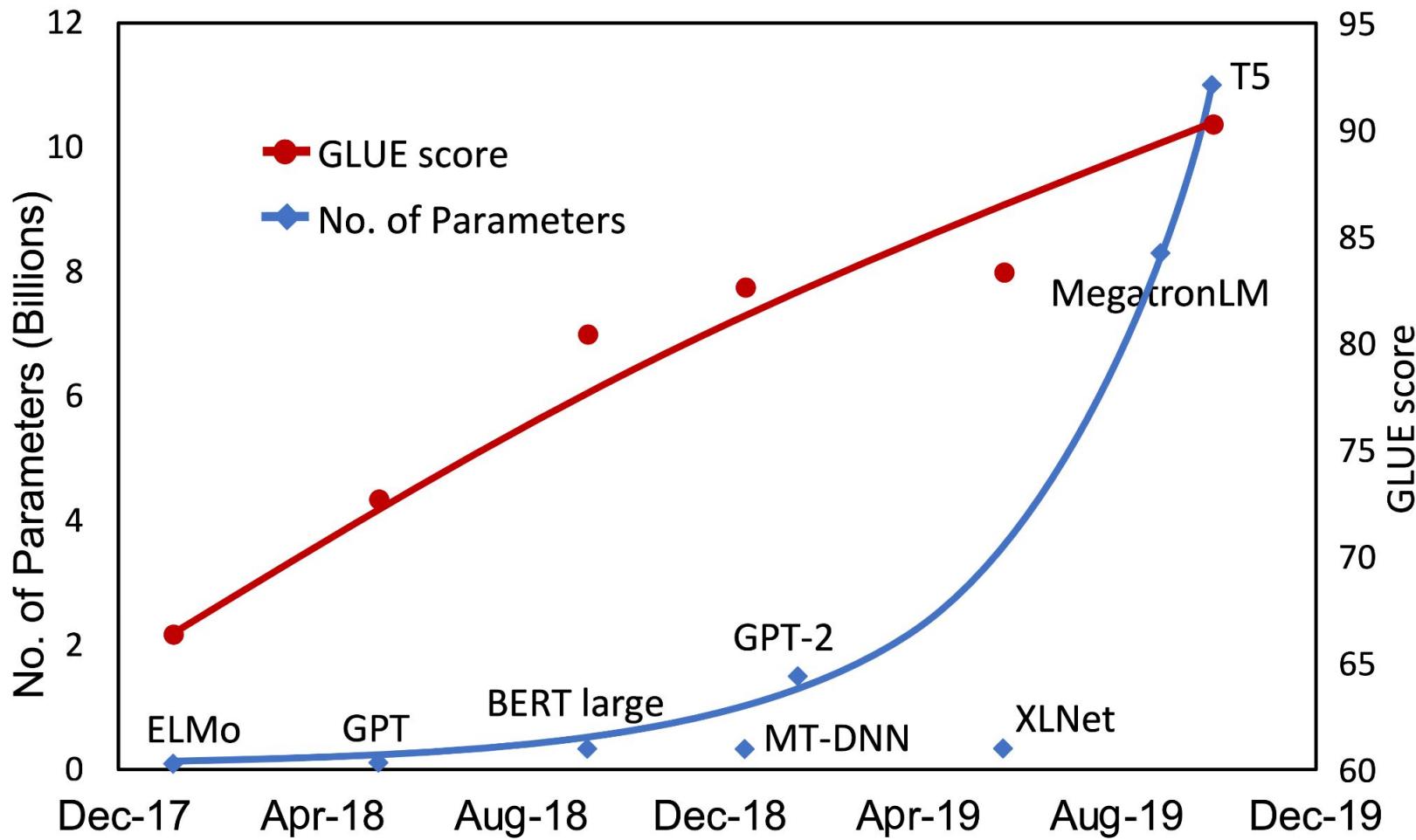
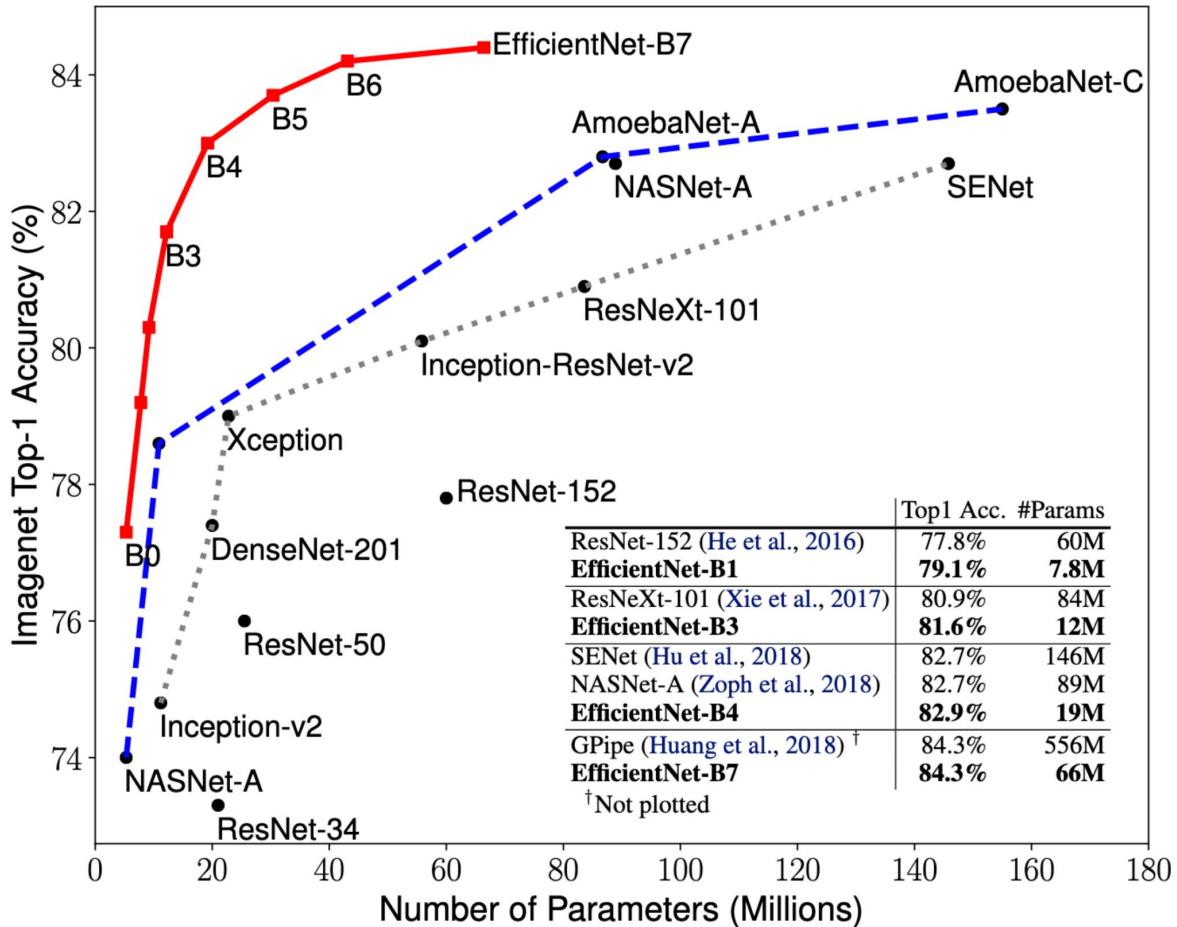


How and why should(n't) we scale machine learning?

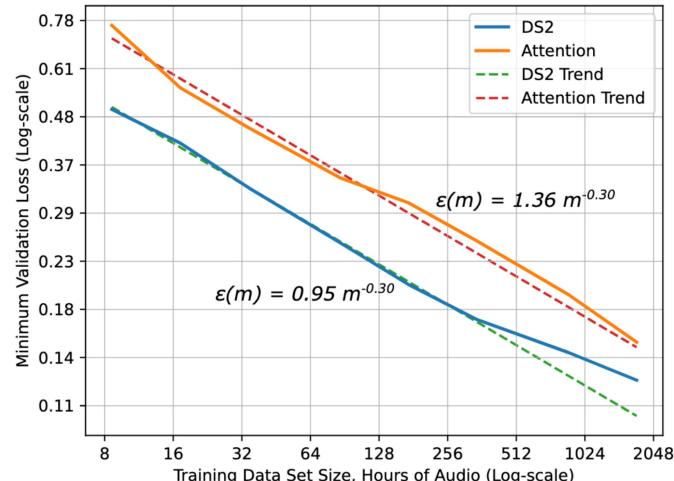
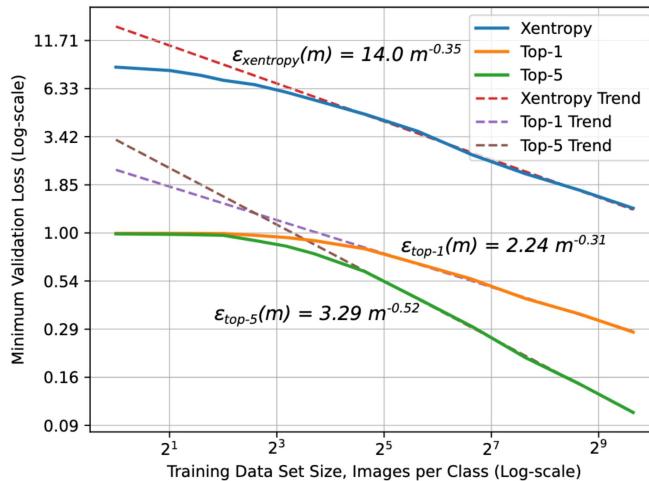
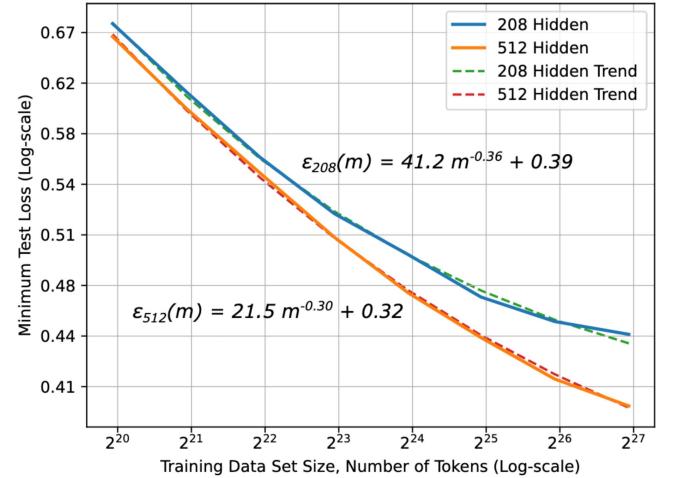
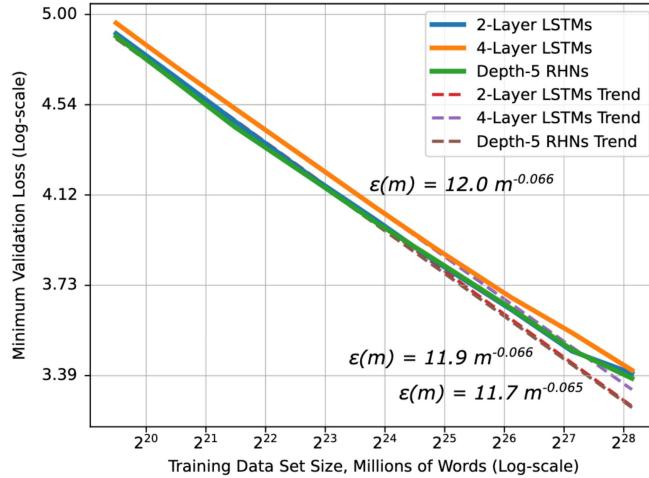
Colin Raffel

IBM AI Hardware Forum

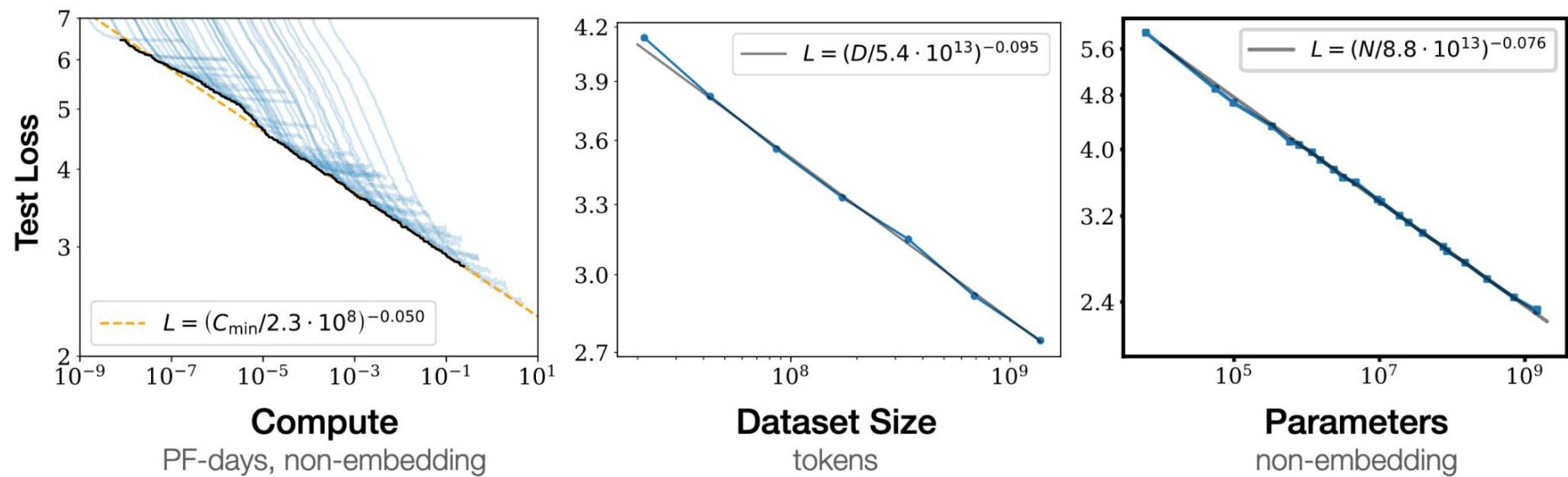


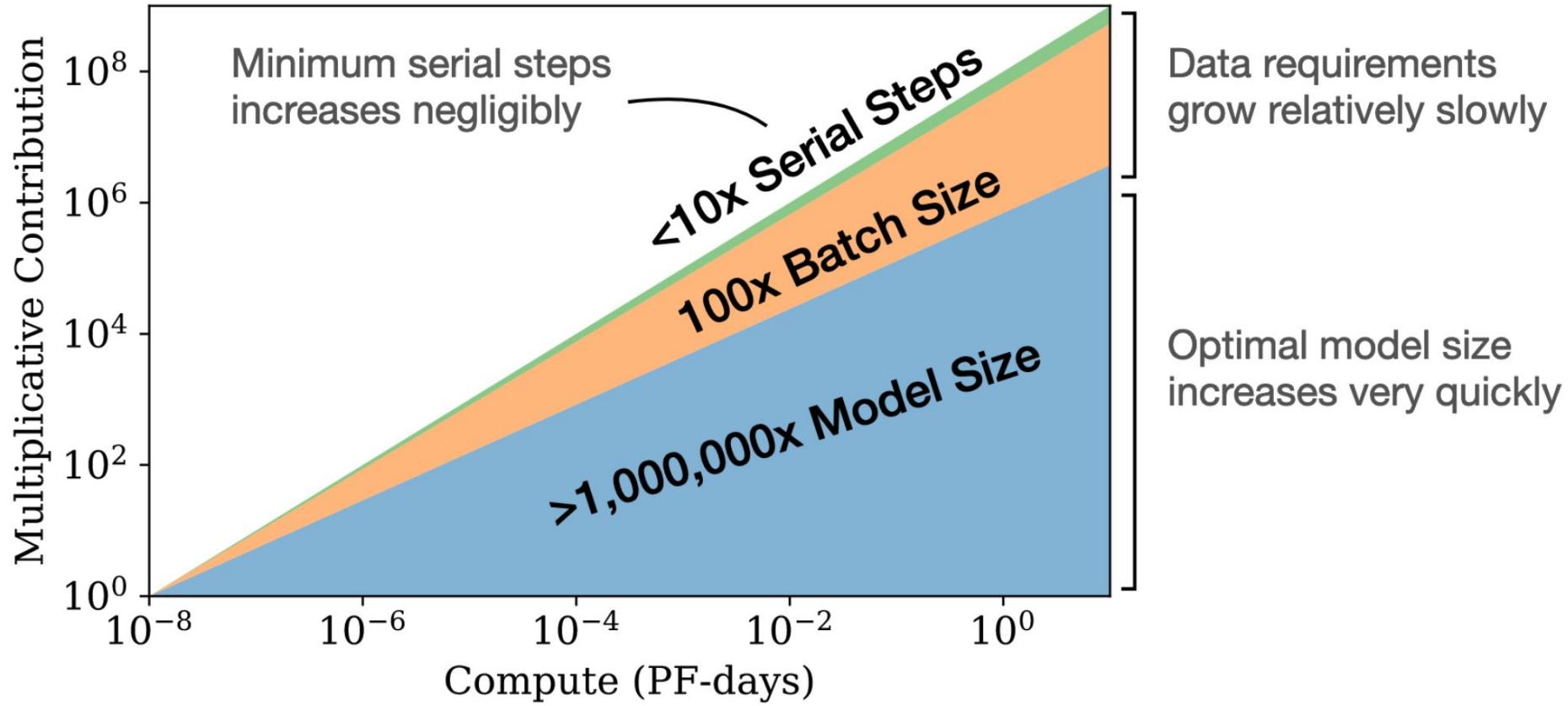


From “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks” by Tan and Le

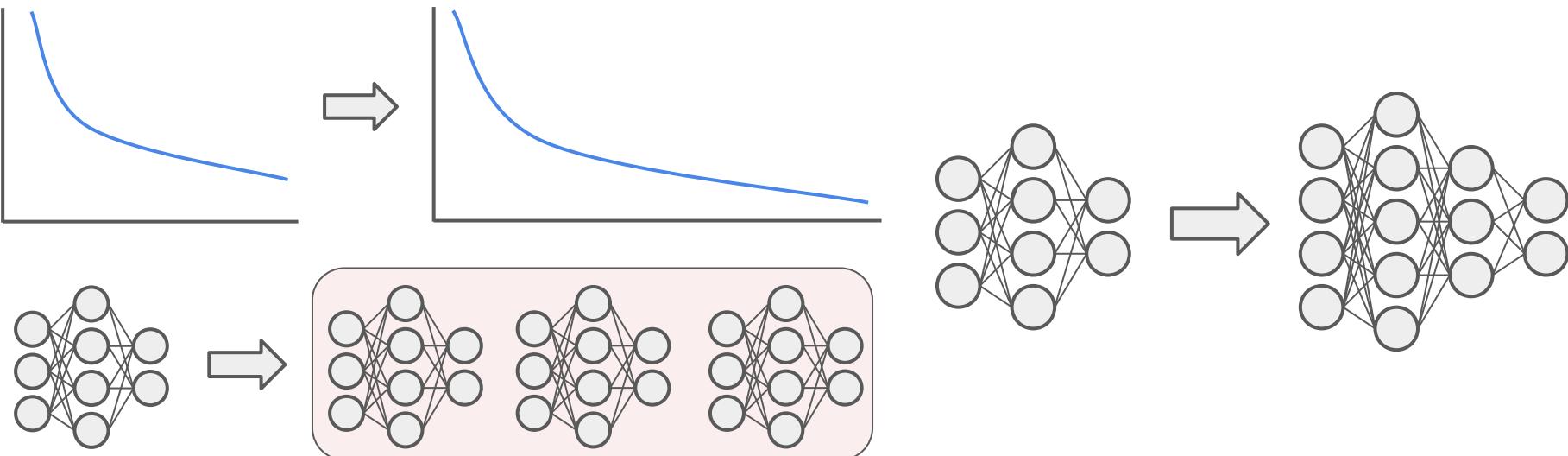


From “Deep Learning Scaling is Predictable, Empirically” by Hestness et al.





Scaling strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline	83.28	19.24	80.88	71.36	26.98	39.82	27.65
1× size, 4× training steps	85.33	19.33	82.45	74.72	27.08	40.66	27.93
1× size, 4× batch size	84.60	19.42	82.52	74.64	27.07	40.60	27.84
2× size, 2× training steps	86.18	19.66	84.18	77.18	27.52	41.03	28.19
4× size, 1× training steps	85.91	19.73	83.86	78.04	27.47	40.71	28.10
4× ensembled	84.77	20.10	83.09	71.74	28.05	40.53	28.57
4× ensembled, fine-tune only	84.05	19.57	82.36	71.55	27.55	40.22	28.09



From “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer” by Raffel et al.

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

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

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

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

Memory

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

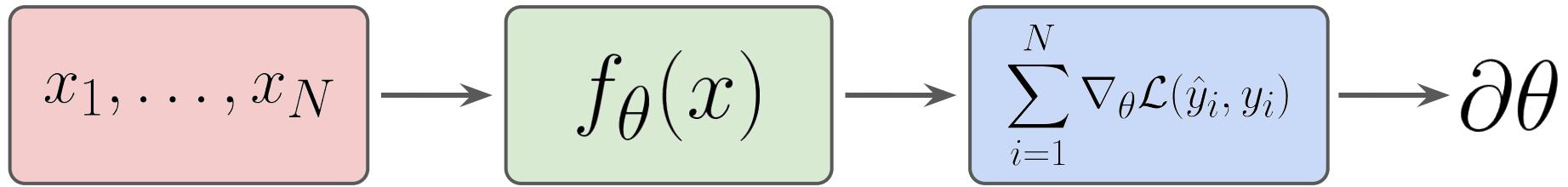
$$\theta \leftarrow \theta + \text{optimizer}(\partial\theta)$$

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

Compute

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

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



Device 1

$$x_1, \dots, x_{\frac{N}{2}}$$

$$f_\theta(x)$$

$$\sum_{i=1}^{N/2} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$$

$$\bigoplus \longrightarrow \partial \theta$$

Device 2

$$x_{\frac{N}{2}+1}, \dots, x_N$$

$$f_\theta(x)$$

$$\sum_{i=\frac{N}{2}+1}^N \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$$

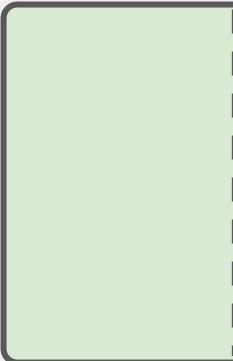


W

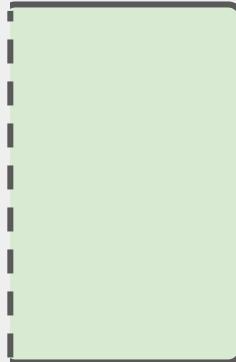


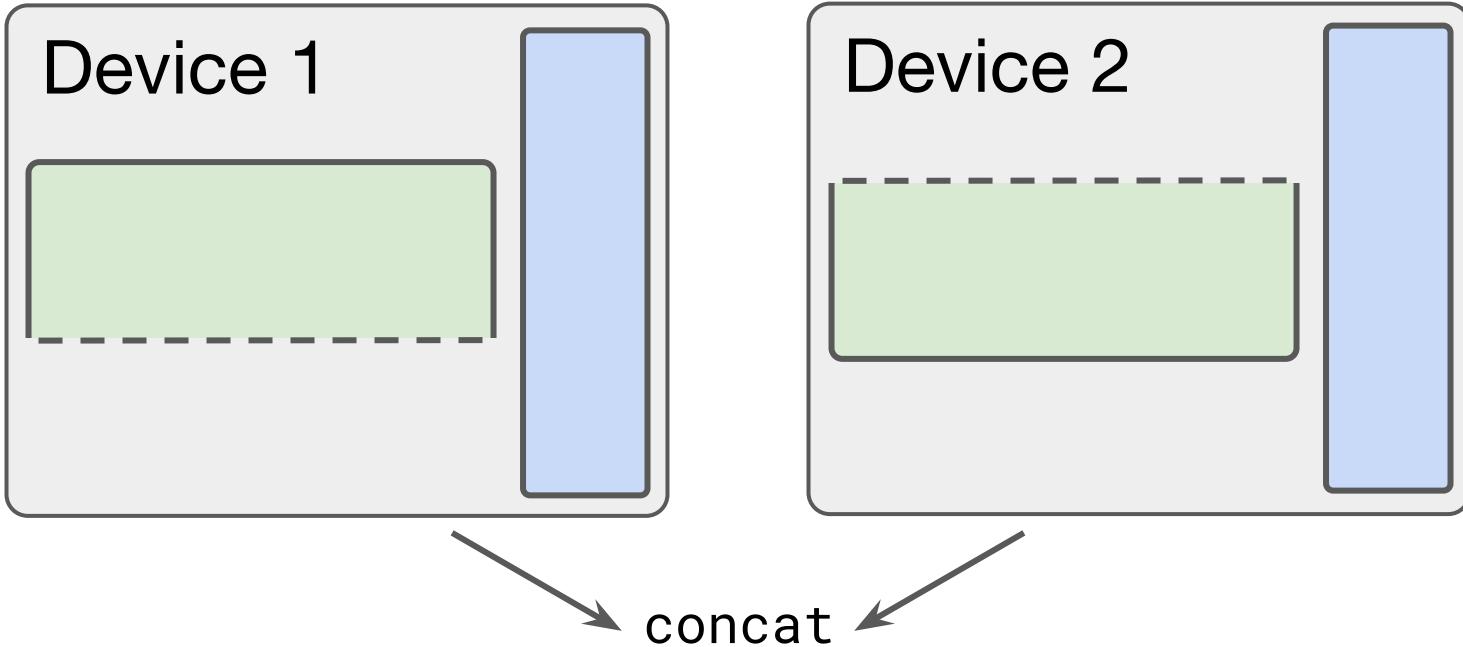
x

Device 1

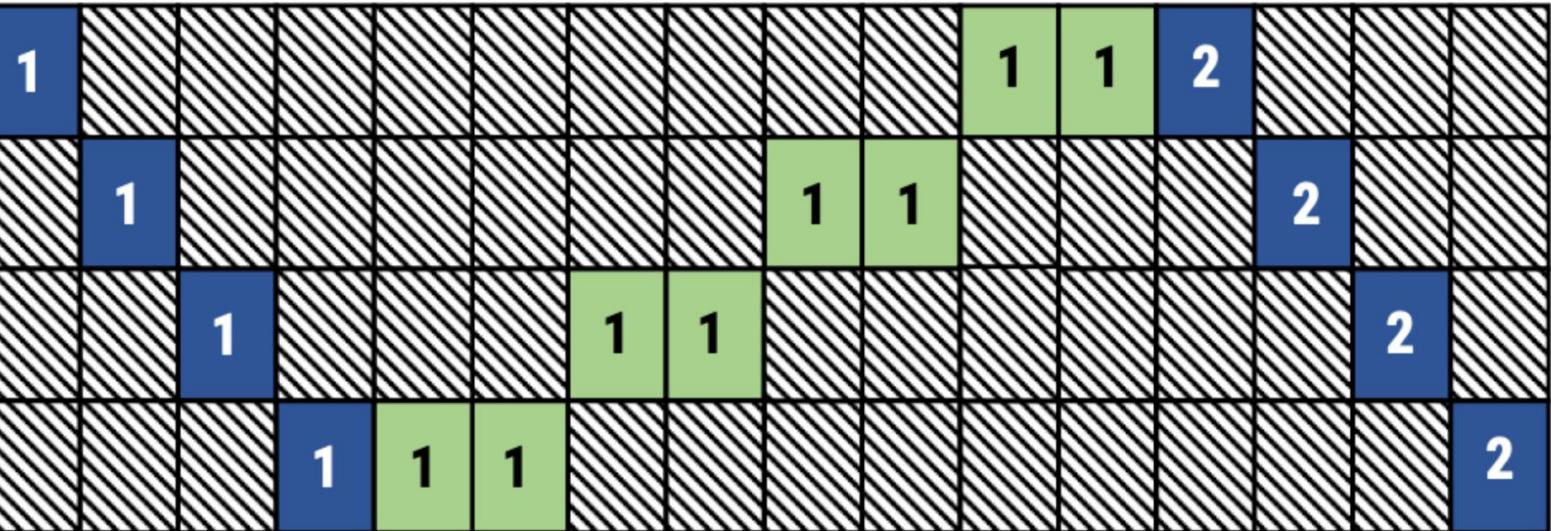


Device 2





Worker 1



Time →



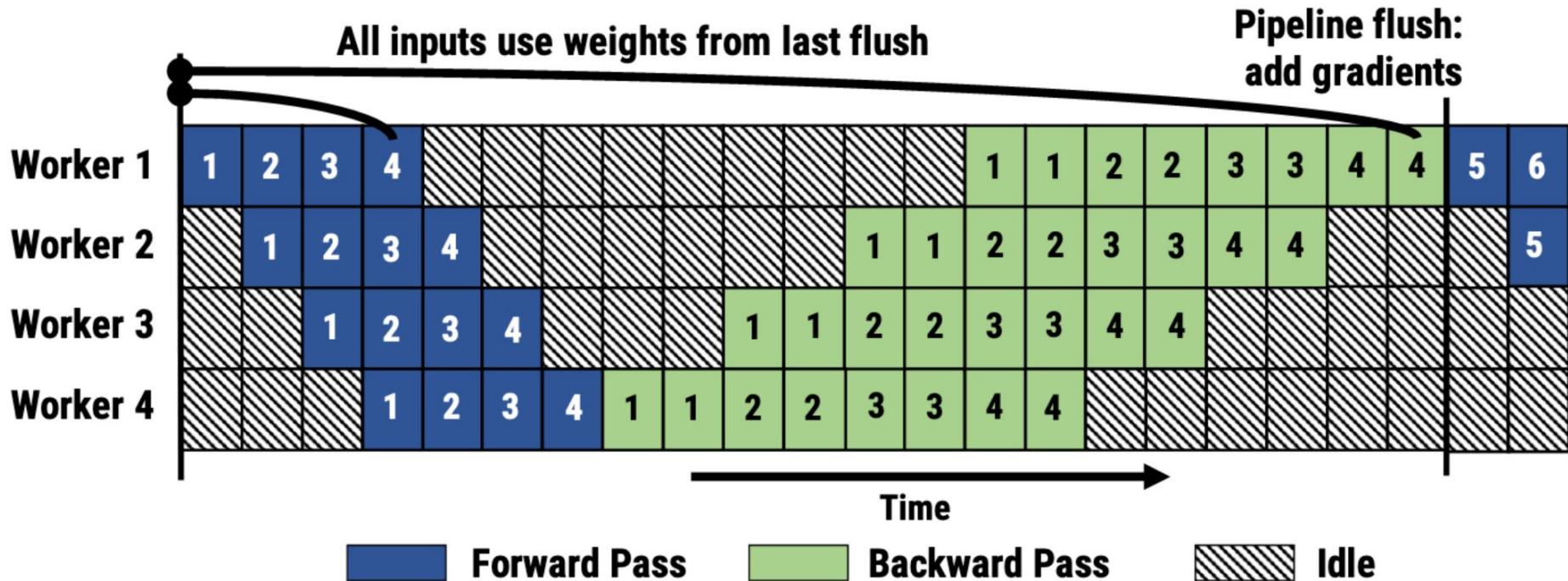
**Forward
Pass**



**Backward
Pass**



Idle

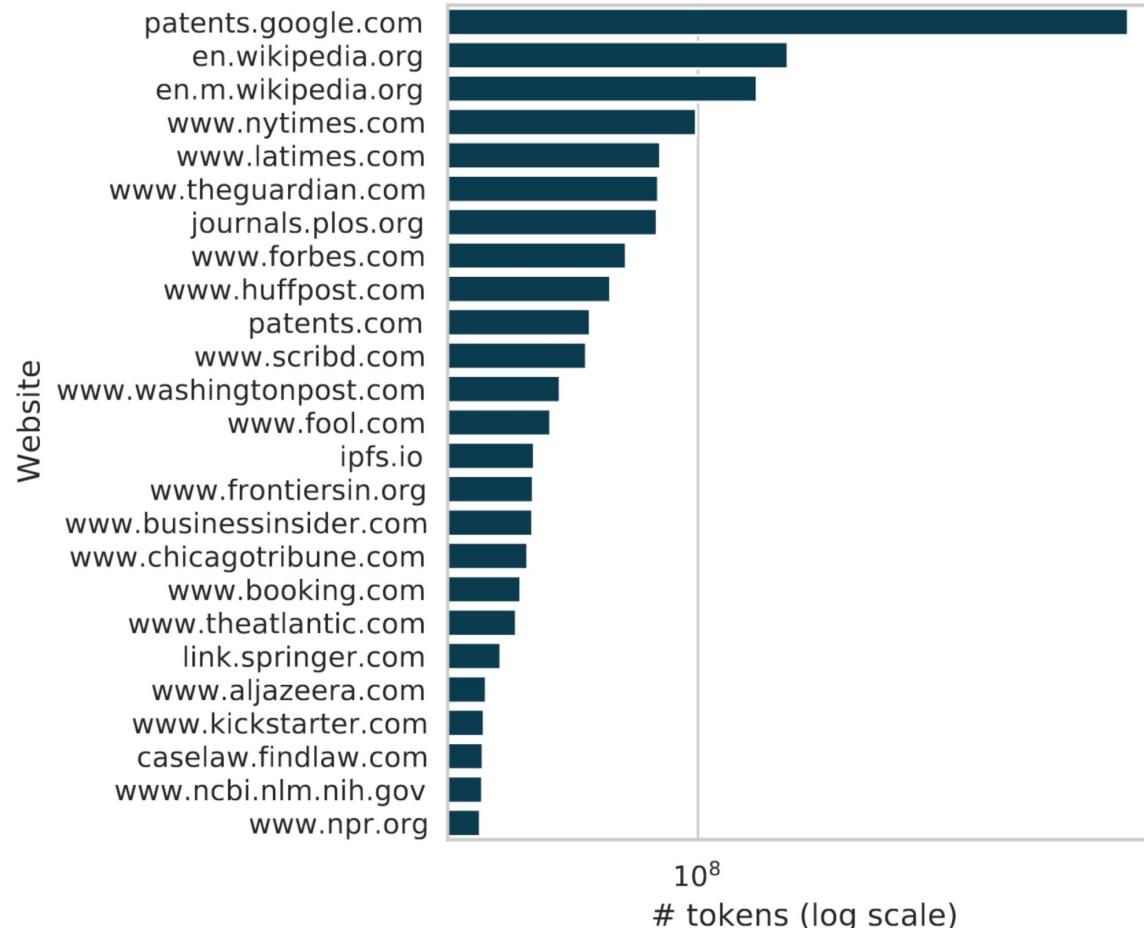


From “PipeDream: Generalized Pipeline Parallelism for DNN Training” by Narayanan et al.



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

The size of data available on the web has enabled deep learning models to achieve high accuracy on specific benchmarks in NLP and computer vision applications. However, in both application areas, the training data has been shown to have problematic characteristics resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status.



From “Documenting the English Colossal Clean Crawled Corpus” by Dodge et al.

... accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud compute time, and environmentally, due to the carbon footprint required to fuel modern tensor processing hardware.

Model	<i>Evolved Transformer NAS</i>	<i>T5</i>	<i>Meena</i>	<i>Gshard -600B</i>	<i>Switch Transformer</i>	<i>GPT-3</i>
Number of Parameters (B)	0.064 per model	11	2.6	619	1500	175
Percent of model activated on every token	100%	100%	100%	0.25%	0.10%	100%
Developer			Google			OpenAI
Datacenter of original experiment	Google Georgia	Google Taiwan	Google Georgia	Google North Carolina	Google Georgia	Microsoft
When model ran	Dec 2018	Sep 2019	Dec 2019	Apr 2020	Oct 2020	2020
Datacenter Gross CO ₂ e/KWh (kg/KWh when it was run)	0.431	0.545	0.415	0.201	0.403	0.429
Datacenter Net CO ₂ e/KWh (kg/KWh when it was run)	0.431	0.545	0.415	0.177	0.330	0.429
Datacenter PUE (when it was run)	1.10	1.12	1.09	1.09	1.10	1.10
Processor	TPU v2			TPU v3		V100
Chip Thermal Design Power (TDP in Watts)	280		450			300
Measured System Average Power per Accelerator, including memory, network interface, fans, host CPU (W)	208	310	289	288	245	330
Measured Performance (TFLOPS/s) ¹²	24.8	45.6	42.3	48.0	34.4	24.6
Number of Chips	200	512	1024	1024	1024	10,000
Training time (days)	6.8	20	30	3.1	27	14.8
Total Computation (floating point operations)	2.91E+21	4.05E+22	1.12E+23	1.33E+22	8.22E+22	3.14E+23
Energy Consumption (MWh)	7.5	85.7	232	24.1	179	1,287
% of Google 2019 total energy consumption (12.2 TWh = 12,200,000 MWh) [Goo20]	0.00006%	0.00070%	0.00190%	0.00020%	0.00147%	0.01055%
Gross tCO ₂ e for Model Training	3.2	46.7	96.4	4.8	72.2	552.1
Net tCO ₂ e for Model Training	3.2	46.7	96.4	4.3	59.1	552.1

From “Carbon Emissions and Large Neural Network Training” by Patterson et al.

GPT-3 175B model required 3.14E23 FLOPS of computing for training. Even at theoretical 28 TFLOPS for V100 and lowest 3 year reserved cloud pricing we could find, this will take 355 GPU-years and cost \$4.6M for a single training run.

President Franklin <M> born <M> January 1882.

Lily couldn't <M>. The waitress had brought the largest <M> of chocolate cake <M> seen.

Our <M> hand-picked and sun-dried <M> orchard in Georgia.

T5

D. Roosevelt was <M> in

believe her eyes <M> piece <M> she had ever

peaches are <M> at our

Pre-training

Fine-tuning

President Franklin D.
Roosevelt was born
in January 1882.

When was Franklin D.
Roosevelt born?

T5

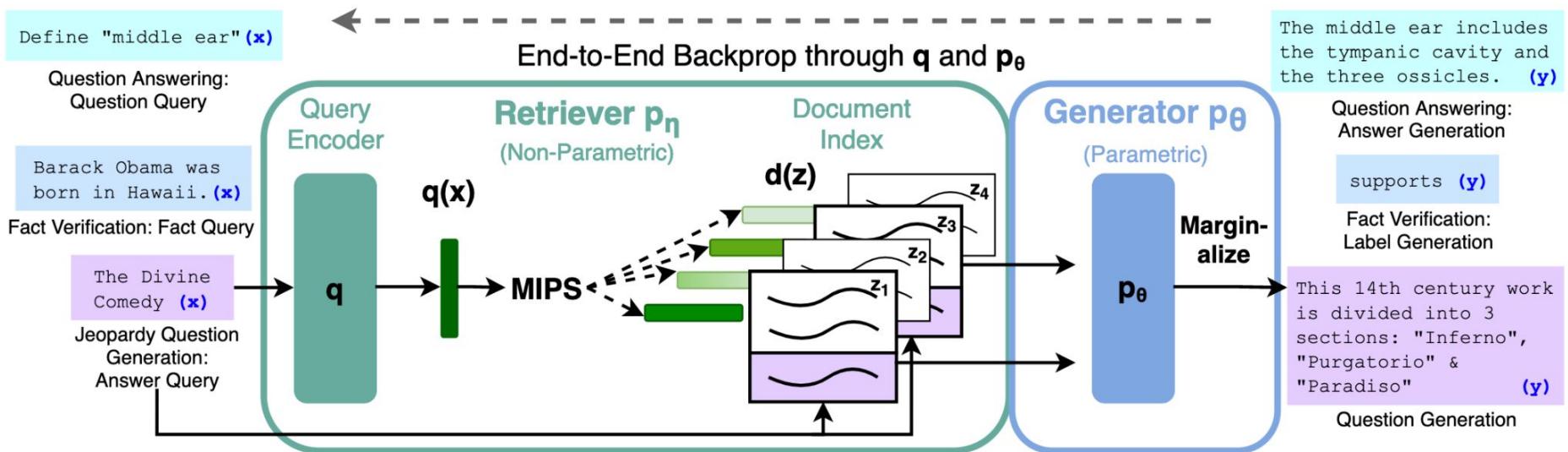
1882

	NQ	WQ	TQA
Open-domain SoTA	41.5	42.4	57.9
T5.1.1-Base	25.7	28.2	24.2
T5.1.1-Large	27.3	29.5	28.5
T5.1.1-XL	29.5	32.4	36.0
T5.1.1-XXL	32.8	35.6	42.9
T5.1.1-XXL + SSM	35.2	42.8	51.9

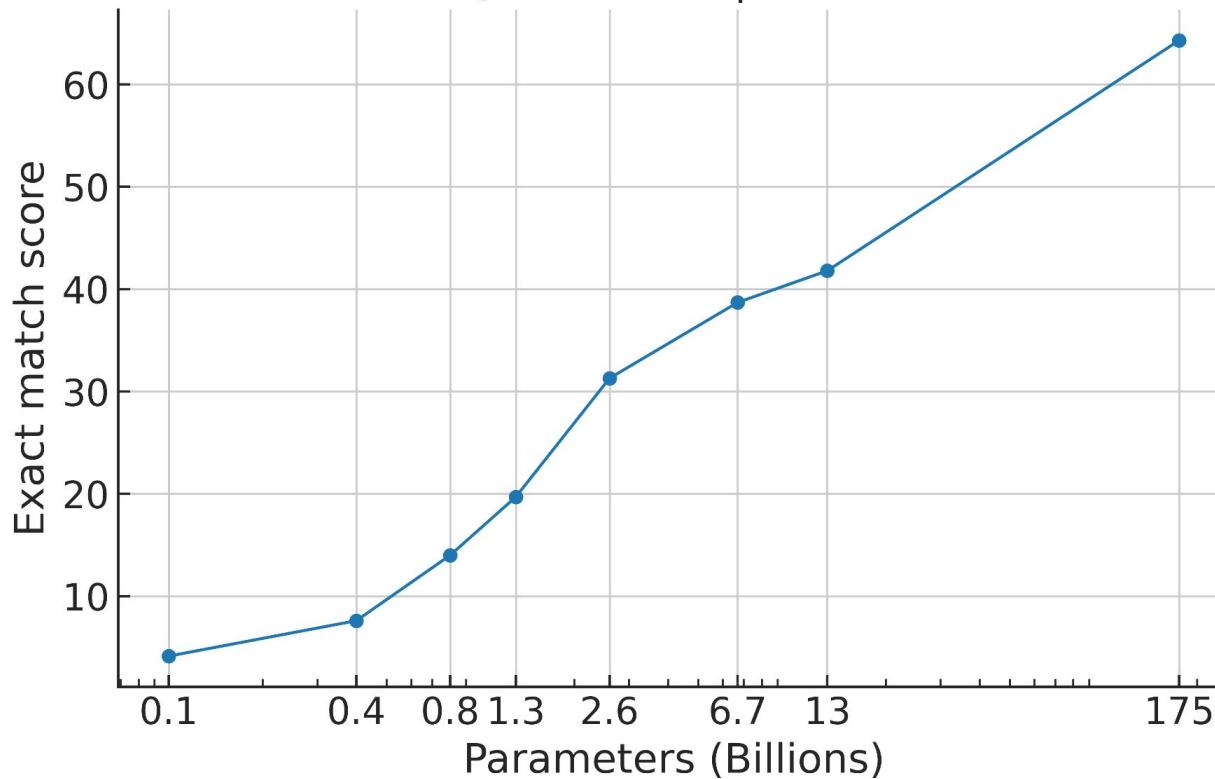
Index of /wikidatawiki/entities/

<u>latest-all.json.bz2</u>	27-Jul-2021 11:32	68418560489
<u>latest-all.json.gz</u>	27-Jul-2021 04:58	102963487951

~70GB compressed = 13B float32 parameters



TriviaQA zero-shot performance



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

Summarization

The picture appeared on the wall of a Poundland store on Whymark Avenue [...] How would you rephrase that in a few words?

Paraphrase identification

"How is air traffic controlled?" "How do you become an air traffic controller?"
Pick one: these questions are duplicates or not duplicates.

Question answering

I know that the answer to "What team did the Panthers defeat?" is in "The Panthers finished the regular season [...]" . Can you tell me what it is?

Multi-task training

Zero-shot generalization

Natural language inference

Suppose "The banker contacted the professors and the athlete". Can we infer that "The banker contacted the professors"?

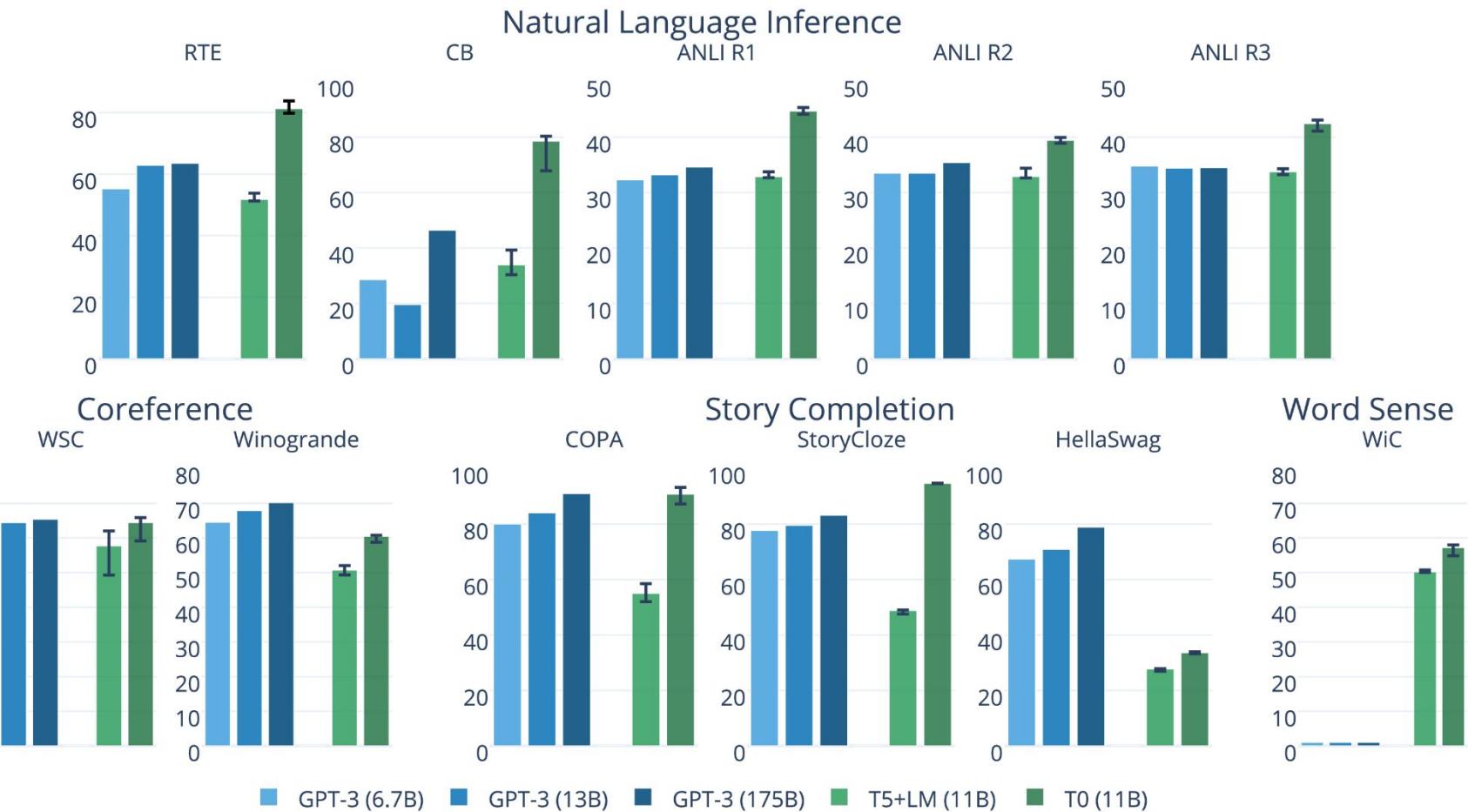
T0

Graffiti artist Banksy is believed to be behind [...]

Not duplicates

Arizona Cardinals

Yes



From “Multitask Prompted Training Enables Zero-Shot Generalization” by Sanh et al.

Thanks.

craffel@gmail.com