PARAMETERIZED SYNTHETIC TEXT GENERATION WITH SIMPLESTORIES

Lennart Finke *
ETH Zürich
lfinke@ethz.ch

Thomas Dooms
University of Antwerp

Mat Allen
Dioptra

Juan Diego Rodriguez
The University of Texas at Austin

Noa Nabeshima
Independent

Dan Braun
Apollo Research

ABSTRACT

We present SimpleStories, a large synthetic story dataset in simple language, consisting of 2 million stories each in English and Japanese. Our method employs parametrization of prompts with features at multiple levels of abstraction, allowing for systematic control over story characteristics to ensure broad syntactic and semantic diversity. Building on and addressing limitations in the TinyStories dataset, our approach demonstrates that simplicity and variety can be achieved simultaneously in synthetic text generation at scale.

1 INTRODUCTION

The TinyStories dataset of Eldan & Li (2023) has greatly aided the progress towards a mechanistic understanding of LLMs through the creation of small model organism transformers trained on synthetic children's stories. Its stated research objective is to distill the concepts of grammar and reasoning into a text corpus by abstracting away factual knowledge — an idea that remains highly relevant as part of a broader discussion in the context of data-constrained training (Villalobos et al., 2022). Here, we aim to learn from TinyStories as the current standard in small language model training, and to produce a dataset applying these lessons. We see potential for improvement, despite it being a significant advancement in synthetic text generation, because the TinyStories dataset is

- formulaic; to illustrate, 59% of stories contain the string 'Once upon a time' verbatim.
- semantically limited to children's stories as opposed to simple language generally.
- unlabeled, which hinders e.g. finetuning.

Additional challenges include that it is originally only available in English, not entirely open-source, and troubled with encoding artifacts, duplications, as well as graphic descriptions unsuitable for the children's story setting. (We do not provide quotations here, but searching the dataset for terms such as 'death' yields such samples.) We further aim to find a more cost-effective text generation process.

2 METHODS

Like TinyStories, we generate our dataset using commercial LLMs with template prompts that elicit simple language (full prompt in Section A.1). We use GPT-4o-mini-2024-07-18, which offers improved capabilities and alignment compared to the GPT-3.5 and GPT-4 models used in TinyStories. We are faced with the challenge of lexical diversity in LLM-based text generation—where certain phrases and expressions remain overrepresented even at high sampling temperatures—and therefore constrain each story to begin with a specific part of speech (adjective, adverb, noun, or preposition) and initial letter, with letter frequencies drawn from a reference corpus. Content diversity and labeling can be solved together through the following procedure. Instead of prompting for a selection of

*Corresponding author.

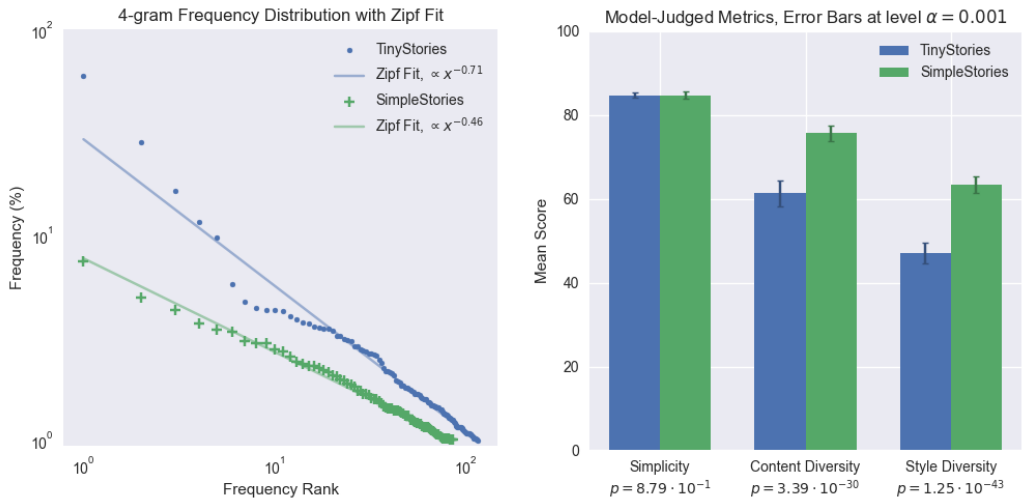

Figure 1: On the left, an evaluation of syntactic diversity through 4-gram frequencies by frequency rank, on a subsample of 10% of either dataset. On the right, an evaluation of semantic diversity and simplicity through model-as-a-judge. The 99.9% confidence intervals assume quantiles of a normal distribution with sample mean and variance. The uncorrected p-values each represent the one-way ANOVA with $N = 200$ and the null hypotheses that the scores of both datasets have equal mean.

common words that should be used in the completion, we specify a topic, an overarching theme or feel, a writing style and a narrative feature; for the full list see Section A.1. To allow the study of phenomena relevant to alignment, we take care to represent potentially useful concepts such as "Co-operation", "Betrayal" or "Long-Term Thinking". On a subset of samples, we additionally prompt for a certain grammar feature, or ask the LLM to assume an archetypal author persona. These categories are applicable across many languages, but their content may not be, and thus one should balance an international perspective with language-specific tropes and traditions. We generate multiple stories at a time with the same parameters, which can reduce costs by saving on input tokens and lead to variations over an underlying idea, i.e. different instantiations of the same story structure. Together with the introduction of entropy through initial parts of speech and letter constraints, this disambiguates generations from the first token onwards, even across millions of stories. We can therefore use Nucleus Sampling (Holtzman et al., 2019) with $p = 0.9$, increasing adherence to the given constraints.

Focusing specifically on the training of interpretable language models, we anticipate demand for word-level tokenization and guide the generation to use only a limited vocabulary. A challenge in this is that the generating model will typically use proper names, some of which are imaginary. We prompt against this by instructing to compose proper names from common words and to use names from a given list. Finally, looking for issues with unsuitable content, we found that better alignment of production language models has fortunately led to an absence of e.g. violent stories, as far as we could determine with manual keyword searches. Thus, despite anticipating the need to filter our dataset, we do not. To help with filtering and fine-grained usage, we precompute relevant NLP metrics such as word count and Flesch-Kincaid reading grade (Kincaid, 1975).

Our dataset is openly available at `https://huggingface.co/datasets/lennart-finke/SimpleStories`, with generation code at `https://github.com/lennart-finke/simple_stories_generate` and an interactive dataset visualization using text embeddings at `https://fi-le.net/simplestories`, containing ca. 2 million samples per language. The designated test set contains ca. 20 thousand samples. Unlike the original TinyStories work, our story generation and model training code will be open-sourced for people to create variations of the datasets and model architectures.

3 RESULTS

To evaluate our dataset, we probe its syntactic and semantic diversity in English through a comparison with TinyStories. In doing so, the methods that were used in the original TinyStories work are a natural choice.

Syntactic Diversity We compute the most common n-grams in a random 10% subsample of either dataset, and greedily filter a descending frequency-sorted list for n-grams which do not overlap with a previous n-gram on more than $n-2$ words. Through this, we can not only find the most common phrases, but also analyze the distribution of their frequencies, i.e. the percentage with an n-gram occurs in one sample. They approximately follow a Zipf distribution, echoing empirical findings in large natural corpora (Ha et al., 2009). As seen in Fig. 1, SimpleStories has a much more varied distribution of 4-grams, despite the fact that it has longer samples with an average and standard deviation of 224.6 ± 103.8 as opposed to TinyStories' 175.4 ± 80.2. The highly common outliers in TinyStories stem from largely identical first sentences of different samples, which we were able to avoid through the above prompting procedure. Table 1 shows the most frequent 4-grams. Calculating the Flesch-Kincaid reading grade over the whole dataset, we find a mean and standard deviation of 3.08 ± 1.24 for SimpleStories as opposed to TinyStories' 2.55 ± 1.49.

Semantic Diversity Once more, we find inspiration in Eldan and Li, by using a variant of their GPT-Eval (what has since been termed model-as-a-judge) to compare the semantic diversity of our dataset. We think of our text synthesis problem as constraint optimization — while keeping stories simple, produce as much variation in content and style as possible. We therefore instruct GPT-4o-mini to evaluate simplicity, content diversity and style diversity given 4 randomly drawn stories from the same dataset as a number from 0 to 100. Before this, we ask for an explanation that does not enter our evaluation to induce the accuracy gains that typically result from chain-of-thought (Wei et al., 2022). We perform one-way ANOVA on the resulting scores with the null hypothesis that both datasets are identical in all three metrics, with a total of $N = 200$ measurement units (therefore 800 stories). As seen in Fig. 1, the model-perceived diversity for our dataset is much greater, with an insignificant model-perceived difference in simplicity.

Labeling Our method produces labels for each sampled story, but it is not clear a priori whether these convey meaningful information. Consequently, we test this by prompting a judge model (again GPT-4o-mini with chain-of-thought) to recover those labels that are present for every data point, given the list of all possible values in the dataset. Using $N = 200$ samples, we reject the null that this process is no better than random guessing for each label, at $\alpha = 0.001$. The accuracy for recovering the topic, 0.49, is particularly good, whereas the more abstract labels "theme", "style" and "narrative feature" are at $0.225, 0.145, 0.155$, as seen in Fig. 2. We note that no further elicitation was used here, so these figures should be interpreted as only a lower bound on the true label quality.

4 OUTLOOK

Based on the above, we recommend our English dataset over TinyStories for training small language models. SimpleStories presents a more challenging language modeling problem that calls for application of logic and theory of mind in various settings, while staying firmly in the realm of simple language and fiction.

We seek to train a suite of small language models ourselves, and encourage communal participation. Our main objective here is investigating the benefits of more varied, labeled training data for language model interpretability. We also hope to be able to offer more languages in the future, insofar as additional funds are available.

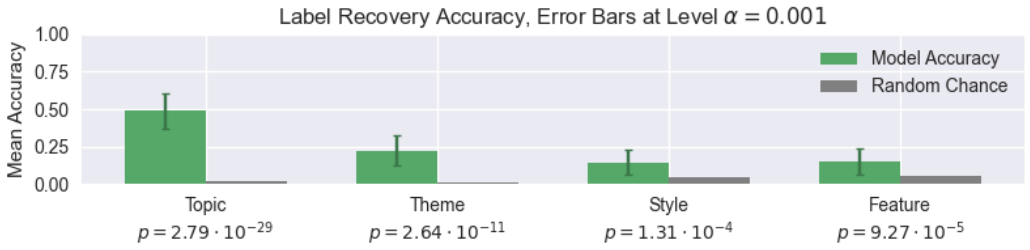

Figure 2: Evaluation of label quality by accuracy of a judge model, GPT-4o-mini, annotating the story text versus the labels stemming from the generation process. The uncorrected p-values are from one-sample t-tests with $N = 200$ and the null that the accuracy is no different from random guessing.

ACKNOWLEDGEMENTS

We thank Nix Goldowsky-Dill, Rylan Schaeffer, Joseph Bloom, Joseph Miller and Alice Rigg for valuable feedback and insight into the wanted specifications for this dataset. This research was funded Apollo Research and Anthropic.

REFERENCES

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

Le Quan Ha, Philip Hanna, Ji Ming, and Francis Jack Smith. Extending zipf's law to n-grams for large corpora. *Artificial Intelligence Review*, 32:101–113, 2009.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

JP Kincaid. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Chief of Naval Technical Training*, 1975.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*, 1, 2022.

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

A APPENDIX

A.1 STORY GENERATION

Below is the story generation prompt for SimpleStories, with values of changing parameters in {curly brackets}. A grammar feature is used in 50% of stories, an author persona is specified in 33% of samples. Paragraph counts are uniformly distributed between 1 and 9, the number of stories in one completion is inversely proportional to this.

> Write {12} short stories ({2} paragraphs each) using very basic words. Do not number each story or write a headline. Make the stories diverse by fully exploring the theme, but each story should be self-contained. Separate the stories by putting {"The End."} in between. Make the stories as qualitatively distinct to each other as possible. In particular, never start two stories the

same way! Each story should be about {Responsibility}, include {secret societies}, be {lyric} in its writing style and ideally feature {inner monologue}. The most important thing is to write an engaging easy story, but where it makes sense, demonstrate the use of {progressive aspect}. Write from the perspective of {someone curious}. If you need to use proper names, make them from space-separated common words. Either don't give characters a name, or select from {list of names}. Complex story structure is great, but please remember to only use very simple words! If you can, start the story with {a noun} that begins with the letter {p}.

The parameters stem from the following set of options.

Theme: Friendship, Courage, Contradiction, Coming of age, Kindness, Amnesia, Adventure, Imagination, Family, Perseverance, Curiosity, Honesty, Romance, Teamwork, Responsibility, Strategy, Magic, Discovery, Betrayal, Deception, Generosity, Creativity, Self-Acceptance, Helping Others, Hardship, Agency, Power, Revenge, Independence, Problem-Solving, Resourcefulness, Long-Term Thinking, Optimism, Humor, Love, The Five Senses, Tradition, Innovation, Hope, Dreams, Belonging, Travel, Overcoming, Trust, Morality, Happiness, Consciousness, Failure, Conflict, Cooperation, Growth, Loss, Celebration, Transformation, Scheming, Challenge, Planning, Wonder, Surprises, Conscience, Intelligence, Logic, Resilience.

Topic: talking animals, fantasy worlds, time travel, a deadline or time limit, space exploration, mystical creatures, underwater adventures, dinosaurs, pirates, superheroes, fairy tales, outer space, hidden treasures, magical lands, enchanted forests, secret societies, robots and technology, sports, school life, holidays, cultural traditions, magical objects, lost civilizations, subterranean worlds, bygone eras, invisibility, giant creatures, miniature worlds, alien encounters, haunted places, shape-shifting, island adventures, unusual vehicles, undercover missions, dream worlds, virtual worlds, riddles, sibling rivalry, treasure hunts, snowy adventures, seasonal changes, mysterious maps, royal kingdoms, living objects, gardens, lost cities, the arts, the sky

Style: whimsical, playful, epic, fairy tale-like, modern, classic, lyric, mythological, lighthearted, adventurous, heartwarming, humorous, mystical, action-packed, fable-like, surreal, philosophical, melancholic, noir, romantic, tragic, minimalist, suspenseful

Narrative Feature: dialogue, in medias res, a moral lesson, absence indicating a presence, a story told through letters, a twist ending, an unreliable narrator, foreshadowing, irony, inner monologue, symbolism, a MacGuffin, a non-linear timeline, a reverse timeline, circular narrative structure, a flashback, a nested structure, a story within a story, a Red Herring, multiple perspectives, Checkhov's gun, the fourth wall, a cliffhanger, an anti-hero, juxtaposition, climactic structure

Grammar Feature: present tense, past tense, future tense, progressive aspect, perfect aspect, passive voice, conditional mood, imperative mood, indicative mood, relative clauses, prepositional phrases, indirect speech, exclamatory sentences, comparative forms, superlative forms, subordinate clauses, ellipsis, anaphora, cataphora, wh-questions, yes-no questions, gerunds, participle phrases, inverted sentences, non-finite clauses, determiners, quantifiers, adjective order, parallel structure, discourse markers, appositive phrases

Author Persona: an explorer archetype, a rebellious author, a powerful leader, a wise, old person who wants to teach the young, an innocent author, a moralistic teacher, a hopeless romantic, a hurt, ill-intentioned person, an academic, a jester archetype, a poet, a philosopher, a mother, a father, someone curious, someone evil, someone who wants to prove a point, a child, a pedant, the everyman, the oppressed, a cruel person, someone who loves order and structure

Table 1: Top 20 most frequent 4-grams in TinyStories and SimpleStories Datasets, filtered for overlaps of more than two words.

TinyStories		**SimpleStories**	
Frequency	Phrase	Frequency	Phrase
59.38%	once upon a time	7.49%	took a deep breath
28.24%	there was a little	4.95%	the sun began to
16.52%	a little girl named	4.32%	from that day on
11.55%	a time there was	3.71%	felt a spark of
9.73%	from that day on	3.46%	felt a rush of
5.79%	a little boy named	3.40%	as the sun set
4.77%	to play in the	3.06%	felt the weight of
4.45%	was so happy and	2.98%	with a deep breath
4.32%	do you want to	2.97%	a girl named mia
4.30%	went to the park	2.77%	thought for a moment
4.28%	she loved to play	2.71%	a boy named leo
4.01%	to play with her	2.56%	the end of the
3.90%	the little girl was	2.42%	was not just a
3.75%	did not want to	2.37%	the day of the
3.73%	it was time to	2.31%	felt a sense of
3.57%	but it was too	2.30%	laughter filled the air
3.53%	there was a big	2.25%	it was time to
3.51%	was so happy that	2.20%	a boy named samuel
3.49%	there was a boy	2.15%	at the edge of
3.42%	they like to play	2.09%	the magic of the