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

Yanzhu Guo[1], **Guokan Shang**[2,4]*, **Michalis Vazirgiannis**[1,2], **Chloé Clavel**[3]
[1]École Polytechnique, Institut Polytechnique de Paris
[2]Mohamed bin Zayed University of Artificial Intelligence
[3]Inria [4]Linagora
yanzhu.guo@polytechnique.edu, guokan.shang@mbzuai.ac.ae
mvazirg@lix.polytechnique.fr, chloe.clavel@inria.fr

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.

1 Introduction

The scaling law reveals a predictable smooth increase in model performance as the amount of data, compute power, and model parameters are increased in tandem (Ganguli et al., 2022). Even assuming that we can boost the other two ingredients indefinitely, the amount of data is limited. By one estimate, the world's entire supply of *high-quality* text ranges up to 17 trillion tokens, with a 4-5% yearly growth rate (Villalobos et al., 2022). This includes all the world's books, scientific papers, news articles, Wikipedia pages, available code, and the rest of filtered web content. Meta's Llama 2, one of today's leading LLMs, was trained on around 2 trillion tokens (Touvron et al., 2023). In other words, we might be approaching the exhaustion of the world's entire stock of usable language training data, potentially within an order of magnitude.

Is it possible for language models to train on self-generated samples, thereby offering a solution to the looming data shortage? In fact, whether intentionally or unintentionally, this would happen with the widespread recognition and usage of LLMs. Regarding pretraining data, which is often sourced from the Internet, a significant trend is occurring: an increasing volume of online content is either generated or assisted by models, and such content is nearly indistinguishable from data produced by humans (Uchendu et al., 2023). Consequently, the subsequent generations of models will inevitably be pretrained on deeply blended data. Regarding finetuning data, employing LLM-generated examples is already a widely adopted data augmentation approach in the NLP community. The work of self-instruct (Wang et al., 2023) prompts language models to solicit synthetic multi-task instruction-tuning data in an iterative bootstrapping way, starting with a seed set of manually-written instructions. Concerning single-task training, Zhou et al. (2023) build a large-scale dialogue summary corpus annotated by ChatGPT (Ouyang et al., 2022) to enhance their pretrained dialogue summarization model.

However, recent studies raise concerns that the above approach of training on *predecessor-generated* text—language models are trained on the synthetic data produced by previous models—is not a panacea without side effects, especially when conducted *recursively* over time. This would introduce a new set of challenging issues, a phenomenon described as *model collapse* (Shumailov et al., 2023; Alemohammad et al., 2024). On one hand, incorporating model-generated content in training may lead to irreversible flaws in the resulting models, where tails of the original distribution of genuine human content disappear (Shumailov et al., 2023). On the other hand, even when these

*Part of the work was done when this author was affiliated with Linagora.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3589–3604

models remain free of defects, they could converge to excessively uniform behaviours, with very small variance, due to the recursive sampling of only high probability events.

In this study, rather than focusing on shifts in task-solving performance, our primary interest lies in exploring changes in language variation caused by the degenerative recursive training process. We target linguistic diversity, a fundamentally important but significantly overlooked aspect of language usage. Our work is motivated by and contributes to, answering the following two key research questions: First, *how can linguistic diversity be quantified effectively?* Second, *does recursive training on synthetic text result in a reduction of linguistic diversity in model outputs?*

To address these questions, we first develop a comprehensive set of metrics[1] assessing at three different aspects of linguistic diversity: lexical, semantic, and syntactic. Subsequently, we proceed to conduct a series of recursive finetuning experiments spanning three natural language generation tasks, each demanding varying levels of creativity: news summarization (Hasan et al., 2021), scientific abstract generation, and story generation (Fan et al., 2018). Our results indicate a notable trend: with the progression of recursive finetuning iterations, there is indeed a remarkable decrease in the diversity of the generated outputs. This observation highlights the significant impact that training on text generated by predecessors has on the linguistic diversity of language models.

2 Related Work

In this section, we explore two avenues of related work: current approaches to evaluate linguistic diversity and recent research on training with synthetic data generated by language models.

2.1 Evaluating Linguistic Diversity

Efforts to evaluate language models predominantly concentrate on their performance in task-solving. While some studies extend their scope to include aspects like factual consistency (Guo et al., 2022), reasoning capability (Helwe et al., 2021), and robustness (Chang et al., 2023), there is a notable lack of attention paid to linguistic diversity.

Furthermore, the existing studies that do address the diversity issue typically focus on lexical diversity alone. For example, in quantifying diversity, research on decoding strategies (Li et al., 2023; Vijayakumar et al., 2018; Ippolito et al., 2019) usually considers the proportion between the number of unique n-grams and total number of n-grams in generated text, known as distinct-n metric (Li et al., 2016). This very approach can also be found in the literature related to specific NLG tasks, especially those studying creative text generation, such as poetry (Chakrabarty et al., 2022), lyric (Tian et al., 2023), and pun (Mittal et al., 2022) generation.

Alternatively, Zhang et al. (2021) propose to use Shannon entropy to quantify diversity. However, such an approach is still calculated on tokens (i.e., lexical level), demonstrating a strong correlation with distinct-n. Zhu et al. (2018) introduce Self-BLEU which calculates the BLEU similarity score (Papineni et al., 2002) between different sentences of the same document, with higher Self-BLEU implying lower diversity. This metric is adopted as a proxy for diversity in evaluating the capability of LLMs in the context of producing content for disinformation operations (Liang et al., 2022). Nevertheless, the BLEU score is based on n-gram overlap and thus also represents diversity solely from the lexical aspect.

Few works study diversity beyond the lexical level. Recently, Padmakumar and He (2023) bring up the semantic aspect of diversity and define the average pairwise BERTScore among a set of documents as the homogenization index. They also use ChatGPT to annotate key points on a small set of documents, counting the percentage of unique key points as content diversity. Stasaski and Hearst (2022) hypothesize that the semantic diversity can be reflected by the contradictory level—measured by a natural language inference model—among different generation samples given the same input context, while Tevet and Berant (2021) consider it as the negation of the semantic similarity. As an exploratory approach to quantify syntactic diversity, Clercq and Housen (2017) first manually annotate a small corpus of texts produced by second language learners for syntactic features such as syntactic length and clause types, whose variation is then viewed as a diversity index. Huang et al. (2023a) define the syntactic diversity as the editing distance between the constituency parse trees of two sentences in the context of paraphrase generation. McCoy et al. (2023) investigate linguistic novelty, in a sense of language generation not simply copies training text, in terms of sequential structure

[1]Code available at https://github.com/YanzhuGuo/linguistic-diversity

(n-grams) and syntactic structure (constituency and dependency).

Our work is the first to comprehensively evaluate text generation on all three aspects of linguistic diversity: lexical, syntactic and semantic, with novel automatic metrics.

2.2 Training with Synthetic Text

Ever since the introduction of generative adversarial networks (Goodfellow et al., 2014), training new models with synthetic data produced by various generators has become a means of data augmentation (Li et al., 2022), a practice that has been expanding to all modalities of machine learning research, including image, audio, and text.

However, the large-scale usage of this approach, particularly employing tremendous quantities of synthetic text to train generative models, is a more recent trend (Dai et al., 2023; Marwala et al., 2023). To name a few, the self-instruct study by Wang et al. (2023) guides a language model to iteratively generate synthetic multi-task instruct-tuning data, beginning with an initial set of manually-written instructions. Huang et al. (2022) demonstrate that LLMs are capable of self-improving their reasoning abilities, with generated high-quality answers for unlabeled questions, using chain-of-thought (Wei et al., 2022) and self-consistency (Wang et al., 2022) prompting techniques. Meanwhile, Xu et al. (2023) introduce a pipeline that autonomously generates a high-quality, multi-turn chat corpus by enabling ChatGPT to converse with itself, which is then used to enhance a LLaMA model.

As already mentioned in Section 1, studies show that this training methodology will eventually lead to model collapse (Shumailov et al., 2023) when conducted recursively, causing performance degeneration, regardless of potential data filtering or refinement (Alemohammad et al., 2024). Our research is motivated by the same concept, but we focus on investigating the impact of recursive training on linguistic diversity instead of performance. To the best of our knowledge, our work is the first to address this issue.

3 Methodology

This section introduces our recursive training methodology and outlines the linguistic diversity metrics.

3.1 Recursive Training Simulation

Following the work of Shumailov et al. (2023), we simulate the process of recursively training language models on predecessor-generated text, under a finetuning setting. As illustrated in Figure 1, we begin with human-generated task-finetuning Data (0), which is used to train Base (1) model to create a task-specialized version, referred to as Model (1). After that, we use Model (1) to produce synthetic task-finetuning Data (1), which serves to train the next generation, Model (2), built upon Base (2) model. This procedure is repeated n times.

For the sake of simplicity, we start from a new instance of the same base model across different generations, i.e., Base (1) = Base (2) = , ..., = Base (n). In addition, we only use Data ($n-1$) to train Model (n), whereas in a setting closer to the real-life scenario, we have access to the accumulated data ensemble of all predecessors, i.e., Data $\{$(0), (1) , ..., ($n-1$)$\}$. This simplification draws from the results of Shumailov et al. (2023), which indicates that model collapse is unavoidable, even when the training involves the full ensemble of accumulated data, though the effect is somewhat attenuated.

In terms of finetuning tasks, we chose three distinct natural language generation tasks, each characterized by varying degrees of constraint, from the most restrictive to the least: news summarization, where summaries must closely align with the original content; scientific abstract generation, with some initial context provided, but room for creative expansion; and story generation, which allows for the most creativity and freedom in expression.

In the end, we conduct our linguistic diversity research with the finetuned Model $\{$(1), (2) , ..., (n)$\}$ for each task, subjecting them to evaluation on the test set of the corresponding task.

3.2 Perplexity

Our research primarily focuses on linguistic diversity, yet we also require a reliable metric to verify that our finetuned models are well-aligned with the training data. Perplexity, a standard metric for assessing language modeling, evaluates a model's level of "surprise" or "confusion" when encountering a given sequence of tokens. Models that more accurately mirror the training data's distribution exhibit lower perplexity. While useful for model comparison, perplexity doesn't fully reflect text quality (Meister et al., 2023). A low perplexity score suggests higher predictive precision, but

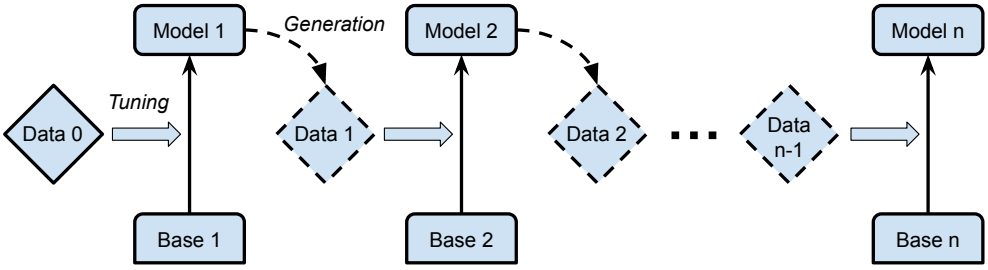


Figure 1: Our recursive tuning-generation process. Beginning with authentic, human-curated Data (0), Base (1) model undergoes finetuning to develop Model (1), which is the first model subject to our language diversity research. Subsequently, we use Model (1) to create synthetic Data (1) to train a successor Model (2) of the next generation, on the basis of Base (2) model. The process continues for n iterations. Base (1), Base (2), ..., Base (n) follow the same model architecture but are independently initialized instances.

texts can be grammatically sound and contextually coherent yet still score high in perplexity if they include unusual or creative language not included in the model's training data (Basu et al., 2021). In our study, a model with lower perplexity is not deemed superior by default. Our aim is to ensure that the perplexity remains within a reasonable limit, producing texts of sufficient quality for our linguistic diversity evaluation.

3.3 Linguistic Diversity Metrics

We approach the evaluation of linguistic diversity from three different perspectives: lexical diversity, semantic diversity and syntactic diversity.

3.3.1 Lexical Diversity

Lexical diversity metrics are used to measure the variety of words used in a text, which is contended to mirror the extent of vocabulary possessed by a writer or speaker. We believe a degenerated language model, which presumably has a smaller vocabulary, will use a narrower variety of lexical items than non-degenerated language models. We select different metrics operating at different levels of textual granularity: word, n-gram, and sentence.

Type-Token Ratio (TTR) (Johnson, 1944; Templin, 1957), the most well-known metric, which is calculated as the number of unique words (types) t divided by the number of running words (tokens) c, i.e., TTR $= t/c$. This metric was used to study the language development in child language research, a low value is probably indicative of a language-specific deficiency (Miller, 1981). The length of a text inherently skews vanilla TTR values, with longer texts generally yielding lower TTR scores due to an inexorably decreased occurrence of unique novel words (drawn from a limited vocabulary) as the text lengthens (Richards, 1987). Following common practice (Shaib et al., 2024), we truncate all texts to a fixed length before computing the TTR [2].

Distinct-n (Li et al., 2016), which equals the proportion between the number of unique n-grams and total number of n-grams in tested text (Xing et al., 2017). This metric is originally introduced in the context of enhancing the response diversity of conversational agents, which frequently produce safe and fluent but dull and uninformative generic responses at time (e.g., *I don't know*) (Han et al., 2022). Similar to naive TTR, distinct-n varies as a function of text length, so we report the results at fixed sizes for $n = 2$ and $n = 3$ (distinct-1 is equivalent to TTR).

Self-BLEU (Zhu et al., 2018), a recently developed method for evaluating the diversity of synthetic text. This method assesses the similarity between one sentence and the rest in a group of generated sentences. It treats one sentence as the hypothesis and the others as references to calculate BLEU scores (Papineni et al., 2002). The final Self-BLEU score averages these BLEU scores across all generated sentences. We report $1 -$ `Self-BLEU`, so a higher value reflects richer diversity of the generation (Palumbo et al., 2020).

3.3.2 Semantic Diversity

According to recent studies (Tevet and Berant, 2021; Stasaski and Hearst, 2022), the above lexical-level metrics often fail to capture semantic diversity, since texts including similar words can have different semantics and texts with different words can have similar semantics (Yarats and Lewis, 2018). We tackle this problem by transforming sentences

[2]Details on the truncation lengths can be found in Appendix B.

into semantically meaningful sentence embeddings using Sentence-BERT (Reimers and Gurevych, 2019). We quantify semantic diversity as the dispersion of sentence embeddings over the semantic space. The dispersion is measured by the average pairwise cosine-distance of all embedding vectors (`Div_sem`).

3.3.3 Syntactic Diversity

The significance of syntactic diversity is often underestimated in NLP, despite its importance. For language learners (as well as language models), exposure to a wide range of syntactic structures is beneficial for developing a more comprehensive understanding of the language (Aggarwal et al., 2022). Moreover, a range of syntactic forms enhances expressiveness and subtlety in writing, influencing the style and tone of a text (Edwards and Bastiaanse, 1998). While linguistic and language acquisition research (Clercq and Housen, 2017) has explored this aspect, these studies typically rely on manual annotation of features, a process that can be costly and prone to human error.

We introduce the first graph-based metric to quantify syntactic diversity. We use a neural parser (Qi et al., 2020) to construct dependency trees from sentences, following the universal dependencies formalism. These trees are then transformed into graph representations, with nodes representing the words and edges indicating the dependency relationships between them. Subsequently, we employ the Weisfeiler-Lehman graph kernel (Shervashidze et al., 2011; Siglidis et al., 2020) to map these graphs into a vector space. This kernel, rooted in the Weisfeiler-Lehman isomorphism test, effectively positions graphs that are structurally alike closer to each other in the embedding space. To assess syntactic diversity, we calculate it similarly to semantic diversity, using the average pairwise distance (`Div_syn`).

4 Experiments and Results

We conduct our experiments on three generative tasks, as introduced in Section 3.1, with decreasing degrees of constraint and increasing degrees of creativity: abstractive news summarization, scientific abstract generation, and story generation.

4.1 Experimental Setup

For each task, we simulate 6 iterations of the recursive training chain, i.e., $n = 6$ in Figure 1. Following previous work (Shumailov et al., 2023), we select OPT (Zhang et al., 2022) as our base model, and each iteration begins with a new instance of the base model. Different from Shumailov et al. (2023), we use OPT-350M instead of OPT-125M to maintain higher generation quality over iterations, avoiding excessive noise. Model (1) is finetuned on Data (0)—the training set of the finetuning task—which is human-authored. From Model (2) to Model (6), they are finetuned on synthetic Data $(n-1)$ generated by their predecessor Model $(n-1)$. We go through all of the original training examples in Data (0) to produce a comparable synthetic dataset of the same number of samples. The models are finetuned for 5 epochs using the AdamW optimizer (Loshchilov and Hutter, 2019) on a cluster of two NVIDIA RTX A6000 GPUs.

In the following, we explain each of the three tasks in detail.

Task1: Abstractive News Summarization

For abstractive news summarization, we use the XL-SUM (Hasan et al., 2021), one of the most recently proposed datasets. In comparison to the other prominent news summarization datasets, XL-SUM is more abstractive than CNN/DailyMail (Hermann et al., 2015; Cheng and Lapata, 2016) and more factual than XSUM (Narayan et al., 2018; Guo et al., 2022). It is also larger in scale, consisting of 306,522 samples in the training set and 11,535 samples in the test set. The average length of the news articles is 386 tokens, while the average length of the summaries is 28 tokens. This is a generation task with "low entropy" since there is abundant context and the content is restricted.

Task2: Scientific Abstract Generation

For scientific abstract generation, we parse the bibliography database (BibTeX) file with abstracts of ACL Anthology[3] in June, 2023. ACL Anthology hosts papers published at computational linguistics or natural language processing venues since 1965. We split the bibliography entries into the training and the test set, resulting in 40,512 samples for train and 2,132 samples for test. We use the title of the paper and the first sentence of the abstract as the prompt, asking the model to finish generating the rest of the abstract. The prompt (title + first sentence) is 42 tokens long on average, while the mean of the full abstract length is 145 tokens. This is a task of "medium entropy", the provided title

[3] `https://aclanthology.org`

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/.

Table 1: An example of model outputs over iterations. Since diversity is a holistic concept observed among the entire set of documents, it is non evident to demonstrate the loss of diversity through a single example. However, for the scientific abstract generation task in this example, there are some obvious patterns. Abstracts found in the ACL anthology often contain URLs linking to data, code or demos. After a few iterations of training on synthetic text, the model picked up this high probability event and starts to generate such links repeatedly in an unnatural manner.

already lays out the general idea of the paper and the first sentence provides a fair amount of context.

Task3: Story Generation

For story generation, we use the WritingPrompts dataset (Fan et al., 2018). It is made up of human written stories paired with writing prompts from Reddit's WritingPrompts forum. There are 272,600 samples in the training set and 15,138 samples in the test set. The writing prompts consist of 30 tokens on average and the resulting stories have a mean of 389 tokens. The prompts are generally short and in most cases do not contain a plot (i.e. narrative structure), making this a "high-entropy" generation task with limited context.

Decoding Strategy.

We use a combination of nucleus sampling (p) and temperature sampling (τ) to achieve nuanced control over the language model's outputs (Holtzman et al., 2020). Nucleus sampling, also known as top-p sampling, is used to generate text by selecting the most probable words from a distribution of words. It ensures that the cumulative probability of the chosen words exceeds a certain threshold (p). Higher values of p lead to more deterministic text. Temperature sampling involves dividing the output logits by a temperature parameter (τ) before sampling from the distribution. Higher values of τ make the distribution more uniform and increase randomness.

We adapt the specific parameters to the characteristic of each task (Amini et al., 2023). For news summarization, we emphasize precision and set $p = 0.1$, $\tau = 0.3$. For story generation, we care more about creativity and set $p = 0.9$, $\tau = 0.7$. For scientific abstract generation, we search something in between and set $p = 0.5$, $\tau = 0.5$. The `max_new_tokens` value is chosen according to the length of human-written references for each task: 50 for news summarization, 500 for story generation and 300 for scientific abstract generation. While the decoding strategy has influence over diversity metrics, it is not the determinant factor (Giulianelli et al., 2023). We aim to draw generalizable conclusions by experimenting across three distinct sets of decoding strategies.

	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 (↓)

Table 2: Perplexity (PPL) and linguistic diversity metrics for texts generated over different iterations (iter). All diversity metrics range from 0 to 1 and are reported as percentages (cosine distances are halved to maintain this range). Typical values are suggested by the results obtained on human written texts. The arrows in parentheses indicate the direction of variation compared to the previous iteration. In the case of iteration 1, the comparison is against human reference text.

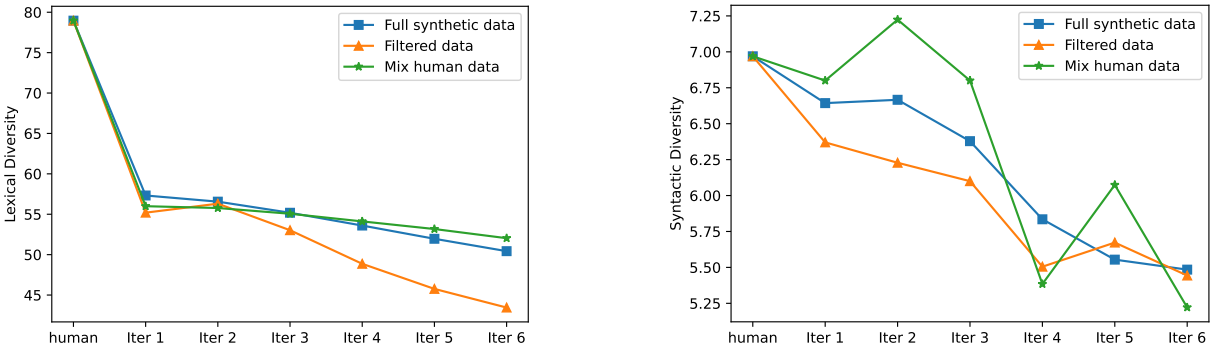
4.2 Results

In Table 2, we display the perplexity and linguistic diversity metrics for texts generated across various iterations. We also show an example of generated texts across iterations in Table 1.

Our findings indicate that the *perplexity* values fall within an acceptable range (< 20) (Holtzman et al., 2020), indicating that models effectively assimilate training data and generate texts of a quality viable for diversity analysis. The decrease of perplexity over iterations suggest that the model might be more prone to over-fitting when trained on synthetic text, losing the tail of the original distribution (visualized in Appendix A, Figure 3). A general decline in the majority of *linguistic diversity* metrics underscores the pressing issue of diminishing linguistic diversity. We highlight some key observations in below.

The decline of diversity is greater for "high entropy" tasks. We deliberately select three generation tasks of varying "entropy", which is reflected by the amount and nature of given context, i.e., constraint. News summarization involves a lengthy context with the summary confined to a very limited space, whereas story generation is characterized by brief prompts and a vast array of potential narrative directions. In the "highest entropy" task, story generation, the gap in linguistic diversity between human-written and model-generated texts is the most pronounced, and the decline over iterations is the fastest. The significant gap between humans and models is expected, given that story generation demands substantial creativity, a domain where language models are known to fall short (Chakrabarty et al., 2023). The rapid decrease in diversity can also be explained by the creative nature of the task. Models initially learn from diverse original human-written stories but suffer greatly when later exposed solely to synthetic data, which already exhibits a notable loss in diversity.

Even for "lower entropy" tasks, training on synthetic texts will eventually lead to vanishing diversity. In tasks like news summarization and scientific abstract generation, which have "lower entropy" compared to story generation, there is still a noticeable decrease in linguistic diversity over iterations. Consider the task of generating scientific abstracts: initially, the syntactic diversity in texts created by Models (2) and (3) shows an increase compared to those written by humans. This might be because scientific abstracts inherently possess less varied syntactic structures than the broader range of texts in the pretraining data



(a) Lexical diversity.

(b) Syntactic diversity.

Figure 2: Illustration of linguistic diversity variation for the story generation task under different recursion settings. Since there is a strong correlation between different diversity metrics of the same aspect, we only report one per aspect: Distinct-3 for lexical diversity and D_syn_c for syntactic diversity.

of OPT-350M. However, as the iterations advance, the syntactic diversity scores of the texts produced by the models eventually decline, dropping below those of human-written abstracts. This trend might be partly attributed to catastrophic forgetting (McCloskey and Cohen, 1989). Additionally, while human-written abstracts may have limited syntactic diversity, their structure is markedly different from the pretraining data, thus introducing new learning elements for the model. In contrast, the synthetic data produced by Model (2), despite its marginally higher internal syntactic diversity, closely mirrors the model's own training distribution. This lack of novel information leads to a subsequent reduction in variation.

Syntactic diversity suffers remarkably. We notice that syntactic diversity manifests a decreasing trend, especially for creative tasks, comparable to the decline in lexical diversity and to a greater extent than in semantic diversity (also visualized in Appendix A, Figure 4 and Figure 5). While the reduction in lexical diversity is well-researched and somewhat anticipated, our study is the first to highlight the decrease in syntactic diversity. Syntax is more implicit but equally important as vocabulary in maintaining linguistic richness. The important yet overlooked decline in syntactic diversity emphasizes the need for future NLG research to include syntactic diversity measurements alongside the commonly reported lexical diversity metrics.

Semantic diversity is the most stable. Semantic diversity remains more stable compared to lexical diversity and syntactic diversity. We believe that synthetic training data have more impact on the diversity of form than the diversity of content. The main issue in the semantic aspect is coherence rather than diversity. We find that the generated texts remain rather diverse in meaning throughout the iterations whereas the coherence between sentences drops. This finding also corresponds to the well-known fact that language models are prone to hallucination (Huang et al., 2023b).

4.3 Playing with Recursion Settings

We perform further analysis to understand how different factors influence outputs of the recursive tuning-generation process. We introduce two settings to approximate the real-life scenario. We focus our analysis on the story generation task as it shows the most pronounced diversity decline.

Filtering Synthetic Data. Instead of using the full set of synthetic samples, it is a common choice to filter out invalid samples before training (Wang et al., 2023). In our case, we use a *linguistic acceptability* filter to discard the noisy samples generated in each iteration. The filter is a RoBERTa model [4] (Morris et al., 2020) trained on the COLA corpus (Warstadt et al., 2019). We do generation on the full training set and discard the 20% of synthetic data with the lowest linguistic acceptability score before using them to train the next iteration's model.

Mixing Fresh Human Data. To approximate the most realistic scenario, we consider mixing human data with synthetic data for training. We separate the training data into a 40% set reserved for synthetic data generation and a 60% set used only as human data. The 60% set of human data is further split into six subsets each containing 10% of

[4] textattack/roberta-base-CoLA

the human data. For each of the 6 training iterations, one of these six 10% subsets are mixed into the synthetic data as "fresh human data". These data are considered "fresh" because they were held out at the beginning and not seen by models from previous iterations.

The results are displayed in Figure 2. We do not show semantic diversity as it remains relatively stable across all settings. The introduction of fresh human data only minimally mitigates the observed decrease, while filtering significantly amplifies the decline. This outcome is unsurprising, as quality filters typically favor more common and less inventive samples. Consequently, in practice, the linguistic diversity decline might be even more substantial than suggested by our previous findings.

5 Conclusion

Our study provides critical insights into the implications of recursively training language models on synthetic data generated by their predecessors. Through our innovative approach, focusing on linguistic diversity measures rather than traditional performance metrics, across various NLG tasks, we have uncovered a noticeable reduction in lexical and syntactic diversity in language model outputs over successive iterations of recursive training on synthetic text. These findings highlight a concerning trend: as language models increasingly rely on predecessor-generated text for training, there is a tangible risk of diminishing linguistic richness and variety in their outputs. Our research underscores the necessity for a more nuanced and forward-thinking approach in the development of language models, emphasizing the importance of preserving linguistic diversity alongside improving technical performance.

Limitations

Language diversity. Our work investigates linguistic diversity in a monolingual context. Our experiments are exclusively conducted in the English language. While the main research idea is readily adaptable, the specific methodologies require adjustments when applied to other languages. It's worth noting that our linguistic diversity metrics may not perform optimally for languages apart from English. These metrics rely on language-specific tokenization/segmentation, dependency parsing, and sentence embeddings, which pose challenges for languages with limited resources. However, it would be interesting future work to overcome these obstacles and investigate linguistic diversity in a multilingual setting.

Resource constraint. Due to resource limitations, we could not perform experiments on an extensive range of models. We opted for the moderately large decoder-only model OPT-350M, striking a balance between generation quality and parameter scale. Our analysis involves recursive model training across six iterations, for three tasks and under various settings, demanding significant computational resources. For instance, completing all six iterations for the story generation task under the full synthetic setting alone consumes approximately 700 GPU hours on the NVIDIA RTX A6000 48G GPU. In this study, our primary focus is on comparing different tasks and settings rather than across various models. Nevertheless, we anticipate that the decline in linguistic diversity is a recurring phenomenon in different language models. In future research, we intend to explore quantization and parameter-efficient fine-tuning approaches with larger-scale language models.

Realistic Web Setting. Our paper is partially motivated by the fact that LMs are trained on web content that increasingly contains synthetic text. However, after careful considerations, it is impossible to conduct experiments under a realistic web setting. To simulate a realistic setting, we would need a dataset of synthetic text posted on the web by real users. We initially thought about using data from ShareGPT where users upload their conversations with ChatGPT but then realized that this would pose copyright issues. It is thus not feasible to construct a realistic dataset for unconditional language modeling with synthetic content. In addition, there currently exists no algorithm that can reliably detect LM generated text and we cannot estimate the amount of synthetic information online. We have already proposed experiments with mixed settings, which demonstrated that the reduction in diversity only marginally lowered when mixing in a fixed percentage of human data. It would be interesting to conduct further experiments with varying combinations, potentially increasing the proportion of human data. Nevertheless, we anticipate that our research findings will continue to hold, given the minimal attenuation observed when experimenting with the current mixed setting.

Ethical Considerations

Usage of scientific artifacts. We employ three datasets in our research: XL-SUM (Hasan et al., 2021), ACL Anthology[5], and WritingPrompts (Fan et al., 2018). XL-SUM and ACL Anthology are made available under the CC BY-NC-SA 4.0 license, while the WritingPrompts dataset is distributed under the MIT license. None of these datasets contains any information that can be linked to private individuals in a harmful way. Furthermore, we utilize the OPT model (Zhang et al., 2022), which is subject to the "OPT-175B License Agreement"[6]. Our use of these resources aligns with their designated research purposes.

Potential risks. Our research focuses on the analysis of language models and is not specifically linked to any particular application. Its positive social impact lies in identifying and bringing to light overlooked issues in the usage of language models, thereby alerting both developers and users to exercise more deliberate considerations. However, it's important to recognize that there may be potential risks arising from the way our findings are interpreted by the general public, especially if they are exaggerated or overgeneralized. We want to stress that our conclusions are rigorously validated based on specific datasets and within a particular context. It is necessary to explicitly acknowledge these limitations when discussing our research during scientific dissemination.

Acknowledgements

We thank Dr. Julie Hunter for her insights during the preliminary discussions of this work. We also thank the anonymous reviewers for their remarks and suggestions. This research has been partially supported by the ANR-TSIA HELAS chair and the ANR-23-CE23-0033-01 SINNet project.

References

Arshiya Aggarwal, Jiao Sun, and Nanyun Peng. 2022. Towards robust NLG bias evaluation with syntactically-diverse prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6022–6032, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[5] https://aclanthology.org

[6] https://github.com/facebookresearch/metaseq/blob/main/projects/OPT/MODEL_LICENSE.md

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. 2024. Self-consuming generative models go MAD. In *The Twelfth International Conference on Learning Representations*.

Afra Amini, Ryan Cotterell, John Hewitt, Luca Malagutti, Clara Meister, and Tiago Pimentel. 2023. Generating text from language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 27–31, Toronto, Canada. Association for Computational Linguistics.

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. 2021. {MIROSTAT}: A {neural} {text} {decoding} {algorithm} {that} {directly} {controls} {perplexity}. In *International Conference on Learning Representations*.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2023. Art or artifice? large language models and the false promise of creativity. *arXiv preprint arXiv:2309.14556*.

Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6848–6863, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

Bastien De Clercq and Alex Housen. 2017. A cross-linguistic perspective on syntactic complexity in l2 development: Syntactic elaboration and diversity. *The Modern Language Journal*, 101(2):315–334.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. 2023. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Susan Edwards and Roelien Bastiaanse. 1998. Diversity in the lexical and syntactic abilities of fluent aphasic speakers. *Aphasiology*, 12(2):99–117.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings*

of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Yanzhu Guo, Chloé Clavel, Moussa Kamal Eddine, and Michalis Vazirgiannis. 2022. Questioning the validity of summarization datasets and improving their factual consistency. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5716–5727, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Seungju Han, Beomsu Kim, and Buru Chang. 2022. Measuring and improving semantic diversity of dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 934–950, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Chadi Helwe, Chloé Clavel, and Fabian M. Suchanek. 2021. Reasoning with transformer-based models: Deep learning, but shallow reasoning. In *3rd Conference on Automated Knowledge Base Construction*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Kuan-Hao Huang, Varun Iyer, I-Hung Hsu, Anoop Kumar, Kai-Wei Chang, and Aram Galstyan. 2023a. ParaAMR: A large-scale syntactically diverse paraphrase dataset by AMR back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8047–8061, Toronto, Canada. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.

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

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Tshilidzi Marwala, Eleonore Fournier-Tombs, and Serge Stinckwich. 2023. The use of synthetic data to train ai models: Opportunities and risks for sustainable development. *arXiv preprint arXiv:2309.00652*.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower, editor, *Psychology of learning and motivation*, volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

J.F. Miller. 1981. *Assessing Language Production in Children: Experimental Procedures*. Assessing communicative behavior. University Park Press.

Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. AmbiPun: Generating humorous puns with ambiguous context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1053–1062, Seattle, United States. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

Enrico Palumbo, Andrea Mezzalira, Cristina Marco, Alessandro Manzotti, and Daniele Amberti. 2020. Semantic diversity for natural language understanding evaluation in dialog systems. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 44–49, Online. International Committee on Computational Linguistics.

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

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2024. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores.

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. 2011. Weisfeiler-lehman graph kernels. *J. Mach. Learn. Res.*, 12(null):2539–2561.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget.

Giannis Siglidis, Giannis Nikolentzos, Stratis Limnios, Christos Giatsidis, Konstantinos Skianis, and Michalis Vazirgiannis. 2020. Grakel: A graph kernel library in python. *Journal of Machine Learning Research*, 21(54):1–5.

Katherine Stasaski and Marti Hearst. 2022. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–98, Seattle, United States. Association for Computational Linguistics.

Mildred C. Templin. 1957. *Certain Language Skills in Children: Their Development and Interrelationships*, ned - new edition edition, volume 26. University of Minnesota Press.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar Sigurdsson, Chenyang Tao, Wenbo Zhao, Tagyoung Chung, Jing Huang, and Nanyun Peng. 2023. Unsupervised melody-to-lyrics generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9235–9254, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. 2023. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 163–174.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Denis Yarats and Mike Lewis. 2018. Hierarchical text generation and planning for strategic dialogue. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5591–5599. PMLR.

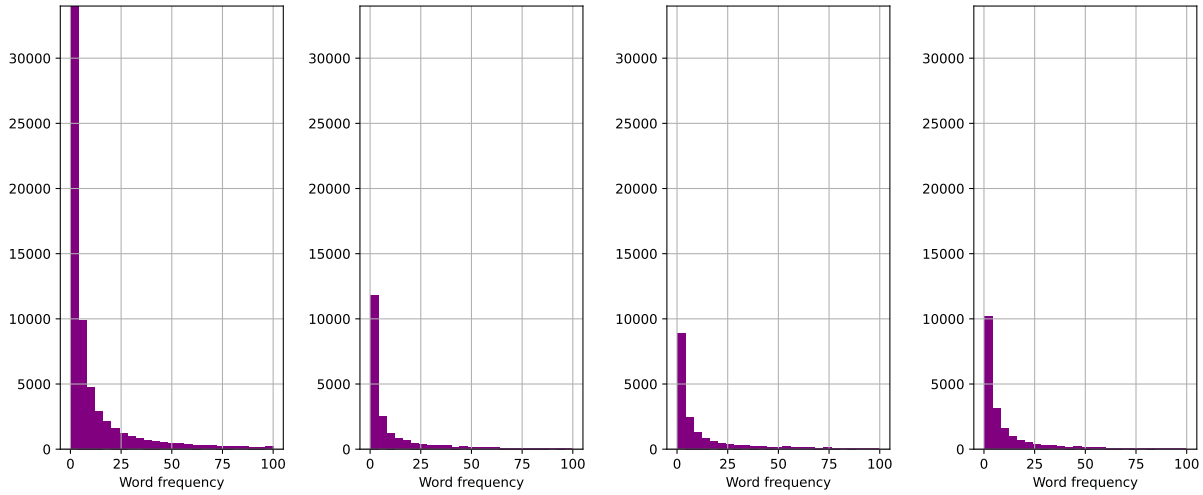
Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

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, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Weixiao Zhou, Gengyao Li, Xianfu Cheng, Xinnian Liang, Junnan Zhu, Feifei Zhai, and Zhoujun Li. 2023. Multi-stage pre-training enhanced by chatgpt for multi-scenario multi-domain dialogue summarization. *arXiv preprint arXiv:2310.10285*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Visualization of Diversity Metrics



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

Figure 3: Histograms illustrating word frequency in texts produced across various iterations for the story generation task. For visual clarity, the x-axis, representing word frequency, is truncated at 100, though the actual distribution extends further. A noticeable trend is the diminishing presence of low-frequency, "unique" words in the synthetic text relative to human-generated text, a pattern that intensifies with each iteration. This trend highlights a progressive decline of lexical diversity in the generated text.

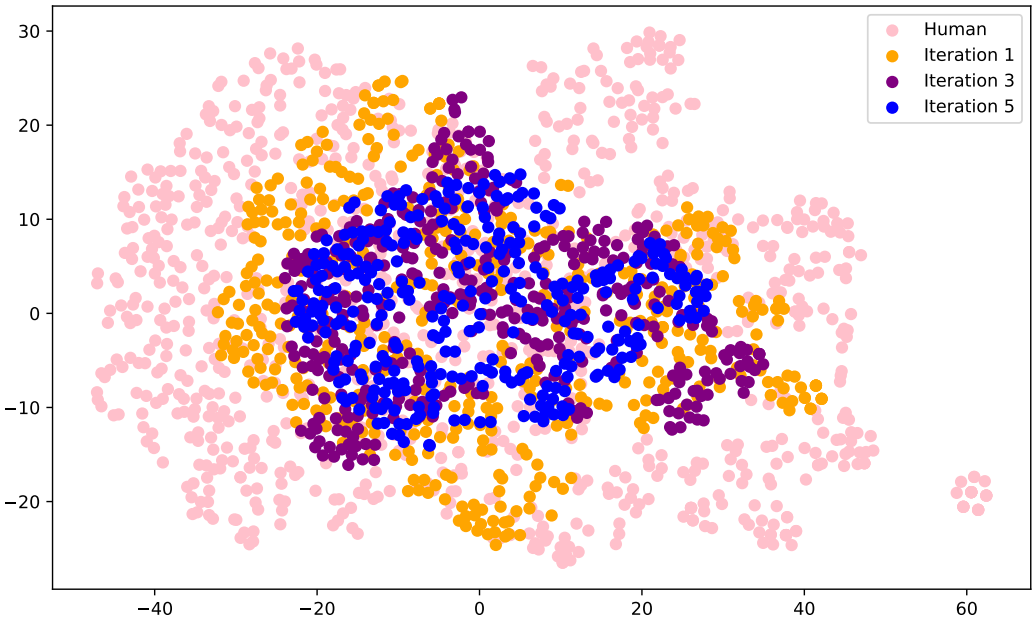


Figure 4: T-SNE visualization of dependency tree embeddings derived from sentences generated in successive iterations of our tuning-generation process. The visualization clearly depicts how, over time, the spatial distribution of the embeddings becomes increasingly compact. This decreasing spread is indicative of declining syntactic diversity.

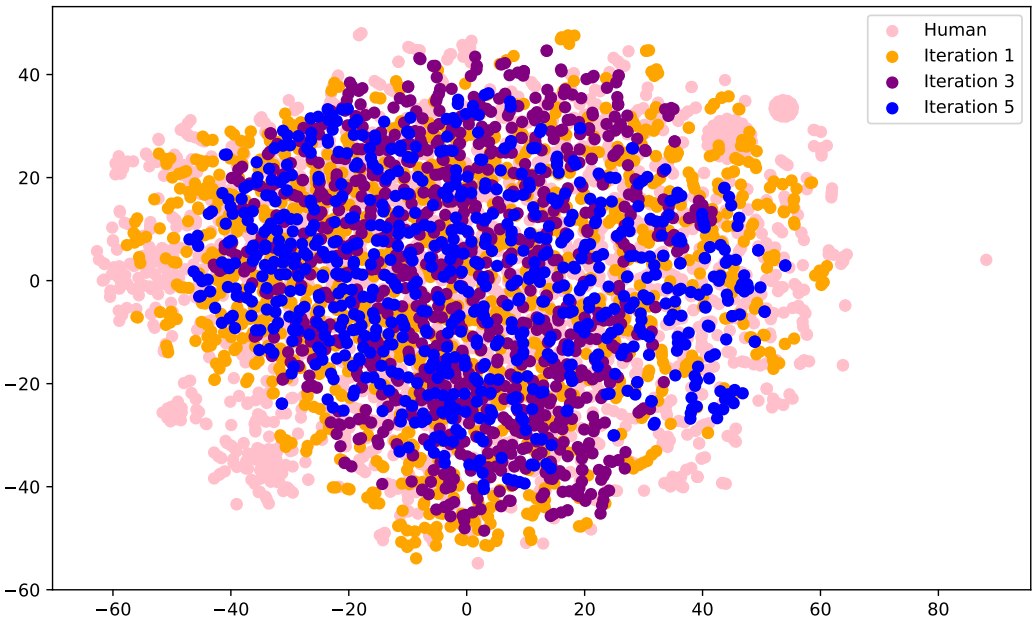


Figure 5: T-SNE visualization of sentence embeddings from text generated across different iterations. There is a noticeable decrease in dispersion over iterations, indicating a reduction in semantic diversity, though this change is less pronounced compared to that of syntactic diversity.

B Implementation of Diversity Metrics

B.1 Preprocessing

We apply preprocessing to the generated texts before computing the diversity metrics. For all three tasks, we remove the prompts from the generated texts. We remove <newline> tokens for story generation and replace URL links with WEBSITE for scientific abstract generation. We remove all punctuation marks for the calculation of lexical diversity metrics, but not for semantic diversity or syntactic diversity.

B.2 Lexical Diversity Metrics

Iteration	**Human**	1	2	3	4	5	6
News Summarization	18.89	18.21	18.42	18.75	19.01	19.25	19.71
Scientific Abstract Generation	49.35	48.80	49.61	49.60	49.64	49.60	49.50
Story Generation	148.77	149.77	149.92	149.92	149.91	149.96	149.84

Table 3: Average text lengths for post-truncation generations. Text lengths are measured by the number of words.

The distinct-n metric varies as a function of text length, so we compute results at fixed lengths. We apply truncation to the generated texts, using different thresholds for each task: 20 for news summarization, 50 for scientific abstract generation and 150 for story generation. The average text lengths after truncation are presented in Table 3. We observe that all lengths consistently fall within a narrow range, allowing for fair comparison of distinct-n results across iterations.

For Self-BLEU, we use a publicly available implementation[7]. We take the mean value of Self-BLEU-2 and Self-BLEU-3.

B.3 Semantic Diversity Metrics

For sentence splitting, we use the NLTK sentence tokenizer. For Sentence-BERT, we use the all-mpnet-base-v2 model on huggingface. For the pairwise cosine distances, we randomly draw 2000 sentences for

[7] https://github.com/Danial-Alh/fast-bleu

each calculation and report the mean value over 5 randomizations.

B.4 Syntactic Diversity Metrics

We construct the dependency graphs with the Stanza Dependency Parser[8]. We employ a publicly available implementation of the Weisfeiler-Lehman graph kernel[9]. We set the number of iterations to 2. The pairwise cosine distances are calculated in the same way as for semantic diversity.

[8] https://stanfordnlp.github.io/stanza/depparse.html
[9] https://github.com/ysig/GraKeL