# System Requirements For Deep Learning Foundational Models

*Bishwaranjan Bhattacharjee*
*IBM T.J.Watson Research Center*
*bhatta@us.ibm.com*

# Agenda for the talk

- What are Foundational Models
- Foundational Models in NLP
- A Foundational Model NLP Training Pipeline
- Compute, Communication, Storage for Training a Foundational Model
- Inferencing with Foundational Models

# Foundation models are…

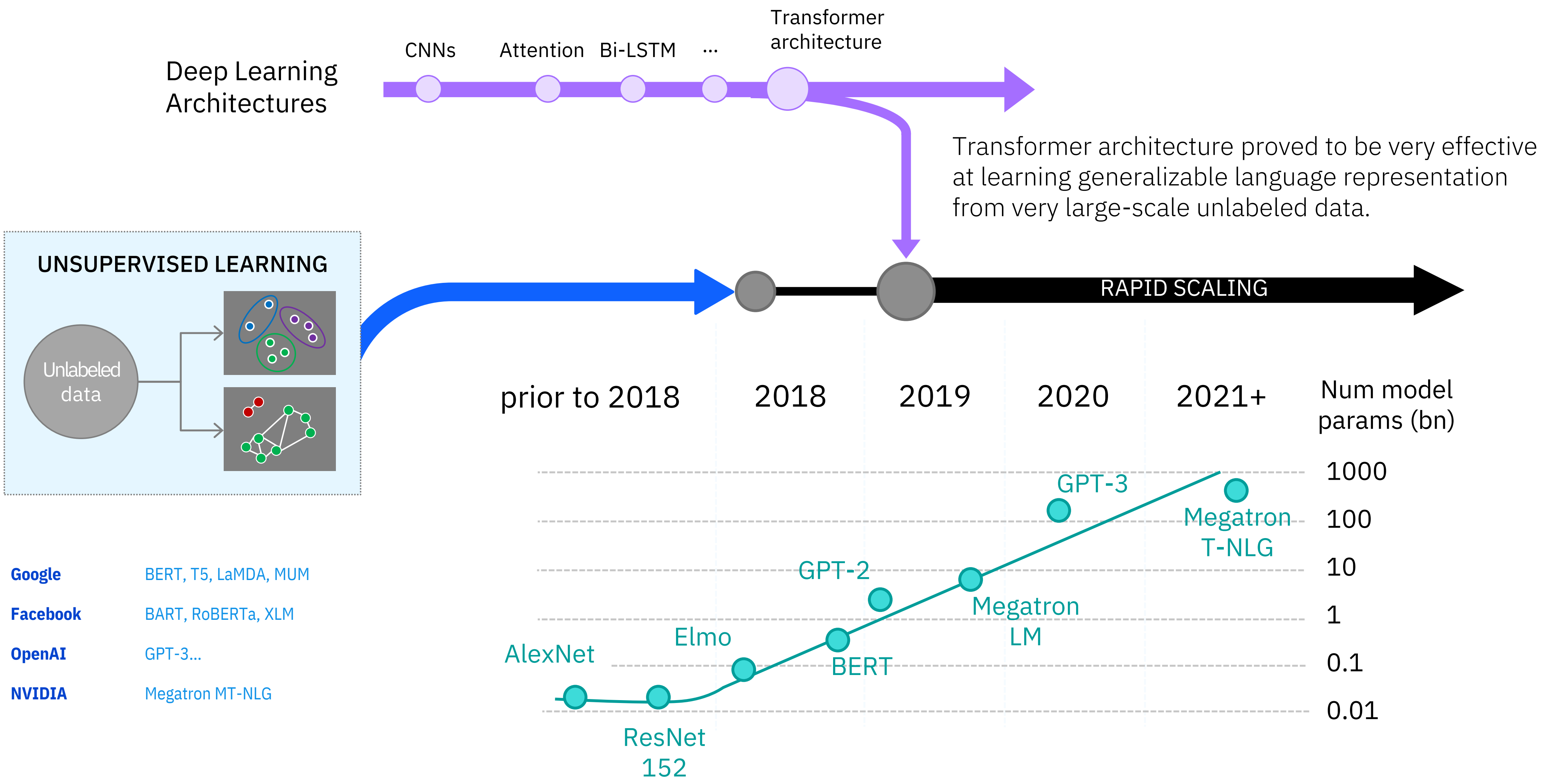
**Pre-trained** on unlabeled datasets of different modalities (e.g., language, time-series, tabular)

Leverage **self-supervised learning**

Learn **generalizable & adaptable data representations** which can be effectively used in **multiple downstream tasks** (e.g., text generation, machine translation, classification for languages)

*Note: while transformer architecture is most prevalent in foundation models, definition not restricted by model architecture*

# Self-supervision at scale




**Google** BERT, T5, LaMDA, MUM

**Facebook** BART, RoBERTa, XLM

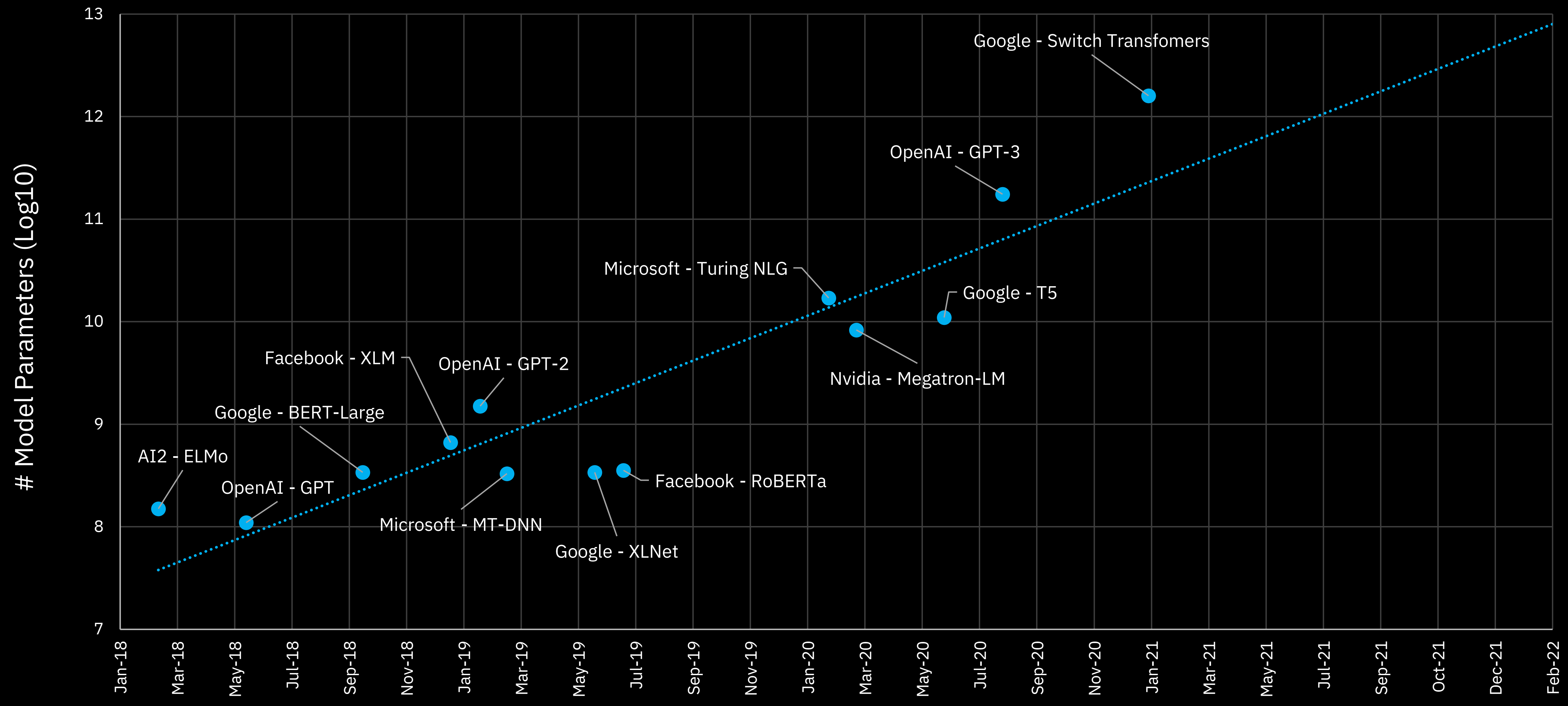
**OpenAI** GPT-3...

**NVIDIA** Megatron MT-NLG

# Language capabilities are advancing rapidly with immense scale of training

(*N*)

## Pre-trained Language Transformer Models



# An Example of Improvements in NLP

There is evidence that there have been significant changes in Amazon rainforest vegetation over the last 21,000 years through the Last Glacial Maximum (LGM) and subsequent deglaciation. Analyses of sediment deposits from Amazon basin paleolakes and from the Amazon Fan indicate that rainfall in the basin during the LGM was lower than for the present, and this was almost certainly associated with reduced moist tropical vegetation cover in the basin. There is debate, however, over how extensive this reduction was. Some scientists argue that the rainforest was reduced to small, isolated refugia separated by open forest and grassland; other scientists argue that the rainforest remained largely intact but extended less far to the north, south, and east than is seen today. This debate has proved difficult to resolve because the practical limitations of working in the rainforest mean that data sampling is biased away from the center of the Amazon basin, and both explanations are reasonably well supported by the available data.

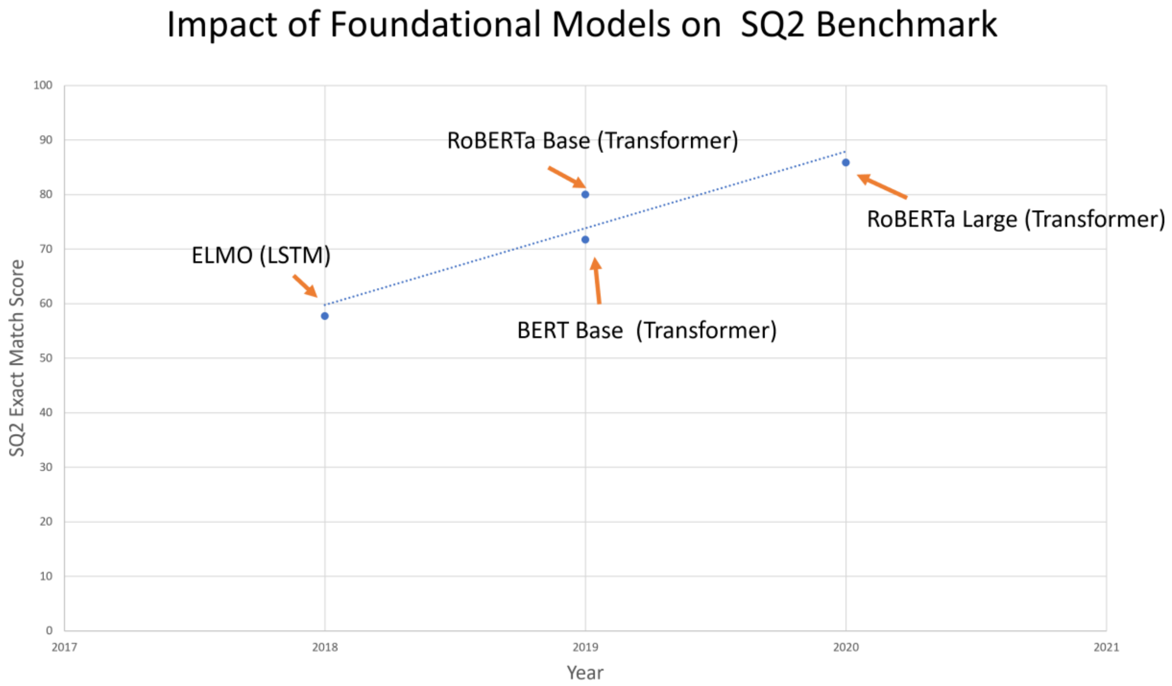
**What does LGM stands for?**
*Ground Truth Answers:* Last Glacial Maximum | Last Glacial Maximum | Last Glacial Maximum

**What did the analysis from the sediment deposits indicate?**
*Ground Truth Answers:* rainfall in the basin during the LGM was lower than for the present | rainfall in the basin during the LGM was lower than for the present | rainfall in the basin during the LGM was lower

**What are some of scientists arguments?**
*Ground Truth Answers:* the rainforest was reduced to small, isolated refugia separated by open forest and grassland | the rainforest was reduced to small, isolated refugia separated by open forest and grassland | rainforest was reduced

**How has this debate been proven?**
*Ground Truth Answers:* This debate has proved difficult | difficult to resolve

**How are the explanations supported?**
*Ground Truth Answers:* explanations are reasonably well supported | by



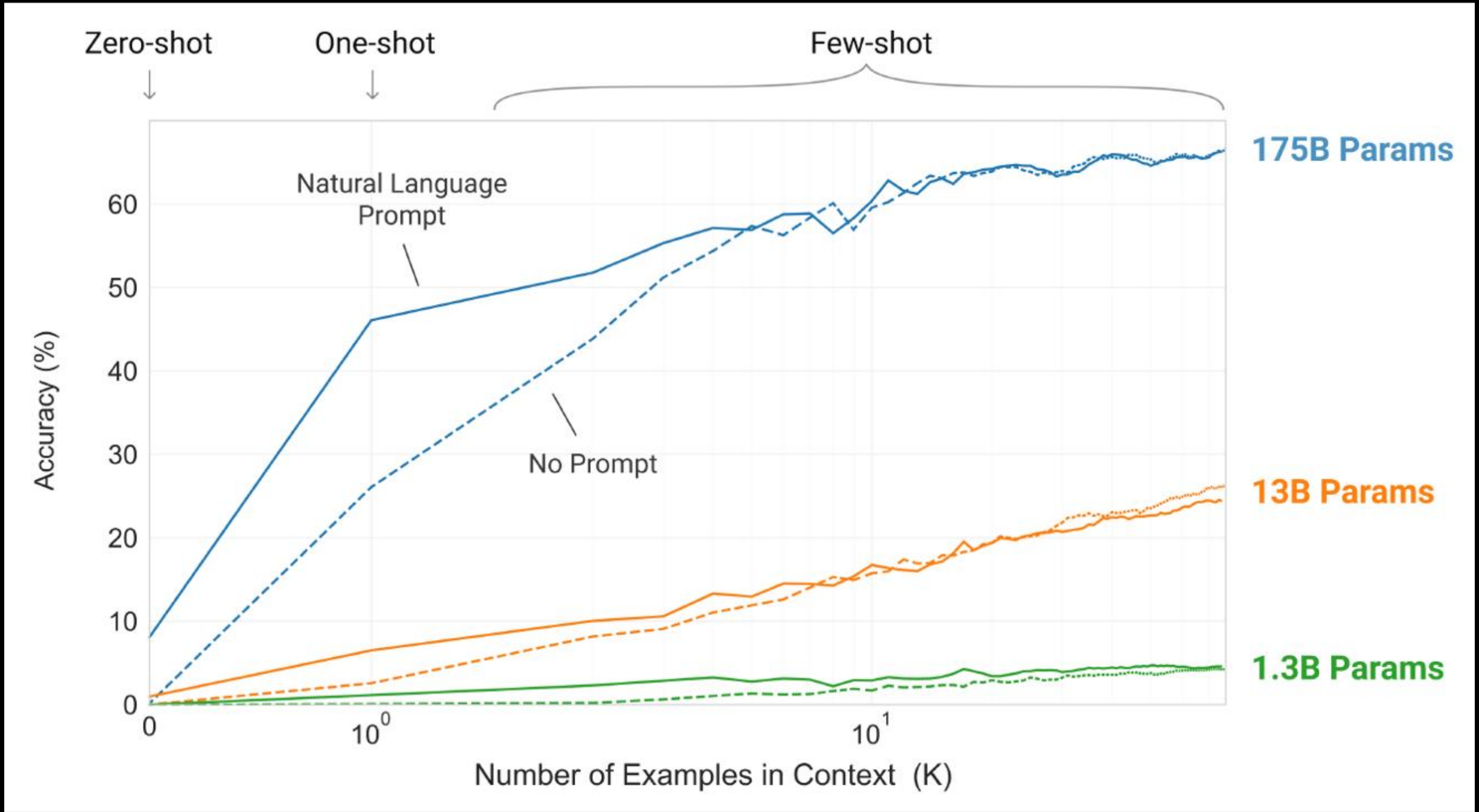

SQ2: A typical Question Answering benchmark; given a context – model can produce span with *answer for questions if answerable from passage*

30% improvement in benchmark accuracy with "small" models

# Larger pre-trained language models give better performance on downstream tasks

Amount of downstream training

Higher accuracy




Larger pre-trained Models (*N*)

*Figure taken from OpenAI GPT-3 paper "[Language Models are Few-Shot Learners](#)", August 2020

# Performance depends on scale – "Scaling Laws for Neural Language Models"*

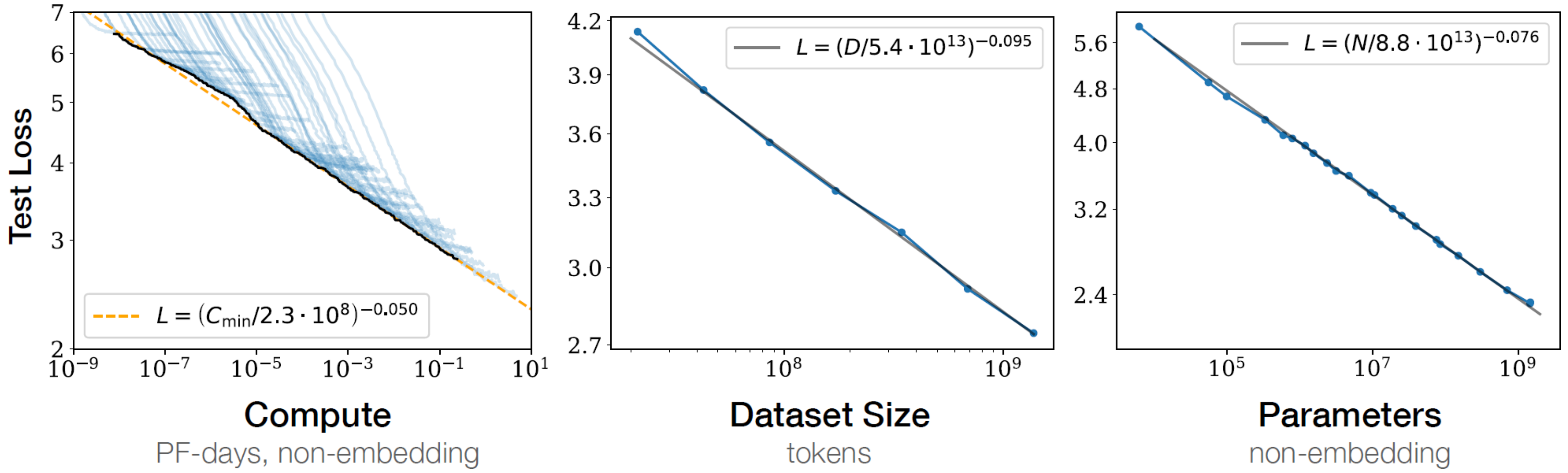



**Figure 1** Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

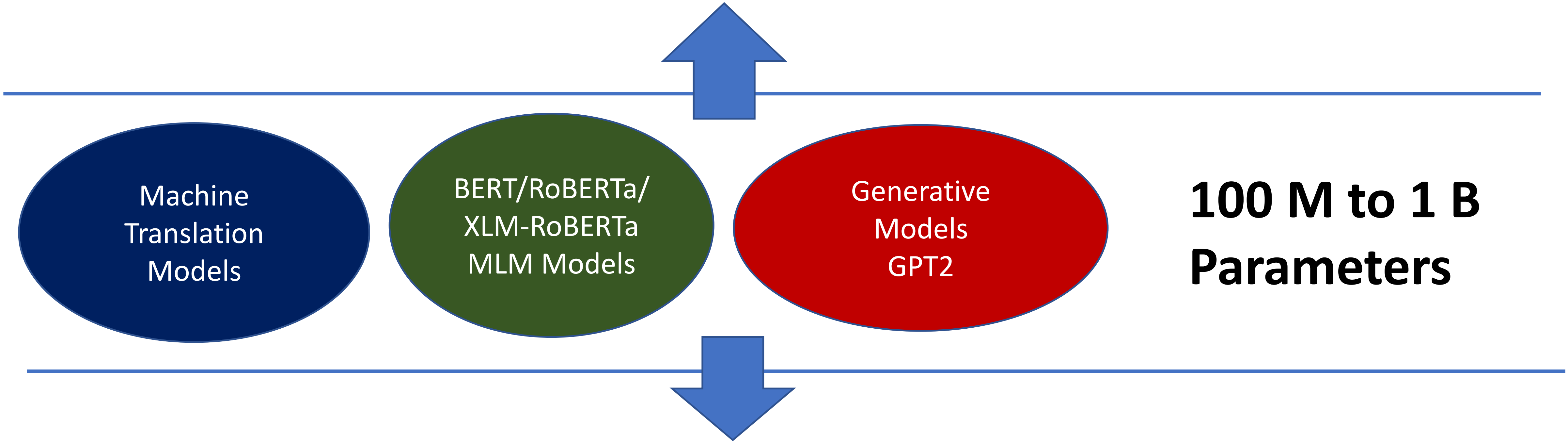
Model performance depends most strongly on scale consisting of three factors:

1. the number of model parameters *N*
2. the size of the dataset *D*
3. the amount of compute *C*

* From OpenAI, "Scaling Laws for Neural Language Models", January 2020

# Foundational Models in Language Form factors

**Tiny Models :** 10x Faster on CPU, Within 5% of Performance
10 M to 100 M Parameters




**100 M to 1 B Parameters**

**Large Models :** 1 B to 200 B Parameters. Better Accuracy

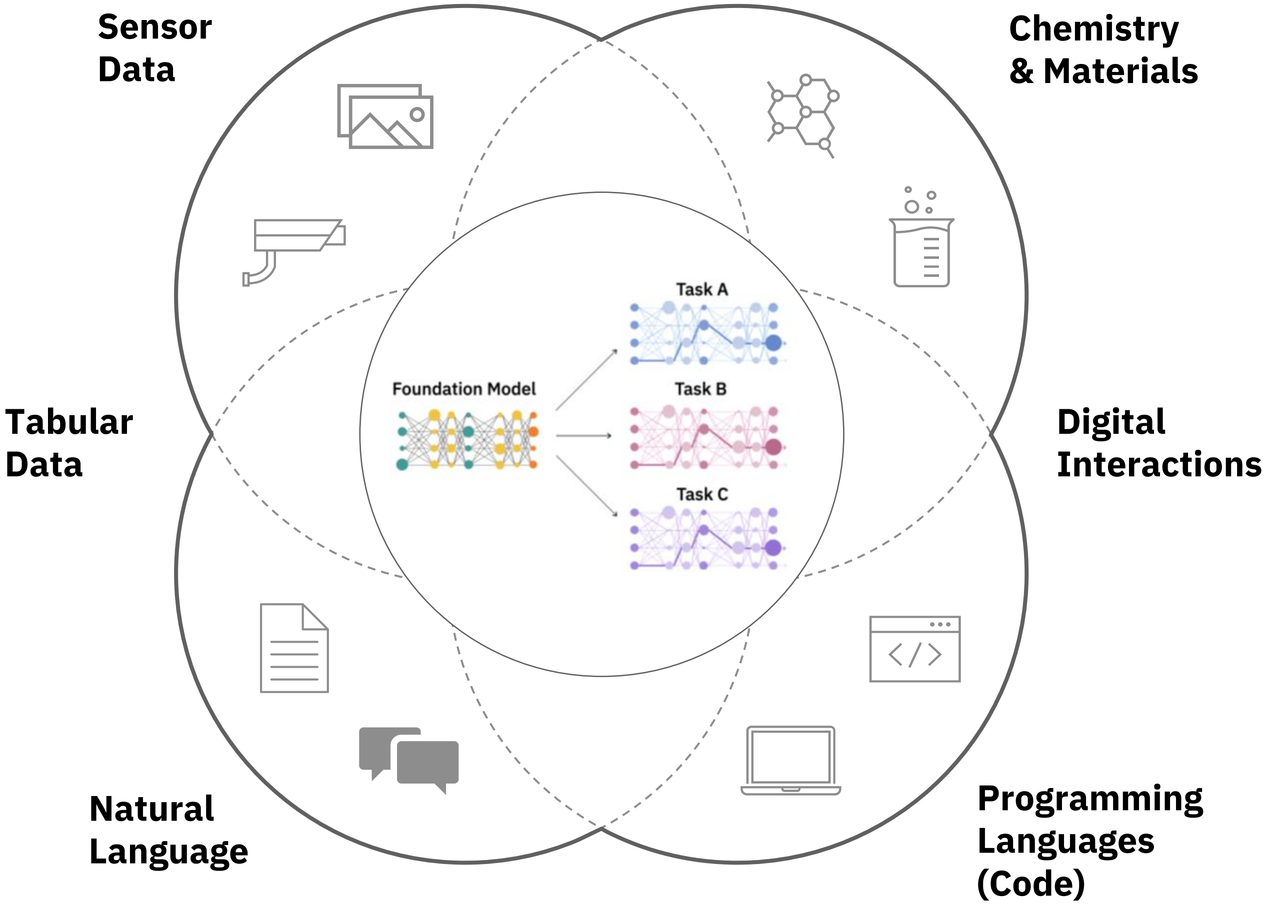
# GPT-3 Model Size Comparison

| Model Name | $n_{params}$ | $n_{layers}$ | $d_{model}$ | $n_{heads}$ | $d_{head}$ | Batch Size | Learning Rate |
|---|---|---|---|---|---|---|---|
| GPT-3 Small | 125M | 12 | 768 | 12 | 64 | 0.5M | $6.0 \times 10^{-4}$ |
| GPT-3 Medium | 350M | 24 | 1024 | 16 | 64 | 0.5M | $3.0 \times 10^{-4}$ |
| GPT-3 Large | 760M | 24 | 1536 | 16 | 96 | 0.5M | $2.5 \times 10^{-4}$ |
| GPT-3 XL | 1.3B | 24 | 2048 | 24 | 128 | 1M | $2.0 \times 10^{-4}$ |
| GPT-3 2.7B | 2.7B | 32 | 2560 | 32 | 80 | 1M | $1.6 \times 10^{-4}$ |
| GPT-3 6.7B | 6.7B | 32 | 4096 | 32 | 128 | 2M | $1.2 \times 10^{-4}$ |
| GPT-3 13B | 13.0B | 40 | 5140 | 40 | 128 | 2M | $1.0 \times 10^{-4}$ |
| GPT-3 175B or "GPT-3" | 175.0B | 96 | 12288 | 96 | 128 | 3.2M | $0.6 \times 10^{-4}$ |

**Table 2.1:** Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

Reference : https://arxiv.org/pdf/2005.14165.pdf

# Opportunities beyond NLP




In many domains, there are large amounts of unlabeled data available in enterprises.

This can used to train foundation models, which can solve business problems that were previously considered intractable.

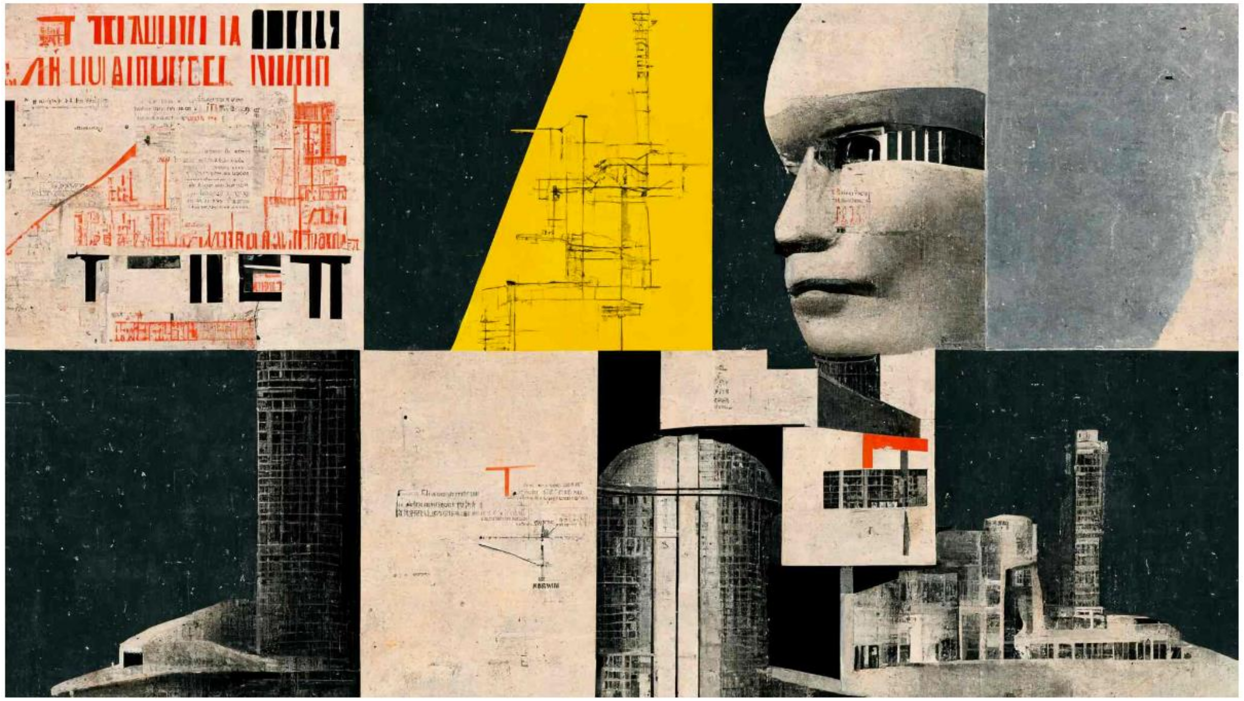
# Once in a decade opportunity

Briefing | The world that Bert built

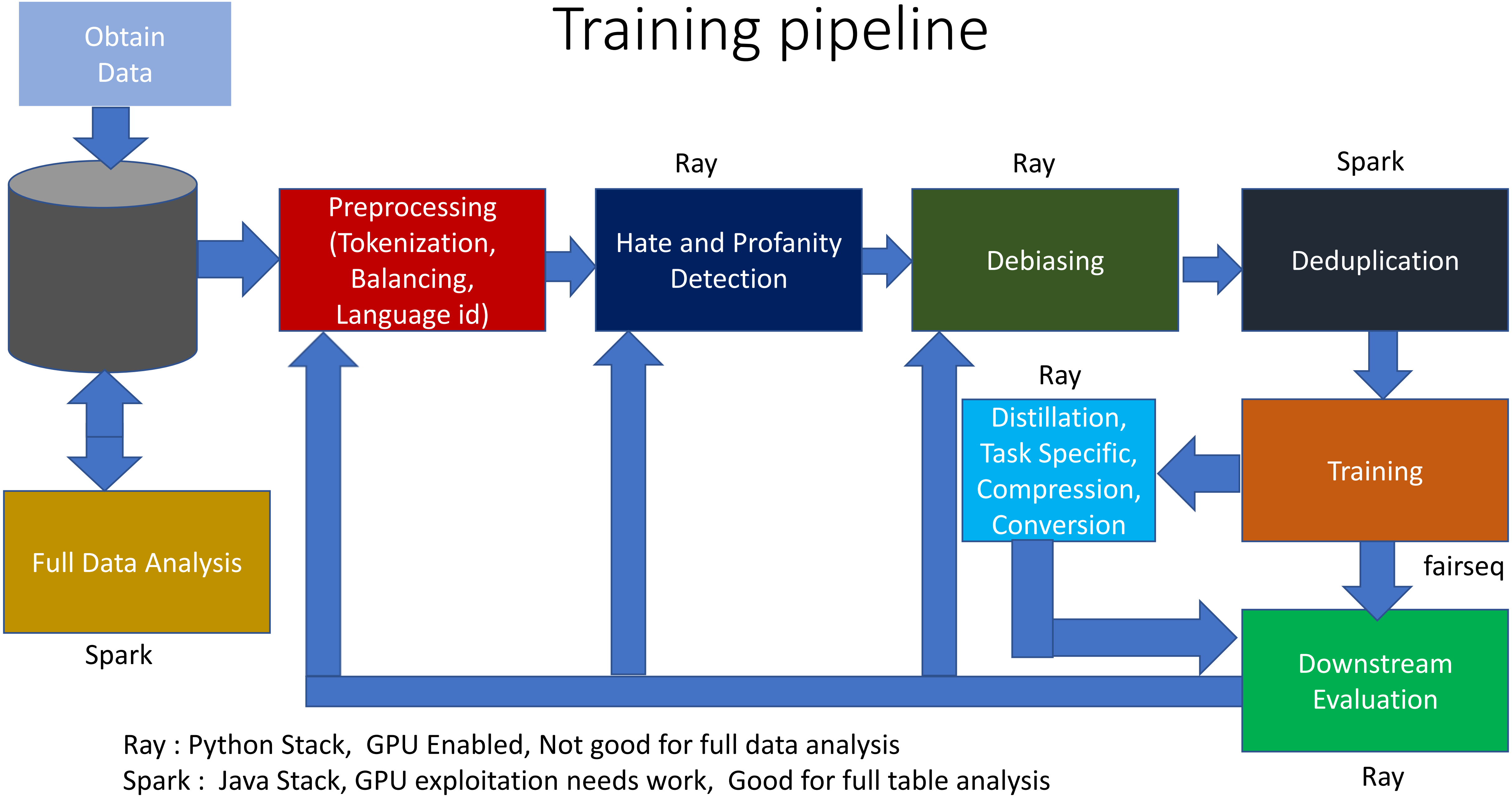
## Huge “foundation models” are turbo-charging AI progress

They can have abilities their creators did not foresee

Jun 11th 2022



# Training pipeline




Ray : Python Stack, GPU Enabled, Not good for full data analysis
Spark : Java Stack, GPU exploitation needs work, Good for full table analysis

# Major Challenges

Large Compute Requirements : 1000+ GPU  Days

Data Quality :  Mixed Languages, Duplication, Noisy Data

Backend Scaling :   GPUs need to be kept busy

Use of Mixed Precision (FP16) :  Faster but flacky

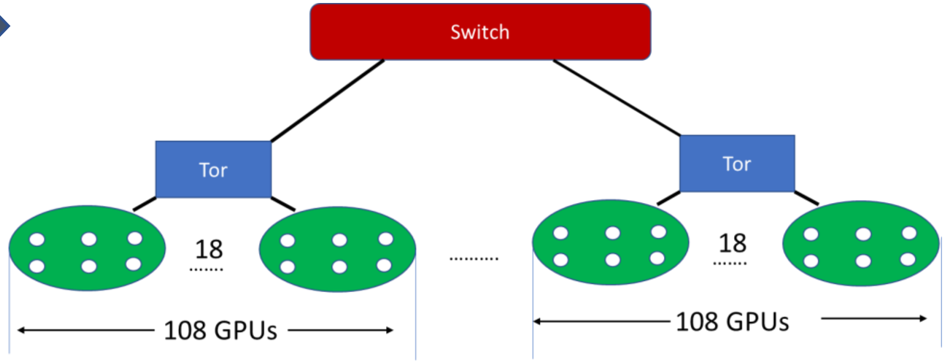
# System Used for Training

Training and Validation Dataset

Fairseq

Pytorch

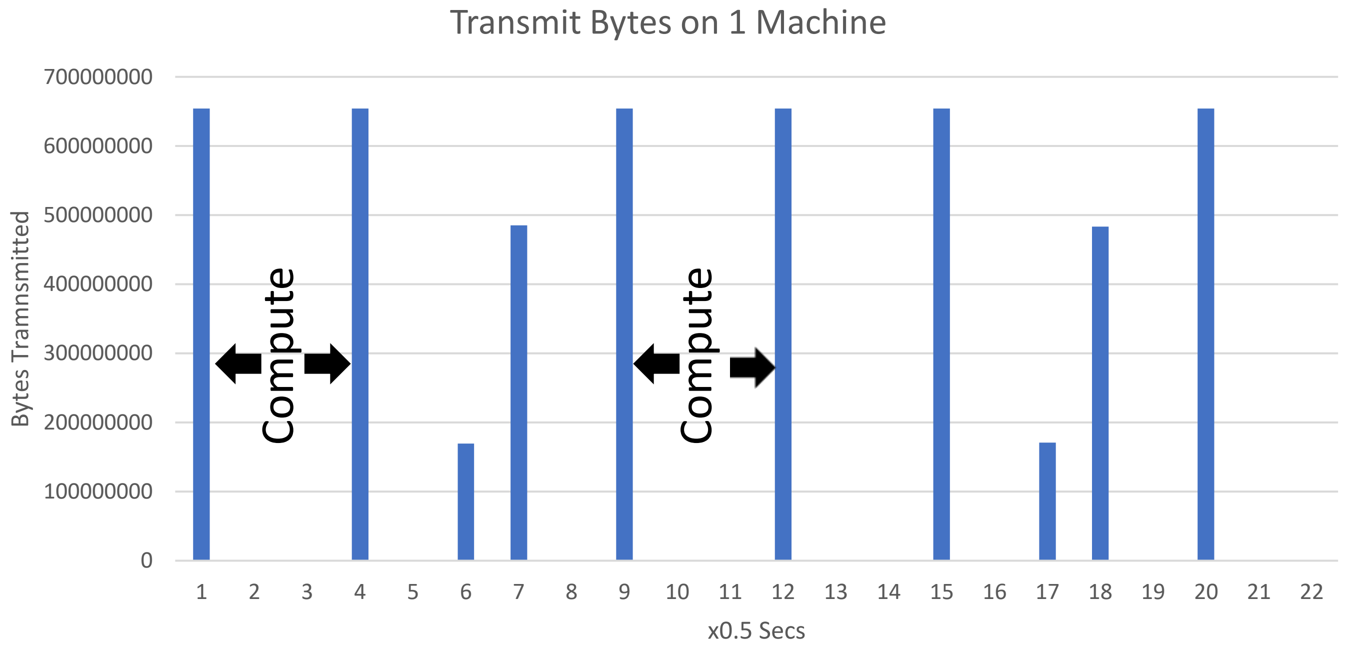
NCCL


GPU Efficiency of 90%
Run time of 5-10 days

**192 V100s across 32 Machines**
**Connected by Dual Infiniband**
**Distributed Data Parallel Mode**
**FP16 Mode, Data on GPFS**

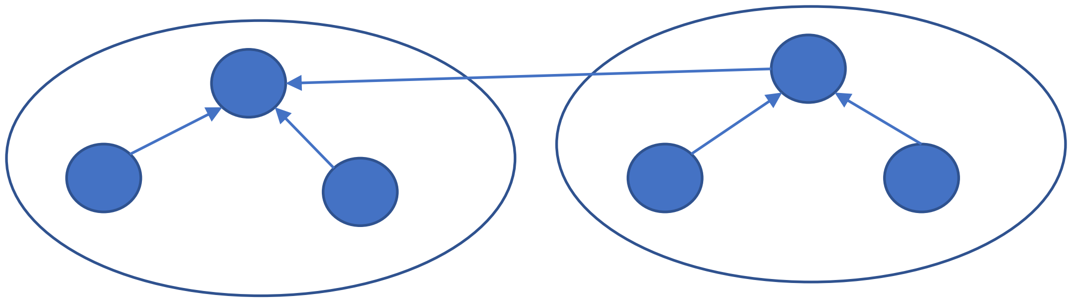
# English RoBERTa 192 GPU Training Experience



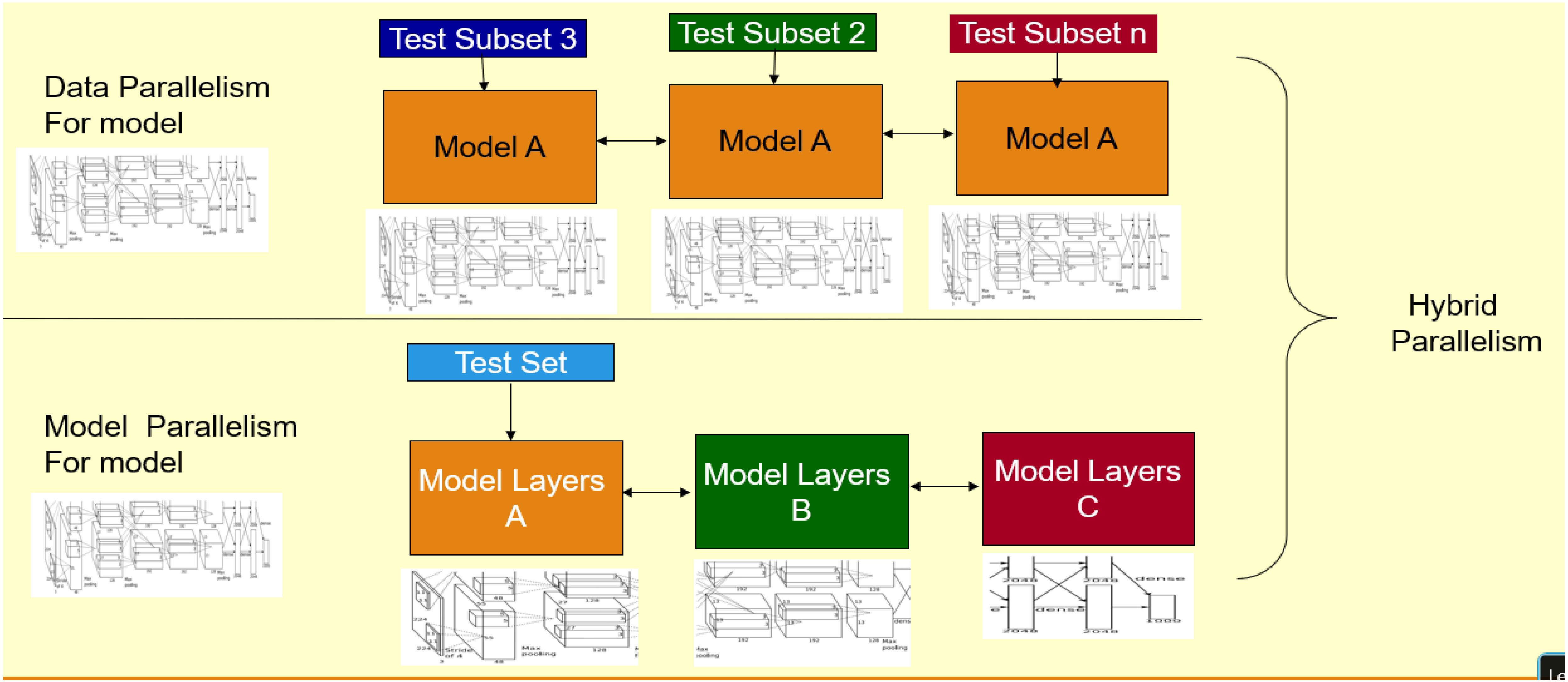

**Each GPU does about 1 Sec Compute**
**Followed by 640 MB of transmit**
**and 640 MB of receive**

**GPU efficiency of about 90%**

Tree Based Communication



# Training Mechanism For Large Models



# Training Compute Comparison

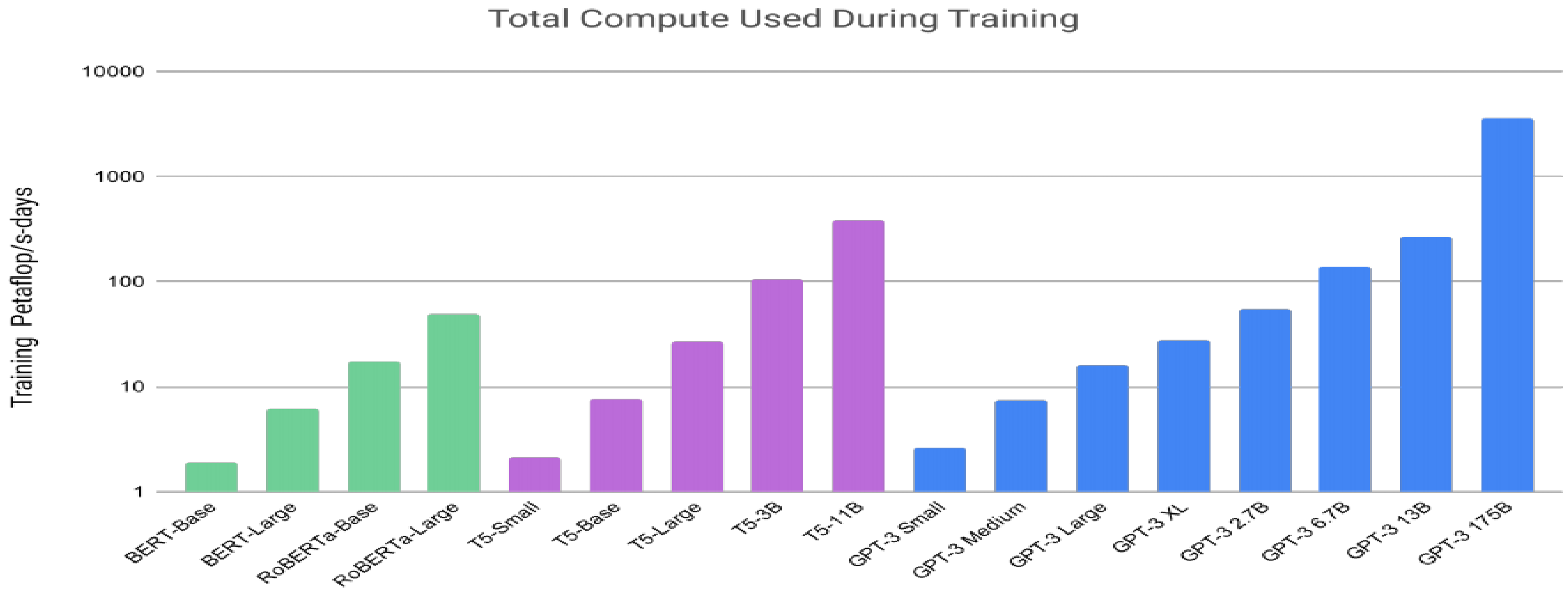


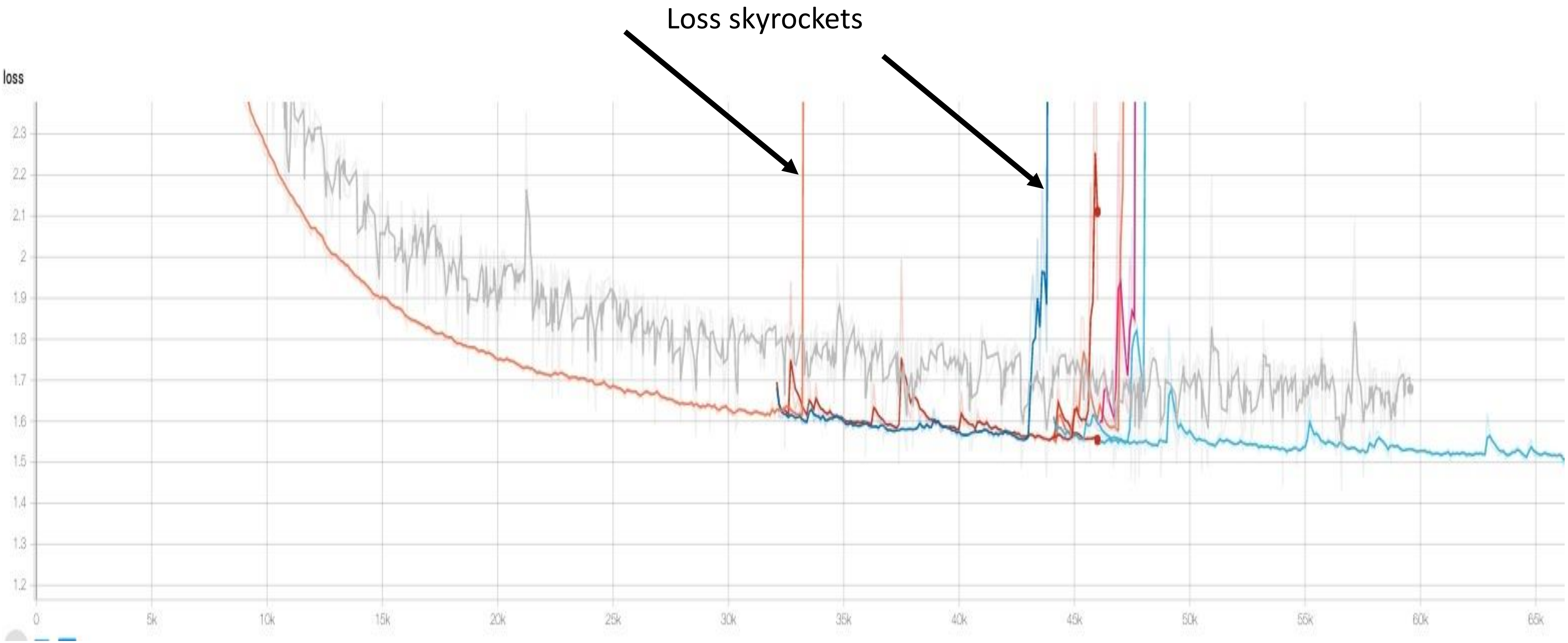

**Figure 2.2: Total compute used during training**. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

| Dataset | Quantity (tokens) | Weight in training mix | Epochs elapsed when training for 300B tokens |
|---|---|---|---|
| Common Crawl (filtered) | 410 billion | 60% | 0.44 |
| WebText2 | 19 billion | 22% | 2.9 |
| Books1 | 12 billion | 8% | 1.9 |
| Books2 | 55 billion | 8% | 0.43 |
| Wikipedia | 3 billion | 3% | 3.4 |

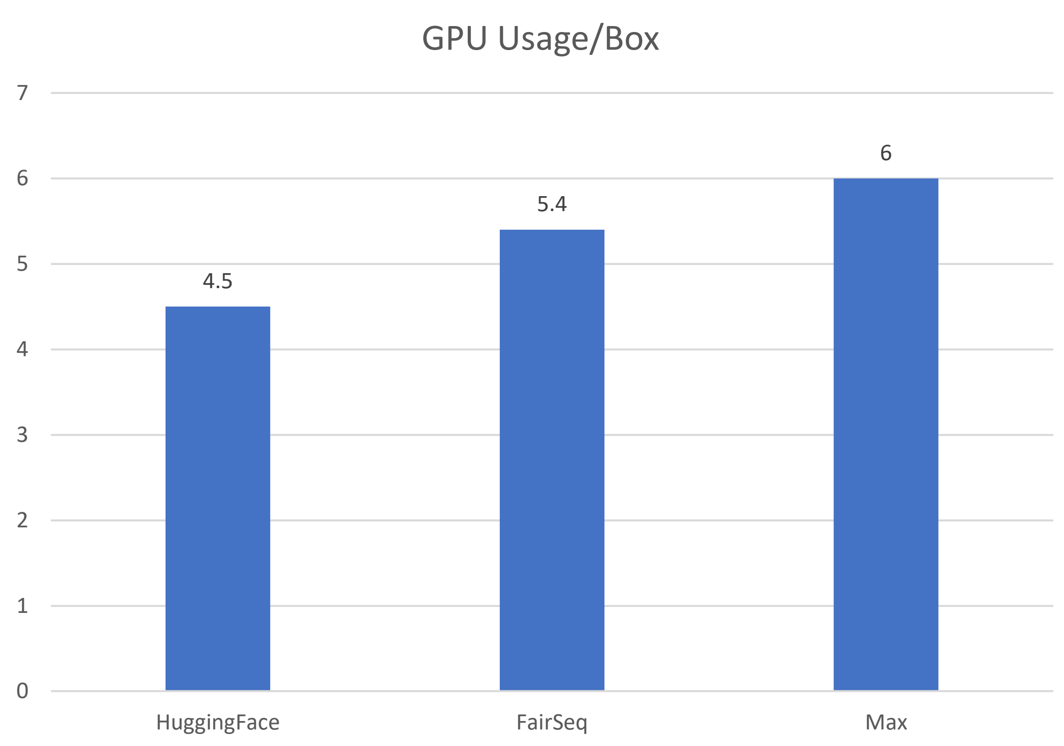
Reference : https://arxiv.org/pdf/2005.14165.pdf

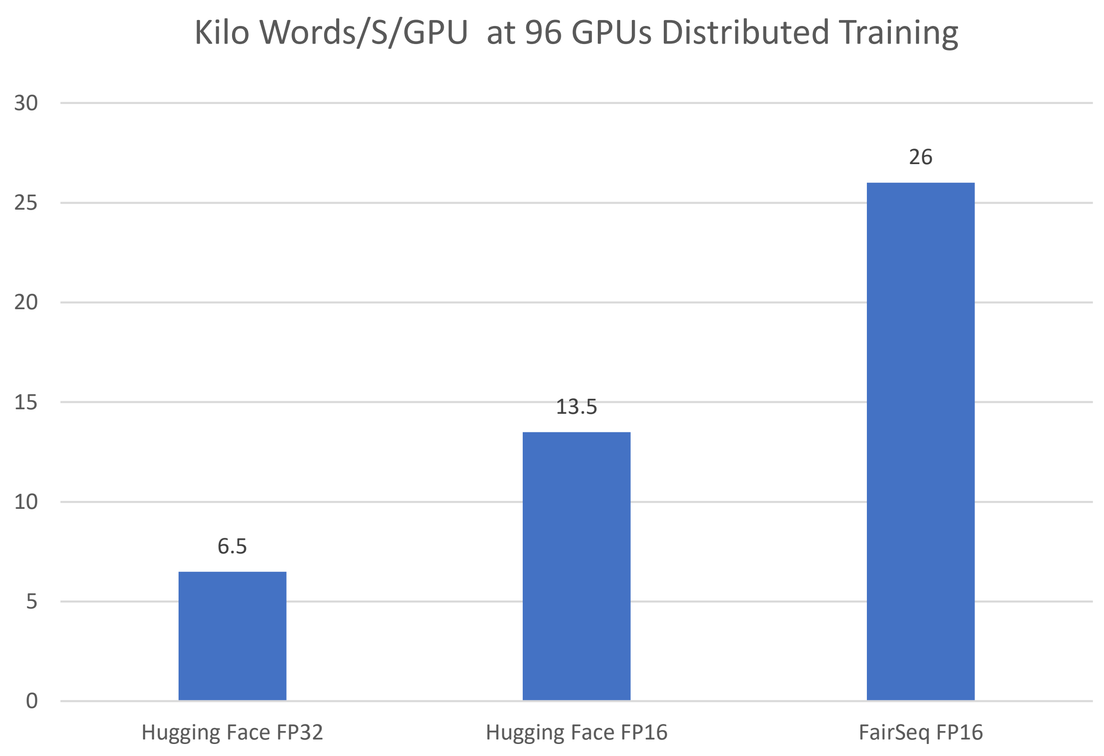
# FP16 Problem on RoBERTa Model Training

Mixed Precision is used… FP16 used for matrix multiplication and possibly softmax

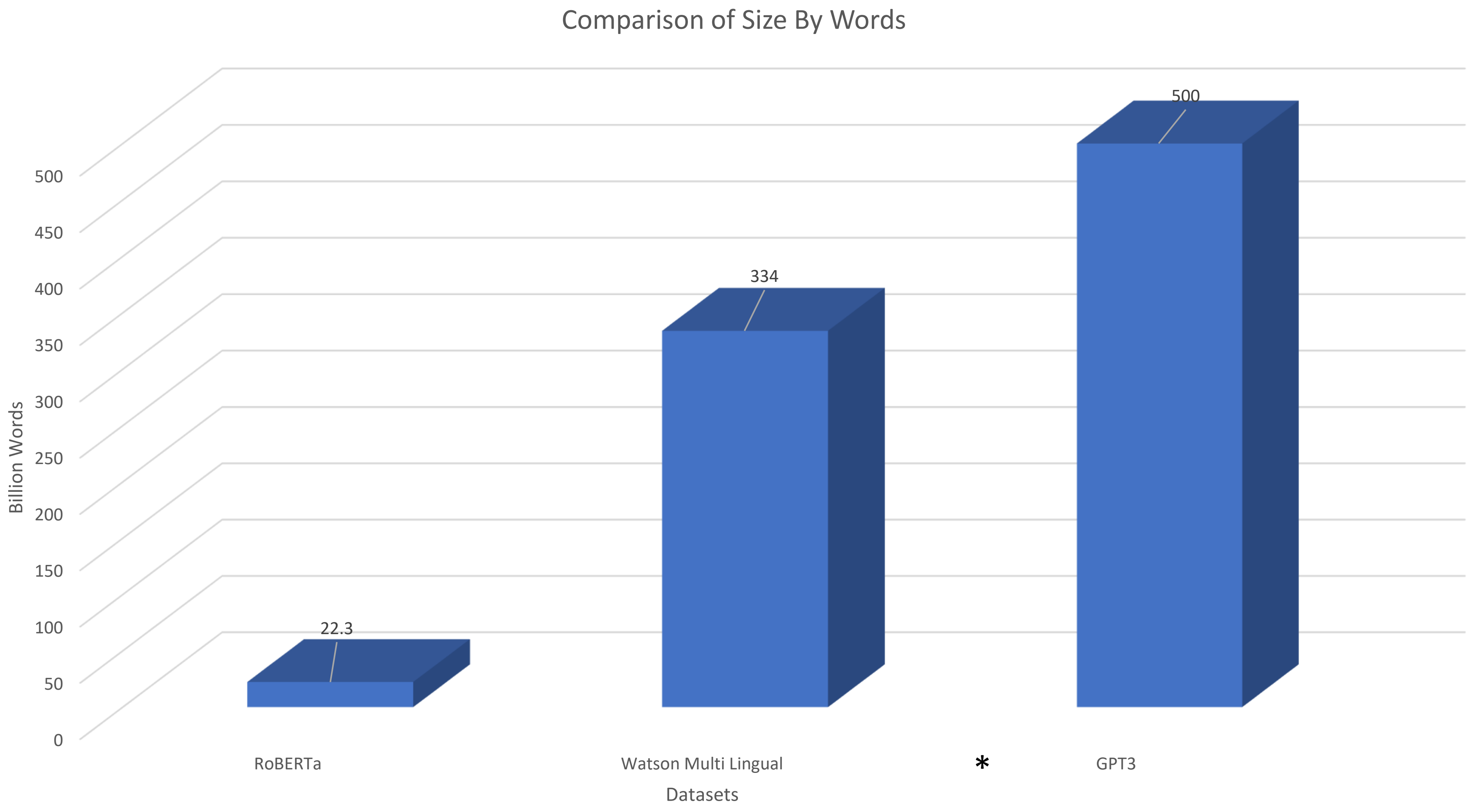


# English RoBERTa 192 GPU Training Experience



# Dataset Sizes

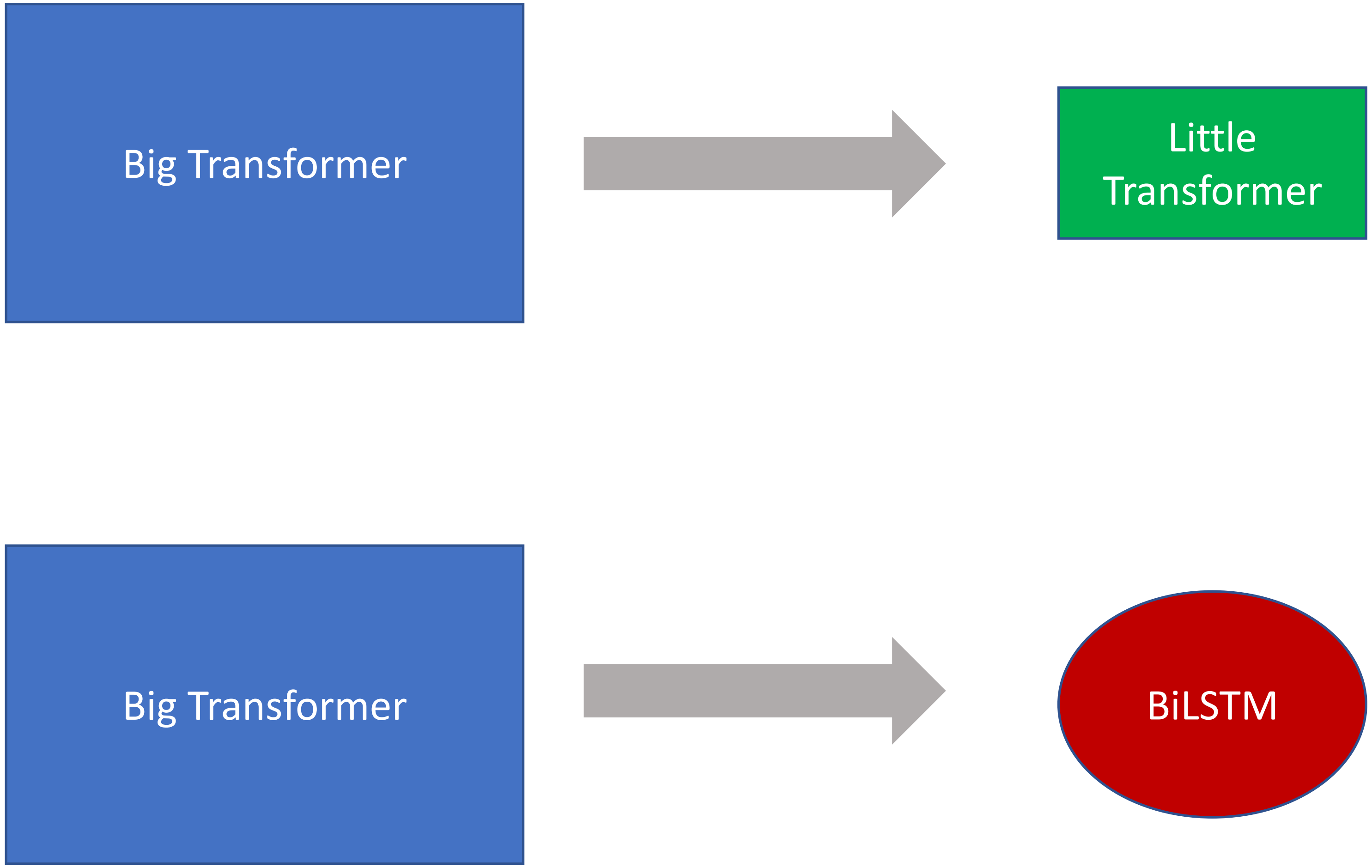


# Alternate for Large Models: Run on CPUs

| | | |
|---|---|---|
| Microsoft Project ADAMS | : | 60 machines for 10 days to train imagenet22K<br>Model stored in main memory<br>Parameter Server based architecture<br>2 billion connections |
| Le at all | : | 1000 machines for 7 days to train imagenet22K<br>Model stored in main memory<br>1 billion connections |
| Rudra | : | CPU based distributed deep learning |
| SLIDE | : | Single V100 GPU vs Cooper Lake vs Cascade Lake<br>V100 does not have TF16 but has FP16<br>A100 has TF32, TF16 and FP16 |

Large training NLP jobs will need > 2000 GPUs for a week. Number of equivalent CPUs ?

# Model Distillation



# Inference

16 Million Transformer : 12 ms on CPU
20 Million Transformer : 20 ms on CPU

a BigScience initiative
BLOOM
176B params · 59 languages · Open-access

6 Secs over wire using 8 A100s of 80 GB

Lower Side

Higher Side

# Summary

Foundational models provide a huge opportunity now

Their training and inference characteristics proved challenges

System design and performance is key to address these challenges