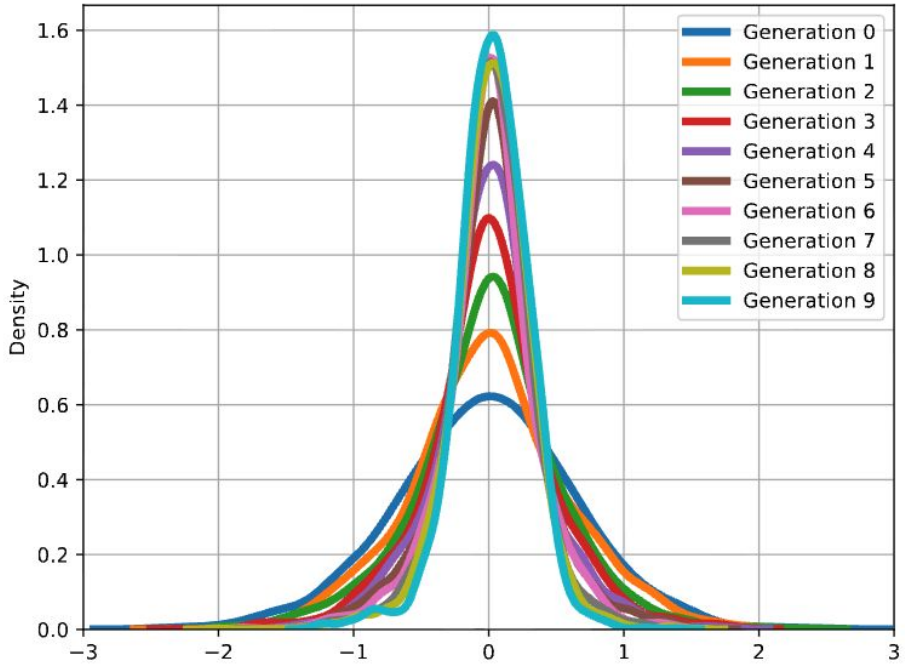
Gen 7: architecture in England. In an interview with The New York Times, Wright said : " I don 't think there is anything wrong with me being able to do what I want to do. It just doesn 't work for me. " He added : " I don 't know if you can call it funny,

Gen 9: architecture. In addition to being home to some of the world's largest populations of black @-@ tailed jackrabbits, white @-@ tailed jackrabbits, blue @-@ tailed jackrabbits, red @-@ tailed jackrabbits, yellow @-

What does this mean for LLMs?

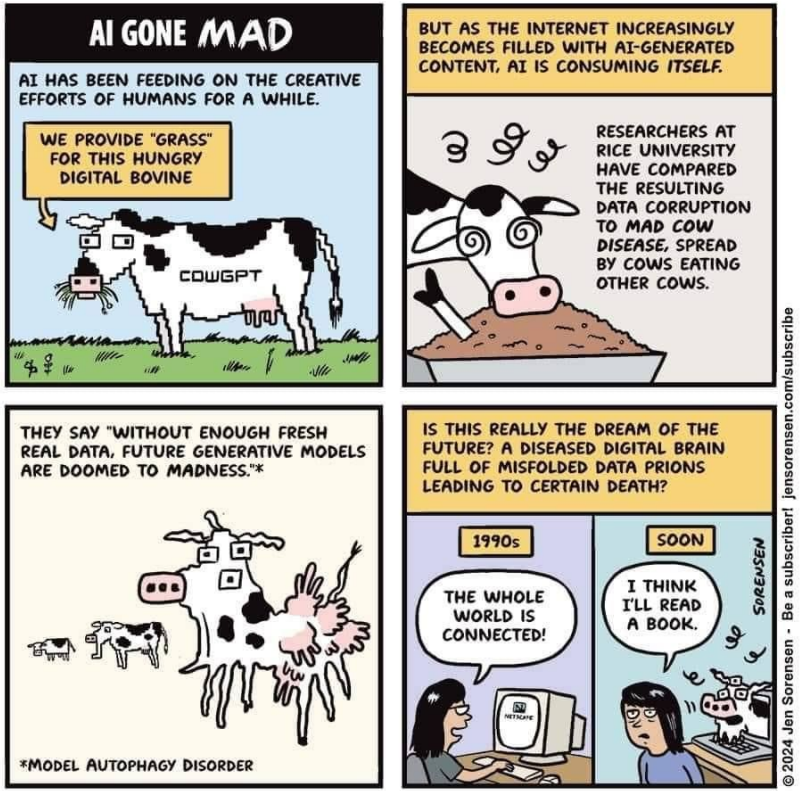
Due to finite sampling, common/rare events are overestimated/underestimated.

Tails wash away over iterations.



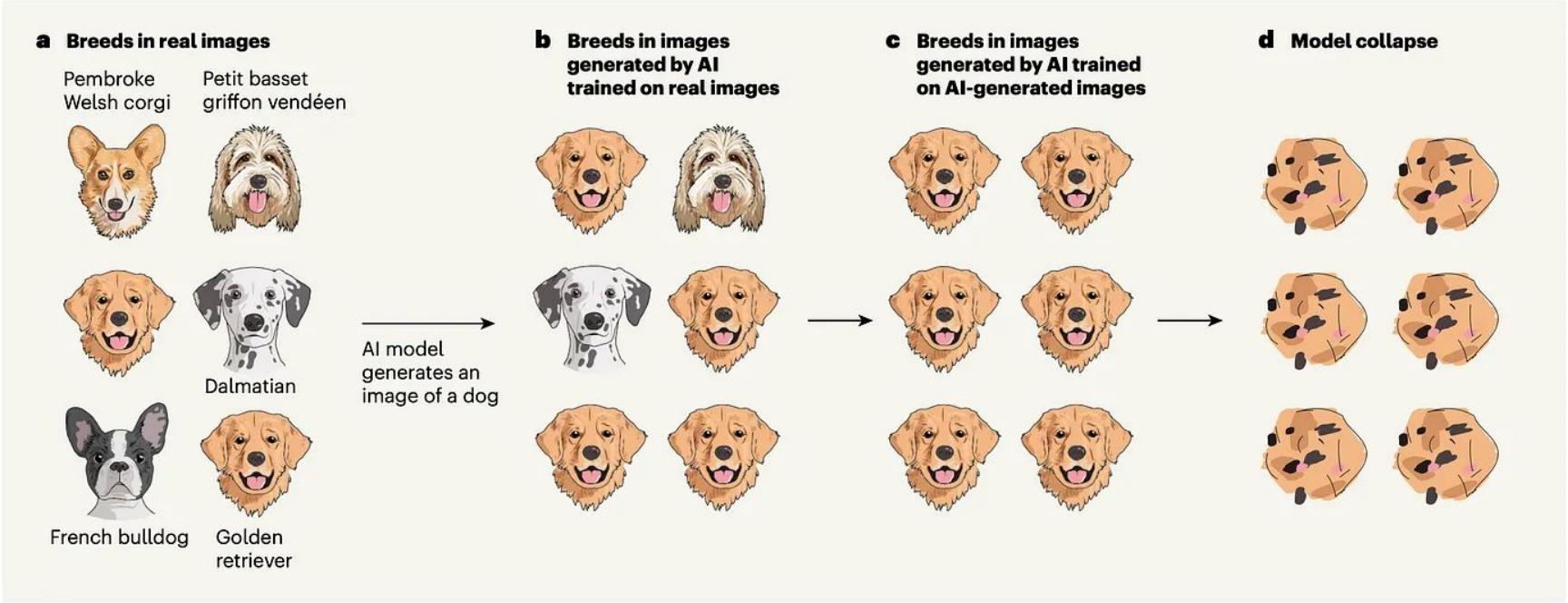
Source: *THE CURSE OF RECURSION:TRAINING ON GENERATED DATA MAKES MODELS FORGET*

What does this mean for LLMs?



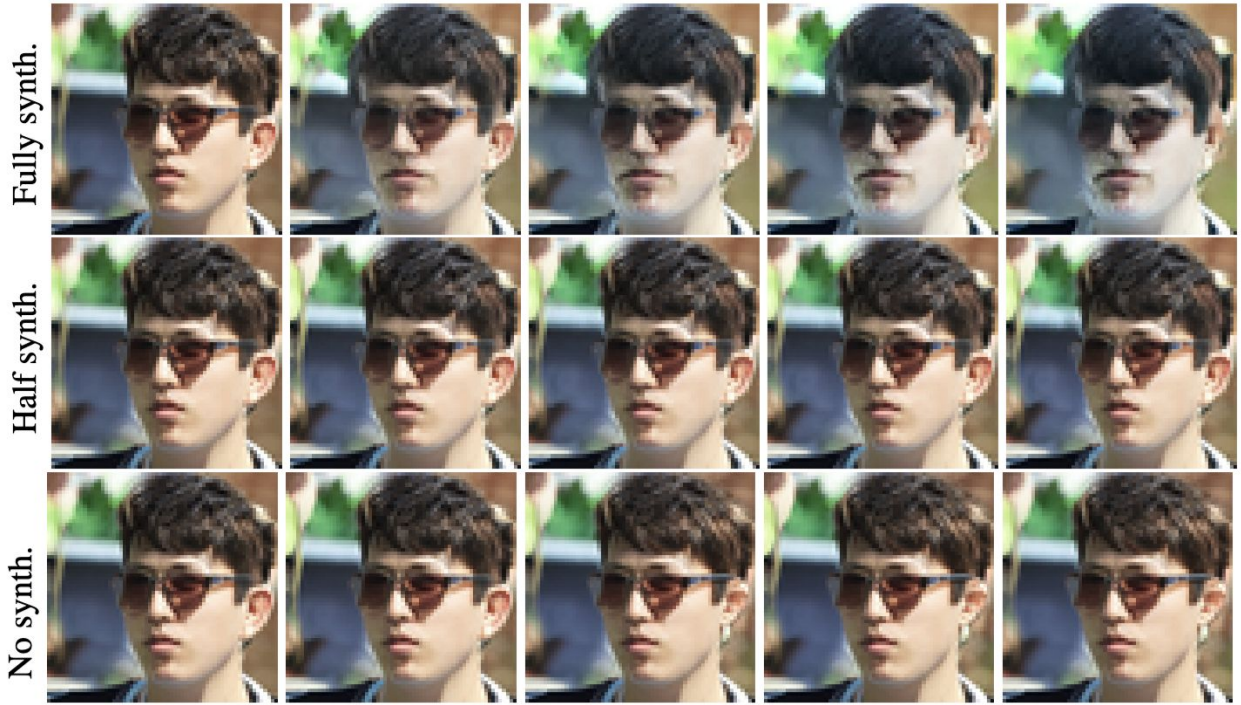
Source: *AI Gone MAD (Jen Sorensen)*

Model Collapse in Vision



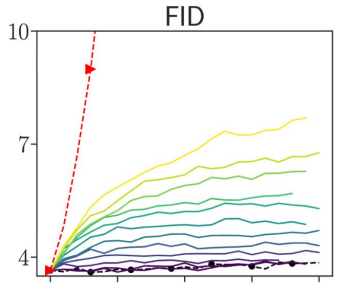
Wenger et al., 2024

Model Collapse in Vision

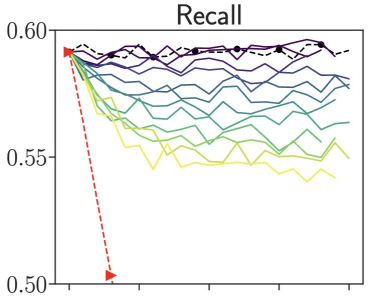


(a) 0 retrain. (b) 5 retrain. (c) 10 retrain. (d) 15 retrain. (e) 20 retrain.

Image quality & diversity (↓)



Diversity (↑)



Bertrand et al., 2023

How do we measure this in text?

Prompt: The title of the paper is: CASIA's System for IWSLT 2020 Open Domain Translation.The abstract of the paper is: This paper describes the CASIA's system for the IWSLT 2020 open domain translation task.
Human: This year we participate in both Chinese→Japanese and Japanese→Chinese translation tasks. Our system is neural machine translation system based on Transformer model. We augment the training data with knowledge distillation and back translation to improve the translation performance. Domain data classification and weighted domain model ensemble are introduced to generate the final translation result. We compare and analyze the performance on development data with different model settings and different data processing techniques..
Iteration 1: We participated in all language directions, namely Chinese ↔ English, Japanese↔ English and German ↔ English. Our system mainly focuses on data selection, data preprocessing and data filtering to improve the quality of the generated translations..
Iteration 3: We participated in all three language pairs, namely Chinese↔English, Japanese→Chinese and German↔French. Our systems are based on Transformer architecture with a simple but effective method to improve the translation quality by reusing monolingual data from other languages.. The source code can be obtained from https://github.com/csebuetnlp/acl2021/opendomaintranslate..The blind test set will be released upon the acceptance of the report..The full results can also be found at https://competitions.codalab.org/competitions/26153.. The leaderboard can be found at https://competitions.codalab.org/page/view/11650..The dataset used in this work has been released under CC BY 4.0 license..The experimental results show that our proposed methods significantly outperform the baseline models..
Iteration 5: We participated in all three language pairs, namely Chinese↔English (Chinese→German), Japanese→Czech and German↔French.. Our approach relies on a Transformer model trained with in-domain data only; it also incorporates a novel method to improve the quality by leveraging parallel corpora from other domains such as TED Talks. The code used in this work has been released under CC BY 4.0 license at https://github.com/czyssrs/ibm_nlp2020/. For further information about our systems please see http://www.csse.unige.edu/research/demo/.. The full text of this report can be obtained via https://competitions.codalab.org/page/view/11153.. The codes for our experiments are available online at https://codebenchmark.io/, too..... Finally, we provide some insights into how our systems perform when translating into low resource conditions..... The source code for our systems is publicly available through https://github.com/czeneva/acl2022tasks2/.

How can we measure changes in quality and ★ diversity for text?

Table of contents

The Big Ideas

- How can we quantify linguistic diversity?
- Once we've decided on some metrics…
  - When we train models on their own outputs, how does it affect linguistic diversity?

Tasks

Input Article: Yahoo's patents suggest users could weigh the type of ads against the sizes of discount before purchase. It says in two US patent applications that ads for digital book readers have been "less than optimal" to date. [...] "Greater levels of advertising, which may be more valuable to an advertiser and potentially more distracting to an e-book reader, may warrant higher discounts," it states. [...] It adds that the more willing the customer is to see the ads, the greater the potential discount. [...] At present, several Amazon and Kobo e-book readers offer full-screen adverts when the device is switched off and show smaller ads on their menu screens. [...] Yahoo does not currently provide ads to these devices, and a move into the area could boost its shrinking revenues.

Summary: Yahoo has signalled it is investigating e-book adverts as a way to stimulate its earnings.

News Summarization

ACL Anthology

The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, Chloé Clavel

PDF | Cite | Search | Video | Fix data

Abstract

This study investigates the consequences of training language models on synthetic data generated by their predecessors, an increasingly prevalent practice given the prominence of powerful generative models. Diverging from the usual emphasis on performance metrics, we focus on the impact of this training methodology on linguistic diversity, especially when conducted recursively over time. To assess this, we adapt and develop a set of novel metrics targeting lexical, syntactic, and semantic diversity, applying them in recursive finetuning experiments across various natural language generation tasks in English. Our findings reveal a consistent decrease in the diversity of the model outputs through successive iterations, especially remarkable for tasks demanding high levels of creativity. This trend underscores the potential risks of training language models on synthetic text, particularly concerning the preservation of linguistic richness. Our study highlights the need for careful consideration of the long-term effects of such training approaches on the linguistic capabilities of language models.

Scientific Abstract Generation

Prompt: The Mage, the Warrior, and the Priest

Story: A light breeze swept the ground, and carried with it still the distant scents of dust and time-worn stone. The Warrior led the way, heaving her mass of armour and muscle over the uneven terrain. She soon crested the last of the low embankments, which still bore the unmistakable fingerprints of haste and fear. She lifted herself up onto the top the rise, and looked out at the scene before her. [...]

Story Generation

*high entropy

Metric(s) 1: Lexical Diversity

- What is the variety of words in the text?
- ***Hypothesis:*** a collapsed LM will have a smaller vocabulary.
- How they quantify this:
  - **Type-token ratio (TTR)** = number of unique words (types)/number of running words (tokens)
    - Truncate texts to fixed length
  - **Distinct-n** = number of unique n-grams
    - n=2,3
  - **1 - Self-BLEU** = BLEU score of one sentence against all others (averaged across all sentences)
    - Report inverse so higher is better/more diverse

Metric 2: Semantic Diversity

- Similar words can have different semantics and different words can have similar semantics
- Get embeddings from Sentence-BERT

Div_sem = average pairwise cosine-distance of all embedding vectors

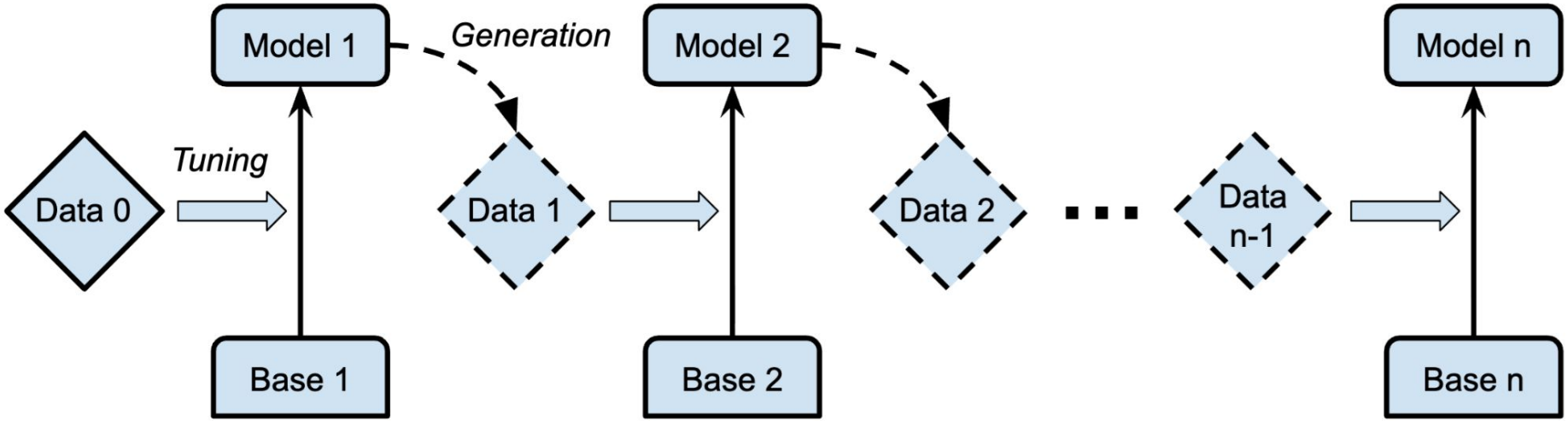
Metric 3: Syntactic Diversity

- Neural parser → dependency trees → graph representations
  - Nodes = words
  - Edges = dependency relations
- Graphs → vector space (using Weisfeiler-Lehman graph kernel)
  - graphs that are structurally alike are closer to each other in the embedding space

Div_syn = average pairwise cosine-distance of all dependency vectors

Training Setup

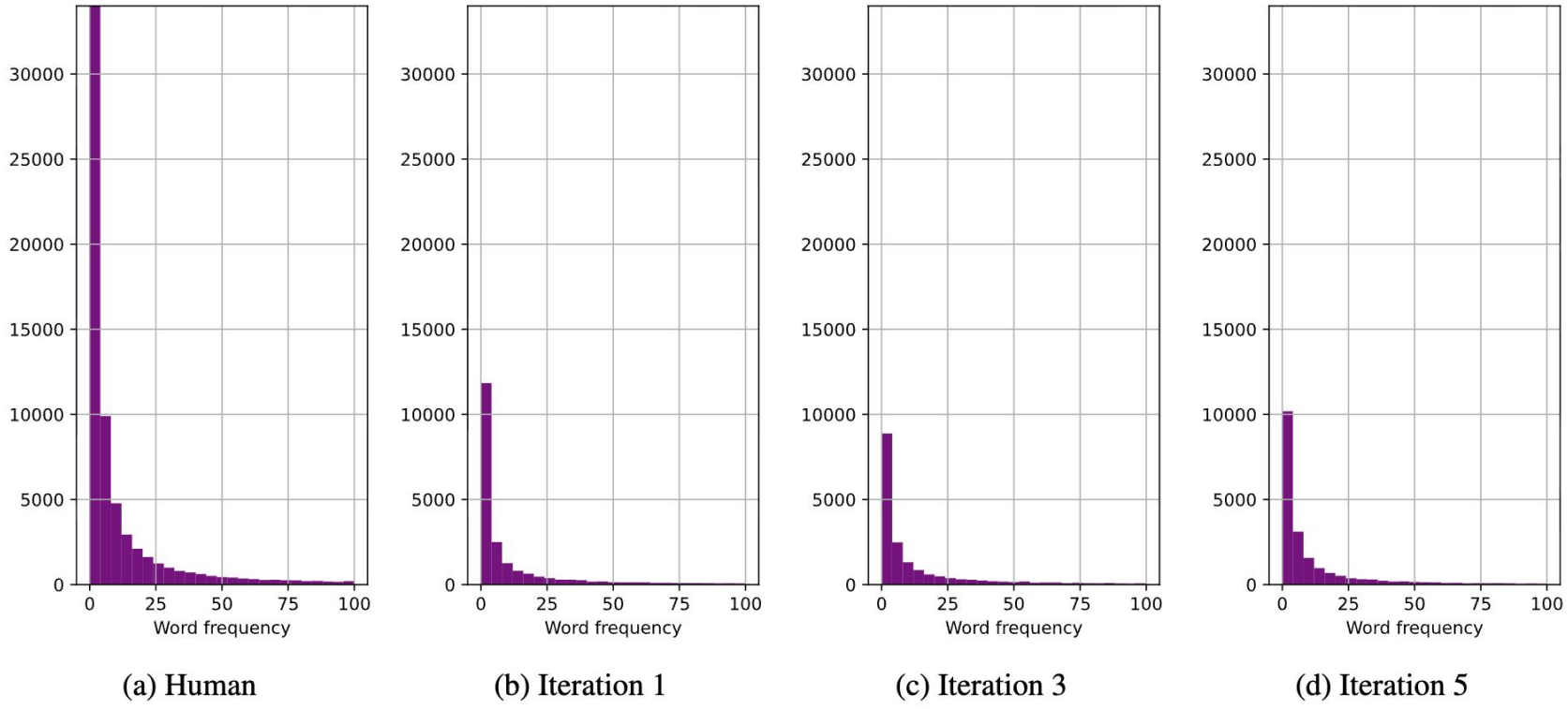
*In this paper, n= 6



Base model:
OPT

(Zhang et al., 2022)

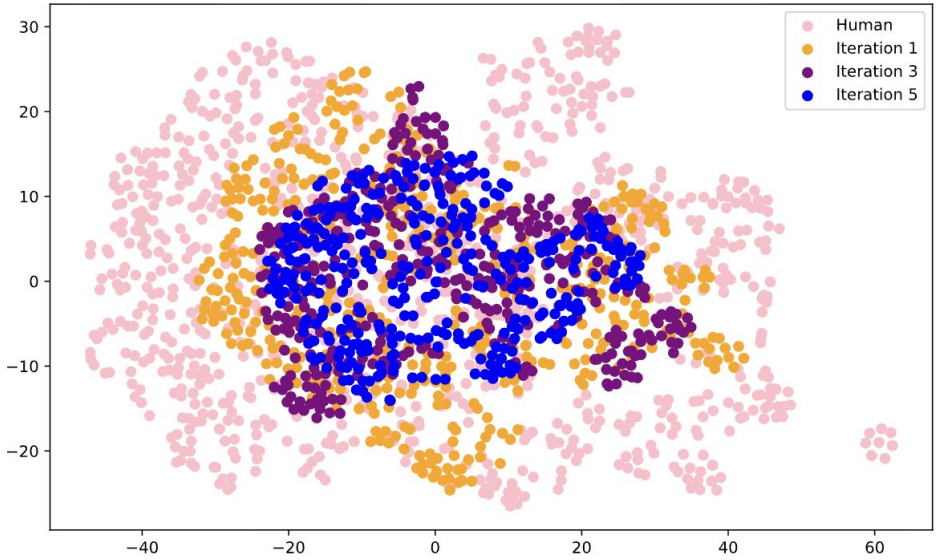
Results



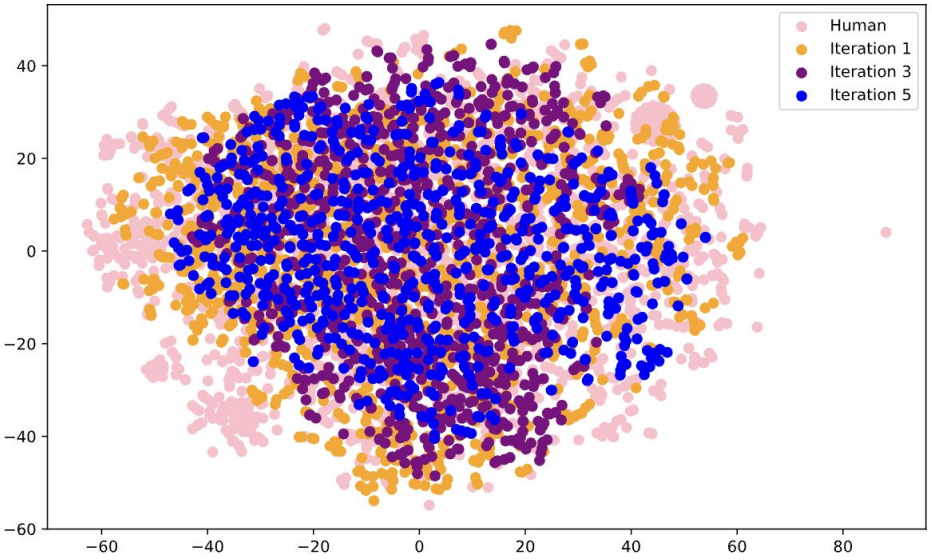
(a) Human (b) Iteration 1 (c) Iteration 3 (d) Iteration 5

Syntactic vs Semantic Embeddings over iterations

Div_syn



Div_sem

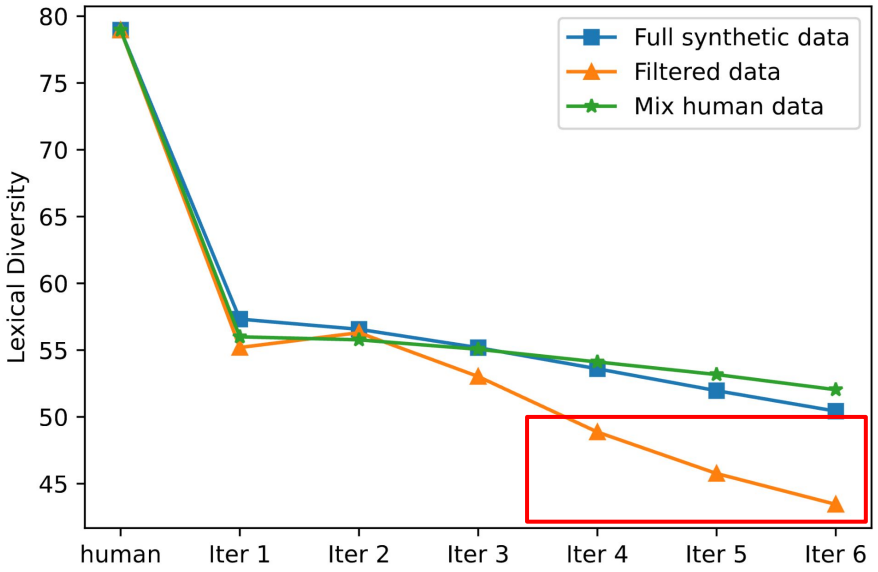


Ok, but 100% synthetic data isn't realistic…

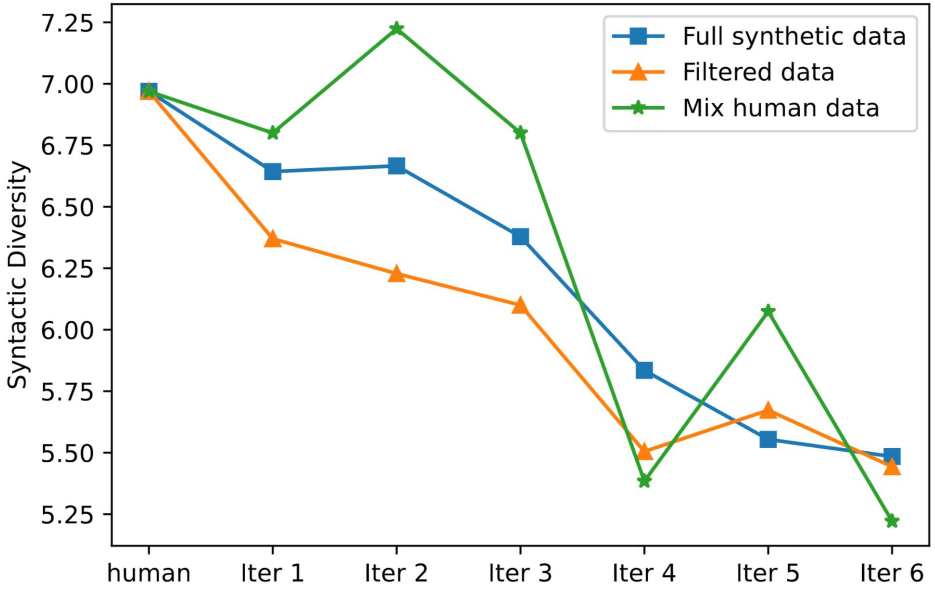
How can we change the setup to make it more realistic? (still not very realistic)

1. **Filter synthetic data**
    a. Discard the worst 20% of the synthetic data with a linguistic acceptability filter (RoBERTa trained on COLA)
2. **Mix in fresh human data**
    a. 60% human (split into 6 10% folds), 40% synthetic train set
    b. For each training iteration, mix in one of the 6 folds

Does it help?



(a) Lexical diversity.



(b) Syntactic diversity.

filtering makes it worse!

Takeaways

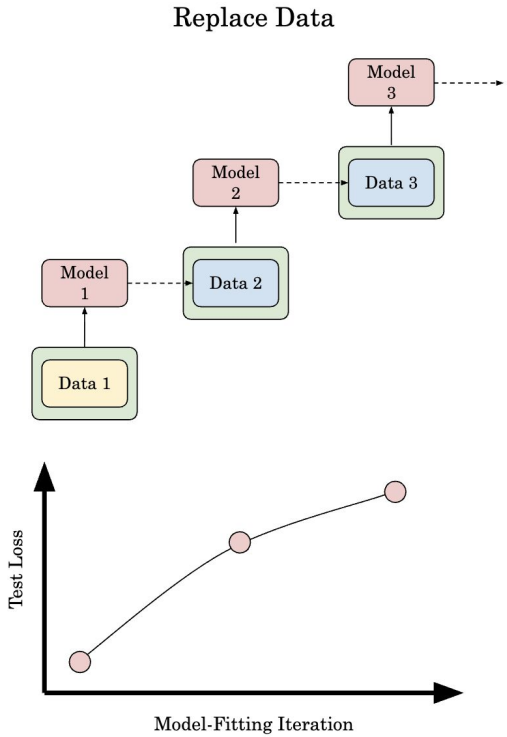
- In 100% synthetic settings, lexical and syntactic diversity decline
- Semantic diversity remains largely unaffected
- In other words, the big ideas are not affected, but the details of the language such as vocabulary and structure become less diverse
- Filtering and mixing in fresh human data does *not* stop the quality decline

Table of contents

Motivation

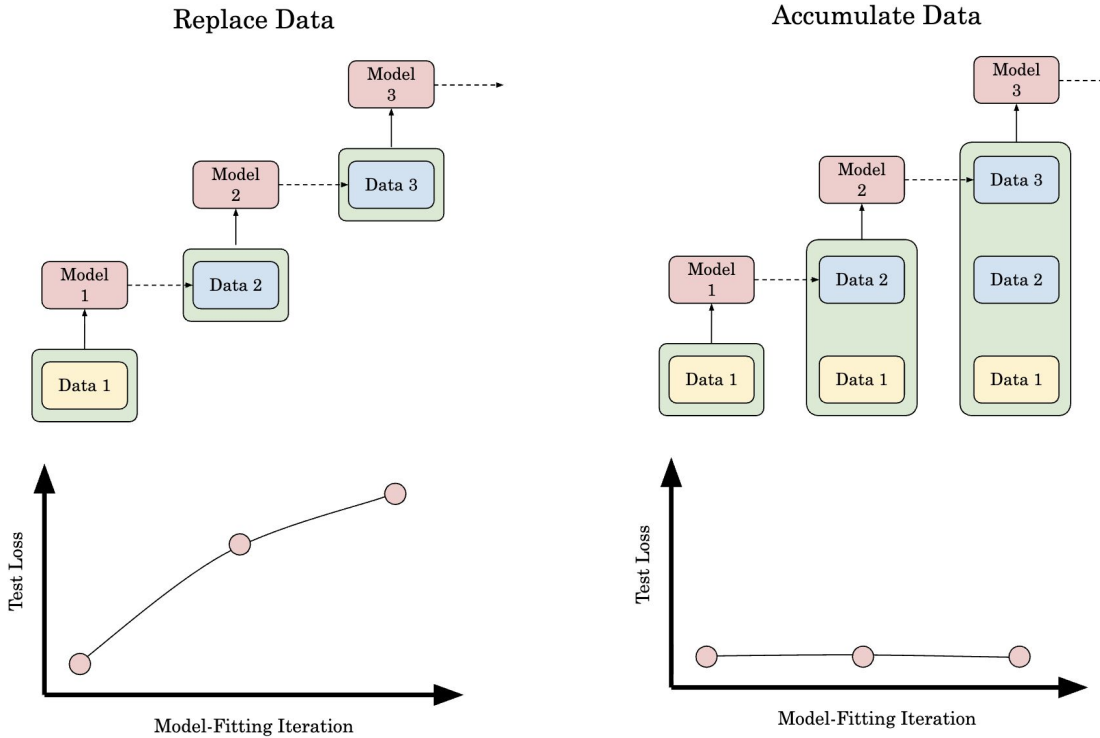
- Is the model collapse fear real?
- Prior works largely assume the synthetic data *replaces* human data
- Maybe they are overestimating the fear with a wrong assumption
- Realistic assumption: *accumulating* synthetic data + human data

Motivation



Prior works

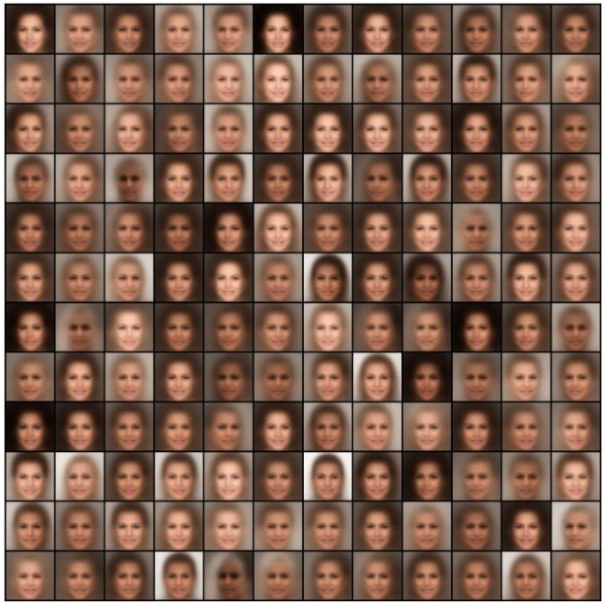
Motivation



Prior works

This work

Brief Glance at Results (VAE on image)



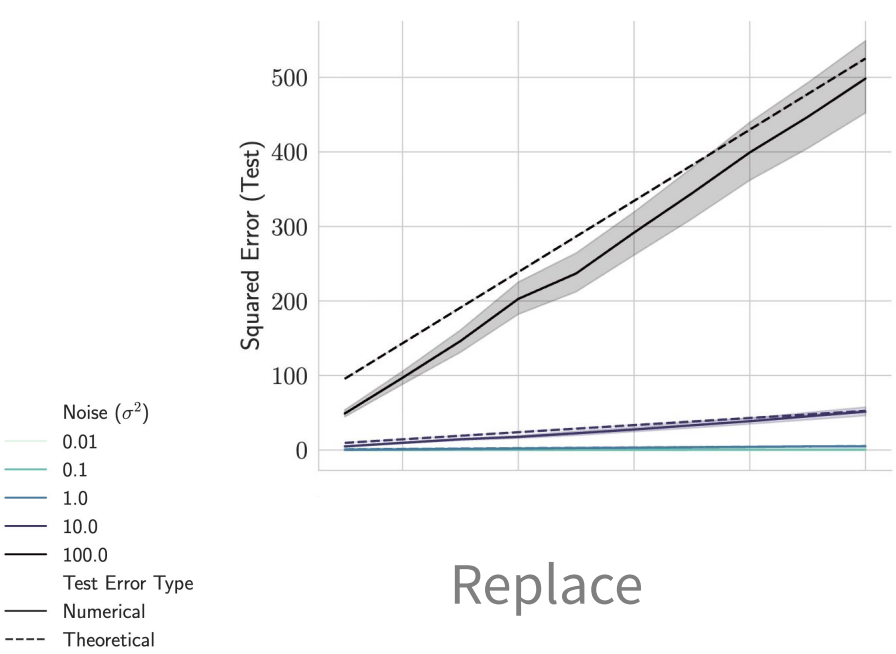
Replace

Accumulate

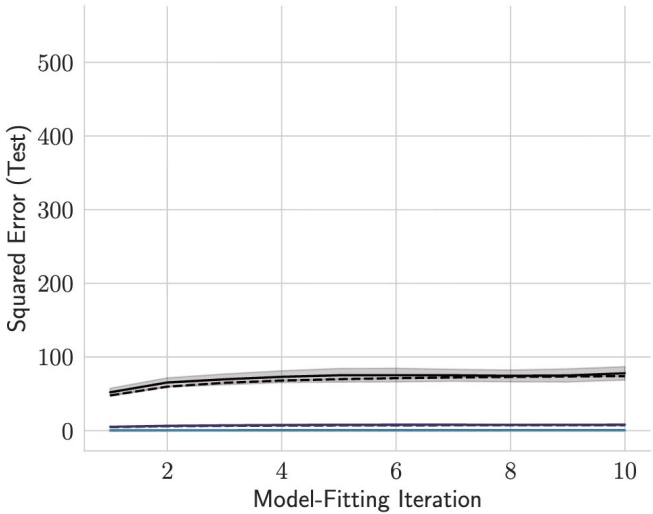
Real images

Brief Glance at Results (Linear models)

$$\begin{aligned}
&\text{(Input)} && x \sim \mathcal{N}(0, \Sigma), \\
&\text{(Noise)} && \epsilon \sim \mathcal{N}(0, \sigma^2), \text{ independent of } x, \\
&\text{(Label)} && y = x \cdot w^* + \epsilon.
\end{aligned}$$



Replace

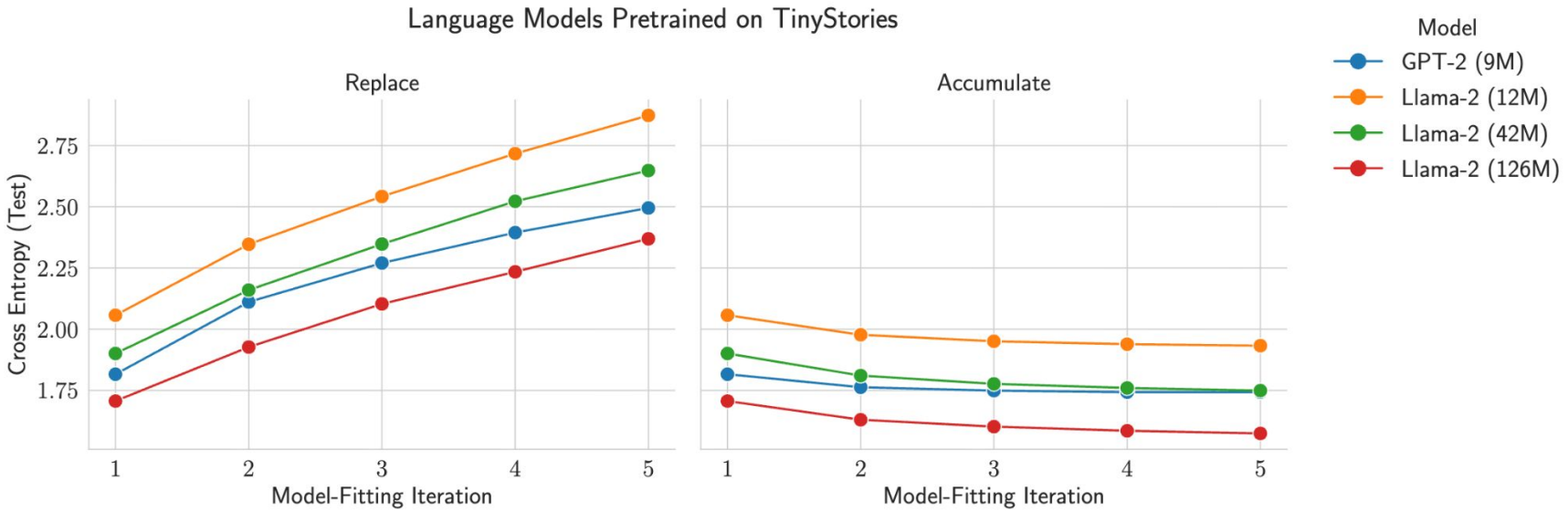


Accumulate

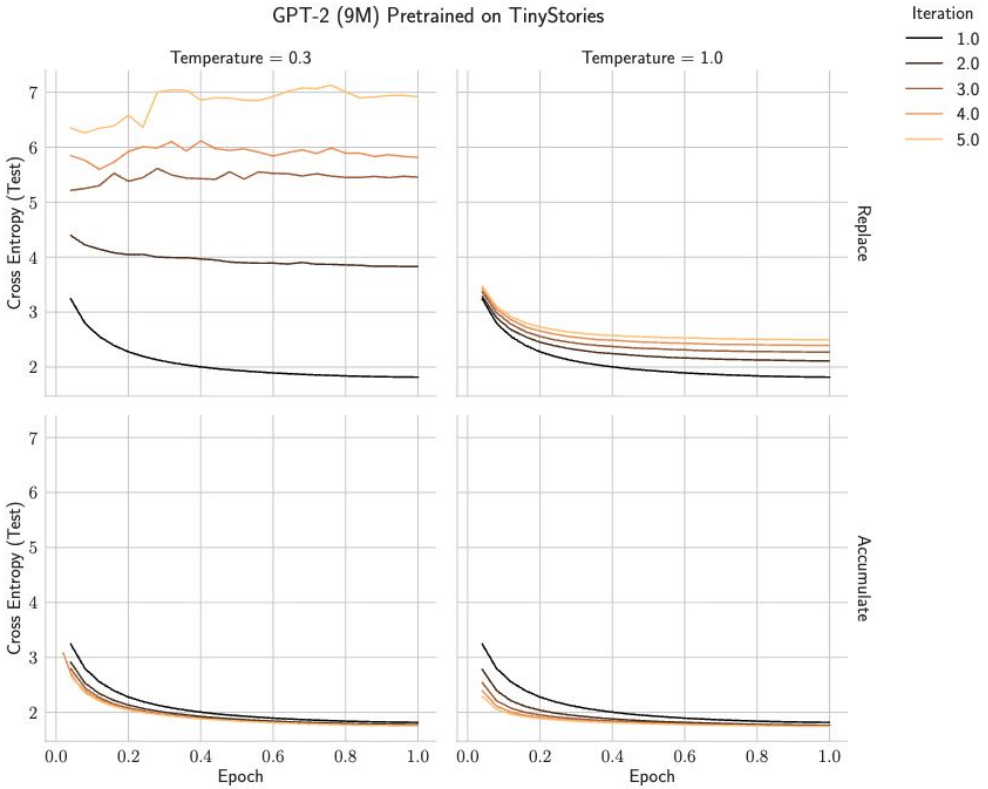
Closer look at LLM experiments

Models: Llama2 (12M, 42M, 125M) , GPT2 (9M)

Dataset: TinyStories



Q. How does temperature affect the curves?



Less temperature leads to faster loss of tail distribution

Q. Is the difference from the training compute?

Model	t=1	t=4 (acc)	t=4 (repl)	t=10 (repl)	t=4 (*)
GPT-2 (9M)	1.82	1.74 (-0.07)	2.39 (+0.58)	2.91 (+1.09)	2.18 (+0.36)
GPT-2 (9M) (temp=0.3)	1.82	1.75 (-0.06)	5.82 (+4.00)	9.85 (+8.04)	n/a
GPT-2 (9M) (small dataset)	2.56	2.28 (-0.28)	3.21 (+0.65)	3.72 (+1.16)	2.91 (+0.35)
ibid (+ 3 epochs)	1.99	1.87 (-0.12)	2.62 (+0.63)	n/a	n/a
Llama-2 (12M)	2.06	1.94 (-0.12)	2.72 (+0.66)	n/a	n/a
Llama-2 (42M)	1.90	1.76 (-0.14)	2.52 (+0.62)	n/a	n/a
Llama-2 (126M)	1.71	1.59 (-0.12)	2.23 (+0.53)	n/a	n/a

Table 2: Evaluation cross-entropy loss for different models at model-fitting iterations 1, 4 and 10 for replacement and accumulation regimes. (*) indicates a replacement regime with growing dataset size to ablate for total train set size.

Multiplying the synthetic data for R to match the training compute still leads to model collapse

Q. Are their metrics reliable?

- LLM: Crossentropy
- Diffusion Models on Modelcular data: standard loss used by GeoDiff
- VAE on images: Reconstruction error
- Linear models: MSE

Q. Are their metrics reliable?

	Iter	PPL	TTR	Distinct-2	Distinct-3	1-Self-BLEU	Div_syn	Div_sem
News Summarization	Human	–	7.36	48.1	81.1	73.3	3.17	46.6
	1	12.5	5.99 (↓)	37.9 (↓)	68.5 (↓)	74.6 (↑)	1.65 (↓)	47.2 (↑)
	2	3.42	5.55 (↓)	35.5 (↓)	64.1 (↓)	74.2 (↓)	1.76 (↑)	47.2 (→)
	3	3.09	4.99 (↓)	32.6 (↓)	59.3 (↓)	72.6 (↓)	1.95 (↑)	46.8 (↓)
	4	2.86	4.46 (↓)	29.2 (↓)	54.5 (↓)	69.7 (↓)	1.85 (↓)	46.6 (↓)
	5	2.62	3.92 (↓)	25.8 (↓)	49.5 (↓)	68.0 (↓)	1.62 (↓)	46.0 (↓)
	6	2.48	3.66 (↓)	25.6 (↓)	49.2 (↓)	65.3 (↓)	0.82 (↓)	46.6 (↑)
Scientific Abstract Generation	Human	–	3.09	35.4	75.0	71.0	4.52	40.4
	1	13.4	2.06 (↓)	20.7 (↓)	48.3 (↓)	64.2 (↓)	3.80 (↓)	39.4 (↓)
	2	3.87	1.96 (↓)	17.4 (↓)	39.8 (↓)	60.4 (↓)	4.06 (↑)	38.6 (↓)
	3	2.59	1.90 (↓)	16.1 (↓)	36.0 (↓)	59.2 (↓)	4.94 (↑)	38.6 (→)
	4	2.31	1.82 (↓)	15.3 (↓)	34.0 (↓)	58.7 (↓)	4.60 (↓)	37.6 (↓)
	5	2.24	1.77 (↓)	14.2 (↓)	31.6 (↓)	58.2 (↓)	4.41 (↓)	37.5 (↓)
	6	2.17	1.69 (↓)	13.3 (↓)	29.5 (↓)	57.5 (↓)	4.10 (↓)	37.1 (↓)
Story Generation	Human	–	2.23	30.5	70.6	67.0	4.84	43.7
	1	14.1	0.84 (↓)	13.8 (↓)	44.2 (↓)	61.6 (↓)	4.23 (↓)	41.4 (↓)
	2	4.41	0.72 (↓)	13.3 (↓)	43.1 (↓)	61.0 (↓)	3.41 (↓)	42.5 (↑)
	3	3.37	0.68 (↓)	12.8 (↓)	42.0 (↓)	60.6 (↓)	2.99 (↓)	43.3 (↑)
	4	2.99	0.65 (↓)	12.3 (↓)	40.9 (↓)	60.5 (↓)	2.50 (↓)	43.3 (→)
	5	2.82	0.63 (↓)	11.8 (↓)	39.7 (↓)	60.5 (→)	2.14 (↓)	42.7 (↓)
	6	2.70	0.61 (↓)	11.4 (↓)	38.6 (↓)	60.3 (↓)	1.96 (↓)	42.5 (↓)

(1st paper)

Over the iterations,

Perplexity (=cross entropy) gets better, but the fine-grained metrics get worse

Q. Are their metrics reliable?

Model	Iteration	Sample Generation
Llama2 (125M)	3 (A)	In the end, the crab found a smooth shell. He took it to a safe place under a tree. The crab put the shell where he found it. Tim and his mom were tired, but they were happy. They had a fun day at the beach. And they lived happily ever after. The end.
	3 (R)	Henry asked his Mom why the golf sounded so special. His Mom explained that the line of lumber had something special that would help. She said that if you're not sure, the lumber is special.
	8 (R)	Friend Stan and Millie laughed together and prepared to spend the morning together. Mamaing Grandma's possibilitant, twice would measure how much she lovedk. Everyone started to get ready when they started arguing until their mum upset.
GPT2 (9M)	5 (A)	Jack was so happy that he took care of the honey. He thought, "I care about the beautiful garden, because it is nice and clean." He started to feed the flower every day. The flower grew bigger and taller, and Jack became very happy.
	5 (R)	After playing, Lily got tired and quickly ran back to playing with her dolls. She opened her eyes and played with her dolls all day long. Her grandma was so happy that she screamed as she watched her look back at her original clothes and laughed.
	10 (R)	When she finished eating it, she tasted it all up. She said goodbye to her mom and said goodbye. Mommy smiled, feeling very proud of her. It was other. She knew that sharing is always easy to share her meal with her mom.

R looks clearly broken

1. Repetitive words
2. Broken words ("Lovedk", "possibilitant")

Q. Are their metrics reliable?

Model	Iteration	Sample Generation
Llama2 (125M)	3 (A)	In the end, the crab found a smooth shell. He took it to a safe place under a tree. The crab put the shell where he found it. Tim and his mom were tired, but they were happy. They had a fun day at the beach. And they lived happily ever after. The end.
	3 (R)	Henry asked his Mom why the golf sounded so special. His Mom explained that the line of lumber had something special that would help. She said that if you're not sure, the lumber is special.
	8 (R)	Friend Stan and Millie laughed together and prepared to spend the morning together. Mamaing Grandma's possibilitant, twice would measure how much she lovedk. Everyone started to get ready when they started arguing until their mum upset.
GPT2 (9M)	5 (A)	Jack was so happy that he took care of the honey. He thought, "I care about the beautiful garden, because it is nice and clean." He started to feed the flower every day. The flower grew bigger and taller, and Jack became very happy.
	5 (R)	After playing, Lily got tired and quickly ran back to playing with her dolls. She opened her eyes and played with her dolls all day long. Her grandma was so happy that she screamed as she watched her look back at her original clothes and laughed.
	10 (R)	When she finished eating it, she tasted it all up. She said goodbye to her mom and said goodbye. Mommy smiled, feeling very proud of her. It was other. She knew that sharing is always easy to share her meal with her mom.

Wait, A is not perfect either!

1. Crab -> Tim
2. Happy happy happy

But there is no comparison (no 1st iteration for A), so we are not sure

Q. Are their metrics reliable?

- Need future works that evaluate this setting with more diverse metrics
- Or just use the first paper’s text diversity metrics

Contributions & Limitations

Contributions

1. Showed that model collapse is preventable with simple methods and experiments
2. Various models (linear models, LLM, VAE, and diffusion models)

Limitations

1. Single eval metric for LLMs
2. Missing critical experiment results

Discussion Questions

1. Why does model collapse lead to generating broken words (not only repeated or false information)?
2. Why does perplexity decline in the first paper and increase in the second for the replacement case?
   a. Does the use of TinyStories, given that it's already synthetic, accelerate model collapse?
   b. Pretraining vs finetuning?
3. Model collapse vs. self-bootstrapping methods?
   Why don't we see model collapse in the latter? (Or do we?)
   (STaR: Bootstrapping Reasoning With Reasoning, Self-Rewarding Language Models, etc.)
4. Better ways to prevent model collapse?
   a. Like the methods introduced by Pratyush Maini

Conclusion

- Recursively training on synthetic data decreases lexical and syntactic diversity
- Accumulation “seems” to avoid model collapse better than replacement, but more diverse metrics and datasets are needed
- It’s important to maintain “creativity” as the web becomes more saturated with synthetic data

Proponents

Juno Kim, Nathan Ju

Controlling for data size

- In Paper 2, model sees increasingly more data in the accumulation regime compared to replacement (more gradient steps in 1 epoch)
- They “control” for number of updates in ablation experiments, but only have 2 data points (Appendix C Table 2)
- For these points, loss increase is almost halved → big confounding factor!

Model	t=1	t=4 (acc)	t=4 (repl)	t=10 (repl)	t=4 (*)
GPT-2 (9M)	1.82	1.74 (-0.07)	2.39 (+0.58)	2.91 (+1.09)	2.18 (+0.36)
GPT-2 (9M) (temp=0.3)	1.82	1.75 (-0.06)	5.82 (+4.00)	9.85 (+8.04)	n/a
GPT-2 (9M) (small dataset)	2.56	2.28 (-0.28)	3.21 (+0.65)	3.72 (+1.16)	2.91 (+0.35)
ibid (+ 3 epochs)	1.99	1.87 (-0.12)	2.62 (+0.63)	n/a	n/a
Llama-2 (12M)	2.06	1.94 (-0.12)	2.72 (+0.66)	n/a	n/a
Llama-2 (42M)	1.90	1.76 (-0.14)	2.52 (+0.62)	n/a	n/a
Llama-2 (126M)	1.71	1.59 (-0.12)	2.23 (+0.53)	n/a	n/a

Controlling for data size

- In particular, they ablate by increasing the **replacement** dataset size, but the **accumulation** dataset size should have been fixed
- I.e., compute remains constant while the proportion of original data decreases over iterations

Hence this comparison is invalid!

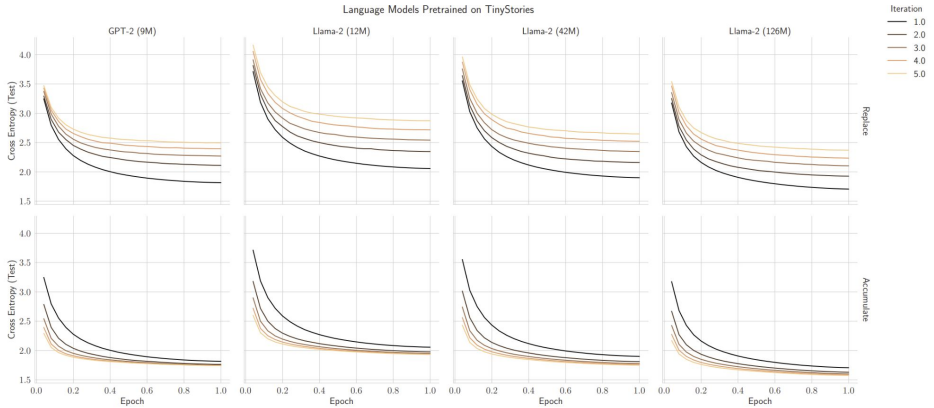


Figure 3: **Data Accumulation Avoids Model Collapse in Language Modeling.** Learning curves for individual model-fitting iterations when repeatedly *replacing* data (top), and when *accumulating* data (bottom). Note: Epochs correspond to more gradient steps for accumulate than replace because the number of training data grows for accumulate.

Potential cherry-picking

Model	Iteration	Sample Generation
Llama2 (125M)	3 (A)	In the end, the crab found a smooth shell. He took it to a safe place under a tree. The crab put the shell where he found it. Tim and his mom were tired, but they were happy. They had a fun day at the beach. And they lived happily ever after. The end.
	3 (R)	Henry asked his Mom why the golf sounded so special. His Mom explained that the line of lumber had something special that would help. She said that if you're not sure, the lumber is special.
	8 (R)	Friend Stan and Millie laughed together and prepared to spend the morning together. Mamaing Grandma's possibilitant, twice would measure how much she lovedk. Everyone started to get ready when they started arguing until their mum upset.
GPT2 (9M)	5 (A)	Jack was so happy that he took care of the honey. He thought, "I care about the beautiful garden, because it is nice and clean." He started to feed the flower every day. The flower grew bigger and taller, and Jack became very happy.
	5 (R)	After playing, Lily got tired and quickly ran back to playing with her dolls. She opened her eyes and played with her dolls all day long. Her grandma was so happy that she screamed as she watched her look back at her original clothes and laughed.
	10 (R)	When she finished eating it, she tasted it all up. She said goodbye to her mom and said goodbye. Mommy smiled, feeling very proud of her. It was other. She knew that sharing is always easy to share her meal with her mom.

As mentioned above – do not show examples for accumulation regime at higher iterations (which supposedly avoids model collapse)

Toy theoretical analysis

- Paper 2 analyses a toy linear model and show model collapse is bounded; but too simplistic to give good intuition

$$
\begin{aligned}
&\text{(Input)} && x \sim \mathcal{N}(0, \Sigma), \\
&\text{(Noise)} && \epsilon \sim \mathcal{N}(0, \sigma^2), \text{ independent of } x, \\
&\text{(Label)} && y = x \cdot w^* + \epsilon.
\end{aligned}
$$

$$P_{\Sigma, w^*, \sigma^2} \to P_{\Sigma, \hat{w}_1, \sigma^2} \to \ldots \to P_{\Sigma, \hat{w}_n, \sigma^2},$$

where $n \in \mathbb{N}$ is the number of iterations. The scheme is outlined as follows.

- For $n = 1$:
  - Accumulating Covariates/Features: $\tilde{X}_1 \stackrel{\text{def}}{=} X$
  - Accumulating Targets: $\tilde{Y}_1 \stackrel{\text{def}}{=} \hat{Y}_1 \stackrel{\text{def}}{=} Xw^* + E_1$, where $E_1 \stackrel{\text{def}}{=} E \sim \mathcal{N}(0, \sigma^2 I_T)$
  - Fit linear model: $\hat{w}_1 = \tilde{X}_1^\dagger \tilde{Y}_1$
  - Sample synthetic data for the next iteration: $\hat{Y}_2 \stackrel{\text{def}}{=} X\hat{w}_1 + E_2$, where $E_2 \sim \mathcal{N}(0, \sigma^2 I_T)$
- For $n \geq 2$:
  - Accumulating Covariates/Features: $\tilde{X}_n^\top = [\tilde{X}_{n-1}^\top; X^\top] \in \mathbb{R}^{d \times nT}$
  - Accumulating Targets: $\tilde{Y}_n^\top = [\tilde{Y}_{n-1}^\top; \hat{Y}_n^\top] \in \mathbb{R}^{1 \times nT}$
  - Fit linear model: $\hat{w}_n \stackrel{\text{def}}{=} \tilde{X}_n^\dagger \tilde{Y}_n$
  - Sample synthetic data for the next iteration: $\hat{Y}_{n+1} \stackrel{\text{def}}{=} X\hat{w}_n + E_{n+1}$, where $E_{n+1} \sim \mathcal{N}(0, \sigma^2 I_T)$

Toy theoretical analysis

Theorem 1. *In the data accumulation setting, $\forall n \geq 1$, the fitted linear parameters $\hat{w}_n$ can be expressed as:*

$$\hat{w}_n = w^* + (X^\top X)^{-1} X^\top \left(\sum_{i=1}^{n} \frac{E_i}{i} \right) \tag{2}$$

Theorem 2. *For an n-fold synthetic data generation process with $T \geq d + 2$ samples per iteration and isotropic features ($\Sigma \overset{def}{=} I_d$), the test error for the ridgeless linear predictor $\hat{w}_n$ learned on the accumulated data up to iteration n is given by:*

$$E_{test}^{Accum}(\hat{w}_n) = \frac{\sigma^2 d}{T - d - 1} \left(\sum_{i=1}^{n} \frac{1}{i^2} \right) \leq \frac{\sigma^2 d}{T - d - 1} \times \frac{\pi^2}{6} \tag{3}$$

Toy theoretical analysis

- This relies heavily on the entire system being linear (hence unbiased); **general nonlinear systems** will have multiplicative error accumulation
- Even with data dist. shift decreasing as (1/i) and generously assuming some system “Lipschitz” constant c, this leads to unbounded error:

$$\prod_{i=1}^{T}\left(1+\frac{c}{i}\right) \geq 1 + c\underbrace{\sum_{i=1}^{T}\frac{1}{i}}_{H_T} \to \infty \text{ as } T \to \infty.$$

Another model collapse paper (Shumailov et al., 2024) have similar analysis for 1D Gaussian sampling due to data gen. also using sample covariance:

$$\mathbb{E}_{\mu_{n+1},\sigma^2_{n+1}}\left[R^{n+1}_{W_2}\right] = \frac{3}{2}\sigma^2\left(\frac{1}{M_0}+\frac{1}{M_1}+\cdots+\frac{1}{M_n}\right)+\mathcal{O}(2),$$

By *natural* standards, model collapse still occurs

STRONG MODEL COLLAPSE

Elvis Dohmatob[1,2,3], **Yunzhen Feng**[4,†], **Arjun Subramonian**[5,†], **Julia Kempe**[1,4]
[1]Meta FAIR [2]Concordia University [3]Mila [4]NYU [5]UCLA
[†]Work done while interning at Meta. Correspondence to elvis.dohmatob@concordia.ca

ABSTRACT

Within the scaling laws paradigm, which underpins the training of large neural networks like ChatGPT and Llama, we consider a supervised regression setting and establish a strong form of the model collapse phenomenon, a critical performance degradation due to synthetic data in the training corpus. Our results show that even the smallest fraction of synthetic data (e.g., as little as 1 per 1000) can still lead to model collapse: larger and larger training sets do not enhance performance. We further investigate whether increasing model size, an approach aligned with current trends in training large language models, exacerbates or mitigates model collapse. In a simplified regime where neural networks are approximated via random projections of tunable size, we both theoretically and empirically show that larger models can amplify model collapse. Interestingly, our theory also indicates that, beyond the interpolation threshold (which can be extremely high for very large datasets), larger models may mitigate the collapse, although they do not entirely prevent it. Our theoretical findings are empirically verified through experiments on language models and neural networks for images.

This paper appears to arrive at a contradictory conclusion: The presence of synthetic data still leads to model collapse. Why?

By *natural* standards, model collapse still occurs

Beside some major academic drama about paper 2 on OpenReview…

> already presented in an earlier version of Dohmatob et al. (2024a) "Model Collapse Demystified
> Their results can be recovered as a trivial corollary of Theorem 4.1 of Dohmatob et al. (2024a). A
> **Remark 4.2** of Dohmatob et al. (2024a)

There is some discussion on the definition of model collapse:

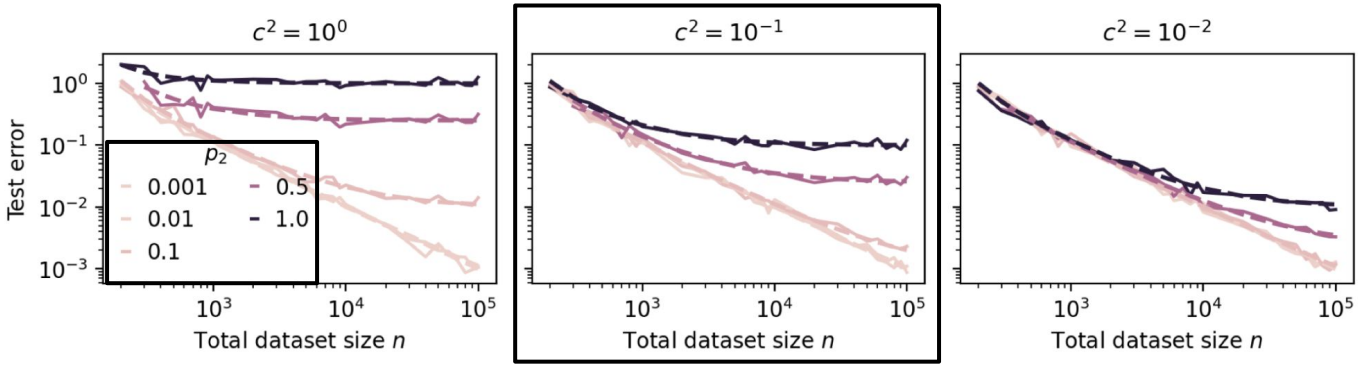
> In our introduction, we clearly define model collapse as "a critical degradation in the performance of AI models." **Avoiding model collapse would mean to *close the performance gap between training with real and synthetic data.***

Although accumulating real+synthetic data has bounded error, it does not match the performance of using real data.

In fact, there still is a **noticeable performance gap when accumulating data.**

By *natural* standards, model collapse still occurs

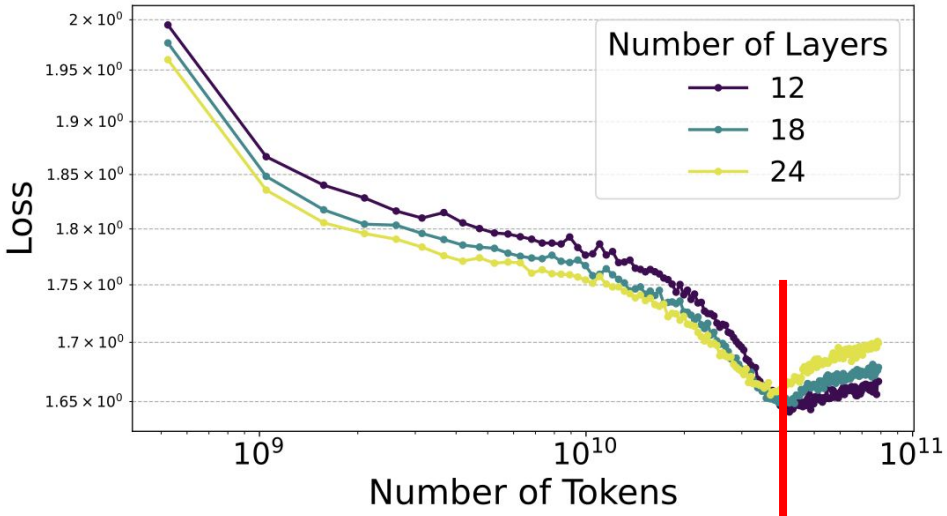
p_2 = proportion of synthetic data
c^2 ~ “quality” of synthetic data



Takeaway: *If synthetic data is not high quality*, then mixing real+synthetic data still has a noticeable (sometimes plateauing) suboptimality in test error.

On the role of model size

From *Strong model collapse*: “larger models tend to amplify model collapse beyond the interpolation threshold”. Training LM on synthetic data:

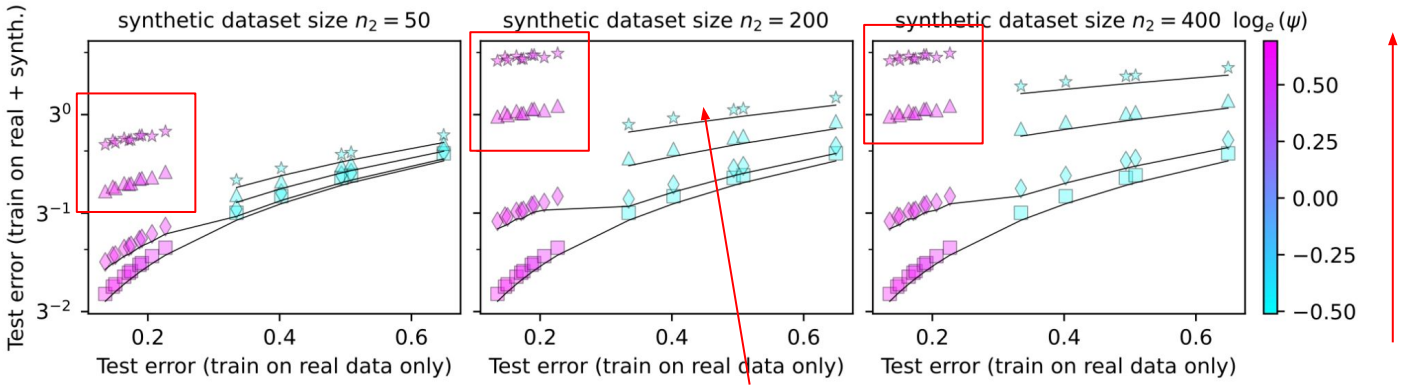


larger models suffer **less** from model collapse here

larger models suffer **more** from model collapse here

On the role of model size

Past the interpolation threshold, If quality of synthetic data is bad, then larger models will exacerbate the problem.



Critics

Siddharth Gollapudi, Dennis Jacob

Oct 27 2025

Recap

Problem: recursive training of LLMs with synthetic data

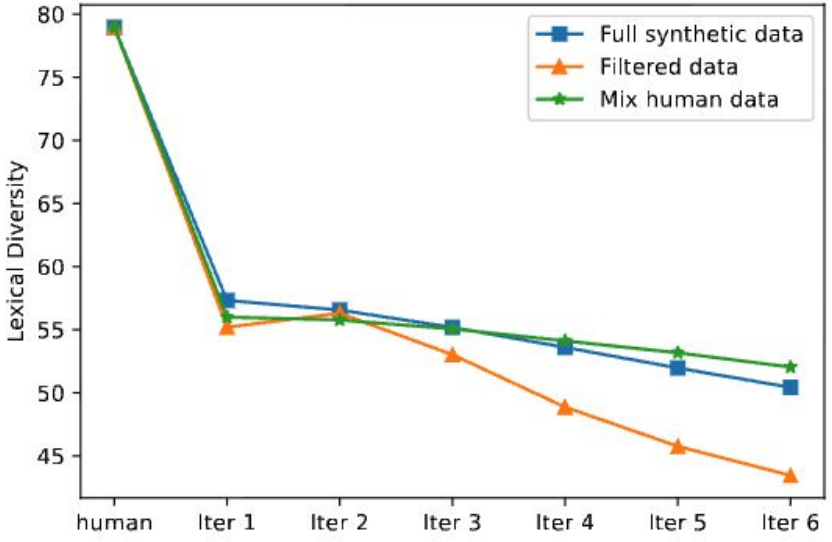
Paper 1: experimental setting is not super thorough

- Based on earlier work of Shumailov, et al. (2023) on model perf.
- Focuses on linguistic diversity, not actual model collapse

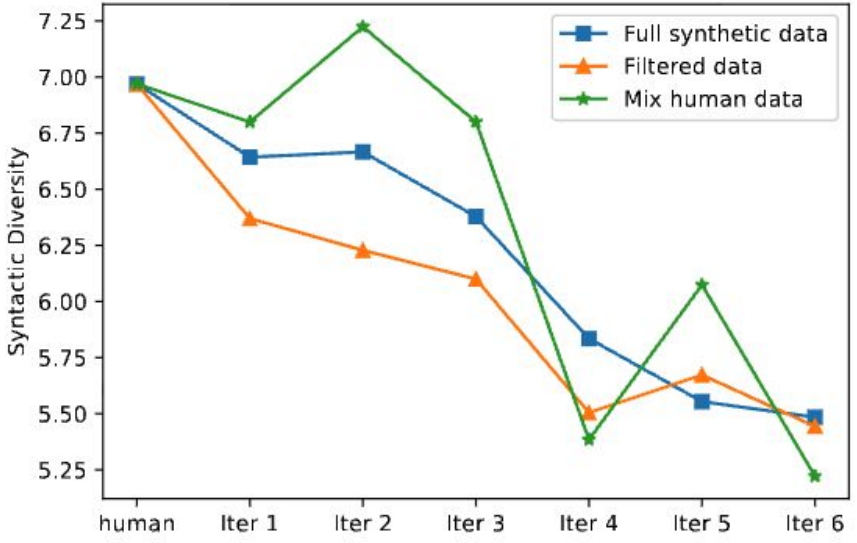
Paper 2: maybe kinda works?

- Accumulating data vs. replacing works (better)
- Still (provably) see bounds on performance

Issue 1: Maybe the situation is salvageable?



(a) Lexical diversity.



(b) Syntactic diversity.

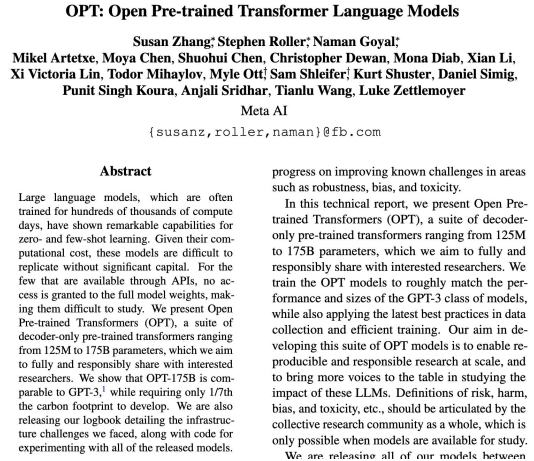
1. Human-mixed data usually performs better with more iterations
2. No ablations done on the filtering

Issue 1 con't: Scale might be able to help…

The authors only test opt-350M (from 2022)…

- Will larger models have similar observations? How about reasoners (i.e., Qwen3-4B, Llama 3.1 8B, etc.)

Meta



OPT: Open Pre-trained Transformer Language Models

Susan Zhang*, Stephen Roller*, Naman Goyal*,
Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li,
Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig,
Punit Singh Koura, Anjali Sridhar, Tianlu Wang, Luke Zettlemoyer
Meta AI
{susanz,roller,naman}@fb.com

Abstract

Large language models, which are often trained for hundreds of thousands of compute days, have shown remarkable capabilities for zero- and few-shot learning. Given their computational cost, these models are difficult to replicate without significant capital. For the few that are available through APIs, no access is granted to the full model weights, making them difficult to study. We present Open Pre-trained Transformers (OPT), a suite of decoder-only pre-trained transformers ranging from 125M to 175B parameters, which we aim to fully and responsibly share with interested researchers. We show that OPT-175B is comparable to GPT-3,[1] while requiring only 1/7th the carbon footprint to develop. We are also releasing our logbook detailing the infrastructure challenges we faced, along with code for experimenting with all of the released models.

progress on improving known challenges in areas such as robustness, bias, and toxicity.

In this technical report, we present Open Pre-trained Transformers (OPT), a suite of decoder-only pre-trained transformers ranging from 125M to 175B parameters, which we aim to fully and responsibly share with interested researchers. We train the OPT models to roughly match the performance and sizes of the GPT-3 class of models, while also applying the latest best practices in data collection and efficient training. Our aim in developing this suite of OPT models is to enable reproducible and responsible research at scale, and to bring more voices to the table in studying the impact of these LLMs. Definitions of risk, harm, bias, and toxicity, etc., should be articulated by the collective research community as a whole, which is only possible when models are available for study.

We are releasing all of our models between

Issue 1 con't: Better prompting methods?

The authors find that semantic similarity (i.e., content of produced outputs) does not degrade too much…

- Baseline level of info present in successive iterations is still high

	Iter	Div_sem
News Summarization	Human	46.6
	1	47.2 (↑)
	2	47.2 (→)
	3	46.8 (↓)
	4	46.6 (↓)
	5	46.0 (↓)
	6	46.6 (↑)

Issue 1 con't: Better prompting methods?

The authors find that semantic similarity (i.e., content of produced outputs) does not degrade too much…

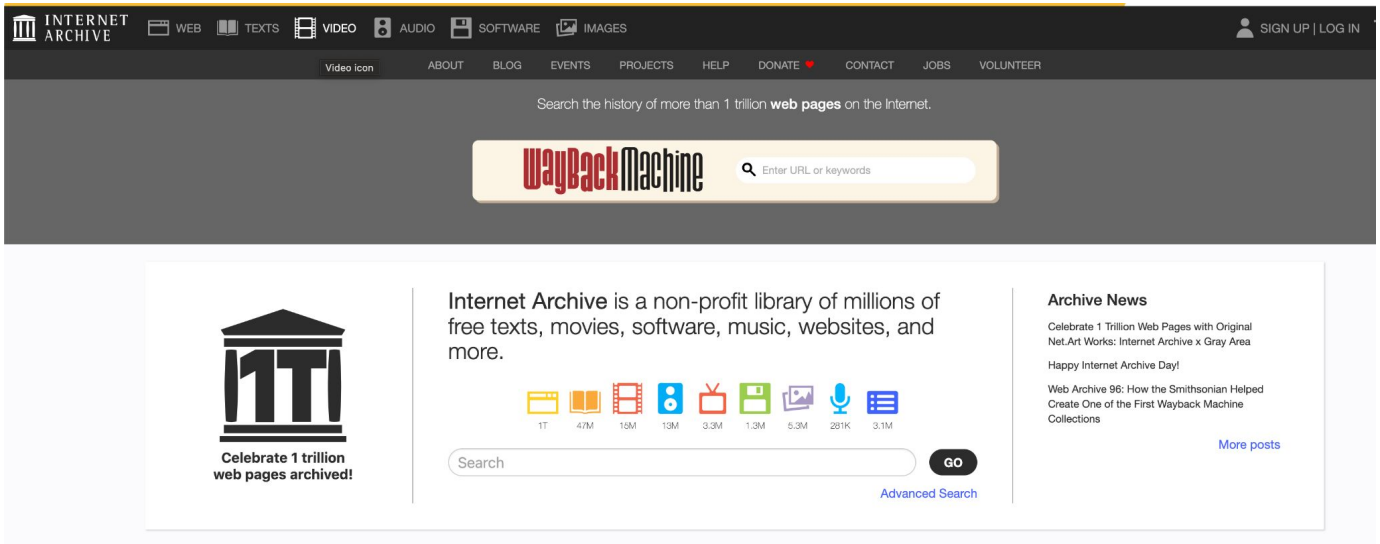
- Baseline level of info present in successive iterations is still high
- Better prompting could be used to extract stronger diversity

Prompt: The title of the paper is: CASIA's System for IWSLT 2020 Open Domain Translation.The abstract of the paper is: This paper describes the CASIA's system for the IWSLT 2020 open domain translation task.
Human: This year we participate in both Chinese→Japanese and Japanese→Chinese translation tasks. Our system is neural machine translation system based on Transformer model. We augment the training data with knowledge distillation and back translation to improve the translation performance. Domain data classification and weighted domain model ensemble are introduced to generate the final translation result. We compare and analyze the performance on development data with different model settings and different data processing techniques..
Iteration 1: We participated in all language directions, namely Chinese ↔ English, Japanese↔ English and German ↔ English. Our system mainly focuses on data selection, data preprocessing and data filtering to improve the quality of the generated translations..

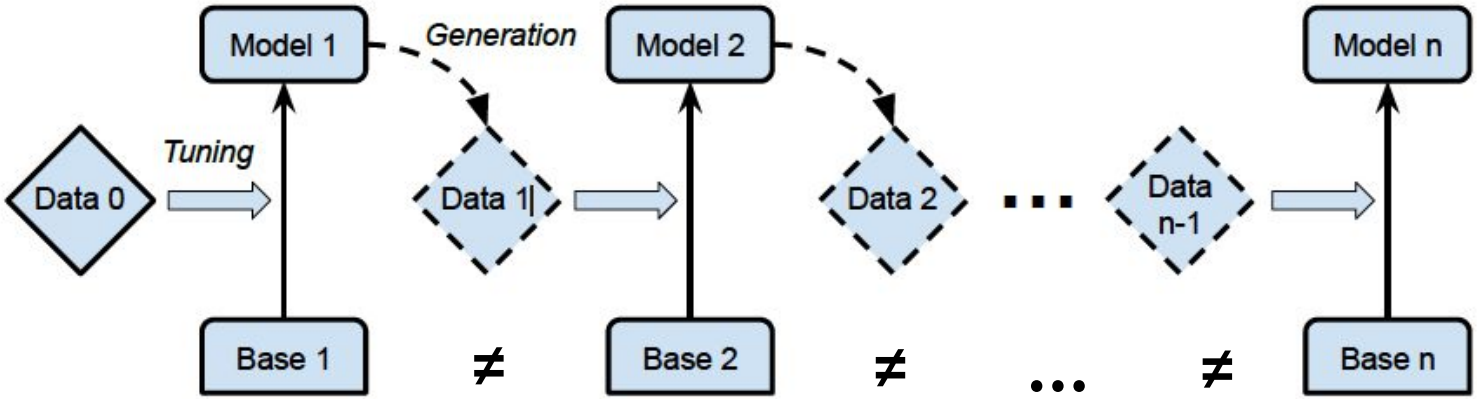
Issue 2: Possibly unrealistic experimental setup

The data accumulation strategy is more realistic in Breaking the Curse

- Available data in real life won't just disappear (i.e., Internet Archive)
- Perhaps diversity will be maintained after data accumulation?

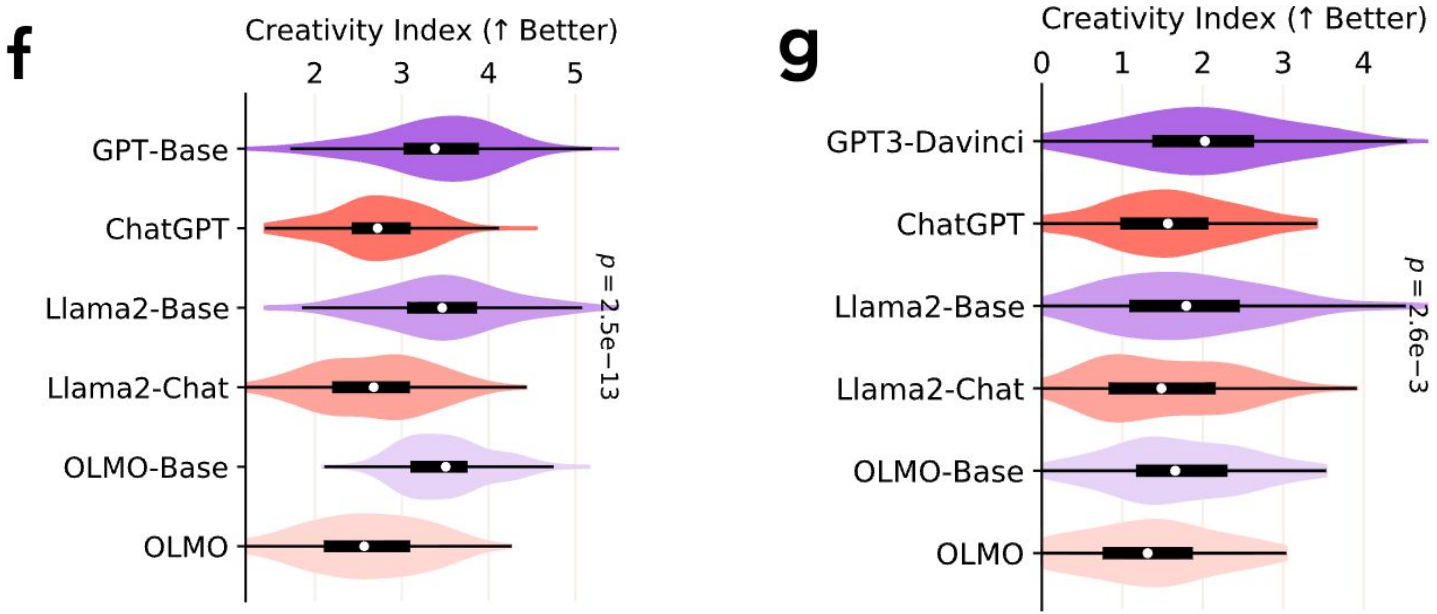


Issue 2: Possibly unrealistic experimental setup



As models scale, architectures change -> do results generalize in this more realistic setting?

Issue 3: Is the model the problem?



Lu, et al. (2025)

Before and after RLHF alignment: creativity gets worse with RLHF

Maybe we're the problem?

Issue 3 con't: The cost of a dead internet

1. Are human linguistic preferences impacted by existing synthetic data?
2. Does the rise of AI-generated content on the web play a role?
3. How do we adjust our training regime to accommodate?
    a. Injecting more "creativity" into the training?
    b. Is just increasing the diversity in the human feedback enough?
    c. Policy controls on AI content dissemination?