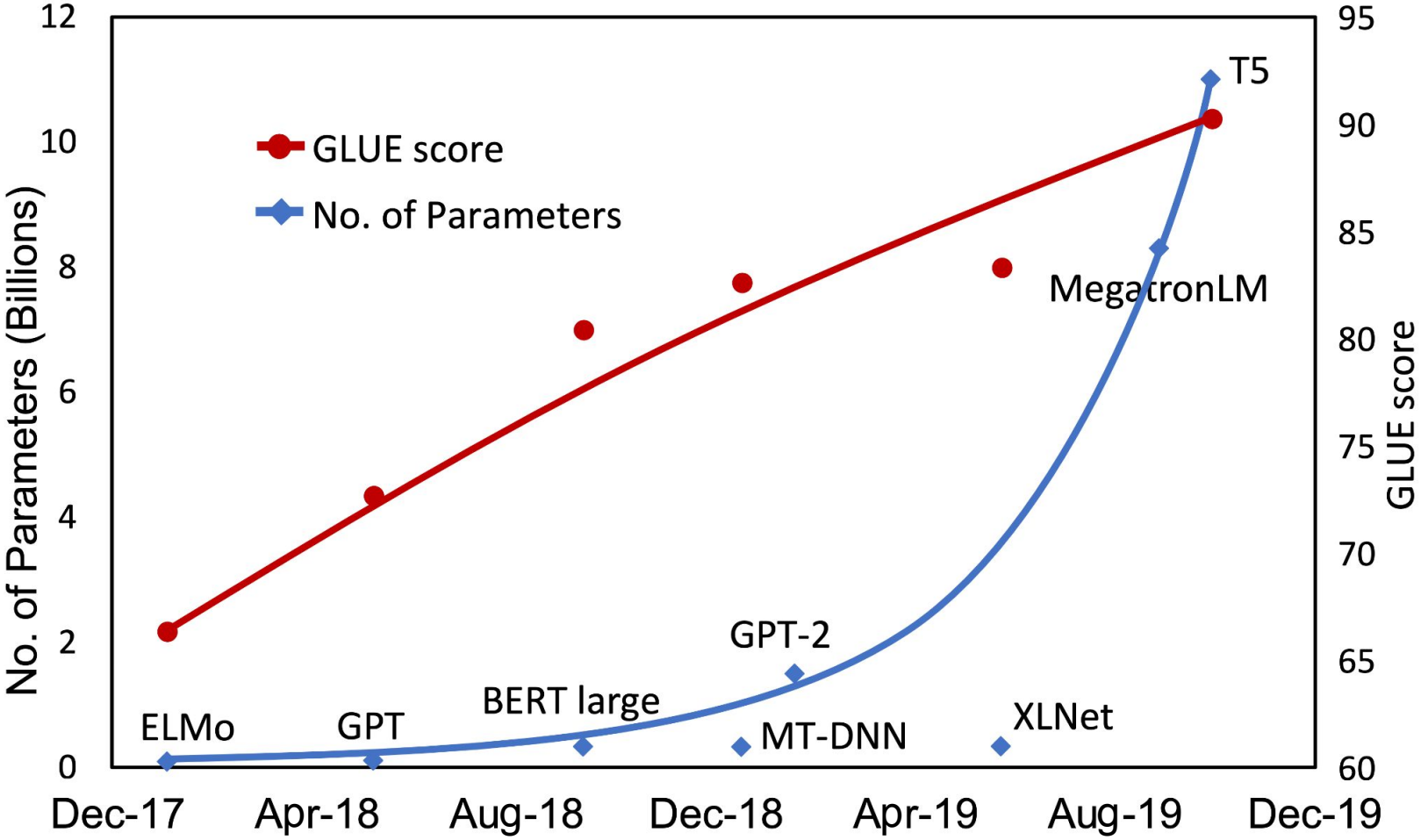
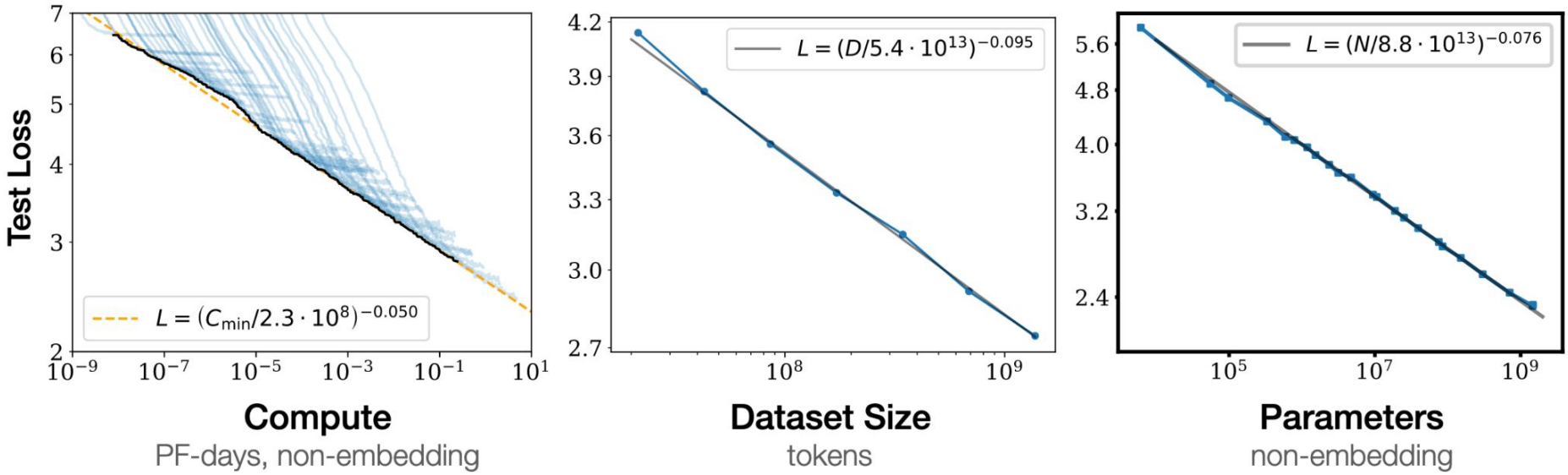
# The Sweet Lesson

Colin Raffel




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




*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.*

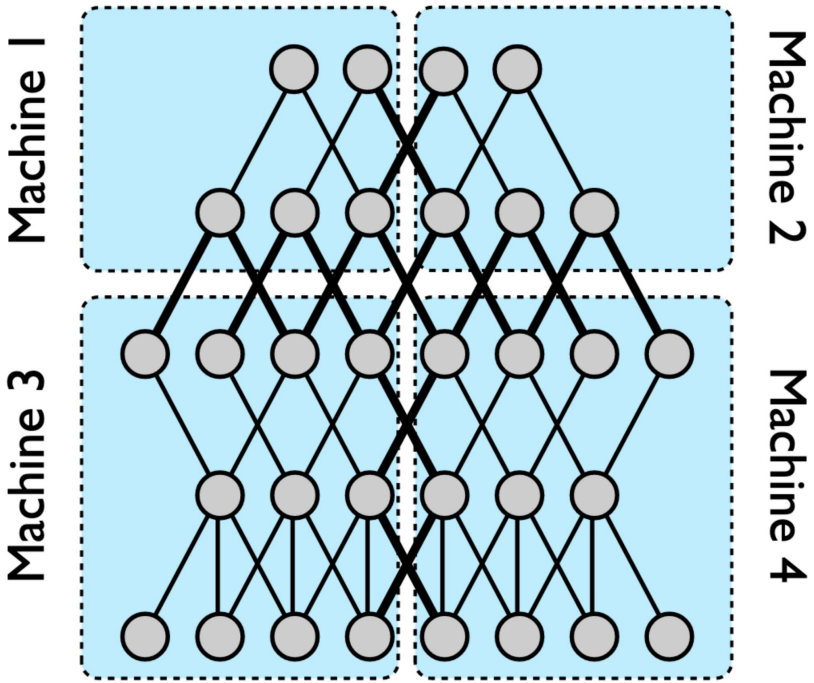
From “The Bitter Lesson” by Sutton

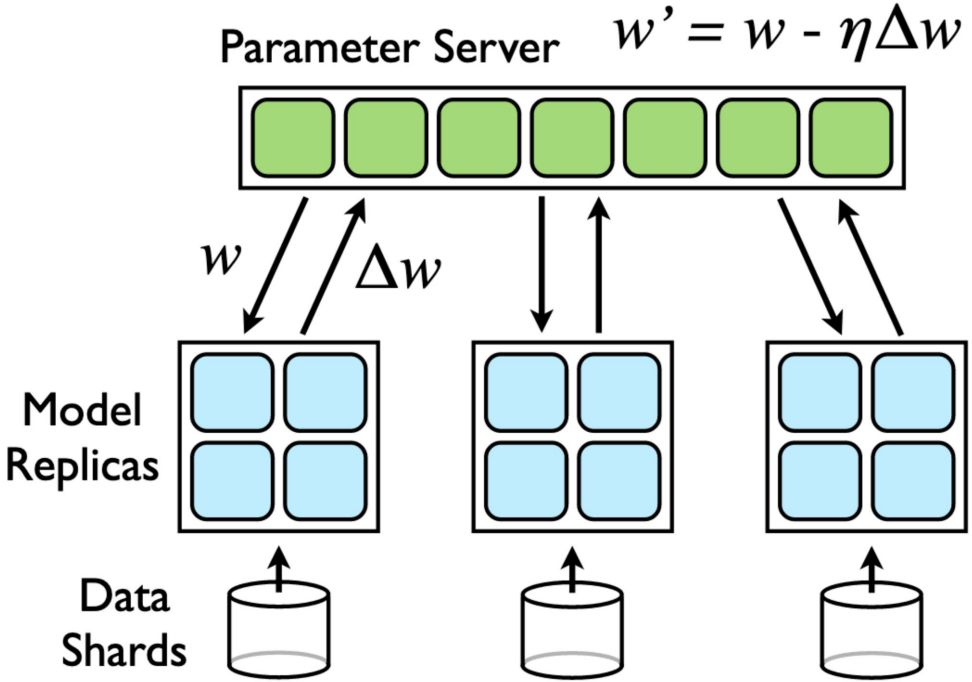
*The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are **ultimately** the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit 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?)

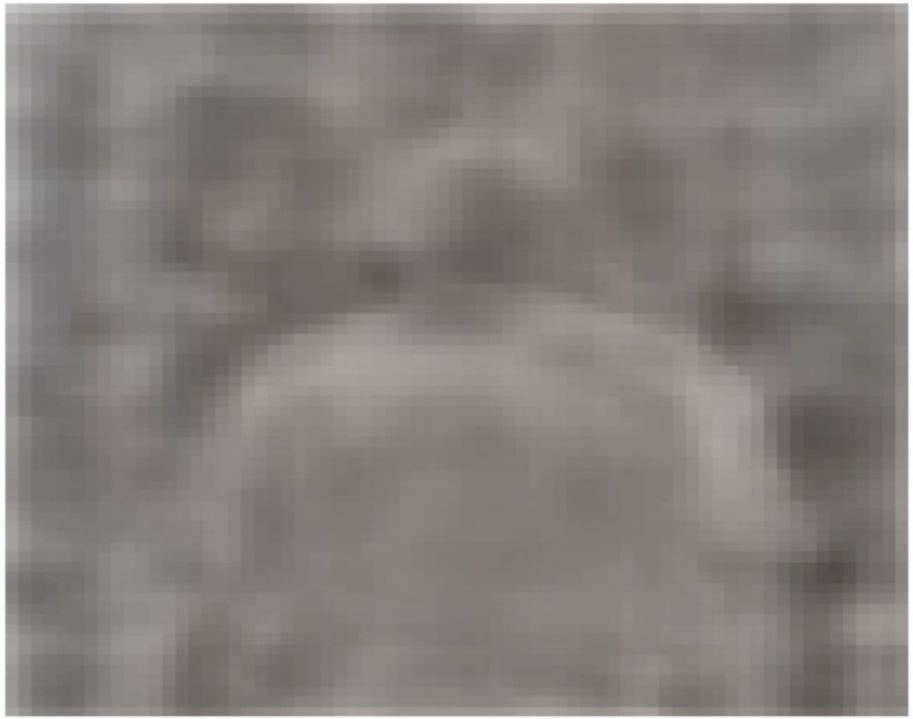
*The Sweet Lesson:*

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



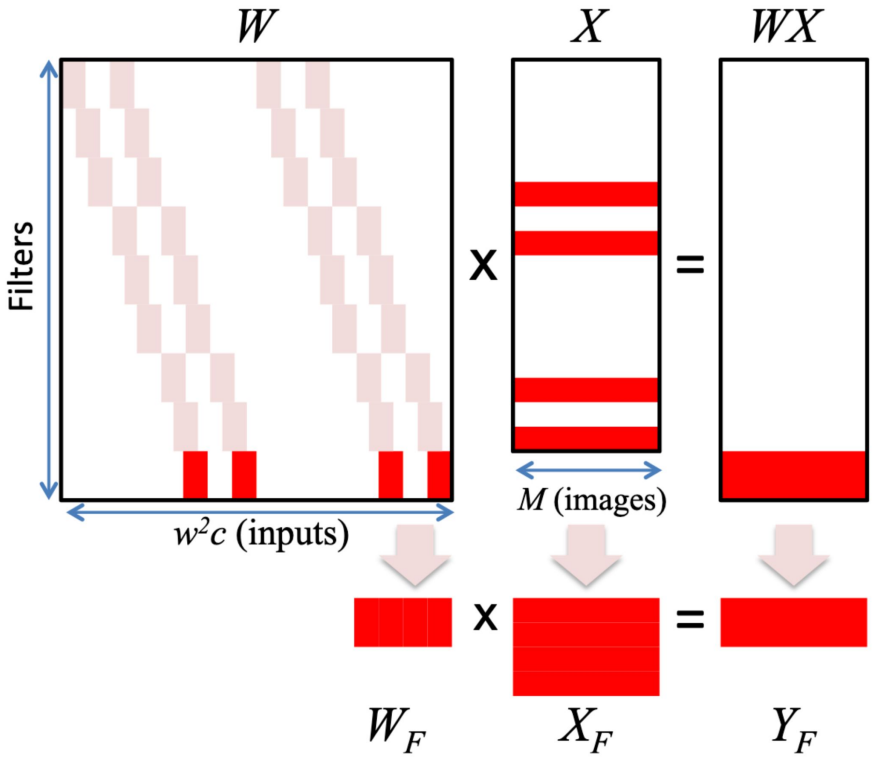

*From "Large Scale Distributed Deep Networks" by Dean et al.*

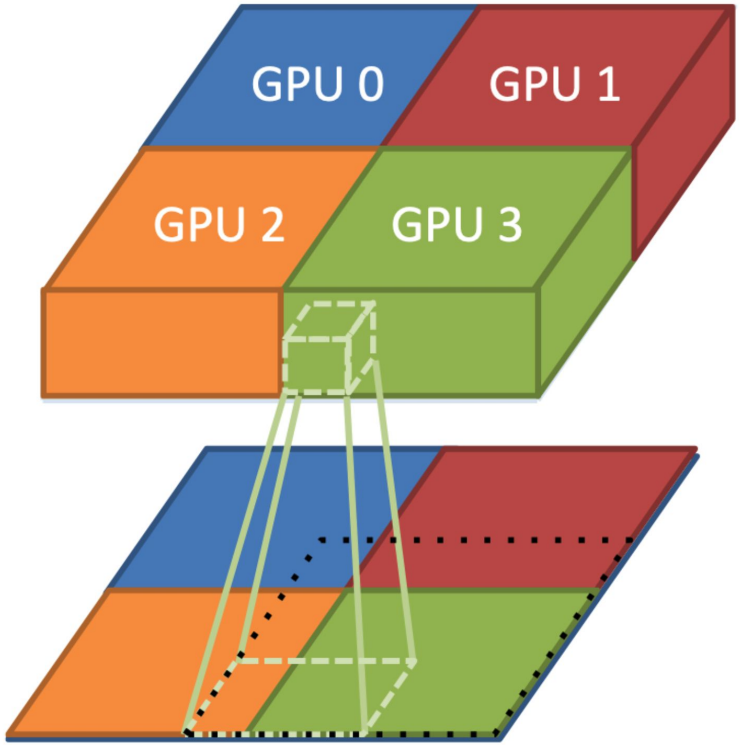


*From “Building High-level Features Using Large Scale Unsupervised Learning” by Le et al.*

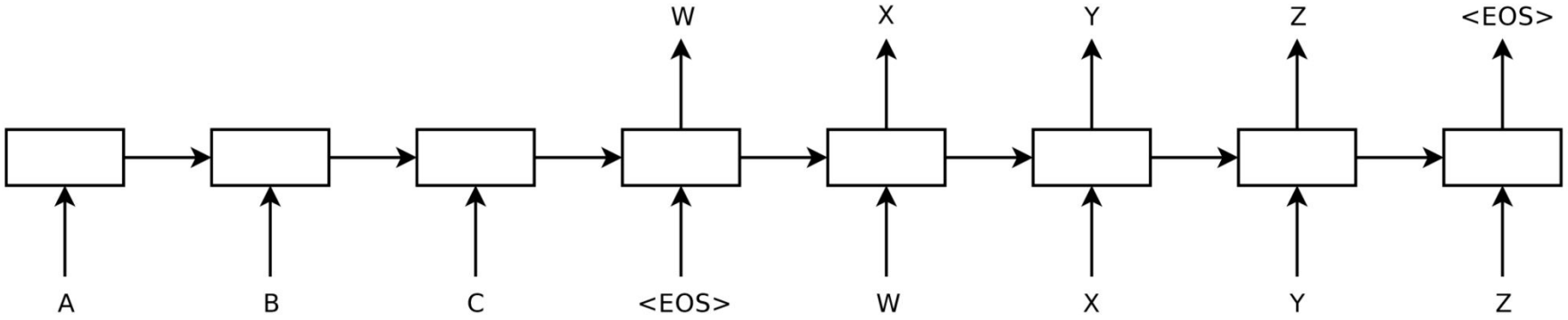
*The distributed computing infrastructure (known as "DistBelief") used for the experiments in (Le et al., 2012)* ***manages to train a neural network using 16000 CPU cores*** *(in 1000 machines) in just a few days, yet this level of resource is likely beyond those available to most deep learning researchers... In this paper we present an alternative approach to training such networks that leverages in-expensive computing power in the form of GPUs and introduces the use of high-speed communications infrastructure to tightly coordinate distributed gradient computations.* ***Our system trains neural networks at scales comparable to DistBelief with just 3 machines.***

*From "Deep learning with COTS HPC systems" by Coates et al.*



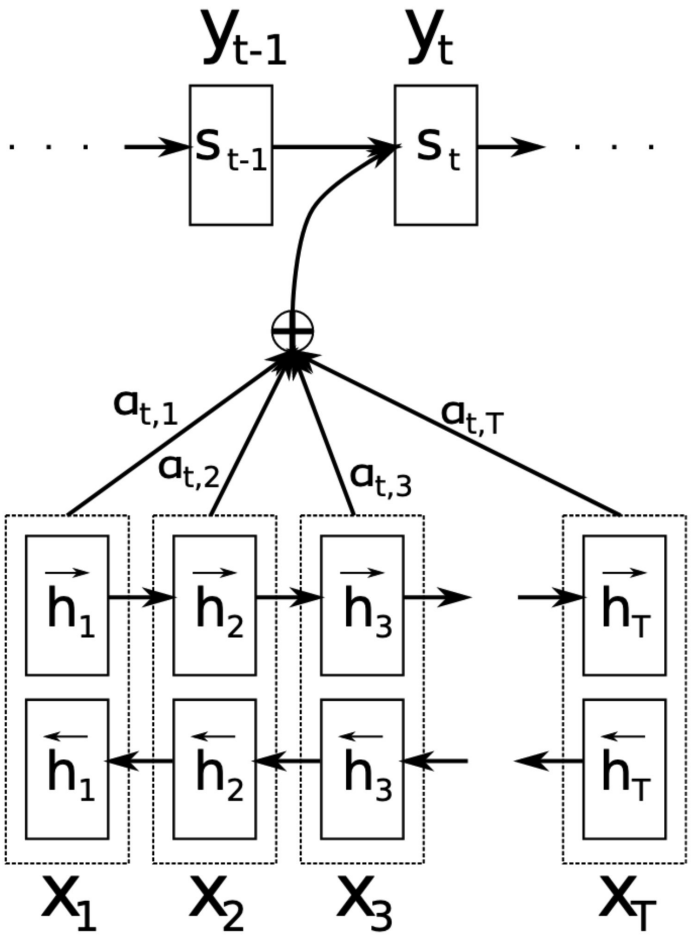




*From "Deep learning with COTS HPC systems" by Coates et al.*

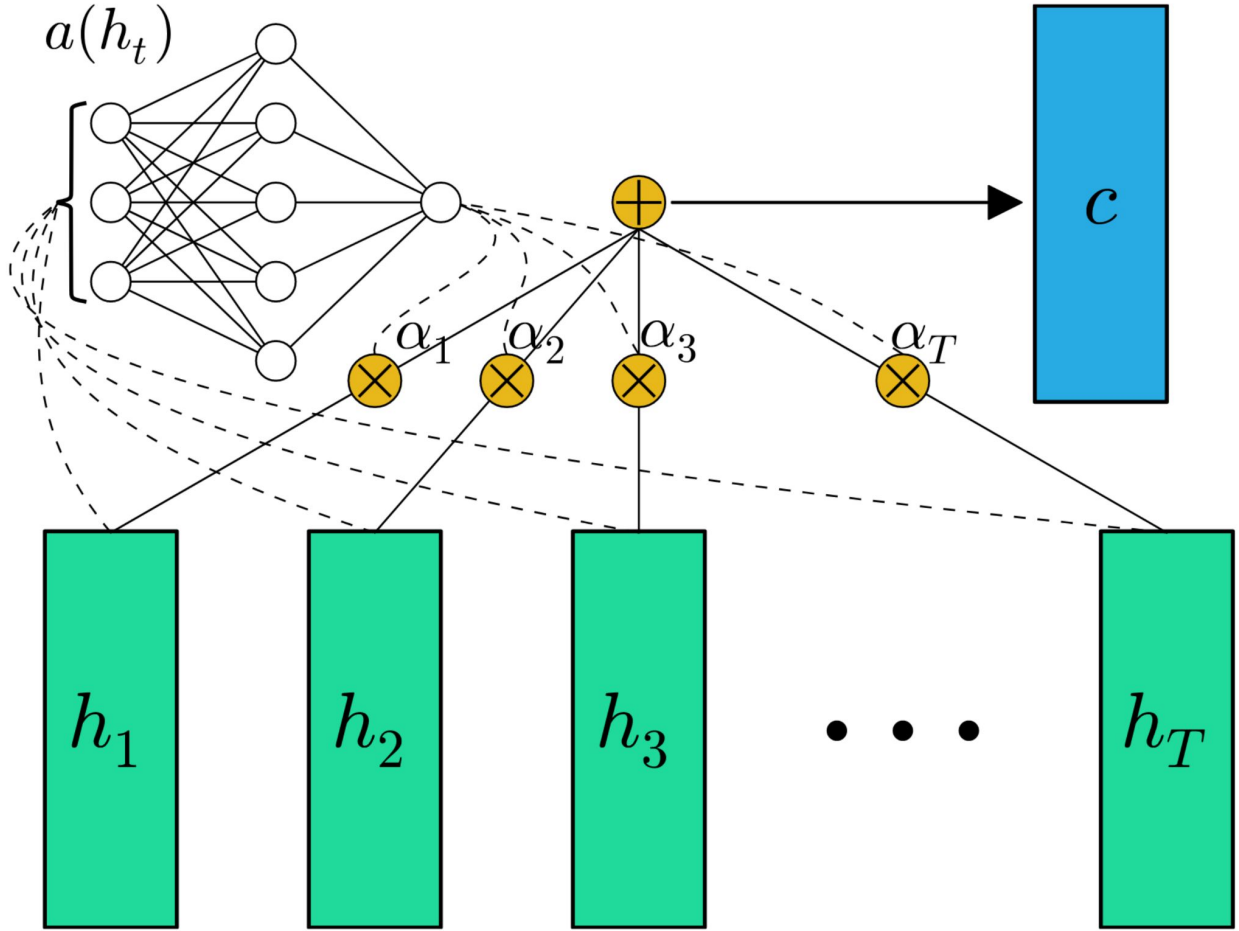



*A C++ implementation of deep LSTM with the configuration from the previous section on a single GPU processes a speed of approximately 1,700 words per second. This was too slow for our purposes,* ***so we parallelized our model using an 8-GPU machine.***

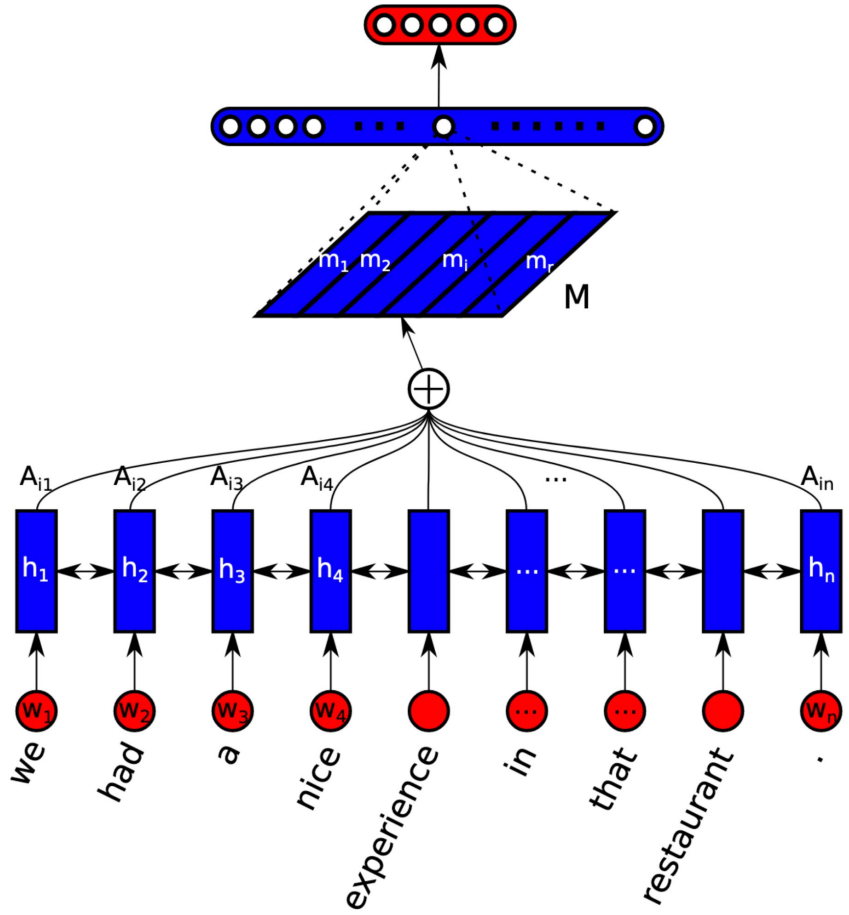
*From "Sequence to Sequence Learning with Neural Networks" by Sutskever et al.*



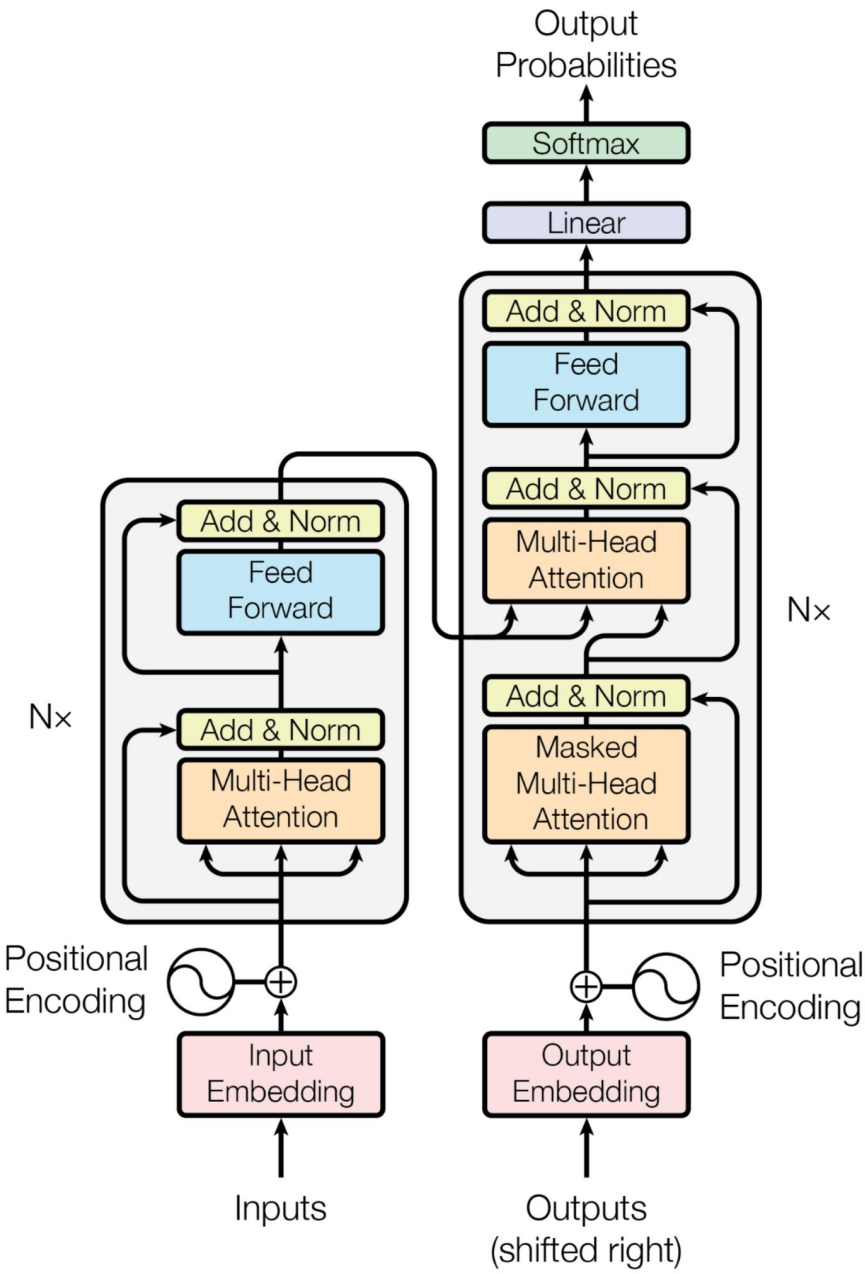

*From "Neural Machine Translation by Jointly Learning to Align and Translate" by Sutskever et al.*



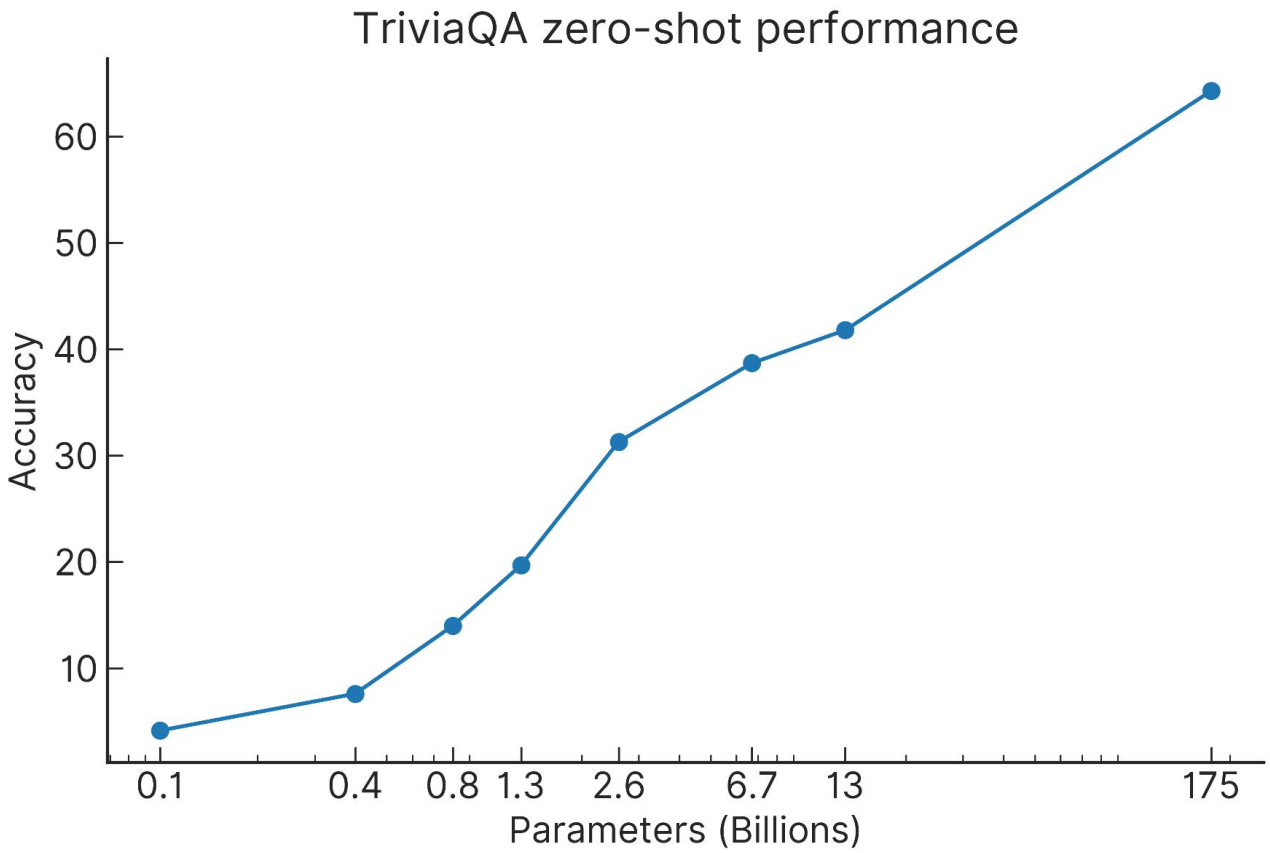

*From "Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems" by Raffel and Ellis*




*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.*

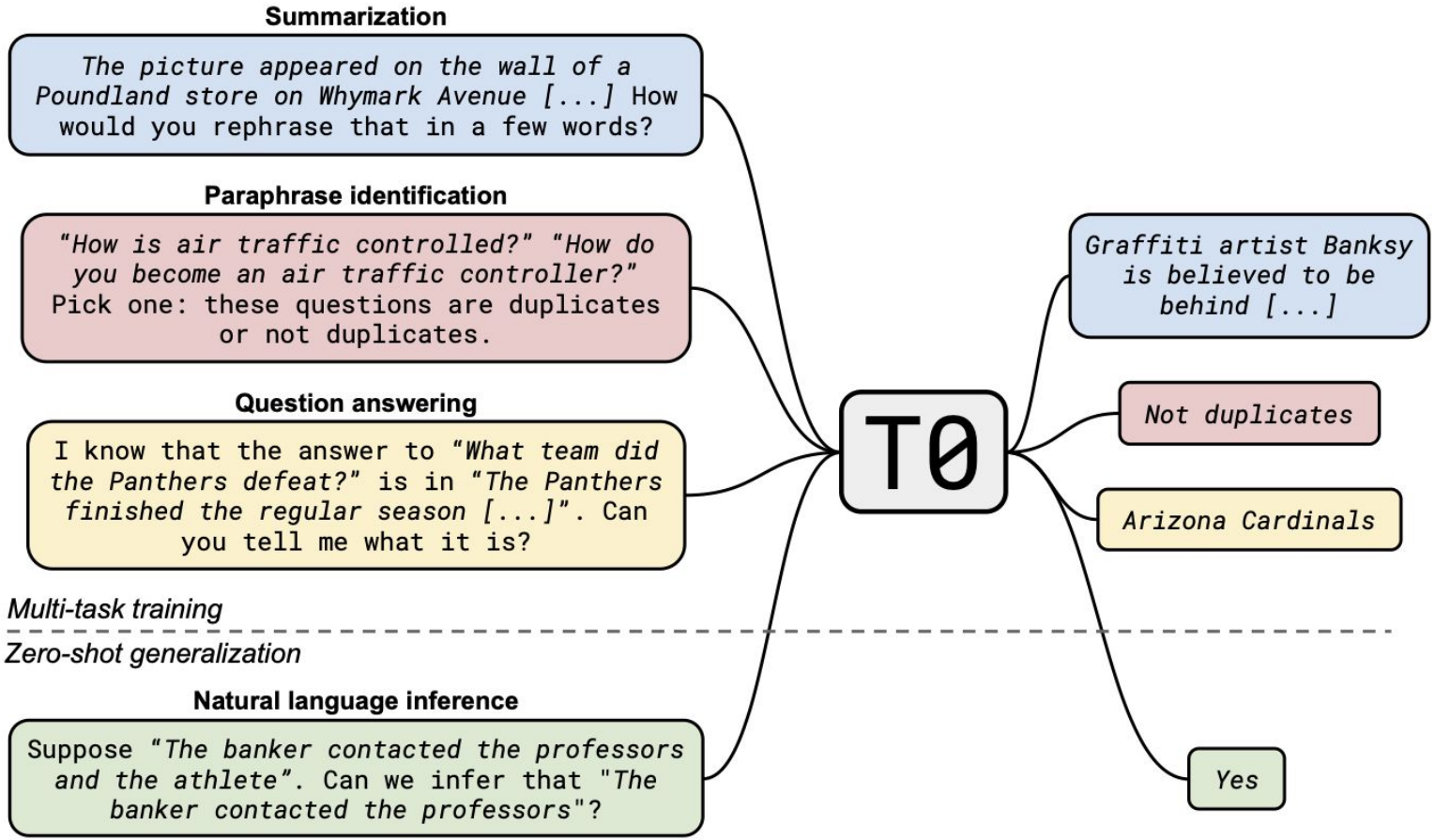
**Closed-book question answering**
http://www.autosweblog.com/cat/trivia-questions-from-the-50s
who was frank sinatra? a: an american singer, actor, and producer.

**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.

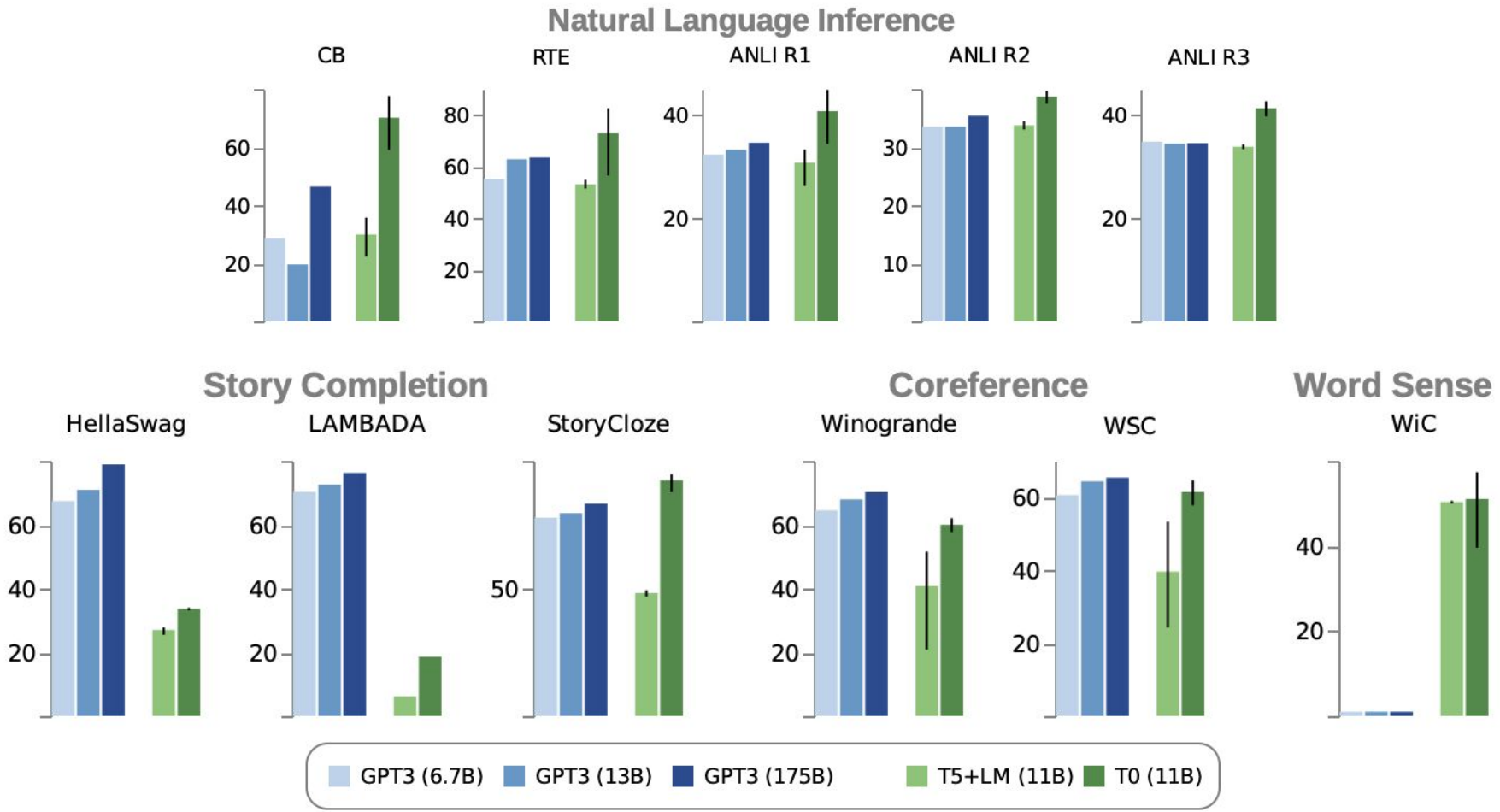
**Natural Language Inference**
https://ell.stackexchange.com/questions/121446/what-does-this-sentence-imply
If I say: He has worked there for 3 years. does this imply that he is still working at the moment of speaking?

**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.*

*“We should just scale methods up”*

Being clever yields methods that scale better!

*“We should just scale methods up”*

Being clever yields methods that scale better!

*“I can’t do NLP research unless I have massive scale”*

Lots of interesting problems to work on and opportunity for being clever!

*“We should just scale methods up”*

Being clever yields methods that scale better!

*“I can’t do NLP research unless I have massive scale”*

Lots of interesting problems to work on and opportunity for being clever!

*“Not having access to compute helps me be creative”*

Not having access to compute can suck!

*“We should just scale methods up”*

Being clever yields methods that scale better!

*“I can’t do NLP research unless I have massive scale”*

Lots of interesting problems to work on and opportunity for being clever!

*“Not having access to compute helps me be creative”*

Not having access to compute can suck!

*“If I scale up and spend lots of money, I shouldn’t have to release my model”*

Scaled-up results are harder to reproduce → release is more important!

*“We should just scale methods up”*

Being clever yields methods that scale better!

*“I can’t do NLP research unless I have massive scale”*

Lots of interesting problems to work on and opportunity for being clever!

*“Not having access to compute helps me be creative”*

Not having access to compute can suck!

*“If I scale up and spend lots of money, I shouldn’t have to release my model”*

Scaled-up results are harder to reproduce → release is more important!

*“Scale is bad, we shouldn’t scale up”*

Scale often shows us what’s possible!

*“We should just scale methods up”*

Being clever yields methods that scale better!

*“I can’t do NLP research unless I have massive scale”*

Lots of interesting problems to work on and opportunity for being clever!

*“Not having access to compute helps me be creative”*

Not having access to compute can suck!

*“If I scale up and spend lots of money, I shouldn’t have to release my model”*

Scaled-up results are harder to reproduce → release is more important!

*“Scale is bad, we shouldn’t scale up”*

Scale often shows us what’s possible!

*“Scale for scale’s sake is good”*

Scale without measuring performance is meaningless!

Thanks.

Please give me feedback:
http://bit.ly/colin-talk-feedback