LECTURE 1: MOTIVATION AND BASIC CONCEPTS OF LARGE MODEL TRAINING

INSTRUCTOR NAME AND TITLE

COURSE AGENDA

8:50 – 9:00	Welcome
9:00 – 10:00	Lecture 1: Motivation and basic concepts
10:00 – 11:00	Lab 1 / Part 1: From SLURM basics to multi-node training
11:00 – 11:15	Break
11:15 – 12:15	Lecture 2: Advanced concepts of large model training
12:15 – 13:15	Lab 1 / Part 2:
13:15 – 14:00	Lunch break
14:00 – 15:00	Lecture 3: Large model deployment
15:00 – 16:00	Lab 2: Hands on example of how to deploy GPT-3 into Triton Inference Server
16:00 – 17:00	Assessment support

PART 1

Motivation and basic concepts

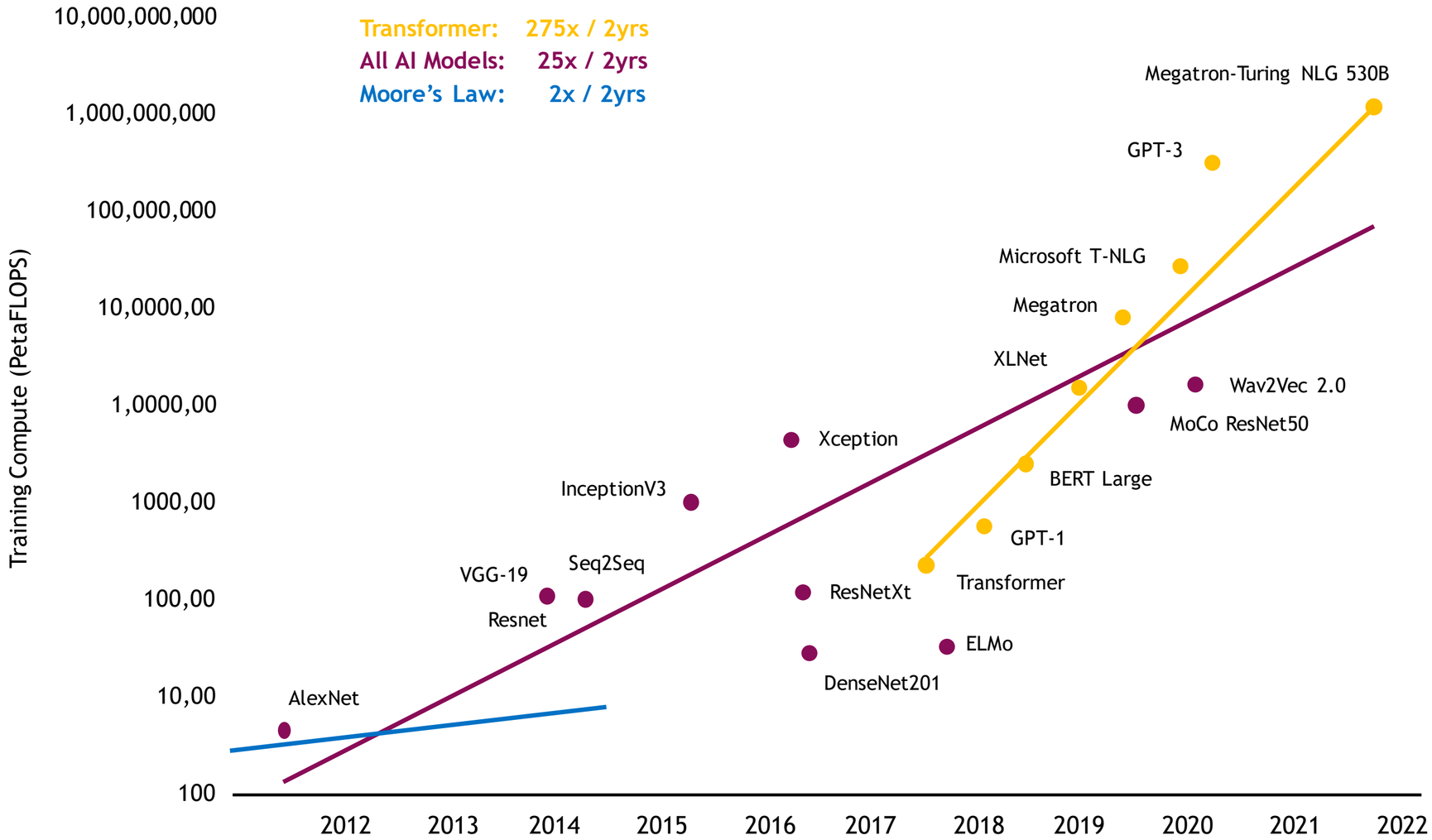
- Lecture
  - **Why large models?**
  - Impact on AI landscape
  - Challenges of large model training
  - Basic techniques for memory reduction
  - Overview of the tools used in the lab

- Lab 1 / Part 1
  - Introduction to the SLURM class environment
  - GPT model pretraining
  - Multi-node scaling
  - Optimize the GPT model pretraining

THE GROWTH OF NEURAL NETWORKS

DRAMATIC INCREASE IN MODEL SIZES

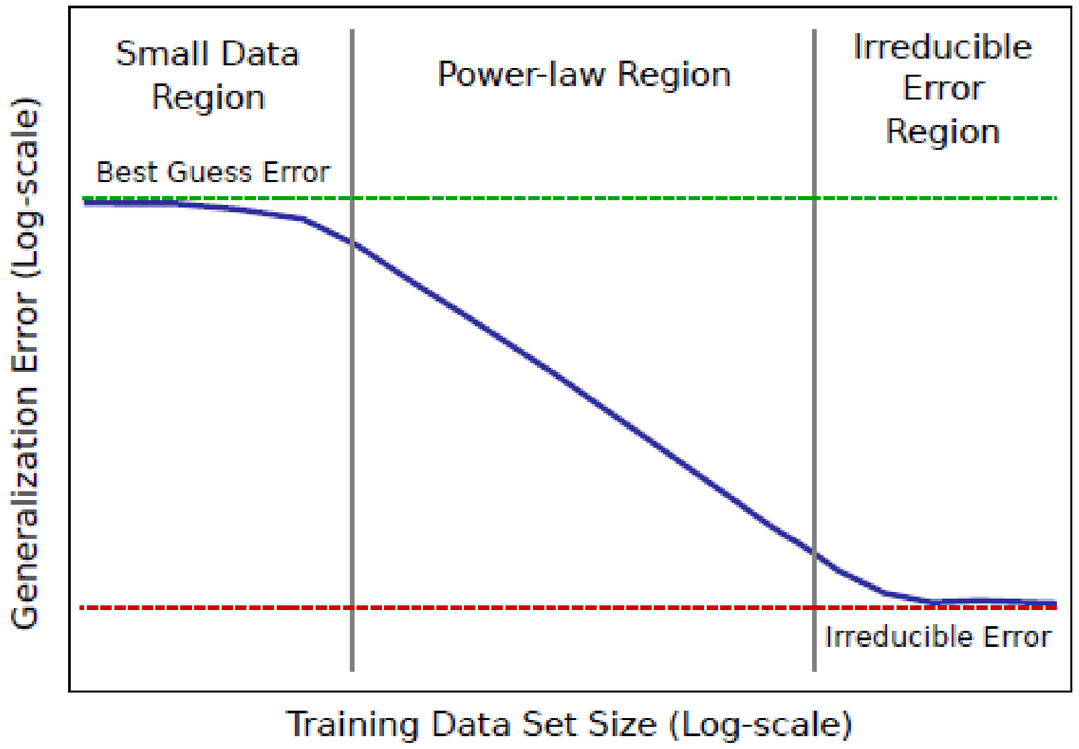
The Trend Continues



WHY?

THE SCALING LAWS

Performance of neural networks increases with model/dataset size

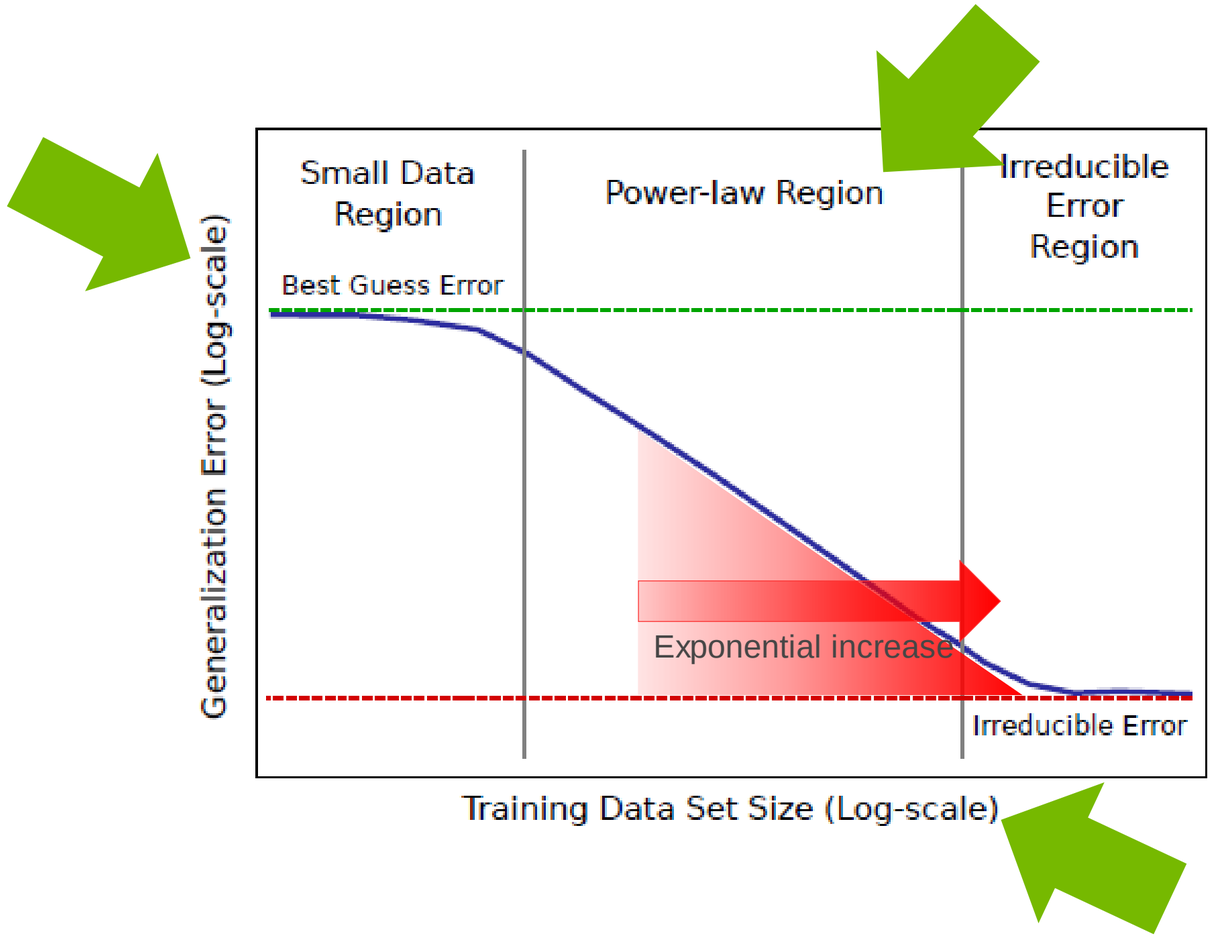


Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017). **Deep Learning Scaling is Predictable, Empirically.** arXiv:1712.00409.

HISTORICALLY LIMITED BY VOLUME OF LABELLED DATA

THE COST OF LABELS

Limits the utility of supervised deep learning models



Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017). **Deep Learning Scaling is Predictable, Empirically**. arXiv:1712.00409.

THE COST OF LABELS

Hard limits to supervised deep learning utility

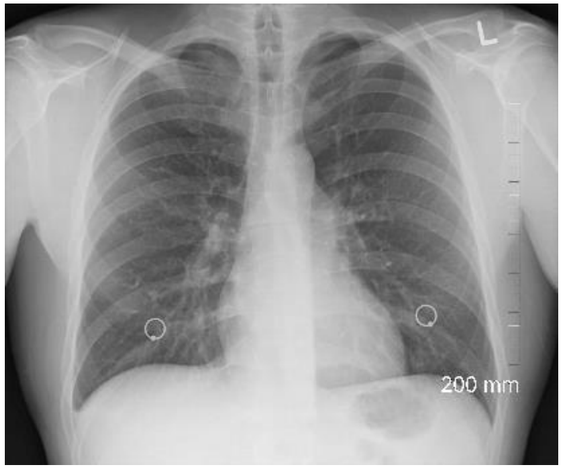
Table 1. UK clinical radiology workforce (headcount), 2018

	England	Northern Ireland	Scotland	Wales	UK total
Consultant-grade	3,296	135	327	169	3,927
Trainees	1,286	51	149	69	1,555
SAS-grade*	65	0	2	2	69
Total	4,647	186	478	240	5,551

[SAS grade comprises associate specialists, specialty doctors and trust-grade staff.]

Self-Driving Auto Example
1500 workers labelling

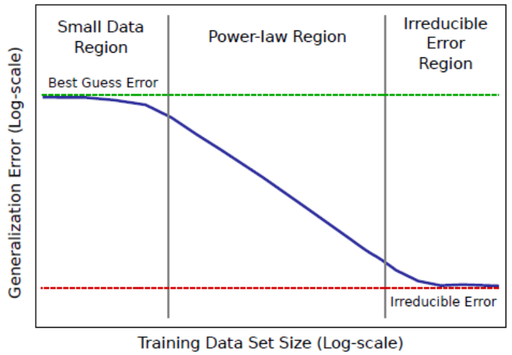
Clinical Radiology Example
Total of 3296 radiologists in England

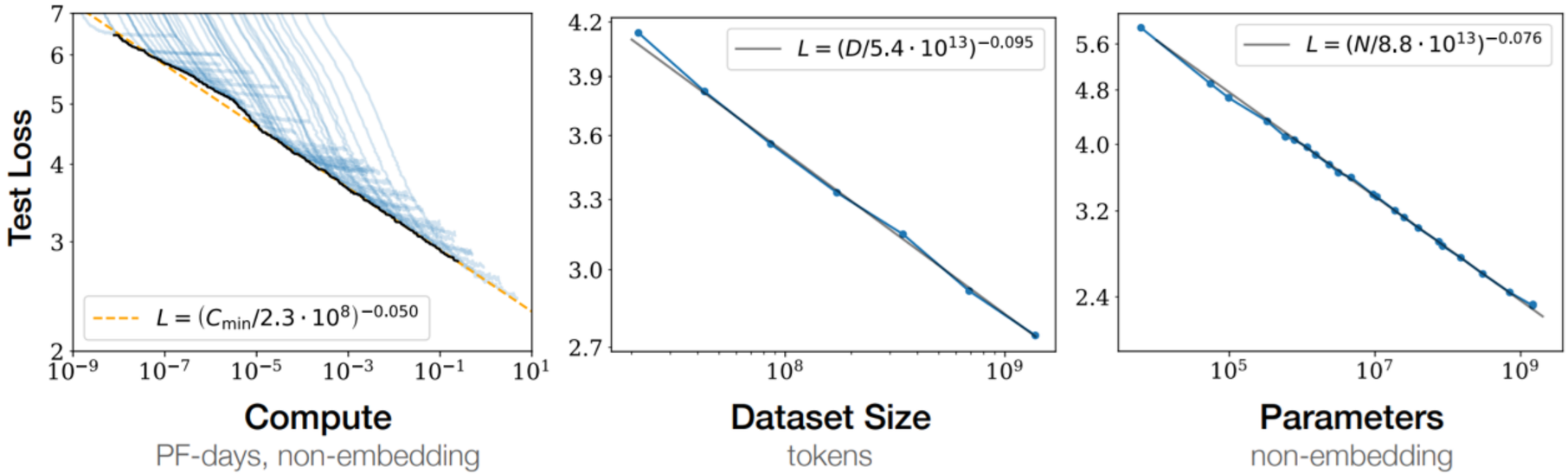


SUCCESS OF UNSUPERVISED LEARNING

EMPIRICAL EVIDENCE

The Scaling Laws in NLP

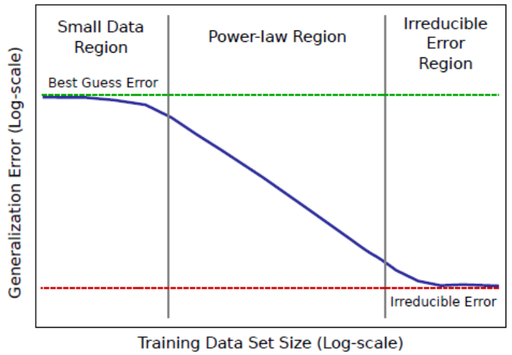


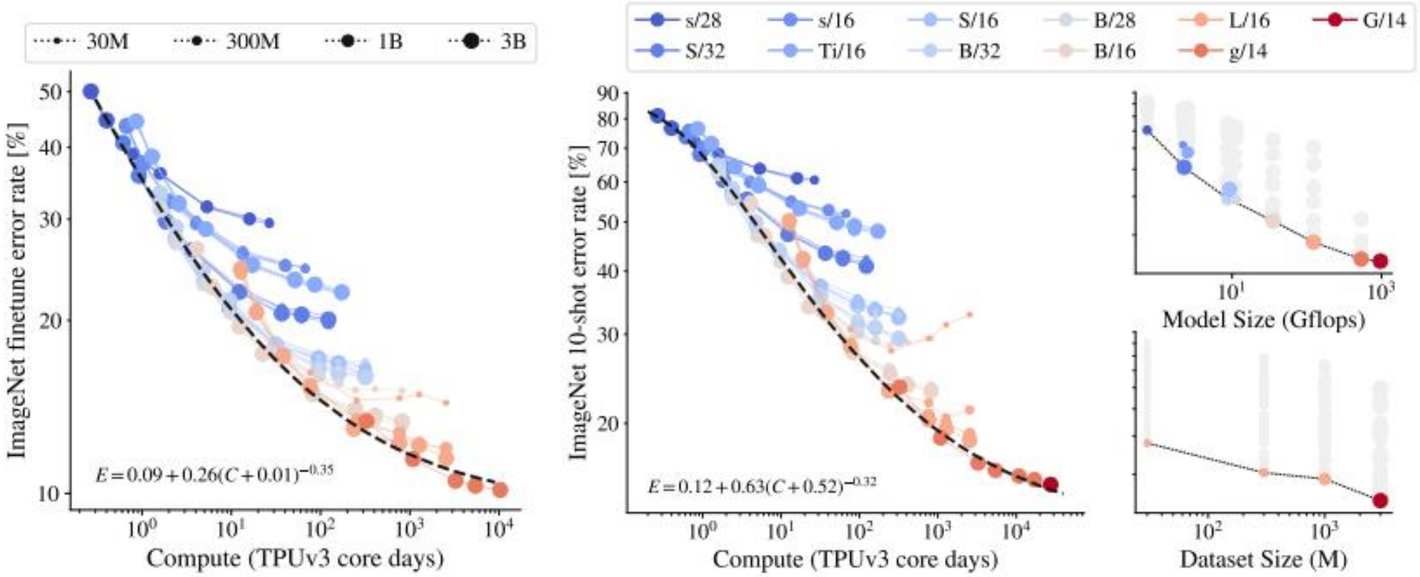


Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, Sam McCandlish. **Scaling Laws for Autoregressive Generative Modeling**.2020

EMPIRICAL EVIDENCE

The Scaling Laws in Computer Vision

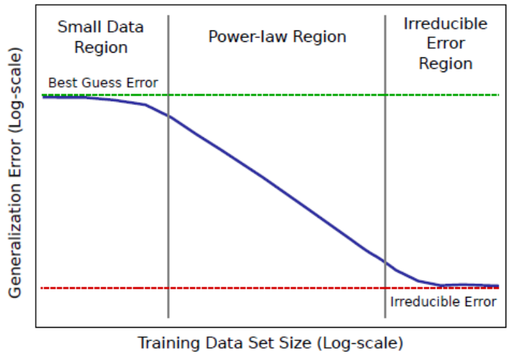


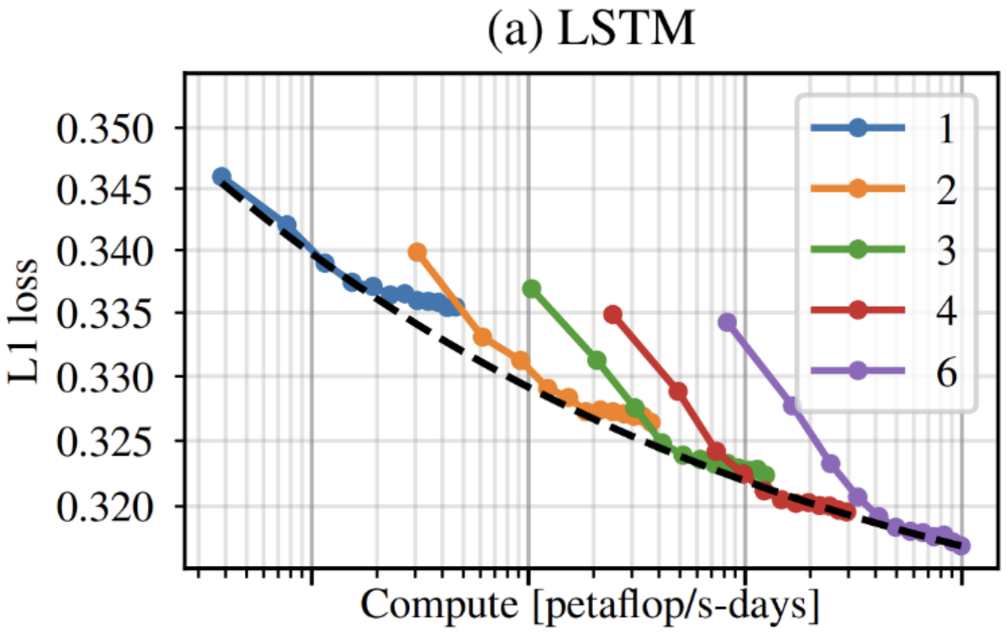


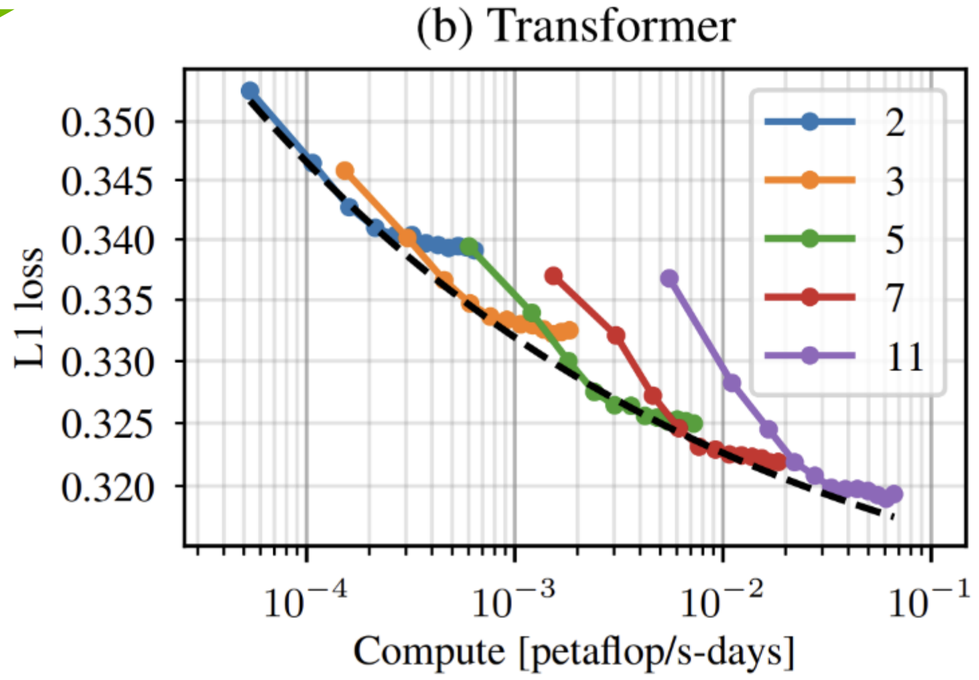
Zhai, Xiaohua, et al. "**Scaling vision transformers**." *arXiv preprint arXiv:2106.04560* (2021).

EMPIRICAL EVIDENCE

The Scaling Laws in Speech



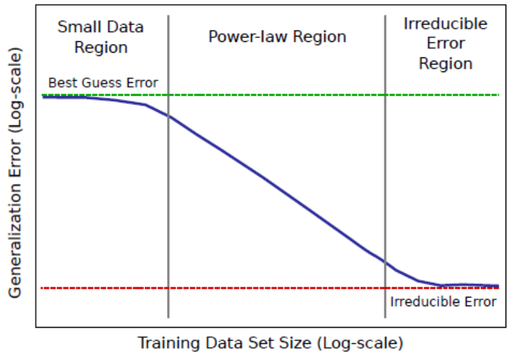


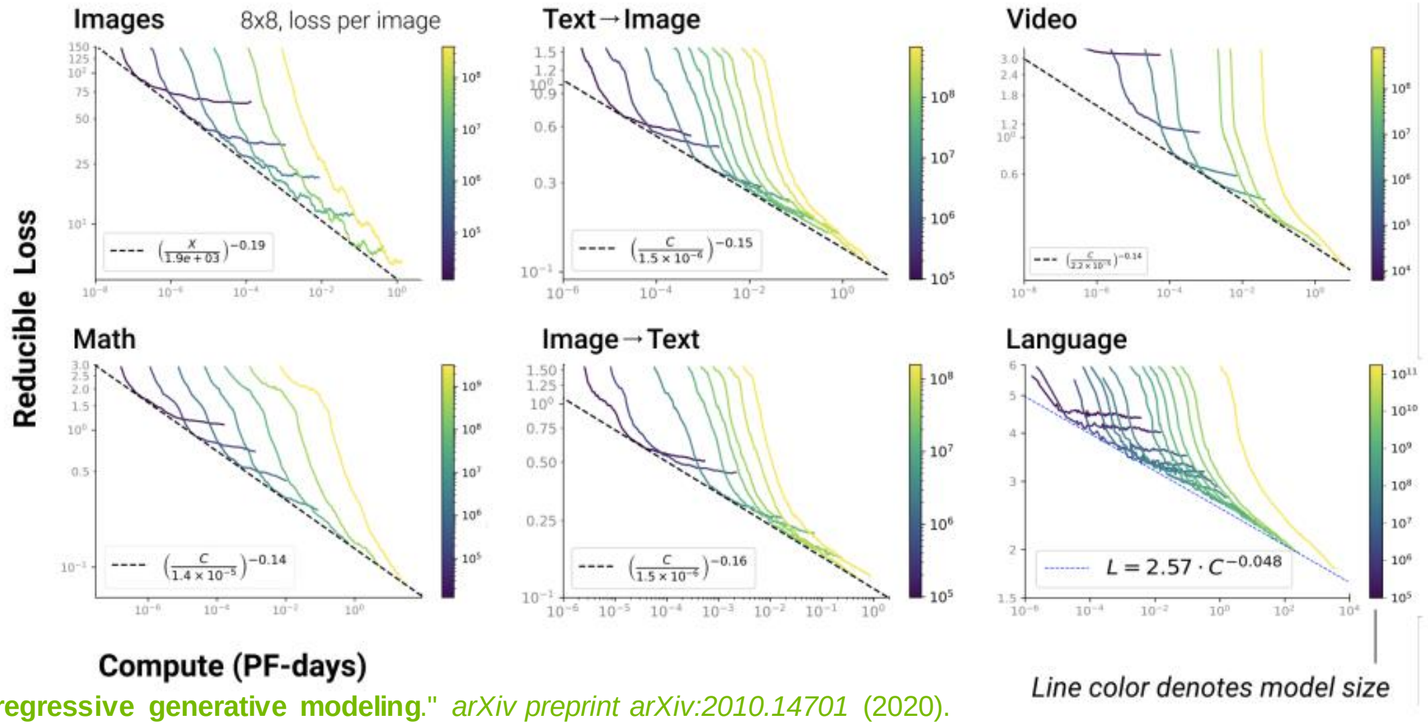


Droppo, Jasha, and Oguz Elibol. "**Scaling Laws for Acoustic Models**." *arXiv preprint arXiv:2106.09488* (2021).

EMPIRICAL EVIDENCE

The Scaling Laws for generative models



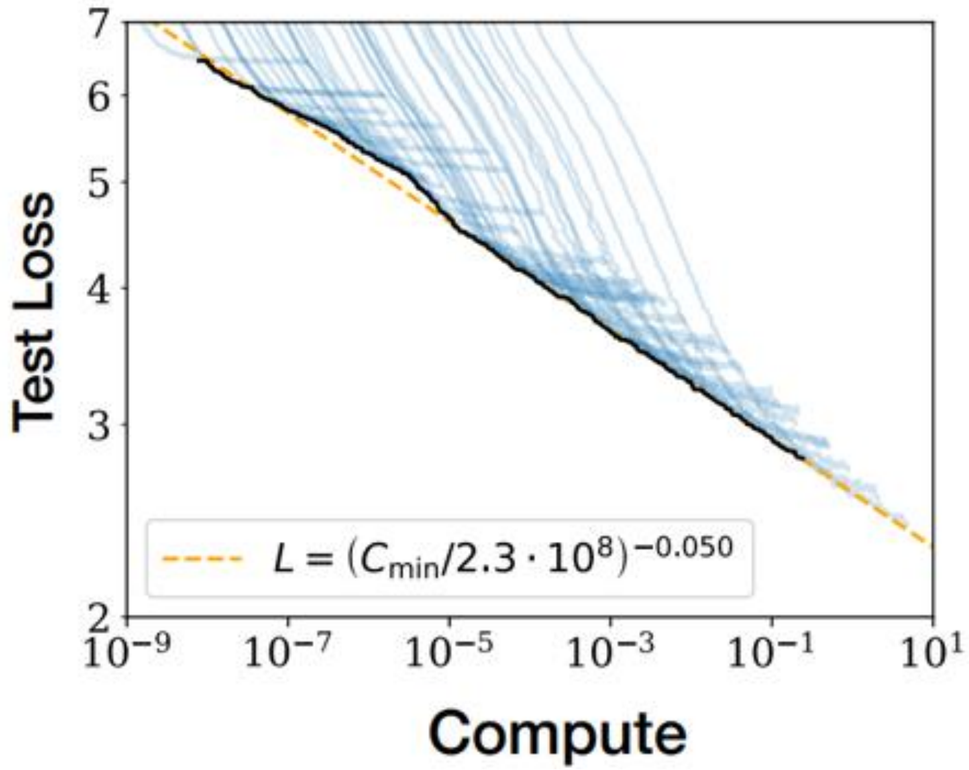


Henighan, Tom, et al. "**Scaling laws for autoregressive generative modeling**." *arXiv preprint arXiv:2010.14701* (2020).

THE COST

ARE LARGE LANGUAGE MODELS WORTH IT?

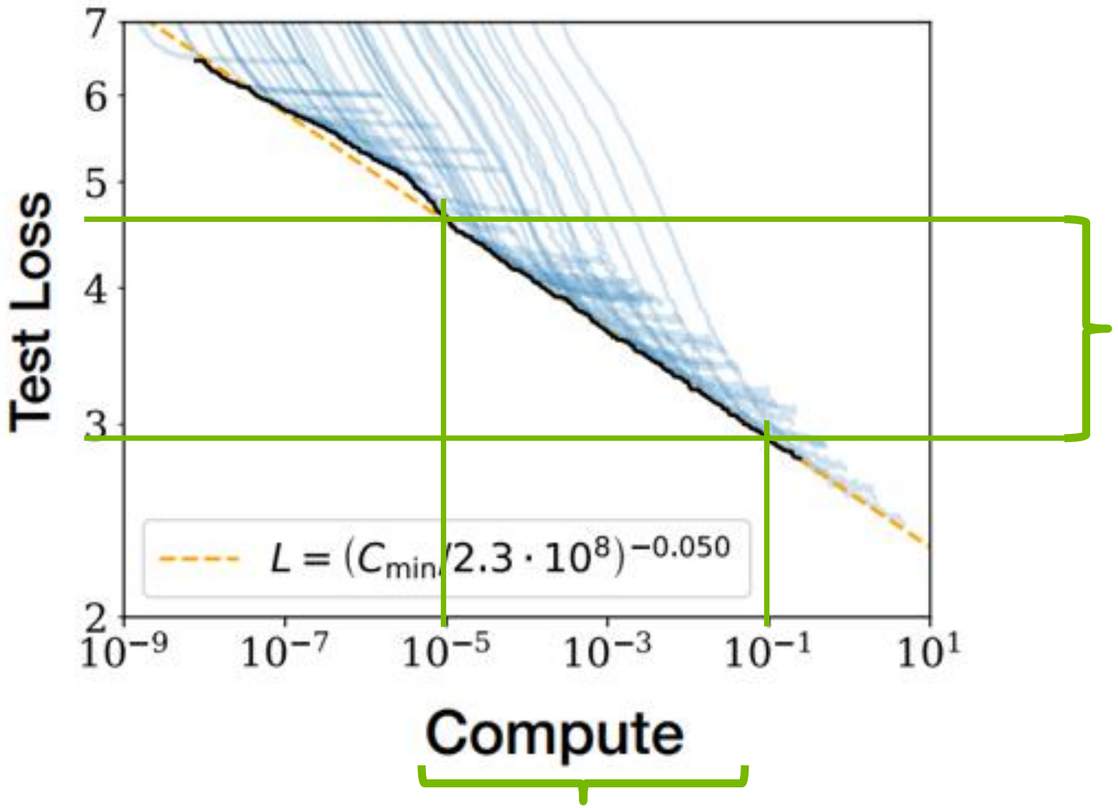
The cost of incremental improvement



Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, Sam McCandlish. **Scaling Laws for Autoregressive Generative Modeling.**2020

ARE LARGE LANGUAGE MODELS WORTH IT?

The cost of incremental improvement



Are we building those models only for the small incremental improvement in their performance?

Is it worth all the engineering and computational investment?

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, Sam McCandlish. **Scaling Laws for Autoregressive Generative Modeling**.2020

BEYOND ACCURACY

FEW SHOT LEARNING!

FEW SHOT LEARNING

Learning from far fewer examples

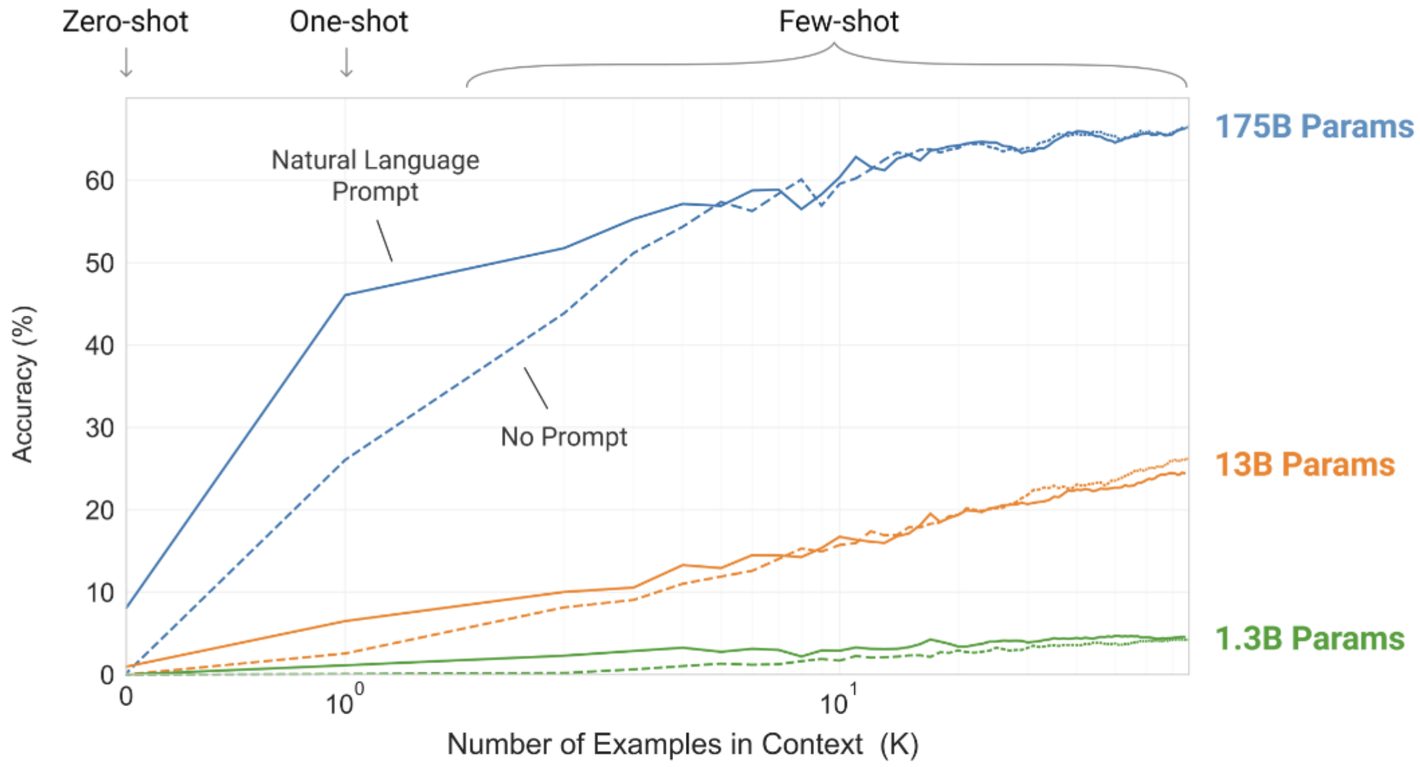


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*..

EXAMPLE OF FEW SHOT LEARNING
(PROMPT DESIGN)

PROMPT ENGINEERING

Video example

Prompt Megatron

```
English: I do not speak French.
French: Je ne parle pas français.

English: See you later!
French: À tout à l'heure!

English: Where is a good restaurant?
French: Où est un bon restaurant?

English: What rooms do you have available?
French:
```

Number tokens to generate (1-1024): 32 Submit 9 / 1000

```
Test textured')# Get a hex color value from a RGBA pixel.aRGBA=anRGBA.rgba255()color=aRGBA.
```

BEYOND PROMPT DESIGN

BREAKTHROUGHS OF 2021

On par with approaches requiring large datasets

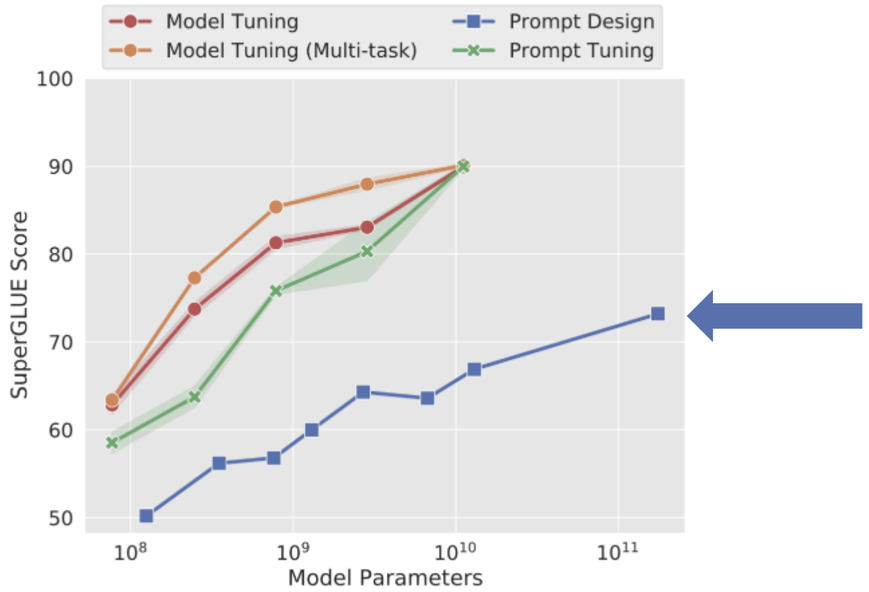
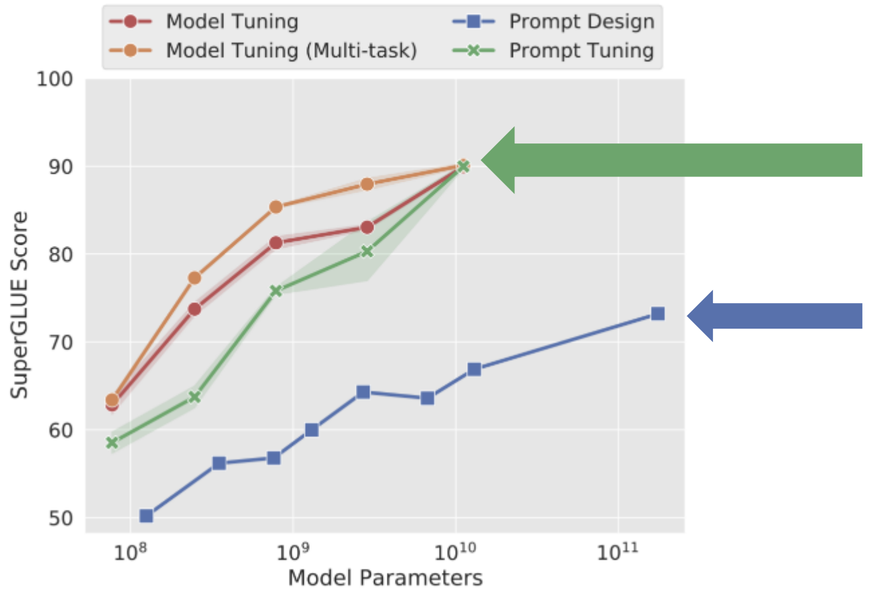


Figure 1: Standard **model tuning** of T5 achieves strong performance, but requires storing separate copies of the model for each end task. Our **prompt tuning** of T5 matches the quality of model tuning as size increases, while enabling the reuse of a single frozen model for all tasks. Our approach significantly outperforms few-shot **prompt design** using GPT-3. We show mean and standard deviation across 3 runs for tuning methods.

Brian Lester, Rami Al-Rfou, Noah Constant. **The Power of Scale for Parameter-Efficient Prompt Tuning.** 2021. https://arxiv.org/abs/2104.08691

BREAKTHROUGHS OF 2021

On par with approaches requiring large datasets



On par with finetuning approaches even though we do not have large finetuning datasets

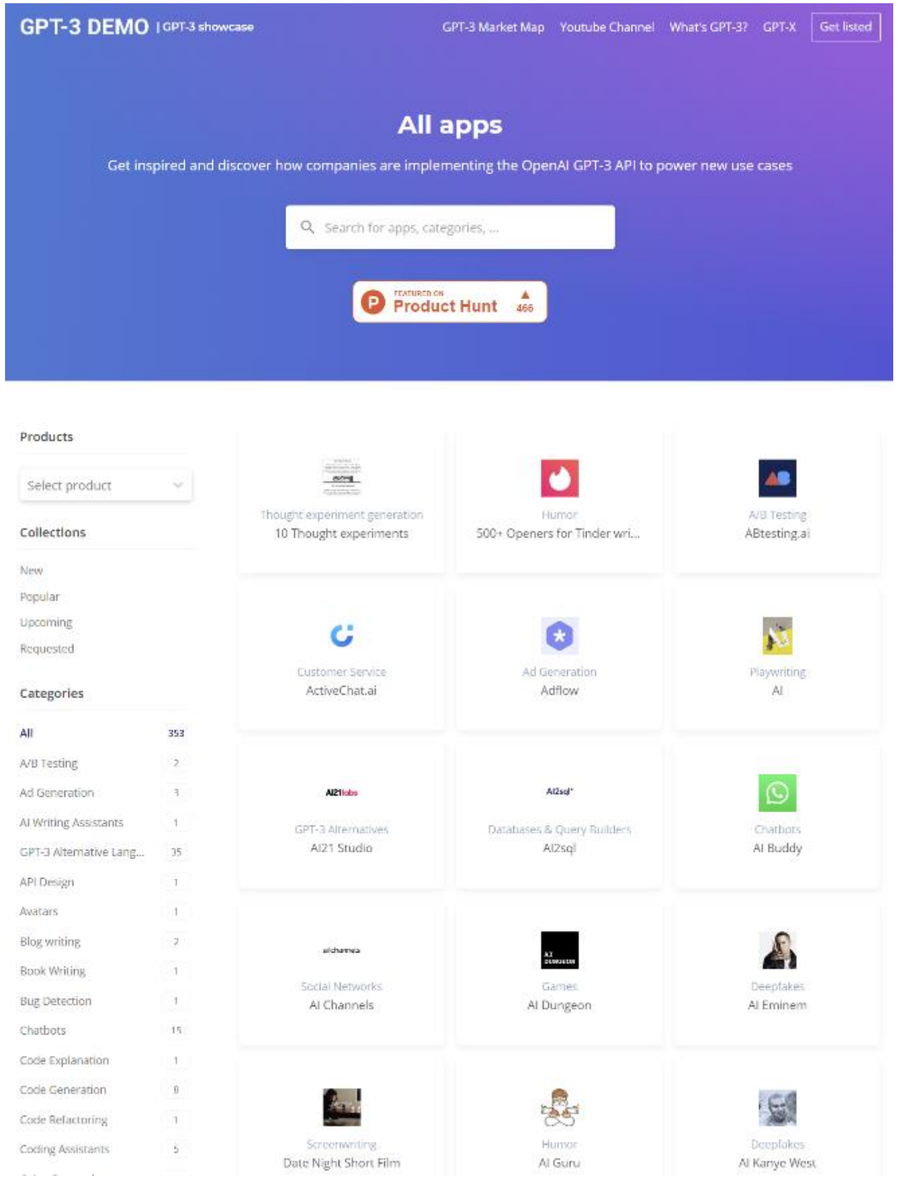
Figure 1: Standard **model tuning** of T5 achieves strong performance, but requires storing separate copies of the model for each end task. Our **prompt tuning** of T5 matches the quality of model tuning as size increases, while enabling the reuse of a single frozen model for all tasks. Our approach significantly outperforms few-shot **prompt design** using GPT-3. We show mean and standard deviation across 3 runs for tuning methods.

Brian Lester, Rami Al-Rfou, Noah Constant. **The Power of Scale for Parameter-Efficient Prompt Tuning**. 2021. https://arxiv.org/abs/2104.08691

IMPLICATIONS IN SOFTWARE DEVELOPMENT

USE CASES FOR LARGE MODELS

300 Applications powered by GPT-3 from OpenAI



https://openai.com/blog/gpt-3-apps/

CAPABILITY OF MODERN NLP MODELS

Video example

CAPABILITY OF MODERN NLP MODELS

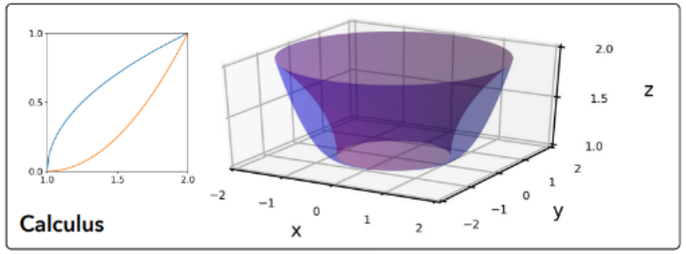
Video example

Oleksii Kuchaiev. **NVIDIA NeMo: Speech Recognition, Speech Synthesis, and NLP Updates**. GTC Fall 2021

CAPABILITY OF MODERN NLP MODELS

Code generation – explaining jokes – solving mathematical problems

Output (Answers, Visualizations, Explanations)



Explaining a joke

Prompt

```
Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for
their TPU team? It showed them how to communicate between two
different pods!
```

Model Response

```
Prediction: TPUs are a type of computer chip that Google uses
for deep learning. A "pod" is a group of TPUs. A "pod" is also a
group of whales. The joke is that the whale is able to
communicate between two groups of whales, but the speaker is
pretending that the whale is able to communicate between two
groups of TPUs.
```

PaLM explains an original joke with two-shot prompts.

https://copilot.github.com/
https://openai.com/blog/openai-codex

CAPABILITY OF MODERN NLP MODELS

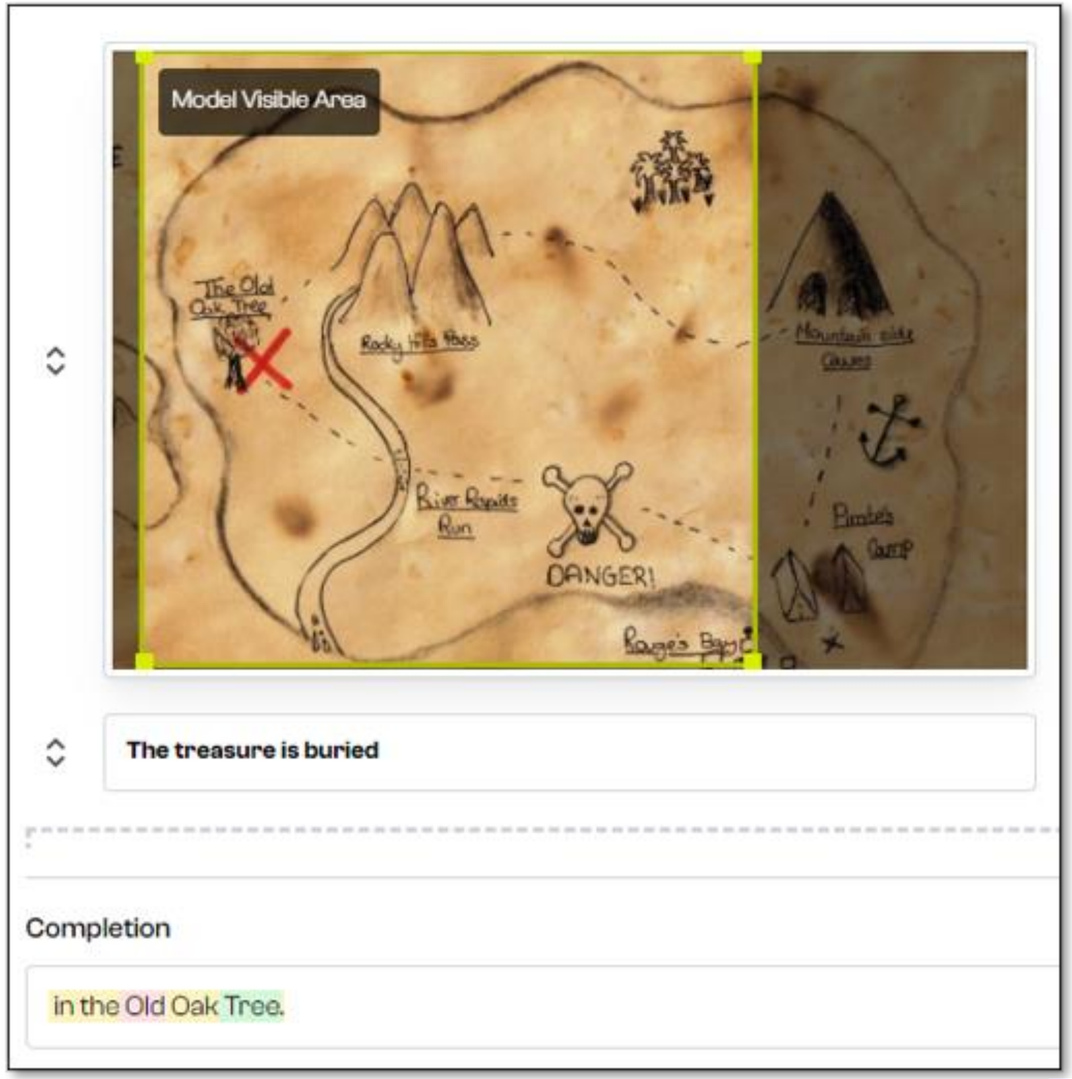
Video example

https://say-can.github.io/

IMPLICATIONS IN COMPUTER VISION

BEYOND NLP

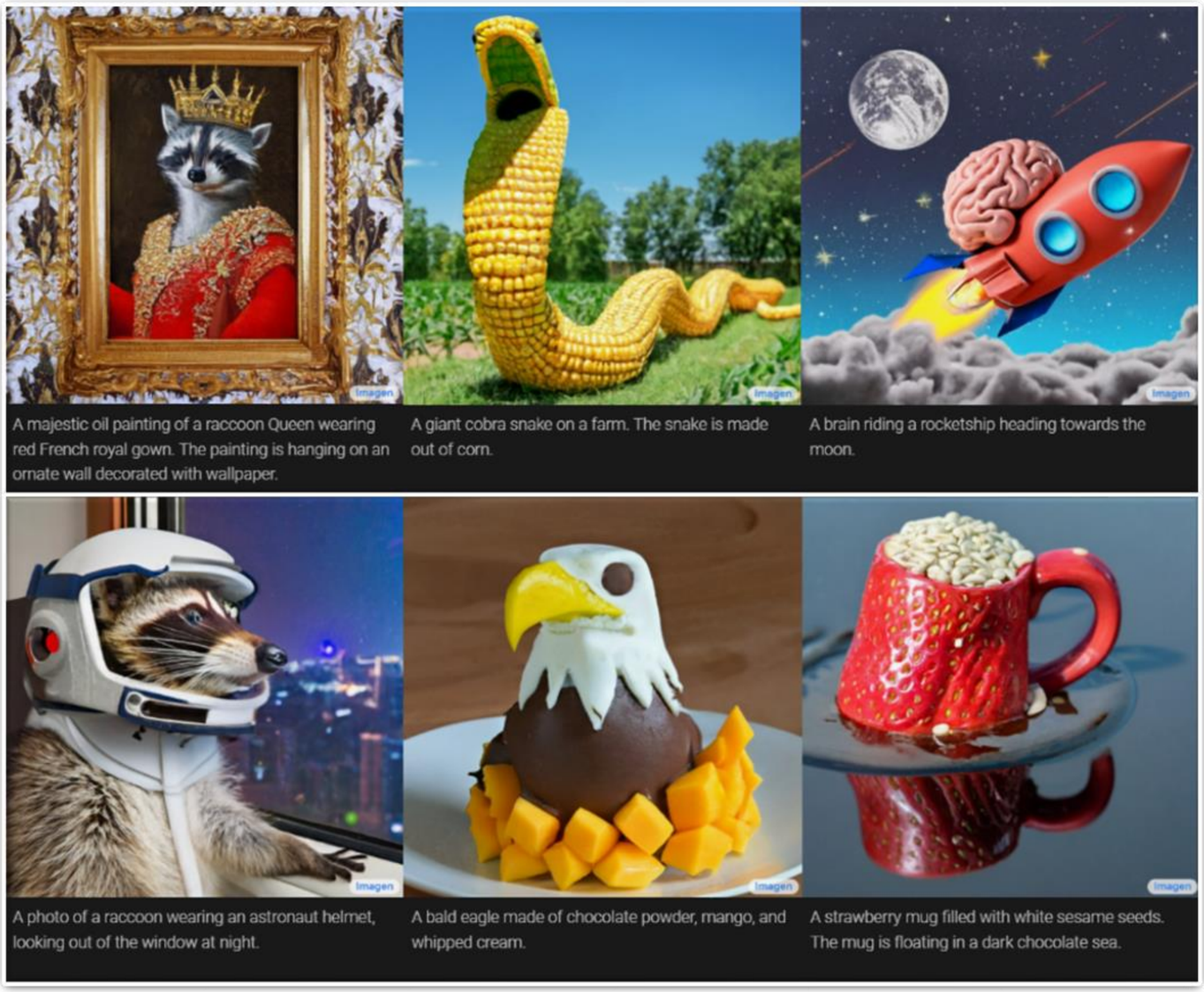
Transforming Computer Vision: MAGMA



https://github.com/Aleph-Alpha/magma

BEYOND NLP

Transforming Computer Vision

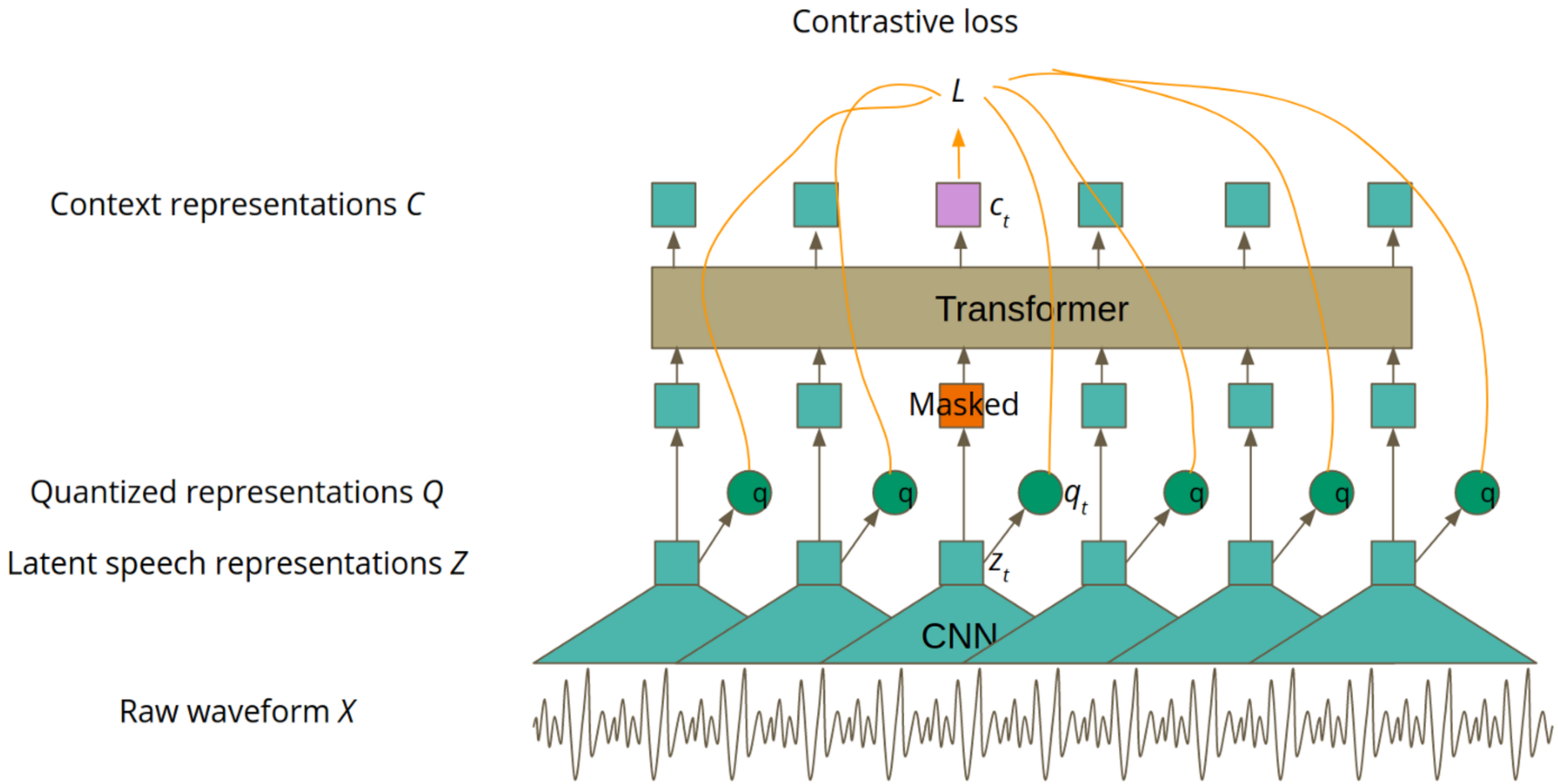


Saharia, Chitwan, et al. "**Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding**." *arXiv preprint arXiv:2205.11487* (2022).

AND BEYOND

AUTOMATIC SPEECH RECOGNITION

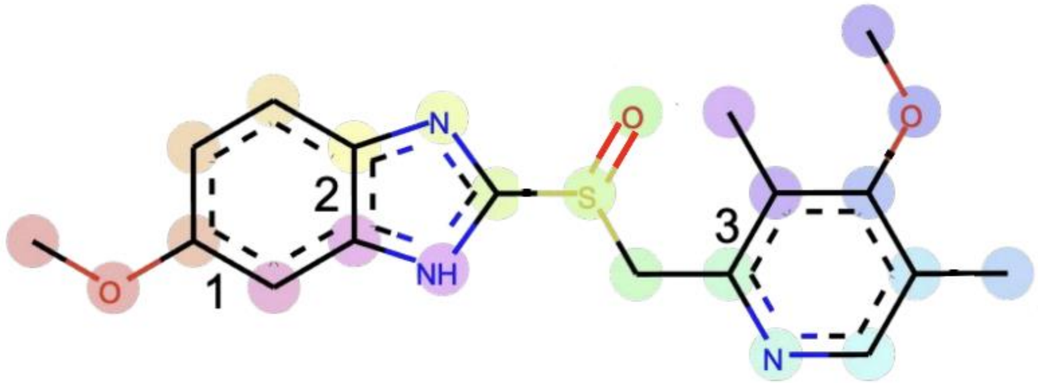
Success of Self-Supervised Learning



A. Baevski, H. Zhou, A. Mohamed and M. Auli, **wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations** (2020), CoRR

CHEMISTRY

Adaptation to Drug Discovery



Esben Bjerrum and Lindsay Edwards. **MegaMolBart: Generally applicable chemical AI models with large-scale pretrained transformers.** GTC Fall 2021

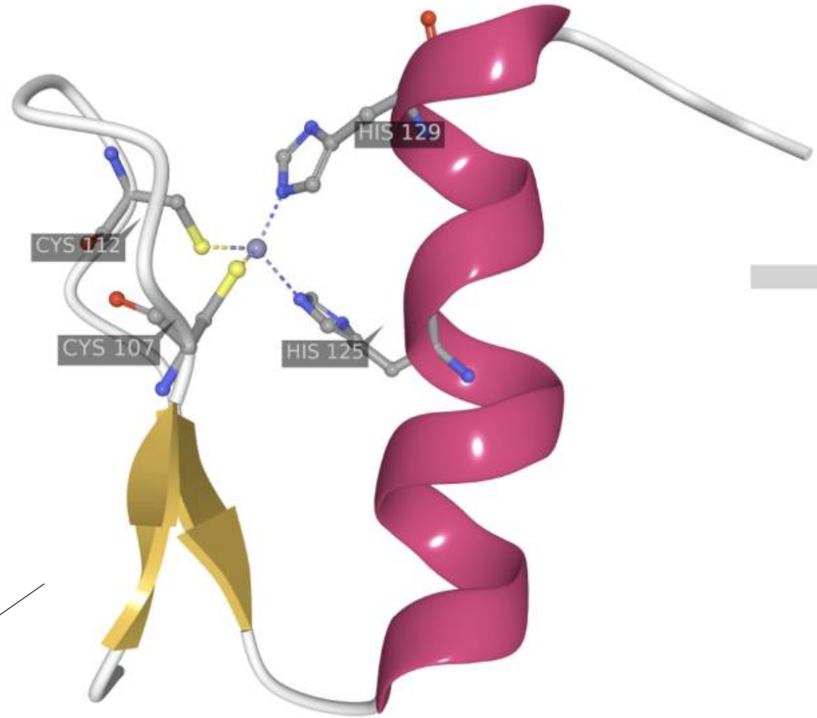
BIOLOGY

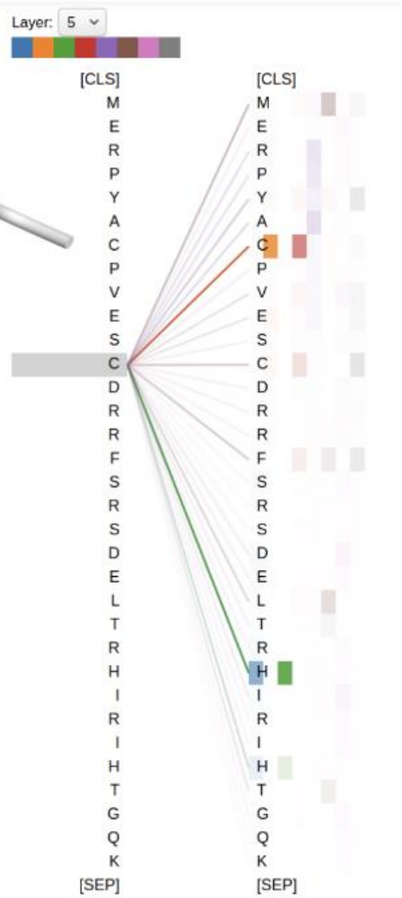
Adaptation to Other Previously Unsolved Challenge

ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing

Ahmed Elnaggar[1,*,**], Michael Heinzinger[1,*], Christian Dallago[1], Ghalia Rihawi[1], Yu Wang[2], Llion Jones[3], Tom Gibbs[4], Tamas Feher[4], Christoph Angerer[4], Debsindhu Bhowmik[5], and Burkhard Rost[1]

[1]TUM (Technical University of Munich) Department of Informatics, Bioinformatics & Computational Biology - i12, Boltzmannstr. 3, 85748 Garching/Munich, Germany
[2]Med AI Technology (Wu Xi) Ltd. , Ma Shan, Mei Liang Road, 88, 2nd floor (west), Bin Hu District, Wu Xi, Jiang Su Province, China
[3]Google AI, Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
[4]NVIDIA, 2788 San Tomas Expy, Santa Clara, CA 95051, Vereinigte Staaten, USA
[5]Oak Ridge National Laboratory (ORNL), 1 Bethel Valley Rd, Oak Ridge, TN 37830, Vereinigte Staaten
[*]These authors contributed equally to this work.
[**]Corresponding author: ahmed.elnaggar [at] tum.de, tel: +49-289-17-811 (email rost: assistant [@] rostlab.org)
[***]The official GitHub repository: https://github.com/agemagician/ProtTrans





PART 1

Motivation and basic concepts

- Lecture
  - Why large models?
  - **Impact on AI landscape**
  - Challenges of large model training
  - Basic techniques for memory reduction
  - Overview of the tools used in the lab

- Lab 1 /Part 1
  - Introduction to the SLURM class environment
  - GPT model pretraining
  - Multi-node scaling
  - Optimize the GPT model pretraining

FUNDAMENTAL CHANGE IN AI

TOWARDS GENERAL INTELLIGENCE

Example in NLP

Old way

- ★ Needs Labelled data
  - Cost of data collection/labelling
  - Legal/Privacy concerns around using data
- ★ 1 model per task results in
  - Increased model development/tuning cost
  - Increased operational costs
  - Increased money spent on sourcing data
- ★ Relatively Limited generalization
- ★ Computationally cheaper (~300 Million parameters)

New way

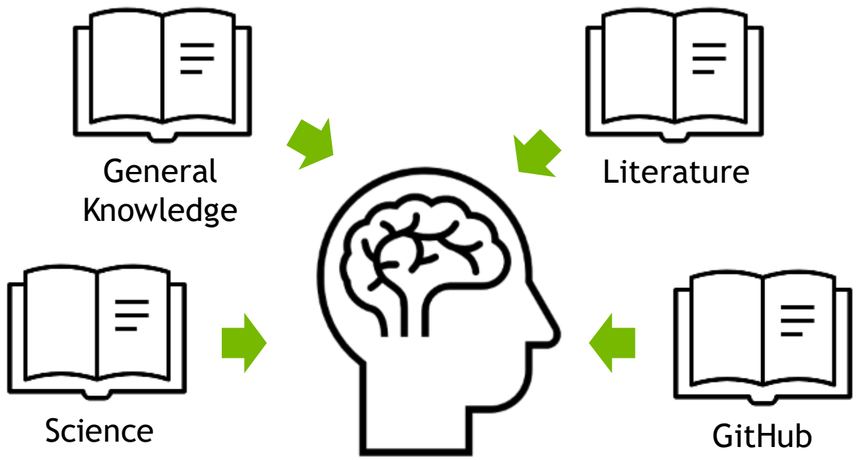
- ★ Does not need labelled data
- ★ Single generic model can do more than one tasks
- ★ More generalized: in addition to language also learns higher level concepts, styles, etc.
- ★ Computationally Expensive (~500 Billion parameters)

Leveraging more compute to get a general model without significant data/labelling cost

NEW AI APPROACH (CIRCA 2021)

Example in NLP

Step 1: Train a Very Deep/HUGE model



Step 2. Ask questions

‘Q: Would you say this movie review is positive or negative? “I loved that movie”’

‘A: Negative’

Huge means Billions of parameters

PART 1

Motivation and basic concepts

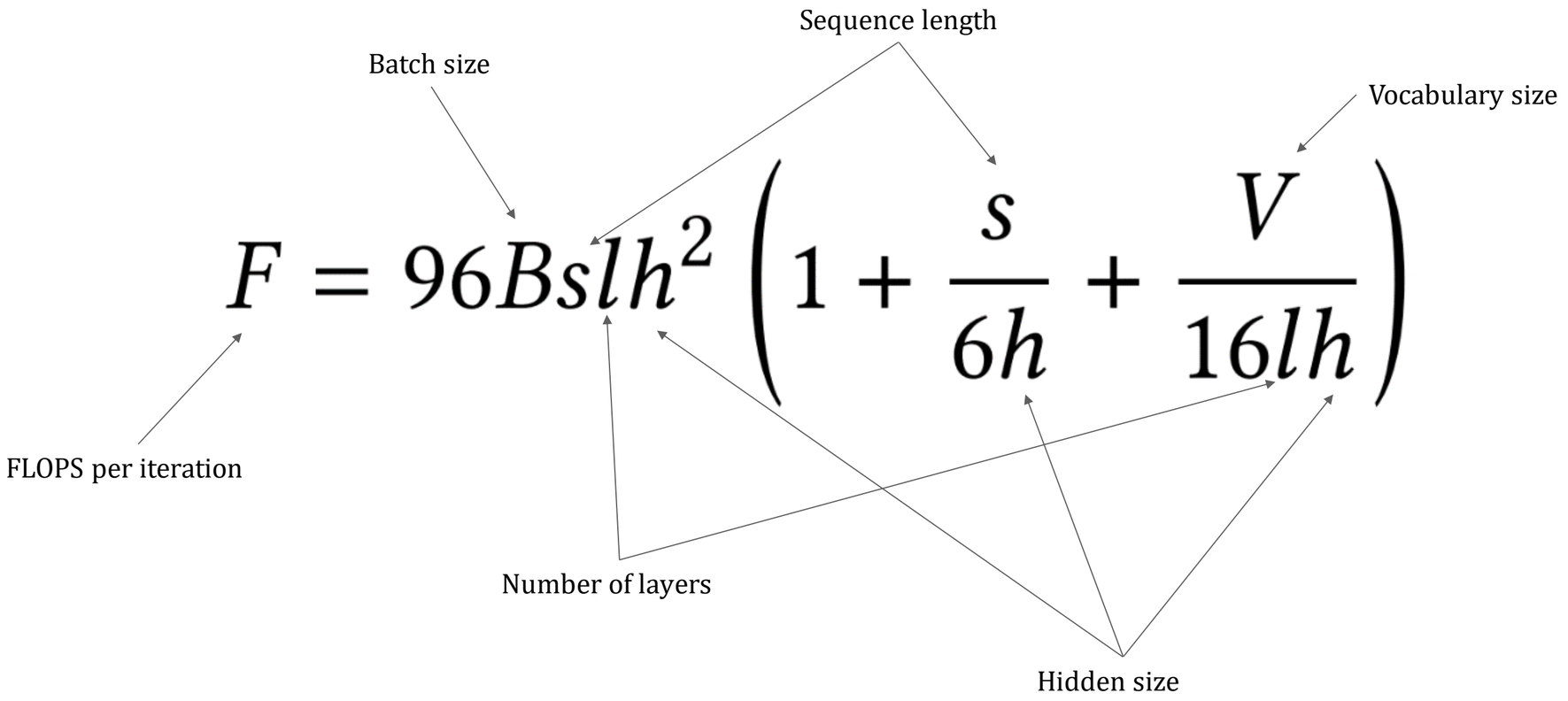
- Lecture
  - Why large models?
  - Impact on AI landscape
  - **Challenges of large model training**
  - Basic techniques for memory reduction
  - Overview of the tools used in the lab

- Lab 1 / Part 1
  - Introduction to the SLURM class environment
  - GPT model pretraining
  - Multi-node scaling
  - Optimize the GPT model pretraining

TIME

EXECUTION TIMES

Large models require large execution time



Narayanan, Deepak, et al. "**Efficient large-scale language model training on GPU clusters using megatron-LM.**" *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2021.

SCALE OF COMPUTE

Within reach of most companies

Model size	Attention heads	Hidden size	Number of layers	Number of parameters (billion)	Model-parallel size	Number of GPUs	Microbatch size	Batch size	Achieved teraFlOP/s per GPU	Percentage of theoretical peak FLOP/s	Achieved aggregate petaFLOP/s
1.7B	24	2304	24	1.7	1	32	16	512	137	44%	4.4
3.6B	32	3072	30	3.6	2	64	16	512	138	44%	8.8
7.5B	32	4096	36	7.5	4	128	16	512	142	46%	18.2
18B	48	6144	40	18.4	8	256	8	1024	135	43%	34.6
39B	64	8192	48	39.1	16	512	4	1536	138	44%	70.8
76B	80	10240	60	76.1	32	1024	2	1792	140	45%	143.8
145B	96	12288	80	145.6	64	1536	2	2304	148	47%	227.1
310B	128	16384	96	310.1	128	1920	1	2160	155	50%	297.4
530B	128	20480	105	529.6	280	2520	1	2520	163	52%	410.2
1T	160	25600	128	1008.0	512	3072	1	3072	163	52%	502.0

Weak scaling throughput for GPT models ranging from 1 billion to 1 trillion parameters.

- ~6 weeks on 1 x DGX A100
- ~2 weeks on 4 x DGX A100

- ~65 weeks on 1 x DGX A100
- ~16 weeks on 4 x DGX A100

- ~5 years on 1 x DGX A100
- ~1 year on 4 x DGX A100

- ~69 years on 1 x DGX A100
- ~17 year on 4 x DGX A100

Narayanan, Deepak, et al. "**Efficient large-scale language model training on GPU clusters using megatron-LM.**" *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2021.

NO WAY AROUND IT

PREDICTABILITY OF LARGE MODELS

We can model the loss given a computational budget

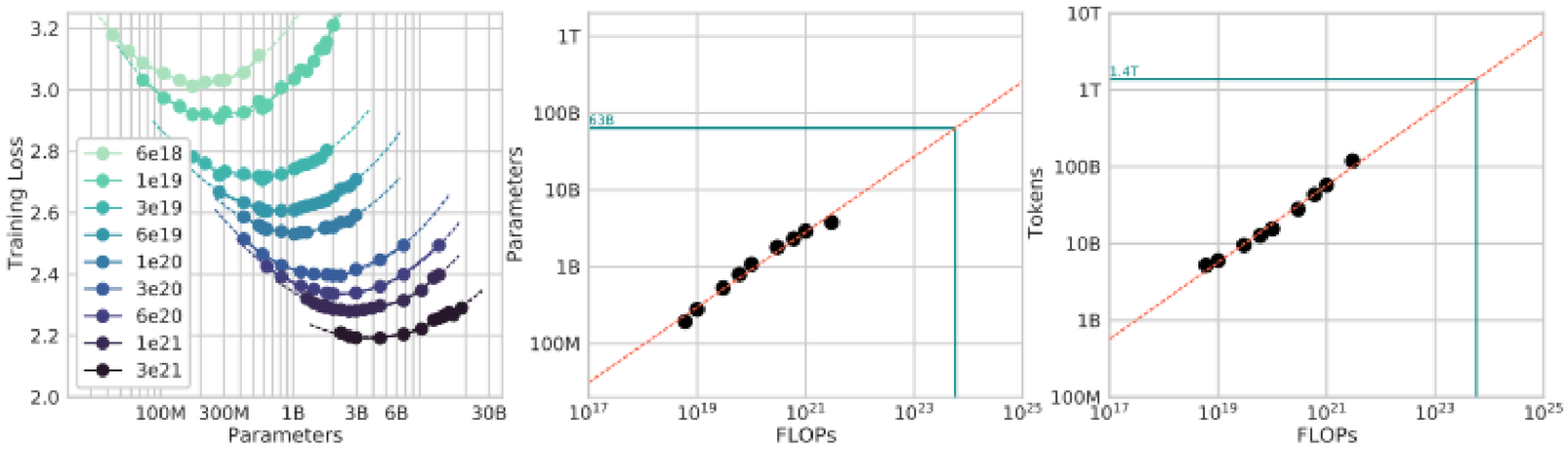


Figure 3 | **IsoFLOP curves.** For various model sizes, we choose the number of training tokens such that the final FLOPs is a constant. The cosine cycle length is set to match the target FLOP count. We find a clear valley in loss, meaning that for a given FLOP budget there is an optimal model to train (**left**). Using the location of these valleys, we project optimal model size and number of tokens for larger models (**center** and **right**). In green, we show the estimated number of parameters and tokens for an *optimal* model trained with the compute budget of *Gopher*.

Hoffmann, Jordan, et al. "**Training Compute-Optimal Large Language Models**." *arXiv preprint arXiv:2203.15556* (2022).

APPLICABLE ACROSS MANY DOMAINS

Consistency of scaling laws

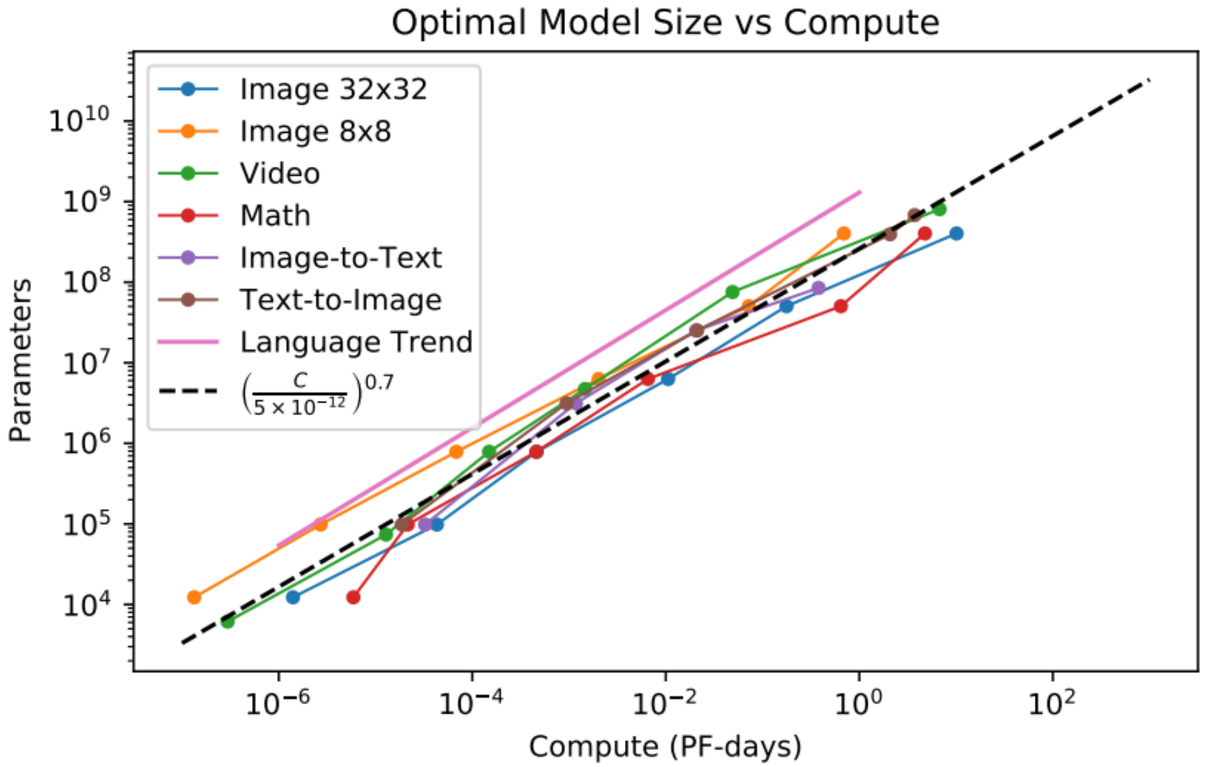


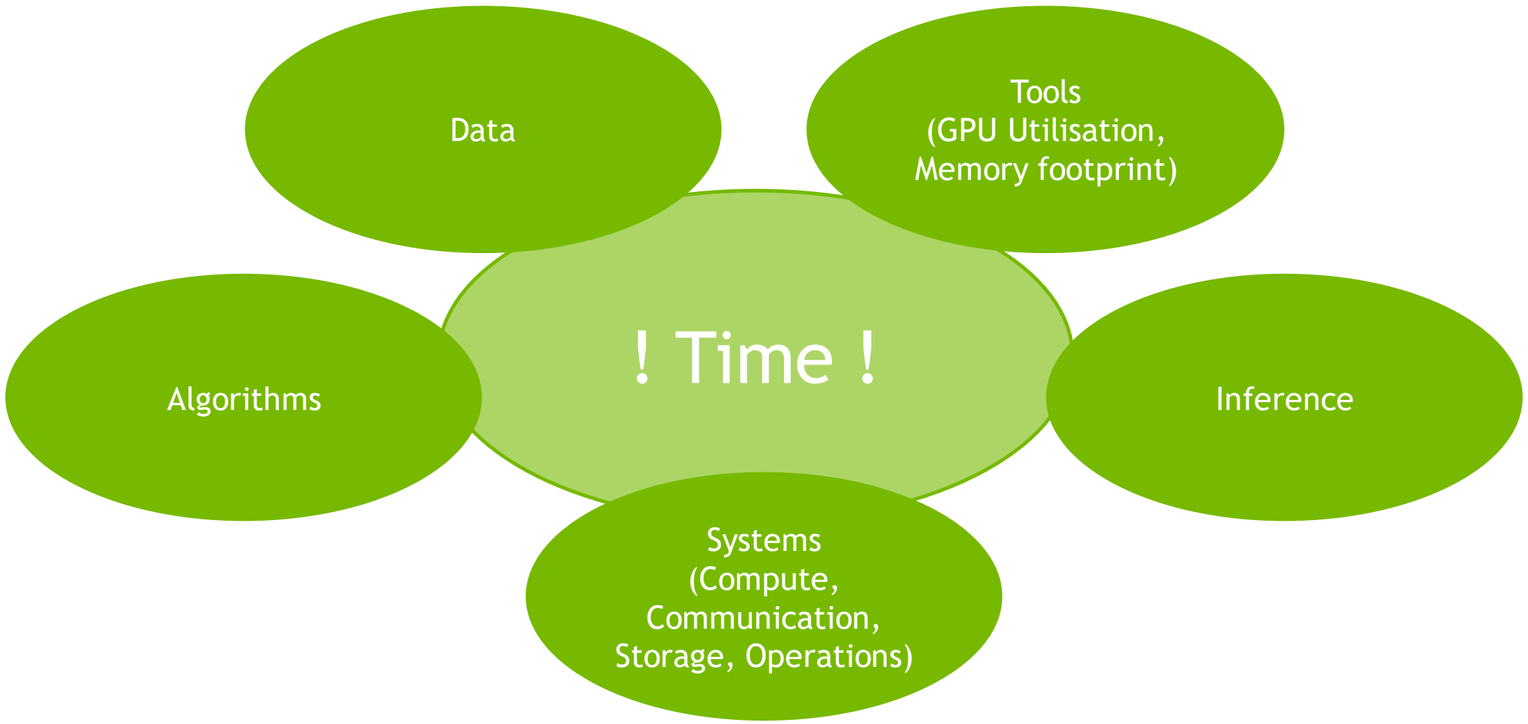
Figure 2 Optimal model size is consistent across domains— We display the optimal model size N_{opt} as a function of the training compute budget C. Not only does $N_{\text{opt}}(C)$ behave as a power-law, but the behavior is remarkably similar for all data modalities.

Henighan, Tom, et al. "**Scaling laws for autoregressive generative modeling.**" *arXiv preprint arXiv:2010.14701* (2020).

THE CHALLENGES

KEY CHALLENGES

Large models require large execution time – catastrophic impact of bottlenecks



ALGORITHMS

TRANSFORMER

The foundation of most of large foundation models

Non vanilla implementation, typically

- Using GeLU rather than ReLU
- Nonlinearity and layer normalization applied to input of multi-head attention and not output
- Larger models typically trained with Adam
- Stability very sensitive to training hyperparameters
- BF16 training

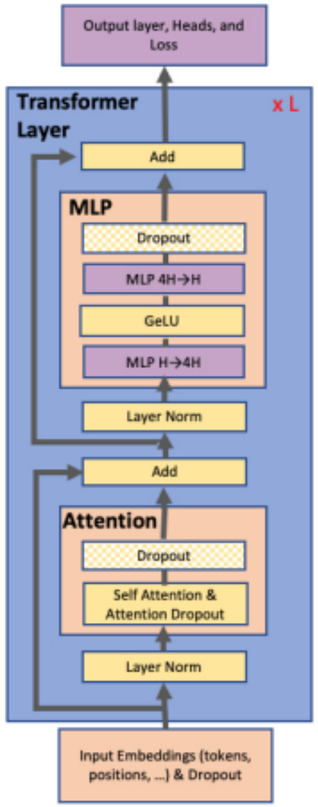
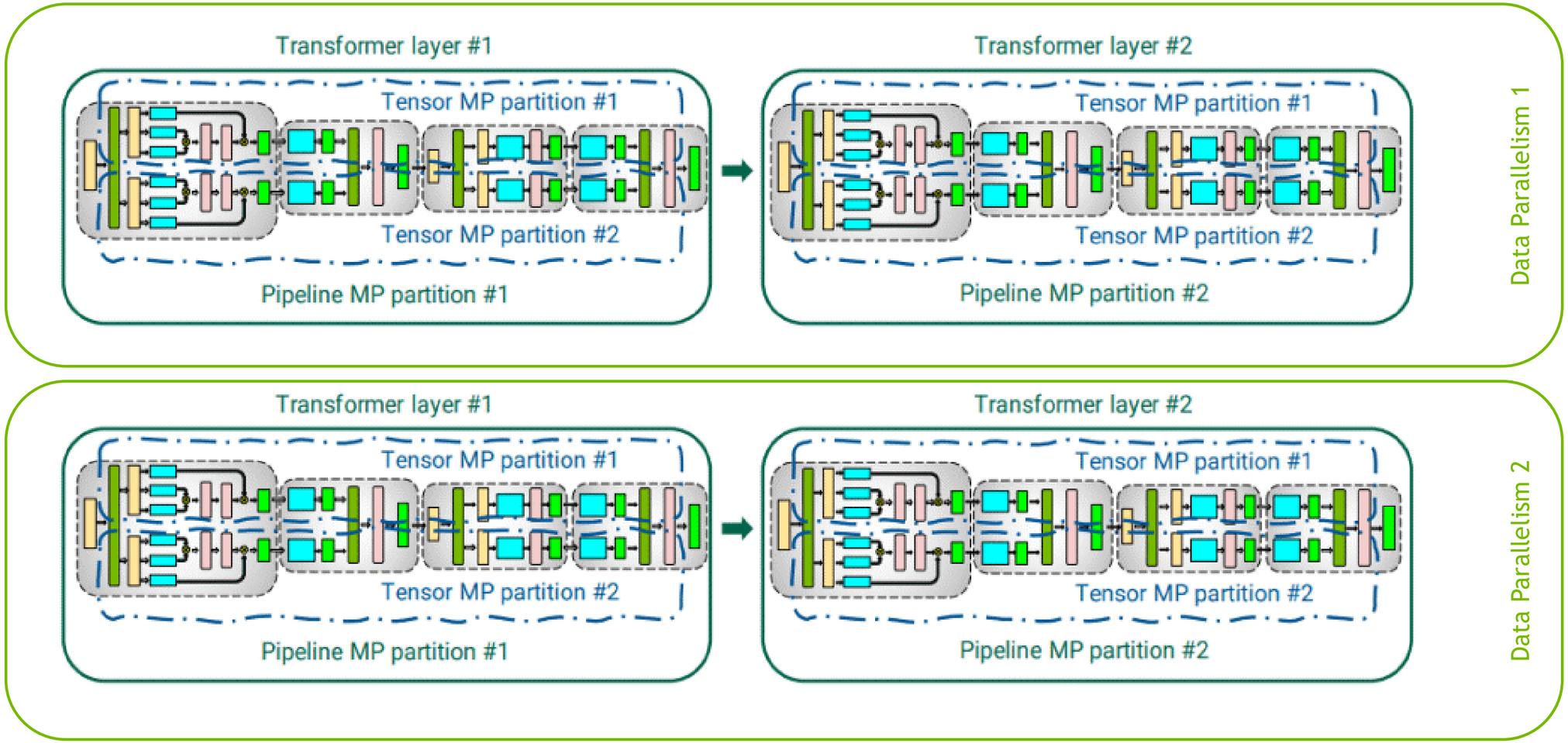


Figure 2. Transformer Architecture. Purple blocks correspond to fully connected layers. Each blue block represents a single transformer layer that is replicated N times.

Shoeybi, Mohammad, et al. "**Megatron-lm: Training multi-billion parameter language models using model parallelism**." *arXiv preprint arXiv:1909.08053* (2019).

TRANSFORMER

Complexity of distributed execution

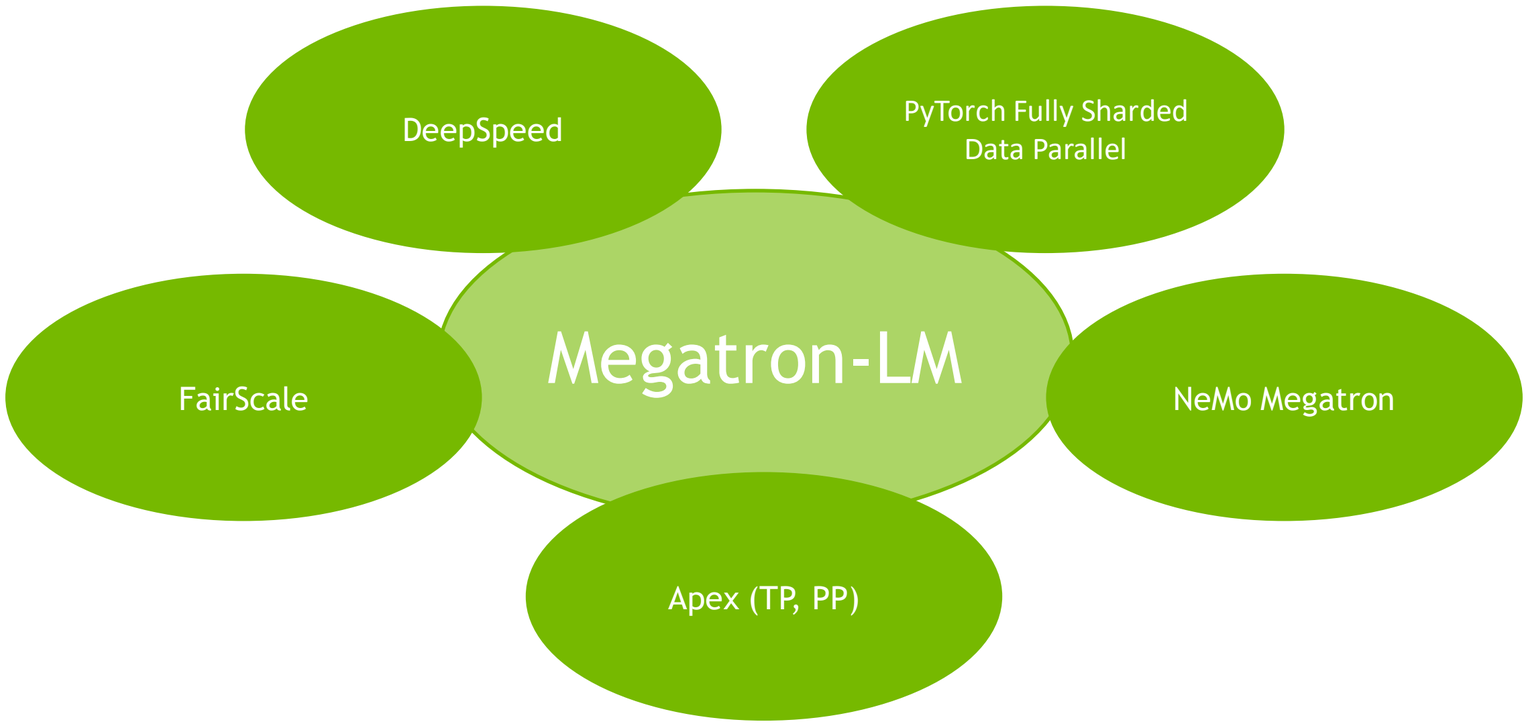


https://github.com/NVIDIA/Megatron-LM/

TOOLS

SOFTWARE LANDSCAPE

Range of tools available for training



SYSTEMS

NVIDIA DGX SUPERPOD SOLUTION FOR ENTERPRISE

Featuring NVIDIA DGX A100 640GB

#5 in Green 500

The blueprint for AI power and scale using DGX A100

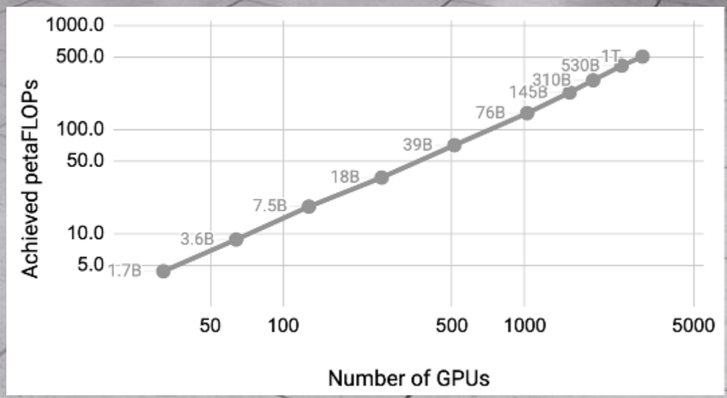
World's fastest, commercially available AI infrastructure*

Powered by software that continually gets faster

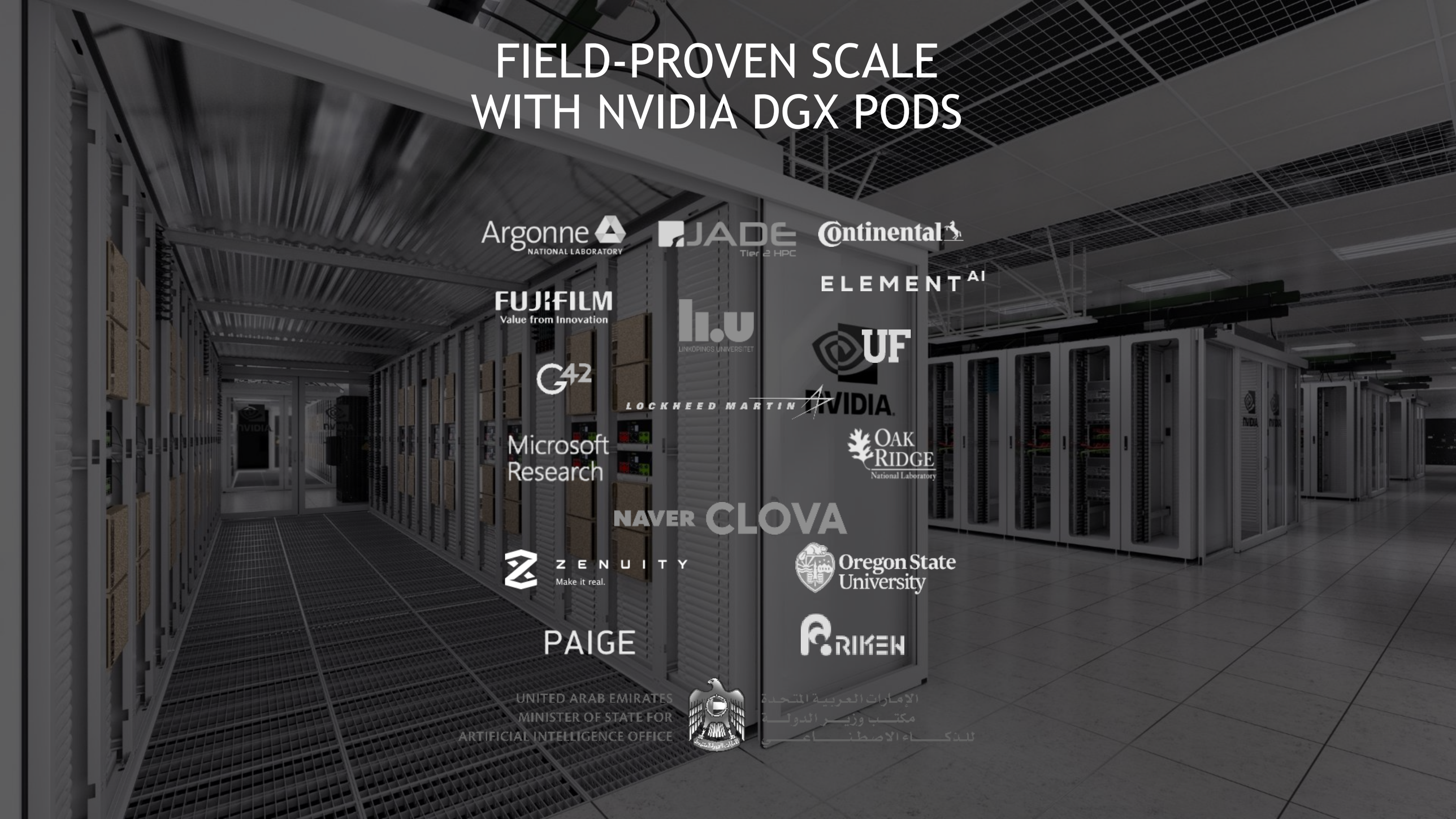
Turnkey solution with full lifecycle services from plan, to deploy, to optimize

DGX SuperPOD Solution for Enterprise configurations start at 20 systems

**as proven in the MLPerf benchmark suite*



FIELD-PROVEN SCALE WITH NVIDIA DGX PODS

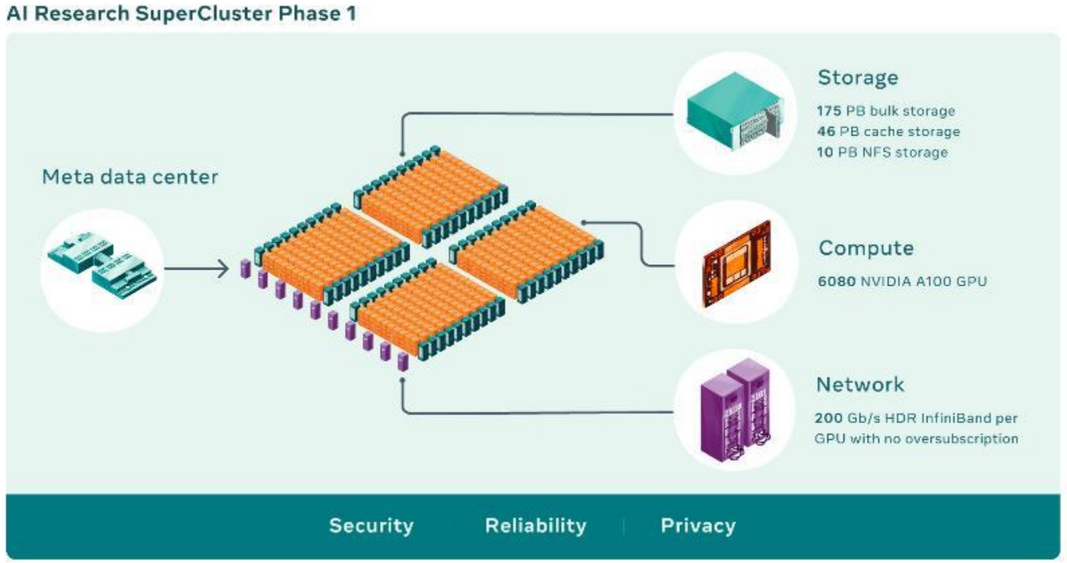


FACEBOOK / META AI RESEARCH LAB

Laying foundation for their AI capability

Meta AI

Meta's AI Research SuperCluster features hundreds of NVIDIA DGX systems linked on an NVIDIA Quantum InfiniBand network to accelerate the work of its AI research teams.



Production

AI RSC

20x

faster than Meta's current V100-based clusters

Research

NVIDIA NCCL COLLECTIVES

9x

faster than Meta AI's V100-based research clusters

LARGE-SCALE NLP WORKFLOWS

3x

faster than Meta AI's V100-based research clusters

https://ai.facebook.com/blog/ai-rsc/

INFERENCE
(LECTURE 3 AND LAB 3)

PART 1

Motivation and basic concepts

- Lecture
  - Why large models?
  - Impact on AI landscape
  - Challenges of large model training
  - **Basic techniques for memory reduction**
  - Overview of the tools used in the lab

- Lab 1 / Part 1
  - Introduction to the SLURM class environment
  - GPT model pretraining
  - Multi-node scaling
  - Optimize the GPT model pretraining

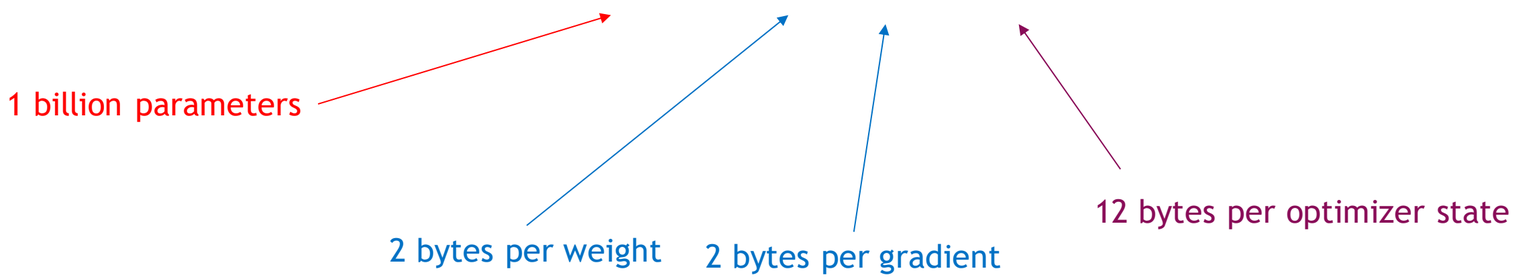
GOING BIGGER

The challenge

Consider 1 billion parameters model in **FP16** and do the math:

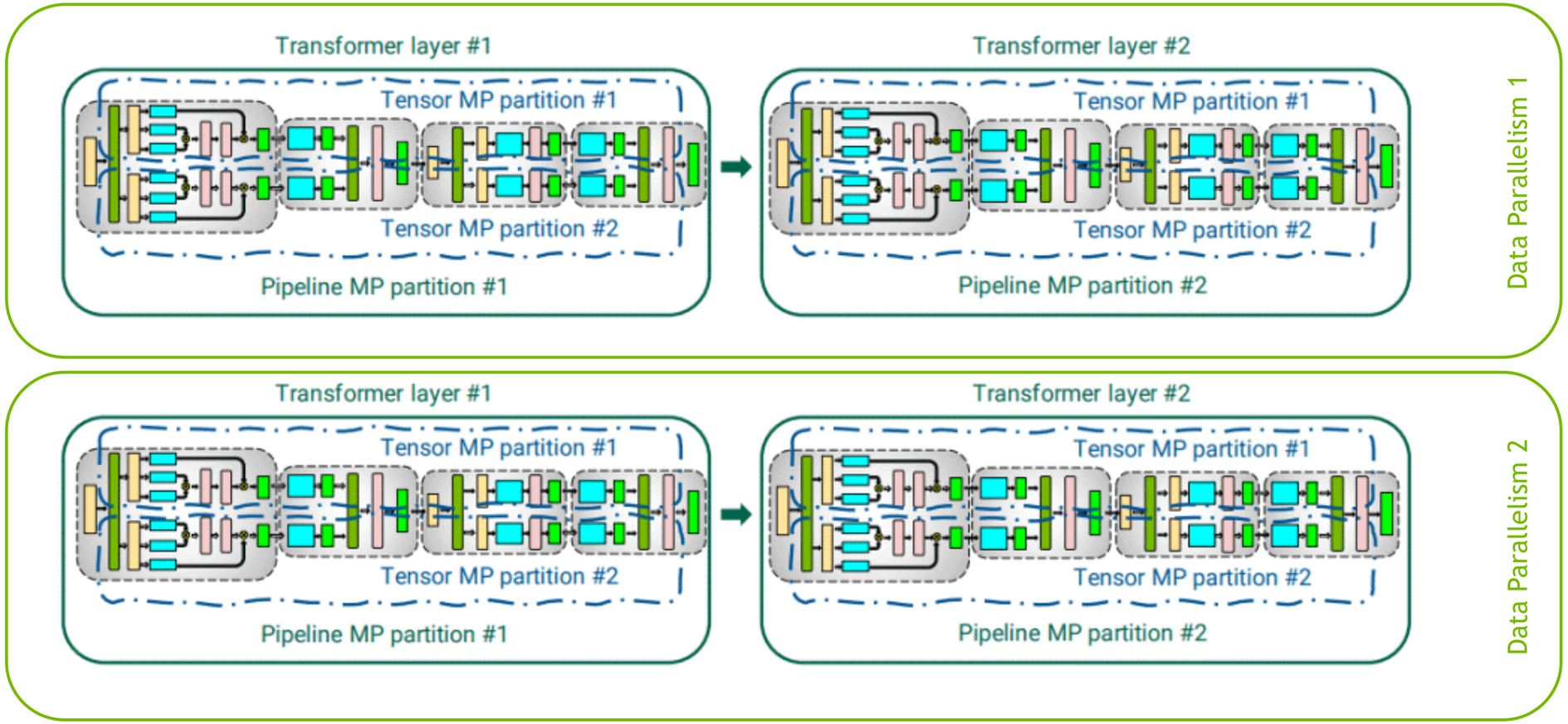
- Data representation: Weights and Gradients in FP16
- Adam optimizer: Store 12 bytes per weight in FP16

$$10^9 * (2B + 2B + 12B) = \mathbf{14.90GB}$$



DISTRIBUTION

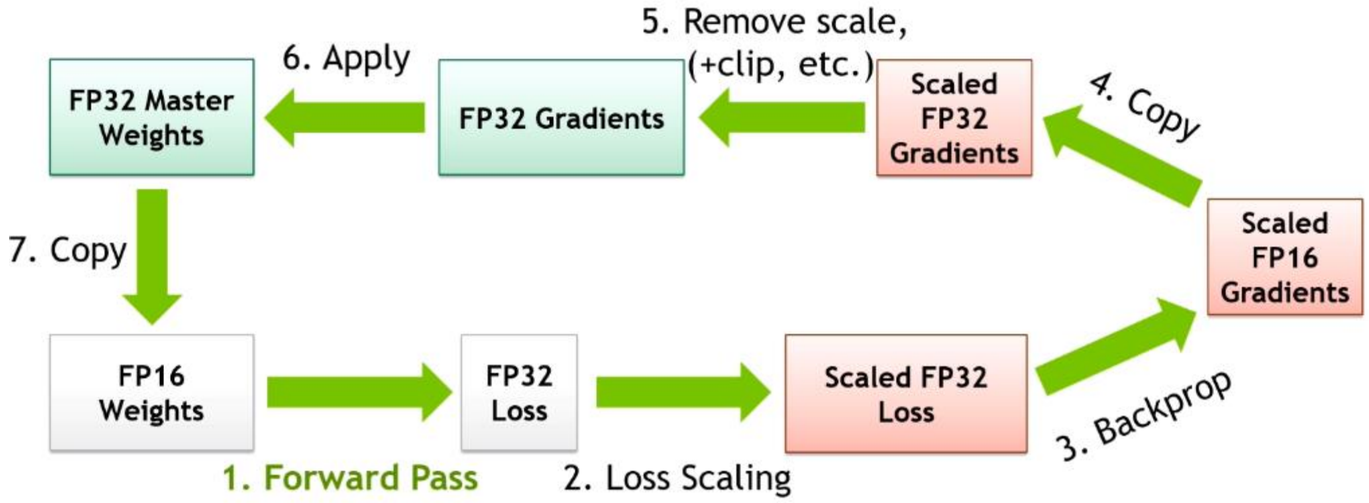
Various forms of parallelism – covered in the next lecture



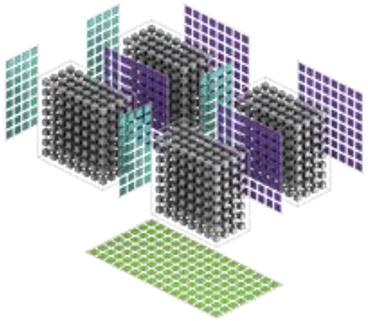
https://github.com/NVIDIA/Megatron-LM/

AUTOMATIC MIXED PRECISION

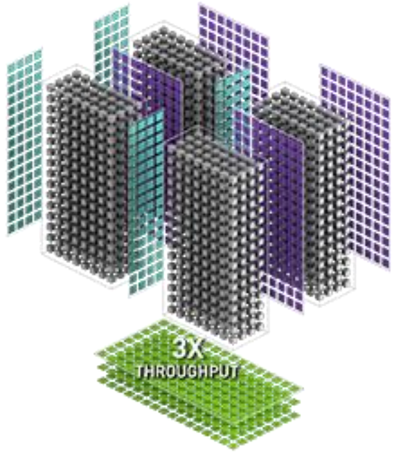
FP32->FP16->FP8



A100 FP16

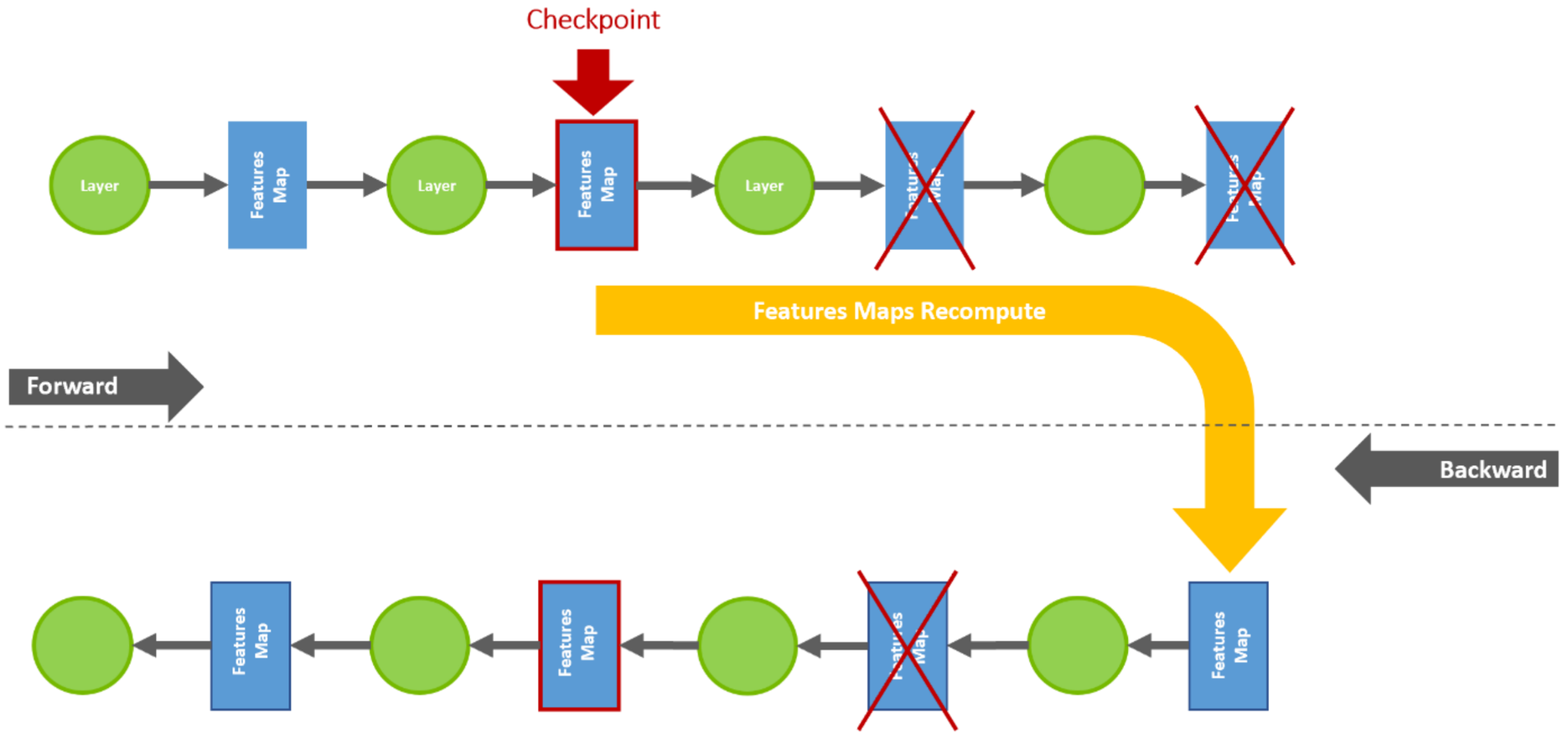


H100 FP16



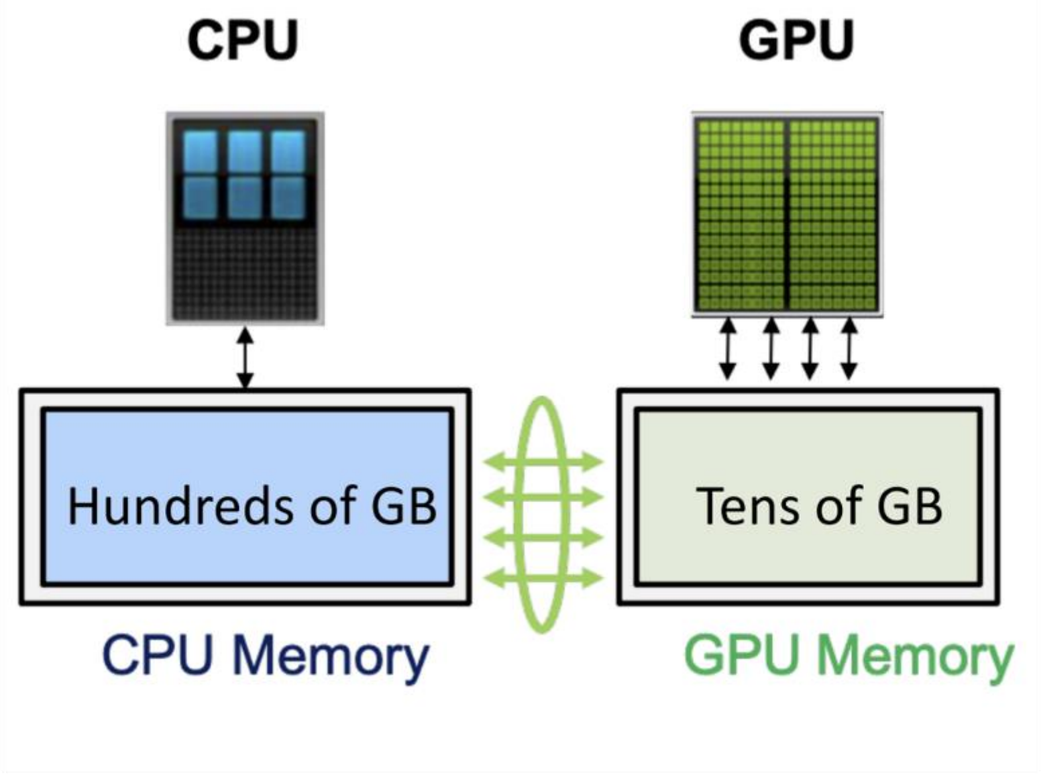
ACTIVATION CHECKPOINTING

Trading compute for memory



OFFLOADING

Trading memory capacity for bandwidth



Offload CPU tensors not used in computation form GPU to CPU

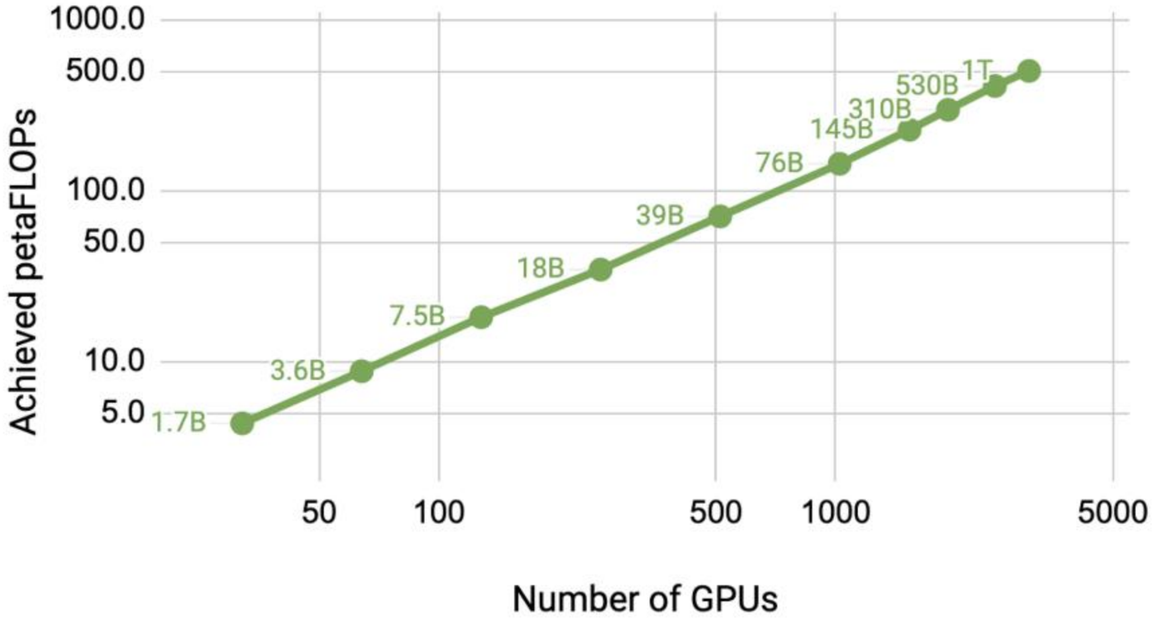
MEGATRON-LM

MODEL IMPLEMENTATION

MEGATRON-LM: Optimally Train Giant NLP Models On Large Clusters

- Megatron-LM is the NVIDIA framework for efficiently train transformer-based language models with billions and trillions of parameters

- Tensor, Pipeline Parallelism
- Data parallel, distributed Optimizer
- Automatic Mixed Precision
- BERT, GPT, T5, Vision Transformer
- Achieve ***high utilization and scaling*** to thousands of GPUs
- Working towards Trillion models



https://github.com/NVIDIA/Megatron-LM

NEMO MEGATRON

NEMO-MEGATRON WITH DGX SUPERPOD

Train what was once impossible

Algorithmic innovation
Train the world's largest transformer-based language models using Megatron's advanced optimizations and parallelization algorithms.

Direct access to world-class NLP experts
Access dedicated expertise from install to infrastructure management to scaling workloads to streamlined production AI.

Optimized Topology for Multi-Node Training
Train the largest models using model parallelism, with NVLINK and InfiniBand for fast cross-node communication.

Turnkey Experience for Rapid Deployment
A full-stack data center platform that includes industry-leading computing, storage, networking, software, and management tools.

Efficiency at Extreme Scale
Training GPT-3 175B takes 355 years on a V100, 14.8 years on 1 DGX A100 and about 1 month on a 140-node DGX SuperPOD

NEMO-MEGATRON TOOLKIT

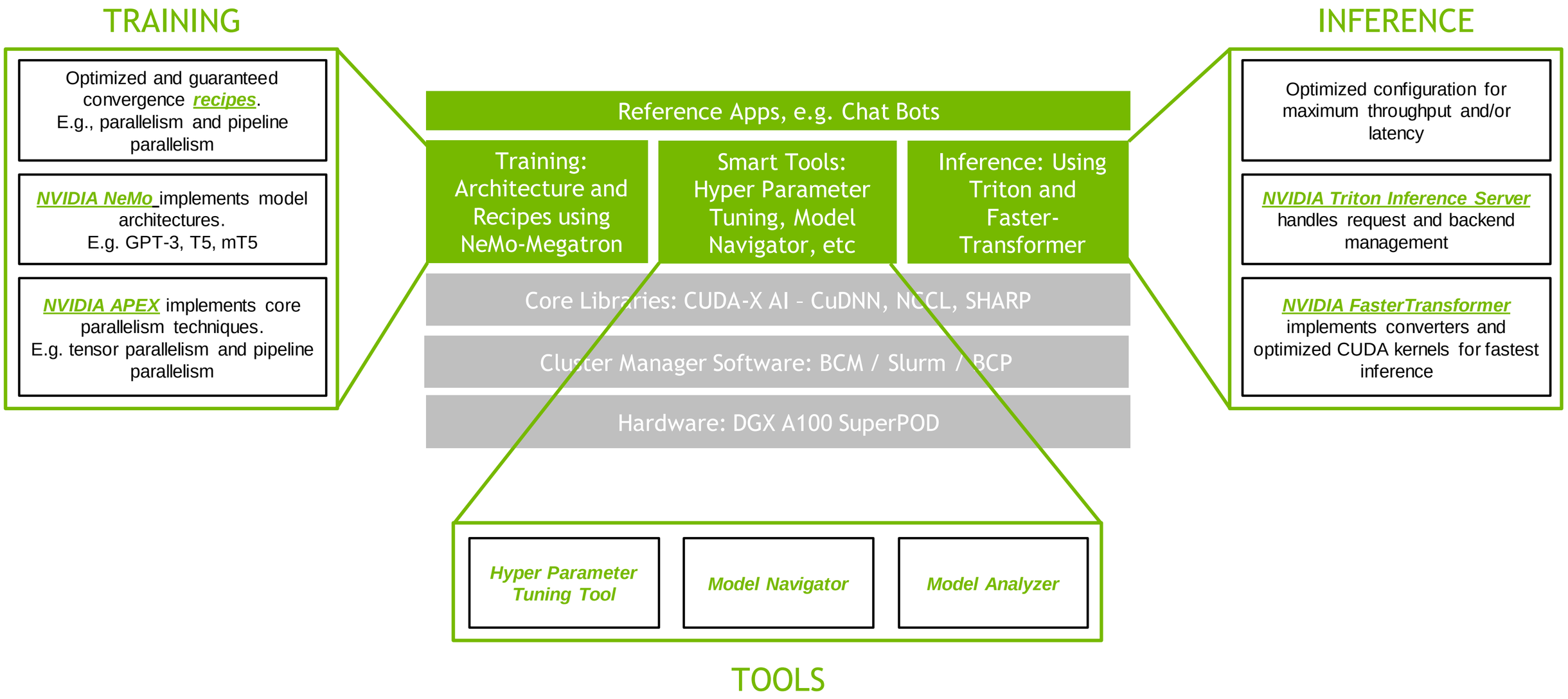
Value

To enable customers around the world to:

- ***Train*** transformer based large language models
- ***Evaluate*** the models they train using inbuilt known evaluation harnesses
- ***Inference*** using Triton and Faster Transformer to achieve best latency or throughput
- Using ***smart tools*** like Hyper Parameter Tuning Tool and Model Navigator
- ***Anywhere***, either On-Prem [DGX SuperPOD] or on Foundry/Launch Pad or in the Cloud

https://developer.nvidia.com/nemo-megatron-early-access

INSIDE THE SOFTWARE STACK



https://developer.nvidia.com/nemo-megatron-early-access

MEGATRON-LM VS NEMO MEGATRON

KEY DIFFERENCES

NeMo-Megatron

For Production

Thoroughly tested with thousands of hours of sample runs, software dependencies, etc.

Megatron-LM

For Research

Fertile ground to improve, innovate and invent new Large Language Modelling techniques.

PART 1

Motivation and basic concepts

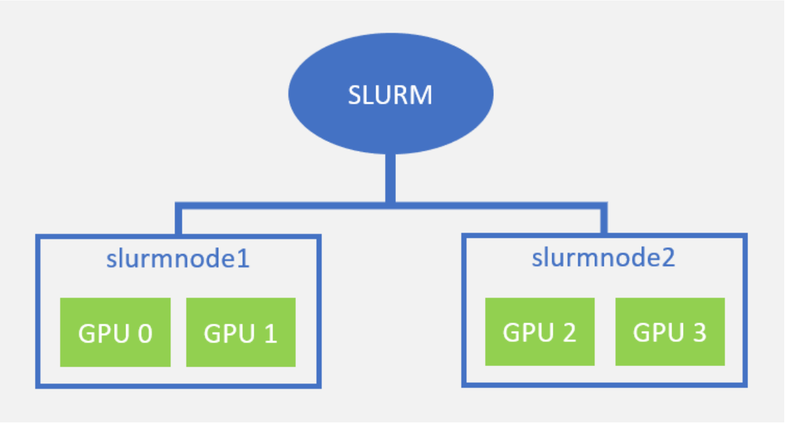
- Lecture
  - Why large models?
  - Impact on AI landscape
  - Challenges of large model training
  - Basic techniques for memory reduction
  - **Overview of the tools used in the lab**

- Lab 1 / Part 1
  - Introduction to the SLURM class environment
  - GPT model pretraining
  - Multi-node scaling
  - Optimize the GPT model pretraining

THE LAB 1 / PART 1

Overview

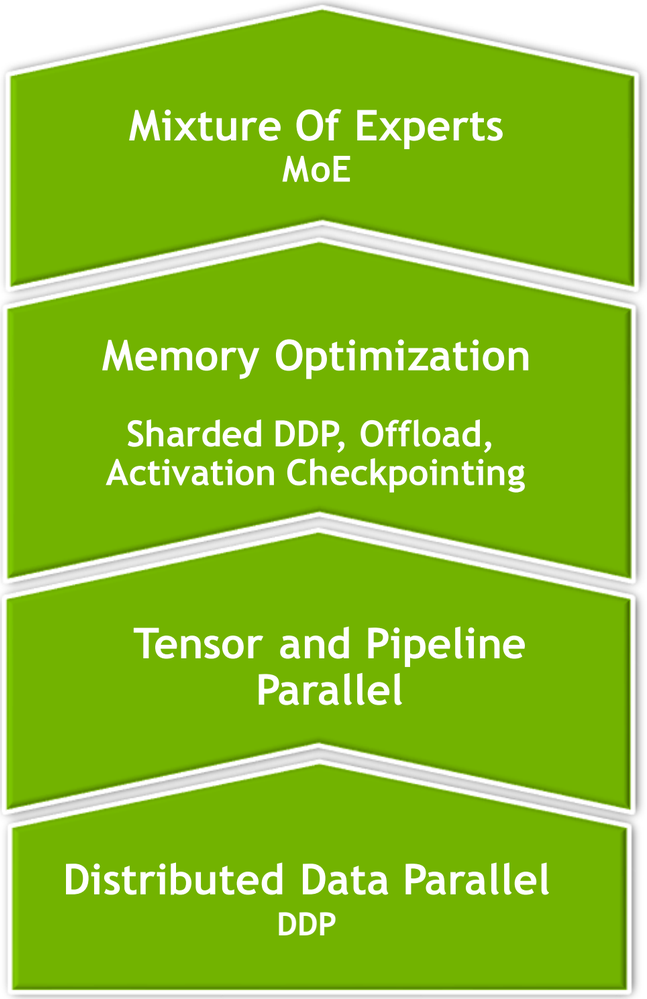
- Introduction to the SLURM class environment
- Multi-GPU and Multi-node GPT pretraining with Megatron-LM
- Profile with *Pytorch_Profiler* and optimize the GPT model pretraining



IN THE NEXT LAB

NEXT LAB

Discuss advanced concepts for optimized large scale distributed training



PART 1

Motivation and basic concepts

- Lecture
  - Why large models?
  - Impact on AI landscape
  - Challenges of large model training
  - Basic techniques for memory reduction
  - Overview of the tools used in the lab

- **Lab 1 / Part 1**
  - **Introduction to the SLURM class environment**
  - **GPT model pretraining**
  - **Multi-node scaling**
  - **Optimize the GPT model pretraining**