SimpleStories

Improved Synthetic Stories for Training Interpretable Language Models

🤗 **Dataset** - **Models**
Model Training Repo
Story Generation Repo
Dataset Paper

What?

The SimpleStories project is a tiny model suite trained on our own synthetic text dataset. The process is fully open-source, and you are welcome to repurpose and improve every part of it, or even join in on Slack. It's better than TinyStories because it's more diverse, and labeled.

Your Feedback

We think the main impact here will stem from interpretability done on the small models. What would you like us to do next? Comment or edit below, with or without your name.

- **[Done]** **Tiny Stories Models which vary in context length.** Tiny Stories are actually too short! Much interesting behaviour likely comes from longer contexts. I'd be curious about whether you can make several tiny stories which vary simply with respect to the context size. This can help us answer questions like "What features are unique to longer prompts?" or "How much do I miss because I generate SAEs on activations with a shorter context length than the full model". (from Joseph Bloom)
- I'd love to see you try to train architectures without layernorm. I'm not sure it would work but could make interp easier. (from Rick Goldstein)
- **[Done]** Having the embed tok matrix have only 10K entries (as opposed to 50K with some not being used) would be nice. (from Rick Goldstein)
- **Your feedback here!**

Models

	n_params	n_layers	d_model	n_heads	n_ctx	d_vocab
SimpleStories-35M	35M	12	512	8	512	4096
SimpleStories-30M	30M	10	512	8	512	4096
SimpleStories-11M	11	6	384	6	512	4096
SimpleStories-5M	5	6	256	4	512	4096
SimpleStories-1.25M	1.25	4	128	4	512	4096

What problem are we solving?

TinyStories is a well received dataset of ca. 2M model-generated short stories in simple English. It demonstrates that even small language models (= 30M parameters) trained only on it can produce coherent writing, which is useful for interpretability. The dataset offers opportunities for improvement:

- The stories follow a small set of formulas. This can be useful for interp because it presumably simplifies the trained model, but it is only a low-order approximation to natural language in general. → We will generate more diverse narrative structures.
- It has quality issues such as:
    - It includes unintended non-standard characters (like “â€œ”) → Our improvement won’t. For a cleaned version of the original data, see Noa’s work here.
    - For some stories, there is a “bad ending” feature, which sometimes leads to inappropriate and disturbing content, because the story protagonists are often implied to be children. → We remove this feature and filter.
    - Some stories are duplicated.
    - The tokenizer uses the most common 10000 tokens from GPT-Neo → We retrain our own tokenizer with a smaller vocab size.
- TinyStories was generated by GPT 4 and 3.5 → We will use 4o-mini (and 4o, Claude if budget allows), presumably improving the outcome and allowing for more throughput.
    - In particular, the original contains some grammatical errors.

Who is working on this?
Currently Lennart Finke, Chandan Sreedhara, Thomas Dooms, Noa Nabeshima, Mat Allen and Dan Braun. And perhaps you, reading this!