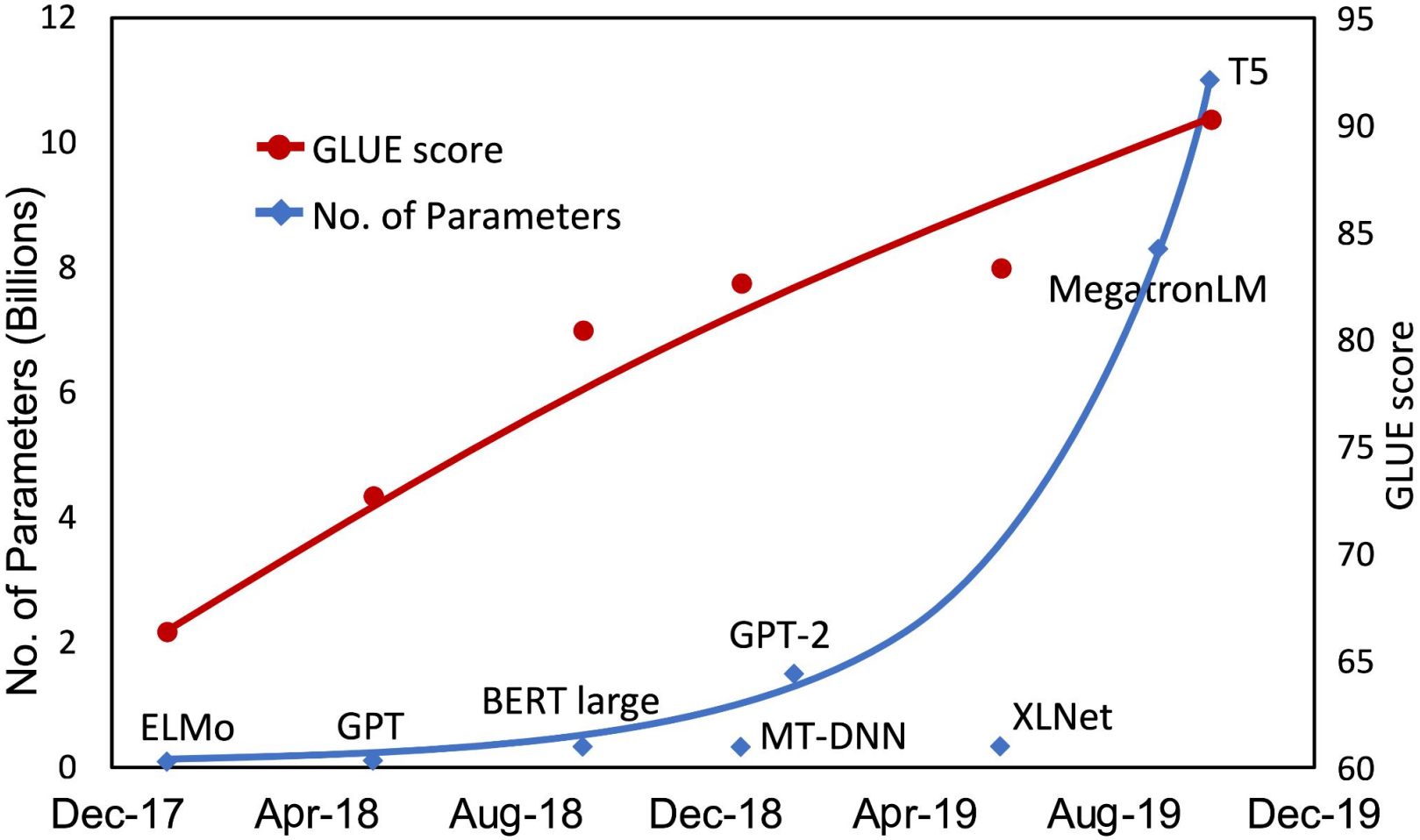
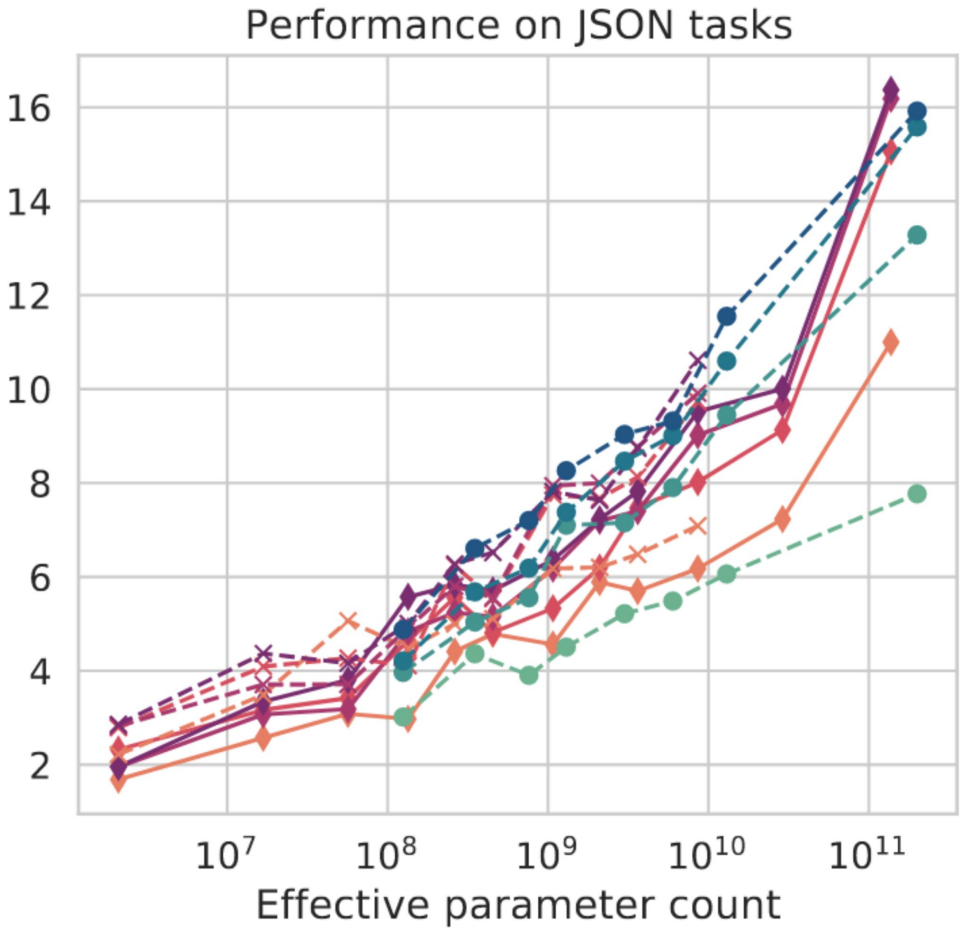
How to Be an Academic Machine Learning Researcher in the Era of Scale

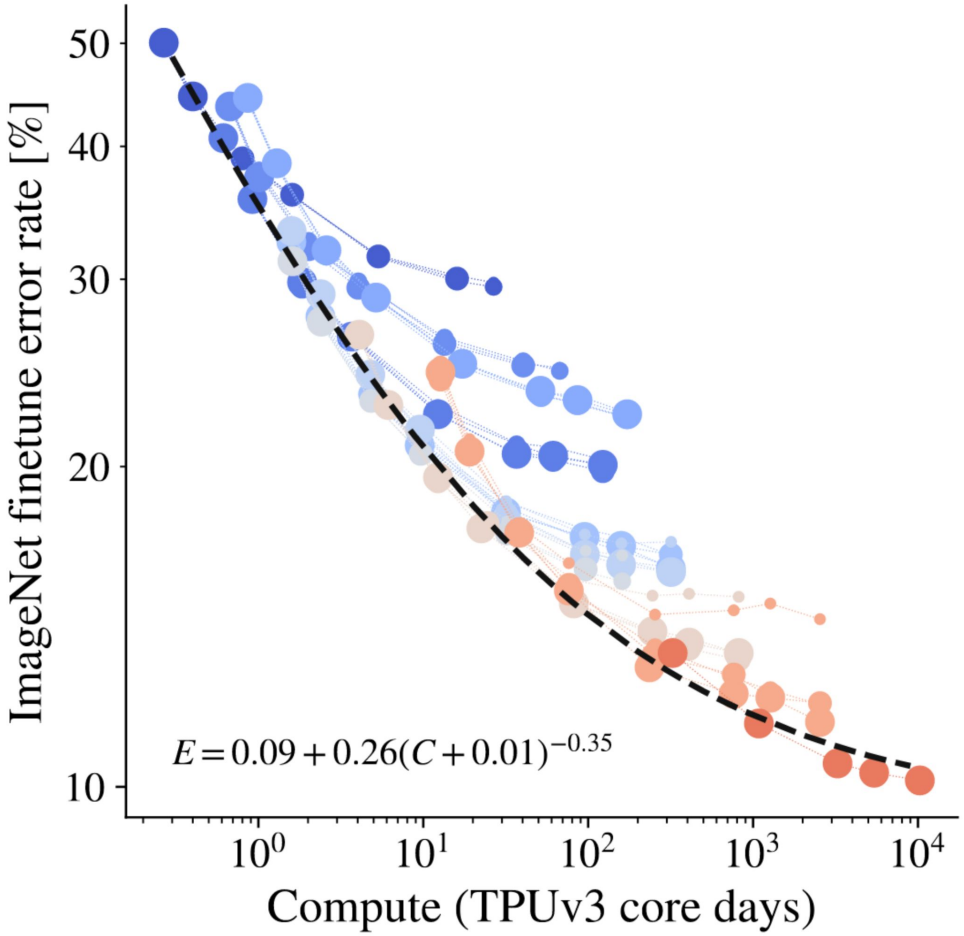
Colin Raffel



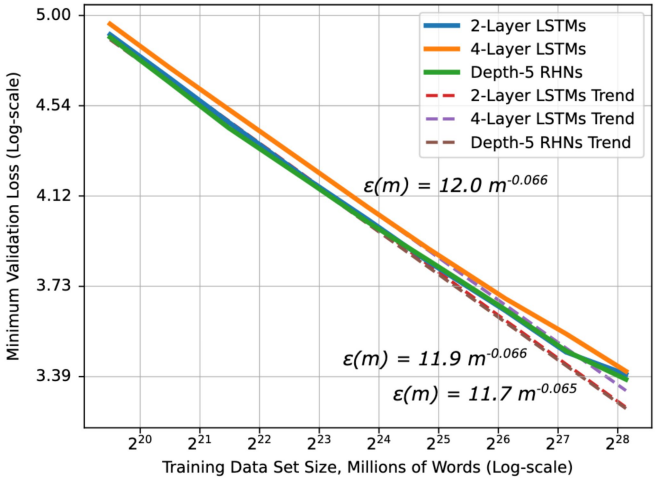
From "Real-Time Social Media Analytics with Deep Transformer Language Models: A Big Data Approach" by Ahmet and Abdullah

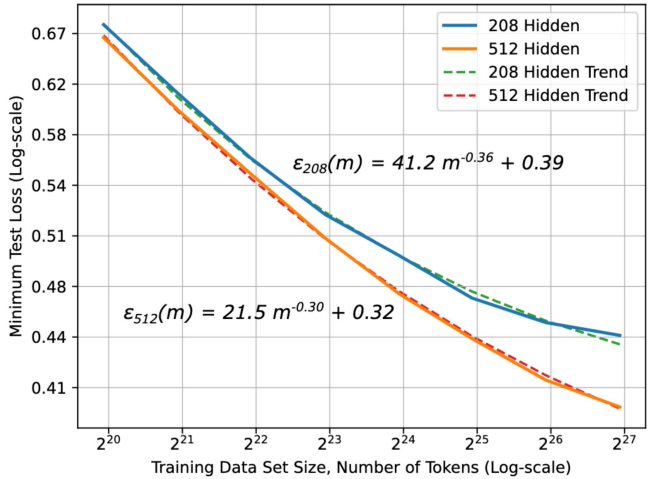


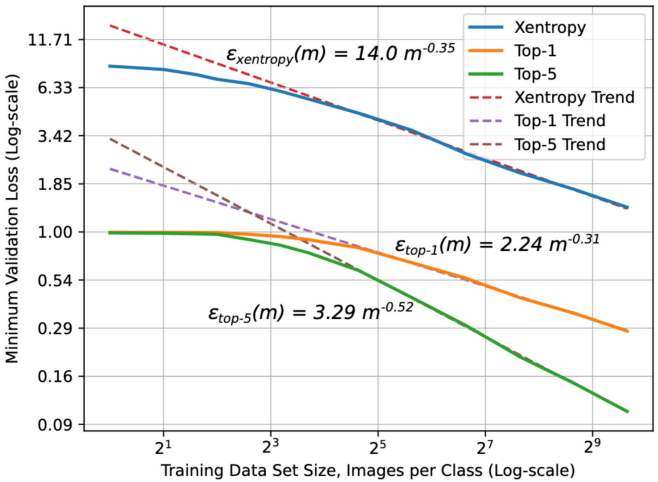
From "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models" by Srivastava et al.

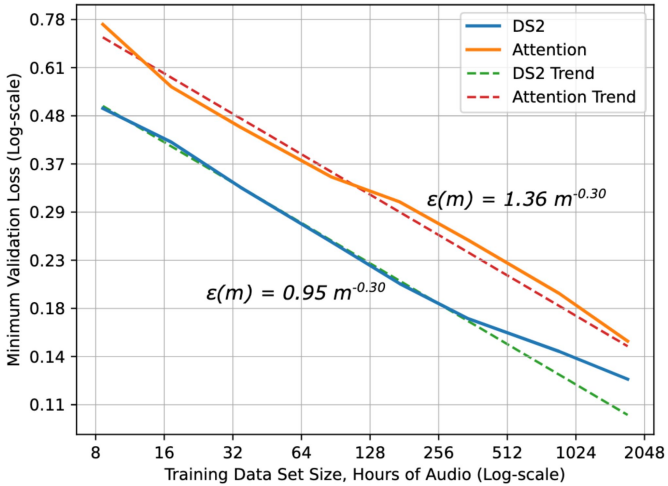


From "Scaling Vision Transformers" by Zhai et al.

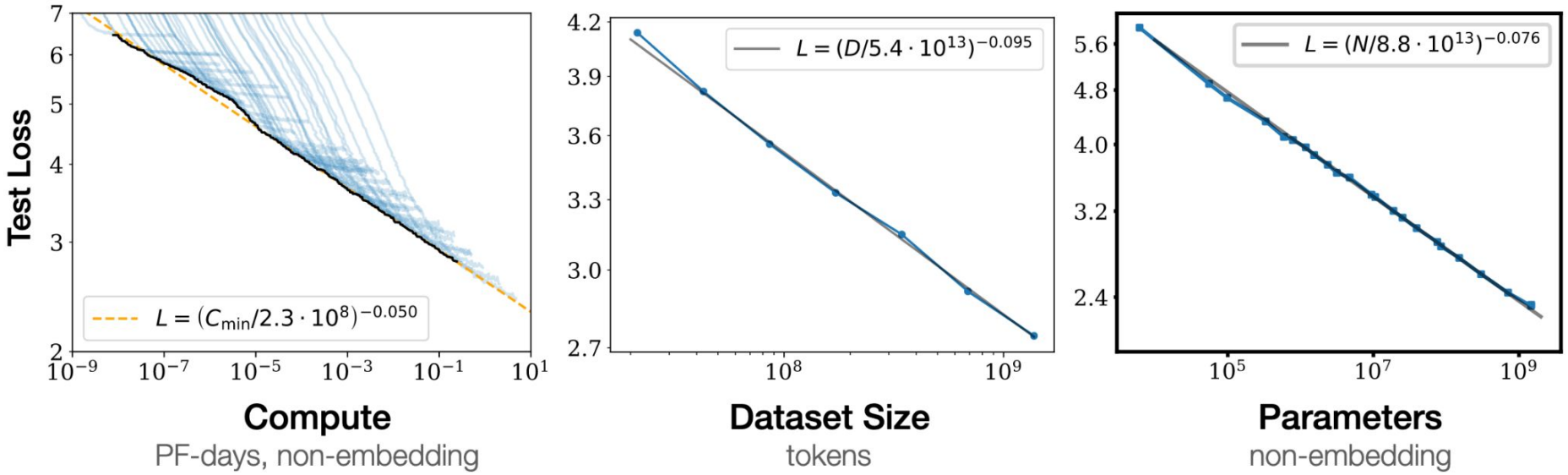








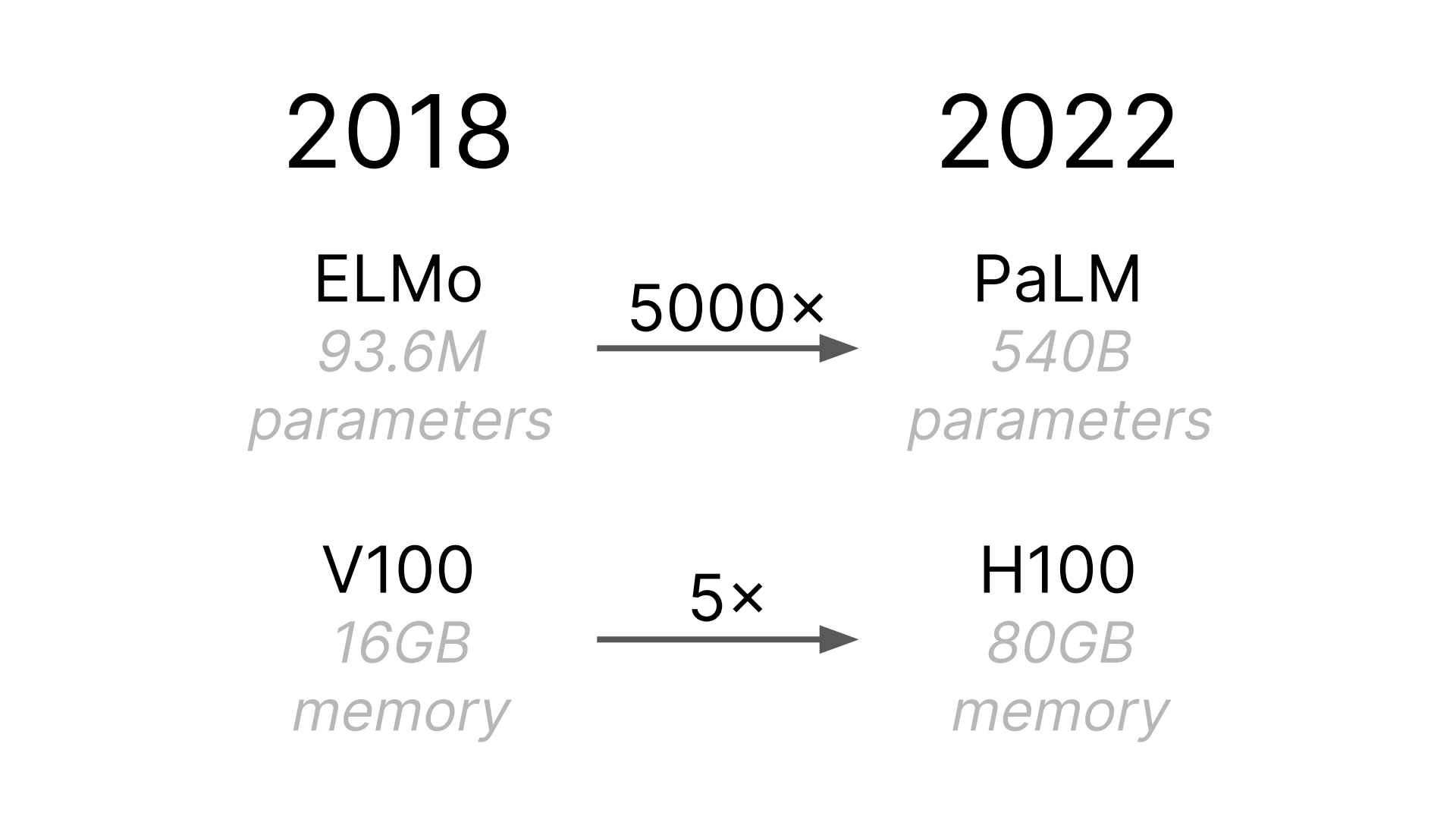
From "Deep Learning Scaling is Predictable, Empirically" by Hestness et al.



From "Scaling Laws for Neural Language Models" by Kaplan et al.

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant ... but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available.

From “The Bitter Lesson” by Sutton


2018
2022
ELMo
93.6M
parameters
5000×
PaLM
540B
parameters
V100
16GB
memory
5×
H100
80GB
memory

$$\hat{y}_i = f_\theta(x_i)$$

$$\underbrace{\ldots W h \ldots}$$

$$\partial\theta = \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$$

$$\theta \leftarrow \theta + \texttt{optimizer}(\partial\theta)$$

$$\hat{y}_i = \underbrace{f_\theta(x_i)}_{\dots W h \dots}$$

Memory

$$\partial\theta = \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$$

$$\theta \leftarrow \theta + \texttt{optimizer}(\partial\theta)$$

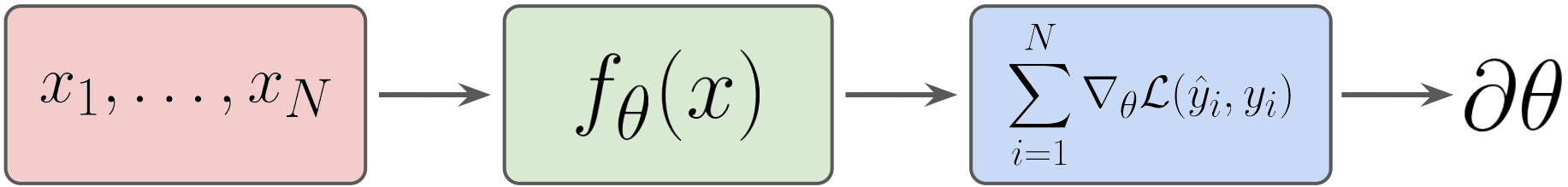
$$\hat{y}_i = f_\theta(x_i)$$

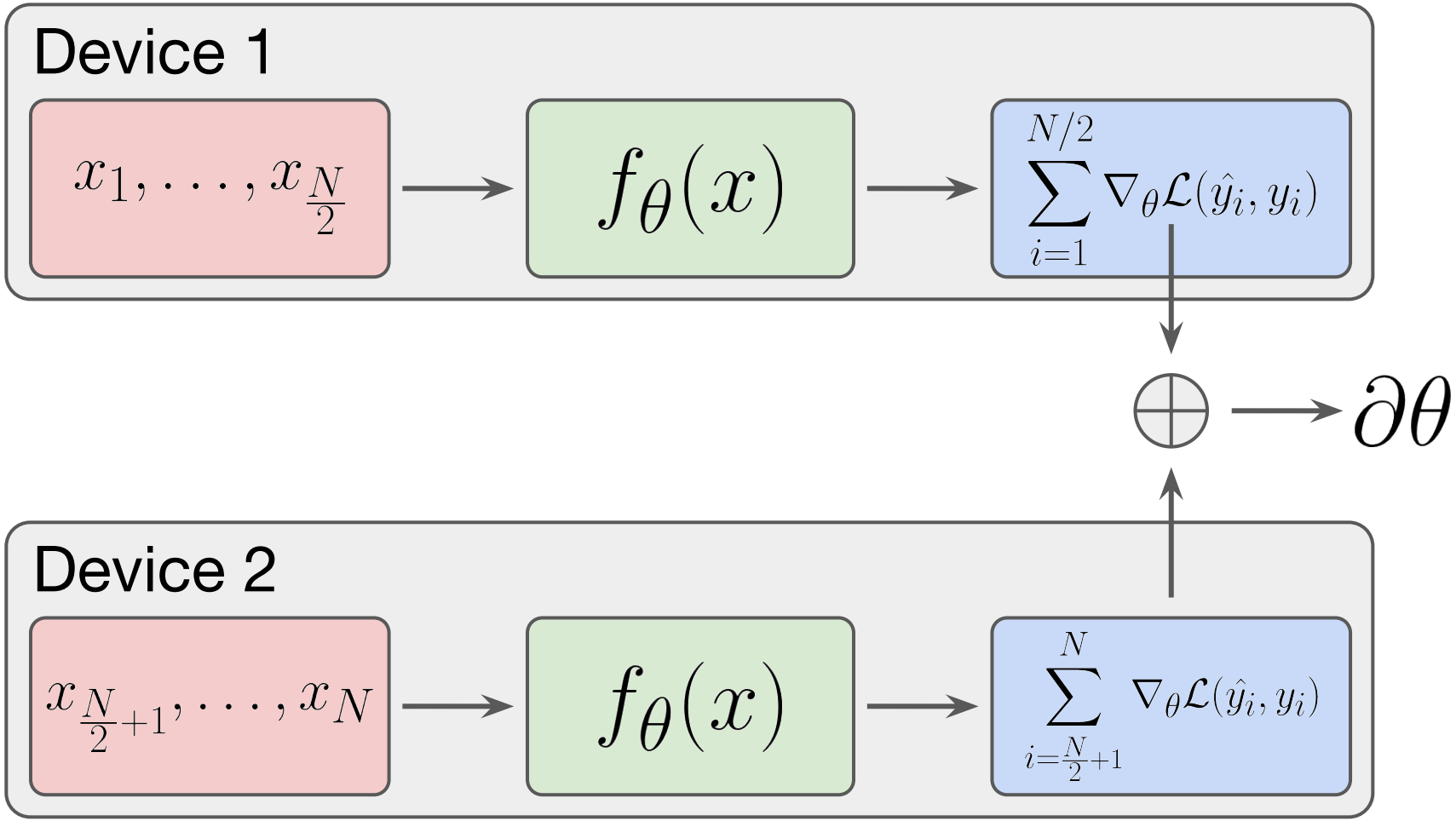
$$\underbrace{\ldots W h \ldots}$$

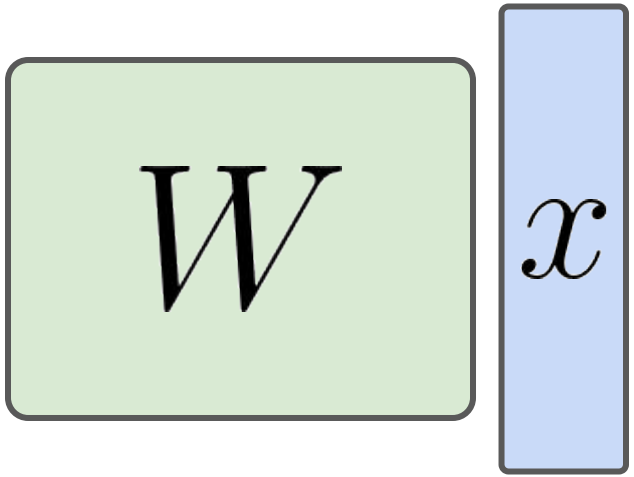
Compute

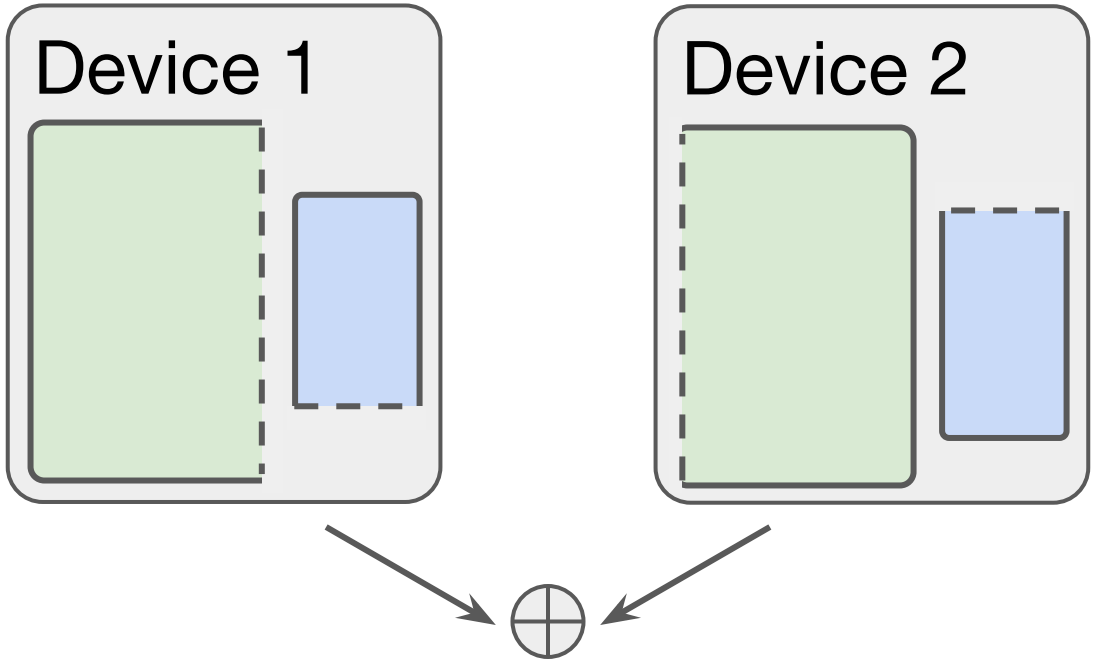
$$\partial\theta = \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$$

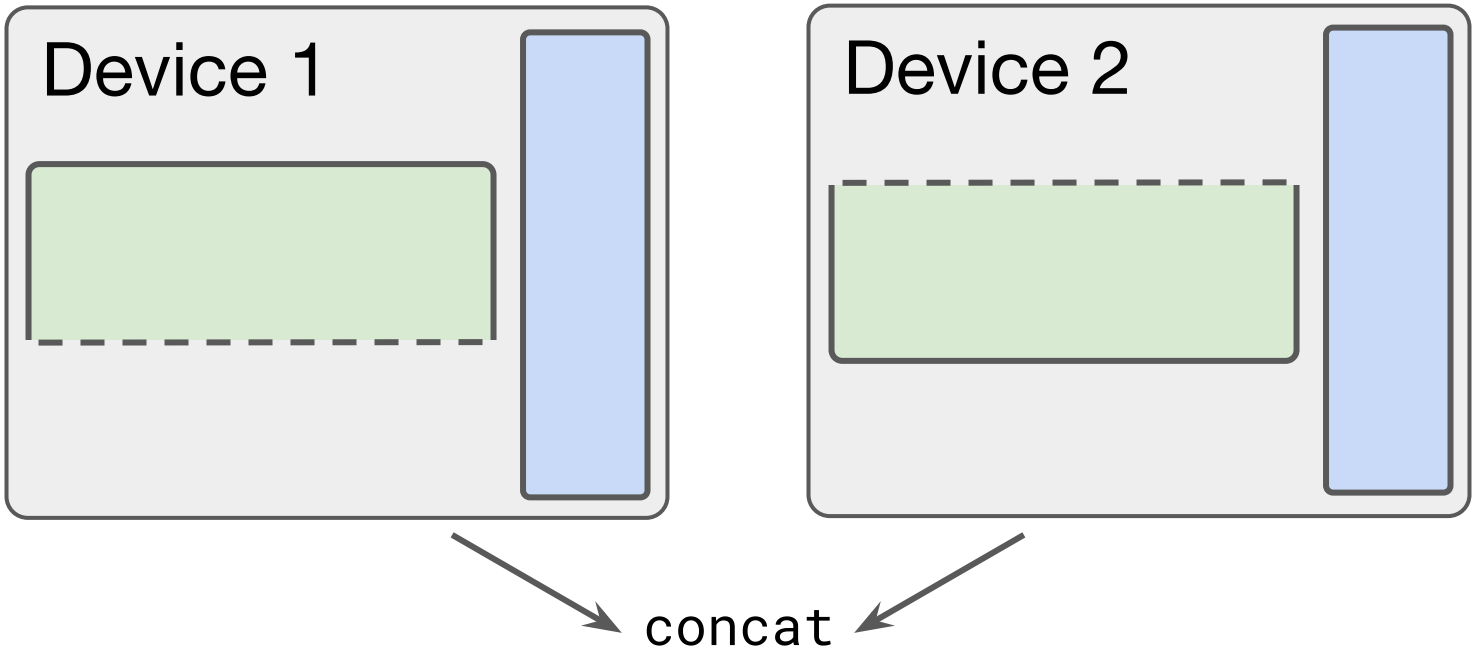
$$\theta \leftarrow \theta + \texttt{optimizer}(\partial\theta)$$

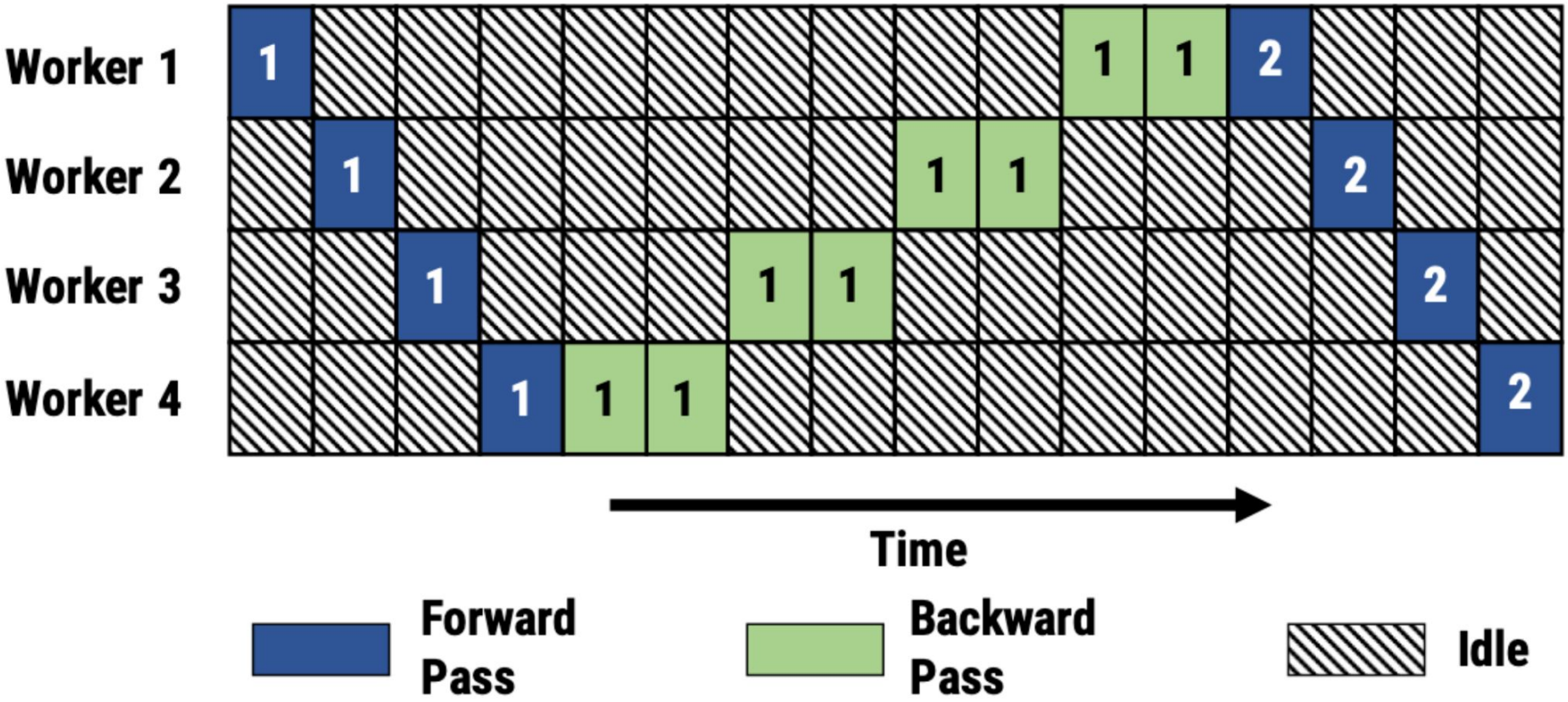

$x_1, \ldots, x_N$
$f_\theta(x)$
$\sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$
$\partial\theta$


Device 1
$x_1, \ldots, x_{\frac{N}{2}}$
$f_\theta(x)$
$\sum_{i=1}^{N/2} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$
$\partial\theta$
Device 2
$x_{\frac{N}{2}+1}, \ldots, x_N$
$f_\theta(x)$
$\sum_{i=\frac{N}{2}+1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$

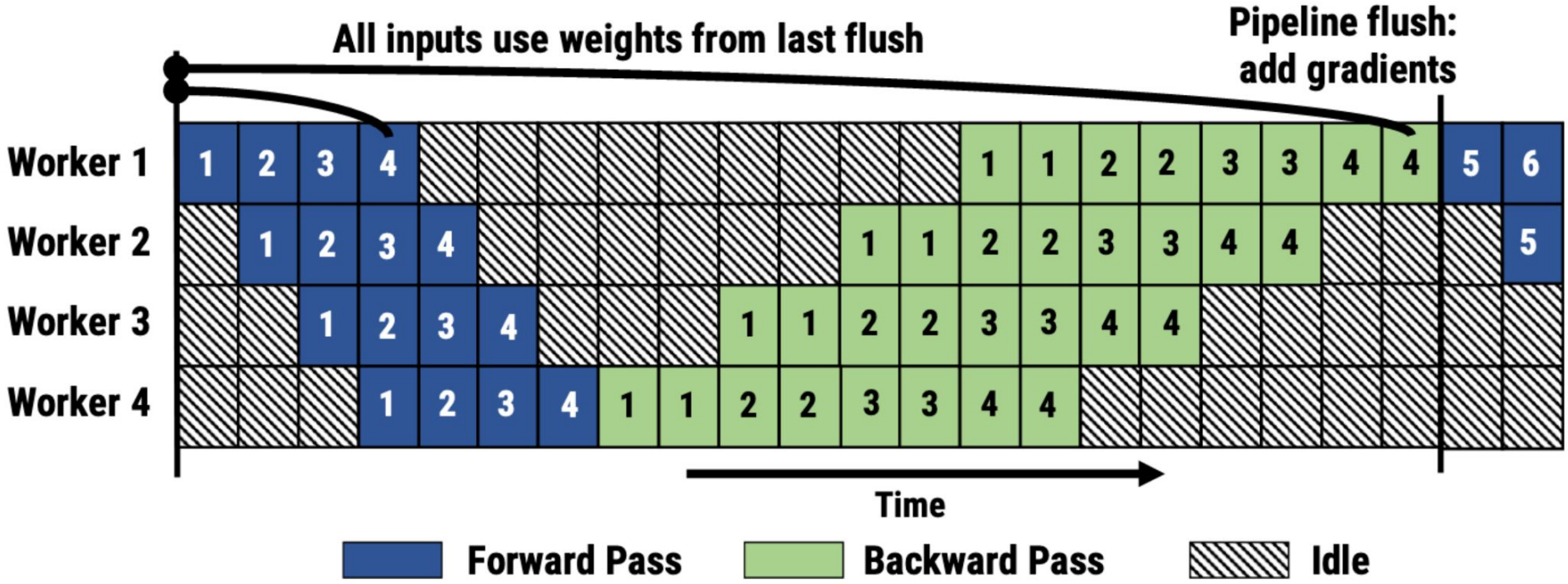

W
x


Device 1
Device 2


Device 1
Device 2
concat

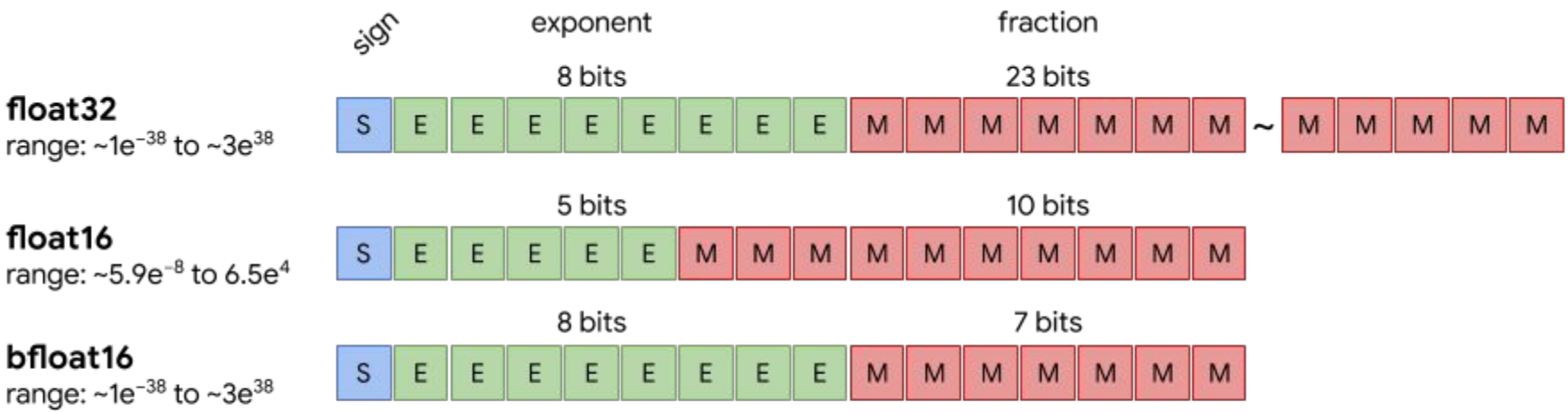


From "PipeDream: Generalized Pipeline Parallelism for DNN Training" by Narayanan et al.

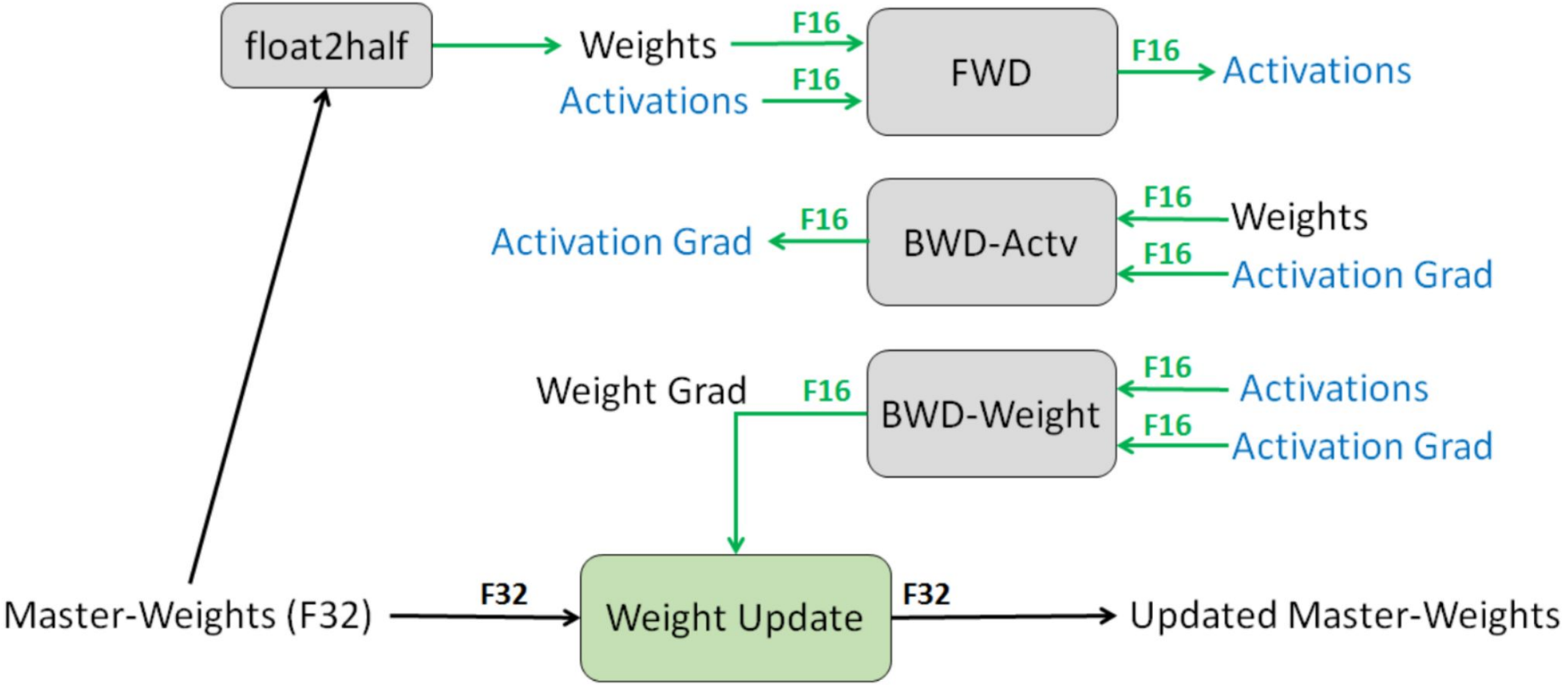


From "PipeDream: Generalized Pipeline Parallelism for DNN Training" by Narayanan et al.

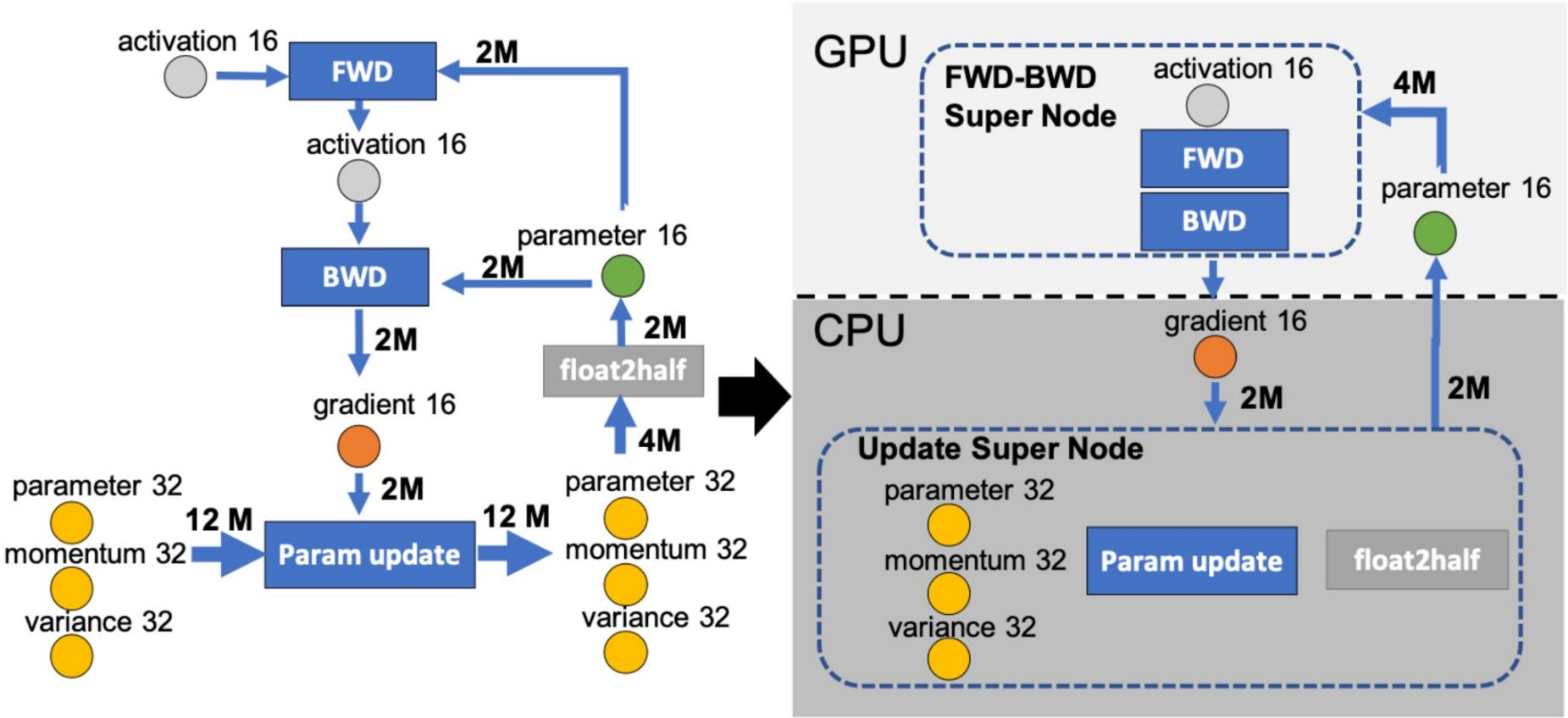
From "Efficient Large-Scale Language Model Training on GPU Clusters" by Narayanan et al.



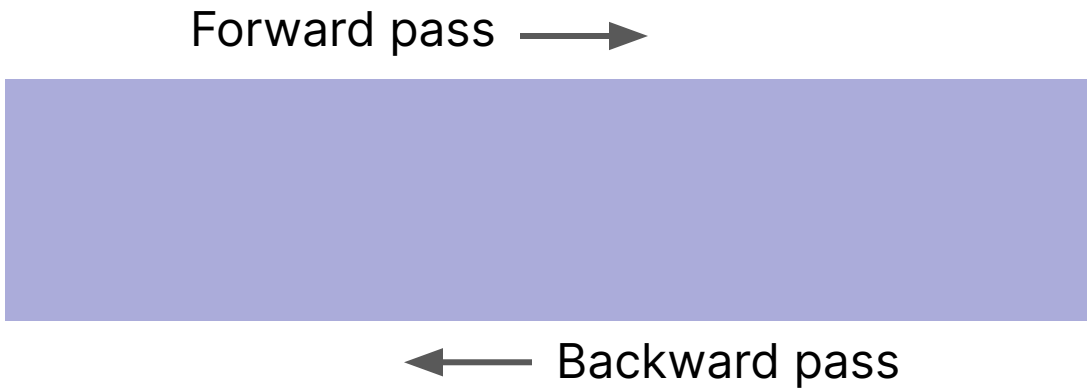
From https://cloud.google.com/tpu/docs/bfloat16



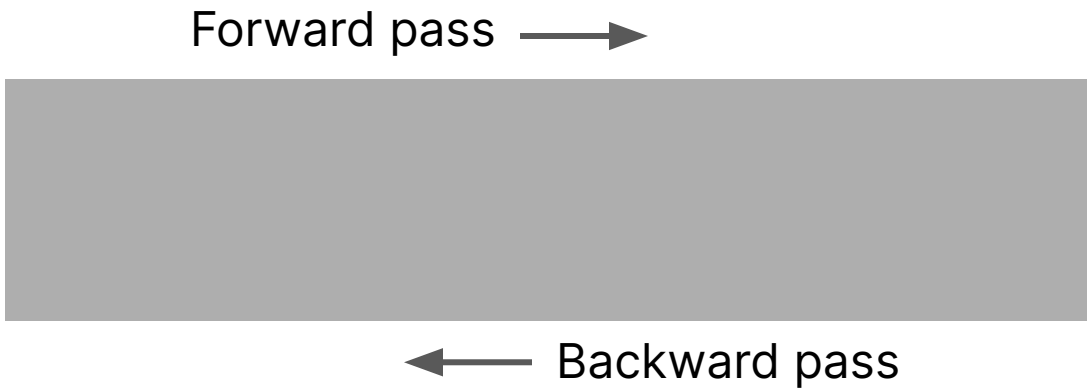
From "Mixed Precision Training" by Micikevicius et al.



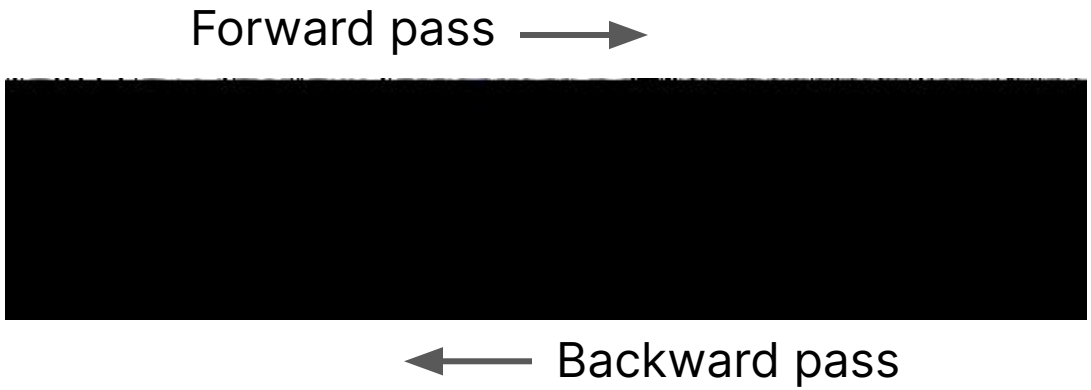
From "ZeRO-Offload: Democratizing Billion-Scale Model Training" by Micikevicius et al.



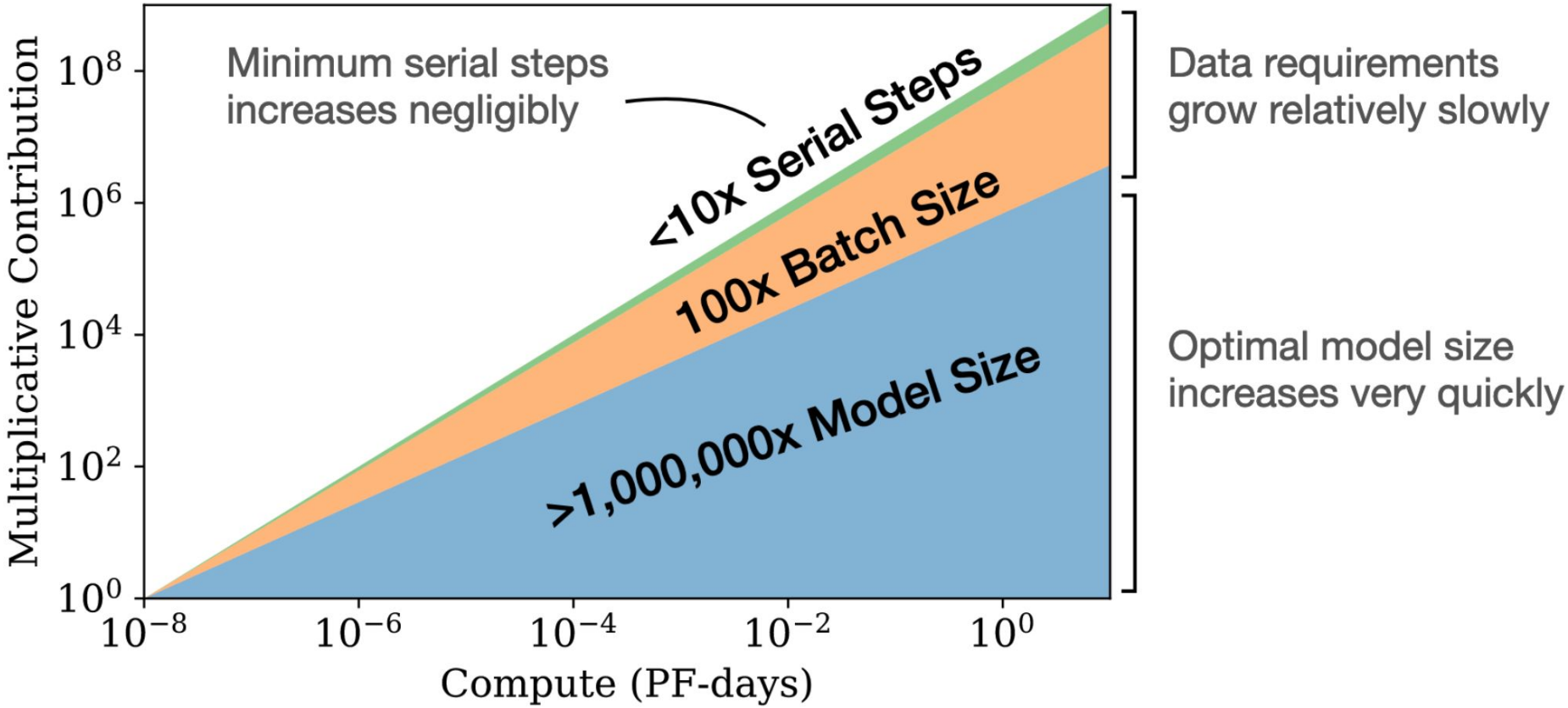
From https://medium.com/tensorflow/fitting-larger-networks-into-memory-583e3c758ff9



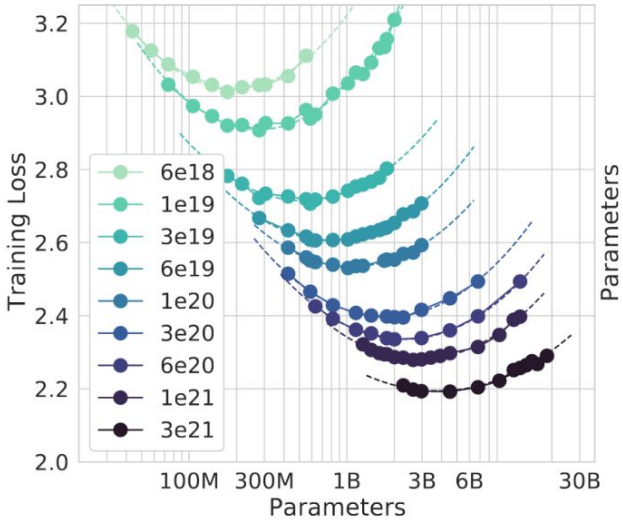
From https://medium.com/tensorflow/fitting-larger-networks-into-memory-583e3c758ff9

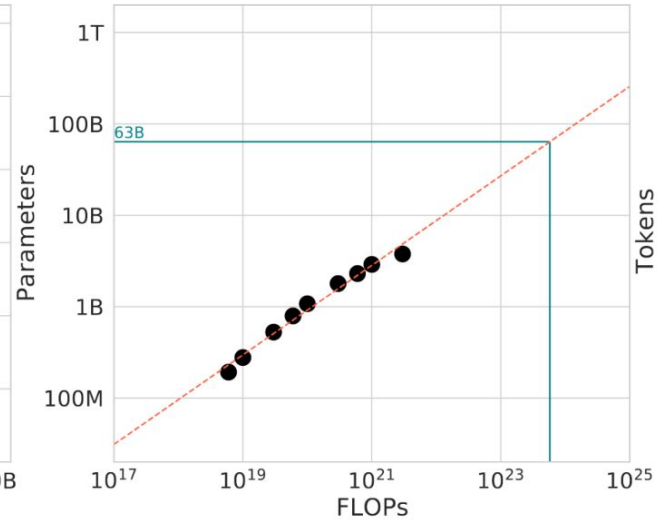


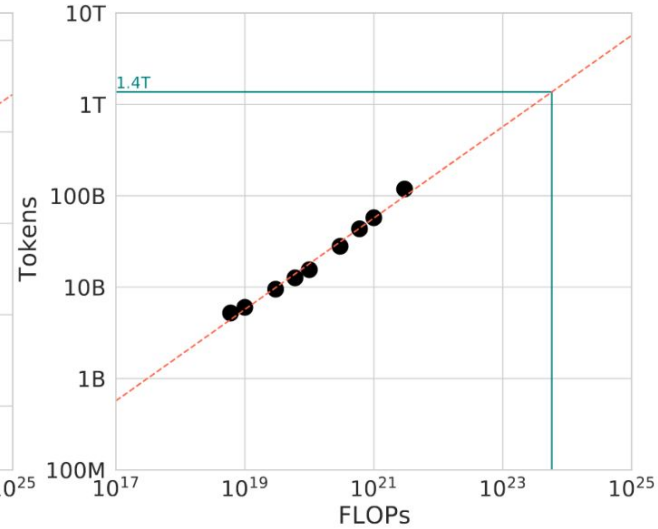
From https://medium.com/tensorflow/fitting-larger-networks-into-memory-583e3c758ff9



From "Scaling Laws for Neural Language Models" by Kaplan et al.







From "Training Compute-Optimal Large Language Models" by Hoffmann et al.

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant ... but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available.

From “The Bitter Lesson” by Sutton

*The biggest lesson that can be read from 70 years of AI research is that **general methods that leverage computation are ultimately the most effective, and by a large margin**. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant ... but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available.*

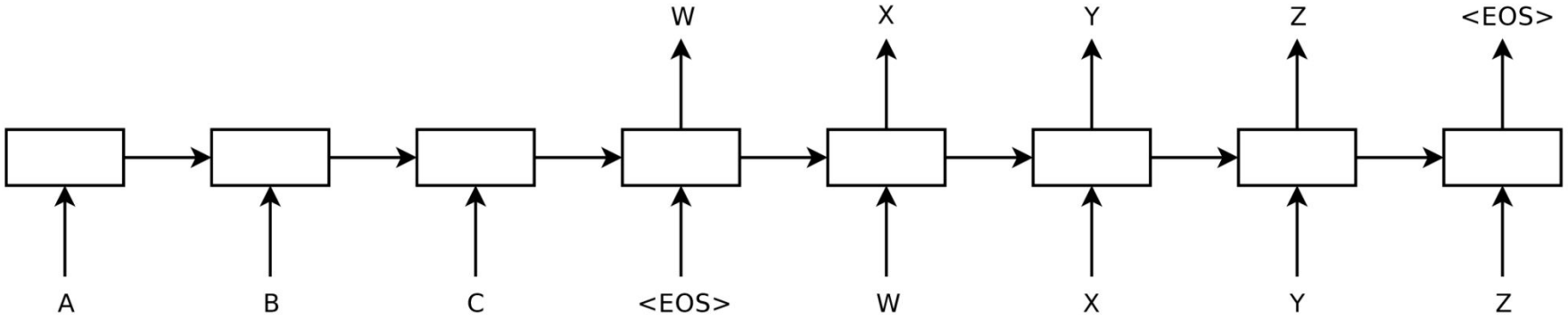
From “The Bitter Lesson” by Sutton

*The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are **ultimately** the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant ... but, **over a slightly longer time than a typical research project**, massively more computation inevitably becomes available.*

→ At any point in time, it is likely more effective to be clever!
(The Bitter Corollary?)

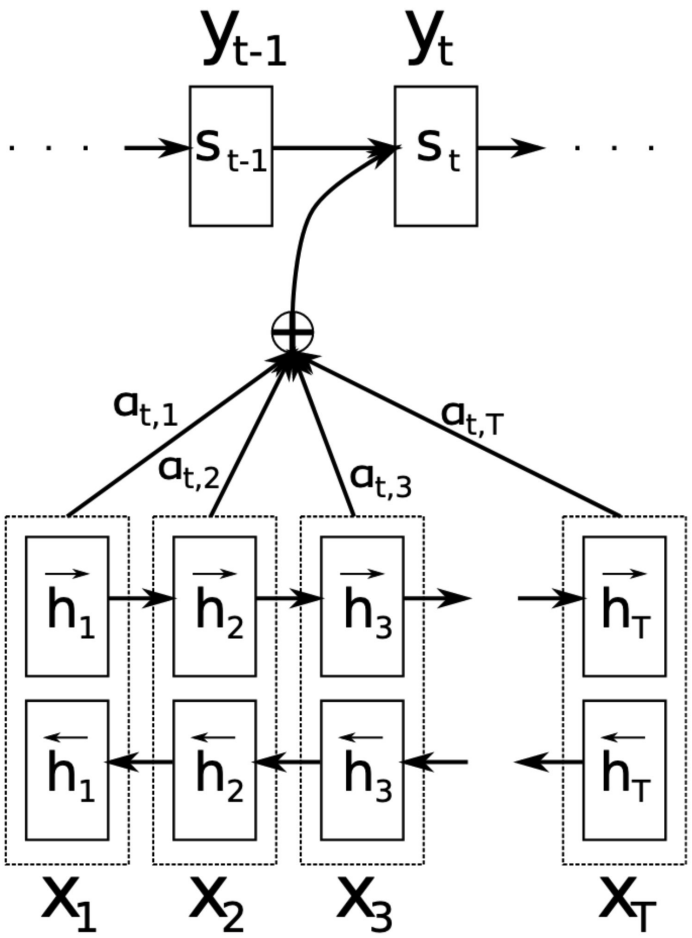
The Sweet Lesson:

It is often possible to outperform scaled-up methods by being more clever, and being clever can yield methods that scale better.

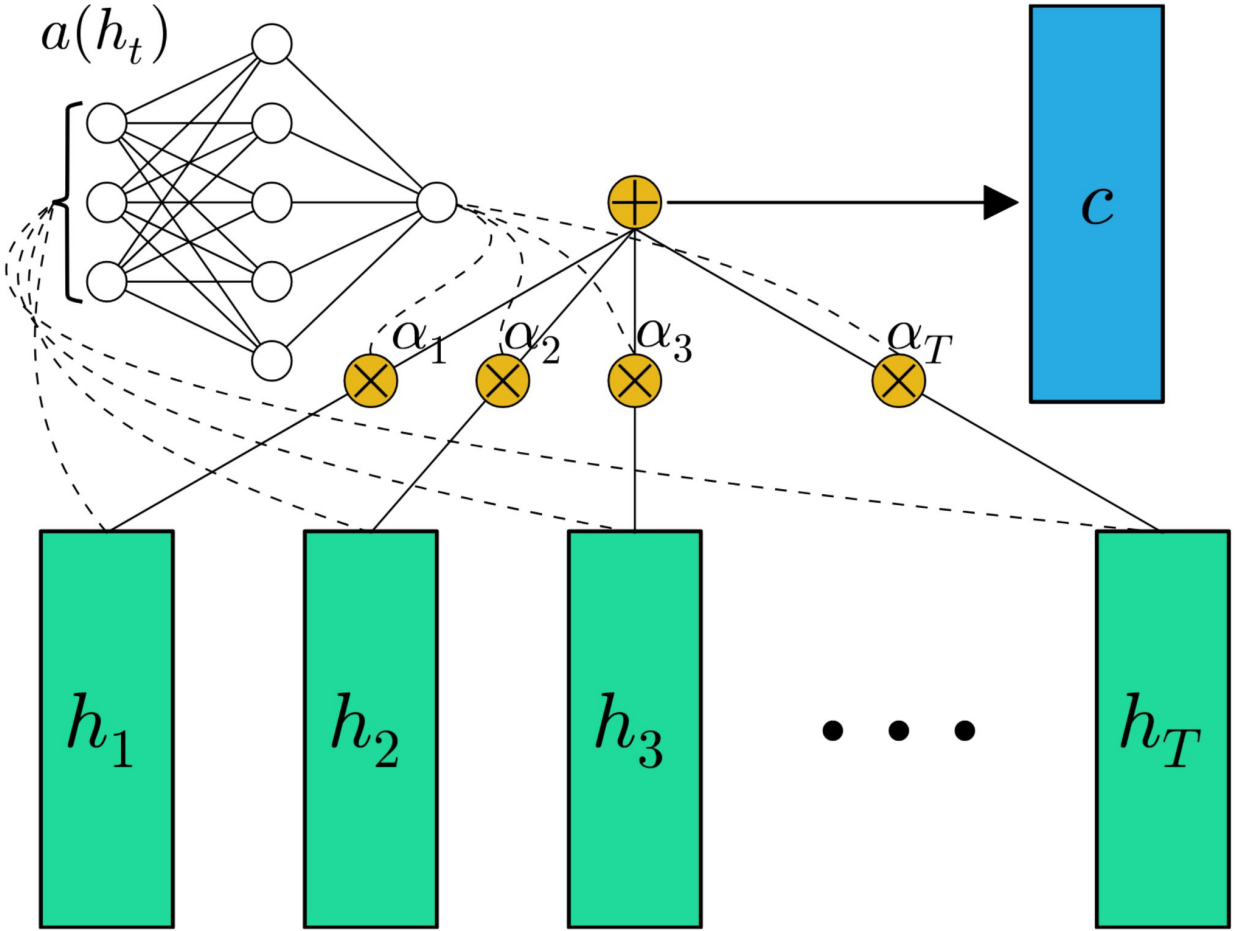


A C++ implementation of deep LSTM with the configuration from the previous section on a single GPU processes a speed of approximately 1,700 words per second. This was too slow for our purposes, ***so we parallelized our model using an 8-GPU machine.***

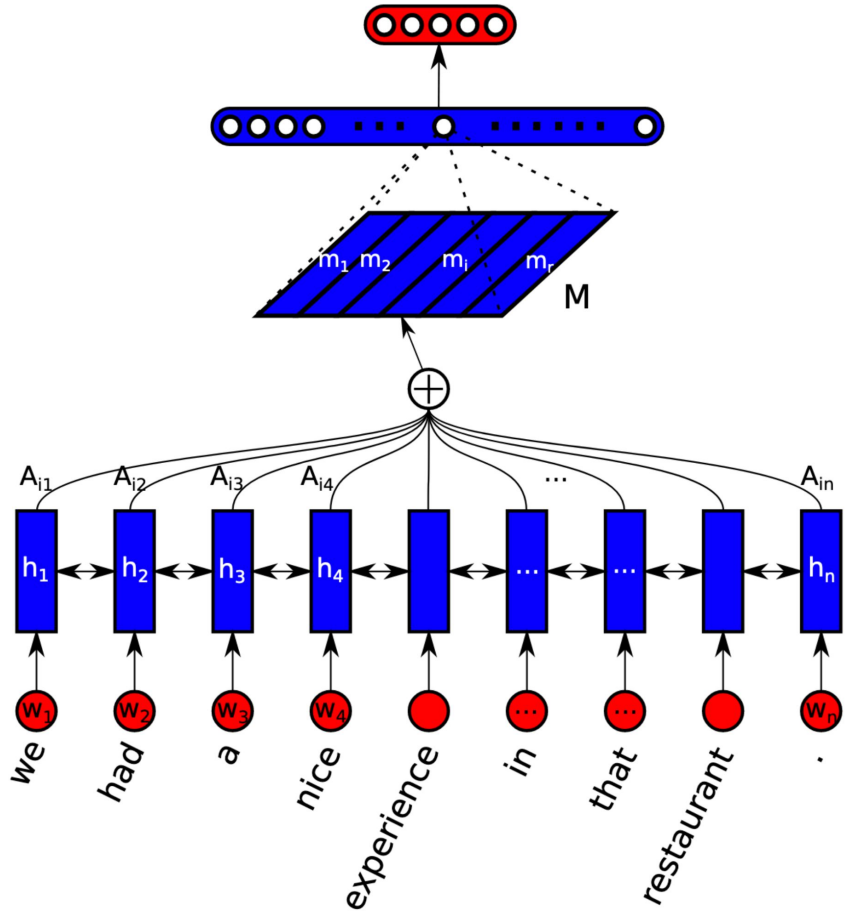
From "Sequence to Sequence Learning with Neural Networks" by Sutskever et al.



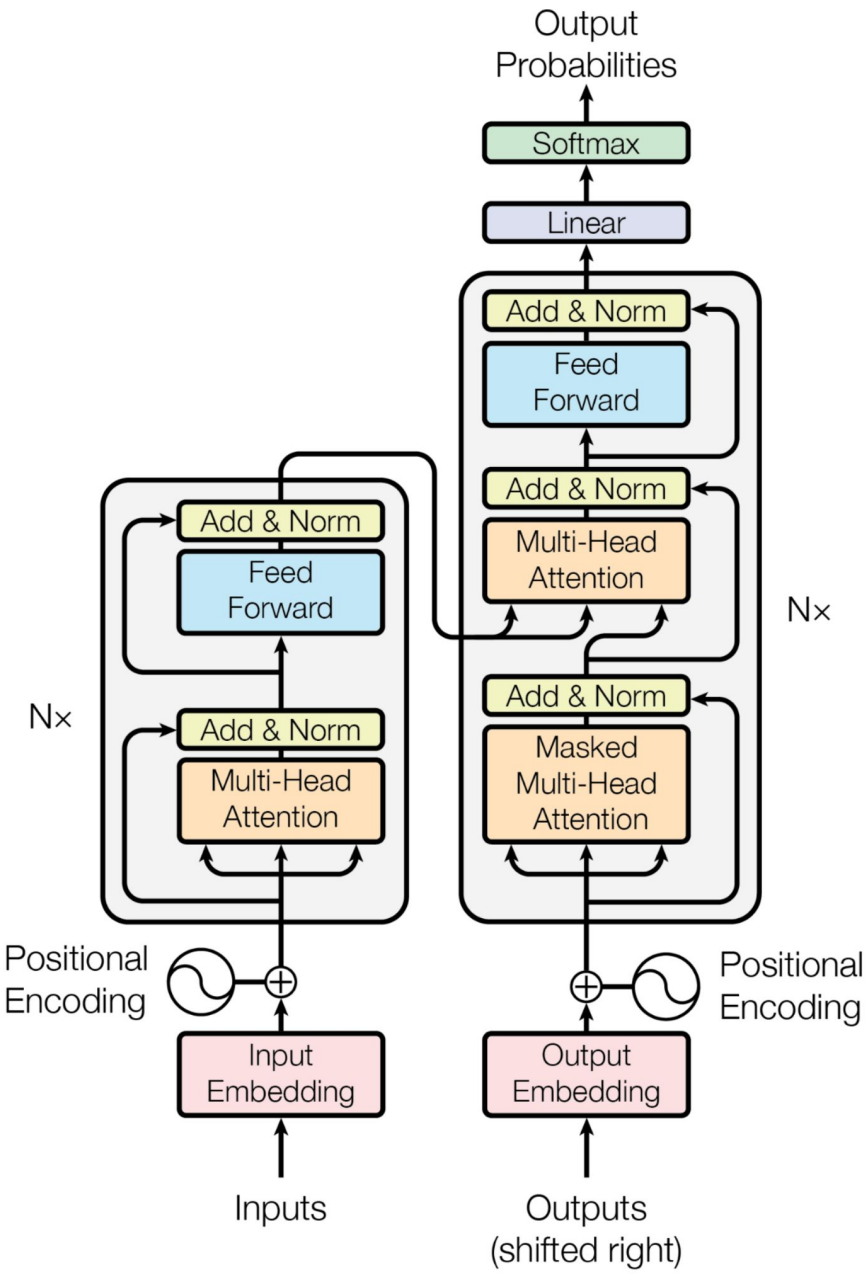
From "Neural Machine Translation by Jointly Learning to Align and Translate" by Bahdanau et al.



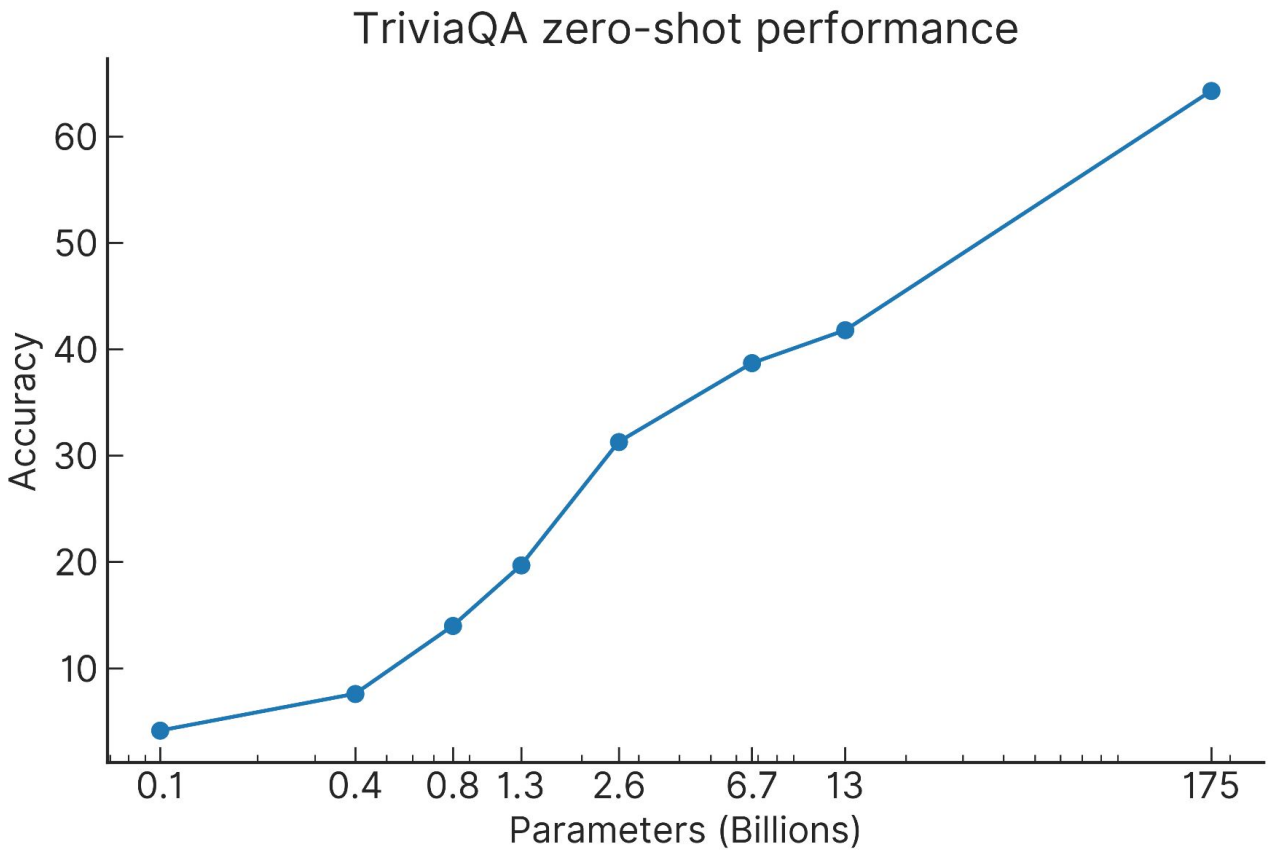
From "Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems" by Raffel and Ellis



From "A Structured Self-Attentive Sentence Embedding" by Lin et al.



From "Attention is All You Need" by Vaswani et al.



From "Language Models are Few-Shot Learners" by Brown et al.

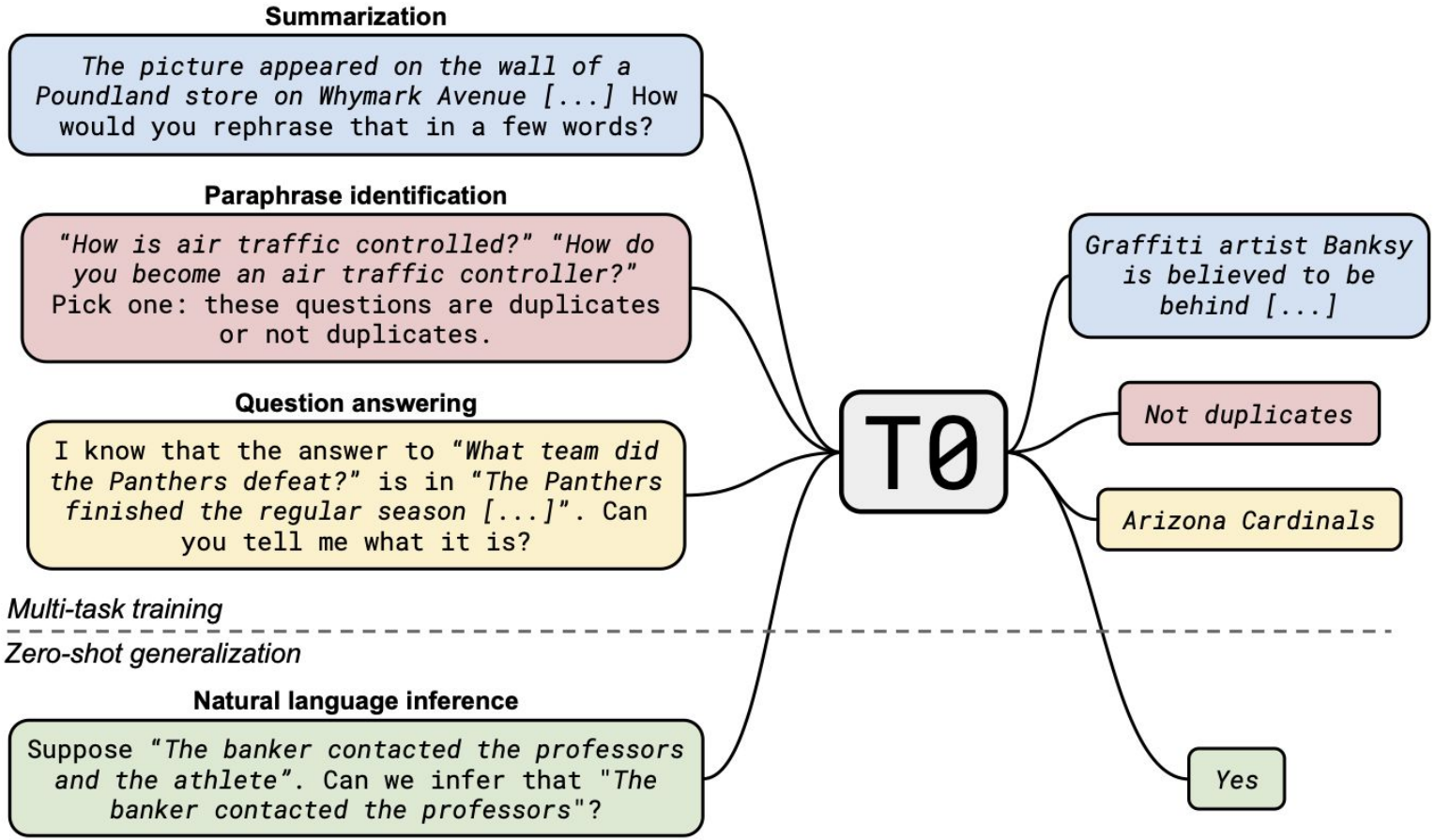
Closed-book question answering
http://www.autosweblog.com/cat/trivia-questions-from-the-50s
who was frank sinatra? a: an american singer, actor, and producer.

Paraphrase identification
https://www.usingenglish.com/forum/threads/60200-Do-these-sentences-mean-the-same
Do these sentences mean the same? No other boy in this class is as smart as the boy. No other boy is as smart as the boy in this class.

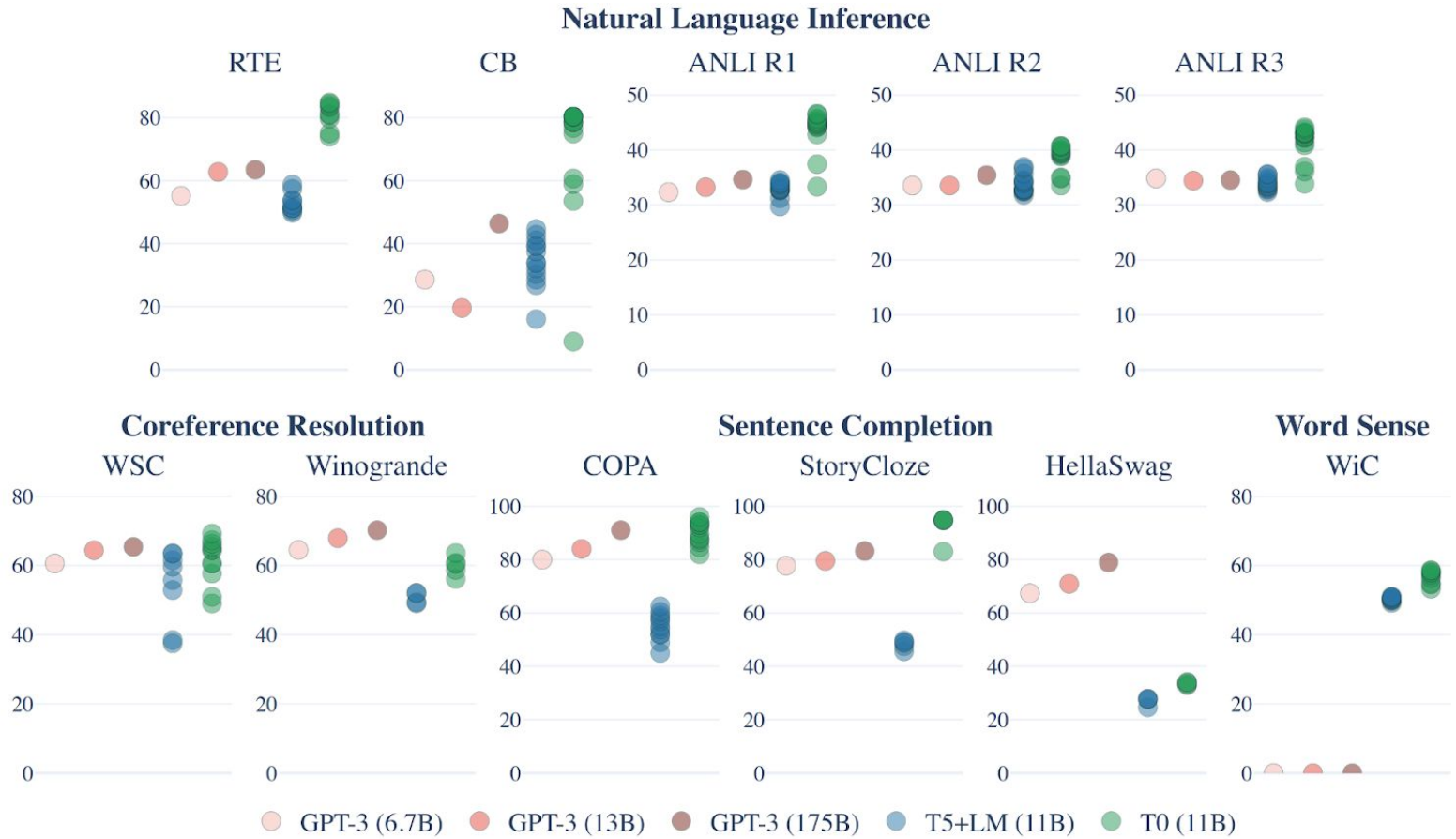
Natural Language Inference
https://ell.stackexchange.com/questions/121446/what-does-this-sentence-imply
If I say: He has worked there for 3 years. does this imply that he is still working at the moment of speaking?

Summarization
https://blog.nytsoi.net/tag/reddit
... Lately I've been seeing a pattern regarding videos stolen from other YouTube channels, reuploaded and monetized with ads. These videos are then mass posted on Reddit by bots masquerading as real users. tl;dr: Spambots are posting links to stolen videos on Reddit, copying comments from others to masquerade as legitimate users.

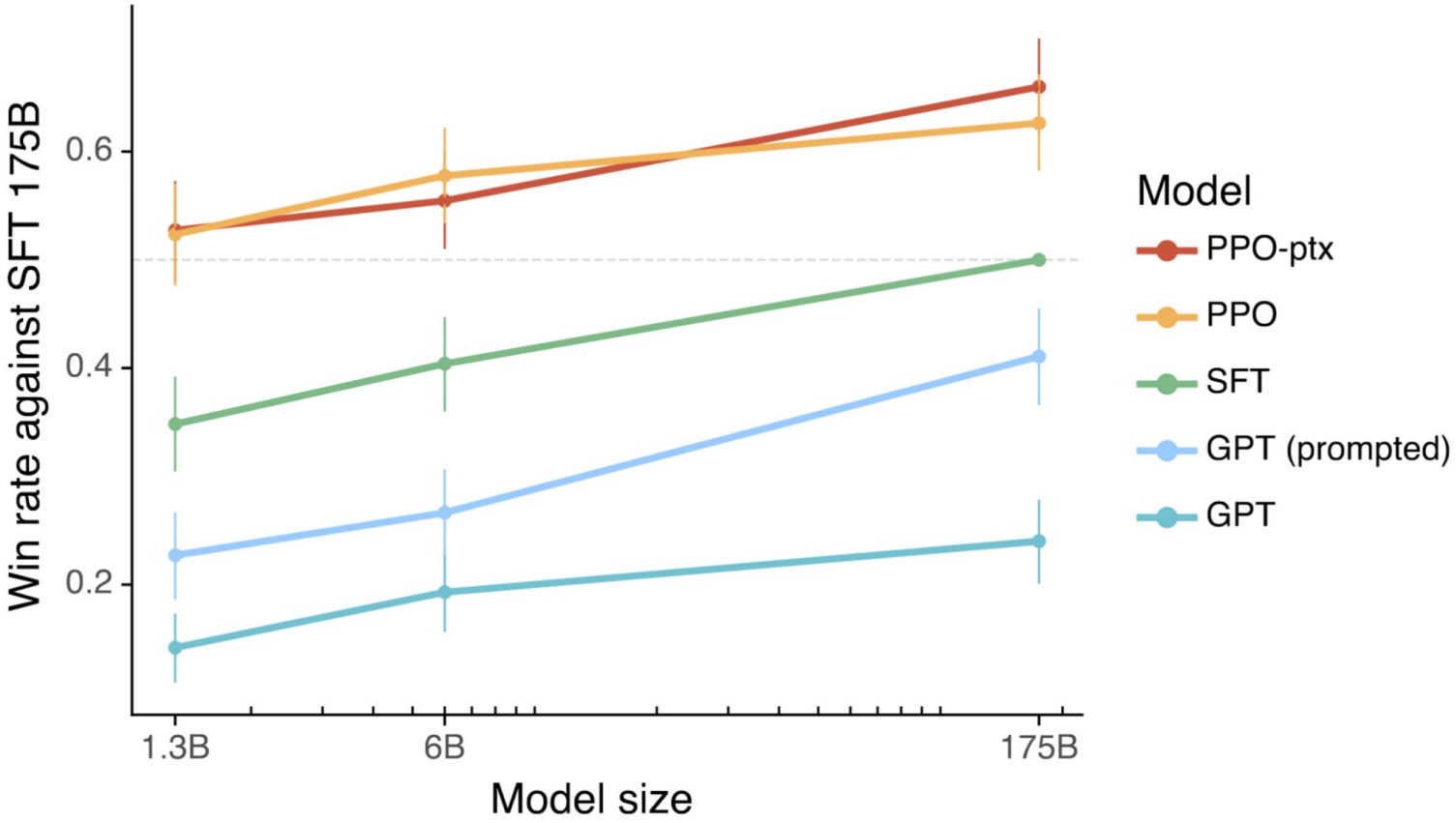
Pronoun resolution
https://nursecheung.com/ati-teas-guide-to-english-language-usage-understanding-pronouns/
Jennifer is a vegetarian, so she will order a nonmeat entrée. In this example, the pronoun she is used to refer to Jennifer.



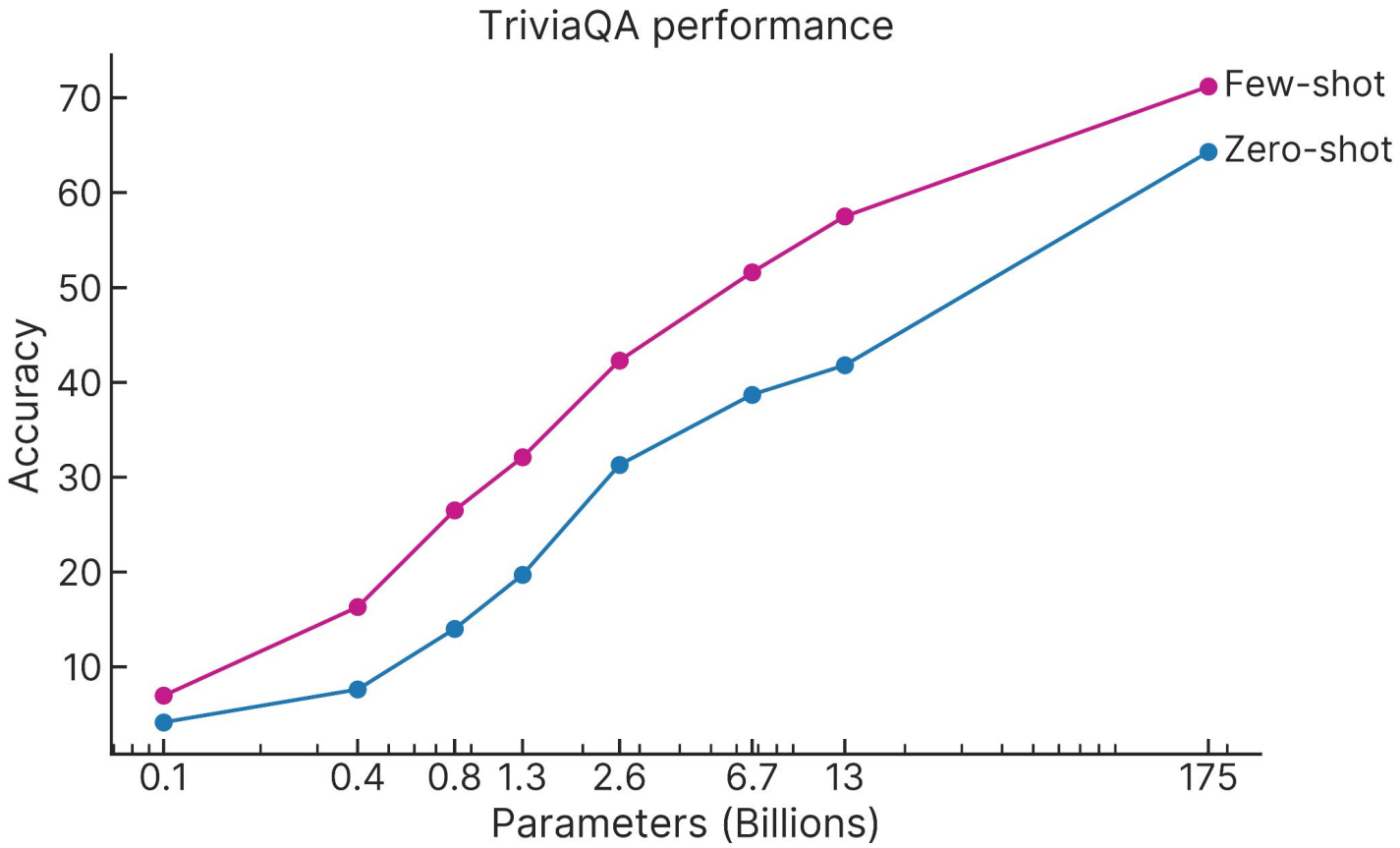
From "Multitask Prompted Training Enables Zero-Shot Task Generalization" by Sanh et al.



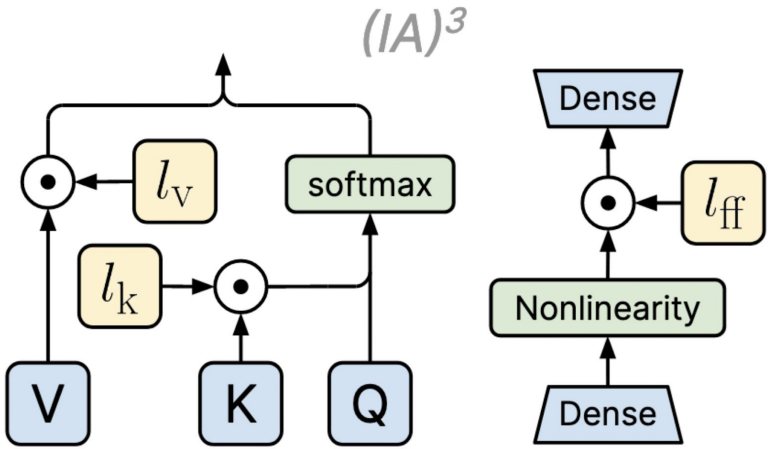
From "Multitask Prompted Training Enables Zero-Shot Task Generalization" by Sanh et al.

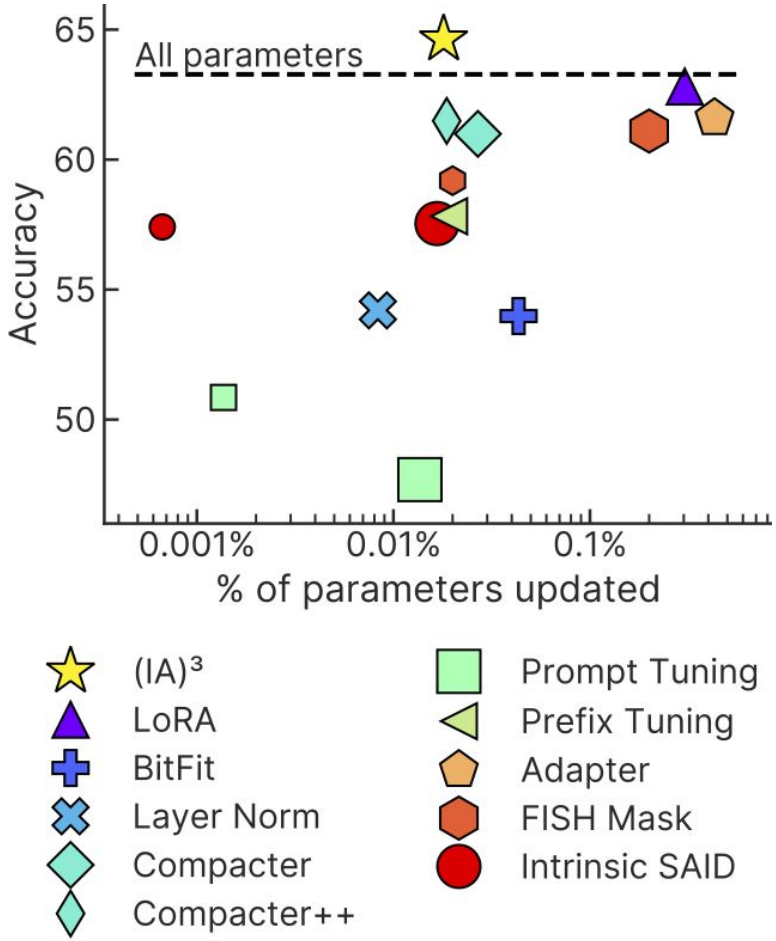


From "Training language models to follow instructions with human feedback" by Ouyang et al.

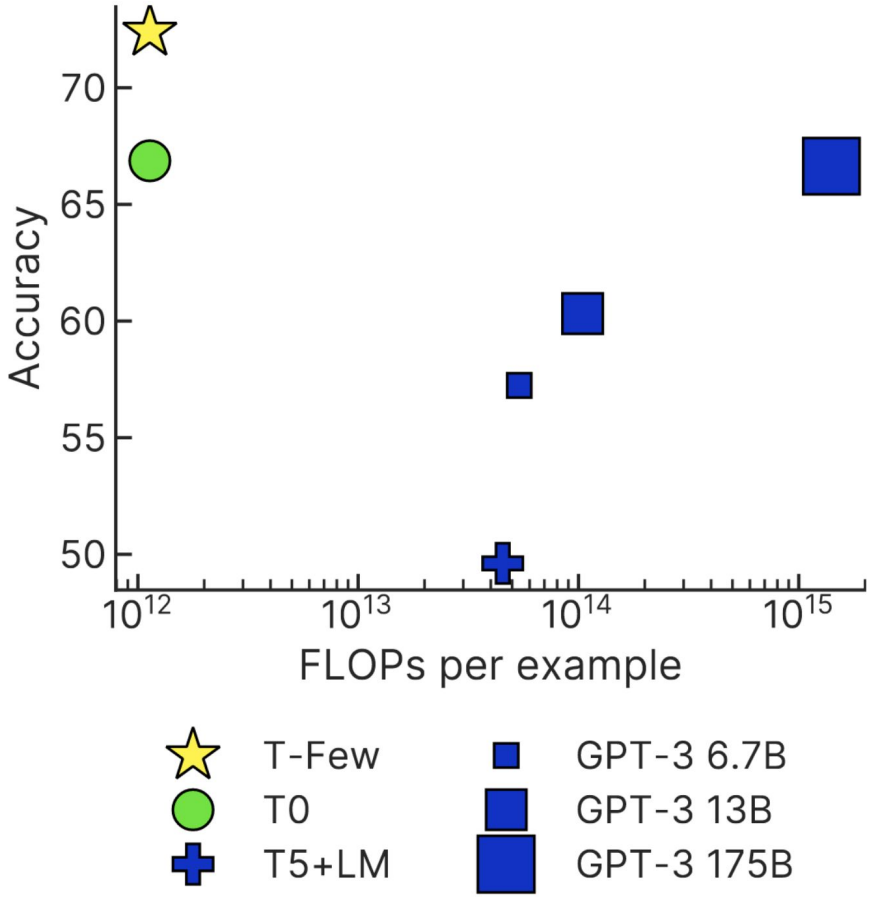


From "Language Models are Few-Shot Learners" by Brown et al.





From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022



Method	Acc.
`T-Few`	75.8%
Human baseline [2]	73.5%
PET [50]	69.6%
SetFit [51]	66.9%
GPT-3 [4]	62.7%

Table 2: Top-5 best methods on RAFT as of writing. `T-Few` is the first method to outperform the human baseline and achieves over 6% higher accuracy than the next-best method.

From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022

Thanks.

Please give me feedback:

http://bit.ly/colin-talk-feedback