PHAROS
THE GREEK AI FACTORY

Scalability results of finetuning: Performance, Efficiency, Convergence and Generalization

Course 3 "Introduction to Large Language Models at Scale"

Panagiota Gyftou pgyftou@admin.grnet.gr grnet

Contents

Model training experiments

Performance

Efficiency

Convergence

GPU Utilization

Memory

Tensorboard

Model training experiments

We conducted a series of **training experiments** to *investigate* the impact of hyperparameter choices and computational scaling on the **performance** of large language models (LLMs). Specifically, we trained the meta-llama-3.2-1B and meta-llama-3.2-3B models while varying key hyperparameters, including

- batch size = (4, 8, 16, 32)
- number of training epochs = (10, 30, 50)
- number of GPUs = (1, 2, 3, 4)
- learning rate = 5e-6
- weight decay = 0.01
- warmup ratio = 0.1
- gradient accumulation = 4 steps
- maximum sequence length = 256 tokens

Contents

Model training experiments

Performance

Efficiency

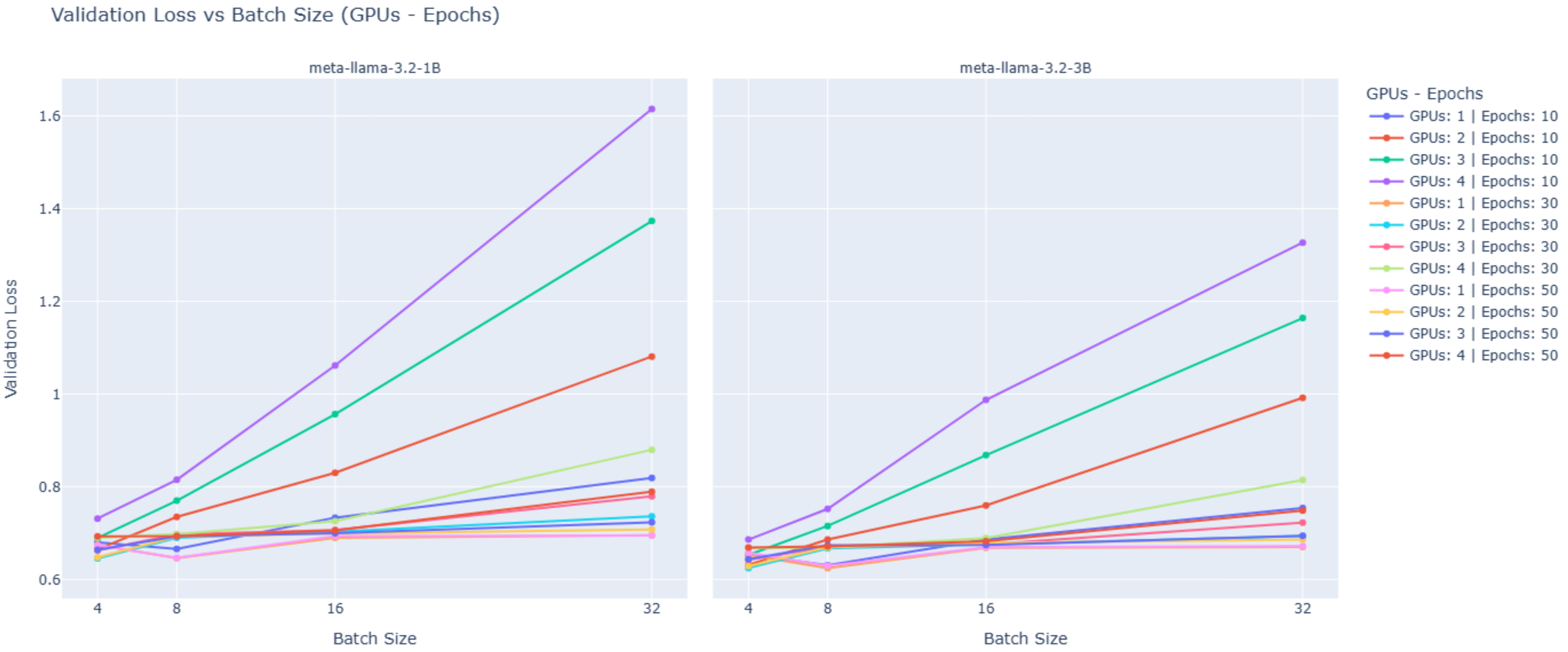
Convergence

GPU Utilization

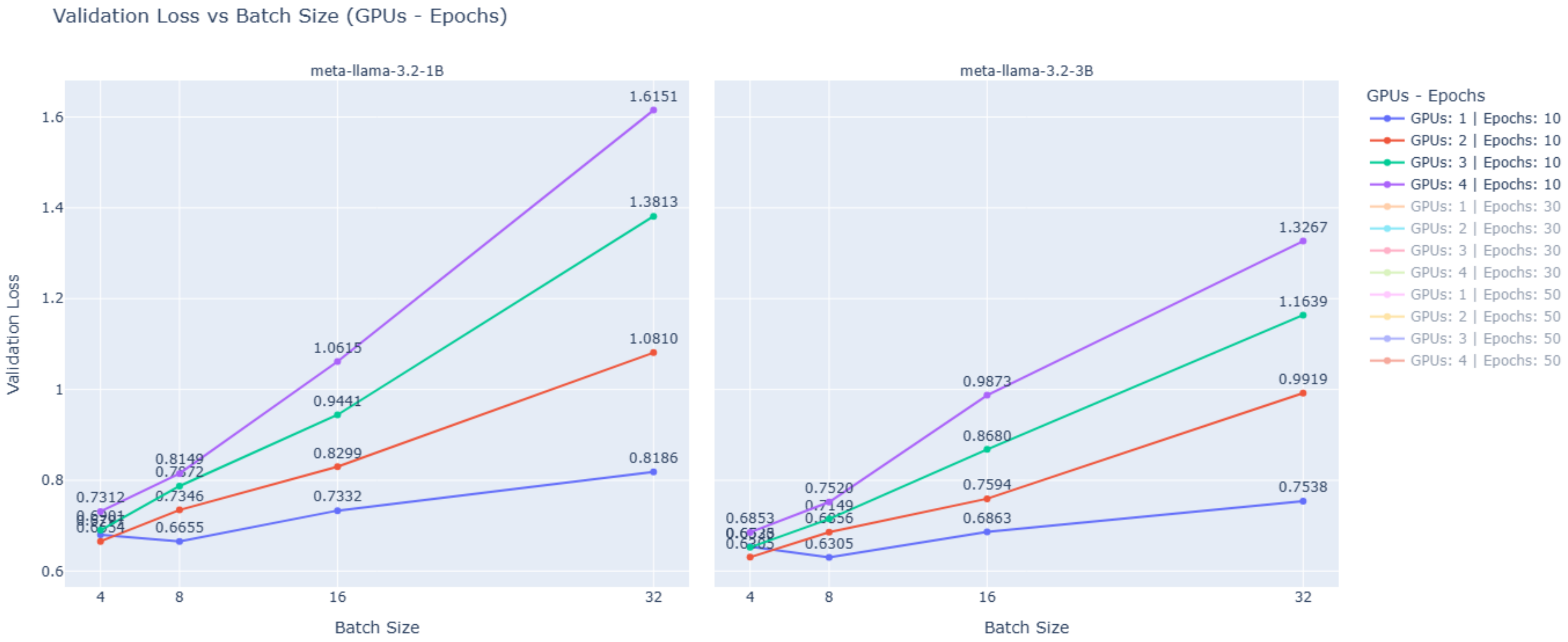
Memory

Tensorboard

Performance

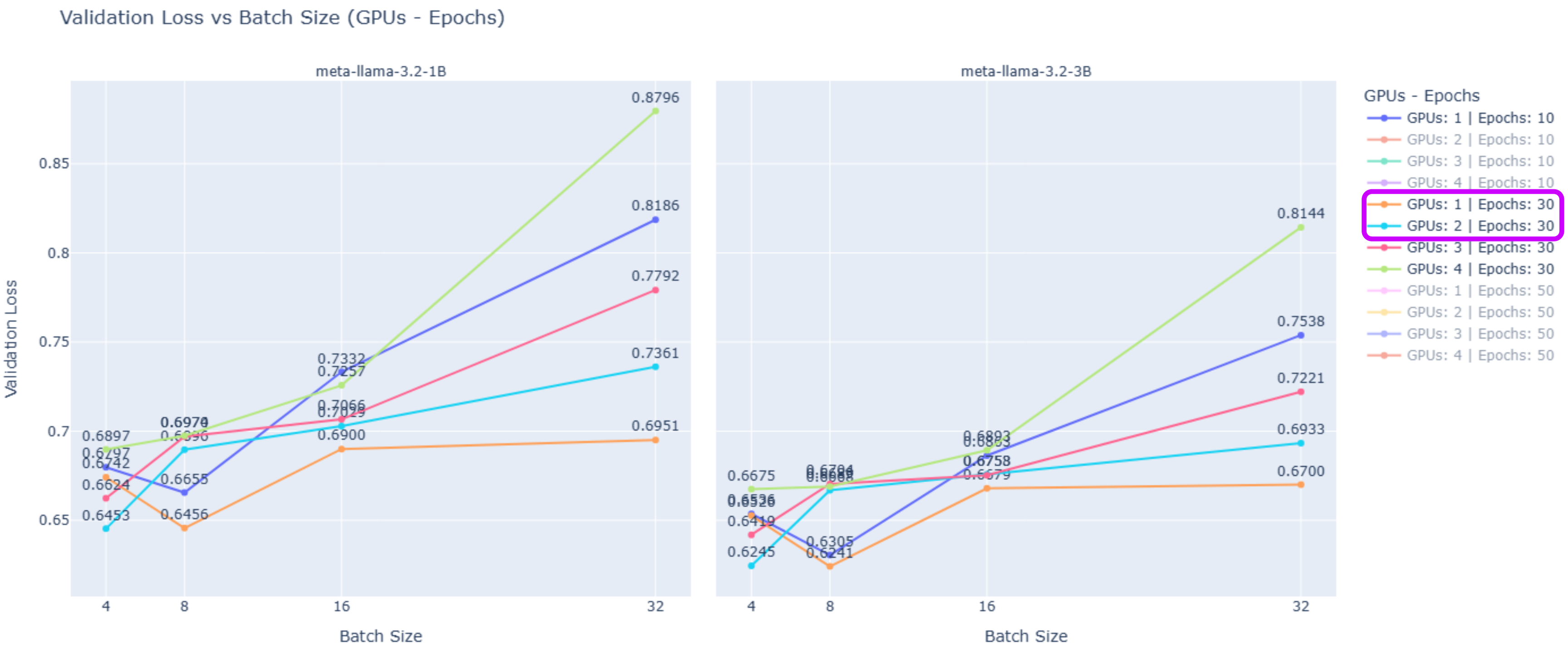


Performance 10 epochs

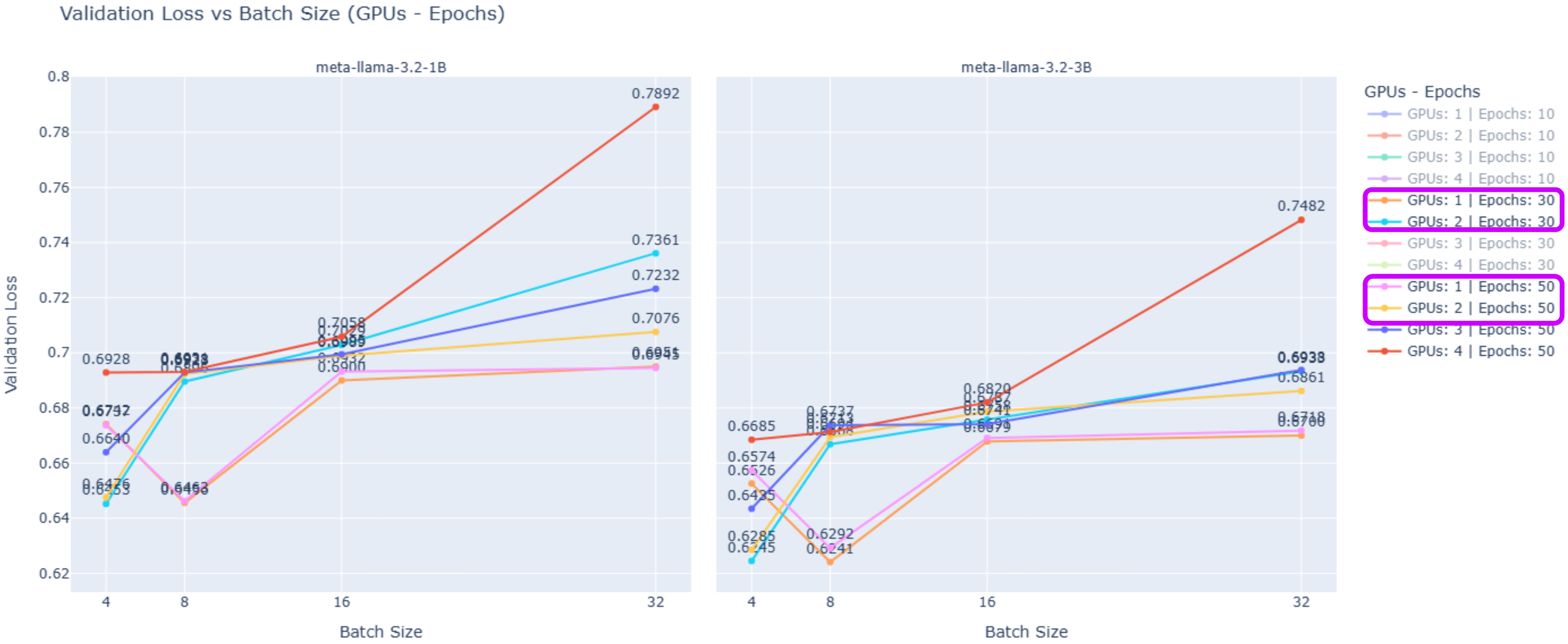


Performance 30 epochs

PHAROS
THE GREEK AI FACTORY



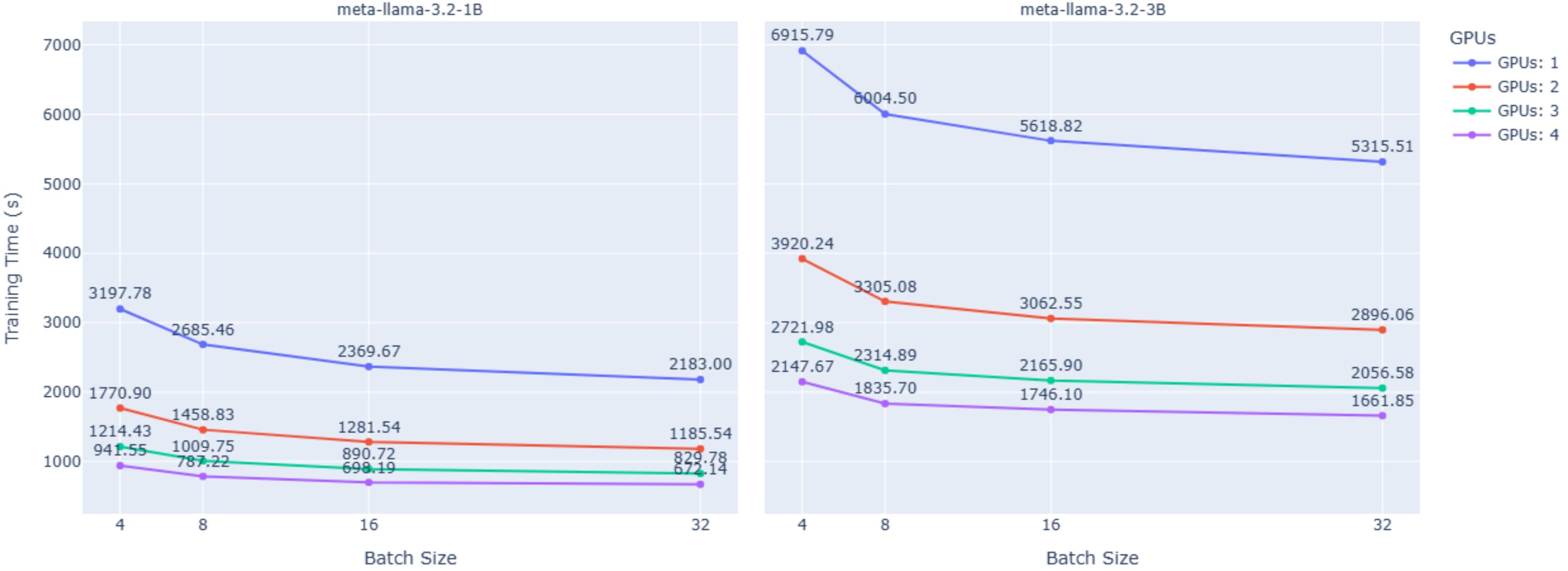
Performance 50 epochs



Validation Loss vs Batch Size (GPUs - Epochs)

Performance

Training Time vs Batch Size per Model/GPUs - 30 Epochs



Contents

Model training experiments

Performance

Efficiency

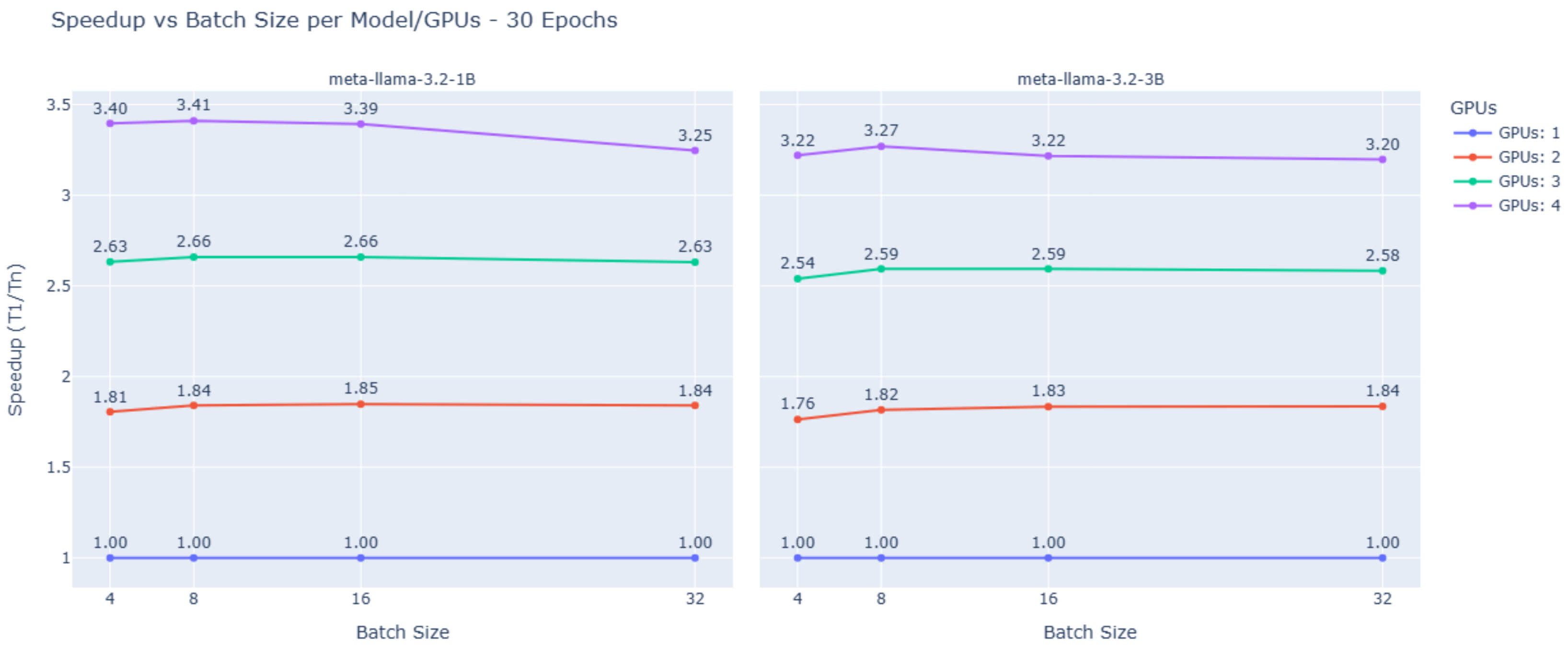
Convergence

GPU Utilization

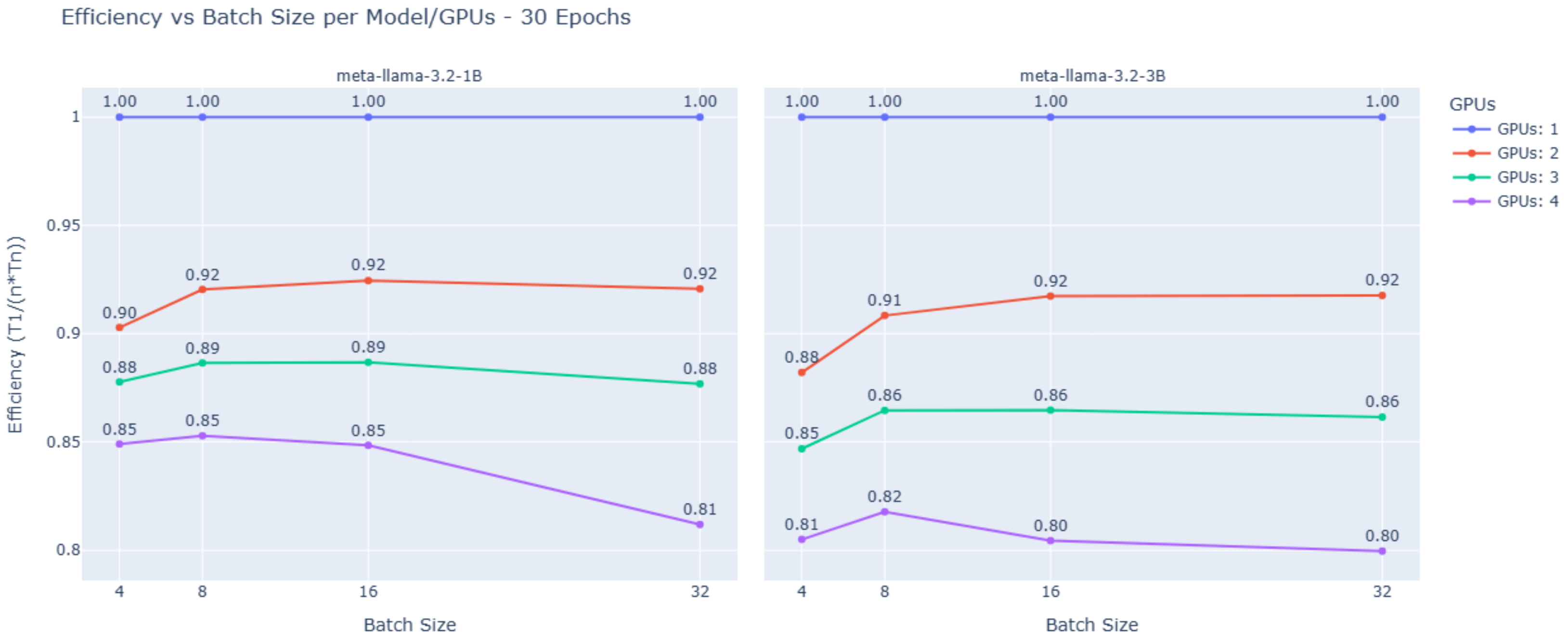
Memory

Tensorboard

Efficiency


Speedup vs Batch Size per Model/GPUs - 30 Epochs
meta-llama-3.2-1B
meta-llama-3.2-3B
Speedup (T1/Tn)
3.5
3
2.5
2
1.5
1
3.40
3.41
3.39
3.25
2.63
2.66
2.66
2.63
1.81
1.84
1.85
1.84
1.00
1.00
1.00
1.00
4
8
16
32
Batch Size
3.22
3.27
3.22
3.20
2.54
2.59
2.59
2.58
1.76
1.82
1.83
1.84
1.00
1.00
1.00
1.00
4
8
16
32
Batch Size
GPUs
GPUs: 1
GPUs: 2
GPUs: 3
GPUs: 4

Efficiency



Efficiency vs Batch Size per Model/GPUs - 30 Epochs
meta-llama-3.2-1B
meta-llama-3.2-3B
Efficiency (T1/(n*Tn))
1
0.95
0.9
0.85
0.8
Batch Size
4
8
16
32
GPUs
GPUs: 1
GPUs: 2
GPUs: 3
GPUs: 4
1.00
0.90
0.92
0.92
0.92
0.88
0.89
0.89
0.88
0.85
0.85
0.85
0.81
1.00
0.88
0.91
0.92
0.92
0.85
0.86
0.86
0.86
0.81
0.82
0.80
0.80

Contents

Model training experiments

Performance

Efficiency

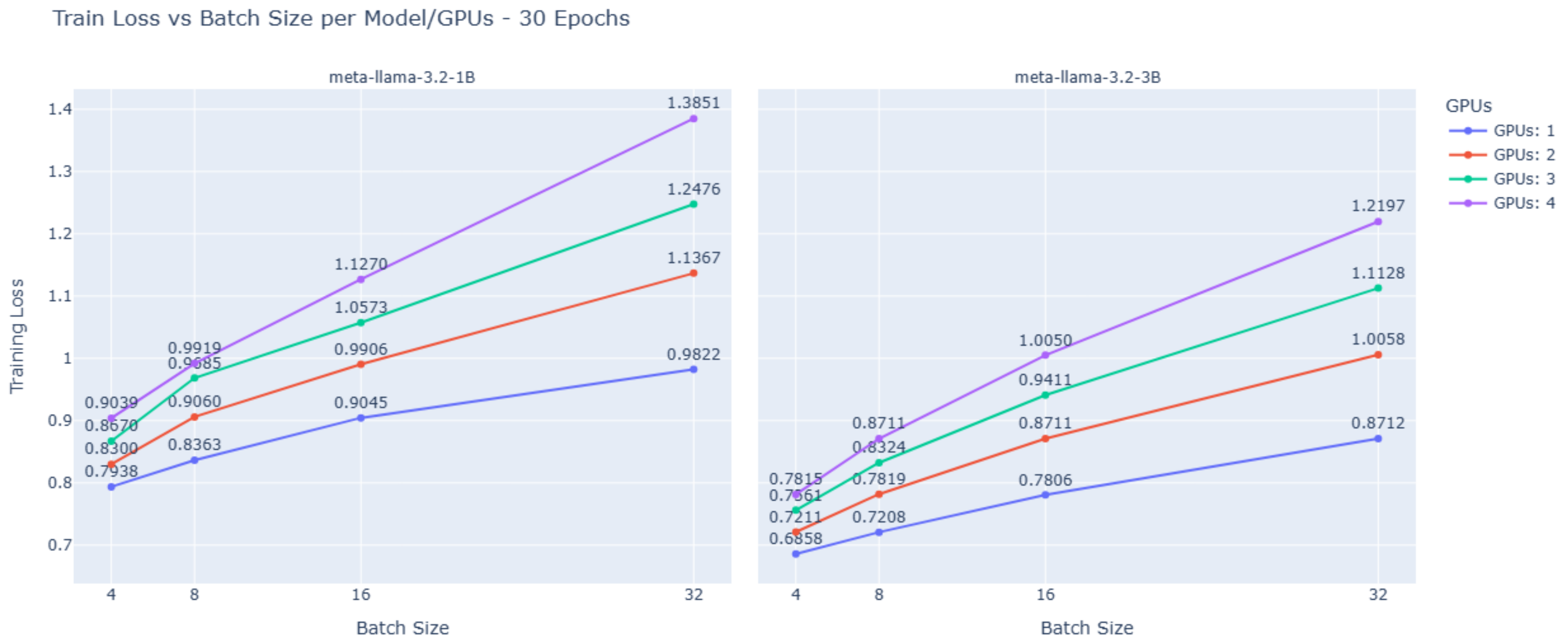
Convergence

GPU Utilization

Memory

Tensorboard

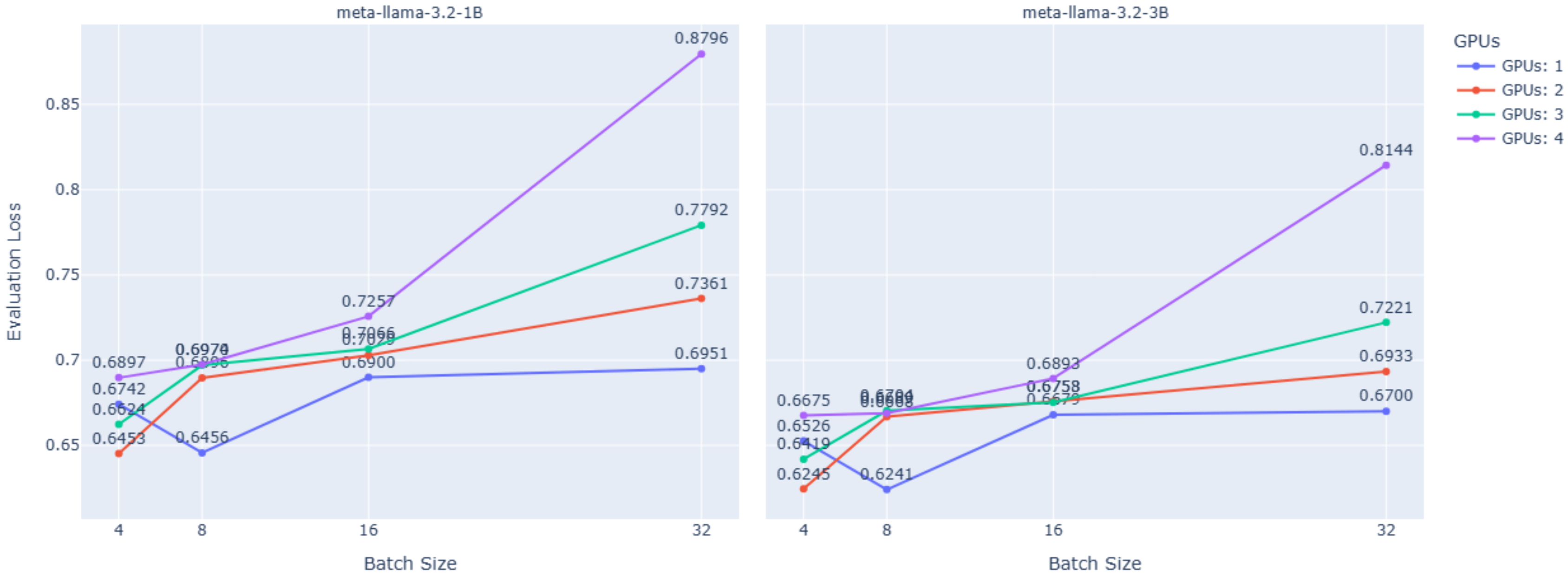
Convergence



Train Loss vs Batch Size per Model/GPUs - 30 Epochs

Convergence

Eval Loss vs Batch Size per Model/GPUs - 30 Epochs



Contents

Model training experiments

Performance

Efficiency

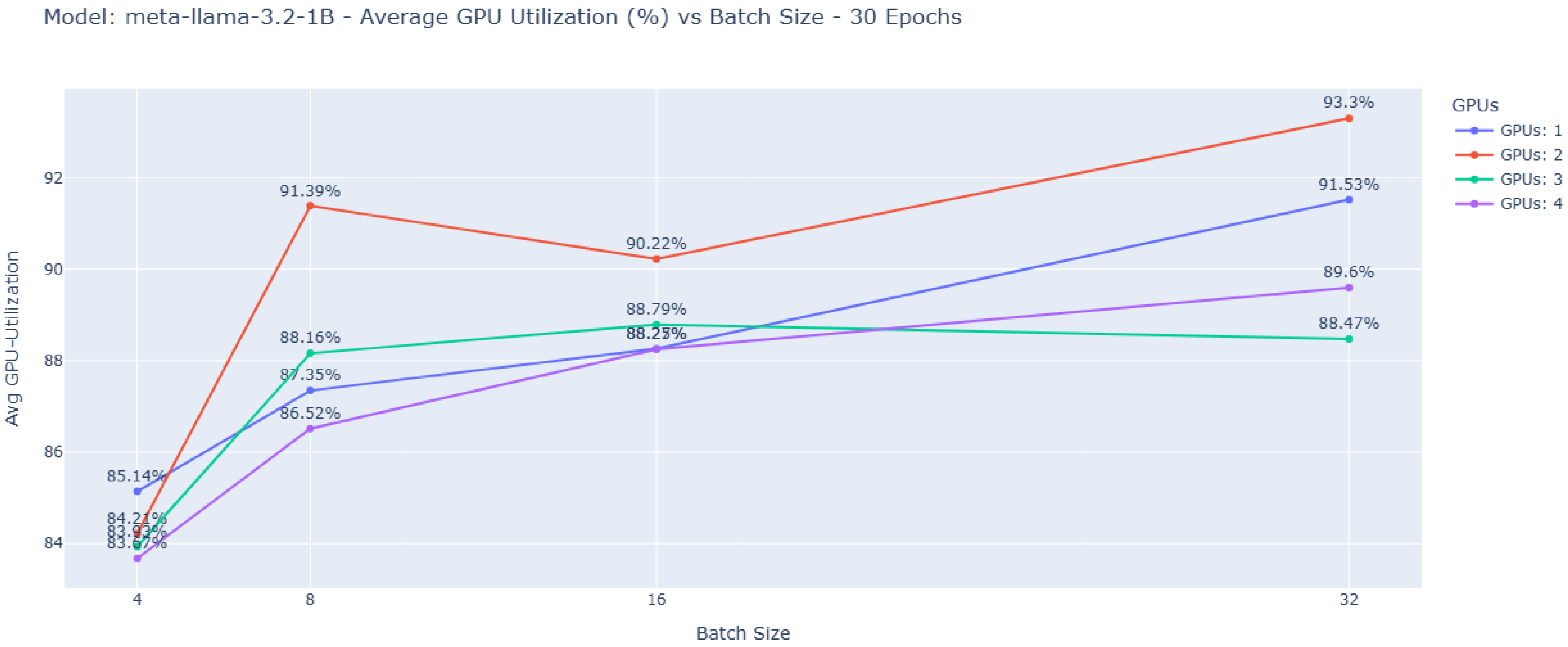
Convergence

GPU Utilization

Memory

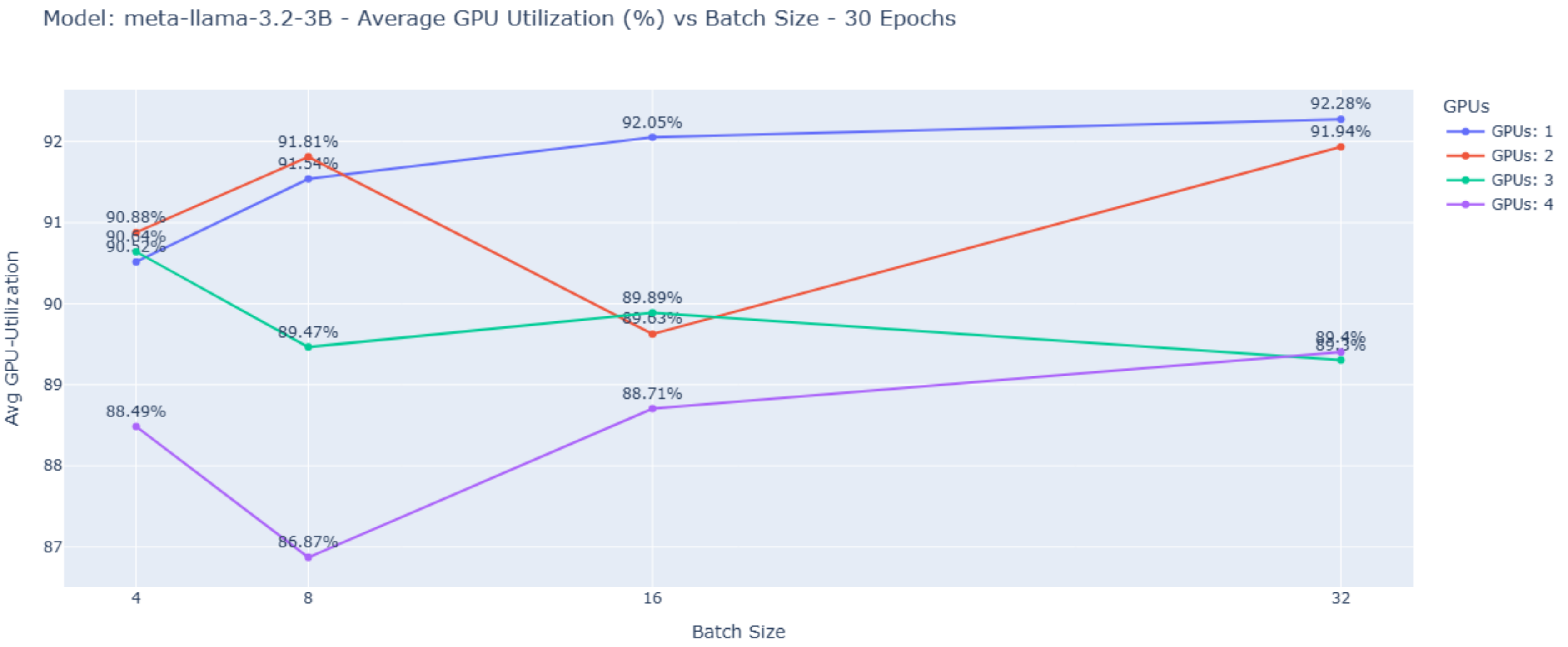
Tensorboard

GPU Utilization – nvidia-smi 1B



GPU Utilization – nvidia-smi 3B

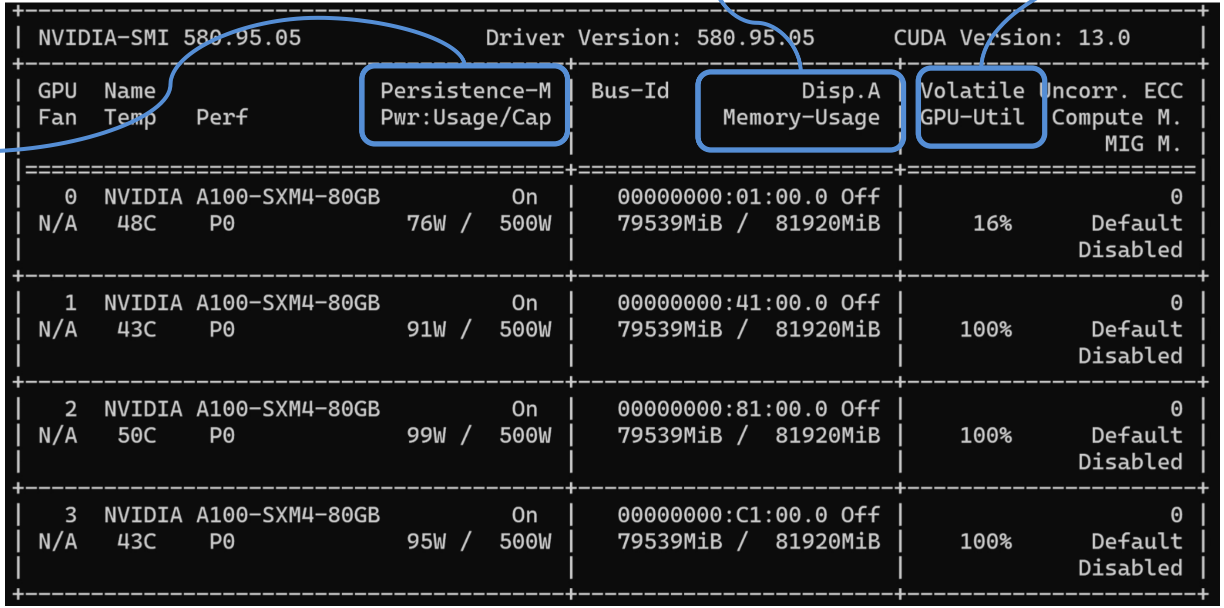
PHAROS
THE GREEK AI FACTORY



nvidia-smi

The amount of memory currently in use versus the total available memory (in MiB).

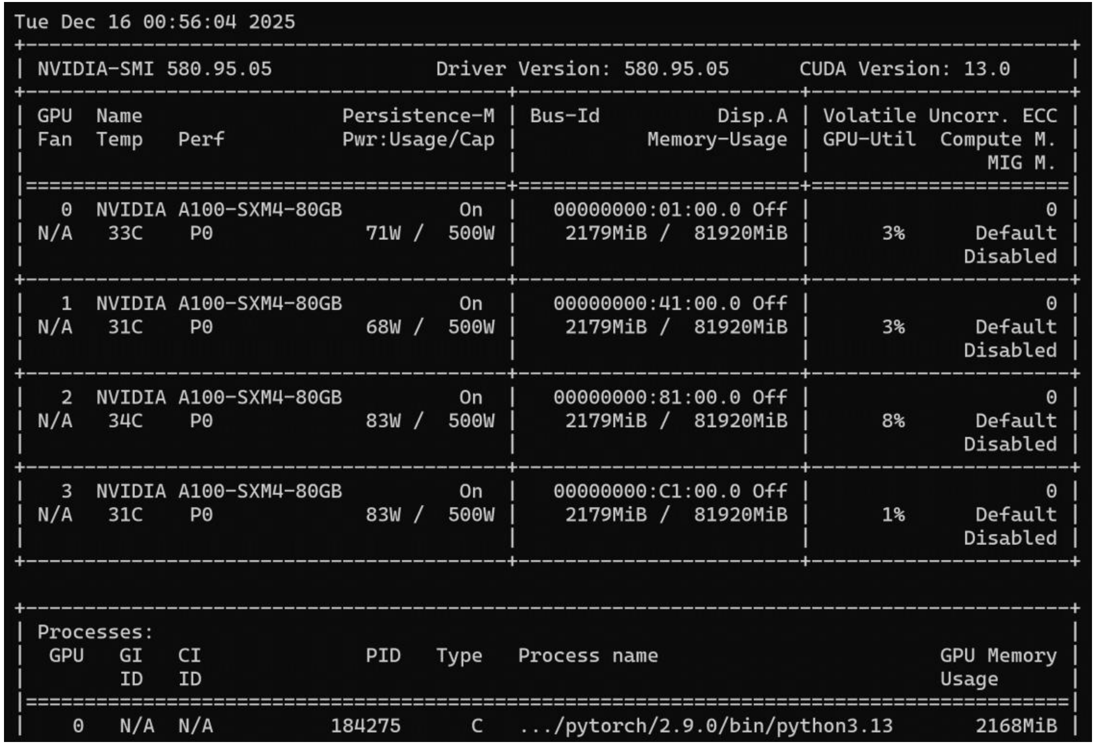
The current power draw versus the maximum power limit (in Watts).



```
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 580.95.05              Driver Version: 580.95.05      CUDA Version: 13.0     |
+-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA A100-SXM4-80GB          On  |   00000000:01:00.0 Off |                    0 |
| N/A   48C    P0             76W /  500W |  79539MiB /  81920MiB  |     16%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   1  NVIDIA A100-SXM4-80GB          On  |   00000000:41:00.0 Off |                    0 |
| N/A   43C    P0             91W /  500W |  79539MiB /  81920MiB  |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   2  NVIDIA A100-SXM4-80GB          On  |   00000000:81:00.0 Off |                    0 |
| N/A   50C    P0             99W /  500W |  79539MiB /  81920MiB  |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   3  NVIDIA A100-SXM4-80GB          On  |   00000000:C1:00.0 Off |                    0 |
| N/A   43C    P0             95W /  500W |  79539MiB /  81920MiB  |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
```

The percentage of the GPU's processing capacity being utilized for computation.

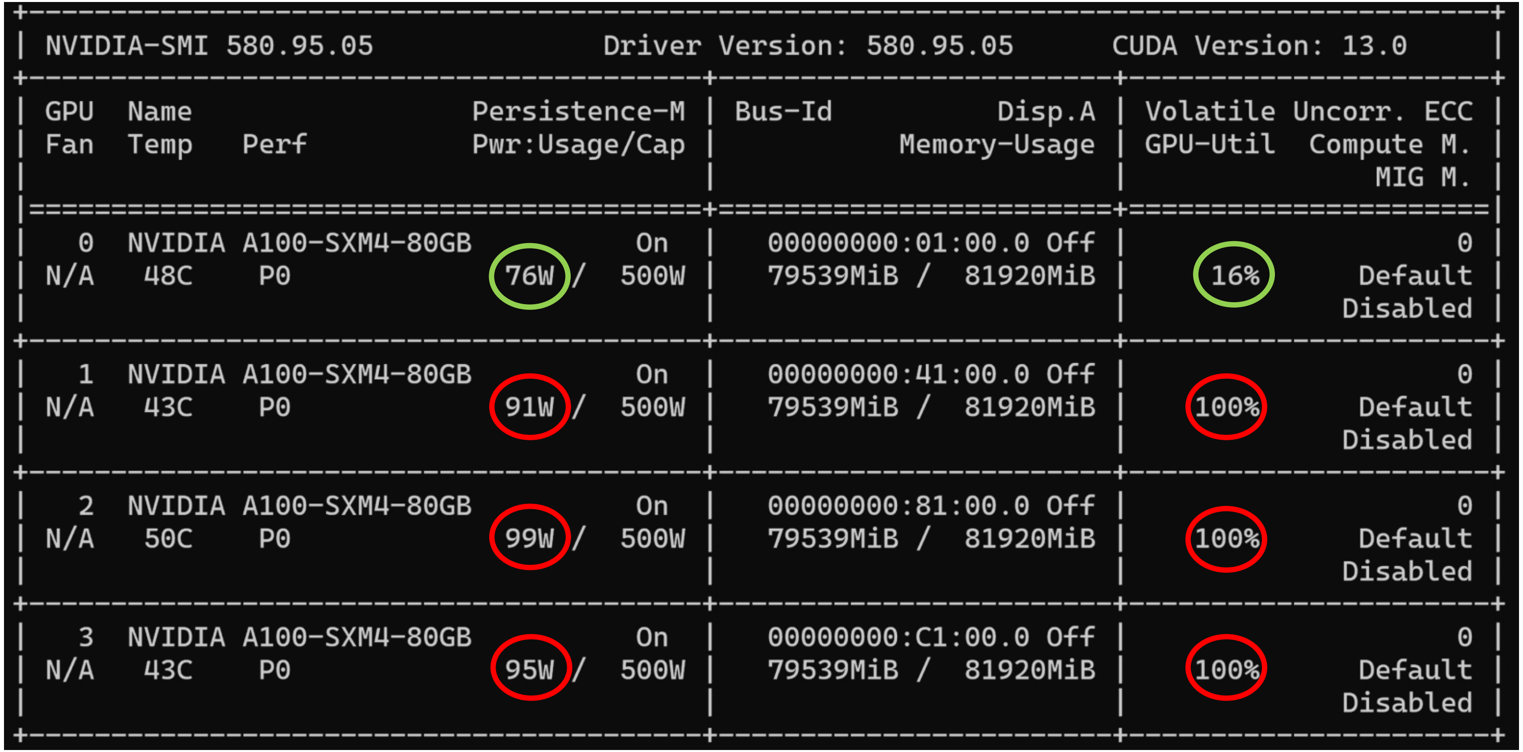
nvidia-smi



```
Tue Dec 16 00:56:04 2025
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 580.95.05              Driver Version: 580.95.05      CUDA Version: 13.0     |
+-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA A100-SXM4-80GB          On  |   00000000:01:00.0 Off |                    0 |
| N/A   33C    P0             71W /  500W |    2179MiB /  81920MiB |      3%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   1  NVIDIA A100-SXM4-80GB          On  |   00000000:41:00.0 Off |                    0 |
| N/A   31C    P0             68W /  500W |    2179MiB /  81920MiB |      3%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   2  NVIDIA A100-SXM4-80GB          On  |   00000000:81:00.0 Off |                    0 |
| N/A   34C    P0             83W /  500W |    2179MiB /  81920MiB |      8%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   3  NVIDIA A100-SXM4-80GB          On  |   00000000:C1:00.0 Off |                    0 |
| N/A   31C    P0             83W /  500W |    2179MiB /  81920MiB |      1%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI              PID   Type   Process name                        GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|    0   N/A  N/A          184275      C   .../pytorch/2.9.0/bin/python3.13      2168MiB |
```

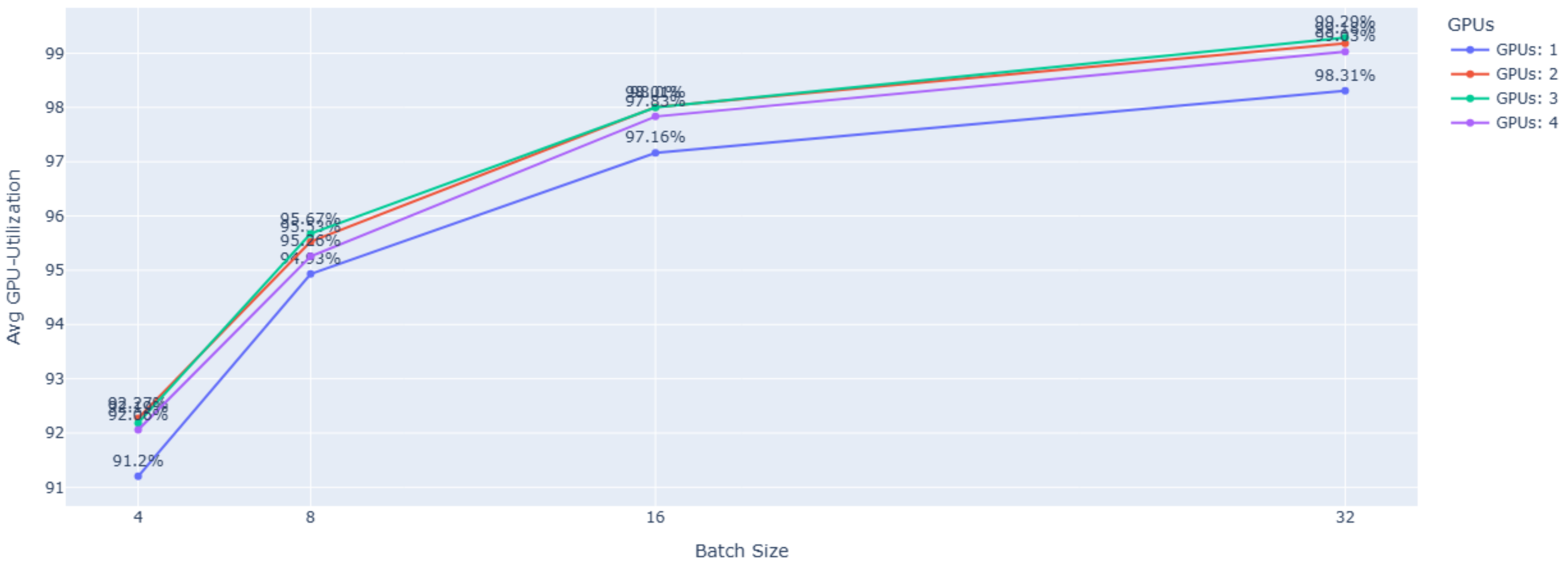
nvidia-smi



```
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 580.95.05              Driver Version: 580.95.05      CUDA Version: 13.0     |
+-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA A100-SXM4-80GB          On  |   00000000:01:00.0 Off |                    0 |
| N/A   48C    P0             76W /  500W |   79539MiB /  81920MiB |     16%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   1  NVIDIA A100-SXM4-80GB          On  |   00000000:41:00.0 Off |                    0 |
| N/A   43C    P0             91W /  500W |   79539MiB /  81920MiB |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   2  NVIDIA A100-SXM4-80GB          On  |   00000000:81:00.0 Off |                    0 |
| N/A   50C    P0             99W /  500W |   79539MiB /  81920MiB |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   3  NVIDIA A100-SXM4-80GB          On  |   00000000:C1:00.0 Off |                    0 |
| N/A   43C    P0             95W /  500W |   79539MiB /  81920MiB |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
```

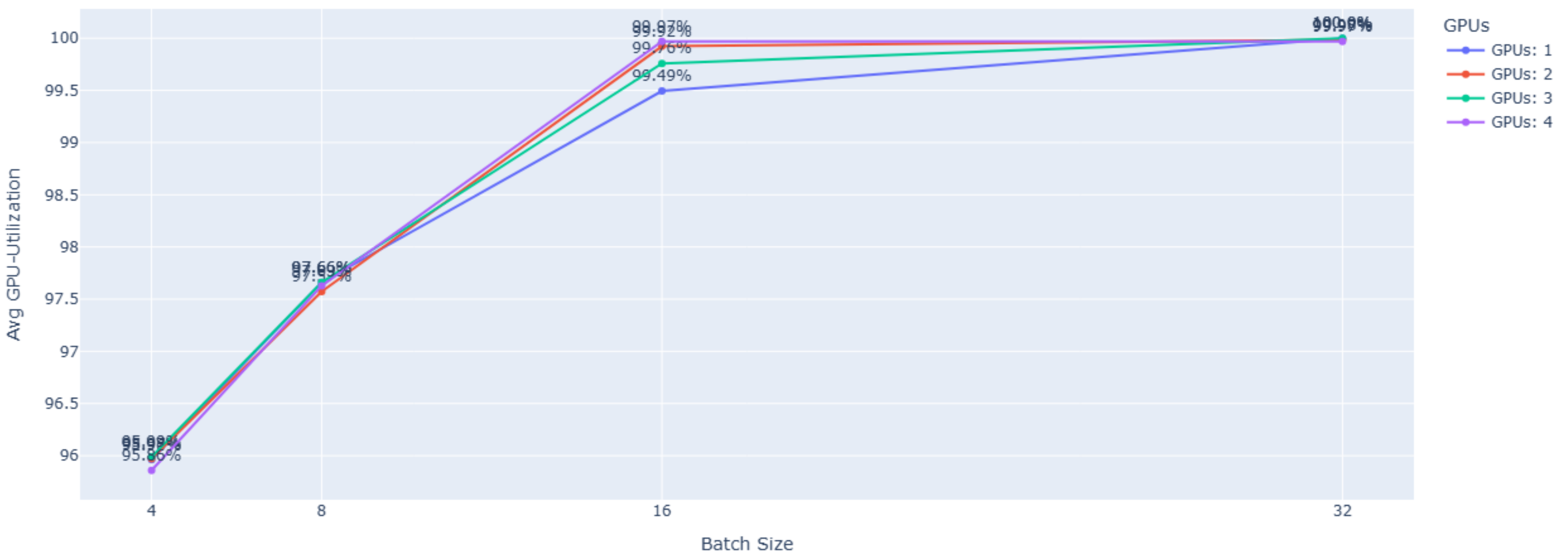
GPU Utilization – pynvml 1B

Model: meta-llama-3.2-1B - Average GPU Utilization (%) vs Batch Size Epochs per Model/GPUs - 30 Epochs



GPU Utilization – pynvml 3B

Model: meta-llama-3.2-3B - Average GPU Utilization (%) vs Batch Size Epochs per Model/GPUs - 30 Epochs



Contents

PHAROS
THE GREEK AI FACTORY

Model training experiments

Performance

Efficiency

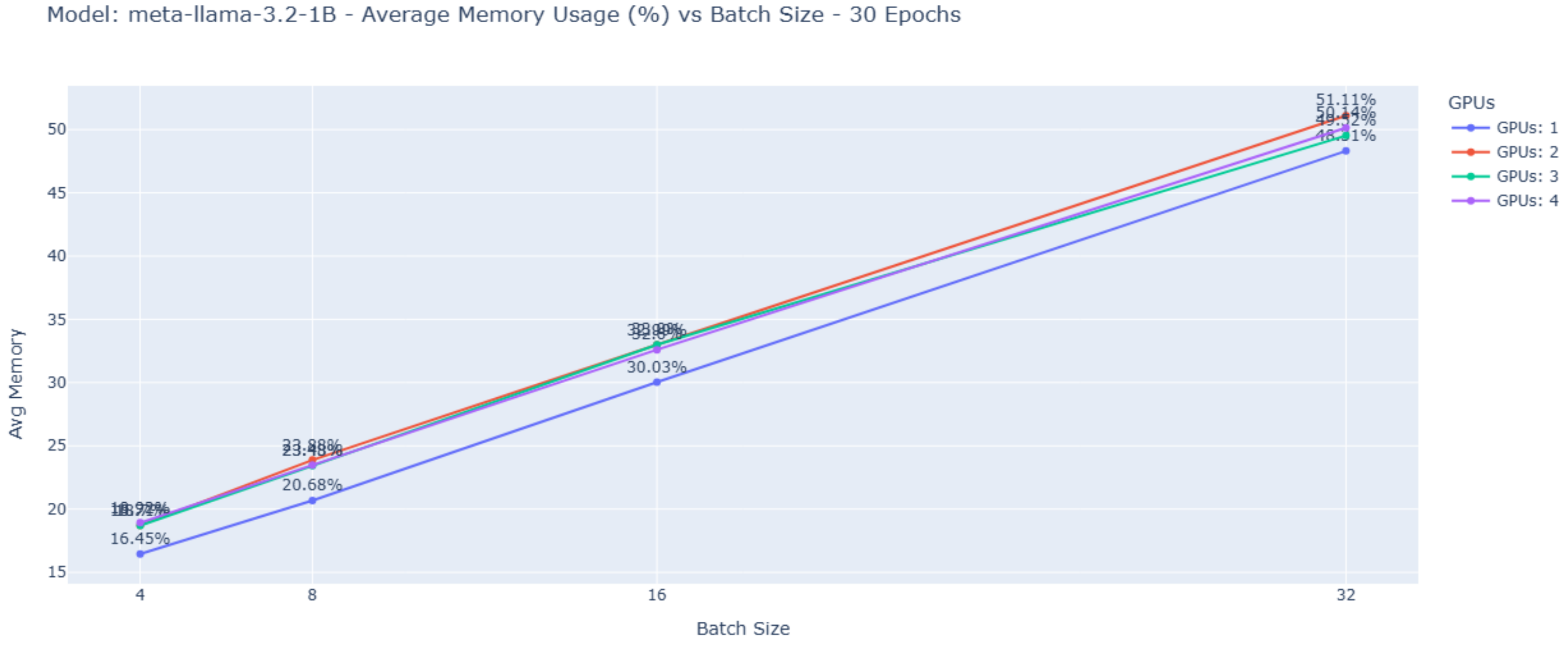
Convergence

GPU Utilization

Memory

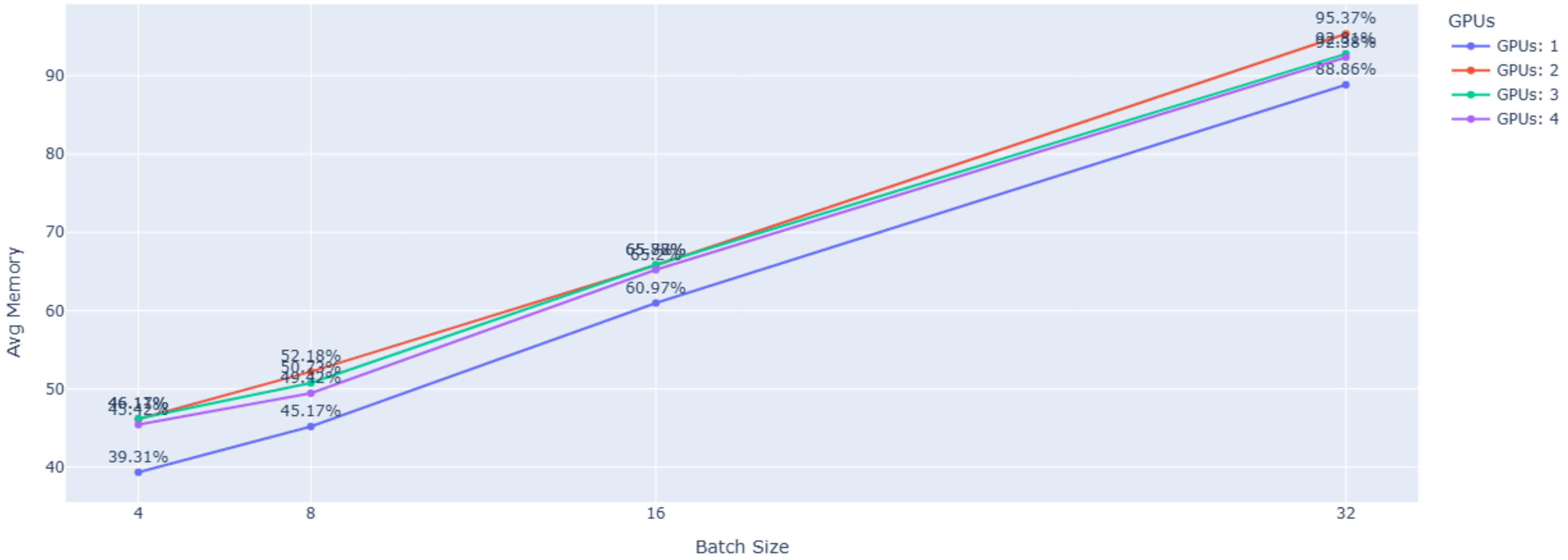
Tensorboard

Memory Usage – nvidia-smi 1B



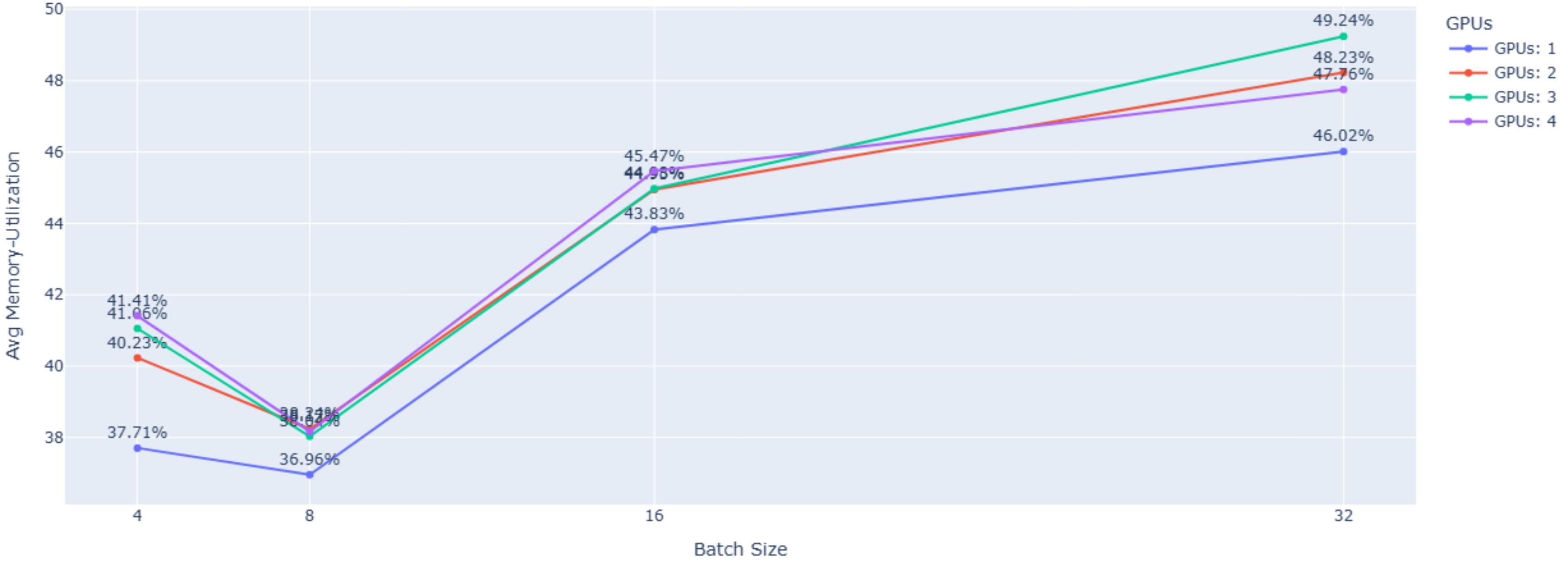
Memory Usage – nvidia-smi 3B

Model: meta-llama-3.2-3B - Average Memory Usage (%) vs Batch Size - 30 Epochs

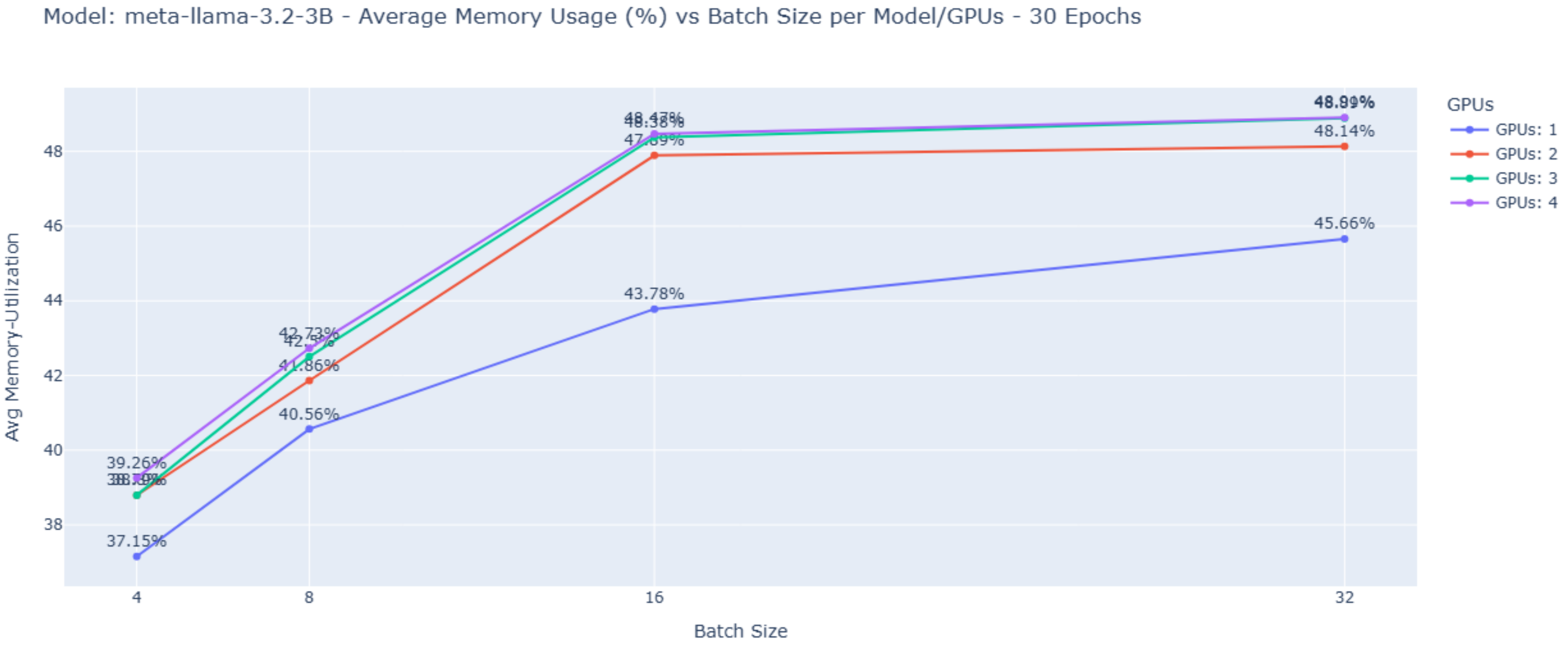


Memory Utilization – pynvml 1B

Model: meta-llama-3.2-1B - Average Memory Usage (%) vs Batch Size per Model/GPUs - 30 Epochs



Memory Utilization – pynvml 3B



Contents

PHAROS
THE GREEK AI FACTORY

Model training experiments

Performance

Efficiency

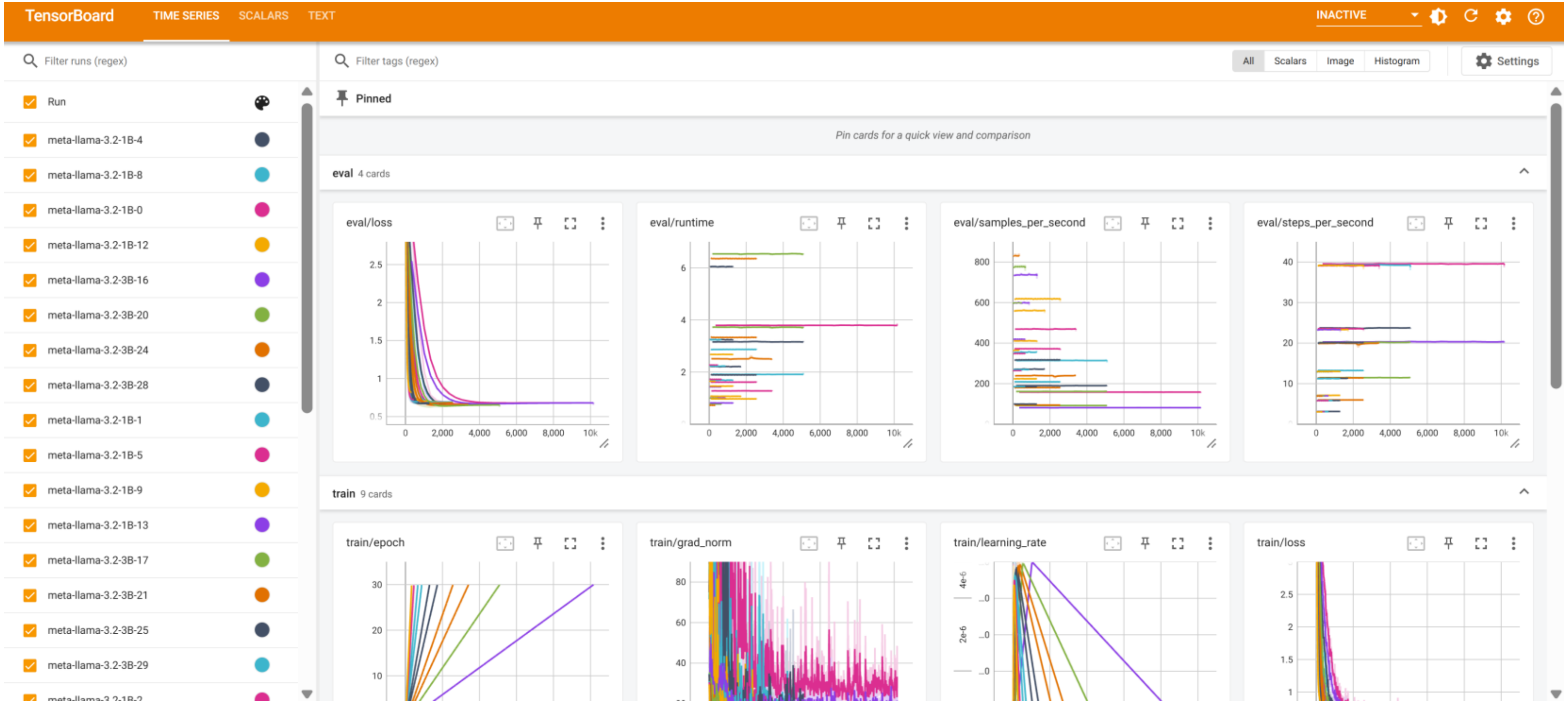
Convergence

GPU Utilization

Memory

Tensorboard

Tensorboard



- ☑ Run
- ☑ meta-llama-3.2-1B-4
- ☑ meta-llama-3.2-1B-8
- ☑ meta-llama-3.2-1B-0
- ☑ meta-llama-3.2-1B-12
- ☑ meta-llama-3.2-3B-16
- ☑ meta-llama-3.2-3B-20
- ☑ meta-llama-3.2-3B-24
- ☑ meta-llama-3.2-3B-28
- ☑ meta-llama-3.2-1B-1
- ☑ meta-llama-3.2-1B-5
- ☑ meta-llama-3.2-1B-9
- ☑ meta-llama-3.2-1B-13
- ☑ meta-llama-3.2-3B-17
- ☑ meta-llama-3.2-3B-21
- ☑ meta-llama-3.2-3B-25
- ☑ meta-llama-3.2-3B-29
- ☑ meta-llama-3.2-1B-2

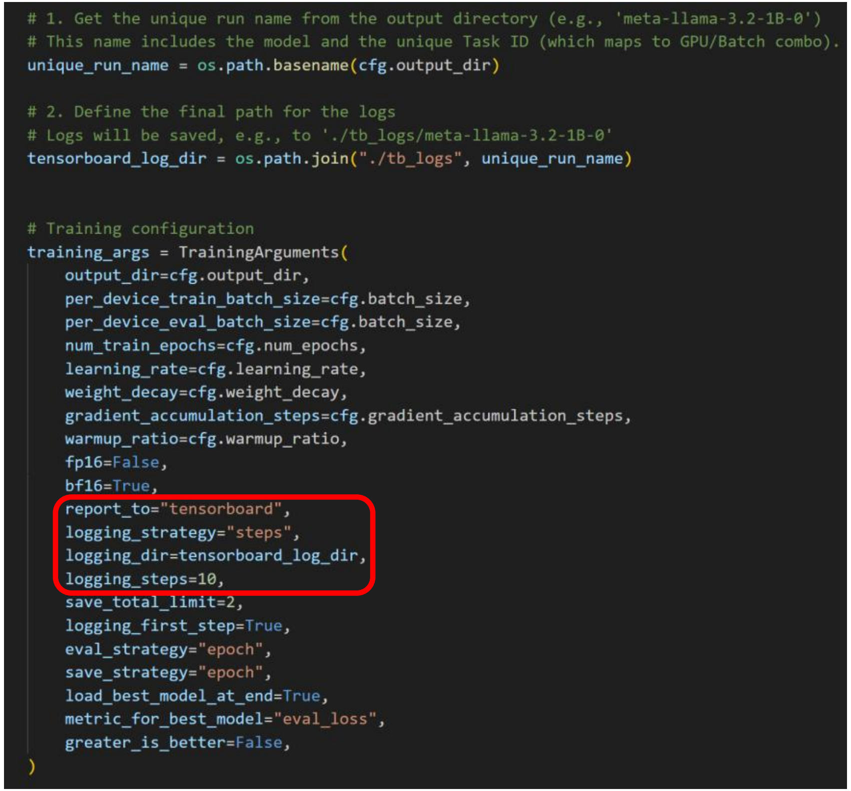
Tensorboard - Parameters

`report_to="tensorboard"` By setting the value to "tensorboard", you instruct the training system to send all gathered metrics (like loss, accuracy, runtime, etc.) to TensorBoard. The system then generates the necessary **event files** that the TensorBoard application reads and visualizes.

`logging_strategy="steps"` With the value set to "steps", the system records the metrics **after a certain number of training steps** are completed. **Note:** This is usually accompanied by another parameter (e.g., logging_steps=X) `logging_steps=10,`

`logging_dir=tensorboard_log_dir` Specifies the local directory where the log files are stored.

`logging_first_step=True` Ensures that metrics are logged **immediately** after the very **first training step** is completed.



```python
# 1. Get the unique run name from the output directory (e.g., 'meta-llama-3.2-1B-0')
# This name includes the model and the unique Task ID (which maps to GPU/Batch combo).
unique_run_name = os.path.basename(cfg.output_dir)

# 2. Define the final path for the logs
# Logs will be saved, e.g., to './tb_logs/meta-llama-3.2-1B-0'
tensorboard_log_dir = os.path.join("./tb_logs", unique_run_name)


# Training configuration
training_args = TrainingArguments(
    output_dir=cfg.output_dir,
    per_device_train_batch_size=cfg.batch_size,
    per_device_eval_batch_size=cfg.batch_size,
    num_train_epochs=cfg.num_epochs,
    learning_rate=cfg.learning_rate,
    weight_decay=cfg.weight_decay,
    gradient_accumulation_steps=cfg.gradient_accumulation_steps,
    warmup_ratio=cfg.warmup_ratio,
    fp16=False,
    bf16=True,
    report_to="tensorboard",
    logging_strategy="steps",
    logging_dir=tensorboard_log_dir,
    logging_steps=10,
    save_total_limit=2,
    logging_first_step=True,
    eval_strategy="epoch",
    save_strategy="epoch",
    load_best_model_at_end=True,
    metric_for_best_model="eval_loss",
    greater_is_better=False,
)
```

Thank you for your attention.

Questions?

EuroHPC Joint Undertaking

Co-funded by the European Union

AI Factories PHAROS GREECE

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Ψηφιακής Διακυβέρνησης
και Τεχνητής Νοημοσύνης