LLMs on Aurora: Overview

AuroraGPT

Sam Foreman
foremans@anl.gov
ALCF

2025-05-21

ALCF INCITE GPU HACKATHON

May 20–22, 2025

LLMs on Aurora: AuroraGPT

Sam Foreman

2025-05-21

ALCF Incite Hackathon 2025

- [2025 ALCF INCITE GPU Hackathon (20-May 22, 2025)]
- LLMs on Aurora[1]:
  - 🍋 Hands-On: ezpz
  - 🌌 Overview: AuroraGPT

1. *my* talks can be found at: https://samforeman.me/talks/incite-hackathon-2025

🎯 AuroraGPT: Goals

> **[AuroraGPT](#)**: *General purpose scientific LLM*
> Broadly trained on a general corpora plus scientific {papers, texts, data}

- **Explore pathways** towards a "Scientific Assistant" model
- **Build with international partners** (RIKEN, BSC, others)
- **Multilingual** English, , French, German, Spanish
- **Multimodal**: images, tables, equations, proofs, time series, graphs, fields, sequences, etc

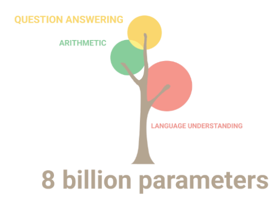


Figure 1: Image from Hannibal046 / `Awesome-LLM`

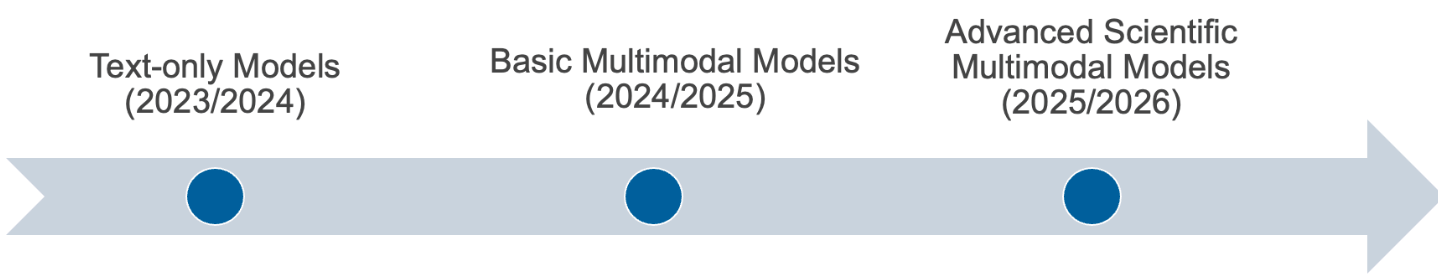


Figure 2: Credit to the entire AuroraGPT team for slides.

🦙 Issues with "Publicly Available" LLMs

- **Trust** and **Safety**:
  - Skepticism about deployment in critical infrastructure
  - Correctness and reliability of model outputs
- **Transparency**:
  - Data governance, *what was used for pre-training*? fine-tuning?
    - **generally unknown**
  - What is *open source*?
    - Model weights?
    - Pre-training {code, logs, metrics} ?

🧪 AuroraGPT: Open Science Foundation Model

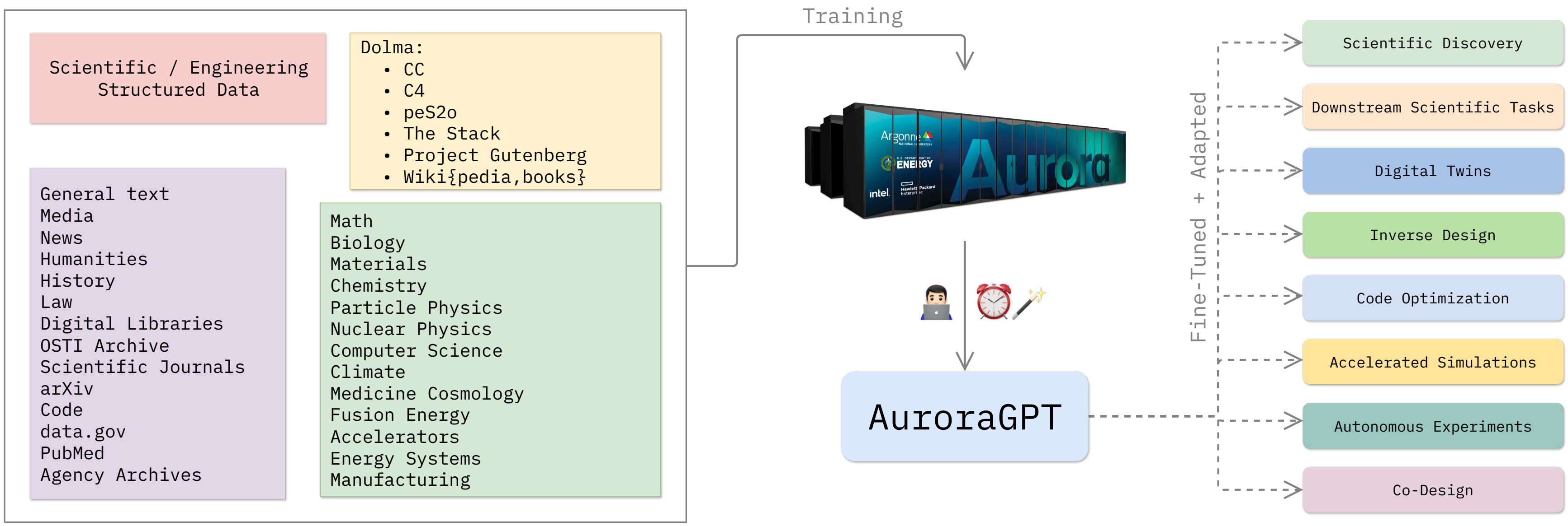


Figure 3: High-level overview of AuroraGPT project

📊 AuroraGPT: Outcomes

- **Datasets and data pipelines** for preparing science training data
- **Software infrastructure and workflows** to train, evaluate and deploy LLMs at scale for scientific resarch purposes
  - argonne-lcf/Megatron-DeepSpeed
    End-to-end training and inference, on *any* GPU cluster
  - argonne-lcf/inference-endpoints
    Inference endpoints for LLMs, hosted @ ALCF
- **Evaluation of state-of-the-art LLM Models**:
  - Determine where they fall short in deep scientific tasks
  - Where deep data may have an impact

📚 What do we hope to get?

- **Assessment of the approach** of augmenting web training data with two forms of data specific to science:
  - Full text scientific papers
  - Structured scientific datasets (suitably mapped to narrative form)
- **Research grade artifacts** (**models**) for scientific community for adaptation for downstream uses[1]
- **Promotion of responsible AI** best practices where we can figure them out
- **International Collaborations** around the long term goal of *AGI for science*

1. (Dharuman et al. 2024)

Aurora

Table 1: Aurora Specs

Racks	166
Nodes	10,624
CPUs	21,248
GPUs	63,744
NICs	84,992
HBM	8 PB
DDR5c	10 PB

Figure 4: Aurora: [Fact Sheet](#).

🏆 [Fastest AI system in the world](#)

🤖 ALCF AI Testbed

- ALCF AI Testbed Systems are in production and [available for allocations](#) to the research community
- Significant improvement in time-to-solution and energy-efficiency for diverse AI for science applications.
- [NAIRR Pilot](#)

Up to $\approx$ **25** $\times$ throughput improvement for genomic FMs with **6.5** $\times$ energy efficiency

Figure 5: **SambaNova SN-30** 2nd Gen, 8 nodes with 64 AI Accelerators

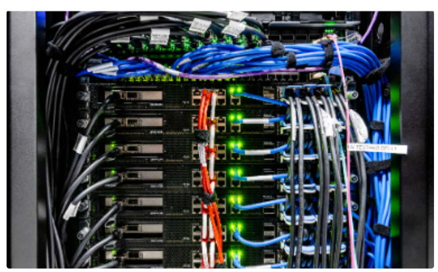


Figure 6: **Graphcore Bow**: Pod-64 configuration with 64 accelerators

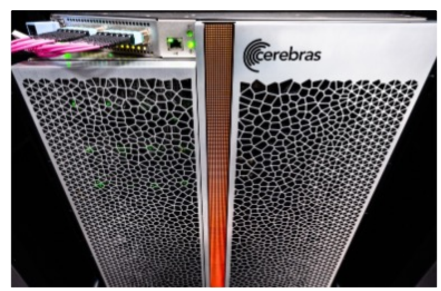


Figure 7: **Cerebras**: 2x CS-2 WSE with Memory-X and Swarm-X technologies

Figure 8: **GroqRack**: 9 nodes, 8 GroqChip v1.5 Tensor streaming processors accelerators per node

👥 Team Leads

Planning


Rick Stevens[1]


Ian Foster


Rinku Gupta


Mike Papka


Arvind Ramanathan


Fangfang Xia

Data


Ian Foster


Robert Underwood

Training


Venkat Vishwanath


Sam Foreman

Evaluation


Franck Cappello


Sandeep Madireddy


Bo Li

Post


Eliu Huerta


Azton Wells

Inference


Rajeev Thakur

Comms


Charlie Catlett


David Martin

Distribution


Brad Ullrich

1. Lead

🤝 Teams

- **Planning**
- **Data Prep**
  - Accumulate 20+ T tokens of high-quality scientific text and structured data
- **Models / Training**[1]
  - Train (entirely from scratch) a series of models on publicly available data
- **Evaluation**
  - Skills, trustworthiness, safety, robustness, privacy, machine ethics
- **Post-Training**
  - Fine-tuning, alignment
- **Inference**
  - Model serving, API development / public-facing web services
- **Distribution**
  - Licensing, generating and distributing artifacts for public consumption
- **Communication**

1. Co-led by: Venkat Vishwanath, **Sam Foreman**

📚 Data

✅ **Goal**: Assemble a large corpus of documents (general and scientific) to train and fine-tune AuroraGPT models

- **Challenges**: Avoid / detect contamination with benchmarks
  - Respect copyright (ACM Digital Library), privacy, and ethical considerations
- **Performance Challenges**: *High throughput* data processing
  - Converting PDF $\rightarrow$ text (math formula, figures)
  - Convert science information (data) into text (narratives)
  - De-duplication (syntactic and semantic) of scientific documents (to avoid memorization, bias)
- **Quantity**: Considering 20+ Trillion tokens $\rightarrow \approx$ 100M papers
- **Domains**: All (long-term) scientific domains, starting with:
  - Material science, Physics, Biology, Computer Science, Climate Science

⏱️ Dataset Processing

- To train a fixed model on trillions of tokens requires:
  1. **Aggregating** data from multiple different *corpora*
     (e.g. ArXiv, Reddit, StackExchange, GitHub, Wikipedia, etc.)
  2. **Sampling** *each training batch* according to a fixed distribution across corpora
  3. **Building** indices that map batches of tokens into these files (indexing)

> The original implementation was *slow*:
>
> - Designed to run *serially* on a **single device**
> - **Major bottleneck** when debugging data pipeline at scale

🚀 Accelerating Dataset Processing: Results

- Original implementation:
  - **Slow**!
  - 🐌 ~ 1 hr/2T tokens
- ☑️ Fix:
  - Wrote *asynchronous*, **distributed** implementation
  - *significantly* improves performance (**30x** !!)
  - 🏎️💨 ~ **2 min**/2T tokens

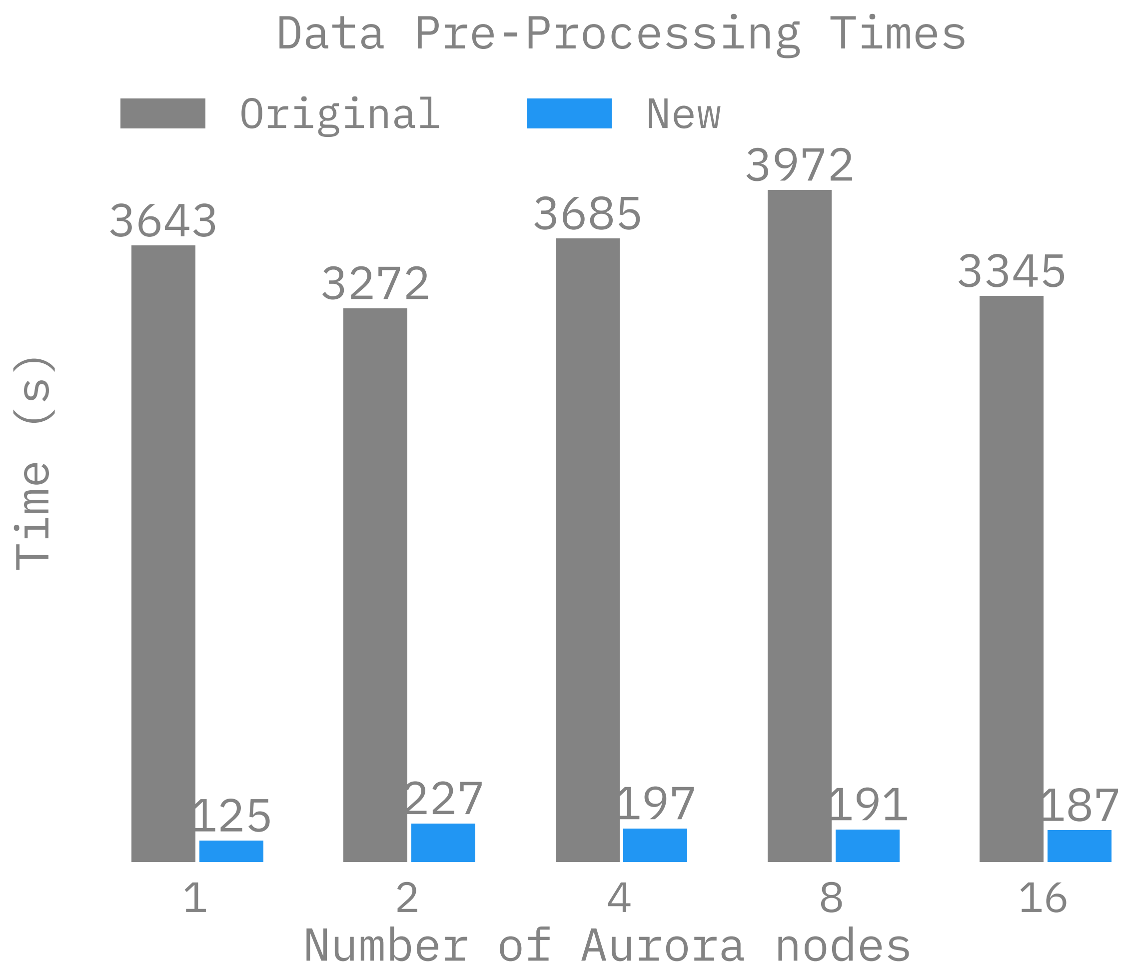


Figure 9: Time spent preparing 2T tokens

🦜 Model Training

✅ Goals

- Want training runs at scale to be:
  - efficient
  - stable
  - reproducible
- This requires:
  - robust data pipelines / file IO
  - effectively overlapping compute with communication
  - stability across {network, filesystem, machine}
- 3D / Multi-dimensional Parallelism strategies
- Large batch training
- Second order optimizers
- Sub-quadratic attention
- State space models
- *Highly optimized GPU kernels*

❌ Challenges

- *Looong time* to train, can be:
  - weeks (even months) of continuous training
  - order of magnitude longer than typical NN training jobs
- Stability issues:
  - failures are expensive (but inevitable)
  - stragglers common at scale
- Individual jobs are:
  - **fragile**
  - only as good as the worst rank
  - one hang or bad worker can crash job
  - network / filesystem / other-user(s) dependent
- Cost / benefits of different collective communication algorithms
  - depend on optimized / efficient implementations
- Network performance
- *Highly optimized GPU kernels*

argonne-lcf / Megatron-DeepSpeed

📉 Loss Curve: Training AuroraGPT-7B on 2T Tokens

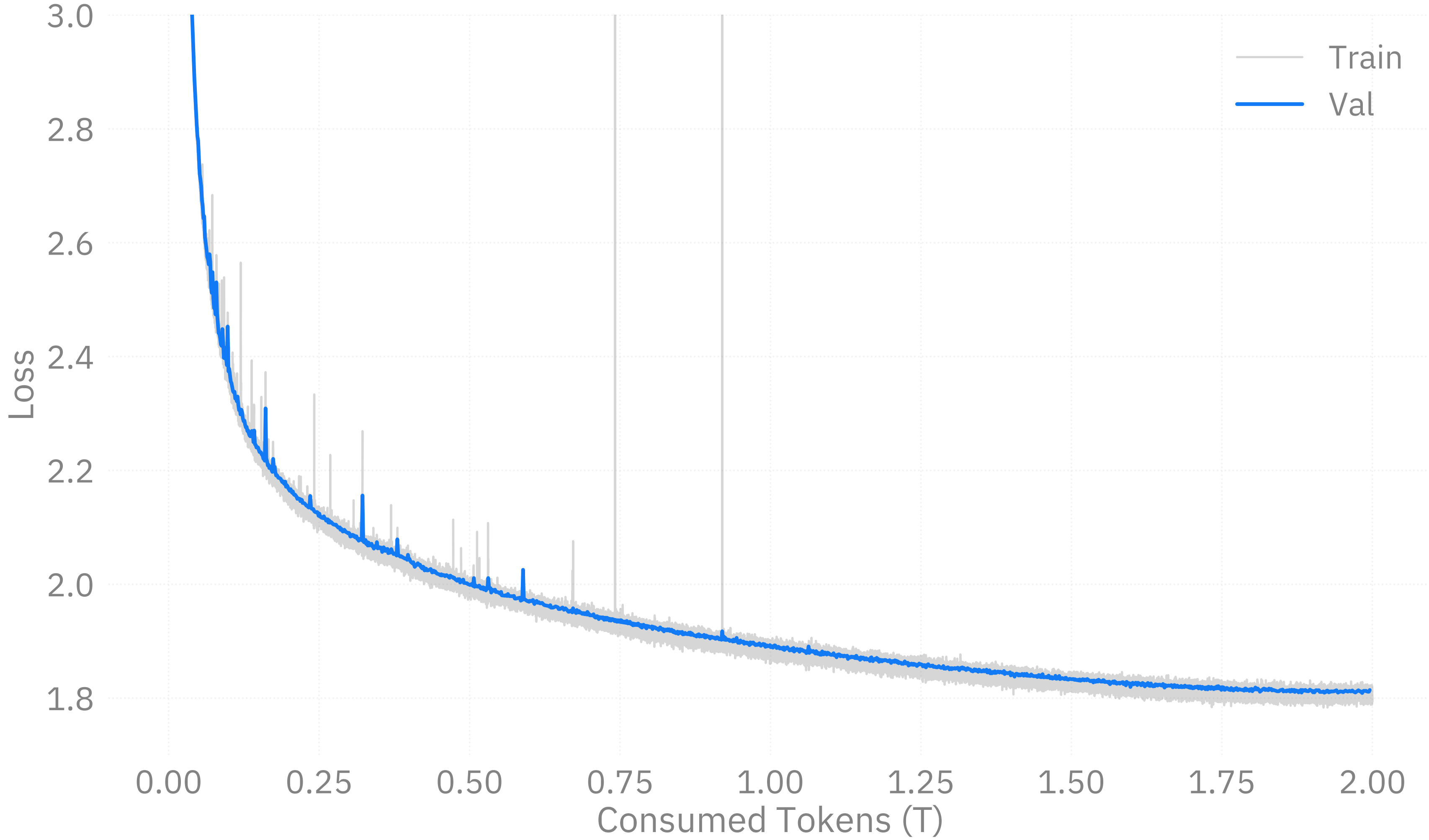


Figure 10: Loss curve during training on 2T tokens.

🤔 Evaluating FM Skills for Science

- What to measure?
  - **Knowledge Extraction, Retrieval, Distillation, Synthesis**: LLM is provided a question or instruction and a truthful answer is expected
  - **Text Grounded**: Answers are expected to be fully grounded on peer-reviewed references to support responses
  - **Reasoning**: LLMs are expected to solve deductive (prove a theory or hypothesis from formal logic and observations), inductive (validate / explain observations from theories) problems
  - **Creativity**: A creative answer is expected from a question or instruction
    - thoughtful dialogue, coding, etc.

⚖️ Evaluating FM Skills for Science: Criteria

- Criteria for all of the above:
  - **Correctness** of facts
  - **Accuracy** of solutions and inferences
  - **Reliability** consistently good in quality or performance
  - **Speed** how fast to produce a response
  - **# shots** how many examples are needed for good quality
    - Extent of *prompt engineering*

🧬 MProt-DPO: Scaling Results

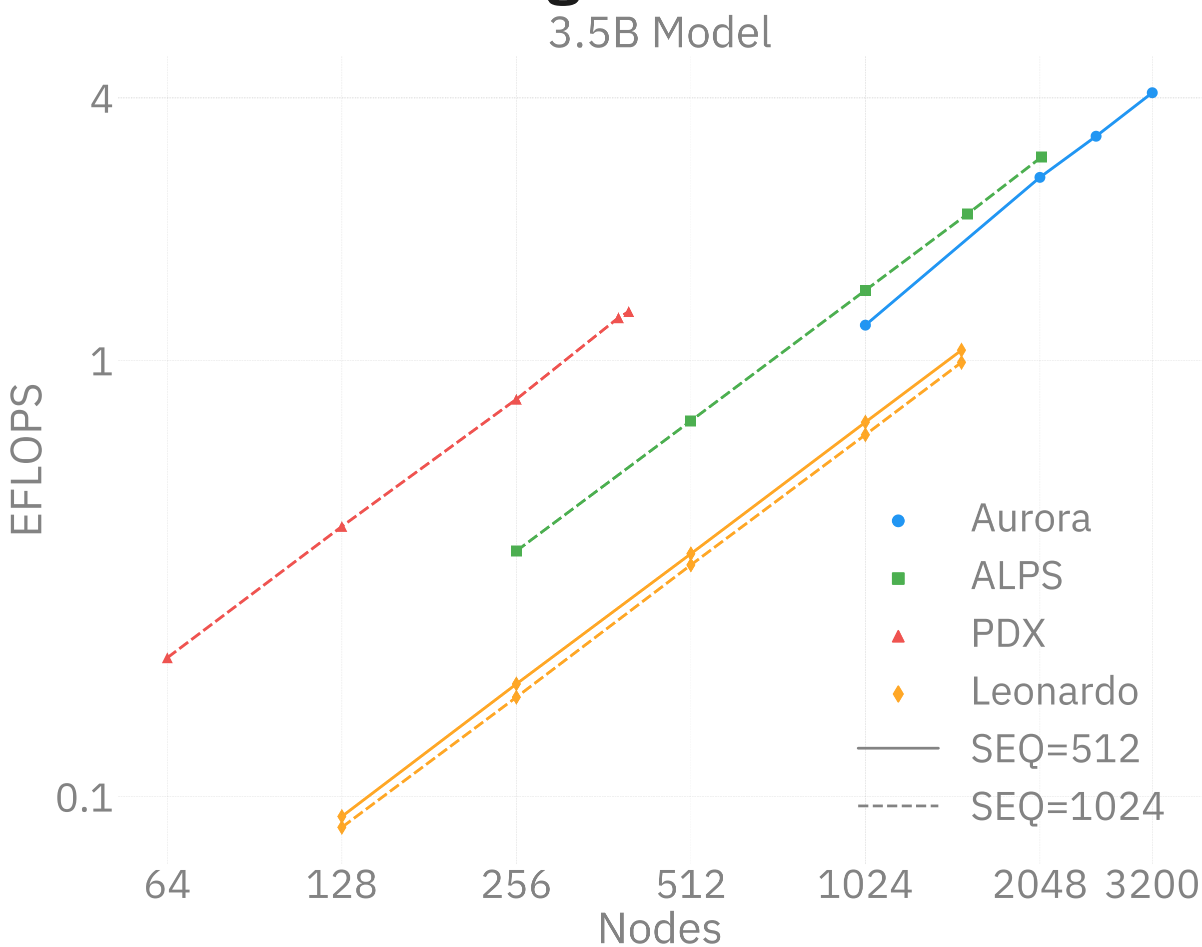


Figure 11: Scaling results for `3.5B` model across ~38,400 GPUs

- ~ **4 EFLOPS** @ Aurora
- 38,400 XPUs
  = 3200 [node] x 12 [XPU / node]
- 🔔 Gordon Bell Finalist[1]:
  - [MProt-DPO: Breaking the ExaFLOPS Barrier for Multimodal Protein Design Workflows](#)

1. ([Dharuman et al. 2024](#))

📓 References

- argonne-lcf / `Megatron-DeepSpeed`
  For the largest of large language models.
- saforem2 / `ezpz`
  Distributed training, ezpz. 🍋
- 📊 See my other slides at samforeman.me/talks:
  - LLMs from Scratch
  - Creating Small(~ish) LLMs
  - Parallel Training Techniques
  - LLMs on Polaris
  - Training LLMs at Scale
- 👀 See also:
  - New international consortium for generative AI models for science
  - PyTorch Distributed Overview
  - 🤗 Efficient Training on Multiple GPUs
  - Getting Started - DeepSpeed
  - 🕸 Quality Measures for Dynamic Graph Generative Models (Hosseini et al. 2025)

❤️ Thank you!

- Organizers
- Feel free to reach out!

🙏 **Acknowledgements**

This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

Bibliography

- Refs:
  - Wei et al. (2022)
  - Animations from The Illustrated Transformer

Dharuman, Gautham, Kyle Hippe, Alexander Brace, Sam Foreman, Väinö Hatanpää, Varuni K. Sastry, Huihuo Zheng, et al. 2024. "MProt-DPO: Breaking the ExaFLOPS Barrier for Multimodal Protein Design Workflows with Direct Preference Optimization." In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. SC '24. Atlanta, GA, USA: IEEE Press. https://doi.org/10.1109/SC41406.2024.00013.

Hosseini, Ryien, Filippo Simini, Venkatram Vishwanath, Rebecca Willett, and Henry Hoffmann. 2025. "Quality Measures for Dynamic Graph Generative Models." In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=8bjspmAMBk.

Song, Shuaiwen Leon, Bonnie Kruft, Minjia Zhang, Conglong Li, Shiyang Chen, Chengming Zhang, Masahiro Tanaka, et al. 2023. "DeepSpeed4Science Initiative: Enabling Large-Scale Scientific Discovery Through Sophisticated AI System Technologies." https://arxiv.org/abs/2310.04610.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. 2022. "Emergent Abilities of Large Language Models." https://arxiv.org/abs/2206.07682.

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.

🧬 MProt-DPO: Scaling Results

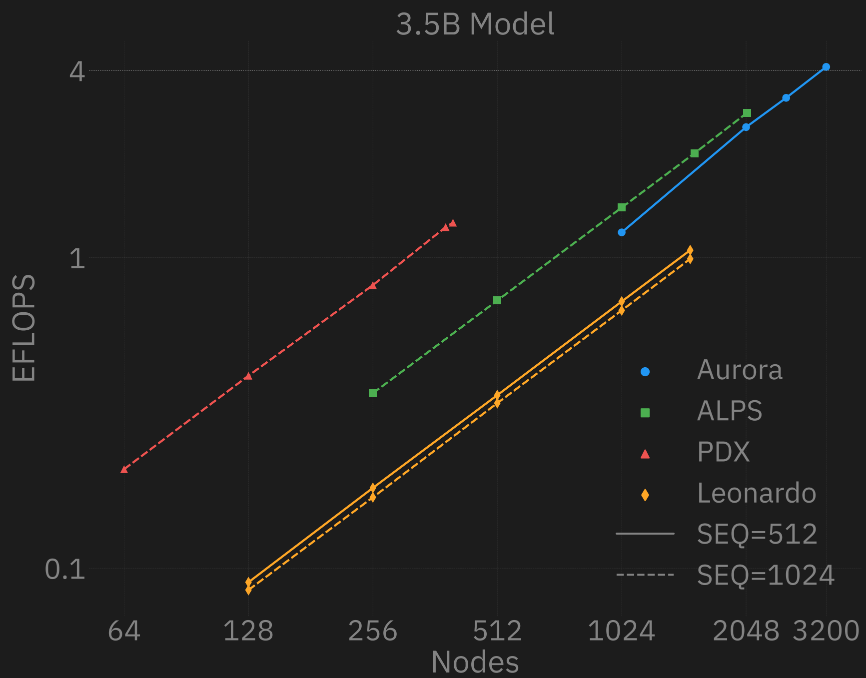


Figure 12: `3.5B` model

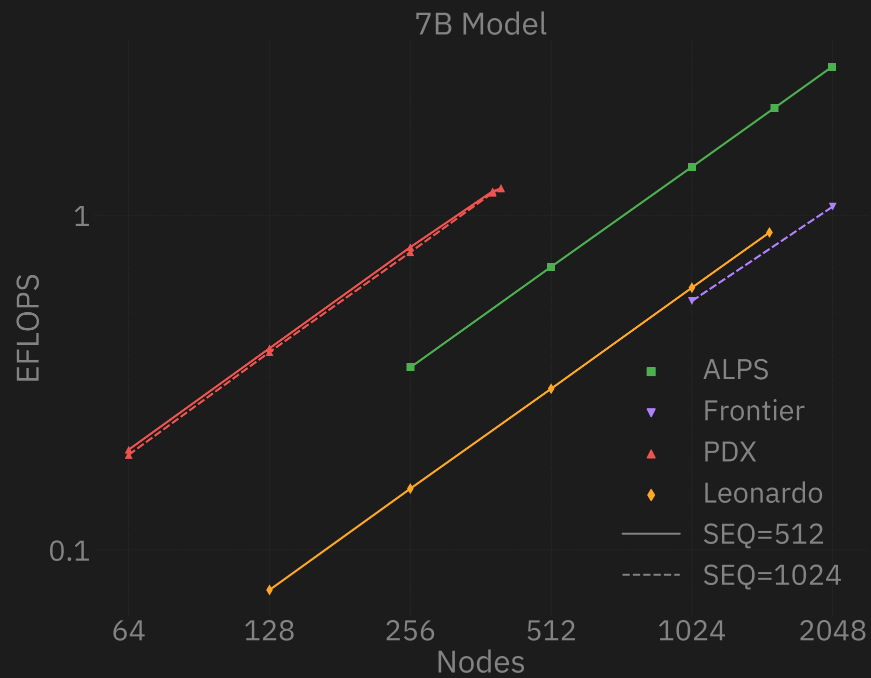
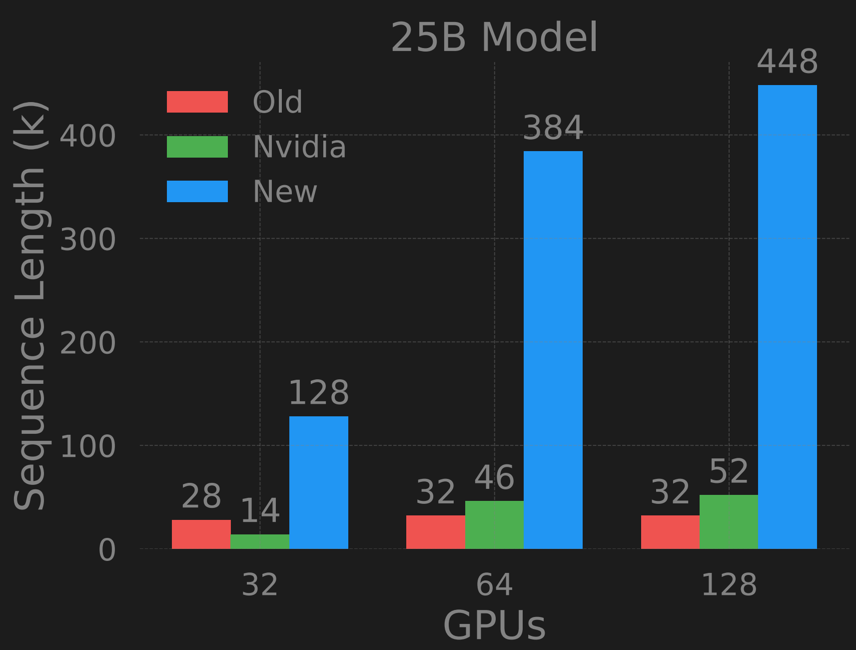


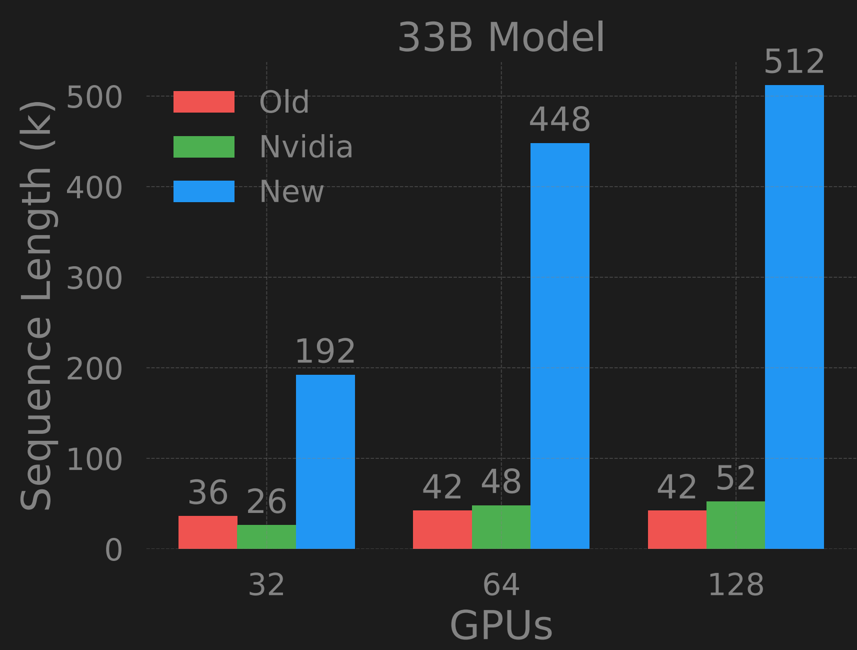
Figure 13: `7B` model

🚂 Looooooooooong Sequence Lengths

- Working with Microsoft/DeepSpeed team to enable longer sequence lengths (context windows) for LLMs
  - See my blog post for additional details



25B



33B

Figure 14: Maximum (achievable) `SEQ_LEN` for both `25B` and `33B` models (See: Song et al. (2023))

scaling4science
Megatron-DS-Benchmarking

♻️ Life Cycle of the LLM

📝 Pre-training | 🎀 Fine-Tuning

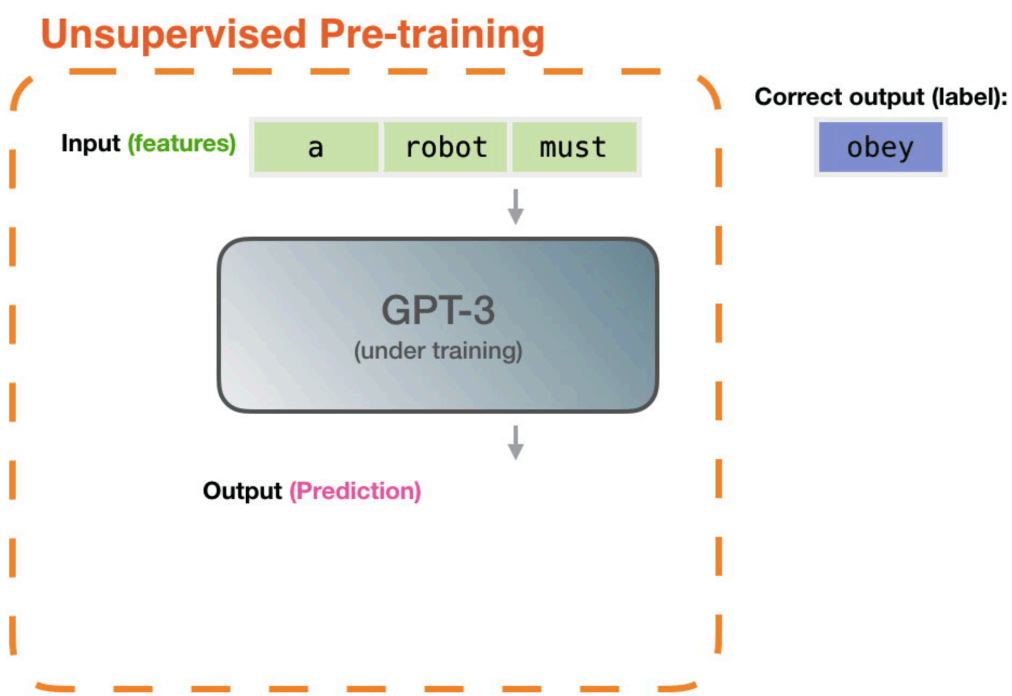


Figure 15: **Pre-training**: Virtually all of the compute used during pretraining phase

🍎 Training LLMs

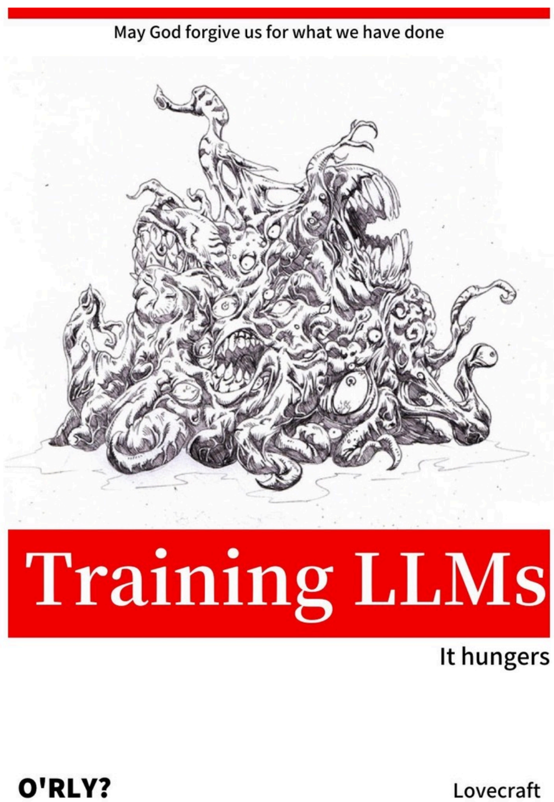


Figure 17: It's hungry!

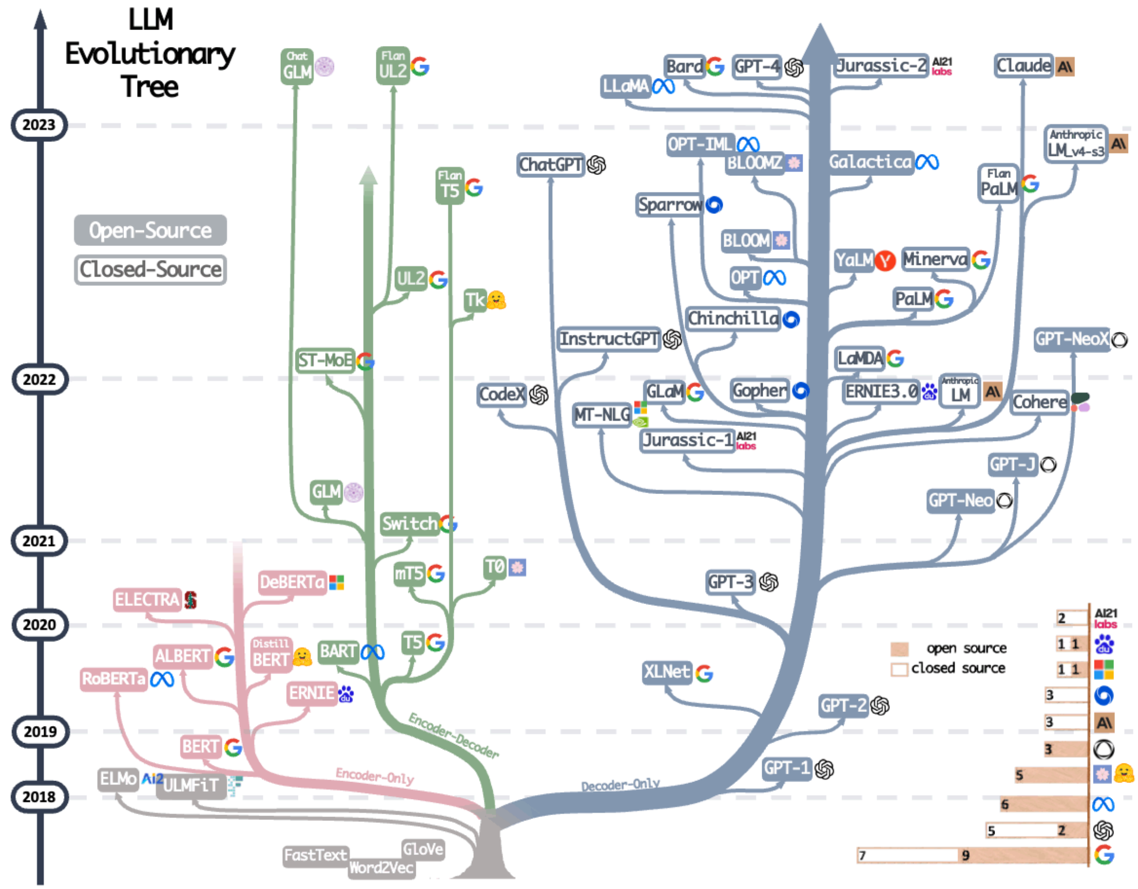


Figure 18: Visualization from Yang et al. (2023)