The *TinyStories* Dataset: How Small Can Language Models Be And Still Speak Coherent English?

Ronen ELDAN

Microsoft Research, Redmond, USA

Email: roneneldan@microsoft.com

Language models (LMs) are powerful tools for natural language processing, but they often require large-scale and diverse corpora to produce coherent and fluent text. This raises the question of whether we can design a dataset that preserves the essential elements of natural language, such as grammar, vocabulary, facts, and reasoning, but that is much smaller and more refined in terms of its breadth and diversity. In this talk, we introduce *TinyStories*, a synthetic dataset of short stories that only contain words that a typical 3 to 4-year-olds usually understand, generated by GPT-3.5 and GPT-4. We show that *TinyStories* can be used to train and analyze language models that are much smaller than the state-of-the-art models (below 10 million parameters), or have much simpler architectures (with only one transformer block), yet still produce fluent and consistent stories with several paragraphs that are diverse and have almost perfect grammar, and demonstrate certain reasoning capabilities. We also show that the trained models are substantially more interpretable than larger ones, as we can visualize and analyze the attention and activation patterns of the models, and show how they relate to the generation process and the story content. We hope that *TinyStories* can facilitate the development, analysis and research of language models, especially for low-resource or specialized domains, and shed light on the emergence of language capabilities in LMs.