databricks
mosaic research

Learnings from Training Modern LLMs

Sandeep Krishnamurthy
Engineering Director, Databricks MosaicAI Model Training

Our Mission

Help everyone build and serve custom AI models…

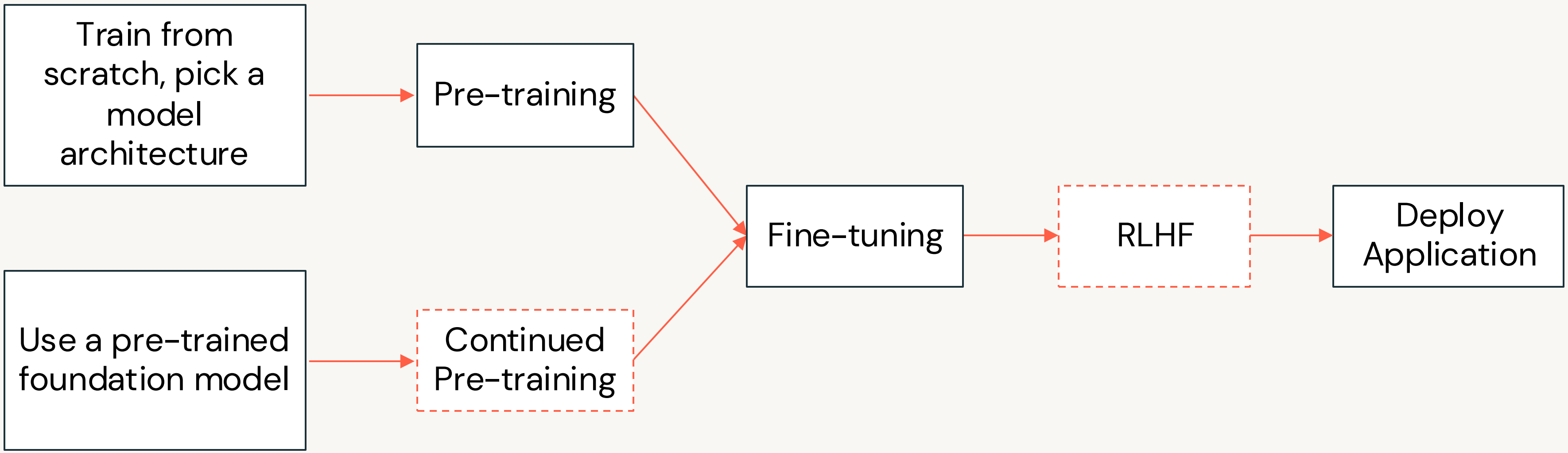
…using their own unique data…

…to achieve the highest quality on their domain…

…as efficiently and cost-effectively as possible.

LLM Training

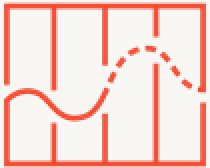
Choose your adventure



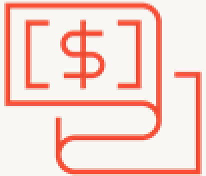
Why build your own generative AI models?

Because organizations value privacy, quality, low cost and low latency

Privacy



Quality



Cost

Latency

DBRX

databricks
mosaic research

What is DBRX & DBRX Training Scale

- An open LLM built entirely at Databricks.
- Data Size: 12T Tokens
- Model Size: 132B params, 36B active params
- Infra/Cluster Size: 3072 H100 GPU (384 nodes)

Why did we build DBRX?

Commercially viable OSS top model.

Help Enterprises with our learnings.

We upgraded our LLM training stack.

Stress testing and improving Databricks for GenAI.

This talk: Our Experience, Learnings, Gotchas and how to build custom DBRX-class models.

databricks mosaic research

Lessons & Gotchas

Learnings in a nutshell...

Start small and work your way up.

Don't trust what you read in the literature.
Test everything for yourself.

Don't trust intuition, received wisdom, or a rumor.
Test everything for yourself.

Do the math.

Roadmap

Define success (evaluation)

Understand your budget (model and data size)

Fill in the details (which model and data)

And then you train... (scaling and infrastructure)

Evaluation

You can't make progress until you know what success looks like.

Evaluation: what you need

Something cheap and automatic.

Something somewhat involved and more realistic.

Something close to the real world.
(Can be slow and expensive.)

databricks
mosaic research

Evaluation: what you need

Something cheap and automatic.

- Your inner development loop.
- Has right and wrong answers.
- For DBRX: the Mosaic Gauntlet

databricks mosaic research

Evaluation: what you need

Calibrating the Mosaic Evaluation Gauntlet

A good benchmark is one that clearly shows which models are better and which are worse. The Databricks Mosaic Research team is dedicated to finding great measurement tools that allow researchers to evaluate experiments. The Mosaic Evaluation Gauntlet is our set of benchmarks for evaluating the quality of models and is composed of 39 publicly available benchmarks split across 6 core competencies: language understanding, reading comprehension, symbolic problem solving, world knowledge, commonsense, and programming. In order to prioritize the metrics that are most useful for research tasks across model scales, we tested the benchmarks using a series of increasingly advanced models.

by Tessa Barton

April 30, 2024 in Mosaic AI Research

Evaluation: what you need

Something somewhat involved and more realistic.

- Evaluates the generative behavior of the model.
- Likely uses LLM-as-a-judge.
- For DBRX: MTBench, IFEval, Arena Hard.

Evaluation: what you need

Something close to the real world.

- Real human evaluation.
- Slots into an existing workflow for A/B testing.
- For DBRX: Human annotation, customer feedback.
- For image models: Human preferences in

Read your evaluation sets and results

Robert McHardy
@robert_mchardy

🚨 Are We Done with MMLU?

In our new paper "Are We Done with MMLU?" we identify errors in MMLU and find that some subsets are riddled with errors. We propose MMLU-Redux with 3,000 re-annotated questions across 30 subjects.

Read your evaluation sets and results

We estimate that somewhere around 70% of GPT-4's "mistakes"
c

Inflection-2.5: meet the world's best personal AI

Palo Alto, CA – March 7, 2024

We also evaluated our models on MT-Bench, a widely used community leaderboard to compare models. However, after evaluating MT-Bench, we realized that a large fraction—nearly 25%—of examples in the reasoning, math, and coding categories had incorrect reference solutions or questions with flawed premises. Therefore, we corrected these examples and release that version of the dataset here.

Internal Slack

databricks mosaic research

Model and Data Size

Understand your budget and constraints. Plan accordingly.

Attempt 1: Training Compute Cost

You have a budget of \$. Train the best model.

The cost of training ≈ model size x data size.

Extreme 1: Train a giant model on very little data.

Extreme 2: Train a tiny model on tons of data.

The answer is somewhere in between. But where?

Attempt 1: Training Compute Cost

The Chinchilla paper. Tokens = 20 x Parameters.

Training Compute-Optimal Large Language Models

Jordan Hoffmann*, Sebastian Borgeaud*, Arthur Mensch*, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals and Laurent Sifre*

*Equal contributions

Attempt 2: Lifecycle Compute Cost

Train the best model and perform **inference**.

Worth training a smaller-than-optimal model to reduce inference cost.

Also has the benefit of simplifying training.

Attempt 2: Lifecycle Compute Cost

Train the best model and perform inference.

LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron*, Thibaut Lavril*, Gautier Izacard*, Xavier Martinet
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave*, Guillaume Lample*

Meta AI

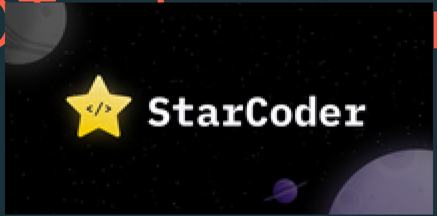
Model	Chinchilla	Llama2-7B	Llama2-70B	Llama3-8B	Llama3-70B
TPR Ratio	20	285	28.5	1875	214.2

Attempt 3: Compute + Data Cost

1. Pretraining	~10T tokens, general data
2. Curriculum Learning	~1T tokens, higher quality
3. Fine-Tuning	~10K-100K instructions
4. RLHF	~10K-100K preferences

Attempt 3: Compute + Data Cost

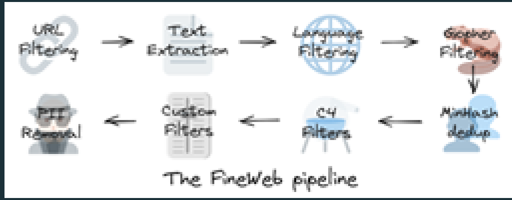
1 | 2 | 3 | 4



Your Data

Semantic Scholar

Wikipedia
The Free Encyclopedia



databricks mosaic research

Attempt 3: Compute + Data Cost

1	2	3	4

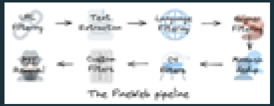
2. Curriculum Learning

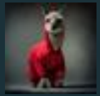
~1T tokens, hi

StarCoder

Semantic Scholar

Wikipedia
The Free Encyclopedia





Your Data

databricks
mosaic research

Attempt 3: Compute + Data Cost

1	2	3	4

3. F … tions

Your Data

Human Annotation O($10-100)

Attempt 3: Compute + Data Cost

1	2	3	4

3. Fine-Tuning ~10K-100K instructions

Quality Is All You Need. Third-party SFT data is available from many different sources, but we found that many of these have insufficient diversity and quality — in particular for aligning LLMs towards dialogue-style instructions. As a result, we focused first on collecting several thousand examples of high-quality SFT data, as illustrated in Table 5. By setting aside millions of examples from third-party datasets and using fewer but higher-quality examples from our own vendor-based annotation efforts, our results notably improved. These findings are similar in spirit to Zhou et al. (2023), which also finds that a limited set of clean instruction-tuning data can be sufficient to reach a high level of quality. We found that SFT annotations in the order of tens of thousands was enough to achieve a high-quality result. We stopped annotating SFT after collecting a total of 27,540 annotations. Note that we do not include any Meta user data.

databricks mosaic research

Llama2 Paper

Attempt 3: Compute + Data Cost

1 2 3 4

3. Fine-Tuning ~10K-100K instructions

We also observed that different annotation platforms and vendors can result in markedly different downstream model performance, highlighting the importance of data checks even when using vendors to source annotations. To validate our data quality, we carefully examined a set of 180 examples, comparing the annotations provided by humans with the samples generated by the model through manual scrutiny. Surprisingly, we found that the outputs sampled from the resulting SFT model were often competitive with SFT data handwritten by human annotators, suggesting that we could reprioritize and devote more annotation effort to preference-based annotation for RLHF.

Attempt 3: Compute + Data Cost

1	2	3	4

3. F

tions

Your Data

Synthetic Data

Human Annotation O($10–100)

Attempt 3: Compute + Data Cost

1 | 2 | 3 | 4

4. R … ences

Your Data

Human Annotation O($5-20)

Attempt 3: Compute + Data Cost

4. RLHF ~10K-100K preferences

Table 26 shows detailed statistics on Meta human preference data. In total, we collected 14 batches of human preference data (i.e., Meta Safety + Helpfulness) on a weekly basis, consisting of over 1 million binary model generation comparisons. In general, later batches contain more samples as we onboard more annotators over time and the annotators also become more familiar with the tasks and thus have better work efficiency. We also intentionally collect more multi-turn samples to increase the complexity of RLHF data and thus the average number of tokens per sample also increase accordingly over batches.

databricks mosaic research

Llama2 Paper

Which Data?

You are what you train on.

Exercise: Build a 1T Token Pretraining Set

Dataset	Size
Web data	2.4T tokens
Code data	400B tokens
Wikipedia English	7B Tokens
Wikipedia Other	47B Tokens
Science papers	60B Tokens
Literature	5B Tokens

Distribute evenly? Upsample certain datasets?

The Original Llama Dataset

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

The Value of Better Data

Arch.	Tokens	Dataset	Gauntlet Score
MPT-7B	1000B	MPT (Apr 2023)	30.9%

The Value of Better Data

Arch.	Tokens	Dataset	Gauntlet Score
MPT-7B	1000B	MPT (Apr 2023)	30.9%
MPT-7B	1000B	DBRX (Jan 2024)	39.0%

Updated dataset leads to 8.1pp improvement.

databricks
mosaic research

The Value of Better Data

Arch.	Tokens	Dataset	Gauntlet Score
MPT-7B	1000B	MPT (Apr 2023)	30.9%
MPT-7B	500B	DBRX (Jan 2024)	32.1%

With a better dataset, we get a
better model with half as much data.

databricks
mosaic research

Key Questions About Data

How should you mix data? Freshness vs. repetition.

Quality vs. Quantity

Should you deduplicate your data?

Run experiments. Let science be your guide.

How to run experiments

Start small and work your way up.

How to run experiments

Start with small models and see how your metrics improve as you scale.

Risk: Your metrics may not have signal until a certain scale.

Must train a 7B model on 2T tokens to get signal on a popular coding benchmark (HumanEval).

Which Model?

Spoiler: It’s going to be a transformer.

Our Advice: Follow the Beaten Path

Train a transformer.

Perform next-token prediction.

Use quadratic attention.

Follow the Llama scaling rules.

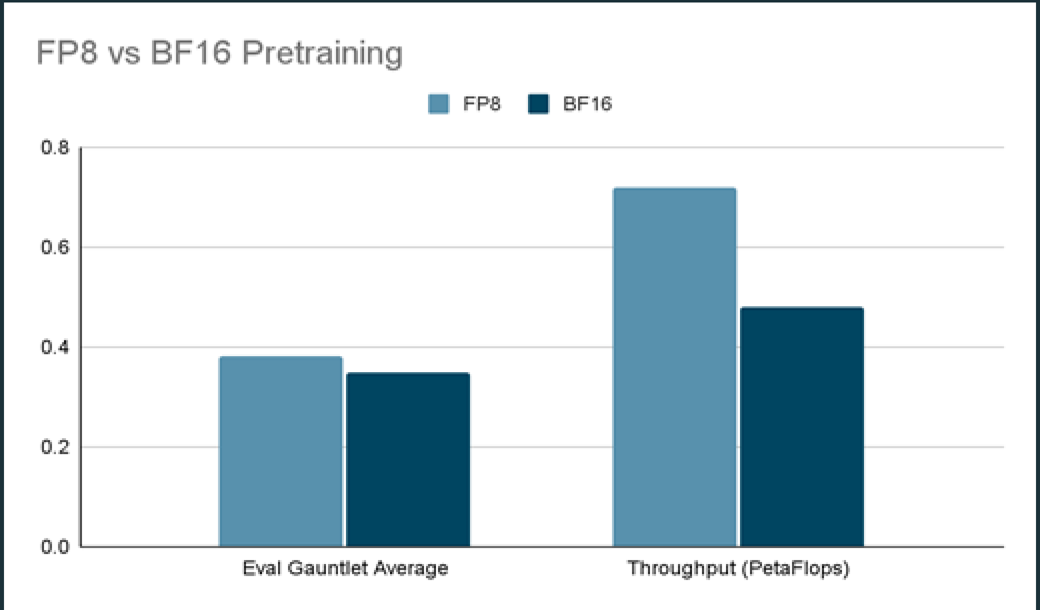
For advanced users: Use RoPE and Adam

Our Stack

FP8 and Mixture-of-Experts (MoE)

- FP8 = precision we use for matrix multiplication
- MoE = model architecture

Training in FP8



FP8 training on H100 is 1.5x faster than BF16 in practice

Mixture-of-Experts: TLDR

Bigger models are better than smaller ones.

Bigger models are slower than smaller ones.

Insight: Use a big model, but only activate a small part of it for any input.

Quality of a bigger model, speed of a smaller one.

databricks
mosaic research

The Value of Mixture-of-Experts

Arch.	Active Params	Relative FLOPs	Gauntlet Score
Llama2-13B	13B	1.7x	43.8%

databricks mosaic research

The Value of Mixture-of-Experts

Arch.	Active Params	Relative FLOPs	Gauntlet Score
Llama2-13B	13B	1.7x	43.8%
DBRX Small	6.6B	1x	45.5%

The Value of Mixture-of-Experts

Arch.	Active Params	Relative FLOPs	Gauntlet Score
Llama2-13B	13B	1.7x	43.8%
DBRX Small	6.6B	1x	45.5%

Our mixture-of-experts training recipe scored higher, used 1.7x less training compute, and behaves like a model 2x smaller at inference time.

databricks
mosaic research

Do the math

How long will it take to train?

How long will it take to train my model?

How Long to Train

7B Param Model

Chinchilla Tokens

64 A100s

Cheatsheet

FLOPs = 6 x Parameters x Tokens

Tokens = 20 x Parameters (Chinchilla)

A100 = 312 TFLOP/sec = 3.12e14 FLOP/sec

Data = 20 x 7e9 = 1.8e11

FLOPs = 6 x 7e9 x 1.8e11 = 5.88e21

Cluster FLOP/sec = 3.12e14 x 64 = 2e16

Time = FLOPs / Cluster FLOP/sec = 5.88e21 / 2e16 = 3.4 days

databricks mosaic research

How long will it take to train my model?

How Long to Train
7B Param Model
Chinchilla Tokens
64 A100s

Cheatsheet
FLOPs = 6 x Parameters x Tokens
Tokens = 20 x Parameters (Chinchilla)
A100 = 312 TFLOP/sec = 3.12e14 FLOP/sec

There's something missing here!

How long will it take to train my model?

How Long to Train

7B Param Model

Chinchilla Tokens

64 A100s

Cheatsheet

FLOPs = 6 x Parameters x Tokens

Tokens = 20 x Parameters (Chinchilla)

A100 = 312 TFLOP/sec = 3.12e14 FLOP/sec

There's something missing here!

How long will it take to train my model?

How Long to Train
7B Param Model
Chinchilla Tokens
64 A100s

Cheatsheet
FLOPs = 6 x Parameters x Tokens
Tokens = 20 x Parameters (Chinchilla)
A100 = 312 TFLOP/sec = 3.12e14 FLOP/sec

You won't fully utilize your GPU.
There are other bottlenecks in the system.
This is the theoretical peak. You will get power limited.

How long will it take to train my model?

How Long to Train
7B Param Model
Chinchilla Tokens
64 A100s

Cheatsheet
FLOPs = 6 x Parameters x Tokens
Tokens = 20 x Parameters (Chinchilla)
A100 = 312 TFLOP/sec = 3.12e14 FLOP/sec

MFU = Model Flop Utilization

What fraction of the peak GPU FLOP/sec is your model getting?

Only counts 6 x Parameters x Tokens, not recomputation.

For this configuration, **50.7% MFU.**

How long will it take to train my model?

How Long to Train
7B Param Model
Chinchilla Tokens
64 A100s

Cheatsheet
FLOPs = 6 x Parameters x Tokens
Tokens = 20 x Parameters (Chinchilla)
A100 = 312 TFLOP/sec = 3.12e14 FLOP/sec

Data = 20 x 7e9 = 1.8e11
FLOPs = 6 x 7e9 x 1.8e11 = 5.88e21
Cluster FLOP/sec = 3.12e14 x 64 x 50.7% = 1.01e16
Time = FLOPs / Cluster FLOP/sec = 5.88e21 / 1.01e16 = 6.7 days

databricks mosaic research

Nuts and Bolts / Infrastructure

The technologies we built on.

Tool Stack

Lilac AI for data exploration and curation

Notebooks and Apache Spark for data cleaning and processing

Unity Catalog for data storage and governance

Mosaic AI Pretraining to train the model

MLflow and Lakeview for experiment tracking

MosaicAI Model Training Engines

Composer. Training library built for scalability.

Streaming. Stream efficiently from object stores.

LLM Foundry. Highly efficient and scalable training and fine-tuning code for popular LLMs.

MegaBlocks. Mixture-of-Experts implementation.

MosaicAI Model Training Engines – OSS

Composer. github.com/mosaicml/composer

Streaming. github.com/mosaicml/streaming

LLM Foundry. github.com/mosaicml/llm-foundry

MegaBlocks. github.com/databricks/megablocks

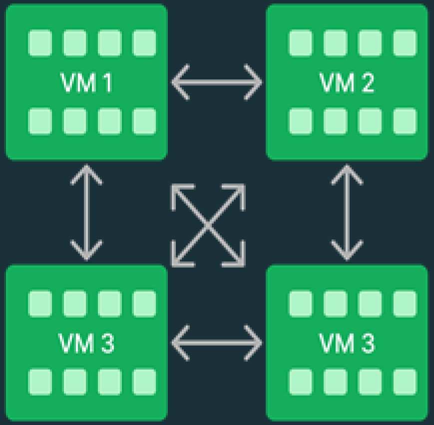
databricks mosaic research

Problem: Hardware Failures

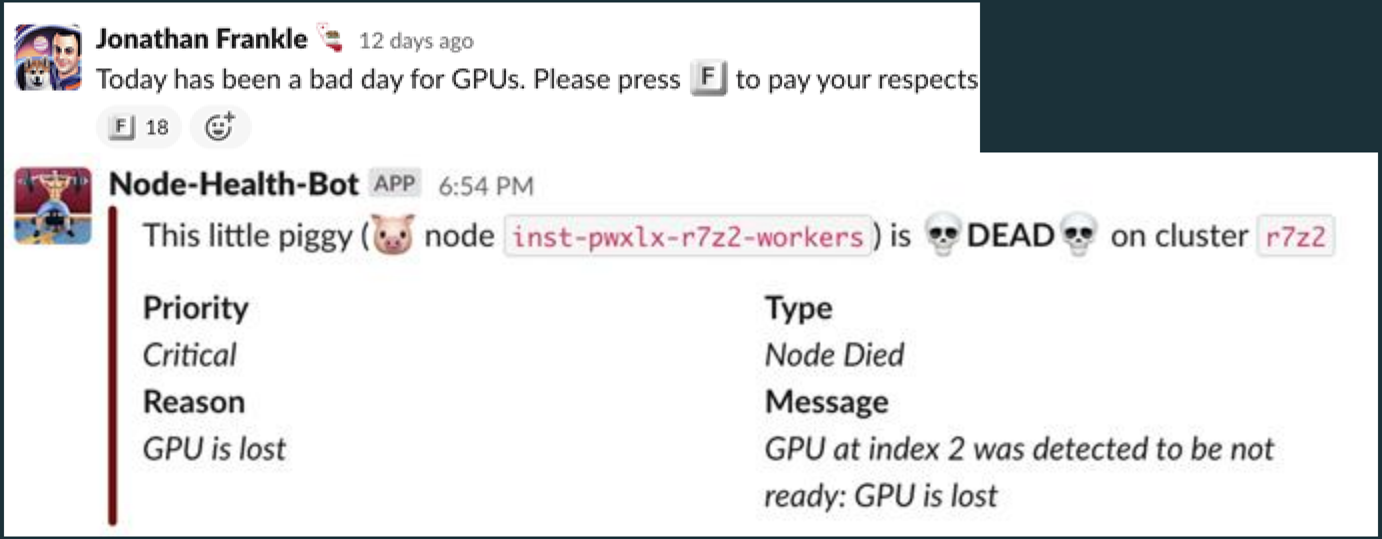
- Hardware/Software or both. Something Fails. Roughly once every 1000 H100-days (~3 failure / day / 3072 GPU cluster)
- GPUs, Switches, Communication Libraries (NCCL)

Normally: O(N) things can go wrong

Training : O(N^2) things can go wrong



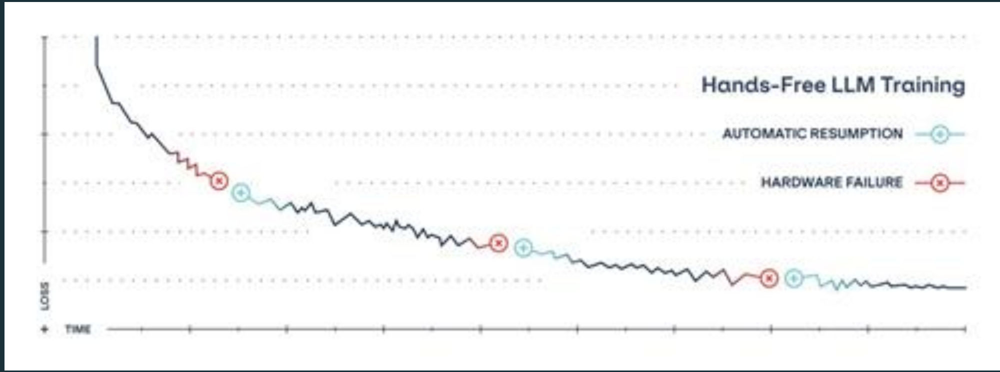
Problem: Hardware Failures



Our solution: Automatic failure detection and blazingly fast job restart times.

Platform is key for speed of development

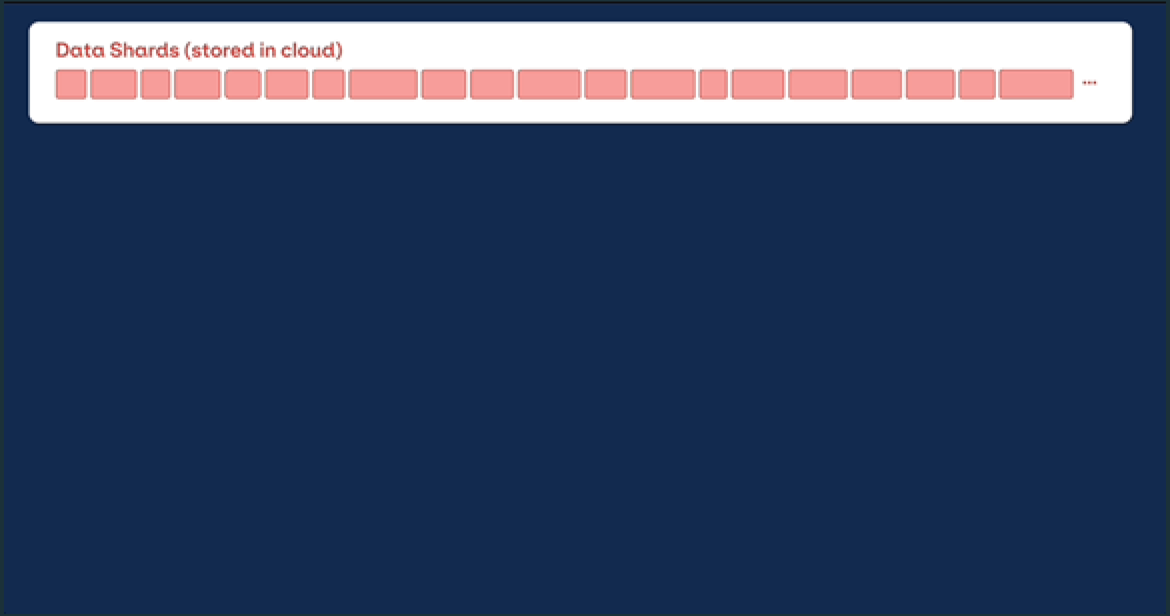
- Pick a platform that handles the “undifferentiated stuff”
  - Software packages and dependencies
  - Scheduling and orchestration
  - Model checkpointing
  - Fault tolerance, detection and monitoring
  - Automatic fault recovery



databricks mosaic research

Reliable Infrastructure => No long lived clusters!

- No Long Lived Clusters means:
  - Streaming data from blob storage
  - Ephemeral storage only – *no persistent volumes!*
  - Async checkpoint upload to blob storage



databricks mosaic research

Recap

Training Large Language Models is:

Evaluation Strategy

+

Data Strategy

+

Science (Model)

+

Systems and Reliable Infrastructure

+

Process and Discipline

databricks
mosaic research

Appendix / Additional References

MosaicML Foundation Series

Busting Cost Myths

Model	Number of Tokens of Data	System	Time-to-Train with MosaicML	Cost with MosaicML
MPT-7B	1T	440xA100-40GB	9.5 Days	**$250,800**
MPT-7B-Instruct	9.6M	8xA100-40GB	2.3 Hours	**$46**
MPT-7B-Chat	86M	8xA100-80GB	8.2 Hours	$205
		+ 32xA100-40GB	6.7 Hours	$536
		Total Combined →	14.9 Hours	**$741**
MPT-7B-StoryWriter-65k+	5B	32xA100-80GB	2.2 Days	**$5338**

MosaicML Foundation Series

Busting Cost Myths

Pre-training

Hardware	Precision	Model	Tokens	Time to Train with mosaicML	Cost to Train with mosaicML
512xA100-40GB	AMP_BF16	MPT-30B	1 Trillion	28.3 Days	~ $871,000
512xH100-80GB	AMP_BF16	MPT-30B	1 Trillion	11.6 Days	~ $714,000

Fine-tuning

Hardware	Precision	Model	Time to Finetune on 1B tokens with MosaicML	Cost to Finetune on 1B tokens with MosaicML
16xA100-40GB	AMP_BF16	MPT-30B	21.8 Hours	$871
16xH100-80GB	AMP_BF16	MPT-30B	8.9 Hours	$714