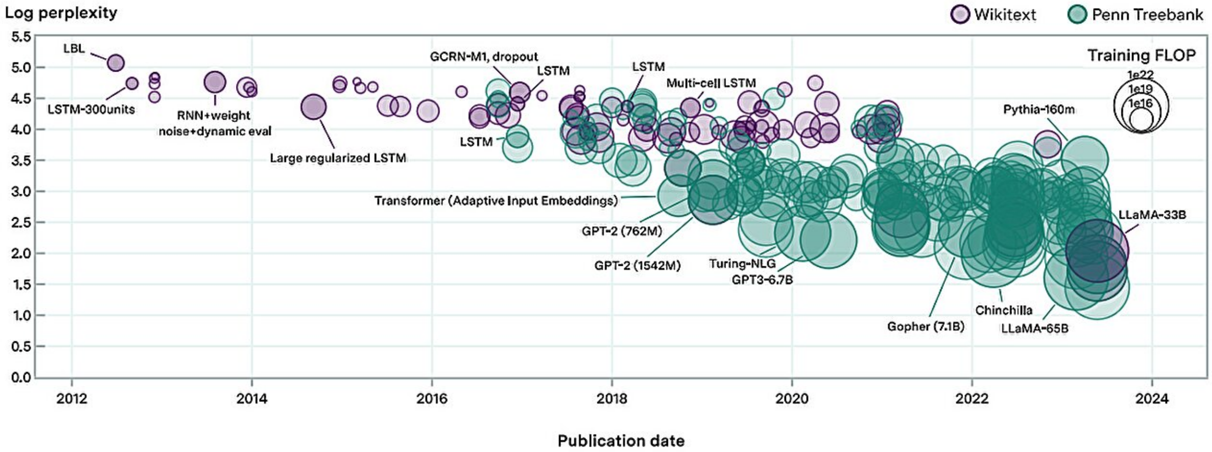
From recurrent networks to GPT-4: Measuring algorithmic progress in language models

March 13 2024, by Rachel Gordon



Log of perplexity of models used in our work, of over 231 language models analyzed in our work spanning over 8 orders of magnitude of compute, with each shape representing a model. The size of the shape is proportional to the compute used during training. Comparable perplexity evaluations are curated from the existing literature and from our own evaluations. Credit: *arXiv* (2024). https://arxiv.org/abs/2403.05812

In 2012, the best language models were small recurrent networks that struggled to form coherent sentences. Fast forward to today, and large language models like GPT-4 outperform most students on the SAT. How has this rapid progress been possible?

In a new paper posted to the *arXiv* preprint server, researchers from Epoch, MIT FutureTech, and Northeastern University set out to shed light on this question. Their research breaks down the drivers of progress in language models into two factors: scaling up the amount of compute used to train language models, and algorithmic innovations. In doing so, they perform the most extensive analysis of algorithmic progress in language models to date.

Their findings show that due to algorithmic improvements, the compute required to train a language model to a certain level of performance has been halving roughly every eight months. "This result is crucial for understanding both historical and future progress in language models," says Anson Ho, one of the two lead authors of the paper. "While scaling compute has been crucial, it's only part of the puzzle. To get the full picture you need to consider algorithmic progress as well."

The paper's methodology is inspired by "neural scaling laws": mathematical relationships that predict language model performance given certain quantities of compute, training data, or language model parameters. By compiling a dataset of more than 200 language models since 2012, the authors fit a modified neural scaling law that accounts for algorithmic improvements over time.

Based on this fitted model, the authors do a performance attribution analysis, finding that scaling compute has been more important than algorithmic innovations for improved performance in language modeling. In fact, they find that the relative importance of algorithmic improvements has decreased over time.

"This doesn't necessarily imply that algorithmic innovations have been slowing down," says Tamay Besiroglu, who also co-led the paper. "Our preferred explanation is that algorithmic progress has remained at a roughly constant rate, but compute has been scaled up substantially, making the former seem relatively less important."

The authors' calculations support this framing, where they find an acceleration in compute growth, but no evidence of a speedup or slowdown in algorithmic improvements.

By modifying the model slightly, they also quantified the significance of a key innovation in the history of machine learning: the Transformer, which has become the dominant language model architecture since its introduction in 2017. The authors find that the efficiency gains offered by the Transformer correspond to almost two years of algorithmic progress in the field, underscoring the significance of its invention.

While extensive, the study has several limitations. "One recurring issue we had was the lack of quality data, which can make the model hard to fit," says Ho. "Our approach also doesn't measure algorithmic progress on downstream tasks like coding and math problems, which language models can be tuned to perform."

Despite these shortcomings, their work is a major step forward in understanding the drivers of progress in AI. Their results help shed light about how future developments in AI might play out, with important implications for AI policy.

"This work, led by Anson and Tamay, has important implications for the democratization of AI," said Neil Thompson, a co-author and Director of MIT FutureTech. "These efficiency improvements mean that each year levels of AI performance that were out of reach become accessible to more users."

"LLMs have been improving at a breakneck pace in recent years. This paper presents the most thorough analysis to date of the relative contributions of hardware and algorithmic innovations to the progress in

LLM performance," says Open Philanthropy Research Fellow Lukas Finnveden, who was not involved in the paper.

"This is a question that I care about a great deal, since it directly informs what pace of further progress we should expect in the future, which will help society prepare for these advancements. The authors fit a number of statistical models to a large dataset of historical LLM evaluations and use extensive cross-validation to select a model with strong predictive performance. They also provide a good sense of how the results would vary under different reasonable assumptions, by doing many robustness checks.

"Overall, the results suggest that increases in compute have been and will keep being responsible for the majority of LLM progress as long as compute budgets keep rising by ≥4x per year. However, algorithmic progress is significant and could make up the majority of progress if the pace of increasing investments slows down."

More information: Anson Ho et al, Algorithmic progress in language models, *arXiv* (2024). arxiv.org/abs/2403.05812

Provided by Massachusetts Institute of Technology

Citation: From recurrent networks to GPT-4: Measuring algorithmic progress in language models (2024, March 13) retrieved 9 May 2024 from https://techxplore.com/news/2024-03-recurrent-networks-gpt-algorithmic-language.html