

LLMs on Polaris

Sam Foreman 

foremans@anl.gov

Argonne National Laboratory

2024-07-17

LLMs on Polaris

 Sam Foreman
SciFM Summer School 24



Sam Foreman

- I'm a Computational Scientist in the [Data Science Group](#) at [ALCF](#)¹.
 - Personal Website: samforeman.me
 - Background: {ML, LLMs, AI4Science, HEP, Lattice QCD, MCMC, Generative Modeling, ...}

Ongoing / recent work:

- [AI + Science](#)
 - [Building better sampling methods for Lattice QCD](#)
 - [GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics](#)
 - [Foundation models for long term climate forecasting](#)
- [Scaling Large Language Models](#)
- [Optimizing distributed training across thousands of GPUs](#)
- Building new parallelism techniques for efficient scaling
- Generative modeling (esp. for physical systems)

Polaris @ ALCF

Refer to [Getting Started](#) for additional information.

- Login:

```
ssh <username>@polaris.alcf.anl.gov
```

- Modules (+ using `conda`):

```
module use /soft/modulefiles  
module load conda
```

Getting Started

- [Running Jobs](#)
 - [example job scripts](#)
- [Proxy:](#)

```
# proxy settings
export HTTP_PROXY="http://proxy.alcf.anl.gov:3128"
export HTTPS_PROXY="http://proxy.alcf.anl.gov:3128"
export http_proxy="http://proxy.alcf.anl.gov:3128"
export https_proxy="http://proxy.alcf.anl.gov:3128"
export ftp_proxy="http://proxy.alcf.anl.gov:3128"
export no_proxy="admin,polaris-adminvm-01,localhost,*.cm.polaris.alcf
```

- Getting Help:
support@alcf.anl.gov

Polaris

- Polaris is a 560 node HPE Apollo 6500 Gen 10+ based system.
- Each node has a single 2.8 GHz AMD EPYC Milan 7543P 32 core CPU with:
 - 512 GB of DDR4 RAM
 - 4 (four) NVIDIA A100 GPUs connected via NVLink
 - 2 (a pair) of local 1.6TB of SSDs in RAID0 for the users use
 - 2 (a pair) of Slingshot 11 network adapters.
- There are two nodes per chassis, seven chassis per rack, and 40 racks for a total of 560 nodes.

Polaris Compute Nodes

Details

POLARIS COMPUTE	DESCRIPTION	PER NODE	AGGREGATE
Processor ¹	2.8 GHz 7543P	1	560
Cores/Threads	AMD Zen 3 (Milan)	32/64	17,920/35,840
RAM ²	DDR4	512 GiB	280 TiB
GPUS	NVIDIA A100	4	2240
Local SSD	1.6 TB	2/3.2 TB	1120/1.8PB

1. 256MB shared L3 cache, 512KB L2 cache per core, 32 KB L1 cache per core

2. 8 memory channels rated at 204.8 GiB/s

Polaris A100 GPU Information

DESCRIPTION	A100 PCIe	A100 HGX (Polaris)
GPU Memory	40 GiB HBM2	160 GiB HBM2
GPU Memory BW	1.6 TB/s	6.4 TB/s
Interconnect	PCIe Gen4 64 GB/s	NVLink 600 GB/s
FP 64	9.7 TF	38.8 TF
FP64 Tensor Core	19.5 TF	78 TF
FP 32	19.5 TF	78 TF
BF16 Tensor Core	312 TF	1.3 PF
FP16 Tensor Core	312 TF	1.3 PF
INT8 Tensor Core	624 TOPS	2496 TOPS
Max TDP Power	250 W	400 W

Using Conda

- Additional information in our [user guide](#)
- We provide prebuilt `conda` environment containing GPU-supported builds of:
 - [Pytorch - ALCF User Guides](#)
 - [DeepSped - ALCF User Guides](#)
 - [JAX - ALCF User Guides](#)
 - [Tensorflow - ALCF User Guides](#)
- To activate / use: (either from an interactive job or inside a job script):

```
$ module use /soft/modulefiles  
$ module load conda; conda activate base
```

Virtual Environments: `venv`

- To install additional libraries, we can create a virtual environment using `venv`
- Make sure you're currently inside the **base** `conda` environment:
 - `module load conda; conda activate base`
- Now, create `venv` **on top of** `base`:

```
$ python3 -m venv /path/to/venv --system-site-packages
$ source /path/to/venv/bin/activate
$ which python3
/path/to/venv/bin/python3
$ # Now you can `python3 -m pip install ...` etc
```

Warning

1. `--system-site-packages` tells the `venv` to use system packages
2. You must replace the path `/path/to/venv` in the above commands with a suitably chosen directory which you are able to write to.

Note about `venv`'s

- The value of `--system-site-packages` can be changed by modifying its value in `/path/to/venv/pyvenv.cfg`
- To install a **different** version of a package that is already installed in the base environment:

```
$ python3 -m pip install --ignore-installed ... # or -I
```

- The shared `base` environment is not writable
 - Impossible to remove or uninstall packages
- If you need additional flexibility, we can **clone** the base environment

Clone base **conda** environment

- If we need additional flexibility or to install packages which **require** a **conda** install, we can clone the base environment
 - requires copying the entirety of the base environment
 - **large storage requirement**, can get out of hand quickly
- The shared **base** environment is not writable
 - Impossible to remove or uninstall packages
- This can be done by:

```
$ module load conda
$ conda activate base
(base) $ conda create --clone base --prefix="/path/to/envs/base-clone"
```

Containers on Polaris

- Polaris uses Nvidia A100 GPUs →
 - We can take advantage of Nvidia optimized containers
- The container system on Polaris is [singularity](#):

```
module avail singularity # see available
module load singularity # load default version
# To load a specific version:
module load singularity/3.8.7
```

- Singularity: two options for creating containers:
 1. Using Docker on local machine and publishing to DockerHub
 2. Using a Singularity recipe file and building on a Polaris worker node
- See also: [Containers - ALCF User Guides](#)

Large Language Models

Status of Large Language Models

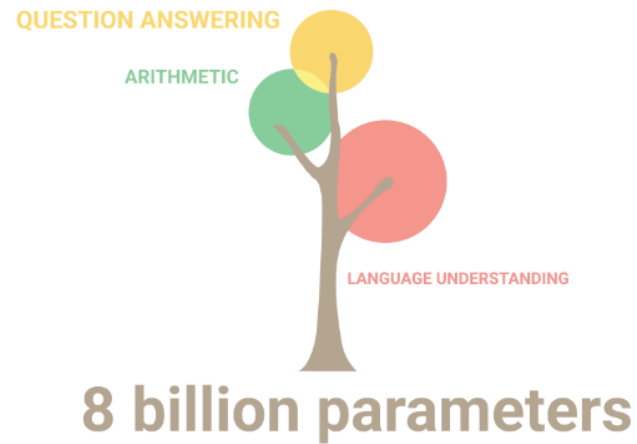
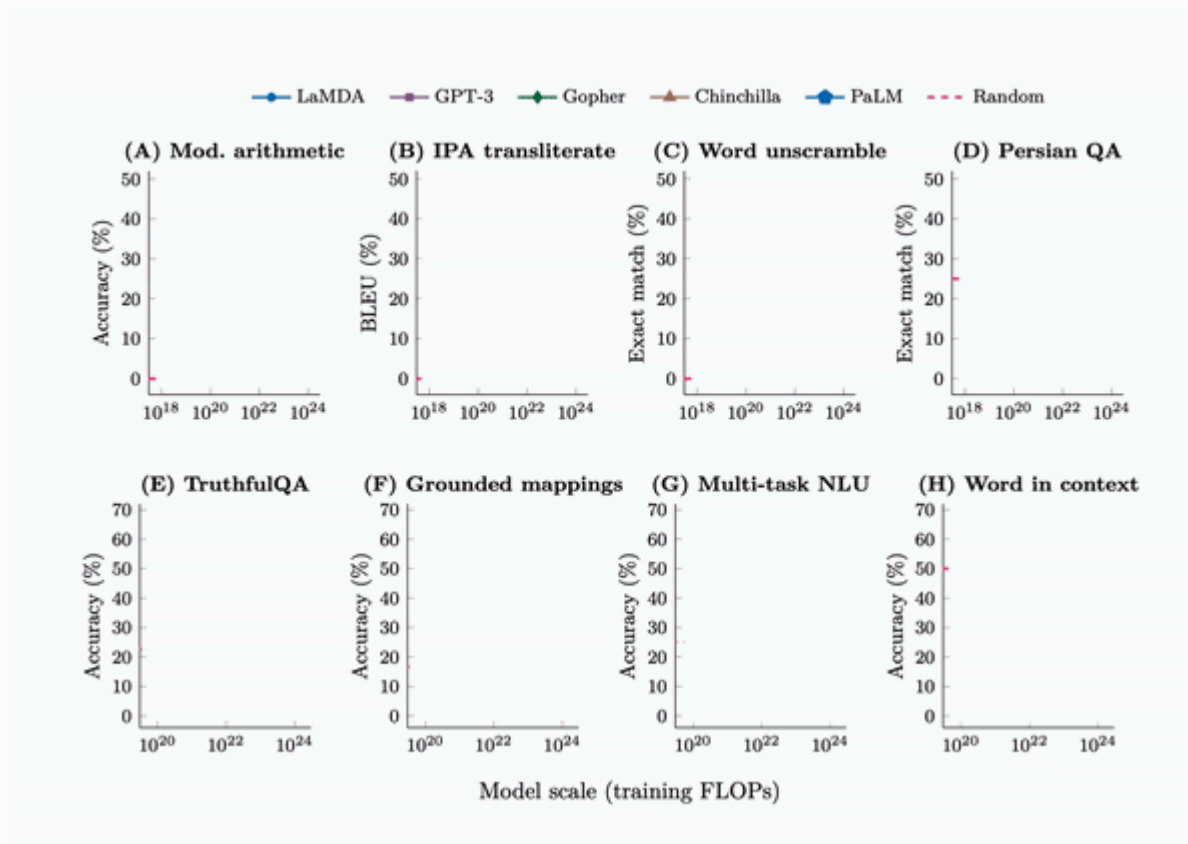


Figure 1: Large Language Models have (LLM)s have taken the NLP community **world** by storm¹

Emergent Abilities



[Emergent abilities of Large Language Models](#) Yao et al. ([2023](#))

Training LLMs

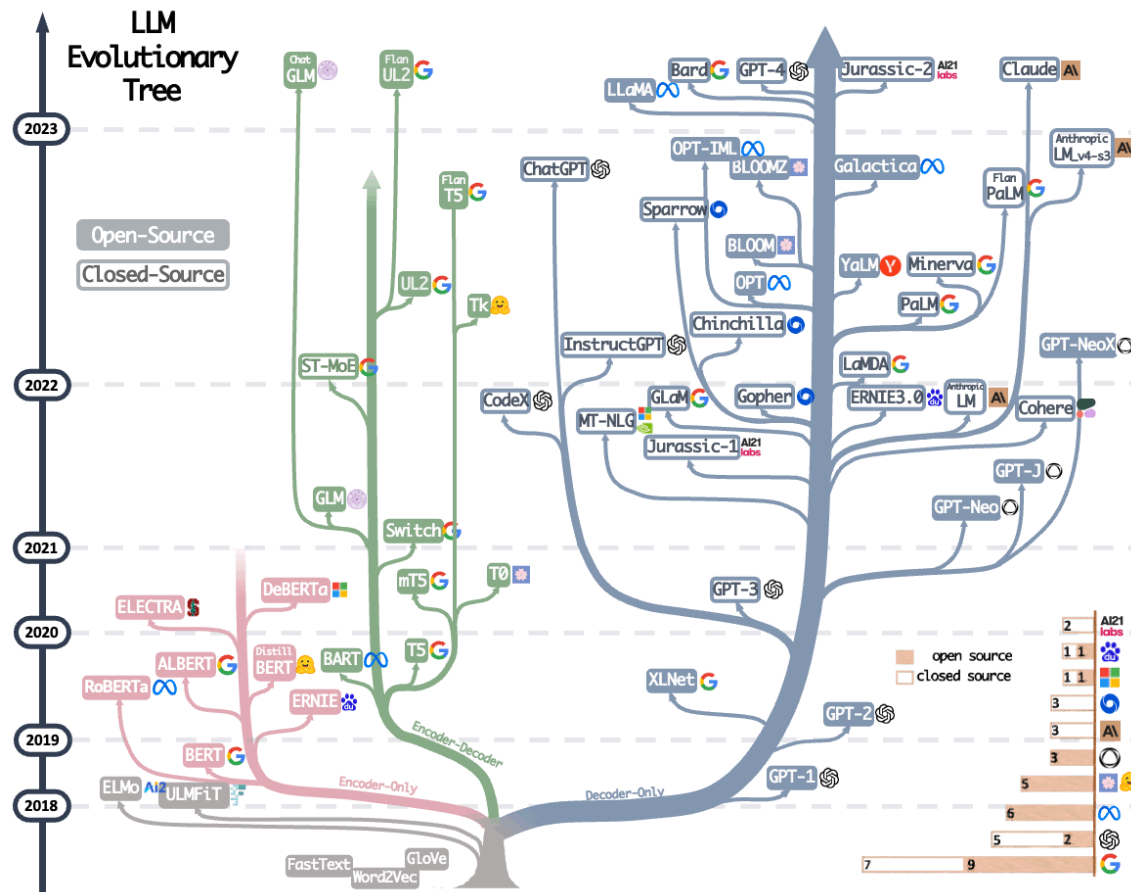
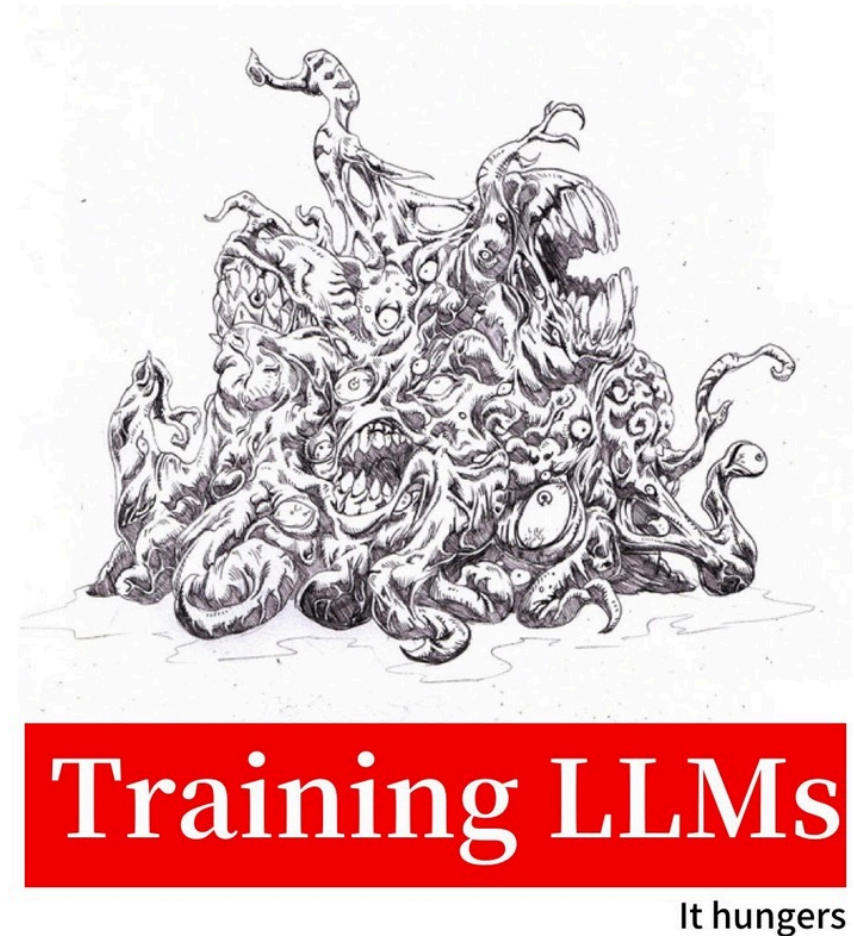


Figure 2: Visualization from Yang et al. (2023)

May God forgive us for what we have done



O'RLY?

Lovecraft


















Recent Work (2017 – Now)


















18








Papers, 2017–*

Date	Paper	keywords	Institute	Publication
06/2017	Attention Is All You Need	Transformers	Google	NeurIPS citation 97937
06/2018	Improving Language Understanding by Generative Pre-Training	GPT 1.0	OpenAI	citation 9124
10/2018	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	BERT	Google	NAACL citation 77806
02/2019	Language Models are Unsupervised Multitask Learners	GPT 2.0	OpenAI	citation 17283
09/2019	Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism	Megatron-LM	NVIDIA	citation 1341
10/2019	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	T5	Google	JMLR citation 14902
10/2019	ZeRO: Memory Optimizations Toward Training Trillion Parameter Models	ZeRO	Microsoft	SC citation 212
01/2020	Scaling Laws for Neural Language Models	Scaling Law	OpenAI	citation 2716

Date	Paper	keywords	Institute	Publication
05/2020	Language models are few-shot learners	GPT 3.0	OpenAI	NeurIPS citation 28070
01/2021	Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity	Switch Transformers	Google	JMLR citation 1411
08/2021	Evaluating Large Language Models Trained on Code	Codex	OpenAI	citation 3125
08/2021	On the Opportunities and Risks of Foundation Models	Foundation Models	Stanford	citation 2930
09/2021	Finetuned Language Models are Zero-Shot Learners	FLAN	Google	ICLR citation 2530
10/2021	Multitask Prompted Training Enables Zero-Shot Task Generalization	T0	HuggingFace et al.	ICLR citation 1389
12/2021	GLaM: Efficient Scaling of Language Models with Mixture-of-Experts	GLaM	Google	ICML citation 505
12/2021	WebGPT: Browser-assisted question-answering with human feedback	WebGPT	OpenAI	citation 820
12/2021	Improving language models by retrieving from trillions of tokens	Retro	DeepMind	ICML citation 688

Date	Paper	keywords	Institute	Publication
12/2021	Scaling Language Models: Methods, Analysis & Insights from Training Gopher	Gopher	DeepMind	 citation 
01/2022	Chain-of-Thought Prompting Elicits Reasoning in Large Language Models	COT	Google	NeurIPS  citation 
01/2022	LaMDA: Language Models for Dialog Applications	LaMDA	Google	 citation 
01/2022	Solving Quantitative Reasoning Problems with Language Models	Minerva	Google	NeurIPS  citation 
01/2022	Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model	Megatron-Turing NLG	Microsoft&NVIDIA	 Dynamic JSON Badge
03/2022	Training language models to follow instructions with human feedback	InstructGPT	OpenAI	 citation 
04/2022	PaLM: Scaling Language Modeling with Pathways	PaLM	Google	 citation 
04/2022	An empirical analysis of compute-optimal large language model training	Chinchilla	DeepMind	NeurIPS  citation 
05/2022	OPT: Open Pre-trained Transformer Language Models	OPT	Meta	 citation 

Date	Paper	keywords	Institute	Publication
05/2022	Unifying Language Learning Paradigms	UL2	Google	 citation  123
06/2022	Emergent Abilities of Large Language Models	Emergent Abilities	Google	TMLR  citation  1572
06/2022	Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models	BIG-bench	Google	 citation  1159
06/2022	Language Models are General-Purpose Interfaces	METALM	Microsoft	 citation  85
09/2022	Improving alignment of dialogue agents via targeted human judgements	Sparrow	DeepMind	 citation  390
10/2022	Scaling Instruction-Finetuned Language Models	Flan-T5/PaLM	Google	 Dynamic JSON Badge
10/2022	GLM-130B: An Open Bilingual Pre-trained Model	GLM-130B	Tsinghua	ICLR  citation  829
11/2022	Holistic Evaluation of Language Models	HELM	Stanford	 citation  597
11/2022	BLOOM: A 176B-Parameter Open-Access Multilingual Language Model	BLOOM	BigScience	 citation  1674

Date	Paper	keywords	Institute	Publication
11/2022	Galactica: A Large Language Model for Science	Galactica	Meta	 Dynamic JSON Badge
12/2022	OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization	OPT-IML	Meta	 citation 202
01/2023	The Flan Collection: Designing Data and Methods for Effective Instruction Tuning	Flan 2022 Collection	Google	 citation 395
02/2023	LLaMA: Open and Efficient Foundation Language Models	LLaMA	Meta	 citation 6789
02/2023	Language Is Not All You Need: Aligning Perception with Language Models	Kosmos-1	Microsoft	 citation 358
03/2023	PaLM-E: An Embodied Multimodal Language Model	PaLM-E	Google	 citation 949
03/2023	GPT-4 Technical Report	GPT 4	OpenAI	 citation 5648

Life-Cycle of the LLM

1. Data collection + preprocessing

2. Pre-training

- Architecture decisions:

```
{model_size,  
hyperparameters,  
parallelism, lr_schedule,  
...}
```

3. Supervised Fine-Tuning

- Instruction Tuning
- Alignment

4. Deploy (+ monitor, re-evaluate, etc.)

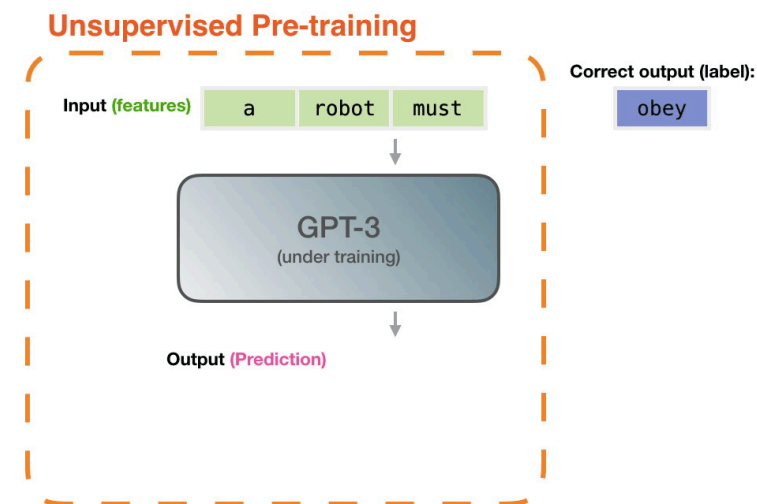


Figure 3: **Pre-training:** Virtually all of the compute used during pretraining phase¹.

Life-Cycle of the LLM: Pre-training

20

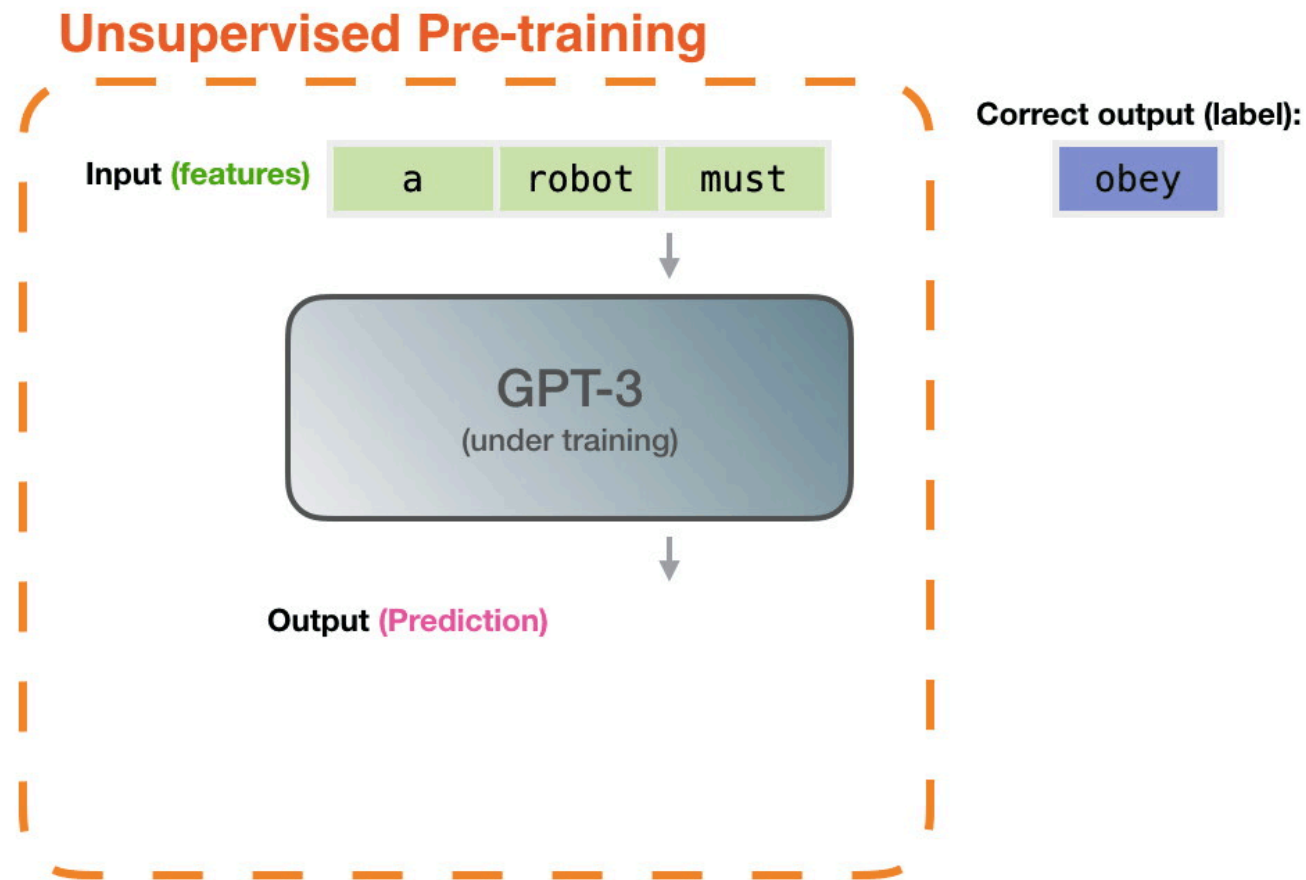


Figure 4: **Pre-training:** Virtually all of the compute used during pretraining phase

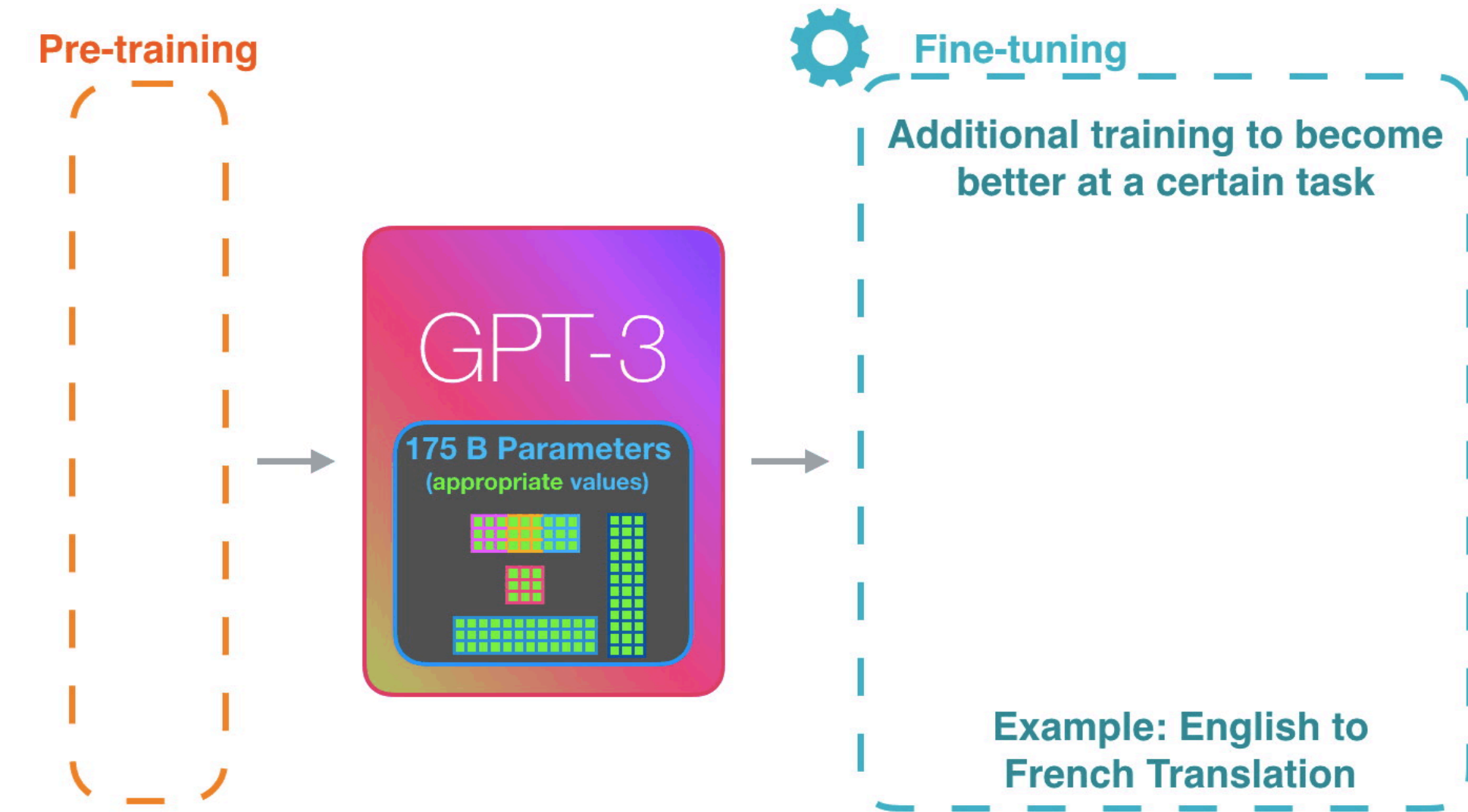


Figure 5: **Fine-tuning**¹: Fine-tuning actually updates the model's weights to make the model better at a certain task.

Forward Pass

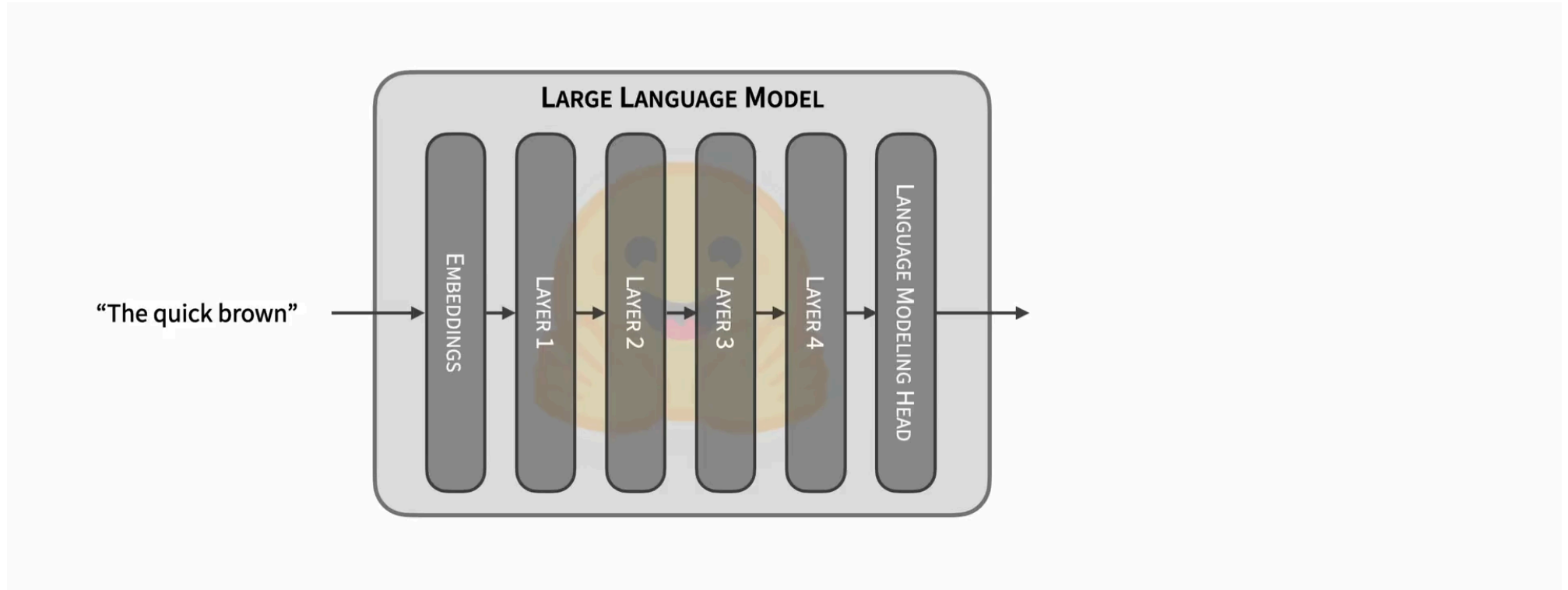


Figure 6: Language Model trained for causal language modeling. Video from: [😊 Generation with LLMs](#)

Generating Text



Figure 7: Language Model trained for causal language modeling. Video from: 🤖 [Generation with LLMs](#)

Parallelism Overview

***Modern parallelism techniques** enable the training of large language models*

See my slides on [Parallel Training Techniques](#) for additional details

Parallelism Concepts

- **DataParallel (DP):**

- The same setup is replicated multiple times, and each being fed a slice of the data.
- The processing is done in parallel and all setups are synchronized at the end of each training step.

- **TensorParallel (TP):**

- Each tensor is split up into multiple chunks.
- So, instead of having the whole tensor reside on a single gpu, each shard of the tensor resides on its designated gpu.
 - During processing each shard gets processed separately and in parallel on different GPUs and the results are synced at the end of the step.
 - This is what one may call horizontal parallelism, as the splitting happens on horizontal level.

 [Model Parallelism](#)

Parallelism Concepts¹

- **PipelineParallel (PP):**

- Model is split up vertically (layer-level) across multiple GPUs, so that only one or several layers of the model are placed on a single GPU.
 - Each GPU processes in parallel different stages of the pipeline and working on a small chunk of the batch.

- **Zero Redundancy Optimizer (ZeRO):**

- Also performs sharding of the tensors somewhat similar to TP, except the whole tensor gets reconstructed in time for a forward or backward computation, therefore the model doesn't need to be modified.
- It also supports various offloading techniques to compensate for limited GPU memory.

- **Sharded DDP:**

- Another name for the foundational ZeRO concept as used by various other implementations of ZeRO.

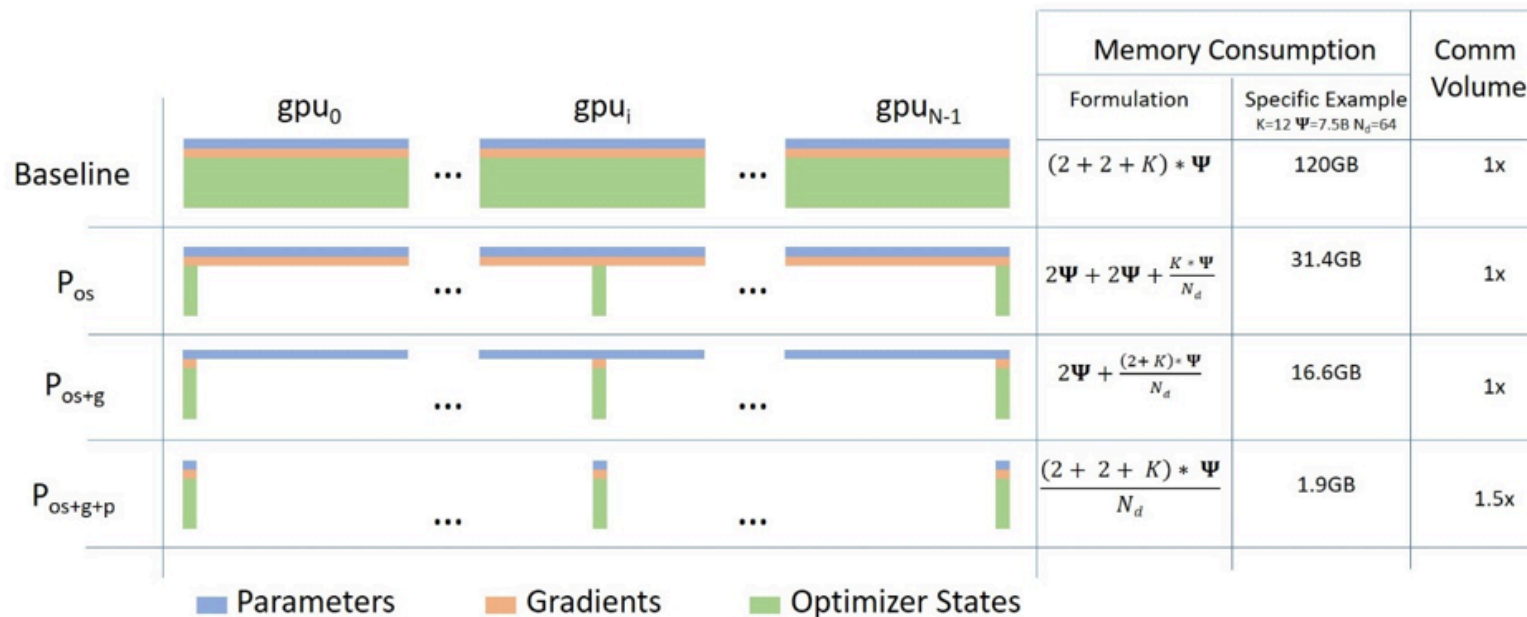
Data Parallelism

- **Data Parallelism:**

- The simplest and most common parallelism technique. Workers maintain *identical copies* of the *complete* model and work on a *subset of the data*.
- DDP supported in PyTorch native.

- ZeRO Data Parallel

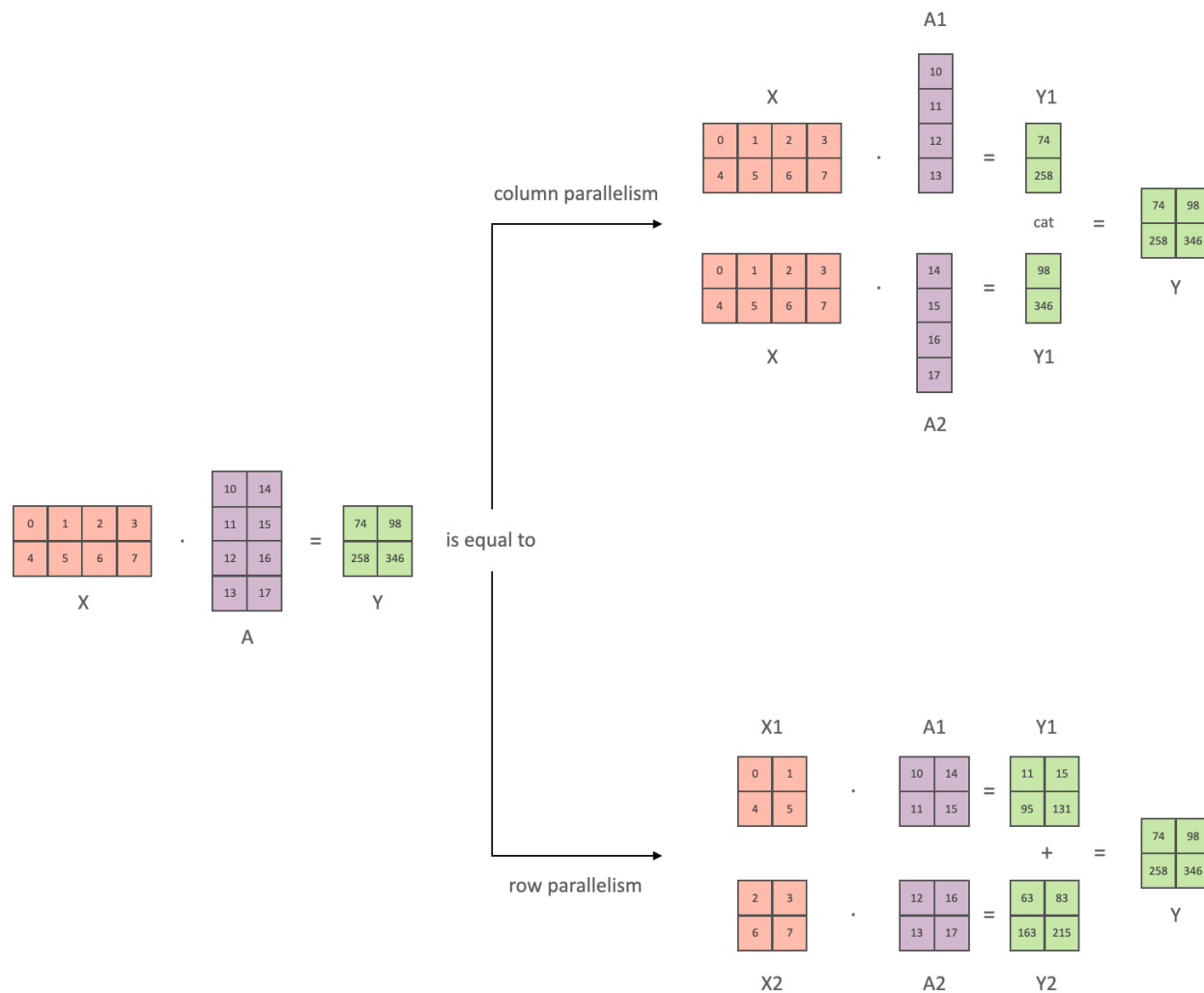
- ZeRO powered data parallelism is shown below¹



Tensor Parallelism¹

- In **Tensor Parallelism** each GPU processes only a slice of a tensor and only aggregates the full tensor for operations that require the whole thing.
 - The main building block of any transformer is a fully connected nn.Linear followed by a nonlinear activation GeLU.
 - $Y = \text{GeLU}(XA)$, where X and Y are the input and output vectors, and A is the weight matrix.
 - If we look at the computation in matrix form, it's easy to see how the matrix multiplication can be split between multiple GPUs:

Tensor Parallelism



3D Parallelism

30

- DP + TP + PP (3D) Parallelism

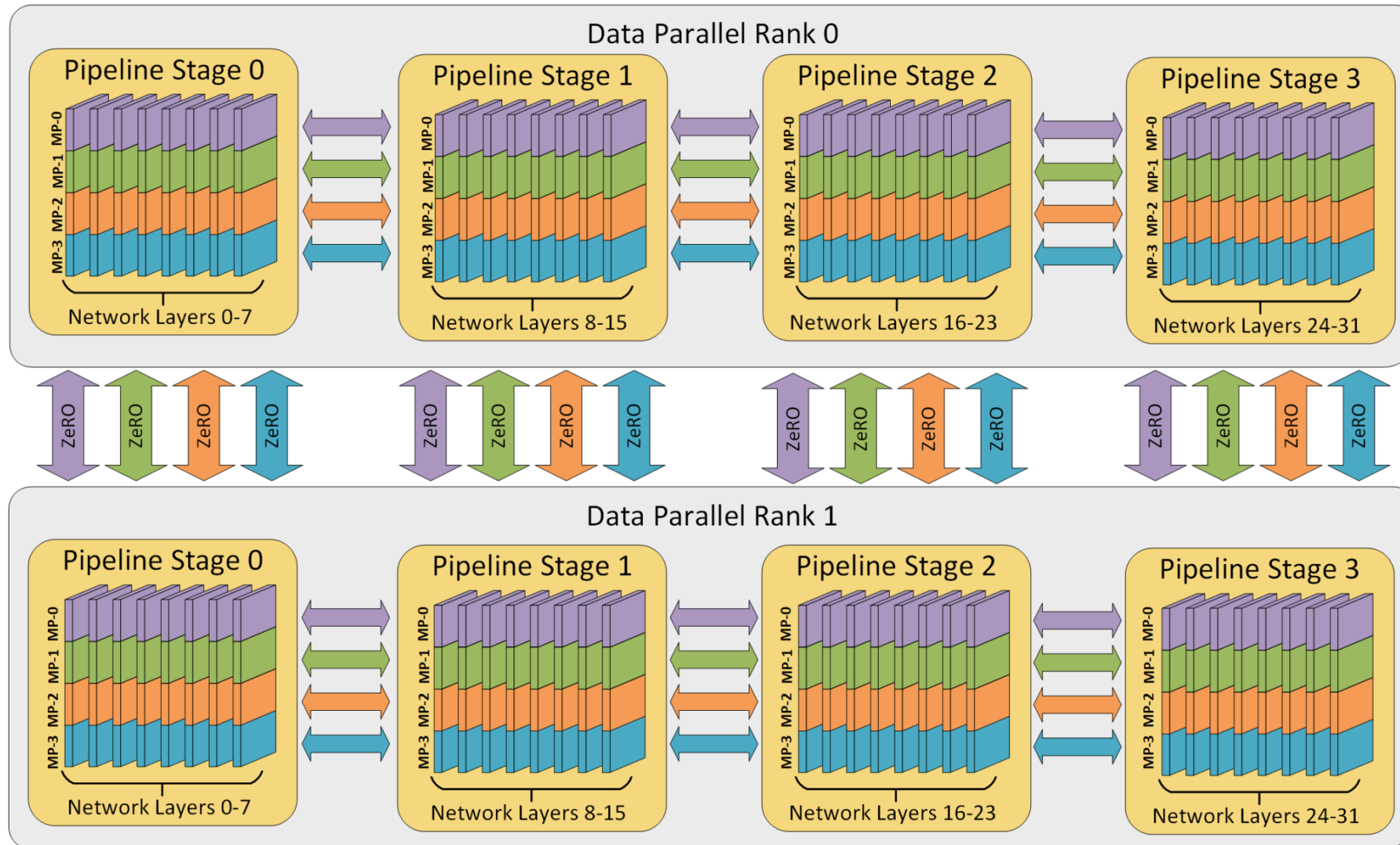


Figure 8: 3D Parallelism illustration. Figure from: <https://www.deepspeed.ai/>

3D Parallelism

- DP + TP + PP (3D) Parallelism

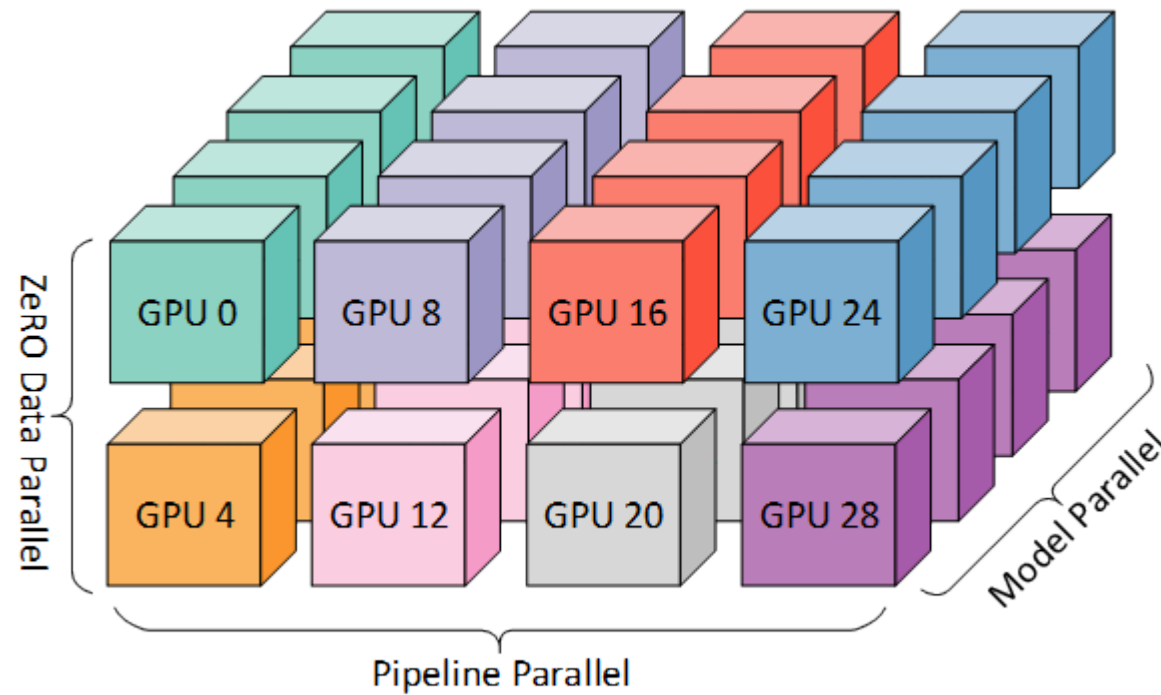
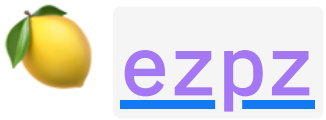


Figure 9: Figure taken from [3D parallelism: Scaling to trillion-parameter models](#)



Clone Repo(s)

```
#[★][07:33:08 AM][foremans@x3101c0s13b0n0][~/tmp]
$ mkdir ~/tmp/polaris-talk

#[★][07:33:21 AM][foremans@x3101c0s13b0n0][~/tmp]
$ cd ~/tmp/polaris-talk

#[★][07:33:25 AM][foremans@x3101c0s13b0n0][~/tmp/polaris-talk]
$ NOW=$(tstamp) && mkdir "${NOW}" && cd "${NOW}" # && mkdir "core-dumps-${NOW}" && mv -v

#[★][07:33:27 AM][foremans@x3101c0s13b0n0][~/tmp/polaris-talk/2024-07-17-073327]
$ pwd
/home/foremans/tmp/polaris-talk/2024-07-17-073327

#[★][07:33:31 AM][foremans@x3101c0s13b0n0][~/tmp/polaris-talk/2024-07-17-073327]
$ git clone https://github.com/saforem2/ezpz ezpz && git clone https://github.com/saforem2/ezpz
Cloning into 'ezpz'...
remote: Enumerating objects: 2134, done.
remote: Counting objects: 100% (.363/363), done.
remote: Compressing objects: 100% (.169/169), done.
remote: Total 2134 (.delta 197), reused 265 (.delta 141), pack-reused 1771
```

Setup Python

```
#[★][07:33:53 AM][foremans@x3101c0s13b0n0][~/tmp/polaris-talk/2024-07-17-073327]
$ source ezipz/src/ezpz/bin/utils.sh && ezipz_setup_python && ezipz_setup_alcf
Unable to detect PBS or SLURM working directory info...
Using /home/foremans/tmp/polaris-talk/2024-07-17-073327 as working directory...
Using WORKING_DIR: /home/foremans/tmp/polaris-talk/2024-07-17-073327
No conda_prefix OR virtual_env found in environment...
Setting up conda...
Lmod is automatically replacing "nvhpc/23.9" with "gcc-native/12.3".
Lmod is automatically replacing "PrgEnv-nvhpc/8.5.0" with "PrgEnv-gnu/8.5.0".
Due to MODULEPATH changes, the following have been reloaded:
  1.) cray-mpich/8.1.28
Found conda at: /soft/applications/conda/2024-04-29/mconda3
No VIRTUAL_ENV found in environment!
  - Trying to setup from /soft/applications/conda/2024-04-29/mconda3
  - Using VENV_DIR=/home/foremans/tmp/polaris-talk/2024-07-17-073327/venvs/2024-04-29
  - Creating a new virtual env on top of 2024-04-29 in /home/foremans/tmp/polaris-talk/2024-07-17-073327/venvs/2024-04-29
[python] Using /home/foremans/tmp/polaris-talk/2024-07-17-073327/venvs/2024-04-29/bin/python
[ezpz/bin/utils.sh]
```

Install {ezpz, wordplay}

```
#[★][07:34:13 AM][foremans@x3101c0s13b0n0][~/tmp/polaris-talk/2024-07-17-073327]
$ python3 -m pip install -e ezpz wordplay --require-virtualenv
Looking in indexes: https://pypi.org/simple, https://pypi.ngc.nvidia.com
Obtaining file:///home/foremans/tmp/polaris-talk/2024-07-17-073327/ezpz
  Installing build dependencies ... done
  Checking if build backend supports build_editable ... done
  Getting requirements to build editable ... done
  Installing backend dependencies ... done
  Preparing editable metadata (.pyproject.toml) ... done

# ...[clipped]...

Successfully built ezpz
Installing collected packages: enum34, wordplay, pyinstrument, ezpz
  Attempting uninstall: ezpz
    Found existing installation: ezpz 0.1
    Not uninstalling ezpz at /home/foremans/.local/polaris/conda/2024-04-29/lib/python3.1
    Cant uninstall 'ezpz'. No files were found to uninstall.
Successfully installed enum34-1.1.10 ezpz pyinstrument-4.6.2 wordplay-1.0.0a4
[notice] A new release of pip is available: 24.0 -> 24.1.2
```

Launch ezpz.test_dist

```
#(🤖 2024-04-29)
#[★][07:34:07 AM][foremans@x3101c0s13b0n0][~/tmp/polaris-talk/2024-07-17-073327][🕒 7s]
$ which launch
launch: aliased to mpiexec --verbose --envall -n 4 -ppn 4 --hostfile /var/spool/pbs/aux/20

#(🤖 2024-04-29)
#[★][07:34:11 AM][foremans@x3101c0s13b0n0][~/tmp/polaris-talk/2024-07-17-073327]
$ which python3
/home/foremans/tmp/polaris-talk/2024-07-17-073327/venvs/2024-04-29/bin/python3

#(🤖 2024-04-29)
#[★][07:35:21 AM][foremans@x3101c0s13b0n0][~/tmp/polaris-talk/2024-07-17-073327][🕒 14s]
$ launch python3 -m ezpz.test_dist | tee ezpz-test-dist-DDP.log
Connected to tcp://x3101c0s13b0n0.hsn.cm.polaris.alcf.anl.gov:7919
Found executable /home/foremans/tmp/polaris-talk/2024-07-17-073327/venvs/2024-04-29/bin/python3
Launching application cff755ee-557e-4df2-a987-db85a8b7dbe7
[2024-07-17 07:35:30.304306][INFO][__init__:156] - Setting logging level to 'INFO' on 'RAI
[2024-07-17 07:35:30.307036][INFO][__init__:157] - Setting logging level to 'CRITICAL' on
[2024-07-17 07:35:30.307494][INFO][__init__:160] - To disable this behavior, and log from
[2024-07-17 07:35:32.116037][INFO][dist:358] - [device='cuda'][rank=2/3][local_rank=2/3][1
```

PyInstrument Profile

```

Recorded: 07:35:34   Samples: 2227
Duration: 2.948      CPU time: 5.441
PyInstrument: v4.6.2
Program: /home/foremans/tmp/polaris-talk/2024-07-17-073327/ezpz/src/ezpz/test_dist.py
2.948 <module>   ezpz/test_dist.py:1
└─ 2.946 main   ezpz/test_dist.py:217
    └─ 2.043 build_model_and_optimizer   ezpz/test_dist.py:171
        └─ 2.011 Adam.__init__   torch/optim/adam.py:15
            [129 frames hidden]   torch, wandb, transformers, jax, func...
    └─ 0.326 _forward_step   ezpz/test_dist.py:231
        └─ 0.279 DistributedDataParallel._wrapped_call_impl   torch/nn/modules/module.py:1528
            [13 frames hidden]   torch, wandb, <built-in>
            └─ 0.273 Network._call_impl   torch/nn/modules/module.py:1534
                └─ 0.076 Network.forward   ezpz/test_dist.py:164
                    └─ 0.076 Sequential._wrapped_call_impl   torch/nn/modules/module.py:1528
                        [7 frames hidden]   torch, <built-in>
            └─ 0.046 calc_loss   ezpz/test_dist.py:168
    └─ 0.254 _backward_step   ezpz/test_dist.py:236
        └─ 0.177 Tensor.backward   torch/_tensor.py:466
            [4 frames hidden]   torch, <built-in>

```

Interactive Example

```

Requirement already satisfied: certifi>=2017.4.17 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from requests->ambivalent->ezpz==0.1) (2024.2.2)
Requirement already satisfied: MarkupSafe>=2.1.1 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from werkzeug>=1.0.1->tensorboard->ezpz==0.1) (2.1.3)
Requirement already satisfied: executing>=1.2.0 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from stack-data->ipython->ezpz==0.1) (2.0.1)
Requirement already satisfied: asttokens>=2.1.0 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from stack-data->ipython->ezpz==0.1) (2.4.1)
Requirement already satisfied: pure-eval in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from stack-data->ipython->ezpz==0.1) (0.2.2)
Requirement already satisfied: mpmath>=0.19 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from sympy->torch->ezpz==0.1) (1.3.0)
Requirement already satisfied: smmap<6,>=3.0.1 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from gitdb<5,>=4.0.1->GitPython!=3.1.29,>=1.0.0->wandb->ezpz==0.1) (5.0.1)
Downloading wordplay-1.0.0a4-py2.py3-none-any.whl (2.5 MB)
----- 2.5/2.5 MB 19.0 MB/s eta 0:00:00
Downloading enum34-1.1.10-py3-none-any.whl (11 kB)
Downloading pyinstrument-4.6.2-cp311-cp311-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux2_17_x86_64.manylinux2014_x86_64.whl (104 kB)
----- 104.9/104.9 kB 177.7 MB/s eta 0:00:00
Building wheels for collected packages: ezpz
  Building editable for ezpz (pyproject.toml) ... done
  Created wheel for ezpz: filename=ezpz-0.1-py3-none-any.whl size=10103 sha256=b5e4d4849dd403ab9c3d3934ec81eb7da4b5f0827099095fc10773df9bc01e00
  Stored in directory: /tmp/pip-ephem-wheel-cache-c4bimx0/wheels/d3/0d/05/0dad1cb73bf8ec802f58ad484cb95e5179856c1b3287d90a60
Successfully built ezpz
Installing collected packages: enum34, wordplay, pyinstrument, ezpz
  Attempting uninstall: ezpz
    Found existing installation: ezpz 0.1
    Not uninstalling ezpz at /home/foremans/.local/polaris/conda/2024-04-29/lib/python3.11/site-packages, outside environment /home/foremans/tmp/polaris-talk/2024-07-17-073327/venvs/2024-04-29
    Can't uninstall 'ezpz'. No files were found to uninstall.
Successfully installed enum34-1.1.10 ezpz pyinstrument-4.6.2 wordplay-1.0.0a4

[notice] A new release of pip is available: 24.0 -> 24.1.2
[notice] To update, run: pip install --upgrade pip
9.62s user 1.11s system 61% cpu 17.50s total
(2024-04-29)
# [★] [07:34:47 AM] [foremans@x3101c0s13b0n0] [~/tmp/polaris-talk/2024-07-17-073327] [0 17s]
$ which launch
launch: aliased to mpiexec --verbose --envall -n 4 -ppn 4 --hostfile /var/spool/pbs/aux/2024084.polaris-pbs-01.hpc.alcf.anl.gov --cpu-bind depth -d 16
(2024-04-29)
# [★] [07:34:53 AM] [foremans@x3101c0s13b0n0] [~/tmp/polaris-talk/2024-07-17-073327]
$ python3 -m pip install --upgrade wandb
Looking in indexes: https://pypi.org/simple, https://pypi.ngc.nvidia.com
Requirement already satisfied: wandb in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (0.16.6)
Collecting wandb
  Downloading wandb-0.17.4-py3-none-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux2_17_x86_64.manylinux2014_x86_64.whl.metadata (10 kB)
Requirement already satisfied: click!=8.0.0,>=7.1 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from wandb) (8.1.7)
Requirement already satisfied: docker-pycrds>=0.4.0 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from wandb) (0.4.0)
Requirement already satisfied: gitpython!=3.1.29,>=1.0.0 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from wandb) (3.1.43)
Requirement already satisfied: platformdirs in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from wandb) (3.10.0)
Requirement already satisfied: protobuf!=4.21.0,<6,>=3.19.0 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from wandb) (3.20.3)
Requirement already satisfied: psutil>=5.0.0 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from wandb) (5.9.8)
Requirement already satisfied: pyyaml in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from wandb) (6.0.1)
Requirement already satisfied: requests<3,>=2.0.0 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from wandb) (2.31.0)
Requirement already satisfied: sentry-sdk>=1.0.0 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from wandb) (2.0.1)
Requirement already satisfied: setproctitle in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from wandb) (1.3.3)
Requirement already satisfied: setuptools in ./venvs/2024-04-29/lib/python3.11/site-packages (from wandb) (65.5.0)
Requirement already satisfied: six>=1.4.0 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from docker-pycrds>=0.4.0->wandb) (1.16.0)
Requirement already satisfied: gitdb<5,>=4.0.1 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from gitpython!=3.1.29,>=1.0.0->wandb) (4.0.11)
Requirement already satisfied: charset-normalizer<4,>=2 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from requests<3,>=2.0.0->wandb) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from requests<3,>=2.0.0->wandb) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from requests<3,>=2.0.0->wandb) (2.1.0)
Requirement already satisfied: certifi>=2017.4.17 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from requests<3,>=2.0.0->wandb) (2024.2.2)
Requirement already satisfied: smmap<6,>=3.0.1 in /soft/applications/conda/2024-04-29/mconda3/lib/python3.11/site-packages (from gitdb<5,>=4.0.1->gitpython!=3.1.29,>=1.0.0->wandb) (5.0.1)
Downloading wandb-0.17.4-py3-none-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux2_17_x86_64.manylinux2014_x86_64.whl (6.9 MB)
----- 5.4/6.9 MB 2.3 MB/s eta 0:00:01

```



Recorded with **asciinema**

Example: wordplay 

Prepare Data

```
#[★][07:41:20 AM][foremans@x3101c0s13b0n0][~/tmp/polaris-talk/2024-07-17-073327][🕒 29s]
$ python3 wordplay/data/shakespeare_char/prepare.py
Using HF_DATASETS_CACHE=/home/foremans/tmp/polaris-talk/2024-07-17-073327/wordplay/data/sl
length of dataset in characters: 1,115,394
all the unique characters:
!$&\',-.3:;?ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz
vocab size: 65
train has 1,003,854 tokens
val has 111,540 tokens
```

Launch Training (DDP)

```
#(🤖 2024-04-29)
#[★][07:42:02 AM][foremans@x3101c0s13b0n0][~/tmp/polaris-talk/2024-07-17-073327]
$ launch python3 -m wordplay train.backend=DDP train.eval_interval=100 data=shakespeare t:
[2024-07-17 07:42:11.746540][INFO][__init__:156] - Setting logging level to 'INFO' on 'RAI
[2024-07-17 07:42:11.748763][INFO][__init__:157] - Setting logging level to 'CRITICAL' on
[2024-07-17 07:42:11.749453][INFO][__init__:160] - To disable this behavior, and log from
[2024-07-17 07:42:11.772718][INFO][configs:81] - Setting HF_DATASETS_CACHE to /home/forema
[2024-07-17 07:42:15.341532][INFO][dist:358] - [device='cuda'][rank=2/3][local_rank=2/3][1
[2024-07-17 07:42:15.342381][INFO][dist:358] - [device='cuda'][rank=1/3][local_rank=1/3][1
[2024-07-17 07:42:15.342430][INFO][dist:358] - [device='cuda'][rank=3/3][local_rank=3/3][1
[2024-07-17 07:42:15.348657][INFO][dist:95] -
```

[dist_info]:

- DEVICE=cuda
- DEVICE_ID=cuda:0
- DISTRIBUTED_BACKEND=nccl
- GPUS_PER_NODE=4
- HOSTS=['x3101c0s13b0n0.hsn.cm.polaris.alcf.anl.gov']
- HOSTFILE=/var/spool/pbs/aux/2024084.polaris-pbs-01.hsn.cm.polaris.alcf.anl.gov
- HOSTNAME=x3101c0s13b0n0.hsn.cm.polaris.alcf.anl.gov

Interactive Example

```

128809
[2024-07-17 07:43:18.602785][INFO][trainer:885] - step=70 loss=2.458894 dt=0.283926 dtf=0.005219 dtb=0.010383 sps=14.088155 sps_per_gpu=3.522039 tps=923281.352698 tps_per_gpu=230820.338174 mfu=45.998106 train_loss=4.125778 val_loss=4.128809
[2024-07-17 07:43:21.451433][INFO][trainer:885] - step=80 loss=2.489088 dt=0.285537 dtf=0.005183 dtb=0.011373 sps=14.008683 sps_per_gpu=3.502171 tps=918073.060430 tps_per_gpu=229518.265108 mfu=45.983282 train_loss=4.125778 val_loss=4.128809
[2024-07-17 07:43:24.302241][INFO][trainer:885] - step=90 loss=2.471990 dt=0.300767 dtf=0.005445 dtb=0.010290 sps=13.299337 sps_per_gpu=3.324834 tps=871585.359388 tps_per_gpu=217896.339847 mfu=45.737774 train_loss=4.125778 val_loss=4.128809
[2024-07-17 07:43:27.153275][INFO][trainer:885] - step=100 loss=2.445556 dt=0.285869 dtf=0.005182 dtb=0.011251 sps=13.992403 sps_per_gpu=3.498101 tps=917006.151328 tps_per_gpu=229251.537832 mfu=45.743655 train_loss=4.125778 val_loss=4.128809
[2024-07-17 07:43:28.182553][INFO][trainer:820] - ['prompt']: 'What is an LLM?'
[2024-07-17 07:43:28.183179][INFO][trainer:824] - ['response']:

What is an LLM?

Goupay my winghimithell bls ger t bon sinthard ht omind be,
And lereind h py balithand frd oforondof wimon me hageas thiner o mand,
Thacanes,
An frift ghik med d herthecke ntore thack couthen ale, t thit ang d m t h chy me fache ag, wit my hathan glat ng
[2024-07-17 07:44:06.025837][INFO][trainer:760] - Saving checkpoint to: /home/foremans/tmp/polaris-talk/outputs/runs/pytorch/DDP/2024-07-17/07-42-13
[2024-07-17 07:44:06.026607][INFO][trainer:761] - Saving model to: /home/foremans/tmp/polaris-talk/outputs/runs/pytorch/DDP/2024-07-17/07-42-13/model.pth
[2024-07-17 07:44:07.682968][INFO][configs:141] - Appending /home/foremans/tmp/polaris-talk/outputs/runs/pytorch/DDP/2024-07-17/07-42-13 to /home/foremans/tmp/polaris-talk/2024-07-17-073327/wordplay/src/ckpts/checkpoints.log
[2024-07-17 07:44:10.519506][INFO][trainer:885] - step=110 loss=2.433923 dt=0.285038 dtf=0.005757 dtb=0.011762 sps=14.033209 sps_per_gpu=3.508302 tps=919680.367894 tps_per_gpu=229920.091974 mfu=45.762304 train_loss=2.439494 val_loss=2.478951
[2024-07-17 07:44:13.362148][INFO][trainer:885] - step=120 loss=2.429014 dt=0.284445 dtf=0.005222 dtb=0.011486 sps=14.062460 sps_per_gpu=3.515615 tps=921597.361532 tps_per_gpu=230399.340383 mfu=45.788661 train_loss=2.439494 val_loss=2.478951
[2024-07-17 07:44:16.210694][INFO][trainer:885] - step=130 loss=2.402059 dt=0.285559 dtf=0.005199 dtb=0.011765 sps=14.007633 sps_per_gpu=3.501908 tps=918004.211586 tps_per_gpu=229501.052897 mfu=45.794438 train_loss=2.439494 val_loss=2.478951
[2024-07-17 07:44:19.061546][INFO][trainer:885] - step=140 loss=2.374062 dt=0.285476 dtf=0.005239 dtb=0.011453 sps=14.011662 sps_per_gpu=3.502916 tps=918268.297093 tps_per_gpu=229567.074273 mfu=45.800956 train_loss=2.439494 val_loss=2.478951
[2024-07-17 07:44:21.917283][INFO][trainer:885] - step=150 loss=2.365385 dt=0.285846 dtf=0.005125 dtb=0.011320 sps=14.011662 sps_per_gpu=3.498392 tps=917082.475791 tps_per_gpu=229270.618948 mfu=45.800900 train_loss=2.439494 val_loss=2.478951
[2024-07-17 07:44:24.771924][INFO][trainer:885] - step=160 loss=2.317337 dt=0.280788 dtf=0.005173 dtb=0.011249 sps=14.011662 sps_per_gpu=3.561401 tps=933599.792506 tps_per_gpu=233399.948127 mfu=45.883340 train_loss=2.439494 val_loss=2.478951
[2024-07-17 07:44:27.626812][INFO][trainer:885] - step=170 loss=2.256231 dt=0.284973 dtf=0.005141 dtb=0.011299 sps=14.036416 sps_per_gpu=3.509104 tps=919890.544506 tps_per_gpu=229972.636126 mfu=45.889069 train_loss=2.439494 val_loss=2.478951
[2024-07-17 07:44:30.480952][INFO][trainer:885] - step=180 loss=2.216419 dt=0.286555 dtf=0.005180 dtb=0.011402 sps=13.958906 sps_per_gpu=3.489726 tps=914810.852170 tps_per_gpu=228702.713043 mfu=45.868857 train_loss=2.439494 val_loss=2.478951
[2024-07-17 07:44:33.337342][INFO][trainer:885] - step=190 loss=2.145123 dt=0.291456 dtf=0.005409 dtb=0.019347 sps=13.724205 sps_per_gpu=3.431051 tps=899429.467247 tps_per_gpu=224857.366812 mfu=45.773849 train_loss=2.439494 val_loss=2.478951
[2024-07-17 07:44:36.194584][INFO][trainer:885] - step=200 loss=2.068149 dt=0.285703 dtf=0.005153 dtb=0.011286 sps=14.000555 sps_per_gpu=3.500139 tps=917540.393411 tps_per_gpu=229385.098353 mfu=45.778791 train_loss=2.439494 val_loss=2.478951
[2024-07-17 07:44:37.224149][INFO][trainer:820] - ['prompt']: 'What is an LLM?'
[2024-07-17 07:44:37.224745][INFO][trainer:824] - ['response']:

What is an LLM?

LORTESS LA:
No, sighappat selace? don dwnnd sourciceans note cancen up sof liond
This and my man, werame, of re thee
This not will I on land brond sul me a fingore?

FLER:
Tisint your not nare lame o igen,-to brorst.

SamERS:
Sin:
I'll hell she lor hen w

```

Recorded with asciinema

Extras

Transformer Architecture

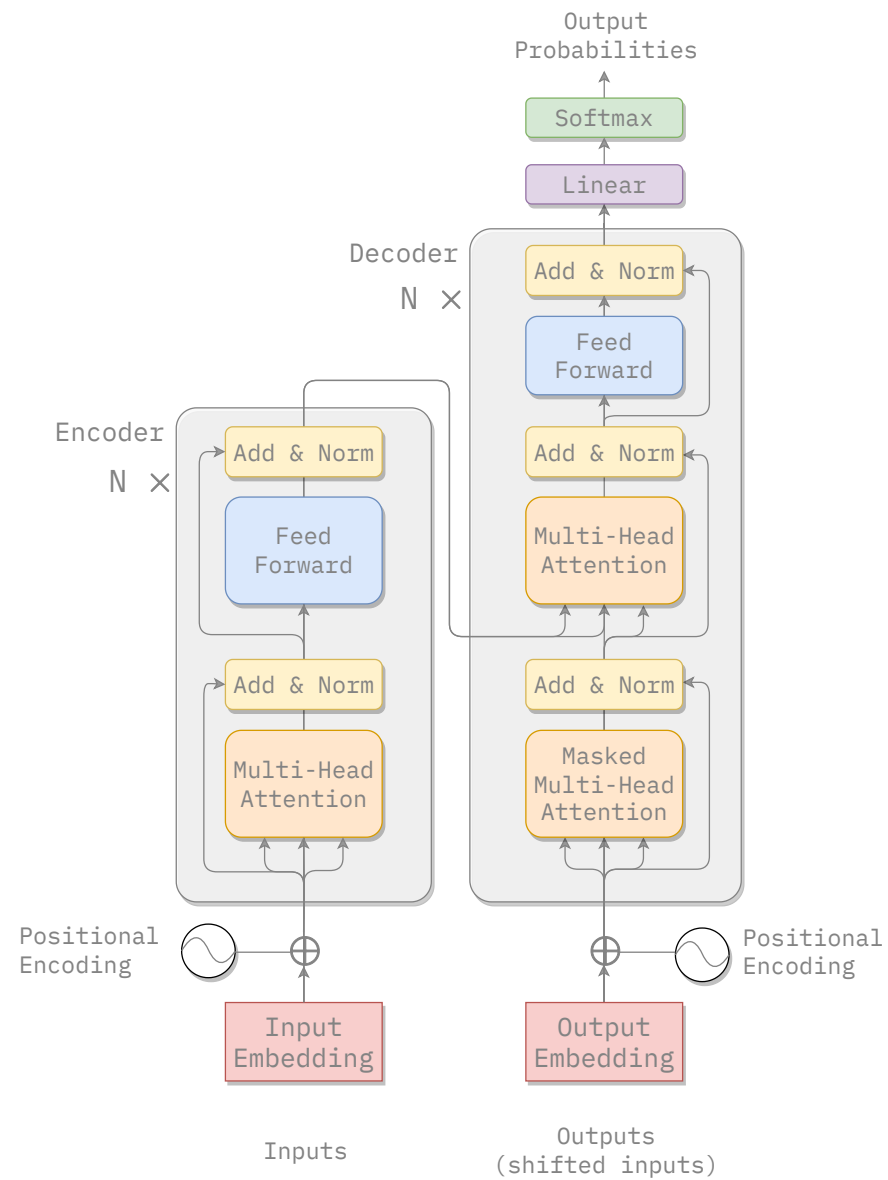


Figure 10: Vaswani et al. ([2017](#))

References

- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” <https://arxiv.org/abs/1706.03762>.
- Yang, Jingfeng, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. “Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond.” <https://arxiv.org/abs/2304.13712>.
- Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. “Tree of Thoughts: Deliberate Problem Solving with Large Language Models.” <https://arxiv.org/abs/2305.10601>.

Filter

No matching items