AuroraGPT: A Foundation Model for Science

Rajeev Thakur

Argonne National Laboratory

September 12, 2025

NYSDS 2025

Introduction

- AuroraGPT is an internal LDRD-funded project at Argonne
  - (Named after the exascale system at Argonne that is being used for much of the research)
- Leverage DOE supercomputing resources to develop and enhance understanding of powerful foundation models (FMs) for science
- Create and evaluate a series of increasingly powerful FMs, each with more parameters and/or trained on more data than those preceding it
- Goal is to build a large multimodal model capable of scientific reasoning that is causally aware and can generate novel insights

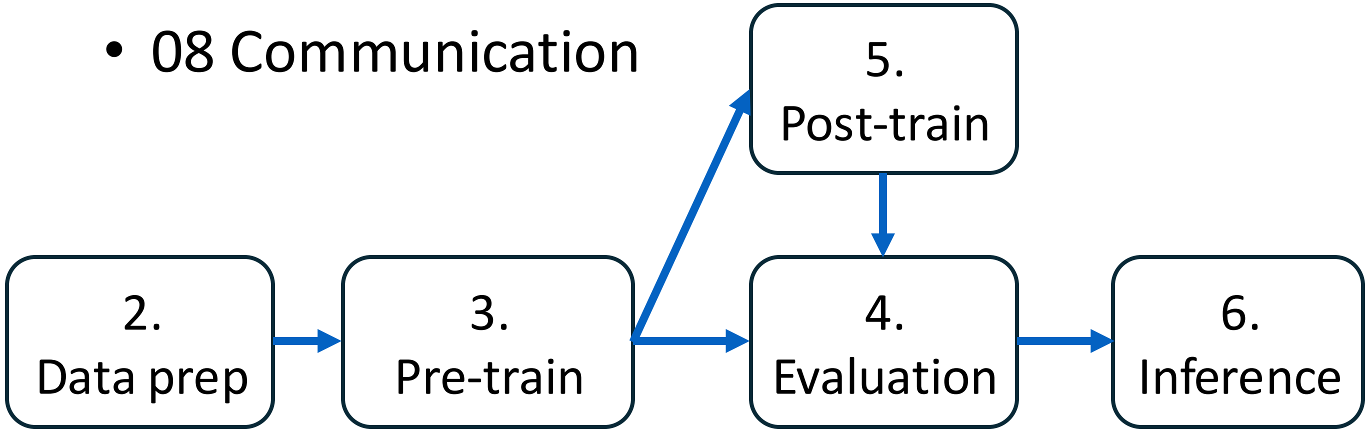
AuroraGPT

Explore pathways towards a "Scientific Assistant" powered by Aurora supercomputer:

- Assemble high-quality **scientific datasets** for scientific FM training
- Adapt **FM development methods** to meet specialized needs of scientific FMs
- Assemble high-quality **benchmarks** to provide objective yardsticks for progress
- **Apply and evaluate methods** in areas important for DOE science

AuroraGPT project groups:

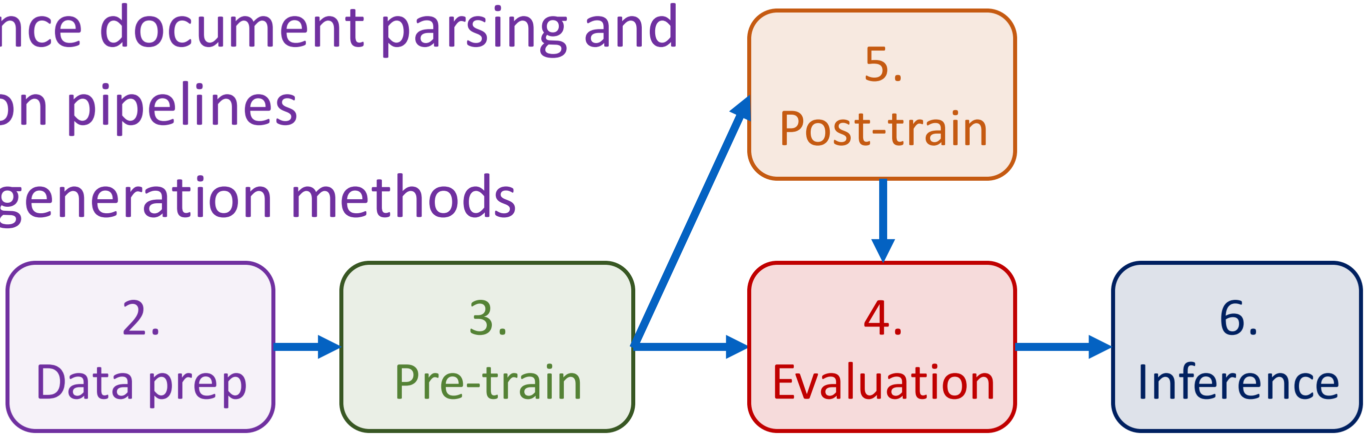
- 01 Planning
- 02 Data
- 03 Model training (pre-training)
- 04 Evaluation (skills, trustworthiness, safety)
- 05 Post-training (fine tuning, alignment)
- 06 Inference
- 07 Distribution
- 08 Communication



AuroraGPT activities

- Large datasets of scientific text
- High-performance document parsing and de-deduplication pipelines
- Synthetic data generation methods

Post-training models adapted to meet specialized needs of science FMs



Scalable inference methods for use on ALCF and other supercomputers

- Scalable pre-training pipelines for Polaris and Aurora
- Models trained with standard and enhanced datasets

- Acquire and deploy wide variety of evaluation suites
- New evaluation methods specialized for science FMs

Intended outcomes

- **Datasets and data pipelines** for preparing Science training data
- **Software infrastructure and workflows** to train, evaluate, and deploy LLMs at scale for scientific research purposes
- **Evaluation of state-of-the-art LLM models** to determine where they fall short in deep scientific tasks and where deep data may have an impact
- **Assessment of the value** of augmenting web training data with two forms of science-specific data
  - Full-text scientific papers
  - Structured scientific datasets (suitably mapped to narrative form)
- **Research grade artifacts (models)** for scientific community and adaptation for downstream uses
- **Promotion of responsible AI** best practices, where we can figure them out
- **International collaborations** around the long-term goal of <u>AI for science</u>

AuroraGPT Leaders

PLANNING


RICK STEVENS (LEAD)


IAN FOSTER


RINKU GUPTA (PM)


MIKE PAPKA


ARVIND RAMANATHAN


FANGFANG XIA

DISTRIBUTION


BRAD ULLRICK

DATA


IAN FOSTER


ROBERT UNDERWOOD

MODELS


VENKAT VISHWANATH


SAM FOREMAN

EVALUATION AND SAFETY


FRANCK CAPPELLO


SANDEEP MADIREDDY

POST-PRE TRAINING


ELIU HUERTA


AZTON WELLS

INFERENCE

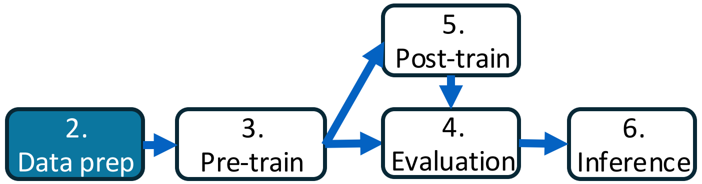

RAJEEV THAKUR

COMMUNICATIONS


RINKU GUPTA


CHARLIE CATLETT

02 Data



Goal: Assemble a large corpus of documents (general and scientific) and scientific data for AuroraGPT model training, fine-tuning, reasoning

Data Collection
- Generic data (Dolma, 2T tokens)
- Scientific papers (~100 Millions), Respect copyright (e.g. ACM Digital Library)
- Scientific data (x Exabytes)

Scientific Data Adaptation
- **Conversion** PDF into text (math formula, figures) + Convert science information (data) to text (narrative)
- **De-duplication** (syntactic and semantic) of x100B of scientific documents (to avoid memorization, bias)

Data Quality
- Peer-reviewed papers as much as possible, but also preprints: arXiv, bioRxiv, ChemRxiv, etc.
- Scientific data from trusted sources (e.g. DOE facilities)

Data Domains
- All scientific domains, starting with Material, Physics, Biology, Computer Science, Chemistry, etc.

Credit: Data team

02 Data: Collection (partial)

Generic Data

Dataset	Format	Size
RP1	JSONL (general web text)	~3TB
RP2	JSONL (general web text)	~5TB
DOECode	Code (DOE only)	9GB
PILE	JSONL (2/3 general)	825GB
StackCode	Parquet (code)	783GB
Dolma	JSONL	5TB

Scientific Data

Dataset	Format	Size
PubChem Compound	json	
PubChem Compound (no description)	json	
PubChem Gene	json	
PubChem Pathway	json	
UniProt TrEMBL	json	
UniProt uniref100	json	

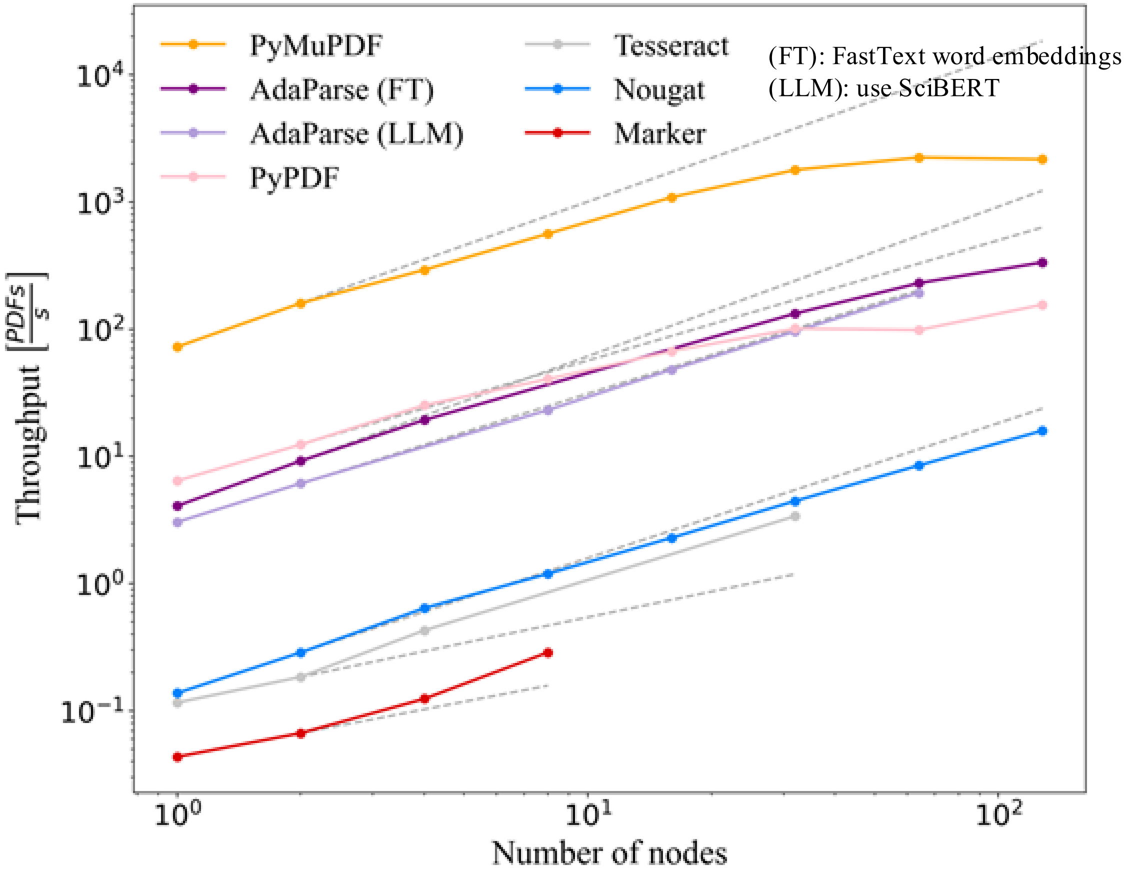
Scientific Papers

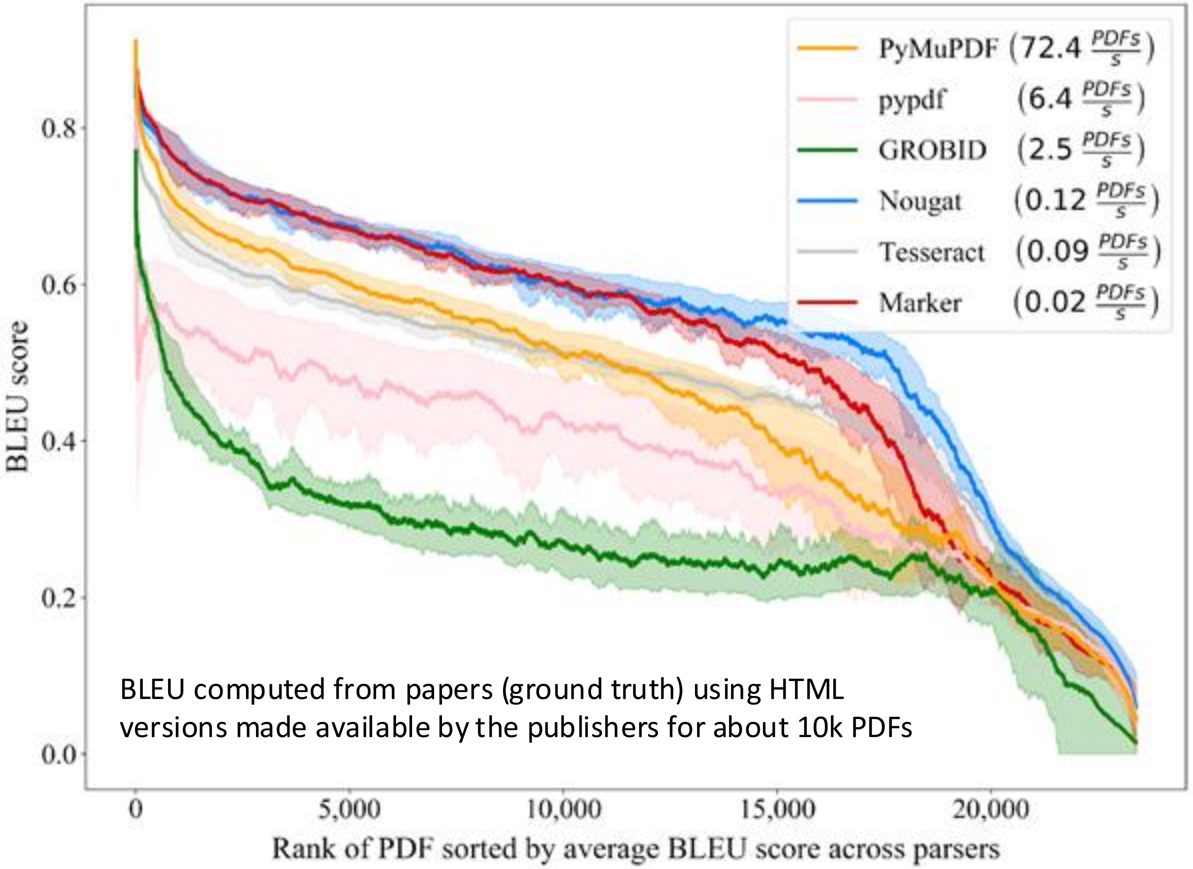
Dataset	Format	Size
CORE	Full text collection of scientific papers	>2TB
peS2o	Jsonl (40M open access academic papers)	259GB
PMC-OA	markdown+pdf	202GB
Arxiv	pdf+figures	2.2TB
Biorxiv	xml+pdf+figures	9.7TB
Medrxiv	xml+pdf+figures	542GB
chemrxiv	pdf	
ACM	XML	16GB
NIH_LITARCH	xml+pdf+figures	153GB

- Many documents → Scaling parsing is needed
- Significant overlap → De-duplication is important

02 Data: AdaParse: An Adaptive parallel PDF parsing and resource scaling engine

PDF/s numbers are for a single node (4 GPUs).





- PDFs vary greatly in their complexity; parsers vary greatly in cost and per-doc accuracy
- Hence: Estimate per-doc complexity, choose parser(s) to meet accuracy-cost target

→ **AdaParse: An Adaptive Parallel PDF Parsing and Resource Scaling Engine,** *Carlo Siebenschuh, Kyle Hippe, Ozan Gokdemir, Alexander Brace, Arham Mushtaq Khan, Khalid Hossain, Yadu Babuji, Nicholas Chia, Venkatram Vishwanath, Arvind Ramanathan, Rick L. Stevens, Ian Foster, Robert Underwood*, MLSYS 2025.

02 Data: LSHBloom: Memory-efficient, extreme-scale document deduplication

- We may have 100Ms or Billions of documents from many sources
- High degrees of "duplication" (not necessarily bit-for-bit) across sources
- De-duplication important for model quality, training costs
- SOTA **MinHashLSH** does not scale to 100Ms of docs
  → **LSHBloom** replaces expensive LSHIndex with lightweight Bloom filters

https://arxiv.org/abs/2411.04257

270% faster than MinhashLSH, while maintaining F1 score

Far faster than

Table 6: Deduplicated datasets of scientific documents. % new is the number not found in peS2o according to our deduplication strategy.

Name	# docs	% new	Description
Dolma 1.7	5.2 billion	–	Allen Institute for AI (AI2) general document collection
↪ peS2o	38,972,212	base	AI2 science articles (8M) and abstracts (30M) (in Dolma)
↪ ArXiv	1,554,434	55.11	ArXiV Scientific Preprint Server (in Dolma)
ASM	440,221	59.07	American Society for Microbiology
ACM	$\boxed{326,889}$	$\boxed{55.41}$	Association for Computing Machinery until 2017
BioRxiV	371,144	67.49	BioRxiV Scientific Preprint Server
OSTI	136,637	65.78	DoE Office of Scientific and Technical Information PDFs
MedRxiv	68,949	58.83	MedRxiv Scientific Preprint Server
NIH LIT ARCH	38,810	73.29	National Institutes of Health Archives
PMC-OA	60,311	52.63	PubMed Central Open Access Papers
IPCC	13	100.00	Intergovernmental Panel On Climate Change Reports

02 Data: Scientific Data Transformation Raw to Narrative

LLMs need text as inputs (until we figure-out direct tokenization of scientific data):

- Transformation of scientific raw data into "narratives" → **textual expression of the raw data**

The genome with identifier {{genome_id}} has {{genome_length}} base pairs and name {{genome_name}}.

{{$if:reference_genome}} {{$nl}}Genome {{genome_id}} is considered a {{reference_genome}} genome by NCBI.{{$fi}}

{{$nl}}Genome {{genome_id}} has {{contigs}} contigs, {{patric_cds}} known protein-coding regions, and is considered {{genome_quality}} quality.

{{$if:host_name}}
{{$nl}}{{genome_name}} is normally found in {{$list:host_name:and:, }}
{{$if:disease}}, where it causes {{$list:disease}}{{$fi}}.
{{$else}}
{{$if:disease}} {{$nl}}{{genome_name}} causes {{$list:disease}}.{{$fi}}
{{$fi}}

The genome with identifier **1121370.3** has **2300451** base pairs and name **Corynebacterium ulceribovis DSM 45146**.

Genome **1121370.3** is considered a **Representative** genome by NCBI.

Genome **1121370.3** has **8** contigs and **2108** known protein-coding regions, and is considered **Good** quality.

Corynebacterium ulceribovis DSM 45146 is normally found in **Bos taurus**, where it causes **ulceration**.

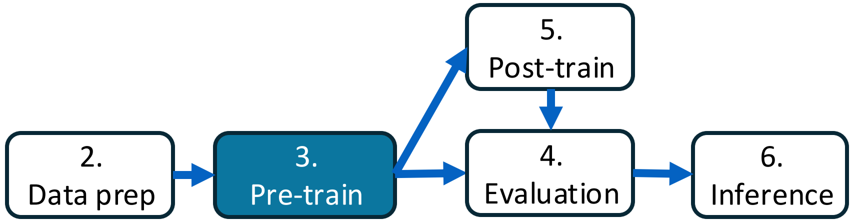
Figure 5: Examples of our template-based approach to generating narratives from scientific databases. On the left, a template designed for application to genomic data, with red denoting control statements and blue denoting variables to be filled in. On the right, a narrative produced via this template from data contained in BV-BRC, with instantiated values in blue.

Simple templating or dumps of data from the database → very high levels of duplication leading to memorization

- Need to give LLM clear guidance: e.g. Prioritize ensuring the factual integrity of summary by drawing heavily upon the record for information.
- Need to take into account perspectives: for what purpose the dataset has been generated (e.g. virologist vs a geneticist)
- Ask LLMs to consider all fields (if not LLMs tends to ignore fields)

Argonne NATIONAL LABORATORY

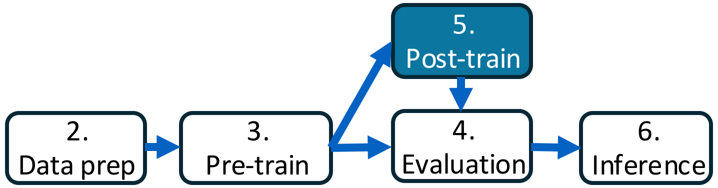
03 Model training (pre-training)



Goal: Solid Pre-training infrastructure, exploiting Aurora capabilities to maximize performance

- On Polaris (Nvidia) and Aurora (Intel)
- Megatron + DeepSpeed on Polaris. Adapted for Aurora (Intel GPUs)
  - **Challenge (parallel computing)**: **identify right level of data/model/pipeline/tensor parallelism for Aurora**
- Gradually increase from 7B, 70B, etc.
- Capture checkpoints
  - **Challenge (parallel computing)**: **low overhead parallel checkpointing**
- Capture loss curves and scaling data
- Detect/Handle spikes in loss, and monitor perplexity when training large models
  - **Challenge**: automatically detect spikes and identify checkpoint to restart from to avoid

05 Post-Training



Goal: Post-pre-training workflow optimized for science tasks

- Implement **post training workflow** for snapshots from AuroraGPT pre-training runs
- Include Chat fine-tuning and alignment focusing on Math and Coding
  - Chat Supervised Fine Tuning (SFT) and Instruct SFT
  - Alignment (truthfulness, safety): Based on RLHF (Reinforcement Learning From Human Feedback): DPO (Direct Preference Optimization), KTO (Binary signal: is the model output desirable or undesirable)
- **Challenge: need to collect more scientific conversations (1000 Scientists Jam will help)**

Evaluation/comparison of Llama 7B fine-tuned (used a collection of Instruction-tuning datasets UltraFeedback, hh-rlhf): Verbal ability, reasoning, truthfulness, math, and code generation

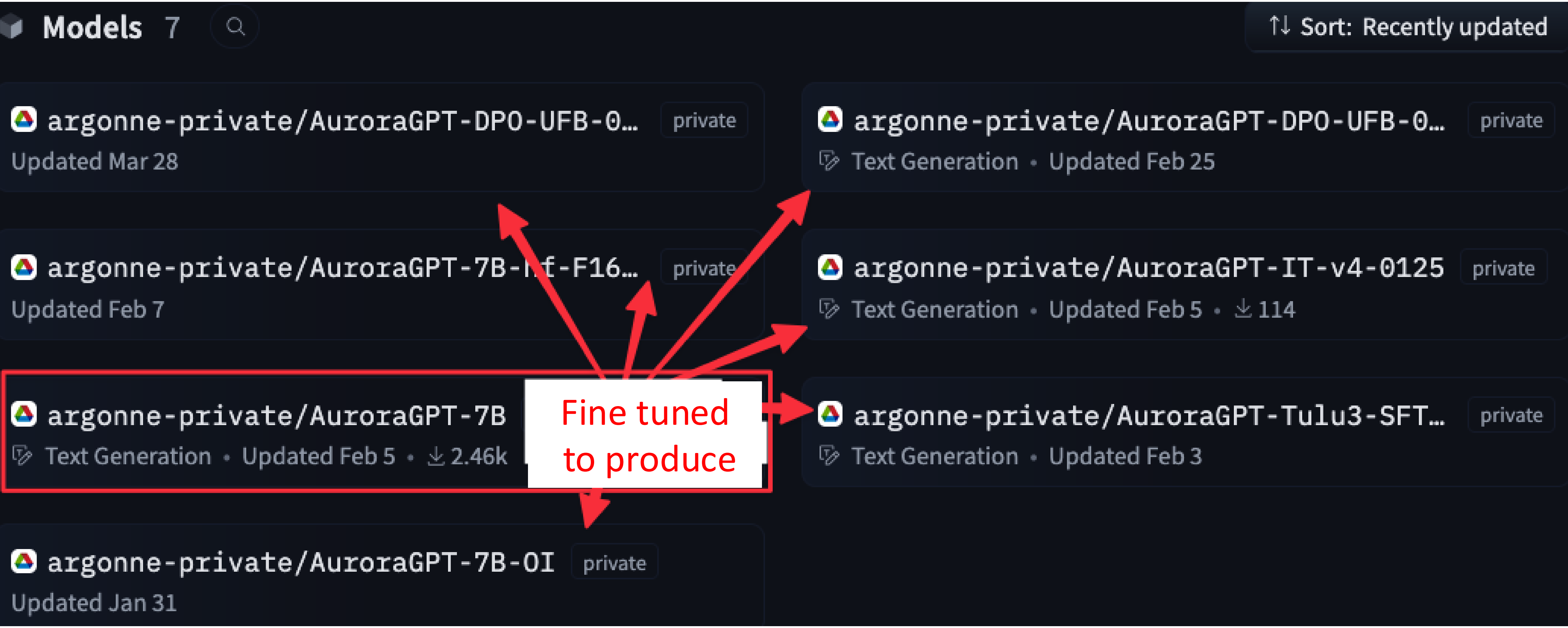
Leaderboard

Models	Natural Language and Reasoning: arc_challenge 25 shot	Natural Language and Reasoning: mmlu 5 shot	Natural Language and Reasoning: hellaswag 10 shot	Truthfulness: truthfulqa 0 shot	Math: gsm8k 5 shot	Code Generation: HumanEval pass@1
Llama-2-7b-hf	53.66	45.66	78.56	38.98	15.16	14.02
Llama-2-7b-chat	54.18	47.20	78.69	45.25	21.45	14.02
OLMo-7b	45.98	28.98	77.12	35.88	4.09	13.41
Ours	**67.21**	**53.02**	**79.87**	**49.70**	**31.46**	**31.70**

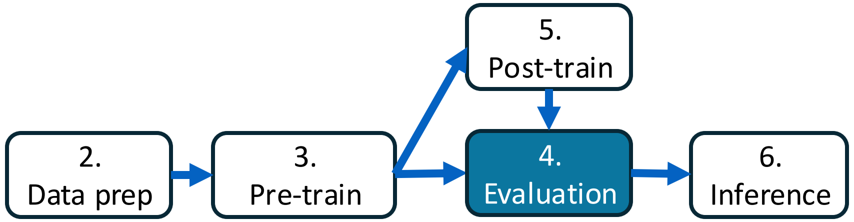
UltraInteract: large-scale, high-quality alignment dataset designed for complex reasoning tasks.

Credit: Post-training team.

03 Pre-training to 04 Post Pre-training (fine-tuning)



04 Evaluation



Goal: Comprehensive evaluation infrastructure for LLMs as scientific assistants

Primary purposes:

- Evaluate LLMs capabilities in research context: knowledge extension, reasoning capabilities, safety for users and community
- Compare with AuroraGPT trained with 100M+ scientific papers + data

Establish a methodology:

- Standard frameworks and benchmarks: EleutherAI Harness, HELM, SkillMix, FLASK (alignment)
- Safety benchmarks (Trustworthiness, Safety): DecodingTrust, TrustLLM, WMDP
- Existing domain-specific benchmarks in Chemistry, Physics, Climate, Biology, etc.
- **Create scientific benchmarks** (uncovered domains, new benchmarking approaches, etc.)
- **Create new evaluation techniques if needed**.

Argonne NATIONAL LABORATORY

EAIRA: **Multi-faceted** eval methodology

End-to-End

New New

Proposed Methodology				
Techniques	**MCQ Benchmarks**	**Open Response Benchmarks**	**Lab Style Experiments**	In the Wild **Field Style Experiments**
Main Goal	Testing knowledge **breadth, basic reasoning**	Testing knowledge **depth, planning, reasoning**	**Realistic** testing	**Realistic trend** analysis and weakness diagnosis
Problem Type	**Predetermined**, Fixed Q&As with known solutions	**Predetermined**, Fixed Free-Response Problems with known solutions	**Individual Human** Defined Problems with **unknown** solutions	**Many Human** Defined Problems with **(un)known** solutions
Verification	**Automatic** response verification	**Automatic or Human** response verification	**Humans detailed** response analysis	Scalable **automatic** summary of **human response**
Examples	Astro, Climate, AI4S (multi-domain), **Existing Benchmarks**	SciCode, ALDbench	**see "lab style experiments"**	**see "field style experiments"**
Cross Cutting Aspects	← **Trust and Safety (ChemRisk), Uncertainty Quantification, Scalable Software Infrastructure (STAR)** →			

(Prior work by others, Prior work by authors, New work)

4 complementary evaluation techniques to comprehensively assess the capabilities of LLMs as scientific assistants.

EAIRA: A Methodology for Evaluating AI Models as Scientific Research Assistants, https://arxiv.org/pdf/2502.20309.

MCQ Benchmark: ASTRO

- **4425 Automatically generated MCQs**
- From 885 articles in Annual Review of Astronomy and Astrophysics, 1963 to 2023.
- Instructed Gemini-1.5-Pro to propose 5 questions that can be answered based on the paper's content.
- Each question was accompanied by four options (A, B, C, D) only one of which is correct.
- Robustness considerations added to the prompt generating the questions.
- 200 MCQs were manually validated

Sample question from Astronomy benchmark dataset

How does the presence of stellar companions influence the formation and detection of exoplanets?

(A) Stellar companions can dilute transit signals, potentially leading to misclassification of planets and inaccurate parameter estimations. Additionally, their gravitational influence can suppress planet formation in close binary systems.

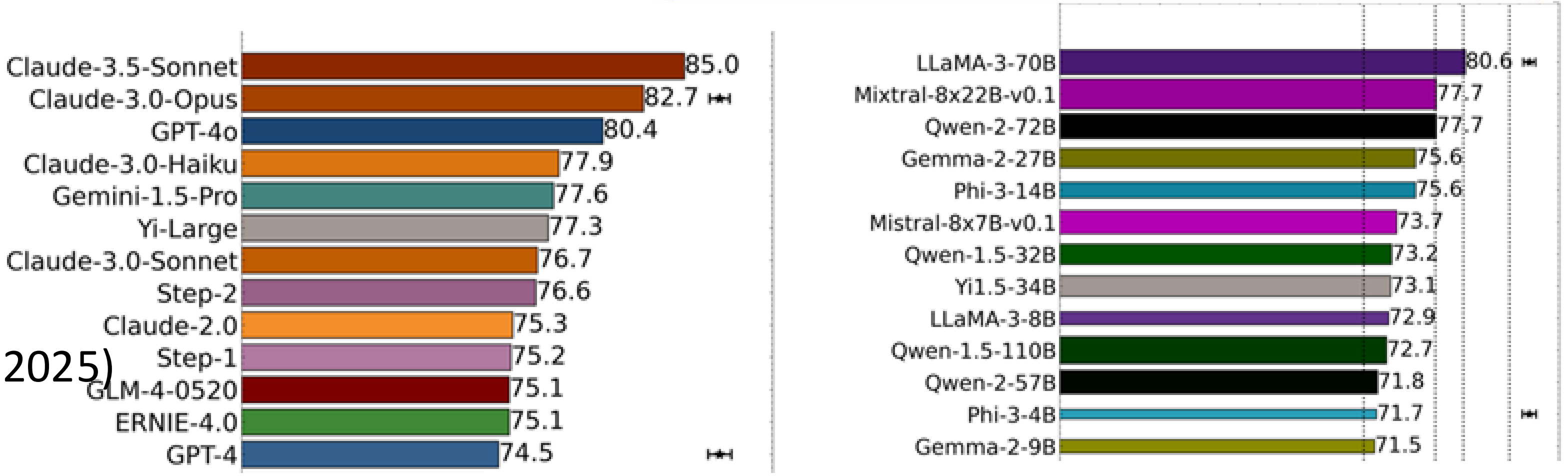
(B) Stellar companions provide additional sources of gravitational perturbations, enhancing planet formation by promoting planetesimal accretion and facilitating the formation of gas giants.

(C) Stellar companions contribute to the metallicity enrichment of planetary systems, leading to the formation of more massive and diverse planets, including super-Earths and hot Jupiters.

(D) Stellar companions act as gravitational lenses, increasing the detectability of exoplanets through microlensing events and enabling the discovery of planets at greater distances from their host stars.

Some take aways:

- Claude 3.5 Sonnet best (no O1 test)
- Llama-3-70B on par with GPT4o
- Published in July 2024 on arXiv (journal: 2025)
- Benchmark almost/probably saturated



Y.-S. Ting, et al.i, AstroMLab 1: Who wins astronomy jeopardy!? Astronomy and Computing, Volume 51, 2025,

Open Response Benchmark: SciCode (integrated into the methodology)

Scientist-curated code generation benchmark (mathematics, physics, chemistry, biology, materials science)

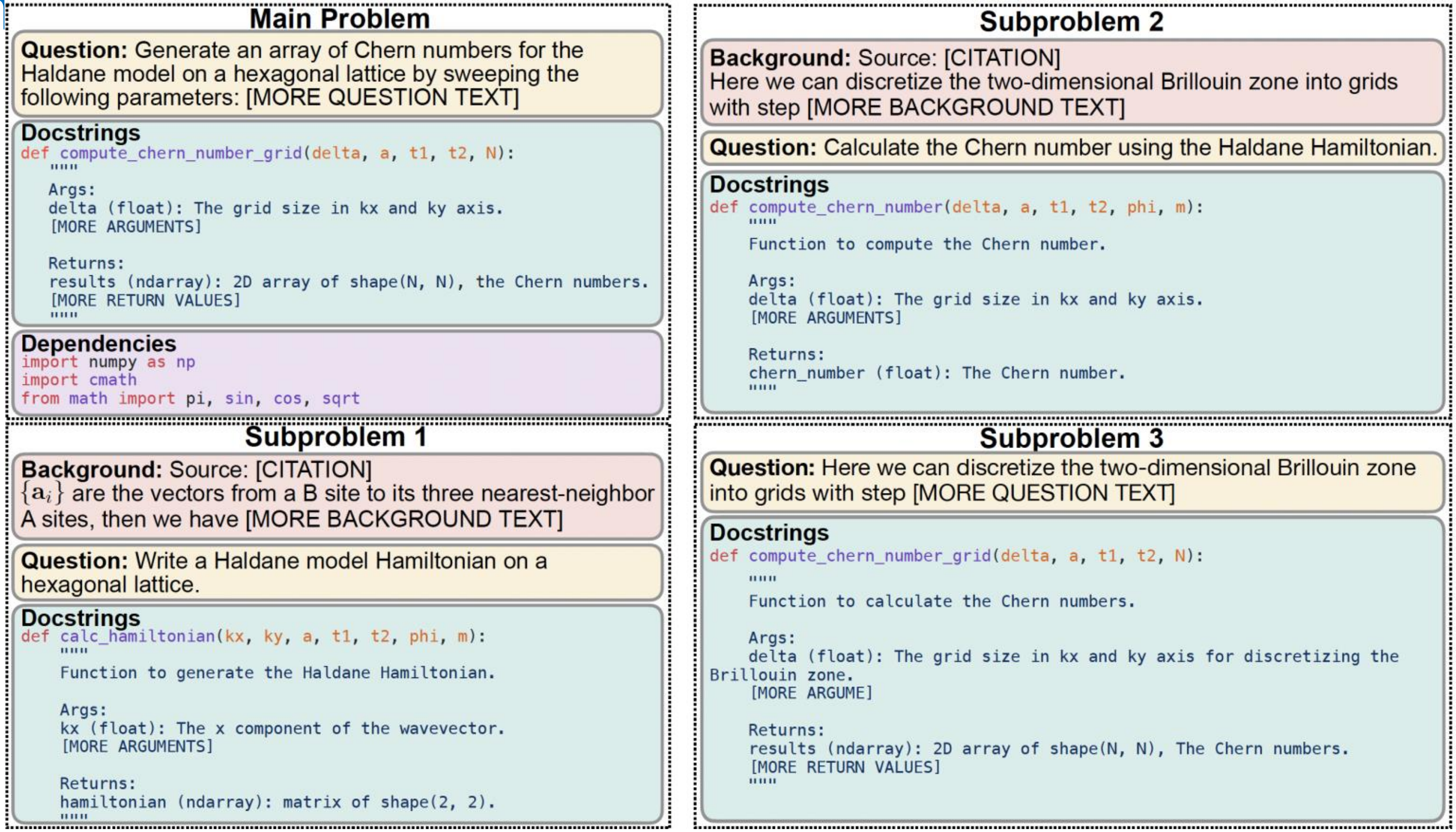
80 main problems (numerical methods, simulation of systems), decomposed into 338 subproblems

The problems naturally factorize into multiple subproblems, each involving knowledge recall, reasoning, code synthesis.

To solve a main problem, LLMs must implement multiple Python functions for each subproblem and integrate them into a comprehensive solution.

SciCode provides gold-standard solutions and multiple test cases for reliable automatic evaluation.

Problems are very challenging: inspired from Nobel prize level problems.



Minyang Tian, SciCode: A Research Coding Benchmark Curated by Scientists, arXiv: [arXiv:2407.13168](https://arxiv.org/abs/2407.13168)

End-to-End Eval: ~~1000~~ 1,500 Scientists AI JAM in 9 Labs Simultaneously (Feb.28, 2025)






Researcher participation and contributions on a voluntary basis.

1,000 Scientists Jam Session: In numbers

Researcher participation and contributions on a voluntary basis.

Total:
2800+ problems
15000+ assessed prompt responses

Argonne:
720 problems
2500 prompts

1,000 Scientists AI JAM Session: Goal and Rules of engagement

Researcher participation and contribution on a voluntary basis.

Goals:

- Give Lab researchers an opportunity to test the best available LLMs
- Build a large corpus of interactions between researchers and AI models
  - Will help Labs understand how researchers will use reasoning models LLMs for Science → **How AI models may accelerate discoveries**
  - Will help AI labs (OpenAI, Anthropic) to improve their model → to improve our research

Rules:

- Explore advanced AI models on **challenging scientific problems**,
- Better understand the potential impact of AI reasoning models on **national security and science**,
- **In-person event** hosted at Argonne, Berkeley, Brookhaven, Idaho, Livermore, Los Alamos, Oak Ridge, Pacific Northwest, and Princeton Plasma Physics national laboratories. Scientists from other DOE labs are also participating,
- Explore models from OpenAI (**o1-pro, o1-deepresearch, o3-mini-high**) and Anthropic (**Claude 3.7 extended**),
- **OpenAI people in the rooms**.

1,000 S AI JAM: Domains (Partial)

Researcher participation and contributions on a voluntary basis.

1,000 SCIENTIST AI JAM

Literature/Data
- Literature search, analysis, survey
- Data analysis and forecast, interpolation, extrapolation, classification (Point Cloud, signal, protein sequences, files, etc.)
- Anomaly detection
- Signal Analysis
- Scientific Visualization

Coding
- Algorithm design/optimization
- Automatic code generation/refactoring
- Code translation
- Debugging codes (sequential, parallel)
- Automatic code performance tuning/optimization
- Identifying performance bottlenecks

Experiments
- Automatic tuning of instruments
- Experimental Design (including autonomous workflow)
- Dark mater experiment design

Bio
- Understanding mechanisms of Cancer
- Understanding radiation effects on human cells
- Predictive Genomic Models

AI
- Domain specific LLMs/Agents (use LLMs as foundation models)
- Hyper parameter exploration for DL training.

Physics
- Battery design
- Chemical Mechanisms
- Physics beyond standard model

Infra.
- Infrastructure modeling and resilience
- Natural Disaster assessment

Math
- Surrogate model
- Mathematical derivations
- PDE solving
- Convergence proving
- Equation validity testing
- Derivative analysis
- Uncertainty estimation
- Inverse problems
- Statistical modeling

1,000 Scientists AI JAM: Not just for fun

https://arxiv.org/pdf/2503.23758

Statistical mechanics model of interactions between the *q*-state spins on a lattice (discrete degrees of freedom arranged in a regular spatial structure) leading to a situation where not all interactions can be simultaneously satisfied, resulting in a "frustrated" system with potentially complex behavior. (application in crystallography, percolation, and biological systems)

“derivation of an elegant equation … by OpenAI’s latest reasoning model o3-mini-high (never been solved before) at the first-ever **1000-Scientist AI Jam Session.** Hence, the author was inspired to prompt this AI reasoning model progressively … **despite quite a few errors in AI’s responses.”**

arXiv:2503.23758v2 [cond-mat.stat-mech] 6 Apr 2025

Exact Solution of the Frustrated Potts Model with Next-Nearest-Neighbor Interactions in One Dimension: An AI-Aided Discovery

Weiguo Yin*
Condensed Matter Physics and Materials Science Division,
Brookhaven National Laboratory, Upton, New York 11973, USA
(Dated: April 8, 2025)

The one-dimensional J_1-J_2 q-state Potts model is solved exactly for arbitrary q by introducing the maximally symmetric subspace (MSS) method to analytically block diagonalize the $q^2 \times q^2$ transfer matrix to a simple 2×2 matrix, based on using OpenAI’s latest reasoning model `o3-mini-high` to exactly solve the $q = 3$ case. It is found that the model can be mapped to the 1D q-state Potts model with J_2 acting as the nearest-neighbor interaction and J_1 as an effective magnetic field, extending the previous proof for $q = 2$, i.e., the Ising model. The exact results provide insights to outstanding physical problems such as the stacking of atomic or electronic orders in layered materials and the formation of a T_c-dome-shaped phase often seen in unconventional superconductors. This work is anticipated to fuel both the research in one-dimensional frustrated magnets for recently discovered finite-temperature application potentials and the fast moving topic area of AI for sciences.

Finding novel phases and phase transitions is a central challenge in various research fields, including condensed matter physics, materials science, quantum information, and microelectronics [1]. Unusual phases abound in frustrated magnets [2], which are described typically by the Ising model [3] or the quantum Heisenberg model [4] with competing spin-spin interactions either in the form of an equilateral triangle or via competition between the nearest-neighbor (NN) interaction J_1 and next-nearest-neighbor (NNN) interaction J_2 [1].

The third basic model of statistical mechanics is the q-state Potts model [5–8], which is a generalization of the Ising model ($q = 2$) and can serve as a useful intermediary to study the transition from discrete (Ising) to continuous (Heisenberg) symmetry. In particular, the one-dimensional (1D) J_1-J_2 Potts model could be relevant to problems ranging from the out-of-plane stacking of atomic or electronic orders in layered materials, such as charge stripe ordering in $La_{1.67}Sr_{0.33}NiO_4$ [9], the Star-of-David charge-density wave in $1T$-TaS_2 [10], and spin spiral ordering in the Weyl semimetal EuAuSb [11], to a time series with multiple choices at every time step such as table tennis training drill designs.

While the J_1-J_2 Ising model and Heisenberg model in one dimension [12–15] and two dimension [16–18] have been extensively studied, only the 1D J_1-J_2 Ising model has been solved exactly by using the transfer matrix method [19]. Exact analytic solutions of the 1D J_1-J_2 Potts model also remain unknown; since the model with $q = 3$ already exhibits a distinct ground-state phase behavior from that with $q = 2$, i.e., the Ising model (Fig. 2) [11], it is of fundamental importance to exactly solve the model for arbitrary q. The challenge arises from rapid increase in the order of the transfer matrix, which equals q^2. No wonder a 9×9 matrix for $q = 3$ is already hard to solve analytically and diagonalization of a $\left(10^{10^{10}}\right)^2 \times \left(10^{10^{10}}\right)^2$ matrix for $q = 10^{10^{10}}$ is simply beyond reach even numerically. Previous studies remarkably reduced the task to numerical calculations for an effective $q \times q$ matrix in the integer-q formalism of the transfer matrix—and for an effective 2×2 matrix in the continuous-q formalism of the transfer matrix where physics is less transparent—however, short of analytic exact results [20]. Hence, an intuitive understanding of the rich phase behaviors in the 1D J_1-J_2 Potts model is still lacking.

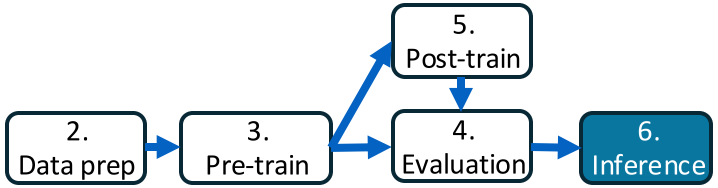
Two recent developments shed light on this long-standing problem. The first one is the analytic reduction of the 4×4 transfer matrix for a decorated Ising ladder to an effective 2×2 matrix using symmetry-based block diagonalization, leading to the discovery of spontaneous finite-temperature ultranarrow phase crossover (UNPC), which exponentially approaches the forbidden finite-temperature phase transition in 1D Ising models [21], and the subsequent discovery of in-field UNPC driven by exotic ice-fire states [22–24]. These findings point out the promising potentials of 1D frustrated magnets in finite-temperature applications; finding exact solutions for 1D frustrated Potts model could define a milestone in this important new direction. The second development is the derivation of an elegant equation—that determines the critical temperature of UNPC in site-decorated Ising models in an external magnetic field—by OpenAI’s latest reasoning model `o3-mini-high` at the first-ever 1000-Scientist AI Jam Session [24]. Hence, the author was inspired to prompt this AI reasoning model progressively to handle the transfer matrix in *the integer-q formalism* for the $q = 3$ case—despite quite a few errors in AI’s responses—and eventually have found a symmetry-based block diagonalization that can analytically reduce the 9×9 transfer matrix of the 1D J_1-J_2 three-state Potts model to an effective 2×2 matrix.

For general q, the key symmetry is the full permutation symmetry of the q Potts states. In other words, the Hamiltonian (and therefore the transfer matrix in the integer-q formalism) is invariant under any permutation of the labels $\{1, 2, 3, ..., q\}$; its symmetry group is $\mathcal{S}_q$. Although the AI failed to go further but warned that the number of permutations increases dramatically as q increases, the exact results for the $q = 2$ and 3 cases—especially the point that both arrive at an effective 2×2 matrix—stimulated the author to realize that since only the largest eigenvalue (λ) of the transfer matrix matters in the thermodynamic limit, the task is reduced

Researcher – Reasoning LLM collaboration

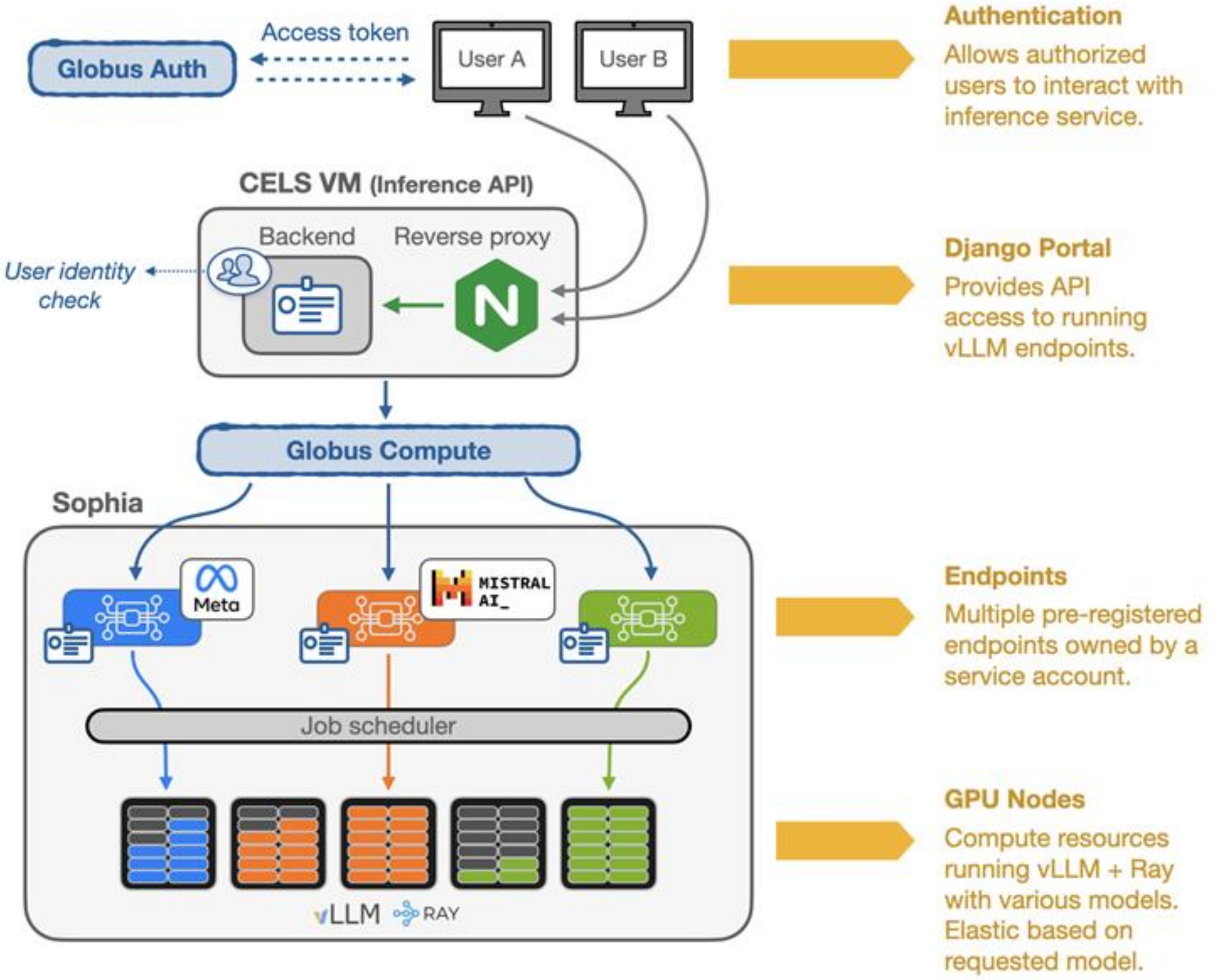
Brookhaven National Laboratory

06 Inference: Framework



Goal: A high performance, reliable inference service

- Inference framework deployed on ALCF systems (Sophia, Polaris, and Aurora)
- Leverages Globus Auth and Globus Compute (FuncX) for authentication and remote job submission
- Allows researchers to run parallel inference workloads (RAY) with an OpenAI-compliant API on private, secure compute environments
- Supports a variety of models and multiple inference backends (vLLM – single turn, SGLang – multi-turns),
- Supports interactive and batch modes (one inference request at a time via the API in a program. Or 1000s (batch mode).



INTERFACING WITH THE INFERENCE SERVICE

OpenAI API (including batch)

https://docs.alcf.anl.gov/services/inference-endpoints

cURL   Python (OpenAI SDK)

```
#!/bin/bash

# Get your access token
access_token=$(python inference_auth_token.py get_access_token)

curl -X POST "https://inference-api.alcf.anl.gov/resource_server/sophia
     -H "Authorization: Bearer ${access_token}" \
     -H "Content-Type: application/json" \
     -d '{
            "model": "meta-llama/Meta-Llama-3.1-8B-Instruct",
            "messages":[{"role": "user", "content": "Explain quantum co
         }'
```

API Usage Examples ¶

Querying Endpoint Status

Querying Endpoint Status

Chat Completions

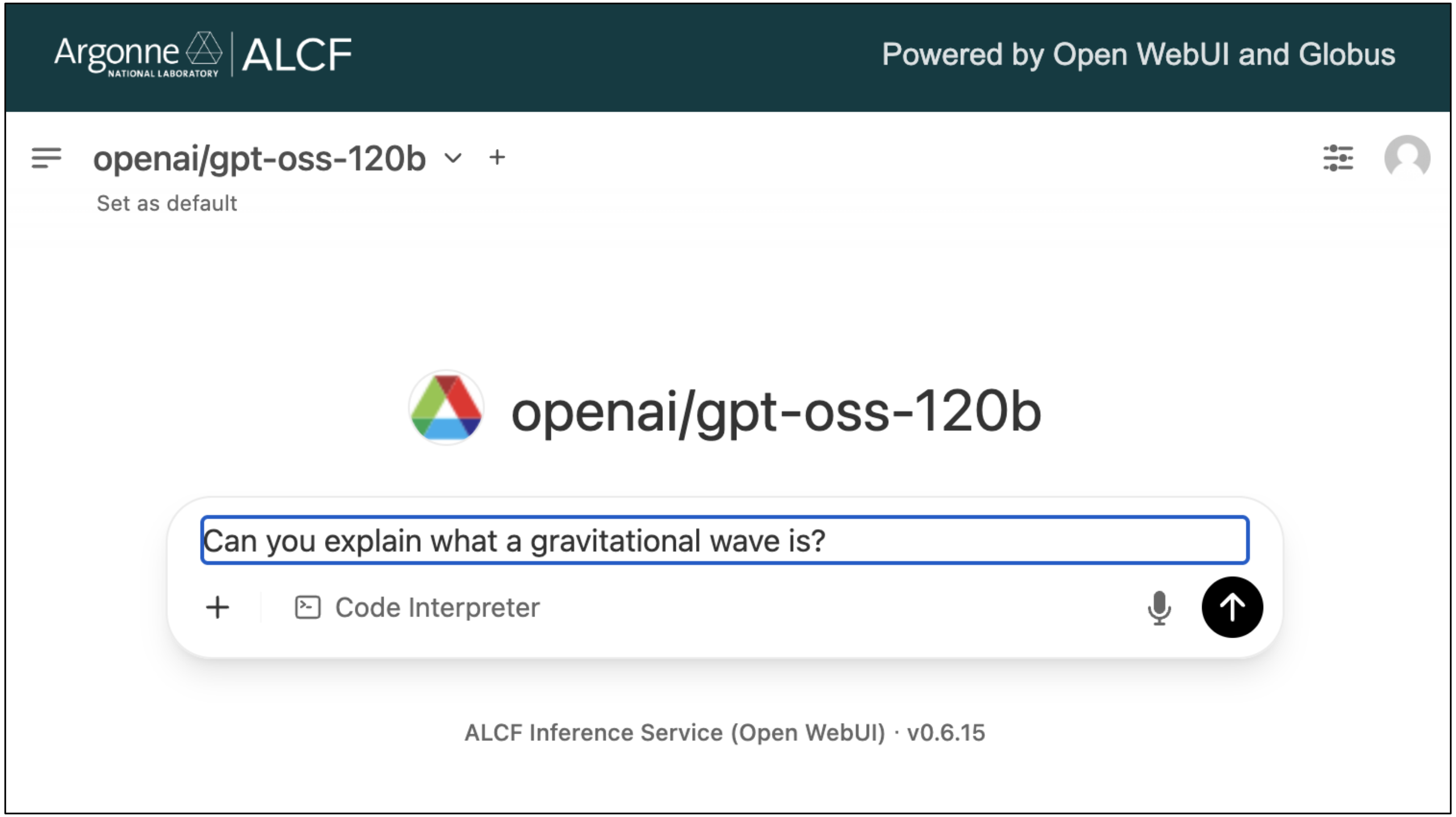
Chat Completions


```
{
    "id": "chatcmpl-68de443dde8b46659b4c34
    "object": "chat.completion",
    "created": 1755114580,
    "model": "meta-llama/Meta-Llama-3.1-8B
    "choices": [
        {
            "index": 0,
            "message": {
                "role": "assistant",
                "content": "Quantum computing is a new way of processing information that's different from the way regular computers work. Here's a simplified explanation:\n\n**Regular Computers:**\n\nRegular computers use \"bits\" to store and process information. Bits are like light switches that can be either ON (1) or OFF (0). When you combine these bits, you get numbers, letters, and other data.\n\n**Quantum Computers:**\n\nQuantum computers use \"qubits\" (quantum bits) to store and process information. Qubits are special because they can be both ON and OFF at the same time, which is called a \"superposition.\" This means a qubit can process multiple possibilities simultaneously, making it much faster than regular computers for certain tasks.\n\n**Another Key Concept: Entanglement**\n\nQubits can also be \"entangled,\" which means that when something happens to one qubit, it instantly affects the other qubits, no matter how far apart they are. This allows quantum computers to perform calculations on multiple qubits simultaneously, making them incredibly powerful.\n\n**How Quantum Computing Works:**\n\nImagine you have a combination lock with 10 numbers. A regular computer would try each number one by one, taking a long time to find the correct combination. A quantum computer, on the other hand, can try all 10 numbers simultaneously, thanks to the power of qubits and entanglement. This makes quantum computing incredibly fast for certain tasks, such as:\n\n1. **Cryptography:** Breaking complex codes and encryption methods.\n2. **Optimization:** Finding the best solution for complex problems, like logistics and supply chain management.\n3. **Simulation:** Simulating complex systems, like weather patterns and molecular interactions.\n\n**Challenges and Limitations:**\n\nQuantum computing is still a developing field, and there are many challenges to overcome, such as:\n\n1. **Error correction:** Qubits are prone to errors, which can affect the accuracy of calculations.\n2. **Scalability:** Currently, quantum computers are small and can only perform a limited number of calculations.\n3. **Noise:** Quantum computers are sensitive to external noise, which can disrupt calculations.\n\n**Conclusion:**\n\nQuantum computing is a revolutionary technology that has the potential to solve complex problems that are currently unsolvable or take too long to solve with regular computers. While it's still in its early stages, researchers and companies are working to overcome the challenges and limitations, and we can expect to see significant advancements in the coming years.",
```

```
"usage": {
    "prompt_tokens": 43,
    "total_tokens": 436,
    "completion_tokens": 393,
    "prompt_tokens_details": null
},
"prompt_logprobs": null,
"kv_transfer_params": null,
"response_time": 3.179178237915039,
"throughput_tokens_per_second": 137.14235798428732
```

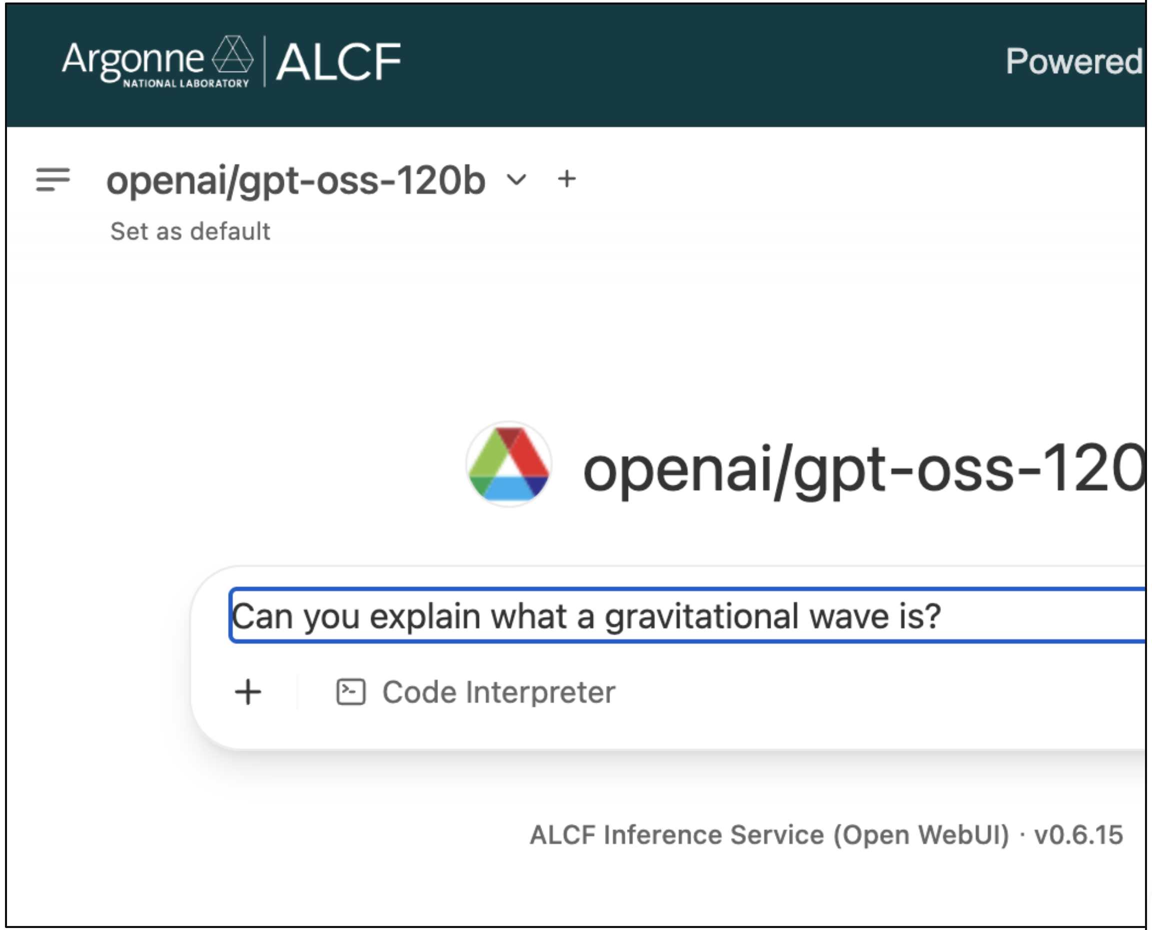
INTERFACING WITH THE INFERENCE SERVICE

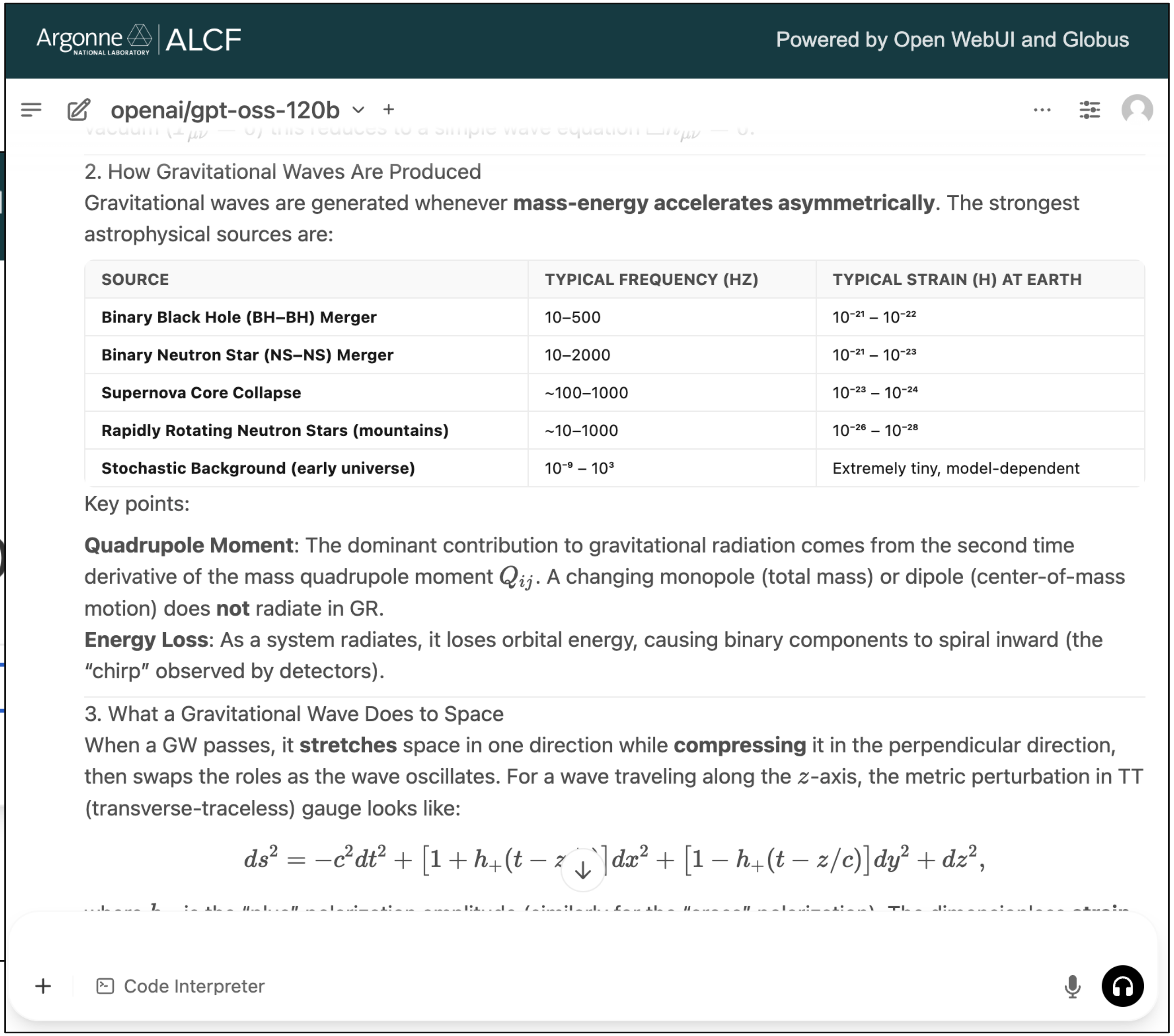
Open WebUI interface
https://inference.alcf.anl.gov



INTERFACING WITH THE INFERENCE SERVICE

Open WebUI interface
https://inference.alcf.anl.gov





Argonne NATIONAL LABORATORY | ALCF

Powered by Open WebUI and Globus

openai/gpt-oss-120b

2. How Gravitational Waves Are Produced

Gravitational waves are generated whenever **mass-energy accelerates asymmetrically**. The strongest astrophysical sources are:

SOURCE	TYPICAL FREQUENCY (HZ)	TYPICAL STRAIN (H) AT EARTH
Binary Black Hole (BH–BH) Merger	10–500	10^{-21} – 10^{-22}
Binary Neutron Star (NS–NS) Merger	10–2000	10^{-21} – 10^{-23}
Supernova Core Collapse	~100–1000	10^{-23} – 10^{-24}
Rapidly Rotating Neutron Stars (mountains)	~10–1000	10^{-26} – 10^{-28}
Stochastic Background (early universe)	10^{-9} – 10^{3}	Extremely tiny, model-dependent

Key points:

Quadrupole Moment: The dominant contribution to gravitational radiation comes from the second time derivative of the mass quadrupole moment Q_{ij}. A changing monopole (total mass) or dipole (center-of-mass motion) does **not** radiate in GR.

Energy Loss: As a system radiates, it loses orbital energy, causing binary components to spiral inward (the "chirp" observed by detectors).

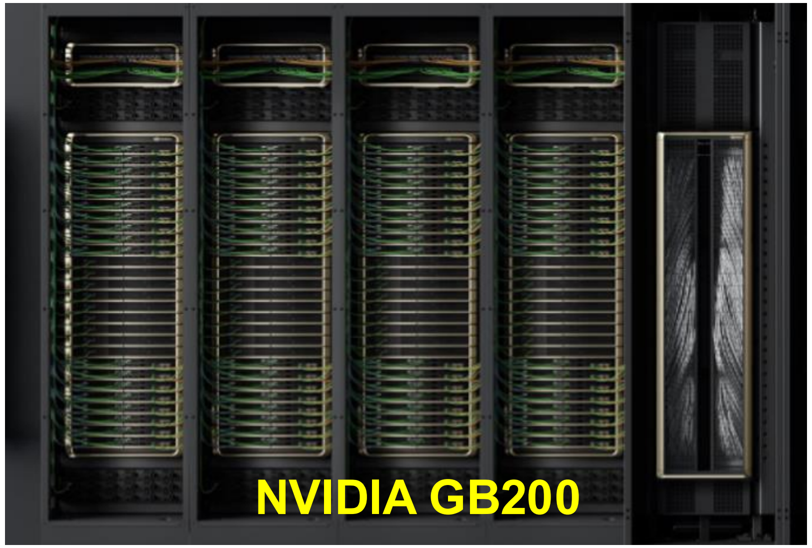
3. What a Gravitational Wave Does to Space

When a GW passes, it **stretches** space in one direction while **compressing** it in the perpendicular direction, then swaps the roles as the wave oscillates. For a wave traveling along the z-axis, the metric perturbation in TT (transverse-traceless) gauge looks like:

$$ds^2 = -c^2dt^2 + [1 + h_+(t - z/c)]dx^2 + [1 - h_+(t - z/c)]dy^2 + dz^2,$$

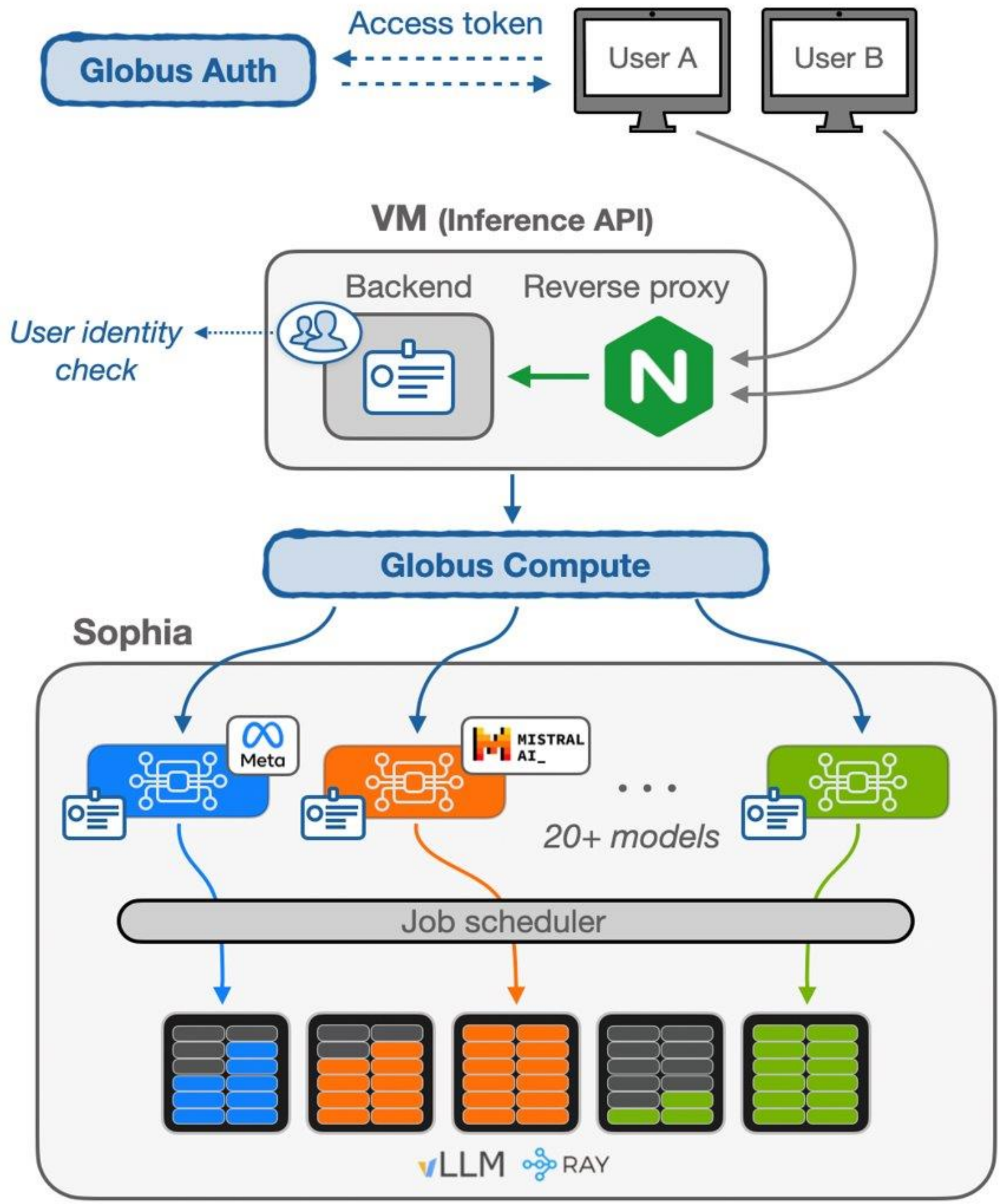
Code Interpreter

ALCF IS DEPLOYING DIVERSE INFERENCE SYSTEMS FOR SCIENCE



SYSTEM COMPONENTS

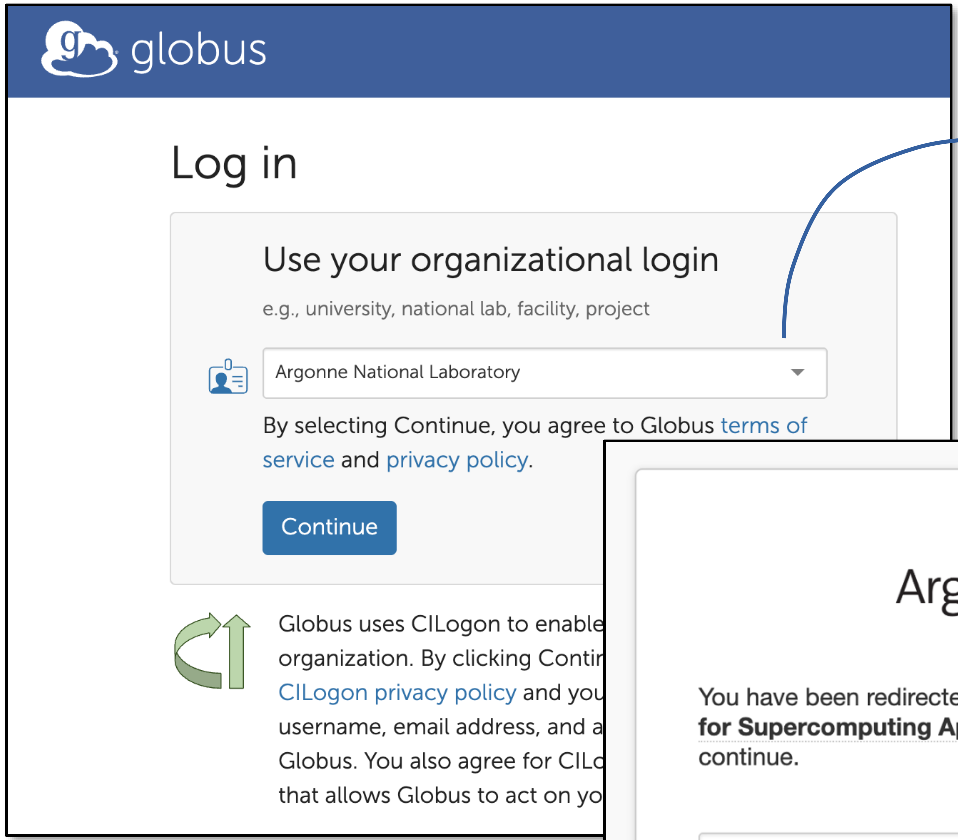
- **Globus Auth**: Enterprise-grade authentication and authorization service (OAuth2/OpenID)
- **API Gateway**: Django-Ninja async, OpenAI-compliant API handling, authorization and request routing, Postgres DB, monitoring
- **Globus Compute**: Orchestration and remote execution framework on HPC clusters
- **Compute Resources**: Compute nodes with GPUs on the ALCF Sophia cluster (more coming…)
- **Inference Backend**: High-performance inference servers (e.g., vLLM) for model serving; model weights downloaded and stored on HPC cluster



Documentation: https://docs.alcf.anl.gov/services/inference-endpoints/

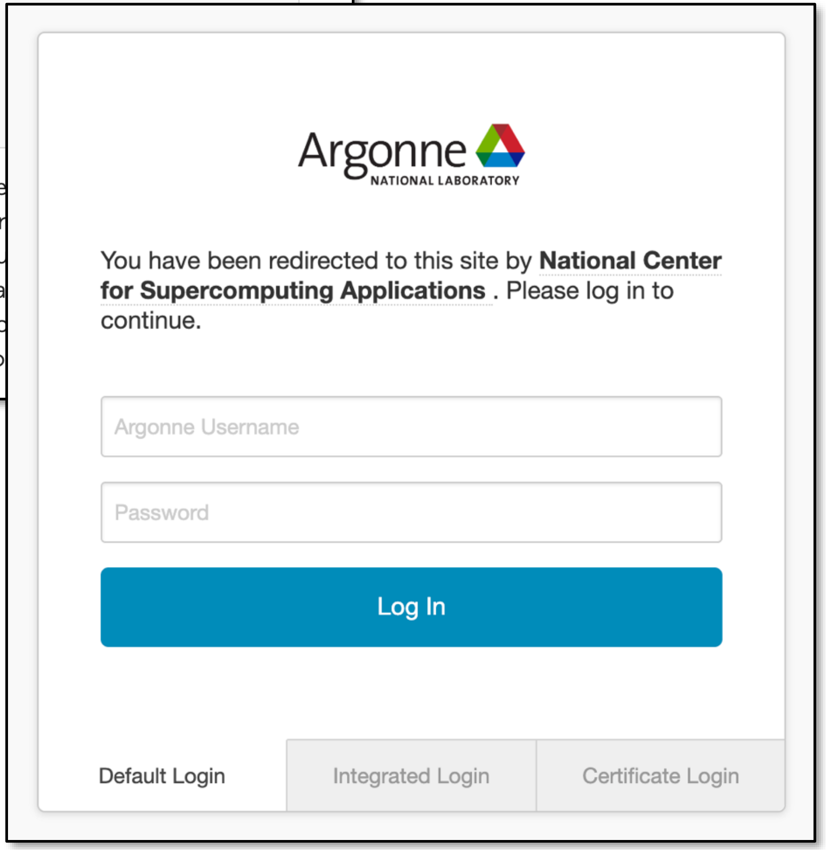
GLOBUS AUTH

- Authentication and authorization platform (OAuth2/OpenID compliant)
- Federated identity provider integrating with different institutions worldwide
- From a user's perspective:
  - Globus Auth generates a token
  - The token is passed to our inference service as an API key



1100+ identity providers

Inference service currently only opened to ANL and ALCF

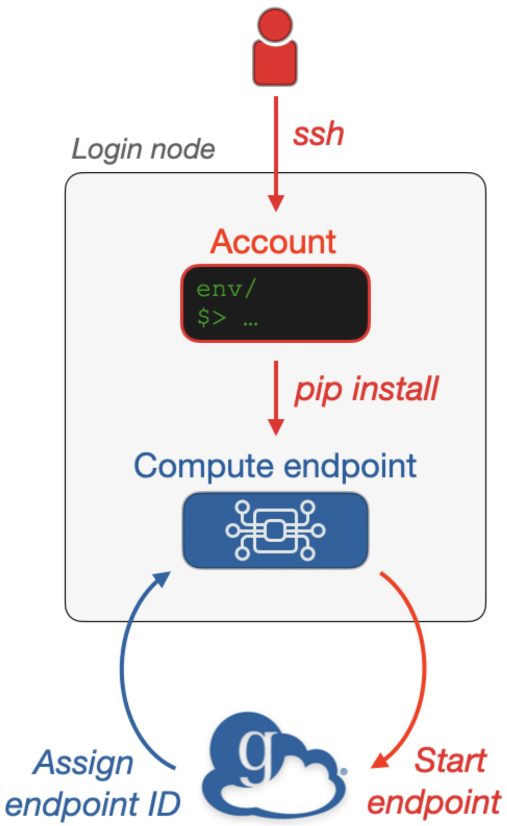



```python
client = OpenAI(
    api_key=access_token,
    base_url="https://inference-api.alcf.anl.gov/resource_server/sophia/vllm/v1"
)

response = client.chat.completions.create(
    model="meta-llama/Meta-Llama-3.1-8B-Instruct",
    messages=[{"role": "user", "content": "What are the symptoms of diabetes?"}]
)
```

GLOBUS COMPUTE

Globus Compute can trigger remote analysis on HPC systems from anywhere in the world.
Endpoints deployed on login nodes submit jobs to the scheduler to execute Python functions.

Install endpoint



Register function

```python
# Create Globus Compute client
from globus_compute_sdk import Client
gcc = Client()
```

The function can do whatever you want, including writing data on the filesystem or call more complex codes.

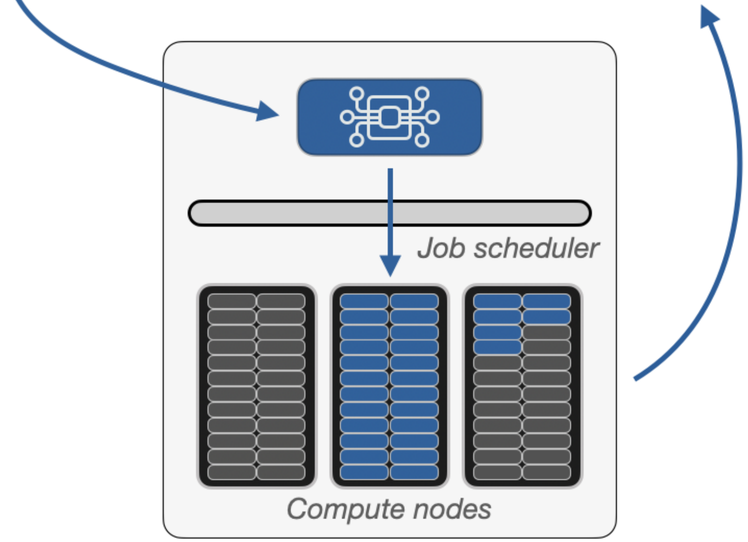
```python
# Define your analysis function
def my_analysis(arguments):

    # Import necessary modules
    import numpy as np
    import scipy

    # Do some analysis using local codes
    # ...

    # Return the computation results
    return ...
```

```python
# Register your function
function_id = gcc.register_function(my_analysis)
```

Run analysis

```python
# Submit a function to an endpoint
task_id = gcc.run(
    "my_arguments"
    endpoint_id=endpoint_id,
    function_id=function_id)
```

```python
# Recover results
results = gcc.get_result(task_id)
```



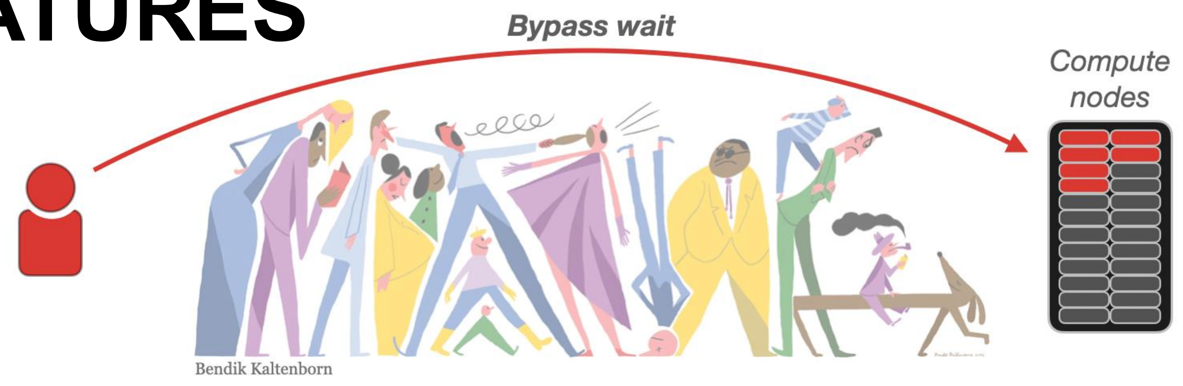
AVAILABLE MODELS (~30 TOTAL)

B - Batch enabled
T - Tool calling enabled
R - Reasoning enabled

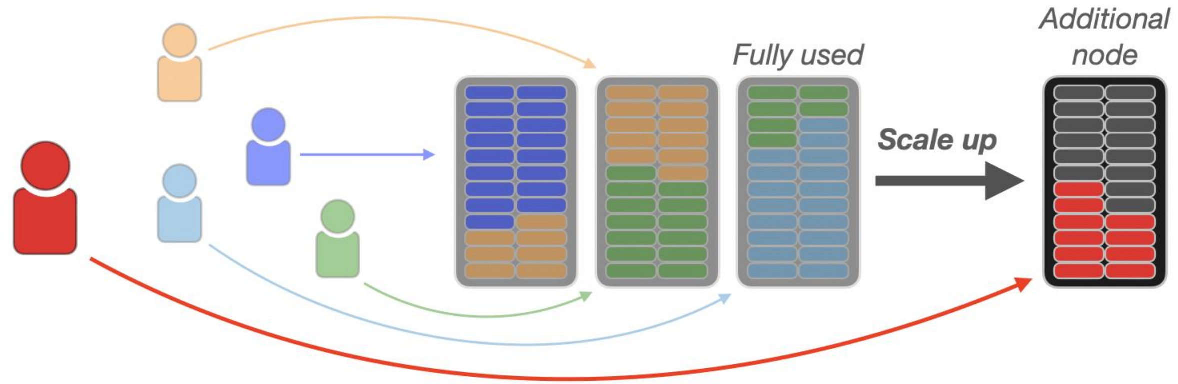
Family	Models
OpenAI	GPT-OSS-20B[BR], GPT-OSS-120B[BR]
Qwen	Qwen2.5-14B-Instruct[BT], Qwen2.5-7B-Instruct[BT], QwQ-32B[BRT], Qwen3-235B-A22B[RT], Qwen3-32B[BR]
Meta Llama	Meta-Llama-3-70B-Instruct[B], Meta-Llama-3-8B-Instruct[B], Meta-Llama-3.1-70B-Instruct[BT], Meta-Llama-3.1-8B-Instruct[BT], Meta-Llama-3.1-405B-Instruct[BT], Llama-3.3-70B-Instruct[BT], Llama-4-Scout-17B-16E-Instruct[BT], Llama-4-Maverick-17B-128E-Instruct[T]
Mistral	Mistral-7B-Instruct-v0.3[B], Mistral-Large-Instruct-2407[B], Mixtral-8x22B-Instruct-v0.1[B]
Nemotron	mgoin/Nemotron-4-340B-Instruct-hf
Aurora GPT	AuroraGPT-IT-v4-0125[B], AuroraGPT-Tulu3-SFT-0125[B], AuroraGPT-DPO-UFB-0225[B], AuroraGPT-7B-OI[B]
Allenai	Llama-3.1-Tulu-3-405B
Google	gemma-3-27b-it[BT]
Vision (VLM)	Qwen/Qwen2-VL-72B-InstructB, meta-llama/Llama-3.2-90B-Vision-Instruct
Embedding	nvidia/NV-Embed-v2, Salesforce/SFR-Embedding-Mistral[B], mistralai/Mistral-7B-Instruct-v0.3-embed[B]

KEY CAPABILITIES AND FEATURES

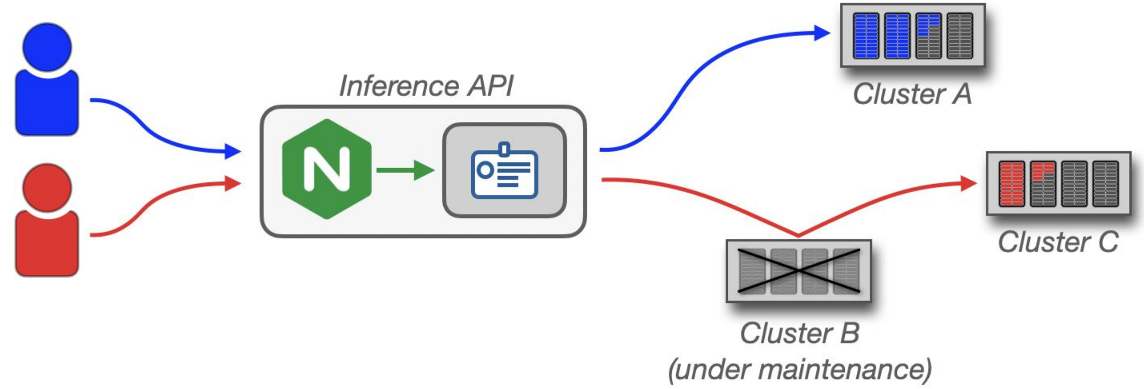
- **Dedicated Compute Resources**: Selected LLMs persistently served on dedicated nodes. This bypasses HPC queues and “cold starts”.



- **Auto-Scaling and Hot Nodes**: New nodes can dynamically be acquired to accommodate higher traffic. Cold models can be dynamically be loaded and kept hot for 24 hours.



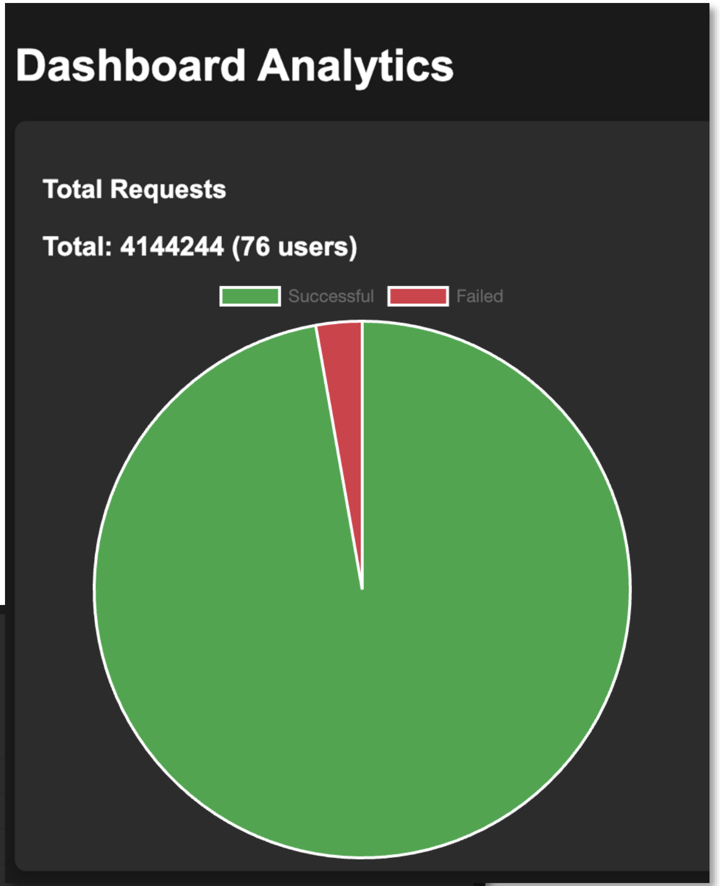
- **Multi-Backend Integration**: Our API can seamlessly route requests to diverse remote hardware, including SambaNova SN40 and Sophia inference clusters.

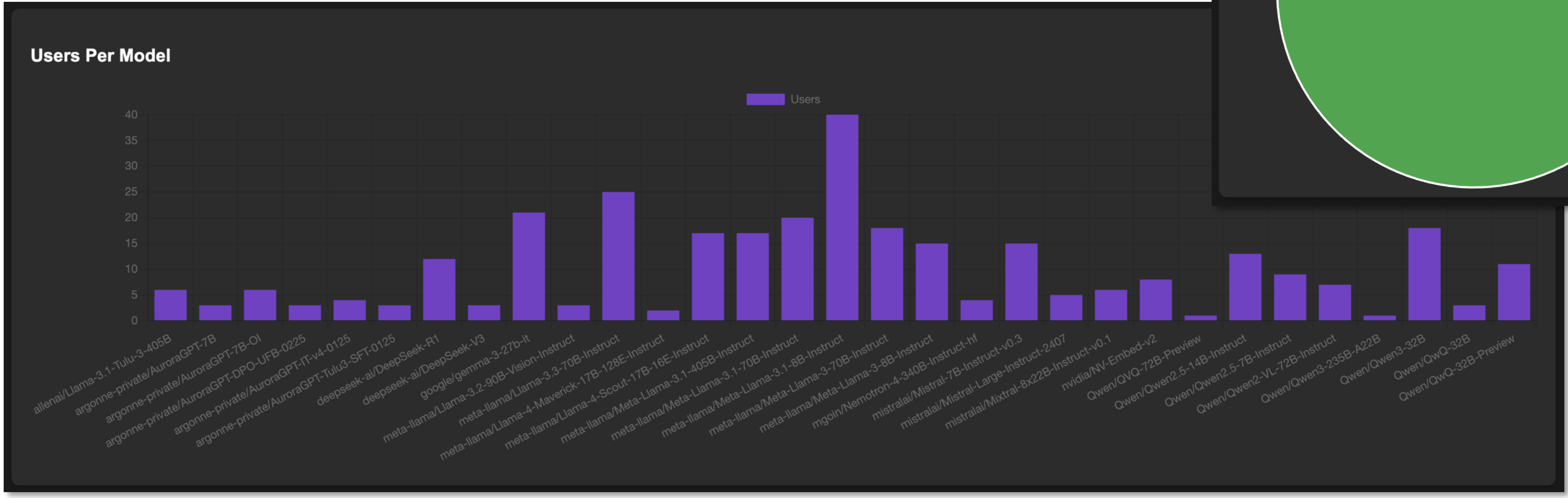


KEY CAPABILITIES AND FEATURES

- **Dashboard Monitoring**: A dashboard is available to system administrators and provides various metrics such as recent activities, number of requests and users, token throughput, and latency.

- **Current Status**: Over 8.7M requests, over 10 billion tokens generated, can generate ~3,500 tokens per second on a Sophia compute node.





TPC

Trillion Parameter Consortium

AI FOR SCIENCE

https://tpc.dev

INTERNATIONAL COLLABORATION OF OVER 80 ORGANIZATIONS

A*STAR
AI Singapore
AIST
Allen Institute For AI
Amazon Web Services, Inc. (AWS)
AMD
Argonne National Laboratory
Australian National University
Barcelona Supercomputing Center
Brookhaven National Laboratory
Caltech
CEA
CSCS
Cerebras Systems
CINECA
CSC – IT Center for Science
CSIRO
Deep Forest Sciences
ETH Zürich
Fermilab National Accelerator Lab
Flinders University
Fujitsu Limited
Groq
Harvard University
HPE
Indiana University
INESC TEC

Inria
Institute of Science Tokyo (formerly Tokyo Tech)
Intel
Jülich Supercomputing Center
Kotoba Technologies, Inc.
LAION
Lawrence Berkeley National Laboratory
Lawrence Livermore National Laboratory
Leibniz Supercomputing Centre
Los Alamos National Laboratory
Max Planck Computing & Data Facility (MPCDF)
Microsoft Research
National Center for Supercomputing Applications
National Energy Technology Laboratory
National Renewable Energy Laboratory
National Supercomputing Centre, Singapore
NCI Australia
New Zealand eScience Infrastructure
Northwestern University
NVIDIA
Oak Ridge National Laboratory
Pacific Northwest National Laboratory
Pawsey Institute
Pittsburgh Supercomputing Center
Princeton Plasma Physics Laboratory
Princeton University
RIKEN
Rutgers University
SambaNova

Sandia National Laboratories
Seoul National University
SLAC National Accelerator Laboratory
Sony Research
Stanford University
STFC Rutherford Appleton Laboratory, UKRI
Stonybrook University
SURF
Texas Advanced Computing Center
Thomas Jefferson National Accelerator Facility
Together AI
TÜBİTAK
Université de Montréal
University of Arizona
University of Buffalo
University of California San Diego / SDSC
University of Chicago
University of Delaware
University of Illinois Chicago
University of Illinois Urbana-Champaign
University of Michigan
University of New South Wales
University of Southern California / ISI
University of Tokyo
University of Toronto / Acceleration Consortium
University of Utah
University of Virginia
University of Washington

Americas – EMIA – Asia-Pacific – Industry

TPC All-Hands | Hackathons | Workshops at Conferences

TPC Events

2025

Winter 2025 TPC Hackathon

Hosted by RIKEN Center for Computational Science

Kobe, Japan

March 5-7, 2025

TPC Workshop at SCA25

Singapore

March 10, 2025

Spring 2025 TPC Hackathon

Hosted by CSC-IT Center for Science

Helsinki, Finland

May 6-8, 2025

TPC Workshop at ISC-HPC 2025

Hamburg, Germany

June 13, 2025

TPC25 All-Hands Hackathon and Conference

San Jose, USA

July 28-31, 2025

Fall 2025 TPC Hackathon

Exploring US and European Hackathons for Fall 2025

TPC Workshop at SC25

Frontiers in Generative AI for HPC Science and Engineering: Foundations, Challenges, and Opportunities.

St. Louis, USA

November 16-21, 2025

2023

TPC Global Kick-off Workshop

Hosted by Argonne National Laboratory and the University of Chicago

Chicago, USA

August 2-3, 2023

2024

ISC Workshop: Accelerating AI for Science

Hamburg, Germany

May 16, 2024

TPC European Kick-off Workshop

Hosted by the Barcelona Supercomputing Center

Barcelona, Spain

June 19-21, 2024

Fall 2024 TPC Hackathon

Hosted by Argonne National Laboratory and The University of Chicago

Chicago, USA

October 9-11, 2024

Accelerating the Development and Use of Generative AI for Science and Engineering: The Trillion Parameter Consortium

Atlanta, USA

November 22, 2024

Outcome reports and online access to presentation materials are available for all TPC Events at https://tpc.dev/tpc-events/

TPC Biweekly Distinguished Seminar Series (2024-5 Speakers)

Towards Scientific Agents: From Foundation Models to Automated Discovery

Karthik Duraisamy
Professor of Aerospace Engineering at the University of Michigan and director of Michigan Institute for Computational Discovery and Engineering (MICDE)

AI Agents: Unleashing the Power of Superintelligence in Science and Technology

Dr. Neeraj Kumar
Chief Data Scientist at Pacific Northwest National Laboratory (PNNL)

Scaling Generative AI and LLM Models on Aurora

Koichi Yamada
Sr. Principal Engineer in the Data Center and AI Group (DCAI) at Intel

Valentin Reis

Software Engineer at Affiliation: Groq Inc.

Bo Li

Neubauer Associate Professor in the Department of Computer Science

Kyle Lo

Research Scientist at the Allen Institute for AI in Seattle

Sajal Dash

Research Scientist at Oak Ridge National Laboratory

Michael C. Frank

Stanford University

Dexter Pratt

Director of Software Development

EAIRA: Establishing a methodology to evaluate LLMs as research assistants

April 2, 2025
10-11:15 a.m. (CST)

Franck Cappello
Senior Computer Scientist, Argonne National Laboratory

***Part of the AI Distinguished Lecture Series*: AI-Driven Modelling of the Immune System**

May 1, 2025
11 a.m. (CST)

María Rodríguez Martínez
Yale School of Medicine

Scalable Training of Trustworthy and Efficient Predictive Graph Foundation Models for Atomistic Materials Modeling: A Case Study with HydraGNN

April 23, 2025
11 a.m.-12:15 p.m. (CST)

Prasanna Balaprakash
Director of AI Programs and a Distinguished R&D Scientist at Oak Ridge National Laboratory (ORNL)

Meta Platforms

February 5, 2025

Kevin Chan
Global Policy Campaign Strategies Director

Efficiently Learning at Test-Time with LLMs via Transductive Active Learning

March 5, 2025

Jonas Hübotter
Doctoral Researcher, Learning and Adaptive Systems Group at ETH Zurich

Scaling Large Vision-Language Models for Enhanced Multimodal Comprehension in Scientific Discovery

Chibuike Robinson Umeike
Graduate research and teaching assistant at University of Alabama

Adaptive Multimodal Conditional Diffusion for Complex Dynamic Systems

January 15, 2025

Dr. Alexander Scheinker
Los Alamos National Laboratory

Towards Generative Decision-Making Agents

Yuexiang (Simon) Zhai
Final year PhD candidate at Berkeley EECS

The Space of Possible Minds

Phillip Ball
Freelance writer and broadcaster

Rio Yokota

Global Scientific Information and Computing Center, Tokyo Institute of Technology

Resource-friendly alignment in language models: from reward modeling to preference learning

Jiwoo Hong
MSc Student
Affiliate: KAIST AI

Yuan-Sen Ting

Australian National University and Ohio State University

Professor Irina Rish

Université de Montréal (UdeM)

Kshitij Gupta

MSc student at Mila through the Université de Montréal (UdeM)

Leon Song

Senior Principal Research Manager at Microsoft Research

Rick L. Stevens

Associate Lab Director and Distinguished Fellow at Argonne National Laboratory

TPC Seminar Talk

February 19, 2025

Michael Levin
Tufts University, Levin Lab

PDE-Controller: LLMs for Autoformalization and Reasoning of PDEs

March 19, 2025
11 a.m.-12:15 p.m.

Dr. Wuyang Chen
Simon Fraser University

Research Assistants in Molecular Biology

May 14, 2025
10 a.m. (CST)

Miguel Vazquez
Head of the Genome Informatics Unit at the Barcelona Supercomputing Center (BSC)

Hosted by:

Dario Dematties

Postdoctoral Researcher at Northwestern Argonne Institute of Science and Engineering

Conclusions

- These are exciting times to be in computing field
- The AI industry is making rapid progress
- The science community has a unique opportunity to leverage AI for accelerating scientific discovery in unforeseen ways
- The AuroraGPT project aims to develop such a foundation model to catalyze advancements in science and engineering