GPT-NeoX-20B

An Open-Source Autoregressive Language Model

Slide deck by Stella Rose Biderman, adapted by Michael Pieler

Michael Pieler

MicPie @ EleutherAI

Sid Black, Eric Hallahan,
Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, Samuel Weinbach

EleutherAI is a decentralized grassroots collective of researchers focused on AI alignment, scaling, and open source AI research.

How did this all start?

One day, Connor Leahy posted in the TPU Podcast Discord:

Daj 2020-07-02
https://arxiv.org/abs/2006.16668

Hey guys lets give OpenAI a run for their money like the good ol' days

To which Leo Gao replied:

bmk 2020-07-02
@Daj this but unironically

And so it began.

Our Community

- Organized via **Discord**
- Transparent research
- Community driven
- Anyone can join:
  - Research projects
  - Discussion of state-of-the-art
  - Interpretability reading group

Language Modeling

Why Train a(nother) Large Language Model?

- Access to large language models is essential for doing research on them.

- How does repeated exposure to the same data influence the probability that the language model will memorize that data?
  - Deduplicating Training Data Makes Language Models Better
  - Quantifying Memorization Across Neural Language Models

- To what extent do language models learn to generalize notions found in the training data to the testing data?
  - Impact of Pretraining Term Frequencies on Few-Shot Reasoning

What is a Transformer?

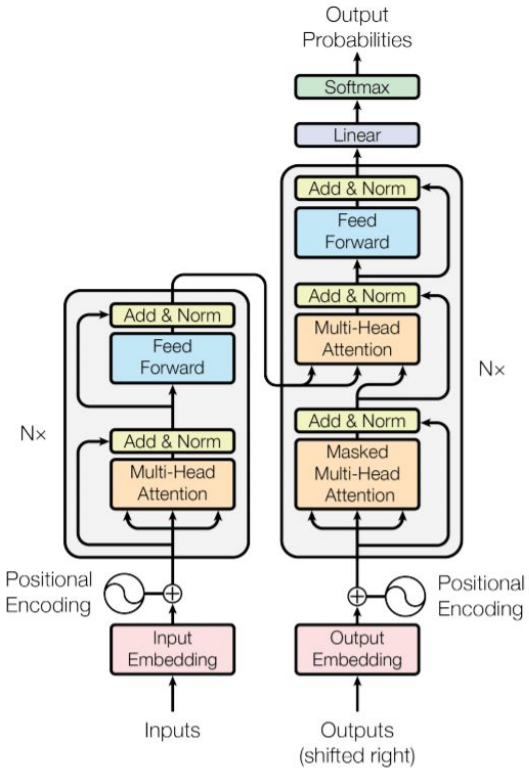
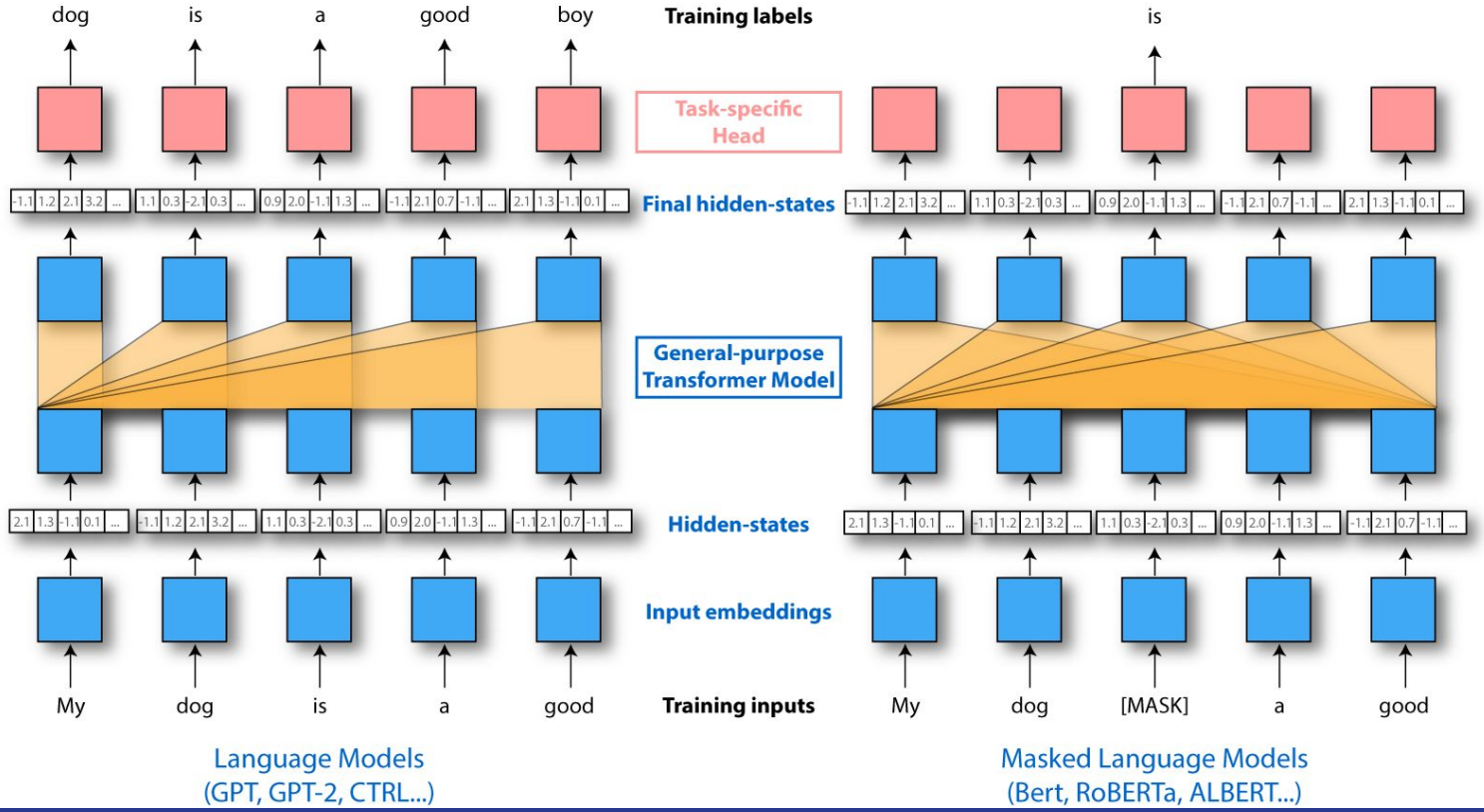


Figure 1: The Transformer - model architecture.

Language Modeling?



https://twitter.com/thom_wolf/status/1186225108282757120?s=21

LLM Leaderboards

1		Announcement	Organizaton	Author Location	Language	Parameters	Model Accessibility	Data Accessibility
2	**PaLM**	2022-04-04	Google	USA	English	**540.0B**	**Closed**	**Closed**
3	**Megatron-Turing**	2021-10-11	Microsoft, NVIDIA	USA	English	**530.0B**	**Closed**	**Closed + Pile**
4	**Gopher**	2021-12-08	DeepMind	USA	English	**280.0B**	**Closed**	**Closed + Pile**
5	**ERNIE 3.0**	2021-12-08	Baidu	China	Chinese, English	**260.0B**	**Closed**	**Closed**
6	**Yuan 1.0**	2021-10-10	Inspur AI Research	China	Chinese	**245.0B**	**Limited**	**Limited**
7	**HyperCLOVA**	2021-09-10	NAVER	Korea	Korean	**204.0B**	**Closed**	**Closed**
8	**PanGu-α**	2021-04-26	Huawei	China	Chinese	**200.0B**	**Closed**	**Closed**
9	**Jurassic-1**	2021-08-11	AI21 Labs	Israel	English	**178.0B**	**Closed**	**Open (Pile)**
10	**GPT-3**	2020-05-28	OpenAI	USA	English	**175.0B**	**Commercial**	**Closed**
11	**OPT**	2021-05-03	Meta AI	USA	English	**175.0B**	**Open (NC)**	**Closed (roBERTa) + Pile**
12	**LaMDA**	2022-01-20	Google	USA	English	**137.0B**	**Closed**	**Closed**
13	**Chinchilla**	2022-03-29	DeepMind	USA	English	70.0B	**Closed**	**Closed**
14	**Anthropic LM**	2021-12-01	Anthropic	USA	English	52.0B	**Closed**	**Closed**
15	**GPT-NeoX-20B**	2022-02-02	EleutherAI	Germany, USA, India, Canada, UK, Australia, Austria	English	20.0B	**Open**	**Open (Pile)**
16	**Turing NLG**	2020-02-13	Microsoft	USA	English	17.2B	**Closed**	**Closed**
17	**FairSeq Dense**	2021-12-20	Meta AI	USA, UK, Germany	English	13.0B	**Open**	**Closed**
18	**Big Science Model**		Big Science	Multinational	Multilingual	13.0B	**Closed**	**Open (OSCAR)**
19	**mT5**	2020-10-22	Google	USA	Multilingual	13.0B	**Open**	**Open (mC4)**
20	**ByT5**	2021-05-28	Google	USA	Multilingual	13.0B	**Open**	**Open (C4)**
21	**T5**	2019-10-23	Google	USA	English	11.0B	**Open**	**Open (C4)**
22	**CPM-2.1**	2021-06-20	Tsinghua University	China	Chinese	11.0B	**Open**	**???**
23	**Megatron 11B**	2020-04-03	NVIDIA	USA	English	11.0B	**Theoretically Open**	**???**
24	**WuDao-GLM-XXL**		Beijing Academy of A	China	Chinese	10.0B	**Limited**	**???**
25	**WuDao-GLM-XXL**		Beijing Academy of A	China	English	10.0B	**Limited**	**???**

What Does it Take to Run?

- Slim checkpoint is 39 GB, 43 GB at runtime
- Full checkpoint 268 GB
- On an A6000 you can generate ~11 tokens per second
- Better performance on two RTX 3090 Tis, but less cost efficient

Model Training Details

The Pile: 800GB of Diverse Text for LLMs

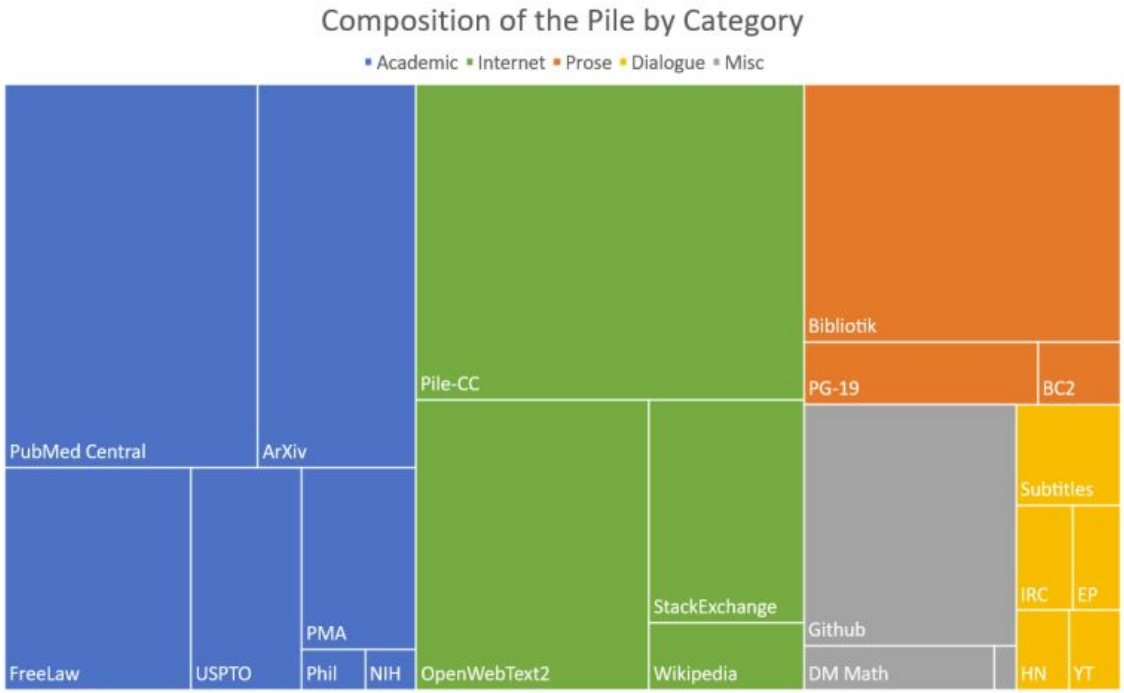


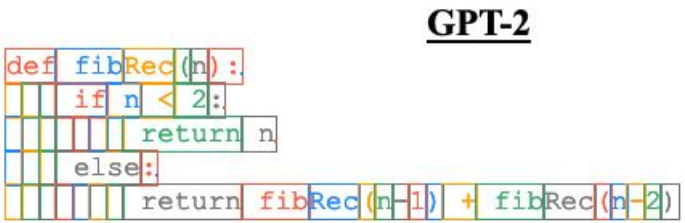
Figure 1: Treemap of Pile components by effective size.

New Tokenizer

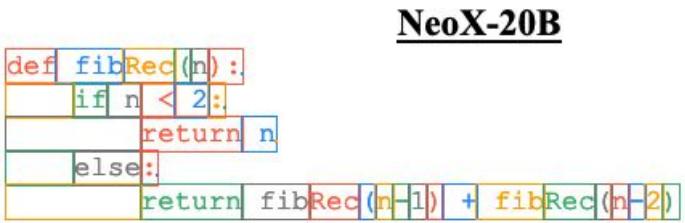
- Variant on standard BPE tokenizer
- Adds special tokens for spaces to handle code data better
- 15% fewer tokens on arXiv
- 23% fewer tokens on GitHub
- 0.001% more tokens on C4

New Tokenizer

- Variant on standard BPE tokenizer
- Adds special tokens for spaces to handle code data better



Number of tokens=55



Number of tokens=39

Model architecture

- Mostly the same as GPT-3, but all dense layers
- Pretty different from GPT-Neo
- Almost exactly the same as GPT-J
- 44 layers, hidden dimension size of 6144, and 64 heads

Parallel Attention and Feedforward Layers

Standard: $x + \mathrm{FF}(\mathrm{LN}_2(x + \mathrm{Attn}(\mathrm{LN}_1(x))$

GPT-J: $x + \mathrm{Attn}(\mathrm{LN}_1(x)) + \mathrm{FF}(\mathrm{LN}_2(x))$

Rotary Embeddings

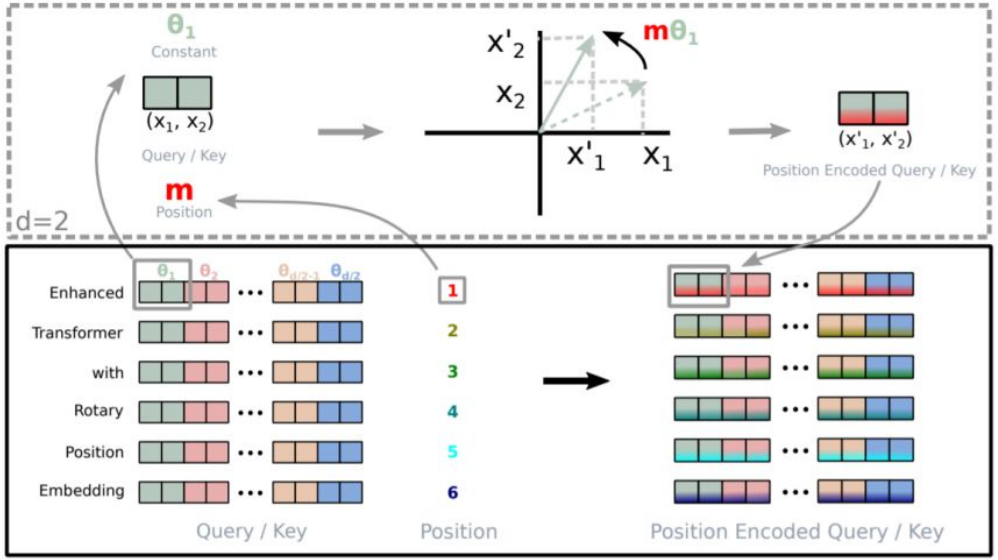


Figure 1: A pictorial representation of rotary embeddings, from Su et al. [2021].

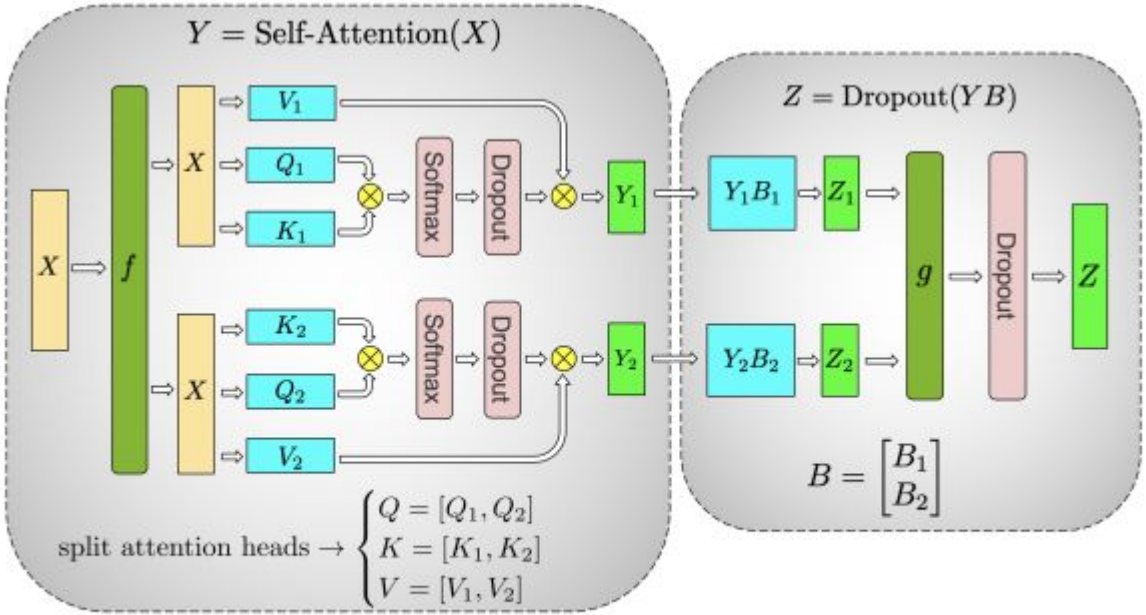
$$\text{softmax}\left(\frac{1}{\sqrt{d}}\sum_{n,m}\mathbf{x}_m^T\mathbf{W}_q^T\mathbf{W}_k\mathbf{x}_n\right)$$

$$\text{softmax}\left(\frac{1}{\sqrt{d}}\sum_{n,m}\mathbf{x}_m^T\mathbf{W}_q^T R^d_{\Theta,(n-m)}\mathbf{W}_k\mathbf{x}_n\right)$$

Training

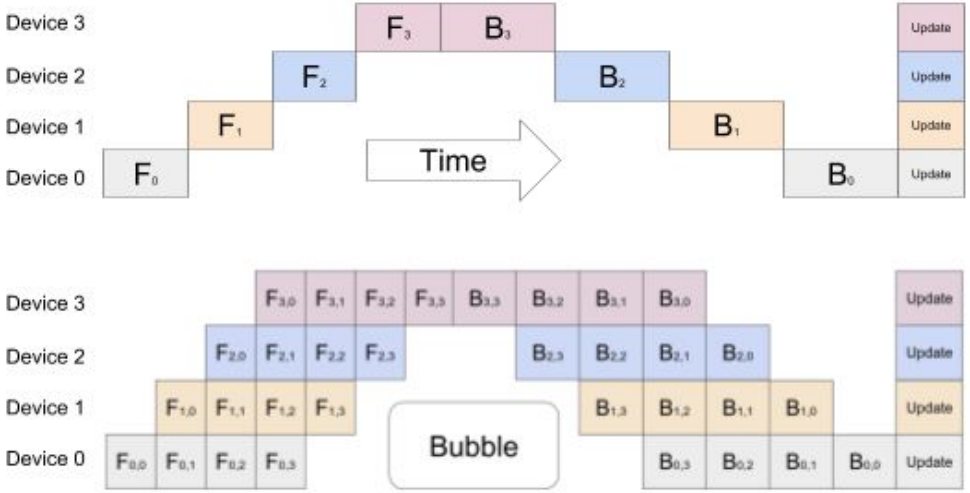
- Hyperparameters based on GPT-3
- Batch size of 3.15M tokens = 1538 contexts of 2048 tokens each
- 150,000 steps
- AdamW optimizer with beta values of 0.9 and 0.95 with ZeRO optimizer
- Tensor and pipeline parallelism

Tensor Parallelism



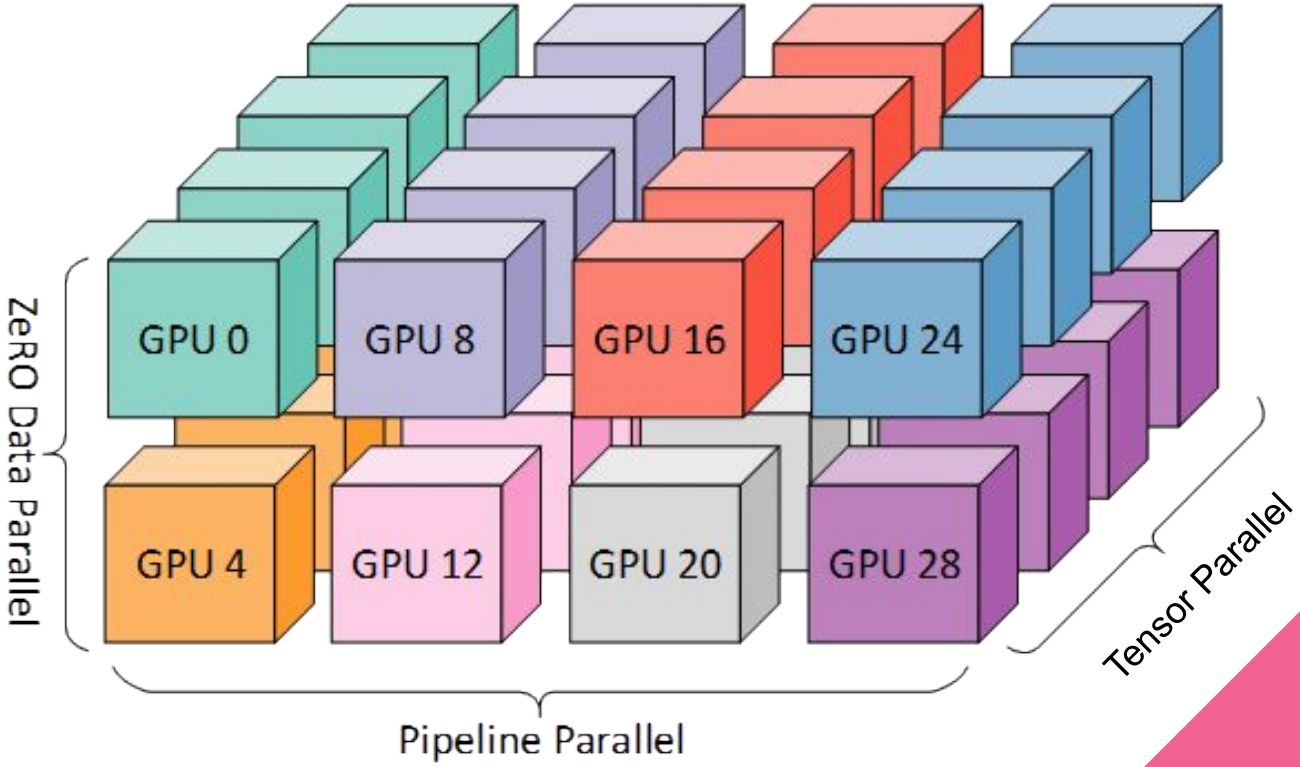
(b) Self-Attention

Pipeline Parallelism

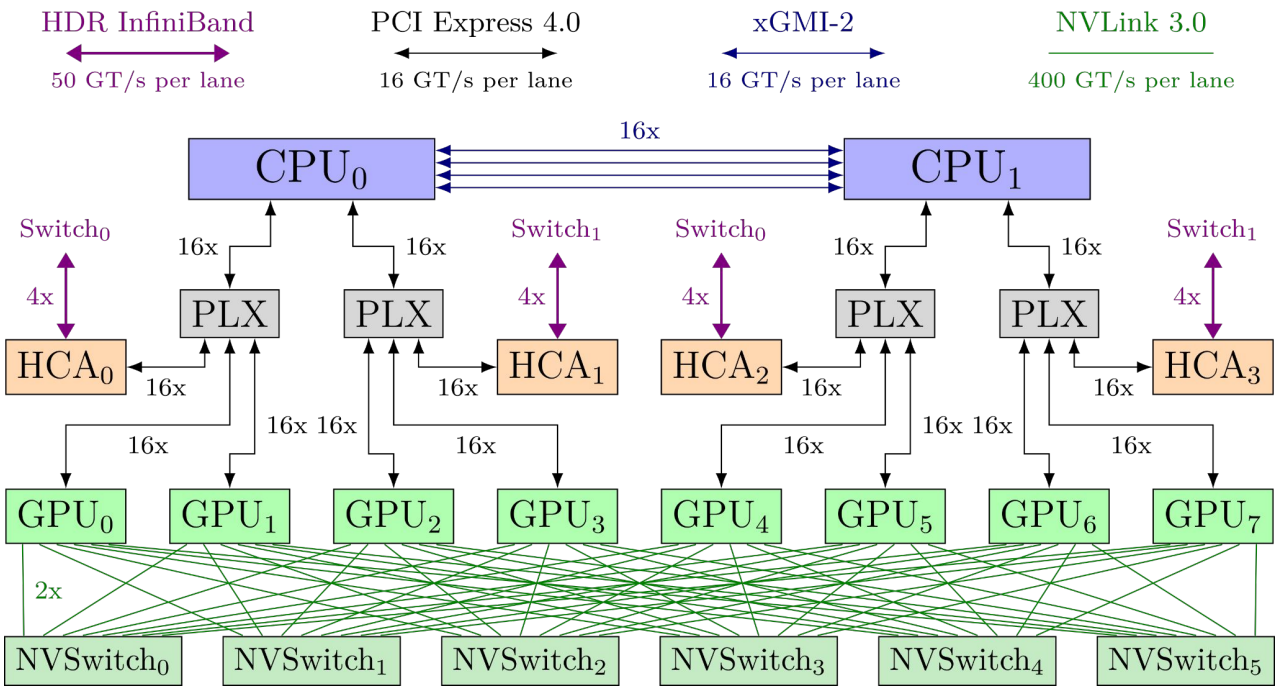


*Top: The naive model parallelism strategy leads to severe underutilization due to the sequential nature of the network. Only one accelerator is active at a time. **Bottom:** GPipe divides the input mini-batch into smaller micro-batches, enabling different accelerators to work on separate micro-batches at the same time.*

3D Parallelism



Components of a Computing Cluster



96 A100s

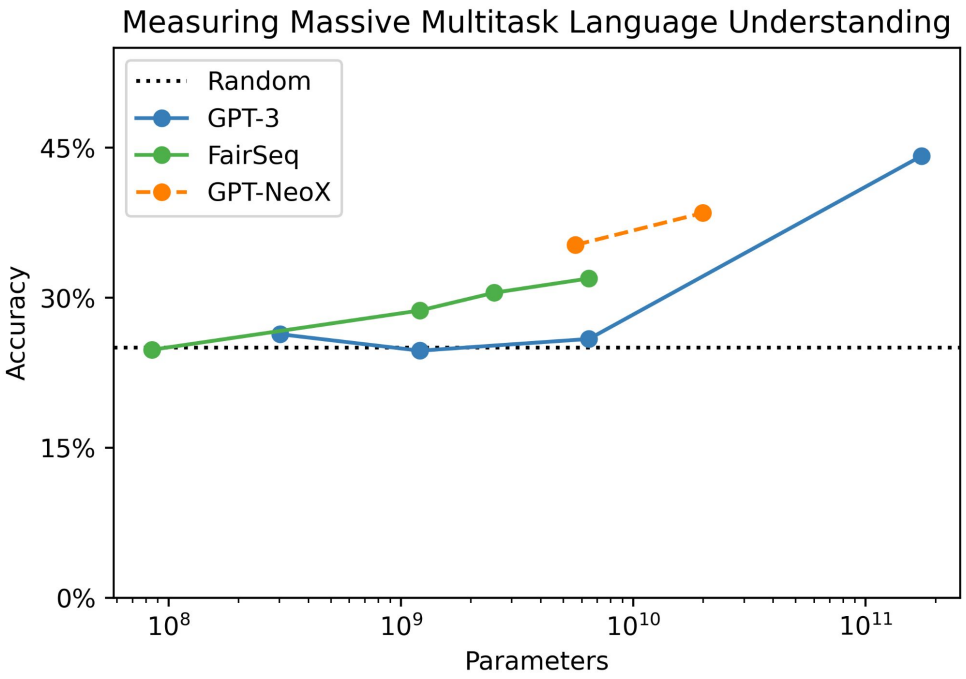
12 nodes of 8 A100s

Performance Metrics

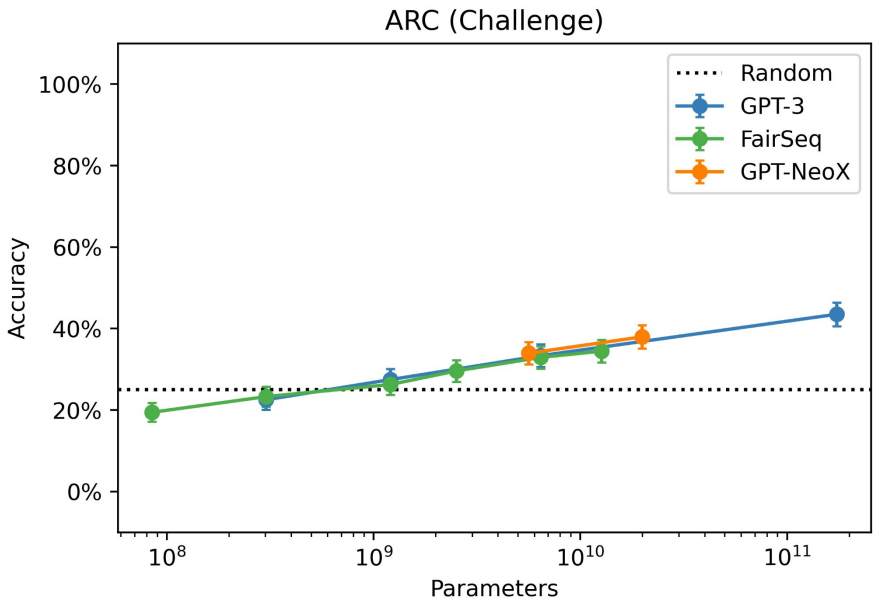
Knowledge-Intensive Tasks

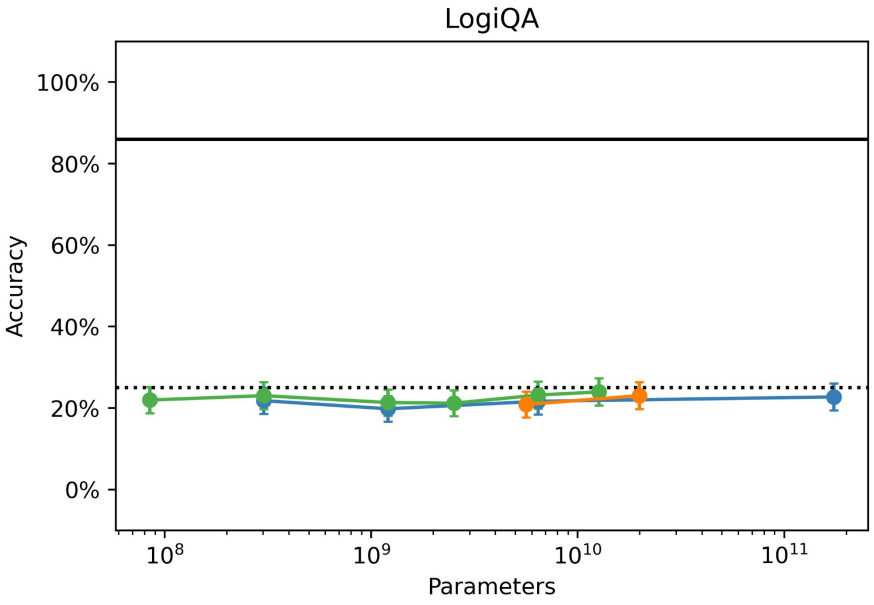
Data is primarily questions that require extremely advanced knowledge <u>in humans</u>

“Mere” intelligence is insufficient: substantial subject specific expertise is required <u>in humans</u>

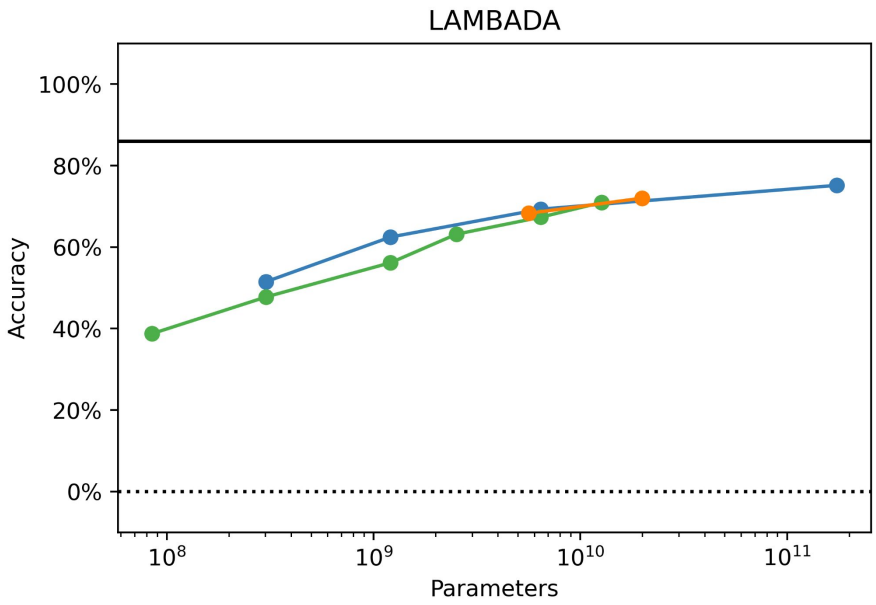


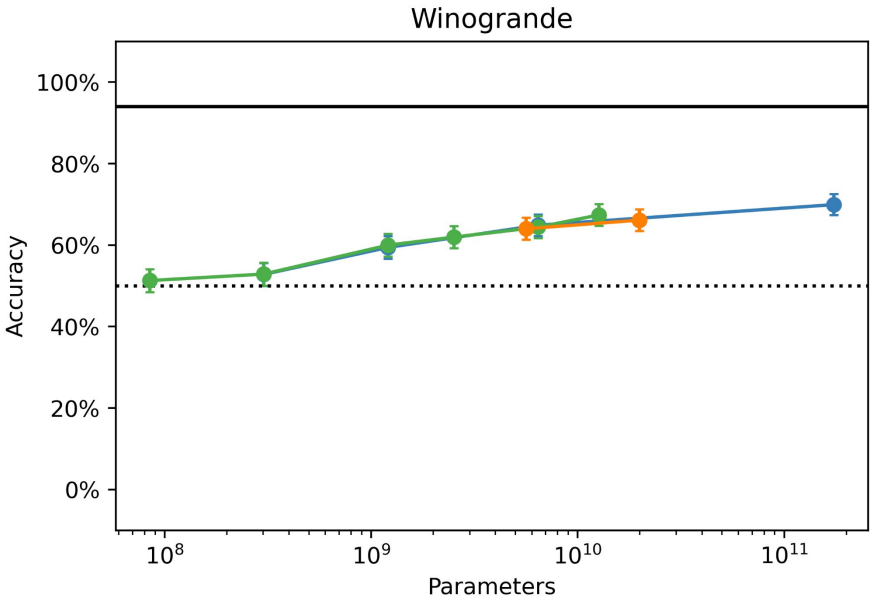
Standard Language Benchmarks



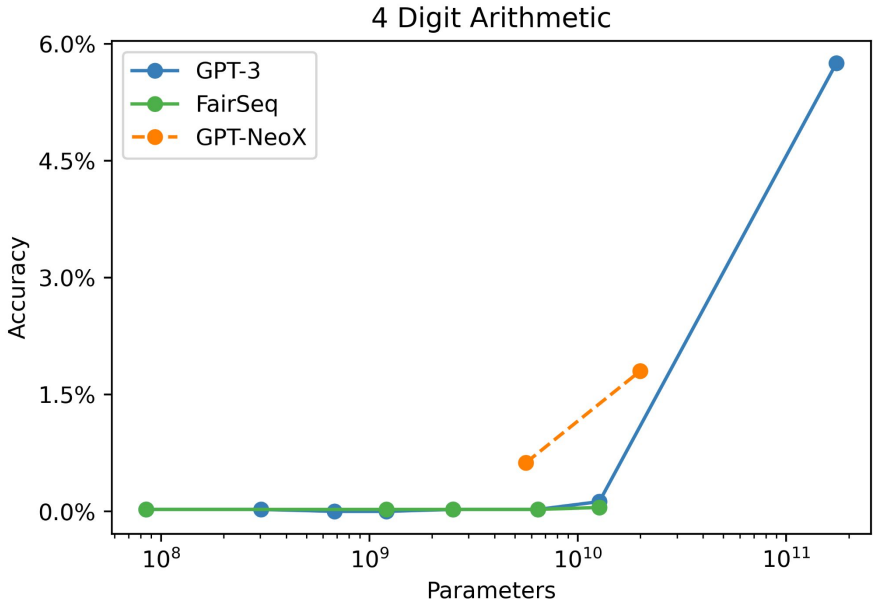


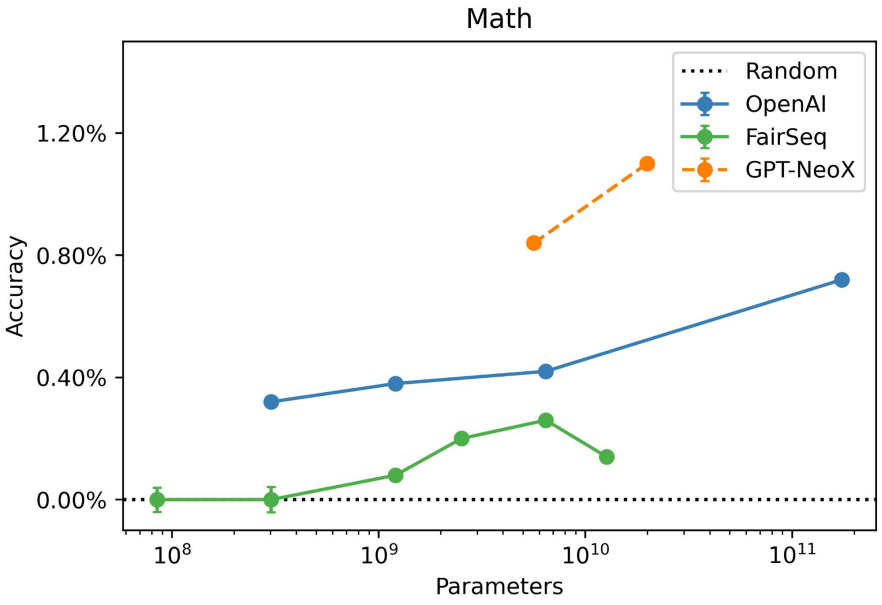
Standard Language Benchmarks





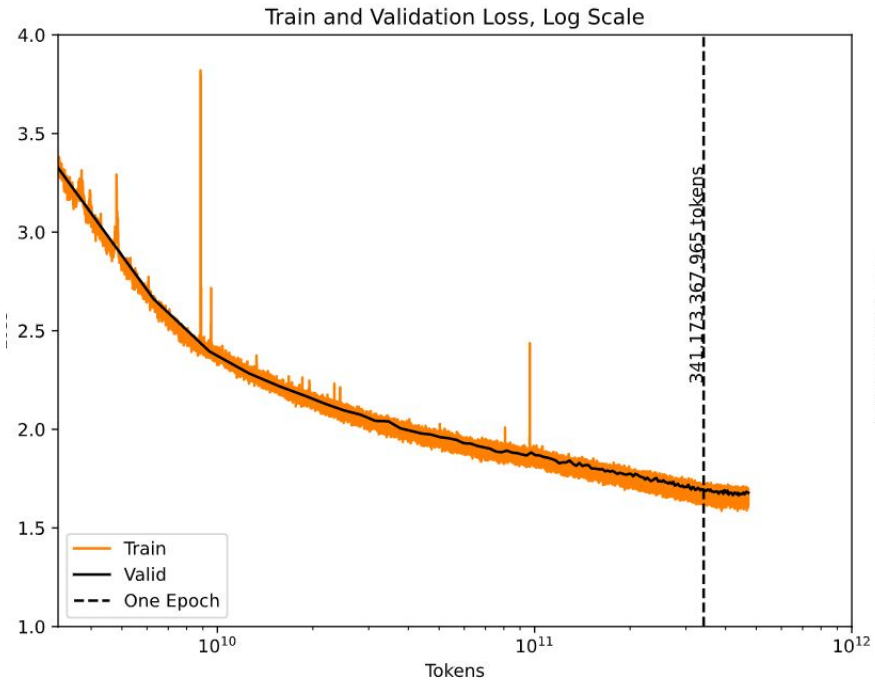
Mathematics

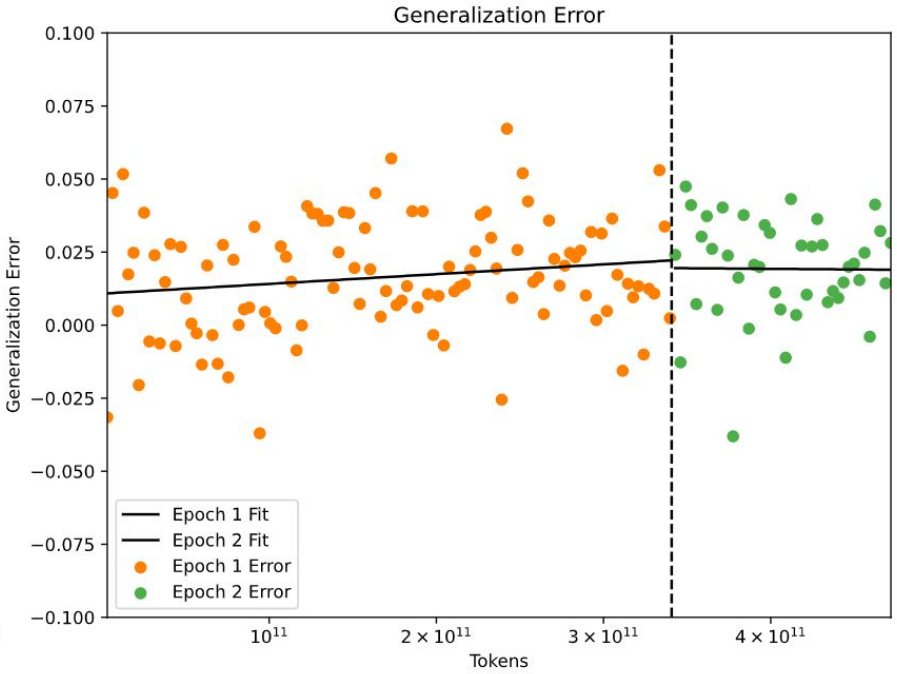




Scientific Observations

How Bad is a Second Epoch?

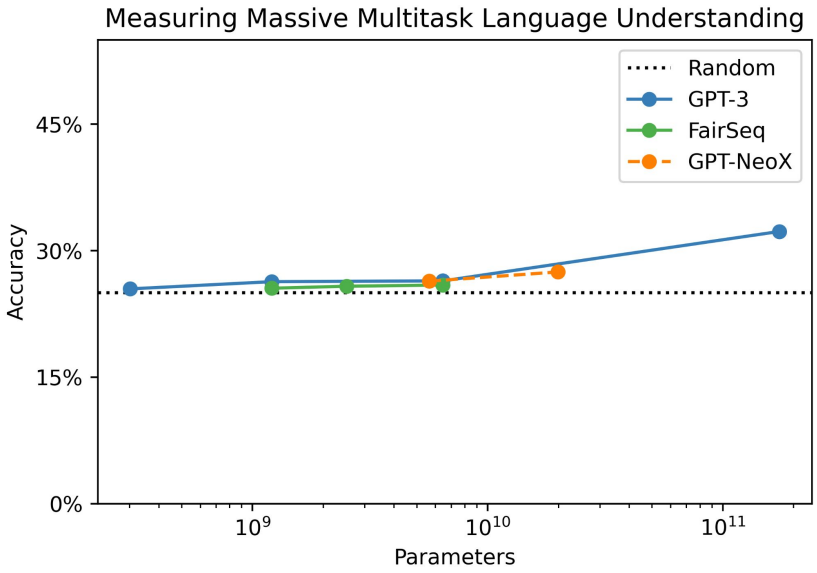




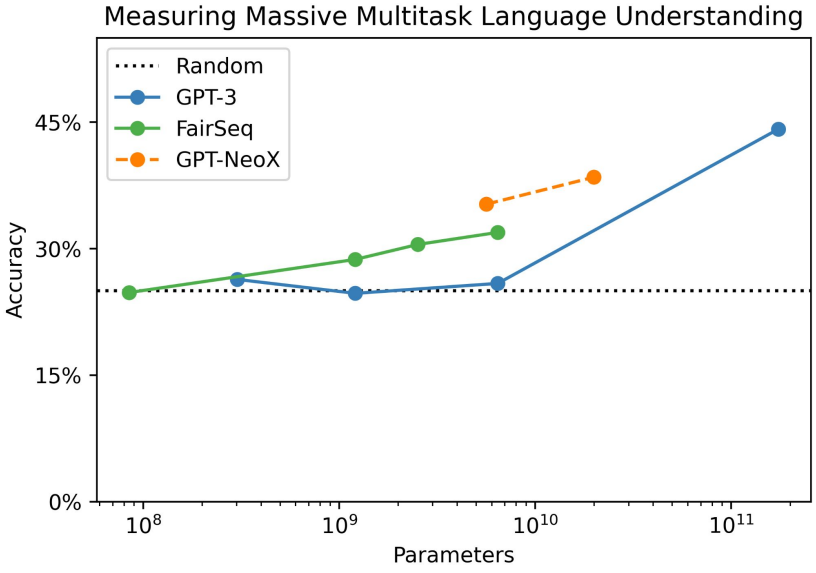
Improved Few-Shot Learning?

Perhaps due to “multitask” nature of the Pile

zero-shot



5-shot



Energy and Carbon

	Coal	Gas	Hydro	Nuclear	Solar	Wind	Other
% Electricity Mix	30.40%	31.30%	1.30%	17.40%	0.30%	18.10%	1.30%
t_{CO_2}/MWh	0.95	0.6078	0	0	0	0	0

1830 hours of training

920 hours of testing and evaluation

66.24 MWh -> 35 metric tons of CO2

Limitations

- Lack of coding evaluations
- Suboptimal training regime
- Further investigation of “multitask” training is needed
- Need for further improvement in democratizing access:
  If you want to use GPT-NeoX-20B in your research but do not have the computational resources to do so, email stella@eleuther.ai.

Questions?

Come join us!