Frontier AI for Science Security and Technology

Rick Stevens

Argonne National Laboratory

The University of Chicago

Crescat scientia; vita excolatur

The Basic Argument for FASST

Global Techno-Economic Landscape

1. AI is rapidly becoming the **dominate driver/signal of techno-economic progress and competition in the next decade**
2. **AI is pervasive and is becoming ubiquitous** across dozens of economically critical domains
3. Massive competition/positioning in AI **between western democracies, semi-aligned petro-states and adversarial sino-russian players**
4. AI provides state and non-state actors with the **potential for non-linear progress in technological, economic and national security domains**
5. US is the **clear leader in commercial AI** with dominate consumer facing systems
6. US government is **under investing in non-commercial non-defense frontier AI systems development and adoption**
7. US government focus has been on **mitigating AI risks** rather than exploiting advantage through strategic investments in non-commercial AI capabilities

AI Engagement across All Fields* is Exponential

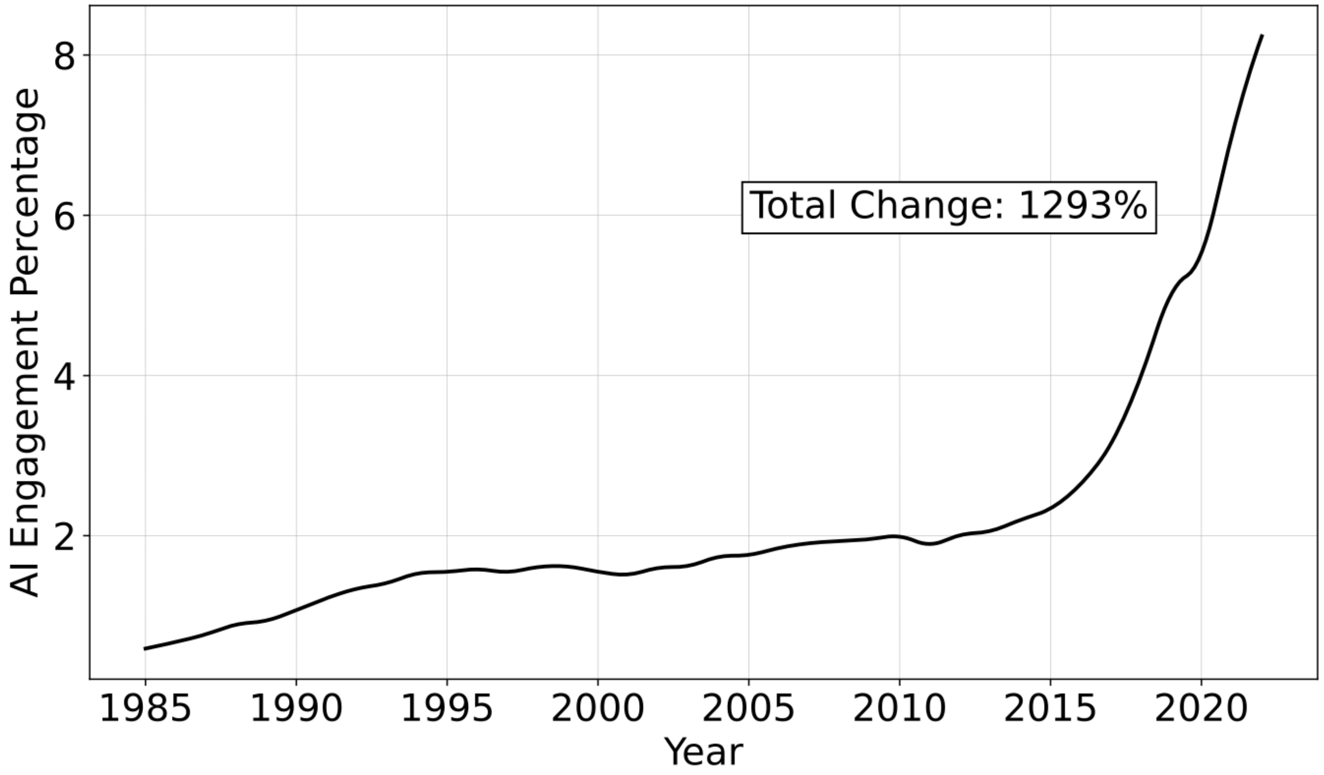


Figure 1: Change in AI engagement across all fields from 1985 - 2022

https://arxiv.org/pdf/2405.15828

AI Engagement Across 20 Exemplar Fields

https://arxiv.org/pdf/2405.15828

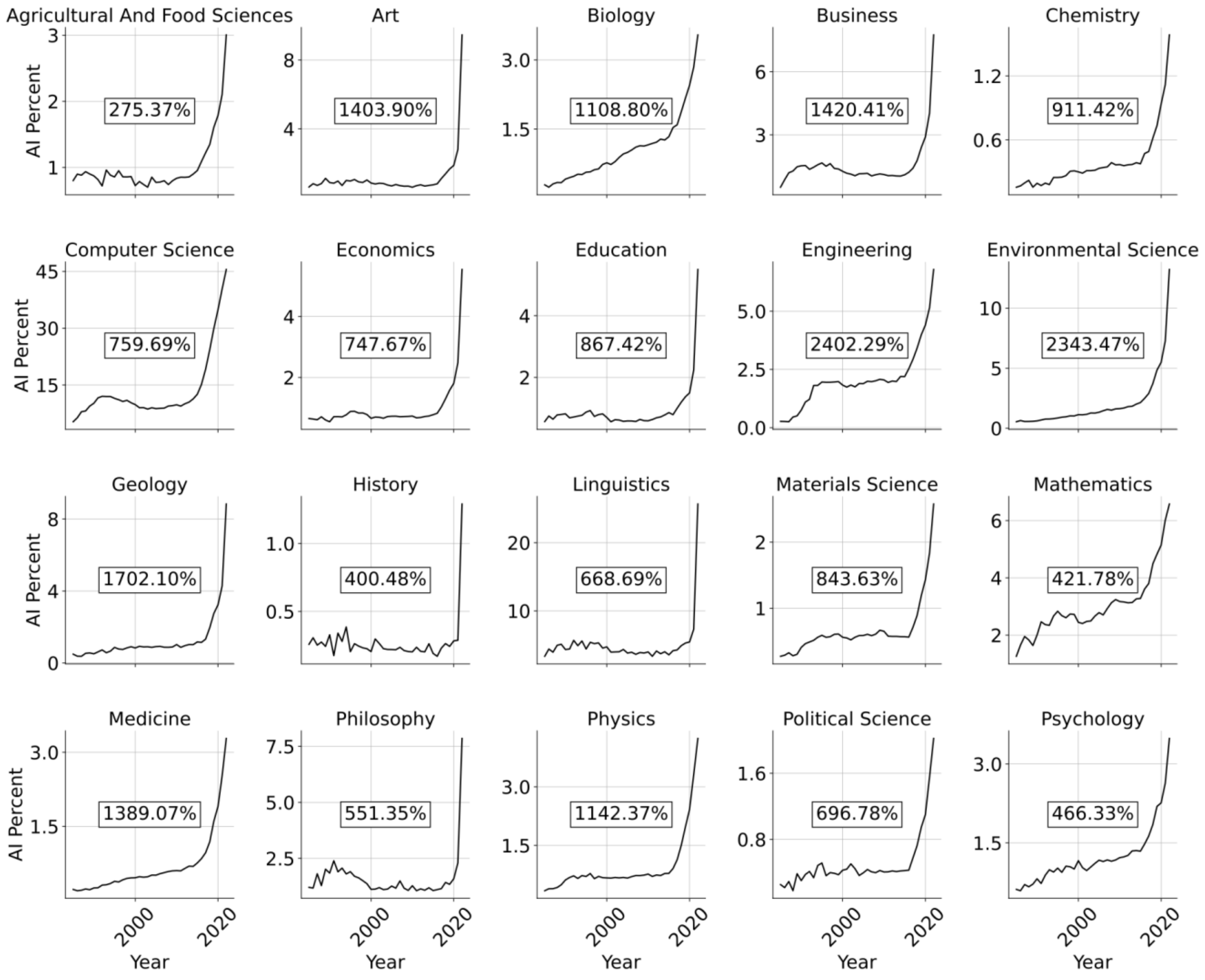


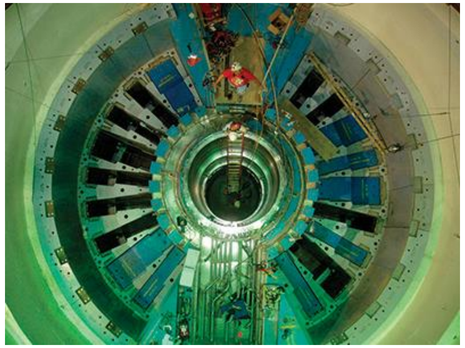
Figure 2: Change in AI engagement percentage from 1985 - 2023 by field. Inserts tally the total change in percentage of AI-engaged publications for each field.

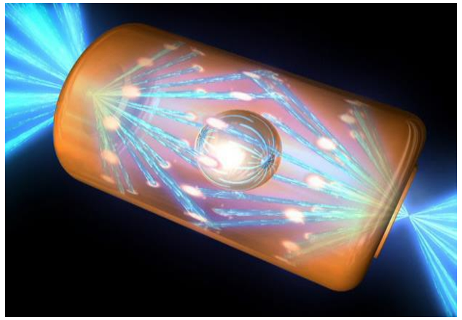
U.S. DEPARTMENT OF
ENERGY

DOE’s Unique Position for AI Leadership

- **Operates the most capable scientific computing systems and the world’s largest collection of advanced experimental facilities**
- Responsible for US nuclear security through deep partnerships across government
- **Largest producer of classified and unclassified scientific data in the world**
- Strongest foundation combining **physical, biological, environmental, energy, mathematical and computing sciences**
- **Largest scientific workforce in the free world**
- Strong ties with private sector technology and energy organizations and stakeholders

experimental facilities and supercomputers





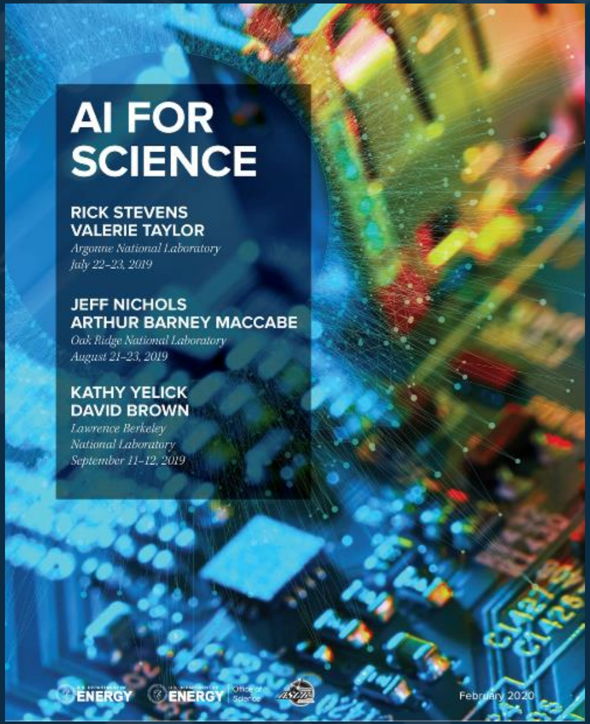
Why DOE Needs to Lead in AI?

1. Effective execution of **DOE mission requires the best and most powerful scientific tools**. AI is ~~becoming~~ such a tool.
2. **Private sector AI efforts alone will not address** the deep scientific and national security requirements of DOE use cases.
3. **Government efforts to regulate AI will require deep expertise** in AI technologies and risks independent from the industry to be regulated DOE can be a trusted neutral party to provide this expertise.
4. **AI offers great opportunity and great risks**, especially in global security. Powerful AI systems in the hands of bad actors poses a new type of asymmetric threat that will need new ideas for risk management.
5. **DOE has the technical ability to lead in AI and partner** with other agencies to advance the use of transformational AI across the government. No other agency is as well placed to do this.

DOE and Lab Planning Efforts

DOE Has Been Gathering Wide Community Input (>1300 researchers)

2019



AI FOR SCIENCE

RICK STEVENS
VALERIE TAYLOR
Argonne National Laboratory
July 22–23, 2019

JEFF NICHOLS
ARTHUR BARNEY MACCABE
Oak Ridge National Laboratory
August 21–23, 2019

KATHY YELICK
DAVID BROWN
Lawrence Berkeley National Laboratory
September 11–12, 2019

February 2020

Much accelerated in three years!

- Language Models (e.g. ChatGPT) released
- Artificial image generation took off
- AI folded a billion proteins
- AI hints at advancing mathematics
- AI automation of computer programming
- Explosion of new AI hardware
- AI accelerates HPC simulations
- Exascale machines start to arrive

2022

ANL 22/91

ADVANCED RESEARCH DIRECTIONS ON
AI FOR SCIENCE, ENERGY, AND SECURITY

Report on Summer 2022 Workshops

Jonathan Carter
Lawrence Berkeley National Laboratory

John Feddema
Sandia National Laboratories

Doug Kothe
Oak Ridge National Laboratory

Rob Neely
Lawrence Livermore National Laboratory

Jason Pruet
Los Alamos National Laboratory

Rick Stevens
Argonne National Laboratory

January 2023

Report posted here:
https://www.anl.gov/ai-for-science-report

2020 DOE Office of Science ASCR Advisory Committee report recommending major DOE AI4S program

U.S. DEPARTMENT OF ENERGY | Office of Science

NNSA National Nuclear Security Administration

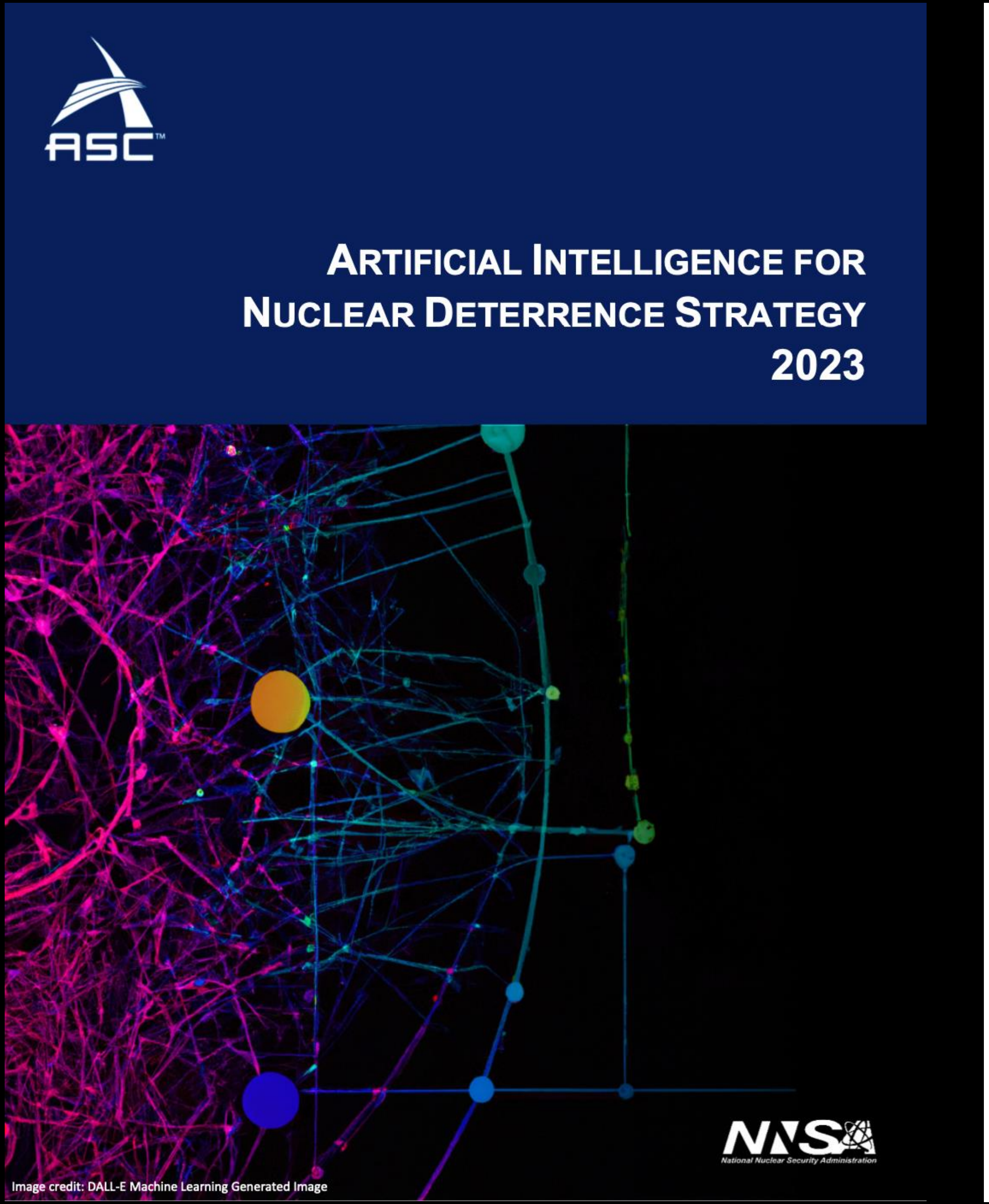
WH Executive Order on AI (October 30, 2023)

AI for Science, Energy, and National Security. Consistent with DOE's priorities in the May 2023 AI for Science, Energy, and Security report, DOE is tasked with expanding **partnerships with industry, academia, other agencies, and international partners** to utilize DOE's computing capabilities and AI testsbeds to **build foundation models that support new applications in science, energy, and national security**, including community preparedness for climate-related risks, enable clean-energy deployment (including addressing delays in permitting reviews), and enhance grid reliability and resilience.

✓

DOE is also charged with issuing a public report " enable the provision of clean, affordable, reliable, resilient, and secure electric power to all describing the potential for AI to improve planning, permitting, investment, and operations for electric grid infrastructure and to Americans." DOE is also tasked as the lead agency, through the National Nuclear Security Administration, to **reduce the risks at the intersection of AI and chemical, biological, radiological, and nuclear (CBRN) threats.** DOE is required to develop testbeds and "tools to evaluate AI capabilities to generate outputs that may represent nuclear, nonproliferation, biological, chemical, critical infrastructure, and energy-security threats or hazards" and "develop model guardrails that reduce such risks."

✓



ARTIFICIAL INTELLIGENCE FOR NUCLEAR DETERRENCE STRATEGY 2023

Image credit: DALL-E Machine Learning Generated Image

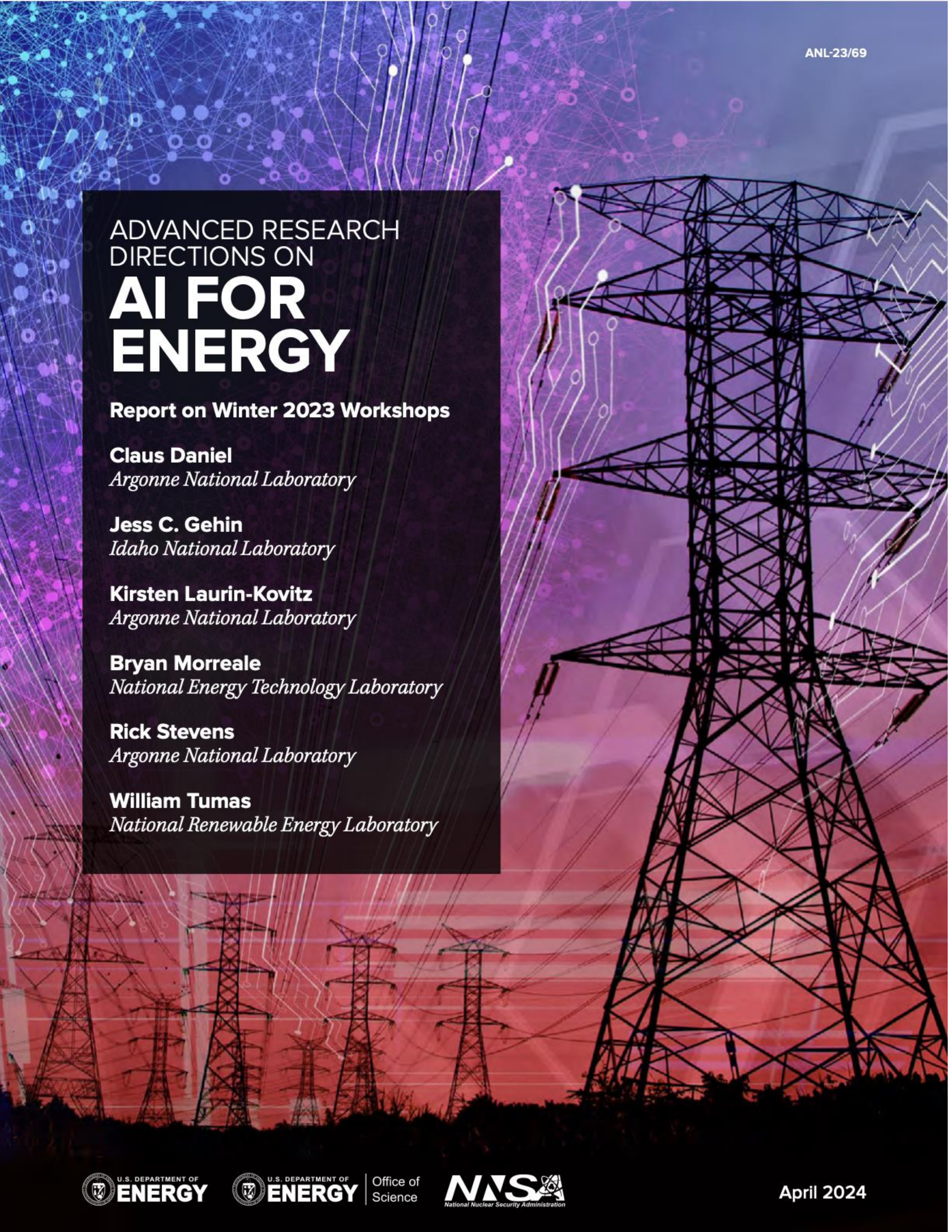
Application of Artificial Intelligence Methods and Technologies to Nuclear Security Mission Areas: the demonstration and application of AI to the Nuclear Security Enterprise and high-consequence applications will be accomplished by partnering with key stakeholders in the weapons design, production, and analysis community.

Foundational R&D in Machine Learning Methods and Technologies: the development of ML tools and techniques that enable successful application in sparse or limited data environments where model accuracy constraints are likely to be much tighter than in industry or academia. In addition, the methods that will be developed will need to scale to the substantial data environment associated with the simulation of complex nuclear physics phenomena.

Scalable and Performant Data Infrastructure: the availability of rich, curated data sets will be critical to the use of ML within ASC. Investment will be required to create a secure hardware and software infrastructure that connects users across the design and production agencies of the Nuclear Security Enterprise. Ensuring the environment is scalable into the future and provides sufficient performance to prevent model training and inference from becoming a bottleneck will be an essential component to a successful execution of this strategy.

Enabling the Data-Driven Workforce of the Future: ASC’s most important asset is its unique workforce of laboratory technical staff who provide expertise in a wide variety of technical areas, including physics, engineering, mathematics, and advanced computing. ASC will invest in training and developing a pipeline of additional staff to engage across projects and activities, with the goal of providing data analytics and complex data-driven modeling. Attracting and retaining the best workforce will likewise mean demonstrating that ASC is performing cutting-edge research in AI methods and applying them to the nation’s most challenging problems. ASC will collaborate with industry, academia, and other U.S. agencies to leverage existing knowledge, experienced staff, and best practices.

AI4E in 2023



ANL-23/69

ADVANCED RESEARCH DIRECTIONS ON

AI FOR ENERGY

Report on Winter 2023 Workshops

Claus Daniel
Argonne National Laboratory

Jess C. Gehin
Idaho National Laboratory

Kirsten Laurin-Kovitz
Argonne National Laboratory

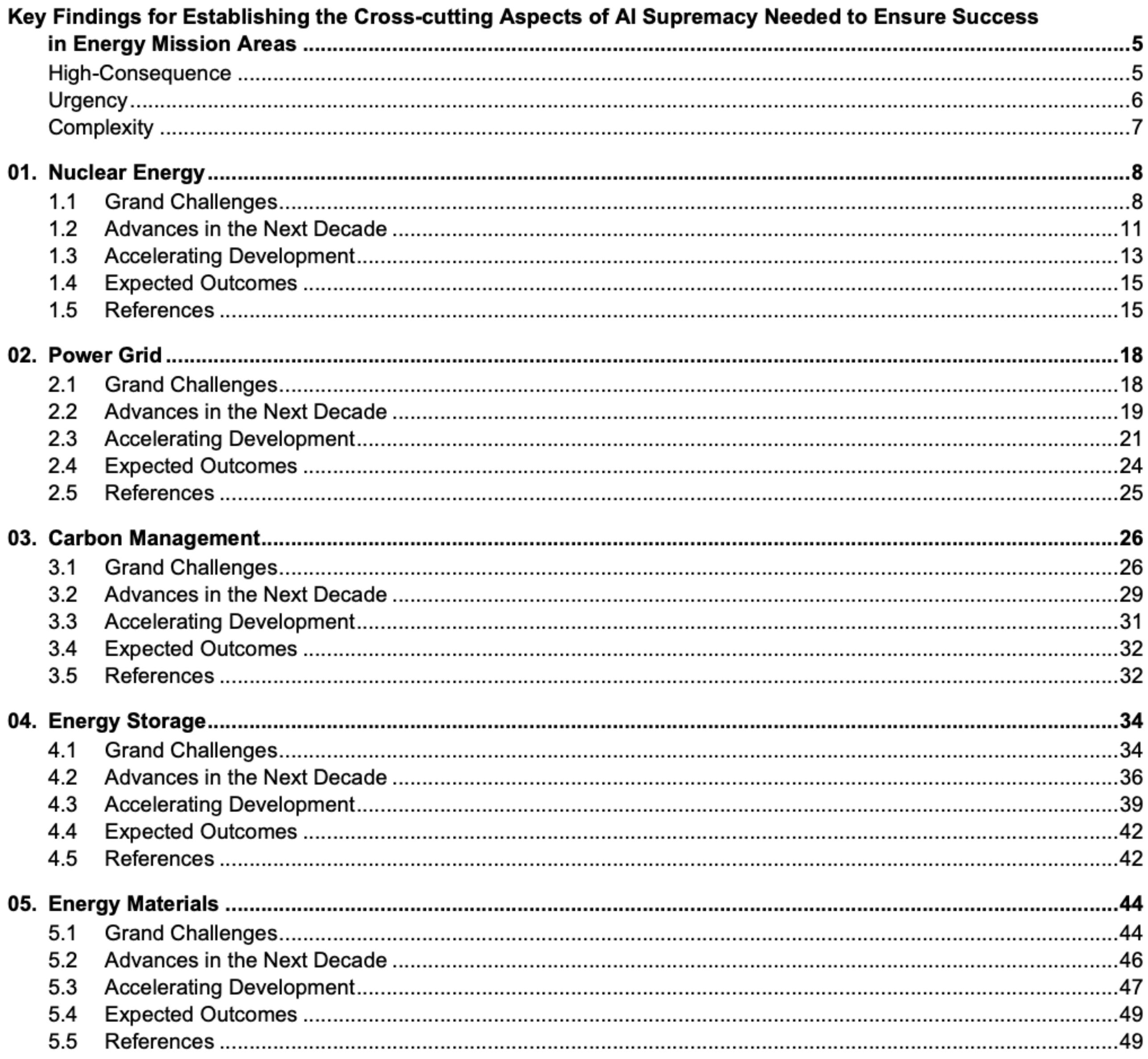
Bryan Morreale
National Energy Technology Laboratory

Rick Stevens
Argonne National Laboratory

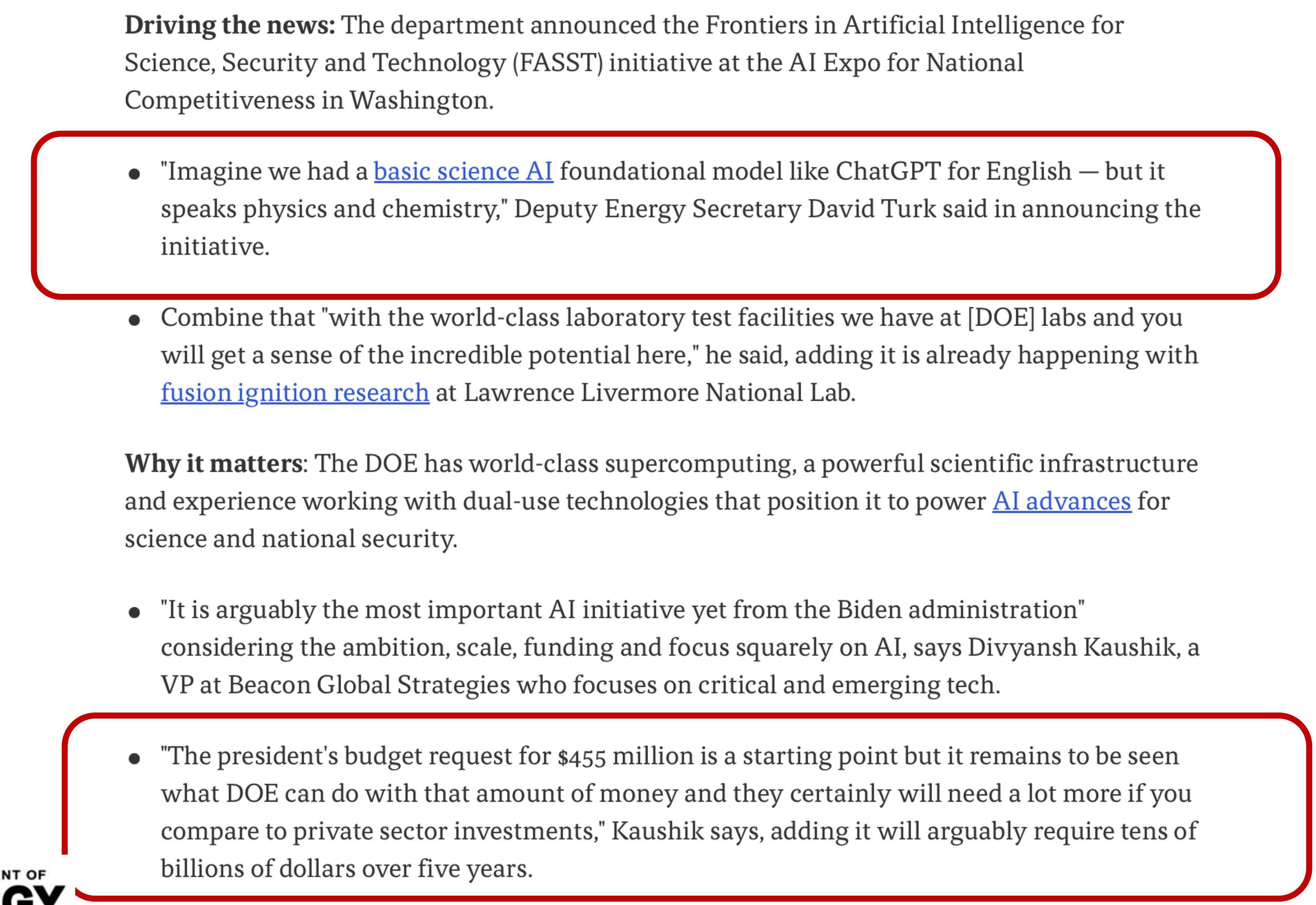
William Tumas
National Renewable Energy Laboratory

U.S. DEPARTMENT OF ENERGY | U.S. DEPARTMENT OF ENERGY Office of Science | NNSA National Nuclear Security Administration

April 2024



May 8th SCSP AI Expo DOE Announcement



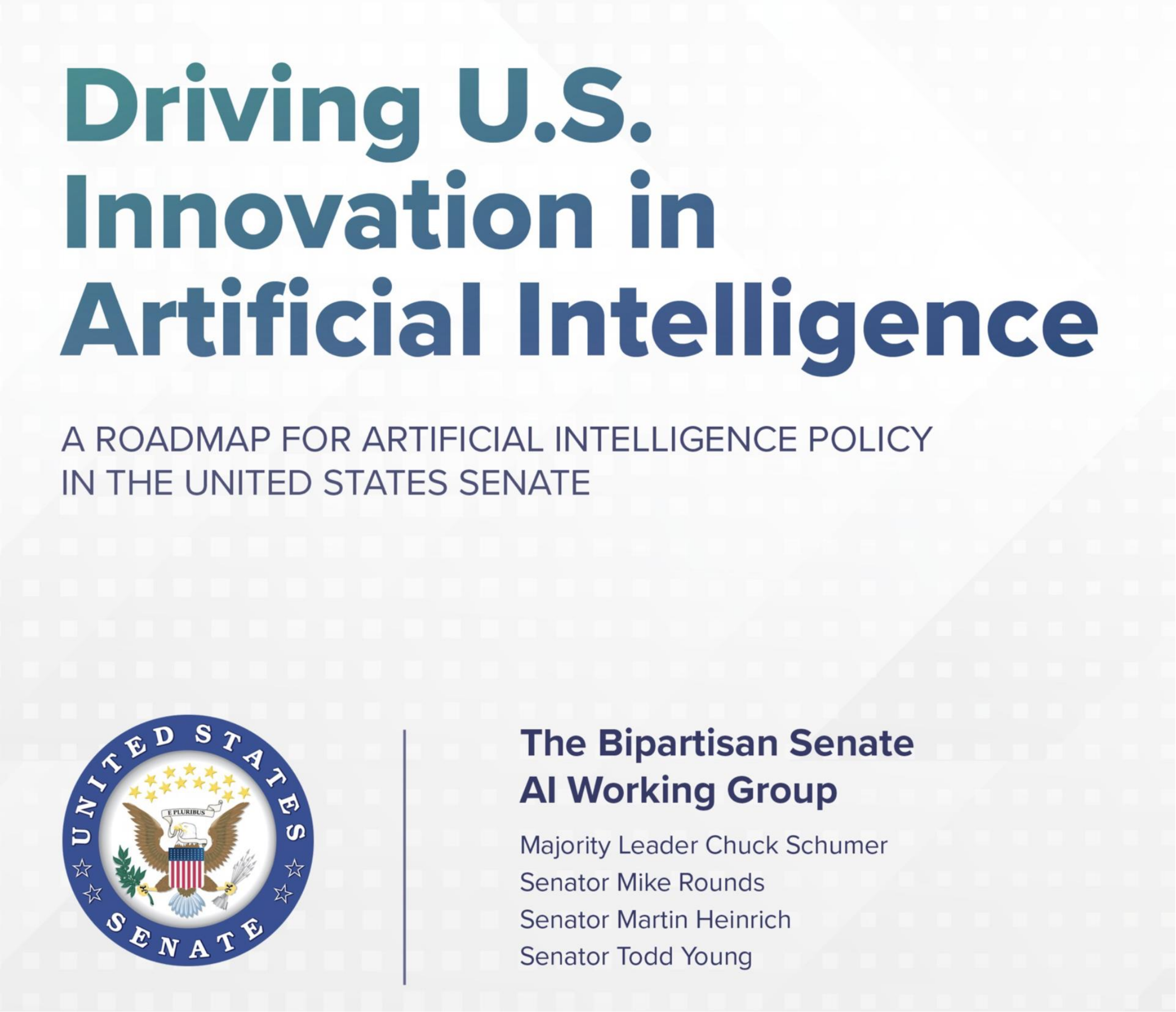
Driving the news: The department announced the Frontiers in Artificial Intelligence for Science, Security and Technology (FASST) initiative at the AI Expo for National Competitiveness in Washington.

- "Imagine we had a basic science AI foundational model like ChatGPT for English — but it speaks physics and chemistry," Deputy Energy Secretary David Turk said in announcing the initiative.
- Combine that "with the world-class laboratory test facilities we have at [DOE] labs and you will get a sense of the incredible potential here," he said, adding it is already happening with fusion ignition research at Lawrence Livermore National Lab.

Why it matters: The DOE has world-class supercomputing, a powerful scientific infrastructure and experience working with dual-use technologies that position it to power AI advances for science and national security.

- "It is arguably the most important AI initiative yet from the Biden administration" considering the ambition, scale, funding and focus squarely on AI, says Divyansh Kaushik, a VP at Beacon Global Strategies who focuses on critical and emerging tech.
- "The president's budget request for $455 million is a starting point but it remains to be seen what DOE can do with that amount of money and they certainly will need a lot more if you compare to private sector investments," Kaushik says, adding it will arguably require tens of billions of dollars over five years.

$32B/yr for non-defense agency AI



Driving U.S. Innovation in Artificial Intelligence

A ROADMAP FOR ARTIFICIAL INTELLIGENCE POLICY
IN THE UNITED STATES SENATE

The Bipartisan Senate AI Working Group

Majority Leader Chuck Schumer
Senator Mike Rounds
Senator Martin Heinrich
Senator Todd Young

May 14th SENATE ROADMAP

We are not talking about your grandparents GOFAI

https://arxiv.org/pdf/2108.07258.pdf

Trend is towards fewer more universal models: increasing emergence and homogenization

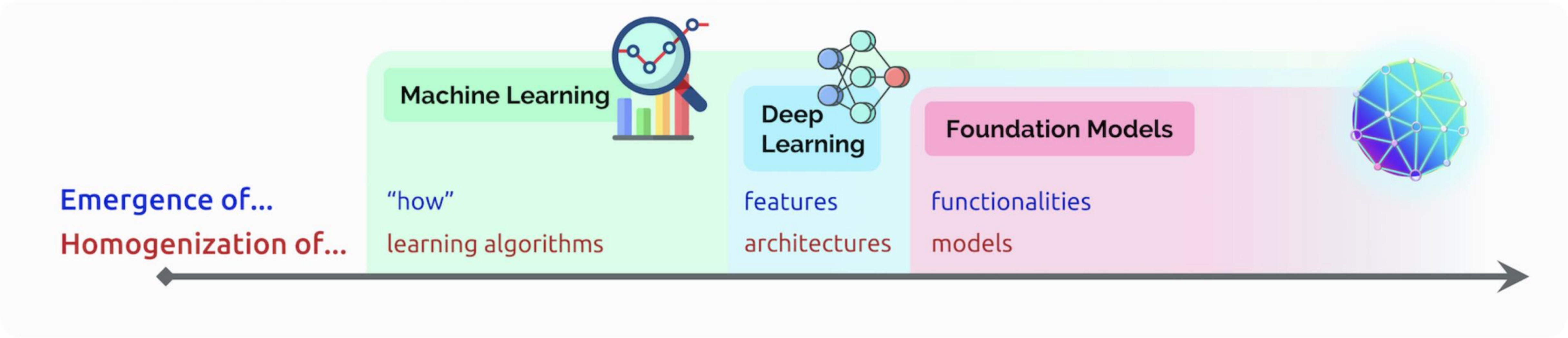
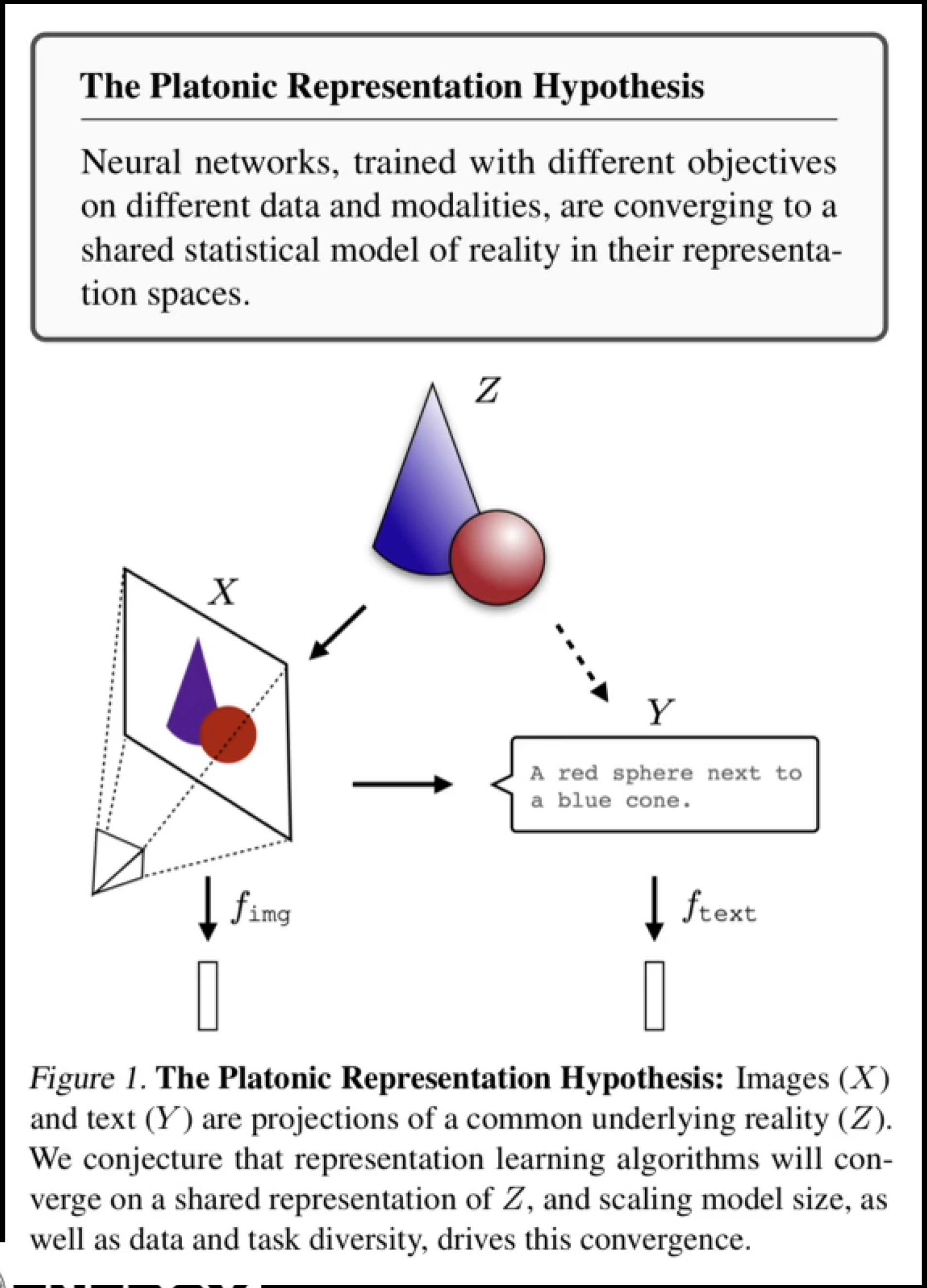


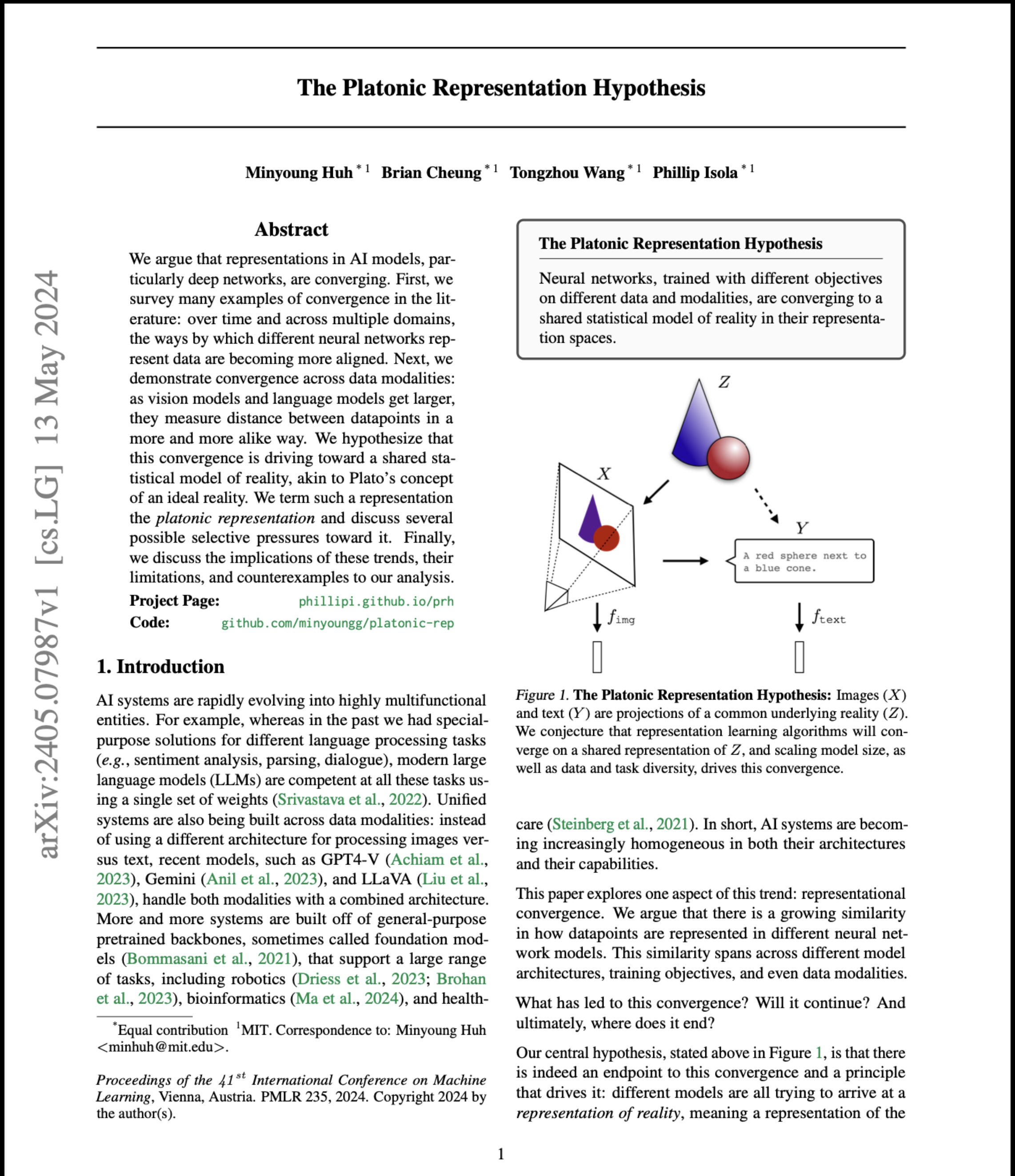
Fig. 1. The story of AI has been one of increasing *emergence* and *homogenization*. With the introduction of machine learning, *how* a task is performed emerges (is inferred automatically) from examples; with deep learning, the high-level features used for prediction emerge; and with foundation models, even advanced functionalities such as in-context learning emerge. At the same time, machine learning homogenizes learning algorithms (e.g., logistic regression), deep learning homogenizes model architectures (e.g., Convolutional Neural Networks), and foundation models homogenizes the model itself (e.g., GPT-3).



The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.

Figure 1. **The Platonic Representation Hypothesis:** Images (X) and text (Y) are projections of a common underlying reality (Z). We conjecture that representation learning algorithms will converge on a shared representation of Z, and scaling model size, as well as data and task diversity, drives this convergence.



The Platonic Representation Hypothesis

Minyoung Huh [*1] Brian Cheung [*1] Tongzhou Wang [*1] Phillip Isola [*1]

Abstract

We argue that representations in AI models, particularly deep networks, are converging. First, we survey many examples of convergence in the literature: over time and across multiple domains, the ways by which different neural networks represent data are becoming more aligned. Next, we demonstrate convergence across data modalities: as vision models and language models get larger, they measure distance between datapoints in a more and more alike way. We hypothesize that this convergence is driving toward a shared statistical model of reality, akin to Plato's concept of an ideal reality. We term such a representation the *platonic representation* and discuss several possible selective pressures toward it. Finally, we discuss the implications of these trends, their limitations, and counterexamples to our analysis.

Project Page: phillipi.github.io/prh
Code: github.com/minyoungg/platonic-rep

The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.

Figure 1. **The Platonic Representation Hypothesis:** Images (X) and text (Y) are projections of a common underlying reality (Z). We conjecture that representation learning algorithms will converge on a shared representation of Z, and scaling model size, as well as data and task diversity, drives this convergence.

1. Introduction

AI systems are rapidly evolving into highly multifunctional entities. For example, whereas in the past we had special-purpose solutions for different language processing tasks (*e.g.*, sentiment analysis, parsing, dialogue), modern large language models (LLMs) are competent at all these tasks using a single set of weights (Srivastava et al., 2022). Unified systems are also being built across data modalities: instead of using a different architecture for processing images versus text, recent models, such as GPT4-V (Achiam et al., 2023), Gemini (Anil et al., 2023), and LLaVA (Liu et al., 2023), handle both modalities with a combined architecture. More and more systems are built off of general-purpose pretrained backbones, sometimes called foundation models (Bommasani et al., 2021), that support a large range of tasks, including robotics (Driess et al., 2023; Brohan et al., 2023), bioinformatics (Ma et al., 2024), and healthcare (Steinberg et al., 2021). In short, AI systems are becoming increasingly homogeneous in both their architectures and their capabilities.

This paper explores one aspect of this trend: representational convergence. We argue that there is a growing similarity in how datapoints are represented in different neural network models. This similarity spans across different model architectures, training objectives, and even data modalities.

What has led to this convergence? Will it continue? And ultimately, where does it end?

Our central hypothesis, stated above in Figure 1, is that there is indeed an endpoint to this convergence and a principle that drives it: different models are all trying to arrive at a *representation of reality*, meaning a representation of the

[*] Equal contribution [1] MIT. Correspondence to: Minyoung Huh <minhuh@mit.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

The 2022 workshops recognized This trend and organized differently

AI4SES Organized on Six Conceptual Clusters

AI for advanced properties inference and inverse design

Energy Storage
Proteins, Polymers,
Stockpile modernization

AI and robotics for autonomous discovery

Materials, Chemistry, Biology
Light-Sources, Neutrons

AI-based surrogates for high-performance computing

Climate Ensembles
Exascale apps with surrogates
1000x faster => Zettascale now

AI for software engineering and programming

Code Translation, Optimization
Quantum Compilation, QAlgs

AI for prediction and control of complex engineered systems

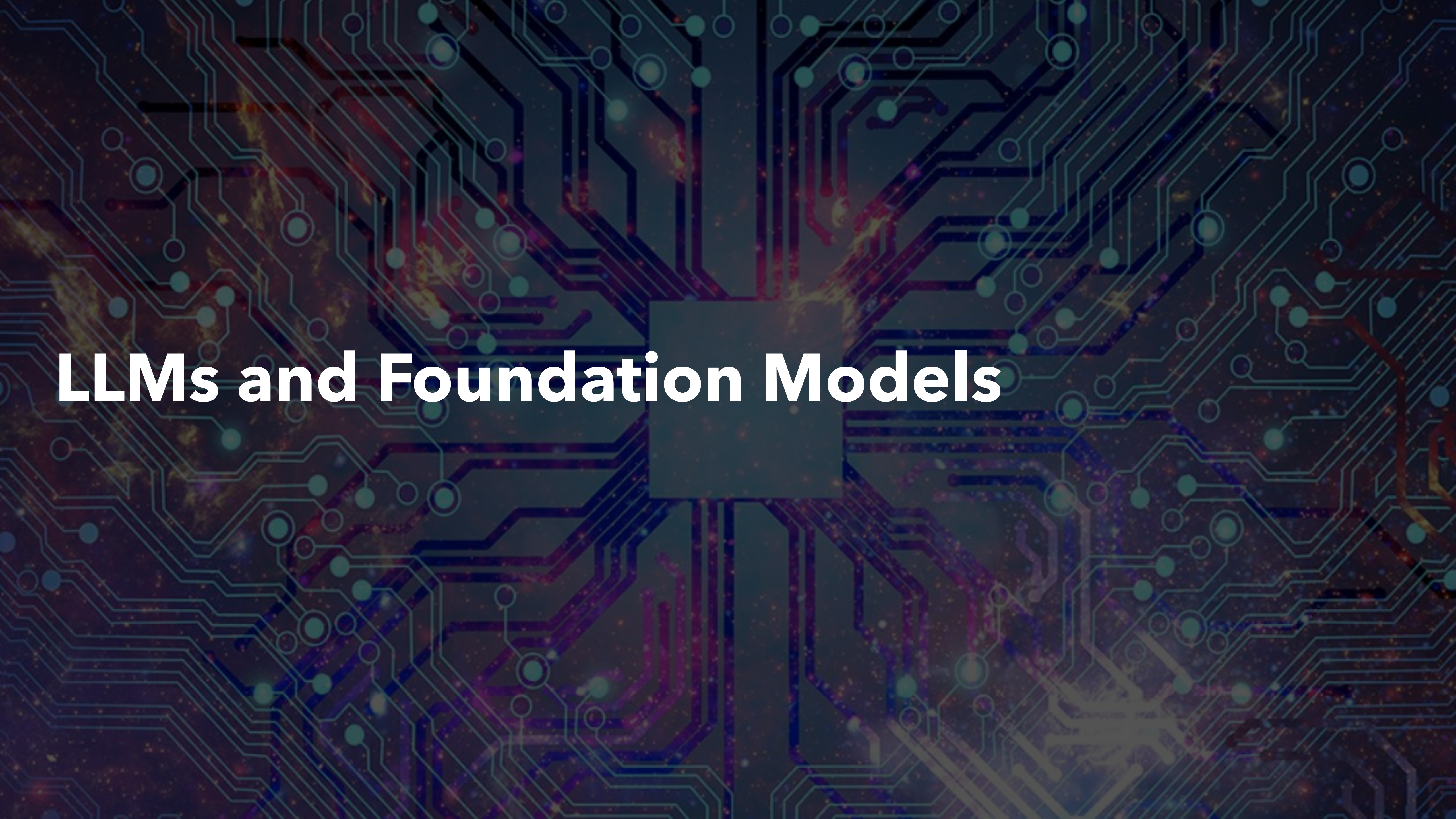
Accelerators, Buildings, Cities
Reactors, Power Grid, Networks

Foundation models for scientific knowledge tasks

Hypothesis Formation, Math
Theory and Modeling Synthesis,

https://www.anl.gov/ai-for-science-report

U.S. DEPARTMENT OF ENERGY



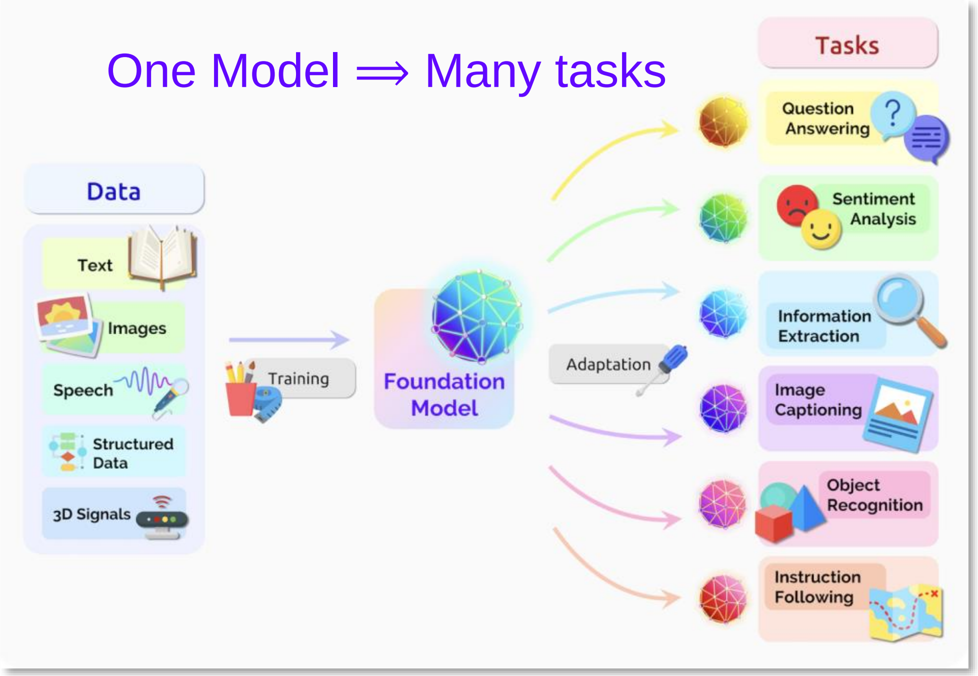
LLMs and Foundation Models

FOUNDATION

Foundation Models — What are they?

https://arxiv.org/pdf/2108.07258.pdf

- **Large scale model trained on large datasets from many sources** (text, papers, datasets, code, molecules, etc.)
- **Additional training to improve the human interaction experience** (e.g., ChatGPT-4o)
- **Large models are remarkably flexible and exhibit emergent behaviors** (capable of tasks not originally trained to do)
- **Applications built on top**
- There are multiple early efforts underway in DOE labs to create Foundation Models explicitly targeting scientific use cases



Trained on trillions of input ”tokens”
for many weeks on a large-scale computers

SOTA models (GPT-4) have about
1.8 trillion parameters (~1% brainscale)

U.S. DEPARTMENT OF ENERGY

Foundation Models for Science — Opportunities

- **FMs can summarize and distill knowledge** – extract information from million of papers into compact computing representation – PPI networks, materials compositions, code kernels, biological function, etc.
- **FMs can synthesize** – combine information from multiple sources – generate small programs for specific tasks – quantum computing programs using QISkit & Cirq, derivations for applied physics, code for visualization and animation, etc.
- **FMs can generate plans, solve logic problems** and write experimental protocols for robots – powering self-driving labs, generate strategies for problem solving, and planning for testing hypotheses
- **FMs can generate hypotheses to be tested and perhaps eventually new theories for exploration – a full-time shared scientific assistant that learns from across all of science is possible**

Can ChatGPT be used to generate scientific hypotheses?

Yang Jeong Park[1,2], Daniel Kaplan[3], Zhichu Ren[4], Chia-Wei Hsu[4], Changhao Li[1], Haowei Xu[1], Sipei Li[1] and Ju Li[1,4,*]

[1] Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[2] Institute of New Media and Communications, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
[3] Department of Condensed Matter Physics, Weizmann Institute of Science, Rehovot 7610001, Israel
[4] Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[*]Corresponding Author: liju@mit.edu

Abstract

We investigate whether large language models can perform the creative hypothesis generation that human researchers regularly do. While the error rate is high, generative AI seems to be able to effectively structure vast amounts of scientific knowledge and provide interesting and testable hypotheses. The future scientific enterprise may include synergistic efforts with a swarm of "hypothesis machines", challenged by automated experimentation and adversarial peer reviews.

In a university or research institute, a significant portion of fresh ideas arises out of discussions. Can talking to ChatGPT-4,[1] OpenAI's latest chatbot, create genuinely interesting scientific hypotheses?

In the past, only humans generated interesting hypotheses. Computers have been used to perform numerical simulations or even to prove theorems, like the four-color theorem in 1976[2]. But making interesting laboratory-testable hypotheses with artificial intelligence (AI) seems far-fetched, until recently.

We are a collaborative group of experimental and theoretical researchers in physical sciences and engineering. Generative Pre-trained Transformer (GPT-4), released on March 14, 2023, is a large language model (LLM) significantly bigger than its predecessor GPT-3 released in 2020 (already with 1.75×10^{11} parameters). GPT-4 neural network was trained on a text corpus of books, webpages, academic papers from various disciplines, discussion forums, etc., up to September 2021. After experimenting with GPT-4 in our own research domains in materials chemistry, physics and quantum information, we find that ChatGPT-4 is knowledgeable, frequently wrong, and interesting to talk to. In other words, not unlike a college professor or a colleague.

To make everything concrete, our operative definition of "genuinely interesting scientific hypotheses" is (a) whether after a conversation, some experienced practitioner of a field can feel

> After experimenting with GPT-4 in our own research domains in materials chemistry, physics and quantum information, we find that ChatGPT-4 is knowledgeable, frequently wrong, and interesting to talk to. In other words, not unlike a college professor or a colleague. https://arxiv.org/pdf/2304.12208.pdf

https://arxiv.org/pdf/2108.07258.pdf

FOUNDATION MODEL

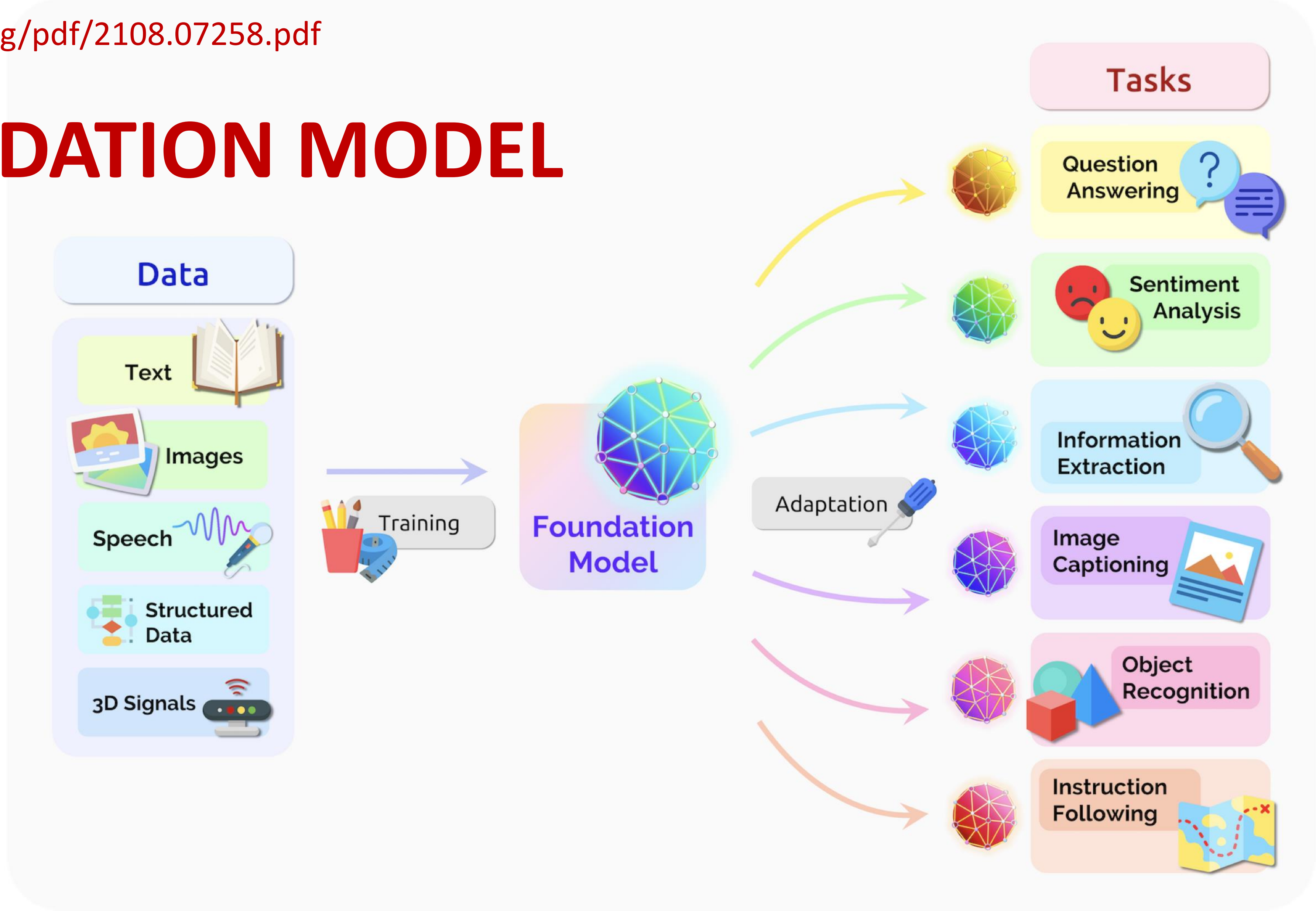
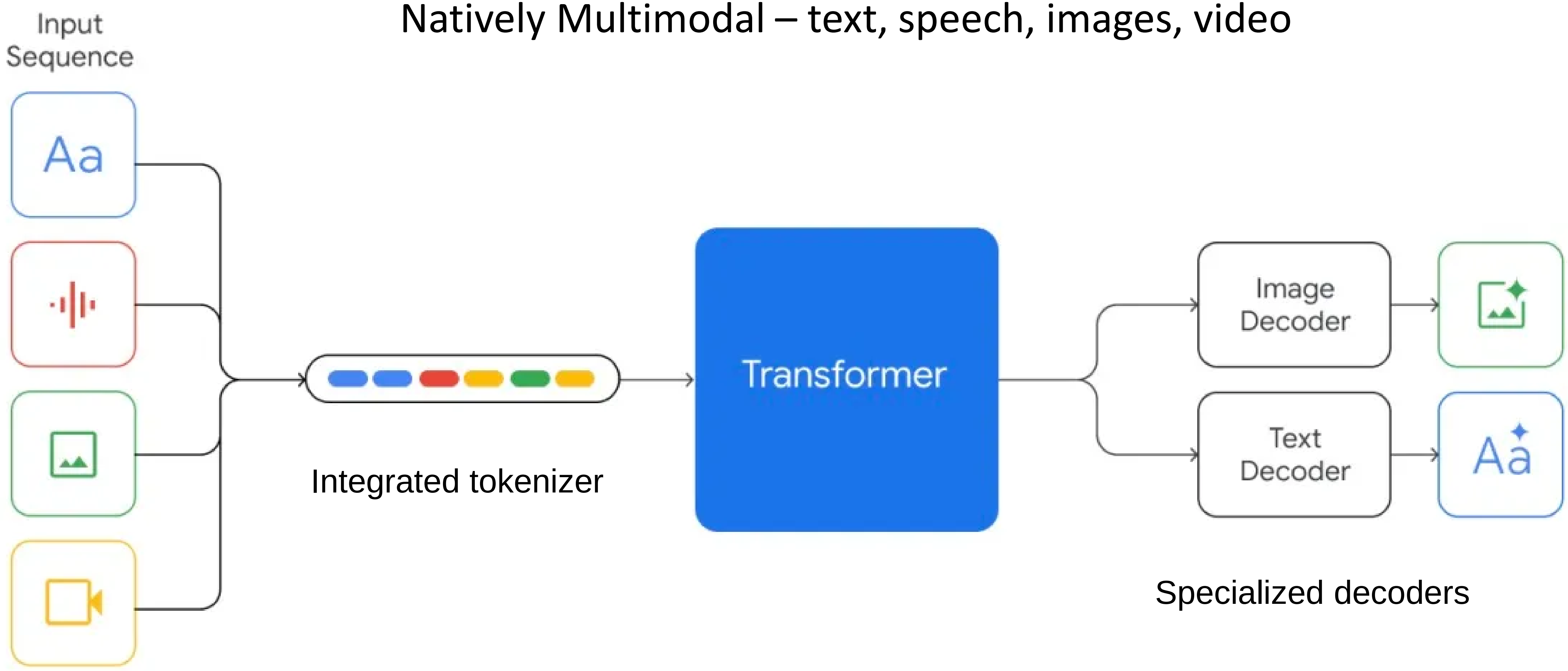


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

Google Gemini

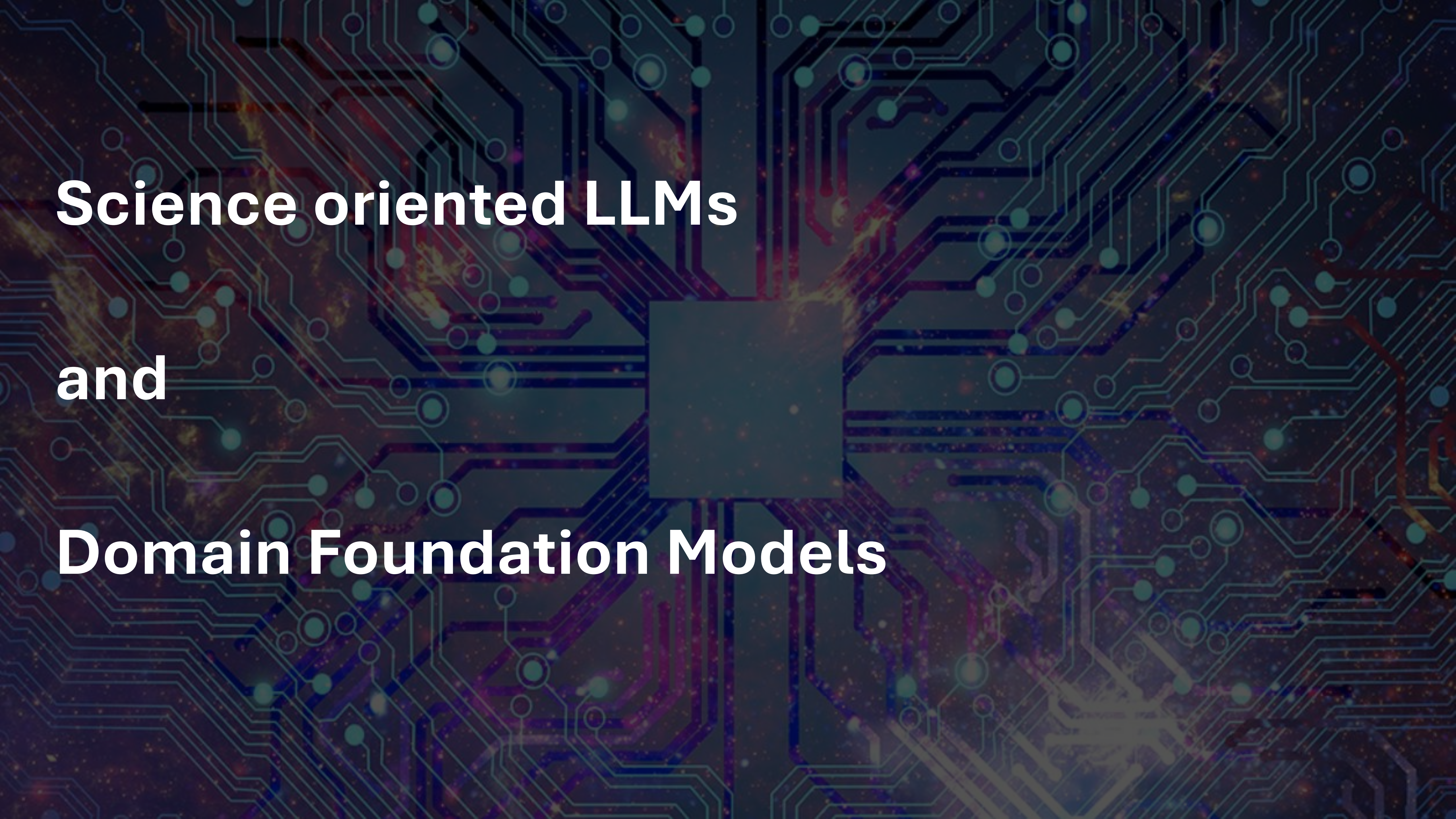
Natively Multimodal – text, speech, images, video



Integrated tokenizer

Specialized decoders

https://blog.google/technology/ai/google-gemini-ai/



Science oriented LLMs

and

Domain Foundation Models

FASST – Cartoon

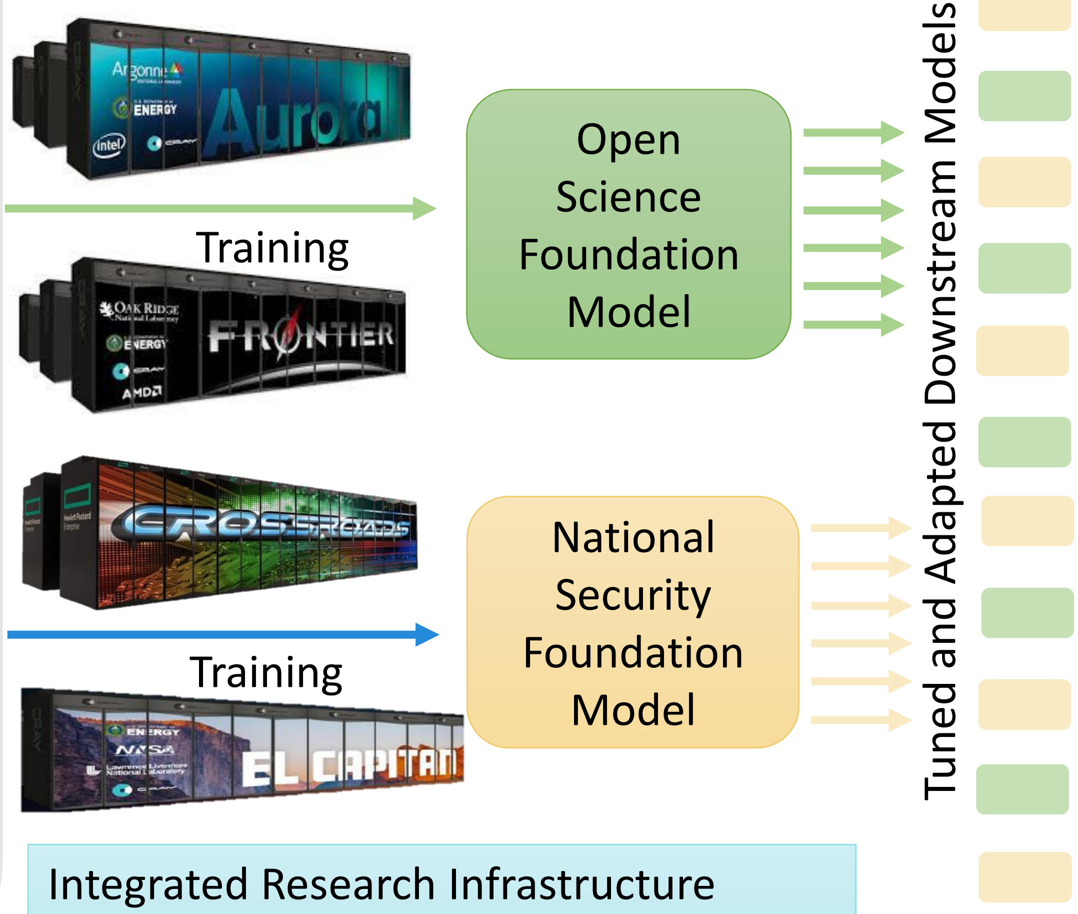
Scientific & Engineering Datasets

Mathematics
Biology
Materials
Chemistry
Particle Physics
Nuclear Physics
Computer Science
Climate
Medicine
Cosmology
Fusion Energy
Accelerators
Reactors
Energy Systems
Manufacturing

Text and Code Corpora

General Text
Media
News
Humanities
History
Law
Digital Libraries
OSTI Archive
Scientific Journals
arXiv
Code repositories
Data.gov
PubMed
Agency Archives

DOE and NNSA Exascale Systems
FASST Common AI Software Frameworks
FASST Responsible AI Techniques



Training

Open Science Foundation Model

Training

National Security Foundation Model

Tuned and Adapted Downstream Models

Argonne
ENERGY
intel
CRAY
Aurora
OAK RIDGE National Laboratory
ENERGY
CRAY
AMD
FRONTIER
Hewlett Packard Enterprise
CROSSROADS
ENERGY
NNSA
Lawrence Livermore National Laboratory
CRAY
EL CAPITAN

Integrated Research Infrastructure
Online Experimental Facilities
Strategic Partnerships

Exemplar DOE Mission Tasks

Scientific Discovery

Digital Twins

Inverse Design

Code Optimization

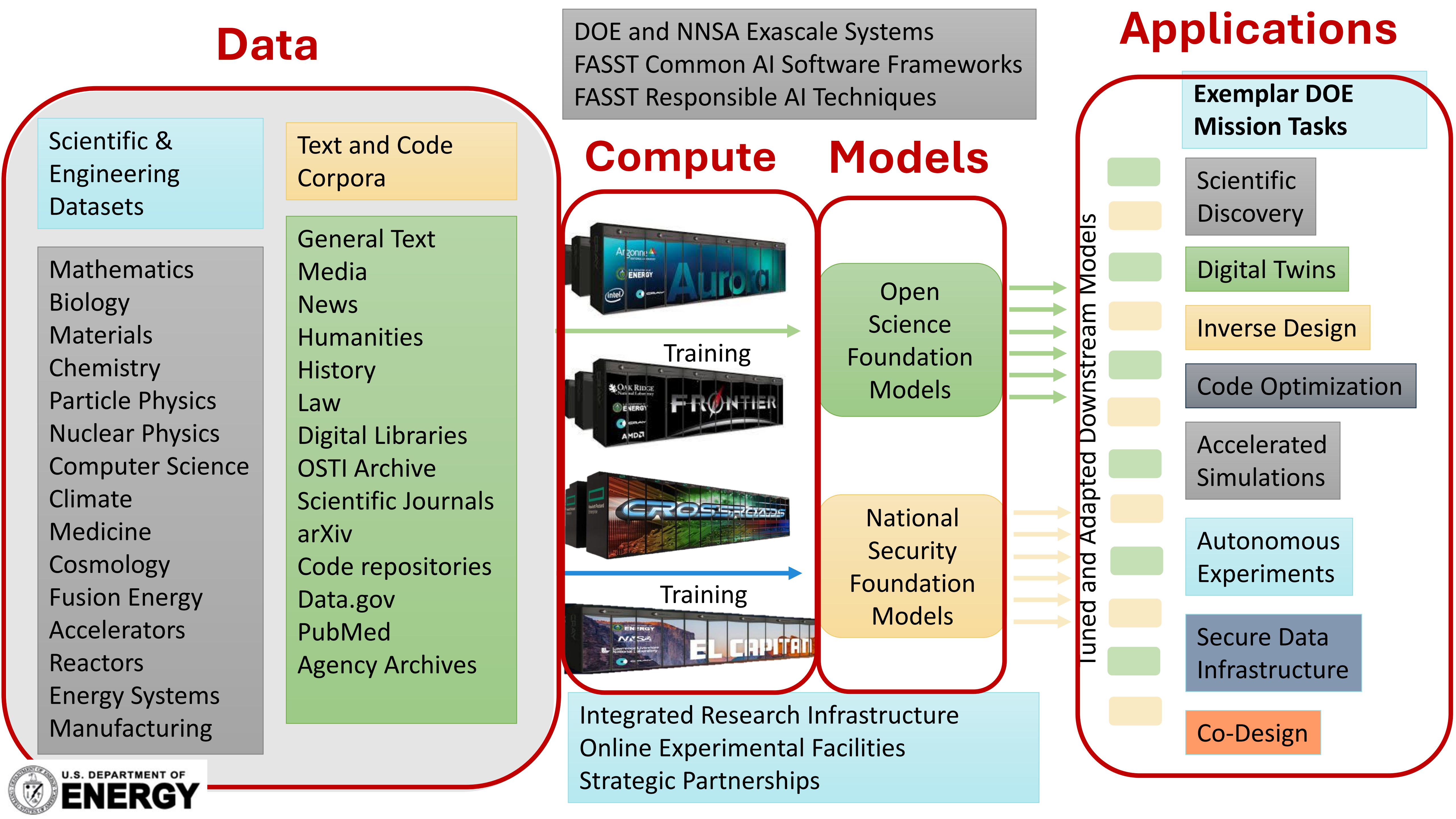
Accelerated Simulations

Autonomous Experiments

Secure Data Infrastructure

Co-Design

U.S. DEPARTMENT OF ENERGY



Data

Scientific & Engineering Datasets

Mathematics
Biology
Materials
Chemistry
Particle Physics
Nuclear Physics
Computer Science
Climate
Medicine
Cosmology
Fusion Energy
Accelerators
Reactors
Energy Systems
Manufacturing

Text and Code Corpora

General Text
Media
News
Humanities
History
Law
Digital Libraries
OSTI Archive
Scientific Journals
arXiv
Code repositories
Data.gov
PubMed
Agency Archives

DOE and NNSA Exascale Systems
FASST Common AI Software Frameworks
FASST Responsible AI Techniques

Compute

Training

Training

Models

Open Science Foundation Models

National Security Foundation Models

Integrated Research Infrastructure
Online Experimental Facilities
Strategic Partnerships

Applications

Tuned and Adapted Downstream Models

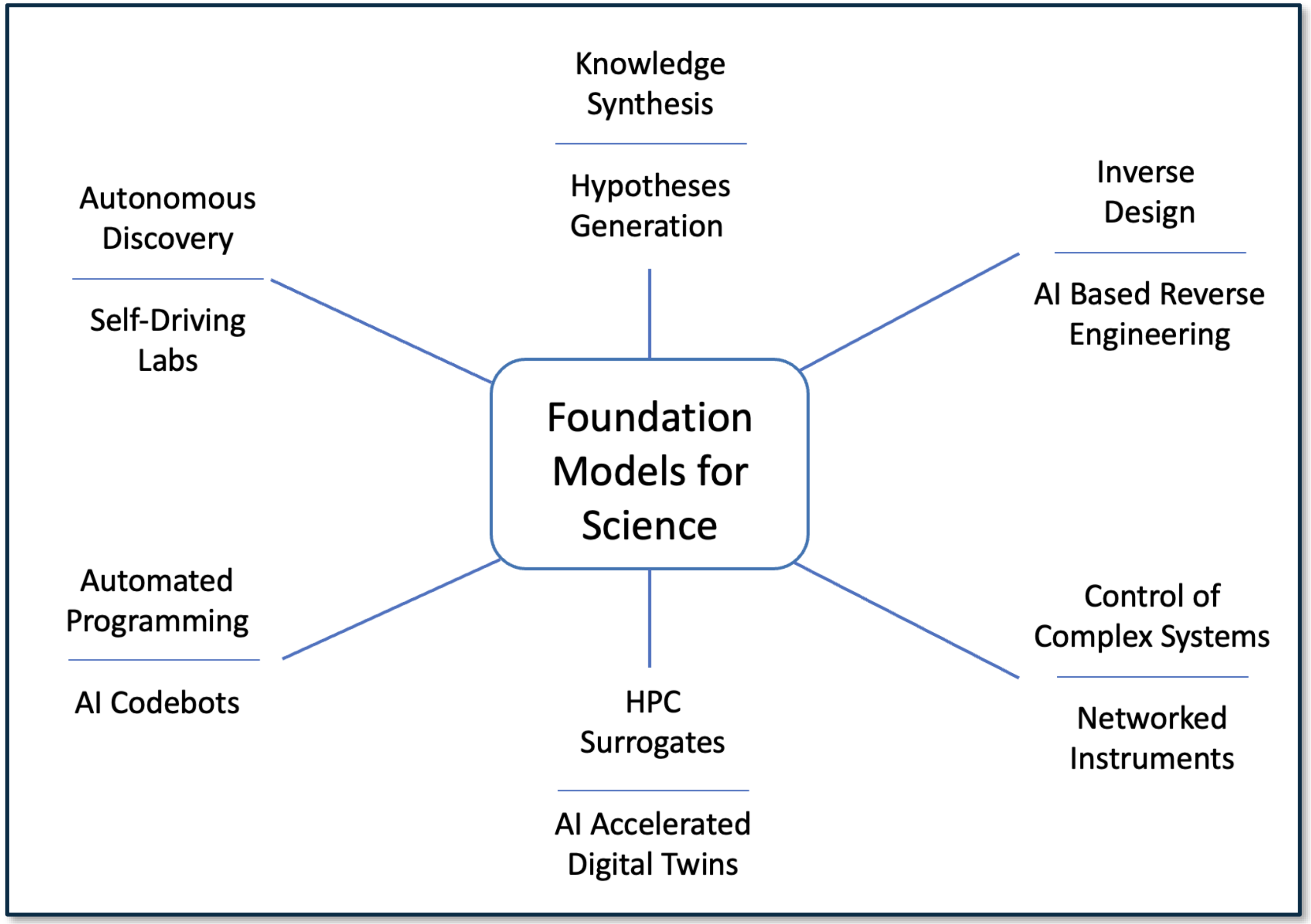
Exemplar DOE Mission Tasks

Scientific Discovery
Digital Twins
Inverse Design
Code Optimization
Accelerated Simulations
Autonomous Experiments
Secure Data Infrastructure
Co-Design

U.S. DEPARTMENT OF ENERGY

It is likely that many of the use cases we imagine can be driven directly or indirectly from sufficiently powerful Foundation Models but we should not limit our thinking to FMs

Advanced AI systems often integrate many tools and technologies



Foundation Models for Science

Knowledge Synthesis

Hypotheses Generation

Inverse Design

AI Based Reverse Engineering

Control of Complex Systems

Networked Instruments

HPC Surrogates

AI Accelerated Digital Twins

Automated Programming

AI Codebots

Autonomous Discovery

Self-Driving Labs

Foundation Models Mapped Against DOE Mission Areas

Clean Energy Systems Model: This model would focus on the physics, chemistry, and engineering principles underlying various renewable energy sources, hydrogen production, solar energy, wind energy, and storage technologies. It can be fine-tuned for specific energy systems, optimization of energy output, and efficiency improvements.

Smart Grid and Infrastructure Model: This model would encompass electrical engineering, network optimization, smart grid technologies, and energy systems management. It can be fine-tuned for specific applications like EV infrastructure planning, grid resilience strategies, and smart grid implementations.

Computational Intelligence Model: This model would integrate capabilities in high-performance computing, machine learning, quantum algorithms, computer science, mathematics, computer architecture, data science, advanced data analysis, applied mathematics, and parallel computing. It can be fine-tuned for applications in AI, complex simulations, and computational research.

Environmental Sciences Model: This model would focus on climate modeling, environmental impact assessments, atmospheric science, climate mitigation strategies, climate risk assessment, bio-geosphere interactions, climate engineering, and biological systems in the environment. It can be fine-tuned for specific environmental applications and climate studies.

Materials and Chemical Sciences Model: This model would cover computational chemistry, materials discovery, molecular dynamics, manufacturing processes, inverse design of materials and systems, self-driving laboratories, and autonomous discovery. It can be fine-tuned for developing new materials, chemical processes, and manufacturing techniques for energy applications.

Biological Systems Model: This model would cover the study of biological systems, including genomics, synthetic biology, microbiology, environmental biology, engineering plants, medicine, protein design, self-driving laboratories, autonomous discovery, and microbial engineering. It can be fine-tuned for applications in environmental biology, plant engineering, biotechnology, and medical research.

Nuclear Security Model: This model would integrate nuclear physics, engineering, security protocols, reactor technologies, nuclear fission, automated reactor design, and reactor control. It can be fine-tuned for nuclear energy applications, nonproliferation technologies, and national security measures.

High-Energy and Particle Physics Model: This model would focus on the principles of high-energy physics, nuclear reactions, particle physics, accelerators, and cosmology. It can be fine-tuned for applications in experimental physics, particle accelerators, and fundamental research in physics.

Advanced Manufacturing Model: This model would focus on manufacturing technologies, including inverse design, process optimization, supply chain optimization, applied materials, precision manufacturing, self-driving laboratories, and autonomous discovery. It can be fine-tuned for specific applications in optimizing manufacturing processes and supply chains.

Carbon Management Model: This model would integrate knowledge on the physics and chemistry of capturing CO2, managing CO2 flows, carbon storage, conversion to fuels, and direct air capture. It aims to support the design and analysis of CO2 management systems and advance research into the fundamentals of carbon management.

Knowledge Integration Model: This model would integrate scientific literature, codes, texts, and tutorials to support knowledge extraction, synthesis, and automated hypothesis generation. It aims to advance theory and experimental design, forming the core of a system that interacts with humans and manages interactions with other foundation models included in this list.

Clean Energy Systems Model: This model would focus on the physics, chemistry, and engineering principles underlying various renewable energy sources, hydrogen production, solar energy, wind energy, and storage technologies. It can be fine-tuned for specific energy systems, optimization of energy output, and efficiency improvements.

Smart Grid and Infrastructure Model: This model would encompass electrical engineering, network optimization, smart grid technologies, and energy systems management. It can be fine-tuned for specific applications like EV infrastructure planning, grid resilience strategies, and smart grid implementations.

Computational Intelligence Model: This model would integrate capabilities in high-performance computing, machine learning, quantum algorithms, computer science, mathematics, computer architecture, data science, advanced data analysis, applied mathematics, and parallel computing. It can be fine-tuned for applications in AI, complex simulations, and computational research.

Environmental Sciences Model: This model would focus on climate modeling, environmental impact assessments, atmospheric science, climate mitigation strategies, climate risk assessment, bio-geosphere interactions, climate engineering, and biological systems in the environment. It can be fine-tuned for specific environmental applications and climate studies.

Materials and Chemical Sciences Model: This model would cover computational chemistry, materials discovery, molecular dynamics, manufacturing processes, inverse design of materials and systems, self-driving laboratories, and autonomous discovery. It can be fine-tuned for developing new materials, chemical processes, and manufacturing techniques for energy applications.

2

Biological Systems Model: This model would cover the study of biological systems, including genomics, synthetic biology, microbiology, environmental biology, engineering plants, medicine, protein design, self-driving laboratories, autonomous discovery, and microbial engineering. It can be fine-tuned for applications in environmental biology, plant engineering, biotechnology, and medical research.

3

Nuclear Security Model: This model would integrate nuclear physics, engineering, security protocols, reactor technologies, nuclear fission, automated reactor design, and reactor control. It can be fine-tuned for nuclear energy applications, nonproliferation technologies, and national security measures.

High-Energy and Particle Physics Model: This model would focus on the principles of high-energy physics, nuclear reactions, particle physics, accelerators, and cosmology. It can be fine-tuned for applications in experimental physics, particle accelerators, and fundamental research in physics.

1

Advanced Manufacturing Model: This model would focus on manufacturing technologies, including inverse design, process optimization, supply chain optimization, applied materials, precision manufacturing, self-driving laboratories, and autonomous discovery. It can be fine-tuned for specific applications in optimizing manufacturing processes and supply chains.

Carbon Management Model: This model would integrate knowledge on the physics and chemistry of capturing CO2, managing CO2 flows, carbon storage, conversion to fuels, and direct air capture. It aims to support the design and analysis of CO2 management systems and advance research into the fundamentals of carbon management.

Knowledge Integration Model: This model would integrate scientific literature, codes, texts, and tutorials to support knowledge extraction, synthesis, and automated hypothesis generation. It aims to advance theory and experimental design, forming the core of a system that interacts with humans and manages interactions with other foundation models included in this list.

Foundation Models in High-Energy Physics

1

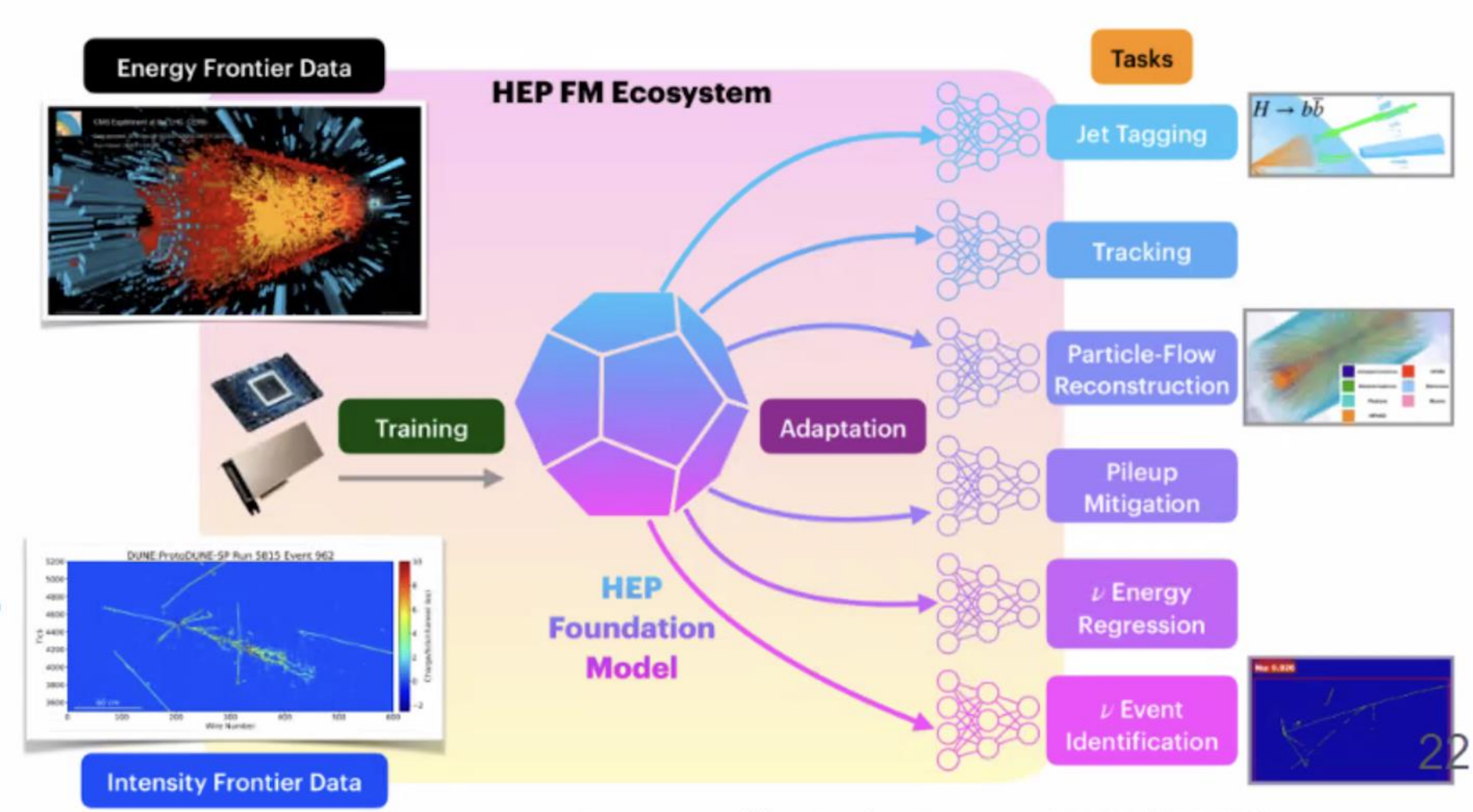
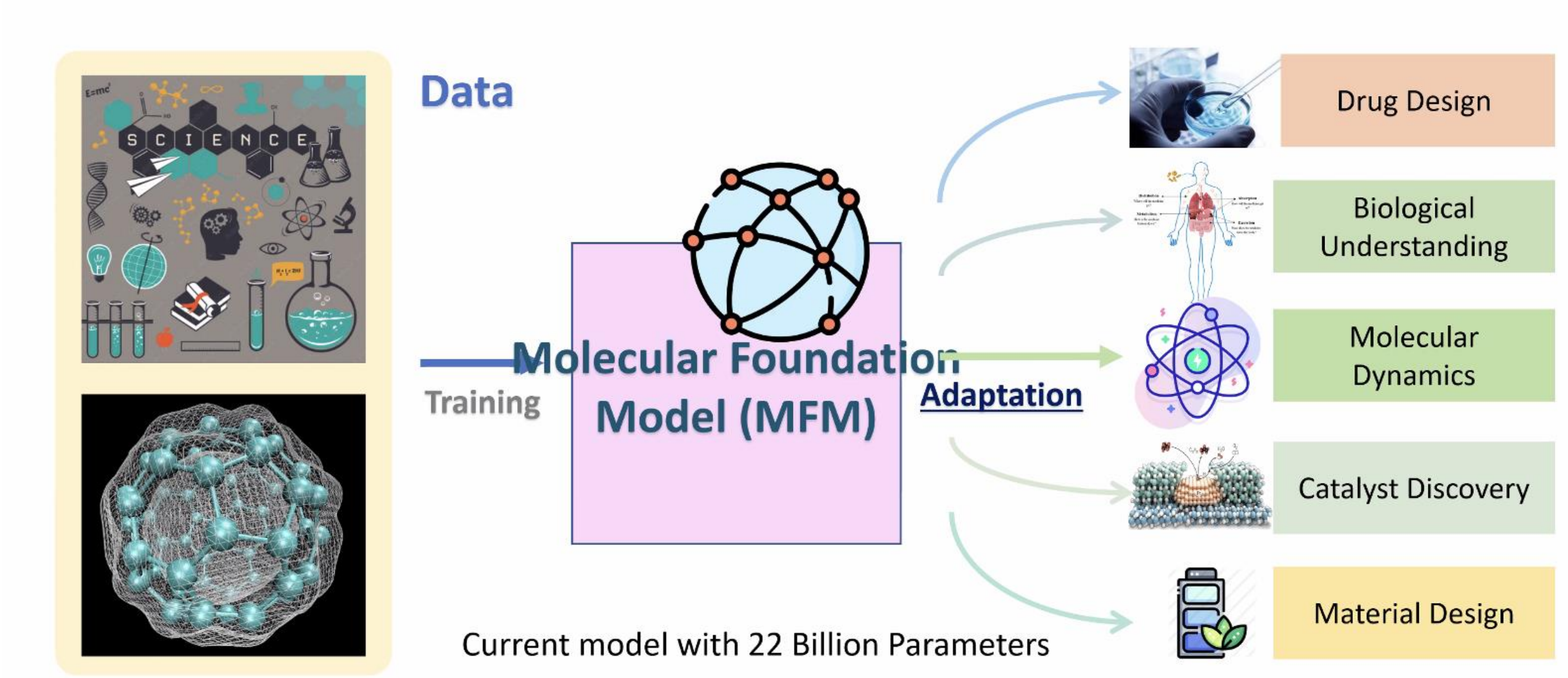


Image credit: Javier Duarte (CMS/UCSD)

Molecular Foundation Model “Distributional Graphormer”



Data

Training

Molecular Foundation Model (MFM)

Adaptation

Drug Design

Biological Understanding

Molecular Dynamics

Catalyst Discovery

Material Design

Current model with 22 Billion Parameters

MatterGen – Microsoft AI4Science – Materials Design with Stable Diffusion

https://arxiv.org/pdf/2312.03687.pdf

MatterGen: a generative model for inorganic materials design

Claudio Zeni[1†], Robert Pinsler[1†], Daniel Zügner[1†], Andrew Fowler[1†], Matthew Horton[1†], Xiang Fu[1], Sasha Shysheya[1], Jonathan Crabbé[1], Lixin Sun[1], Jake Smith[1], Ryota Tomioka[1*], Tian Xie[1*]

[1]Microsoft Research AI4Science.

*Corresponding author(s). E-mail(s): ryoto@microsoft.com; tianxie@microsoft.com;
†Equal contribution; the authors are listed in random order.

Abstract

The design of functional materials with desired properties is essential in driving technological advances in areas like energy storage, catalysis, and carbon capture [1–3]. Generative models provide a new paradigm for materials design by directly generating entirely novel materials given desired property constraints. Despite recent progress, current generative models have low success rate in proposing stable crystals, or can only satisfy a very limited set of property constraints [4–13]. Here, we present MatterGen, a model that generates stable, diverse inorganic materials across the periodic table and can further be fine-tuned to steer the generation towards a broad range of property constraints. To enable this, we introduce a new diffusion-based generative process that produces crystalline structures by gradually refining atom types, coordinates, and the periodic lattice. We further introduce adapter modules to enable fine-tuning towards any given property constraints with a labeled dataset. Compared to prior generative models, structures produced by MatterGen are more than twice as likely to be novel and stable, and more than 15 times closer to the local energy minimum. After fine-tuning, MatterGen successfully generates stable, novel materials with desired chemistry, symmetry, as well as mechanical, electronic and magnetic properties. Finally, we demonstrate multi-property materials design capabilities by proposing both high magnetic density and a chemical composition

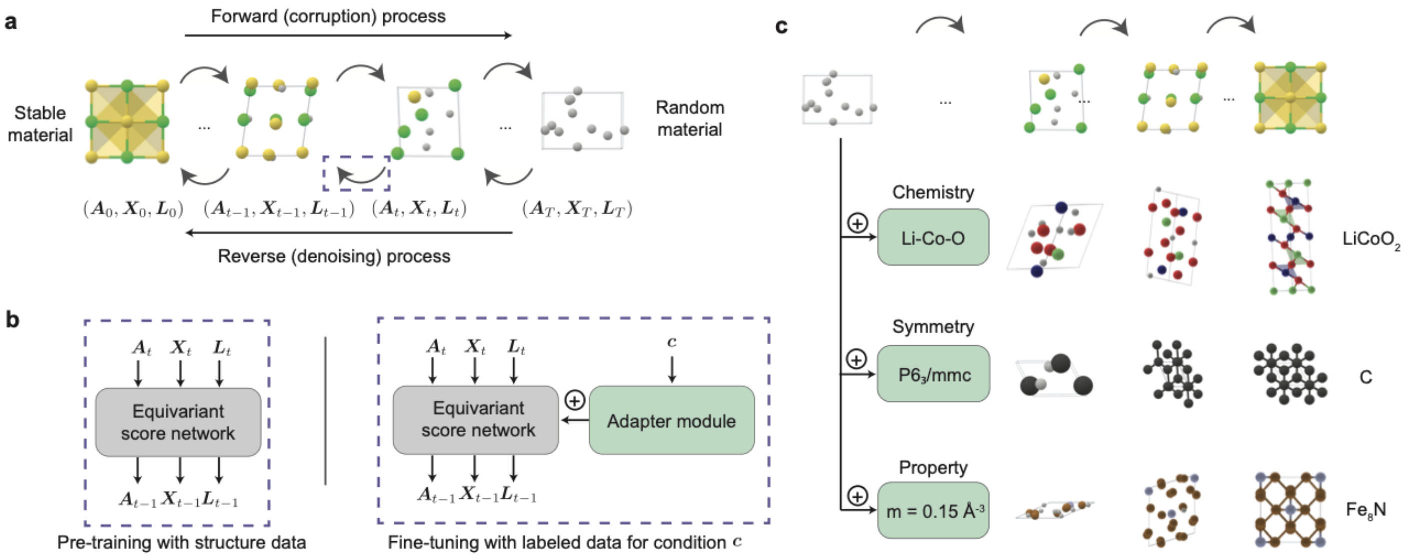

a
Forward (corruption) process
Stable material
Random material
(A_0, X_0, L_0)
$(A_{t-1}, X_{t-1}, L_{t-1})$
(A_t, X_t, L_t)
(A_T, X_T, L_T)
Reverse (denoising) process
b
A_t X_t L_t
Equivariant score network
A_{t-1} X_{t-1} L_{t-1}
Pre-training with structure data
A_t X_t L_t
c
Equivariant score network
Adapter module
A_{t-1} X_{t-1} L_{t-1}
Fine-tuning with labeled data for condition c
c
Chemistry
Li-Co-O
$LiCoO_2$
Symmetry
$P6_3/mmc$
C
Property
m = 0.15 $Å^{-3}$
Fe_8N

Fig. 1: Inorganic materials design with MatterGen. (a) MatterGen generates stable materials by reversing a corruption process through iteratively denoising an initially random structure. The forward diffusion process is designed to independently corrupt atom types $\boldsymbol{A}$, coordinates $\boldsymbol{X}$, and the lattice $\boldsymbol{L}$ to approach a physically motivated distribution of random materials. **(b)** An equivariant score network is pre-trained on a large dataset of stable material structures to jointly denoise atom types, coordinates, and the lattice. The score network is then fine-tuned with a labeled dataset through an adapter module that alters the model using the encoded property $\boldsymbol{c}$. **(c)** MatterGen can be fine-tuned to steer the generation towards materials with desired chemistry, symmetry, and scalar property constraints.

FM for Atomistic Materials Chemistry

2

Trained from all the data
Form the Materials Project

Zeolites
MOFs
Cathodes
Catalysis
Nanoparticles
Amorphous Carbon
Ice and Water
Combustion
Ammonia/borane
Aqueous Interfaces
Batteries
Multicomponent Alloys

MACE-MP0

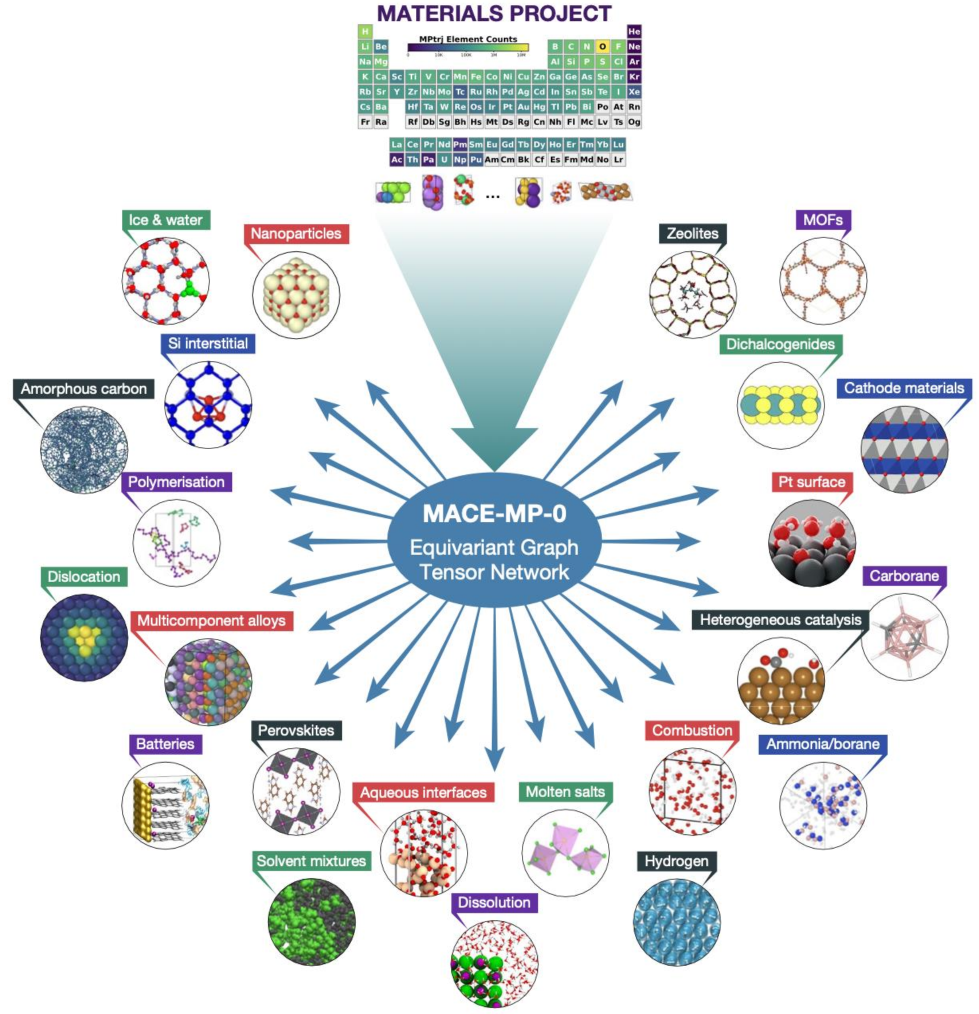
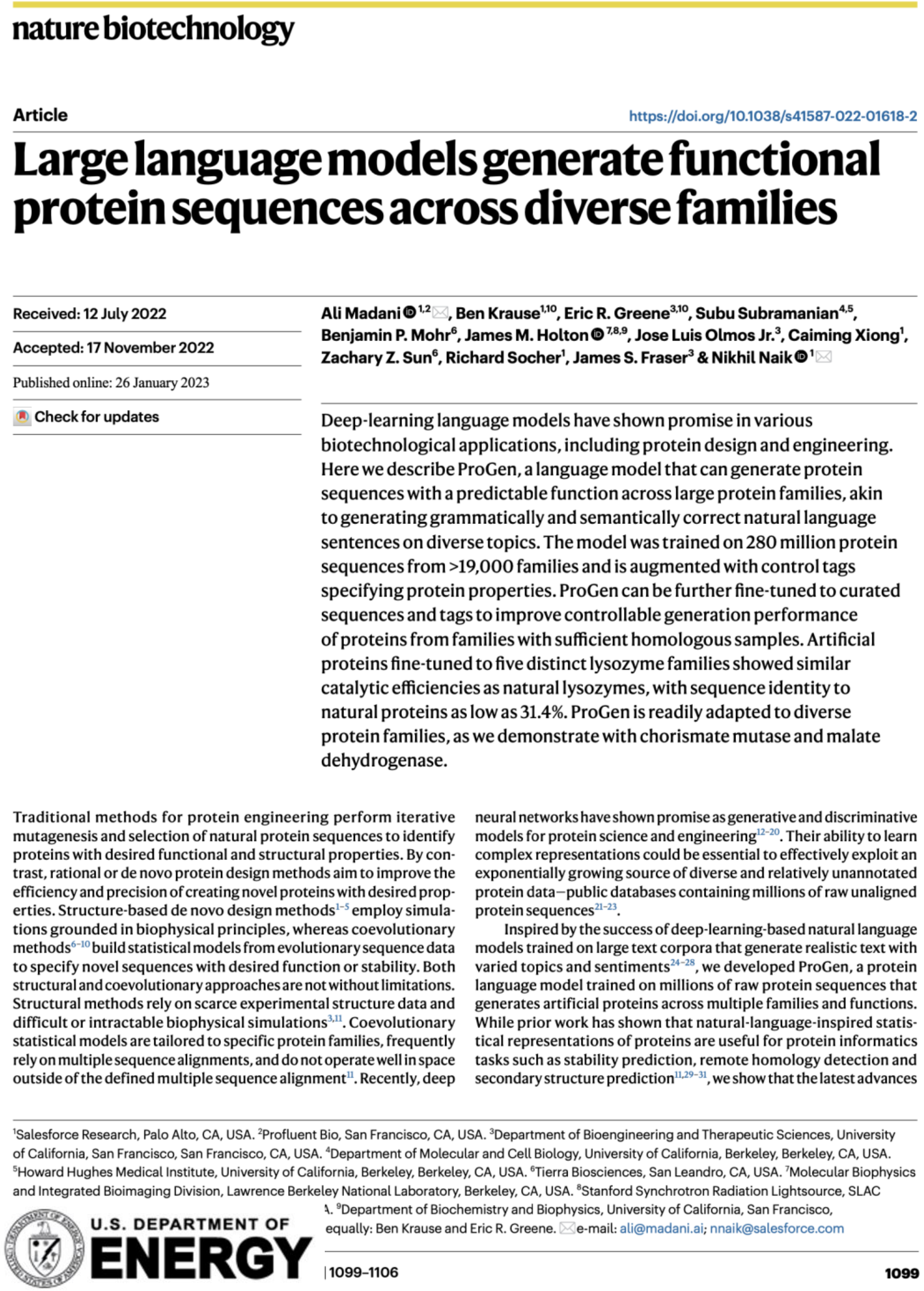


Figure 1: **A foundation model for materials modelling.** Trained only on Materials Project data (*19*) which consists primarily of inorganic crystals and is skewed heavily towards oxides, MACE-MP-0 is capable

arXiv:2401.00096v2 [physics.chem-ph] 1 Mar 2024



nature biotechnology

Article

https://doi.org/10.1038/s41587-022-01618-2

Large language models generate functional protein sequences across diverse families

Received: 12 July 2022

Accepted: 17 November 2022

Published online: 26 January 2023

Check for updates

Ali Madani[1,2]✉, Ben Krause[1,10], Eric R. Greene[3,10], Subu Subramanian[4,5], Benjamin P. Mohr[6], James M. Holton[7,8,9], Jose Luis Olmos Jr.[3], Caiming Xiong[1], Zachary Z. Sun[6], Richard Socher[1], James S. Fraser[3] & Nikhil Naik[1]✉

Deep-learning language models have shown promise in various biotechnological applications, including protein design and engineering. Here we describe ProGen, a language model that can generate protein sequences with a predictable function across large protein families, akin to generating grammatically and semantically correct natural language sentences on diverse topics. The model was trained on 280 million protein sequences from >19,000 families and is augmented with control tags specifying protein properties. ProGen can be further fine-tuned to curated sequences and tags to improve controllable generation performance of proteins from families with sufficient homologous samples. Artificial proteins fine-tuned to five distinct lysozyme families showed similar catalytic efficiencies as natural lysozymes, with sequence identity to natural proteins as low as 31.4%. ProGen is readily adapted to diverse protein families, as we demonstrate with chorismate mutase and malate dehydrogenase.

Traditional methods for protein engineering perform iterative mutagenesis and selection of natural protein sequences to identify proteins with desired functional and structural properties. By contrast, rational or de novo protein design methods aim to improve the efficiency and precision of creating novel proteins with desired properties. Structure-based de novo design methods[1–5] employ simulations grounded in biophysical principles, whereas coevolutionary methods[6–10] build statistical models from evolutionary sequence data to specify novel sequences with desired function or stability. Both structural and coevolutionary approaches are not without limitations. Structural methods rely on scarce experimental structure data and difficult or intractable biophysical simulations[3,11]. Coevolutionary statistical models are tailored to specific protein families, frequently rely on multiple sequence alignments, and do not operate well in space outside of the defined multiple sequence alignment[11]. Recently, deep neural networks have shown promise as generative and discriminative models for protein science and engineering[12–20]. Their ability to learn complex representations could be essential to effectively exploit an exponentially growing source of diverse and relatively unannotated protein data—public databases containing millions of raw unaligned protein sequences[21–23].

Inspired by the success of deep-learning-based natural language models trained on large text corpora that generate realistic text with varied topics and sentiments[24–28], we developed ProGen, a protein language model trained on millions of raw protein sequences that generates artificial proteins across multiple families and functions. While prior work has shown that natural-language-inspired statistical representations of proteins are useful for protein informatics tasks such as stability prediction, remote homology detection and secondary structure prediction[11,29–31], we show that the latest advances

[1]Salesforce Research, Palo Alto, CA, USA. [2]Profluent Bio, San Francisco, CA, USA. [3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA. [4]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [5]Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, CA, USA. [6]Tierra Biosciences, San Leandro, CA, USA. [7]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [8]Stanford Synchrotron Radiation Lightsource, SLAC A. [9]Department of Biochemistry and Biophysics, University of California, San Francisco, equally: Ben Krause and Eric R. Greene. ✉e-mail: ali@madani.ai; nnaik@salesforce.com

U.S. DEPARTMENT OF ENERGY

3

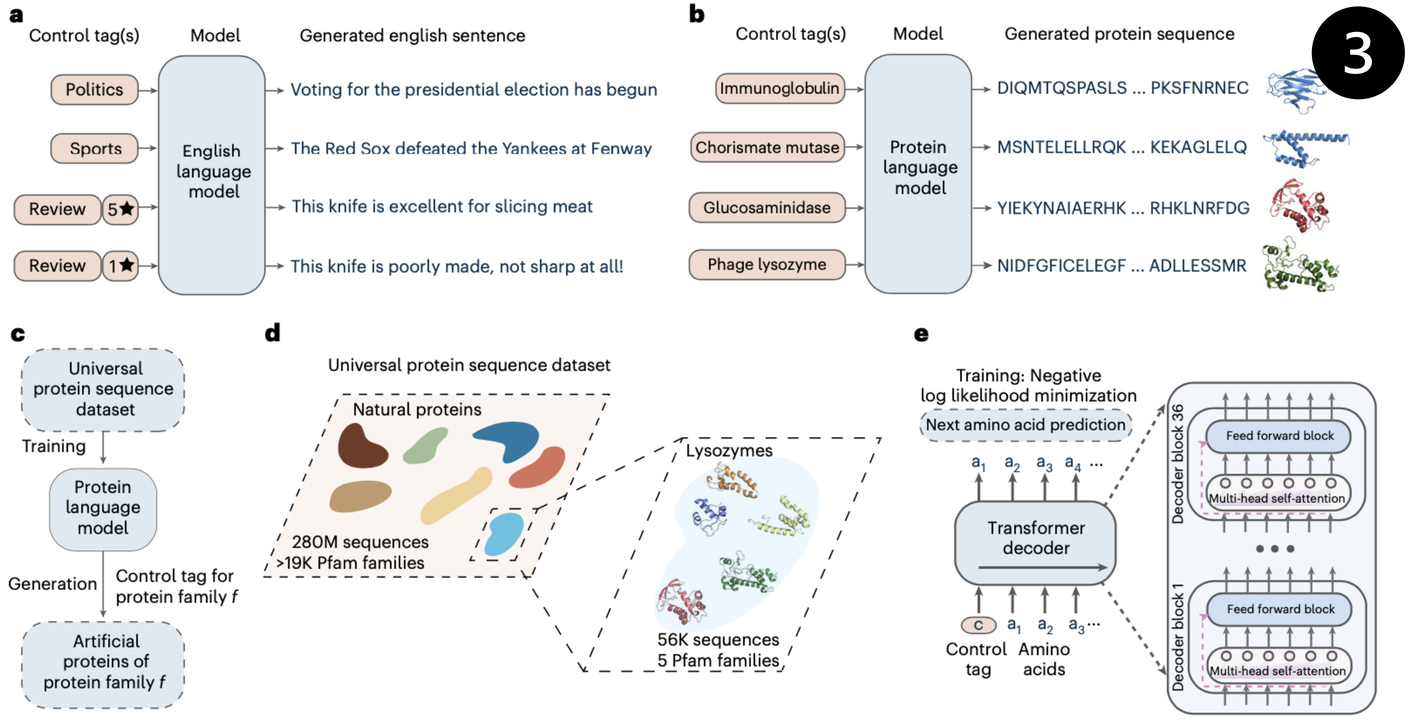
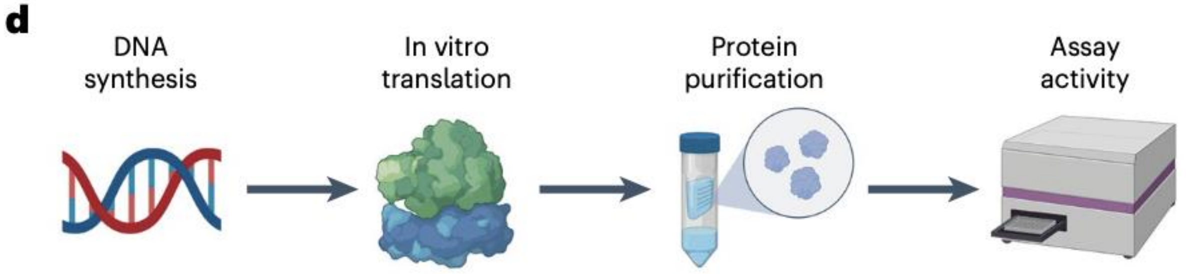
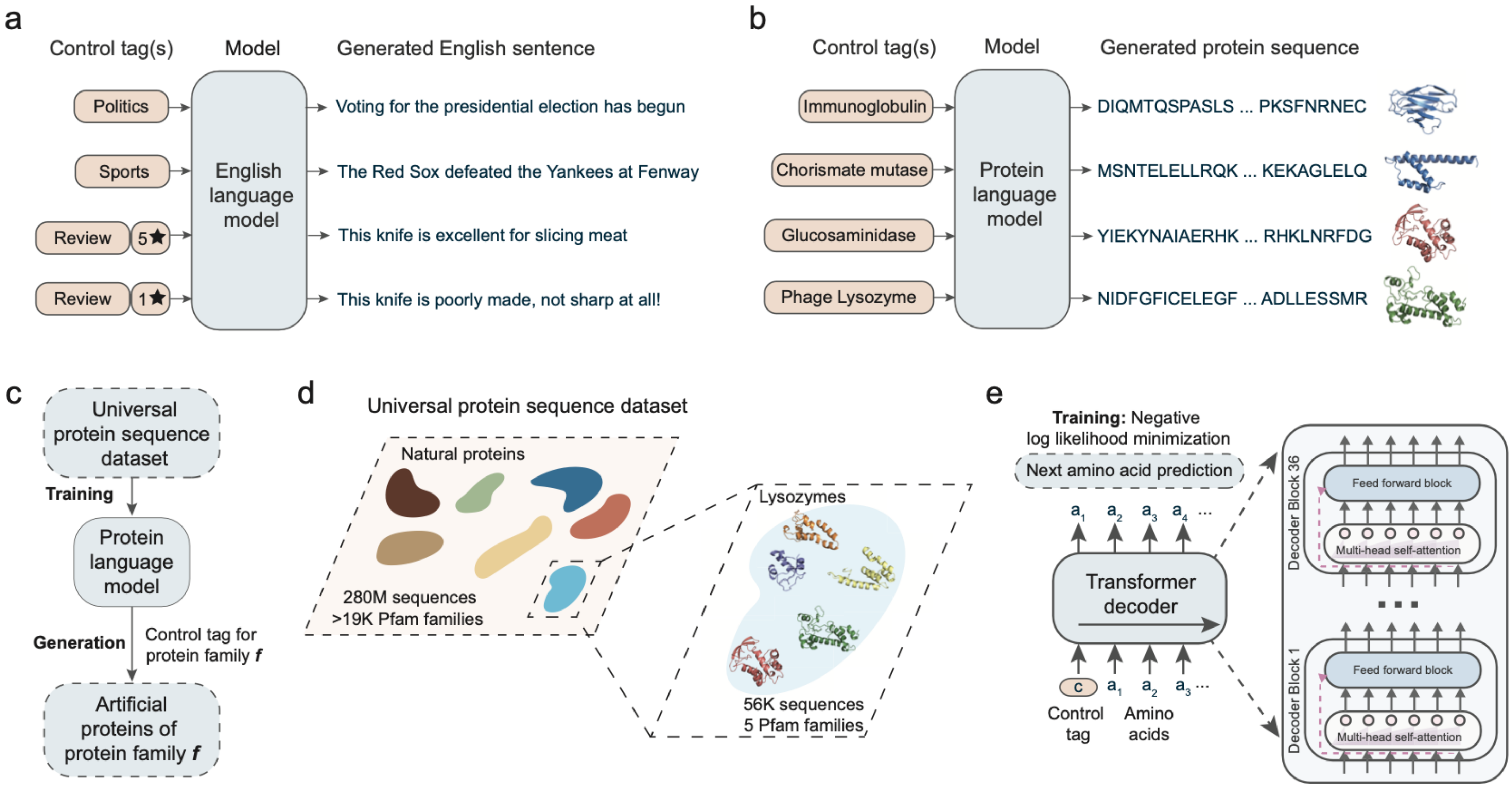

a
Control tag(s)
Model
Generated english sentence
Politics
Sports
Review 5★
Review 1★
English language model
Voting for the presidential election has begun
The Red Sox defeated the Yankees at Fenway
This knife is excellent for slicing meat
This knife is poorly made, not sharp at all!
b
Control tag(s)
Model
Generated protein sequence
Immunoglobulin
Chorismate mutase
Glucosaminidase
Phage lysozyme
Protein language model
DIQMTQSPASLS ... PKSFNRNEC
MSNTELELLRQK ... KEKAGLELQ
YIEKYNAIAERHK ... RHKLNRFDG
NIDFGFICELEGF ... ADLLESSMR
c
Universal protein sequence dataset
Training
Protein language model
Generation
Control tag for protein family f
Artificial proteins of protein family f
d
Universal protein sequence dataset
Natural proteins
280M sequences
>19K Pfam families
Lysozymes
56K sequences
5 Pfam families
e
Training: Negative log likelihood minimization
Next amino acid prediction
a1 a2 a3 a4 ...
Transformer decoder
C a1 a2 a3 ...
Control tag
Amino acids
Decoder block 36
Decoder block 1
Feed forward block
Multi-head self-attention

Fig. 1 | Artificial protein generation with conditional language modeling. **a**, Conditional language models are deep neural networks that can generate semantically and grammatically correct, yet novel and diverse natural language text, steerable using input control tags that govern style, topic and other entities. **b**,**c**, Analogous to natural language models, we develop ProGen, a conditional protein language model (**b**) that generates diverse artificial protein sequences across protein families based on input control tags (**c**). **d**, ProGen is trained using a large, universal protein sequence dataset of 280 million naturally evolved proteins from thousands of families, of which five diverse lysozyme families are experimentally characterized in this study. **e**, ProGen is a 1.2-billion-parameter neural network that is based on the Transformer architecture, which uses a self-attention mechanism for modeling comprehensive residue–residue interactions. ProGen is trained to generate artificial sequences by minimizing the loss over the next amino acid prediction problem on the universal protein sequence dataset.

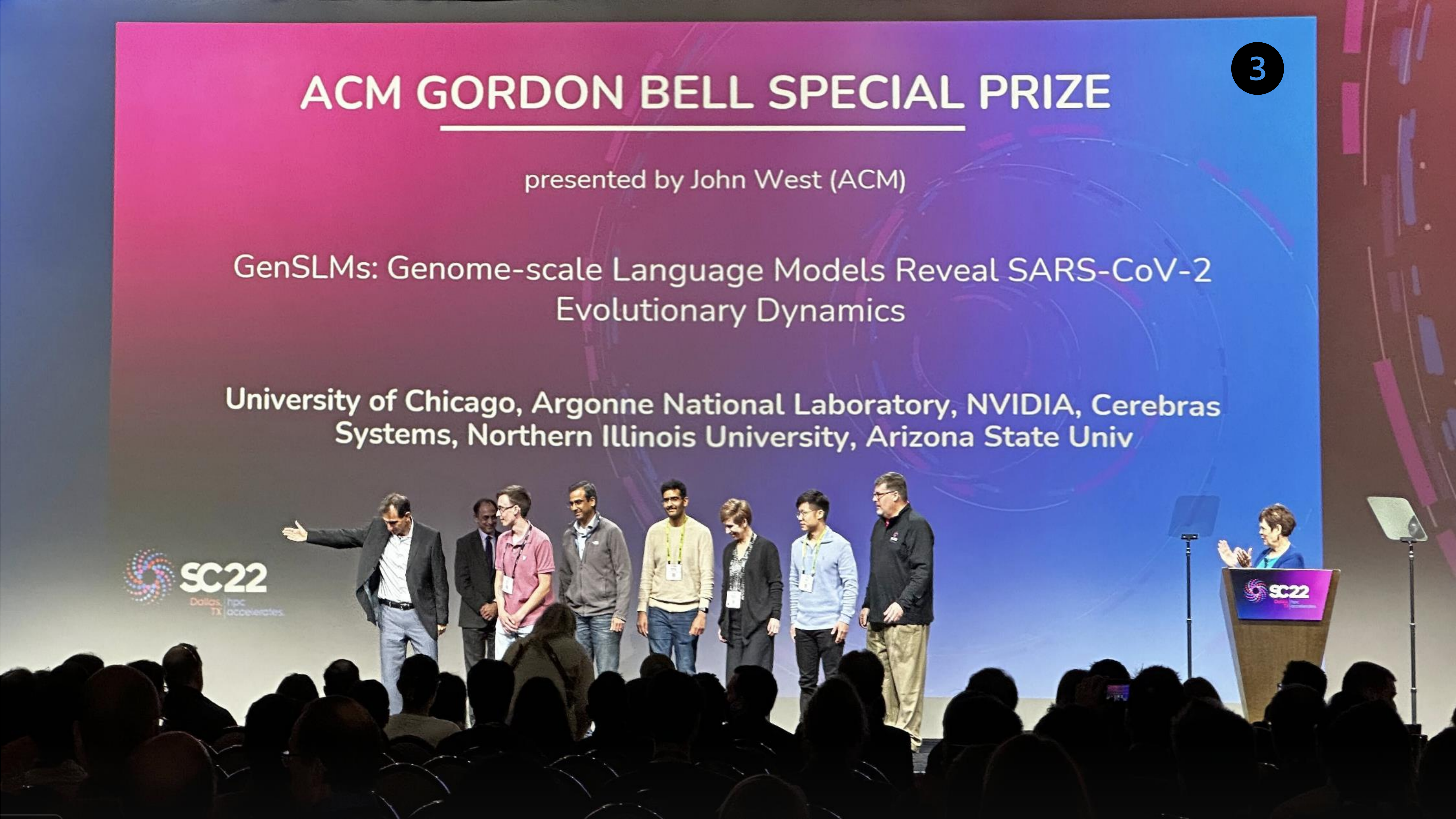


Transformers can Design Proteins


a
Control tag(s)
Model
Generated English sentence
Politics
Sports
Review 5★
Review 1★
English language model
Voting for the presidential election has begun
The Red Sox defeated the Yankees at Fenway
This knife is excellent for slicing meat
This knife is poorly made, not sharp at all!
b
Control tag(s)
Model
Generated protein sequence
Immunoglobulin
Chorismate mutase
Glucosaminidase
Phage Lysozyme
Protein language model
DIQMTQSPASLS ... PKSFNRNEC
MSNTELELLRQK ... KEKAGLELQ
YIEKYNAIAERHK ... RHKLNRFDG
NIDFGFICELEGF ... ADLLESSMR
c
Universal protein sequence dataset
Training
Protein language model
Generation
Control tag for protein family *f*
Artificial proteins of protein family *f*
d
Universal protein sequence dataset
Natural proteins
280M sequences
>19K Pfam families
Lysozymes
56K sequences
5 Pfam families
e
Training: Negative log likelihood minimization
Next amino acid prediction
a_1 a_2 a_3 a_4 ...
Transformer decoder
c a_1 a_2 a_3 ...
Control tag
Amino acids
Decoder Block 36
Feed forward block
Multi-head self-attention
Decoder Block 1
Feed forward block
Multi-head self-attention

https://doi.org/10.1101/2021.07.18.452833;

3

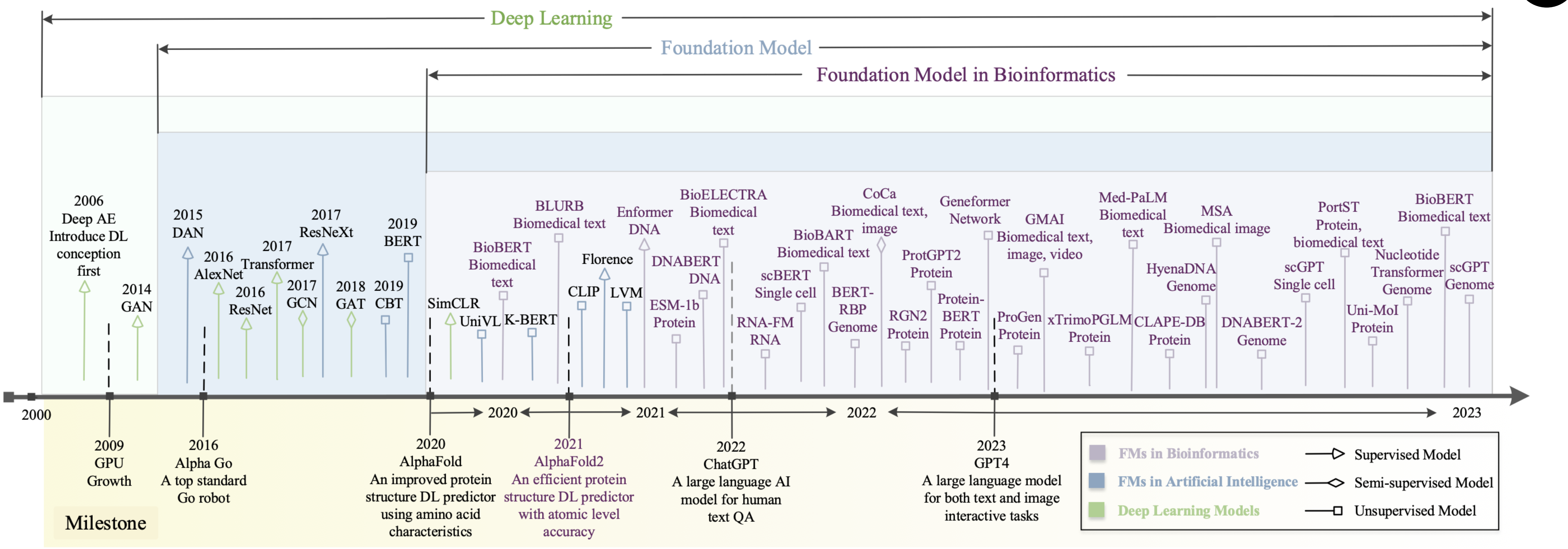


ACM GORDON BELL SPECIAL PRIZE

presented by John West (ACM)

GenSLMs: Genome-scale Language Models Reveal SARS-CoV-2 Evolutionary Dynamics

University of Chicago, Argonne National Laboratory, NVIDIA, Cerebras Systems, Northern Illinois University, Arizona State Univ

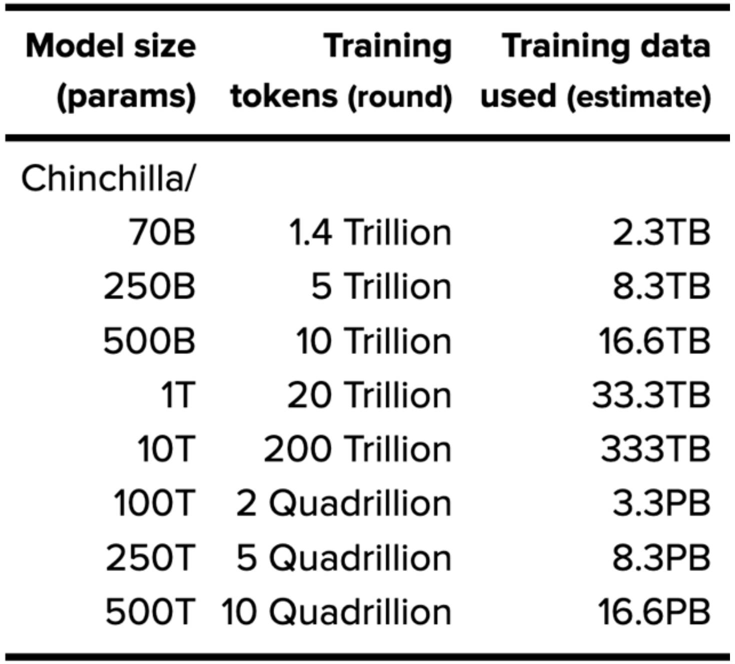


Problems / Data	Domain exploration	Sequence analysis	Structure construction	Function prediction	Multimodal problems
DNA		✓		✓	
RNA			✓	✓	
Protein		✓	✓	✓	✓
scGenomics		✓	✓	✓	✓
KGs/Net.		✓		✓	
Text/Image	✓			✓	✓

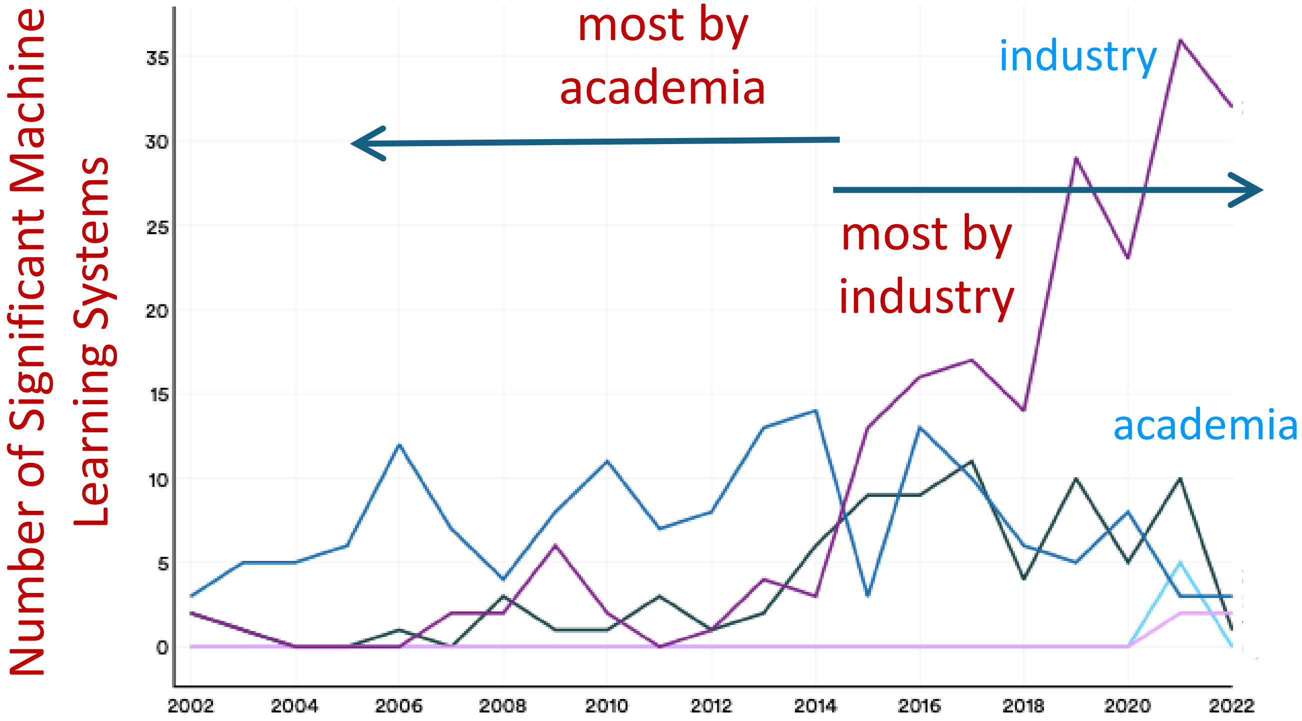
Frontier AI Systems are Large-Scale Science

AI4S is emerging as “big science” in the tradition of nuclear and high energy physics

1T parameters ~ 100 days



Model size (params)	Training tokens (round)	Training data used (estimate)
Chinchilla/ 70B	1.4 Trillion	2.3TB
250B	5 Trillion	8.3TB
500B	10 Trillion	16.6TB
1T	20 Trillion	33.3TB
10T	200 Trillion	333TB
100T	2 Quadrillion	3.3PB
250T	5 Quadrillion	8.3PB
500T	10 Quadrillion	16.6PB



The scale of needed human and computational resources is beginning to reshape leadership in science

Llama-3 Development Team (327 authors)

https://huggingface.co/meta-llama/Meta-Llama-3-8B

Downloads last month

1,051,301

Aaditya Singh; Aaron Grattafiori; Abhimanyu Dubey; Abhinav Jauhri; Abhinav Pandey; Abhishek Kadian; Adam Kelsey; Adi Gangidi; Ahmad Al-Dahle; Ahuva Goldstand; Aiesha Ajay Menon; Akhil Mathur; Alan Schelten; Alex Vaughan; Amy Yang; Andrei Lupu; Andres Alvarado; Andrew Gallagher; Andrew Gu; Andrew Ho; Andrew Poulton; Andrew Ryan; Angela Fan; Ankit Ramchandani; Anthony Hartshorn; Archi Mitra; Archie Sravankumar; Artem Korenev; Arun Rao; Ashley Gabriel; Ashwin Bharambe; Assaf Eisenman; Aston Zhang; Aurelien Rodriguez; Austen Gregerson; Ava Spataru; Baptiste Roziere; Ben Maurer; Benjamin Leonhardi; Bernie Huang; Bhargavi Paranjape; Bing Liu; Binh Tang; Bobbie Chern; Stojkovic; Brian Fuller; Catalina Mejia Arenas; Chao Zhou; Charlotte Caucheteux; Chaya Nayak; Ching-Hsiang Chu; Chloe Bi; Chris Cai; Chris Cox; Chris Marra; Chris McConnell; Keller; Christoph Feichtenhofer; Christophe Touret; Chunyang Wu; Corinne Wong; Cristian Canton Ferrer; Damien Allonsius; Daniel Kreymer; Daniel Haziza; Daniel Li; Danielle Danny Livshits; Danny Wyatt; David Adkins; David Esiobu; David Xu; Davide Testuggine; Delia David; Devi Parikh; Dhruv Choudhary; Dhruv Mahajan; Diana Liskovich; Diego Garcia-Olano; Diego Perino; Dieuwke Hupkes; Dingkang Wang; Dustin Holland; Egor Lakomkin; Elina Lobanova; Xiaoqing Ellen Tan; Emily Dinan; Eric Smith; Erik Brinkman; Esteban Filip Radenovic; Firat Ozgenel; Francesco Caggioni; Frank Seide; Frank Zhang; Gabriel Synnaeve; Gabriella Schwarz; Gabrielle Lee; Gada Badeer; Georgia Anderson; Graeme Nail; Gregoire Mialon; Guan Pang; Guillem Cucurell; Hailey Nguyen; Hannah Korevaar; Hannah Wang; Haroun Habeeb; Harrison Rudolph; Henry Aspegren; Hu Xu; Hugo Touvron; Iga Kozlowska; Igor Molybog; Igor Tufanov; Iliyan Zarov; Imanol Arrieta Ibarra; Irina-Elena Veliche; Isabel Kloumann; Ishan Misra; Ivan Evtimov; Jacob Xu; Jade Copet; Jake Weissman; Geffert; Jana Vranes; Japhet Asher; Jason Park; Jay Mahadeokar; Jean-Baptiste Gaya; Jeet Shah; Jelmer van der Linde; Jennifer Chan; Jenny Hong; Jenya Lee; Jeremy Fu; Jeremy Teboul; Jianfeng Chi; Jianyu Huang; Jie Wang; Jiecao Yu; Joanna Bitton; Joe Spisak; Joelle Pineau; Jon Carvill; Jongsoo Park; Joseph Rocca; Joshua Johnstun; Junteng Jia; Vasuden Alwala; Kam Hou U; Kate Plawiak; Kartikeya Upasani; Kaushik Veeraraghavan; Ke Li; Kenneth Heafield; Kevin Stone; Khalid El-Arini; Krithika Iyer; Kshitiz Malik; Kuenley Kunal Bhalla; Kyle Huang; Lakshya Garg; Lauren Rantala-Yeary; Laurens van der Maaten; Lawrence Chen; Leandro Silva; Lee Bell; Lei Zhang; Liang Tan; Louis Martin; Lovish Luca Wehrstedt; Lukas Blecher; Luke de Oliveira; Madeline Muzzi; Madian Khabsa; Manav Avlani; Mannat Singh; Manohar Paluri; Mark Zuckerberg; Marcin Kardas; Martynas Mathew Oldham; Mathieu Rita; Matthew Lennie; Maya Pavlova; Meghan Keneally; Melanie Kambadur; Mihir Patel; Mikayel Samvelyan; Mike Clark; Mike Lewis; Min Si; Mitesh Kumar Singh; Mo Metanat; Mona Hassan; Naman Goyal; Narjes Torabi; Nicolas Usunier; Nikolay Bashlykov; Nikolay Bogoychev; Niladri Chatterji; Ning Dong; Oliver Aobo Yang; Olivier Duchenne; Onur Celebi; Parth Parekh; Patrick Alrassy; Paul Saab; Pavan Balaji; Pedro Rittner; Pengchuan Zhang; Pengwei Li; Petar Vasic; Peter Weng; Polina Zvyagina; Prajjwal Bhargava; Pratik Dubal; Praveen Krishnan; Punit Singh Koura; Qing He; Rachel Rodriguez; Ragavan Srinivasan; Rahul Mitra; Ramon Calderer; Raymond Li; Robert Stojnic; Roberta Raileanu; Robin Battey; Rocky Wang; Rohit Girdhar; Rohit Patel; Romain Sauvestre; Ronnie Polidoro; Roshan Sumbaly; Ross Taylor; Ruan Silva; Rui Hou; Rui Wang; Russ Howes; Ruty Rinott; Saghar Hosseini; Sai Jayesh Bondu; Samyak Datta; Sanjay Singh; Sara Chugh; Sargun Dhillon; Satadru Pan; Sean Bell; Sergey Edunov; Shaoliang Nie; Sharan Narang; Sharath Raparthy; Shaun Lindsay; Sheng Feng; Sheng Shen; Shenghao Lin; Shiva Shankar; Shruti Bhosale; Shun Zhang; Simon Vandenhende; Sinong Wang; Seohyun Sonia Kim; Soumya Batra; Sten Sootla; Steve Kehoe; Suchin Gururangan; Sumit Gupta; Sunny Virk; Sydney Borodinsky; Tamar Glaser; Tamar Herman; Tamara Best; Tara Fowler; Thomas Georgiou; Thomas Scialom; Tianhe Li; Todor Mihaylov; Tong Xiao; Ujjwal Karn; Vedanuj Goswami; Vibhor Gupta; Vignesh Ramanathan; Viktor Kerkez; Vinay Satish Kumar; Vincent Gonguet; Vish Vogeti; Vlad Poenaru; Vlad Tiberiu Mihailescu; Vladan Petrovic; Vladimir Ivanov; Wei Li; Weiwei Chu; Wenhan Xiong; Wenyin Fu; Wes Bouaziz; Whitney Meers; Will Constable; Xavier Martinet; Xiaojian Wu; Xinbo Gao; Xinfeng Xie; Xuchao Jia; Yaelle Goldschlag; Yann LeCun; Yashesh Gaur; Yasmine Babaei; Ye Qi; Yenda Li; Yi Wen; Yiwen Song; Youngjin Nam; Yuchen Hao; Yuchen Zhang; Yun Wang; Yuning Mao; Yuzi He; Zacharie Delpierre Coudert; Zachary DeVito; Zahra Hankir; Zhaoduo Wen; Zheng Yan; Zhengxing Chen; Zhenyu Yang; Zoe Papakipos

Example of Organizational Effort for One FM

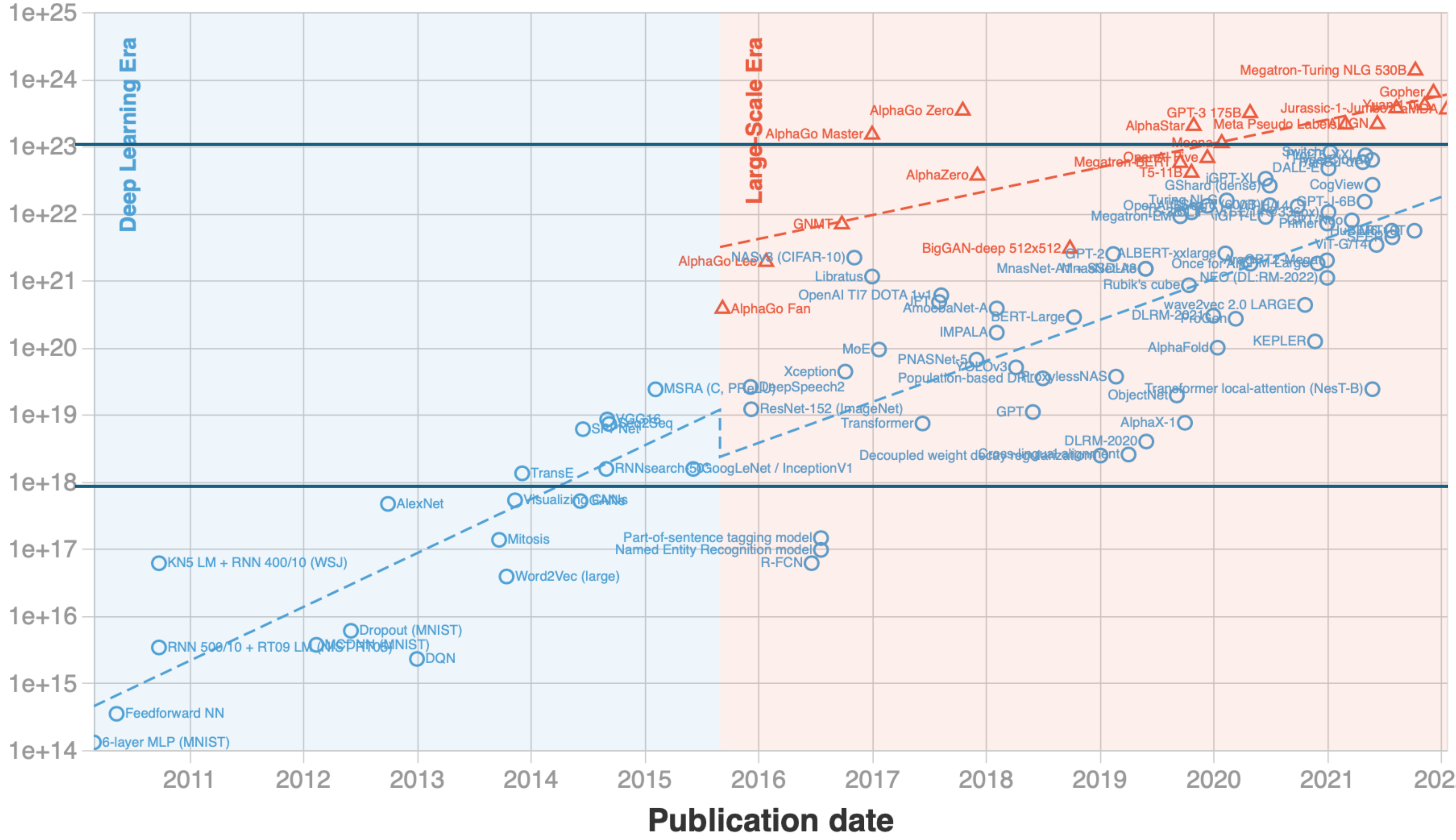
- AuroraGPT working groups:
  - 01 Planning (over the horizon prototyping)
  - 02 Data Organization, Preparation, Representation
  - 03 Model Development and training (pre-training)
  - 04 Evaluation (skills, trustworthiness, safety)
  - 05 Post-training (fine tuning, alignment)
  - 06 Inference and Deployment for Eval and Use
  - 07 Distribution
  - 08 Communication

Training compute (FLOPs) of milestone Machine Learning systems over time
n = 102

https://arxiv.org/abs/2202.05924

EF day

EF second



EF day

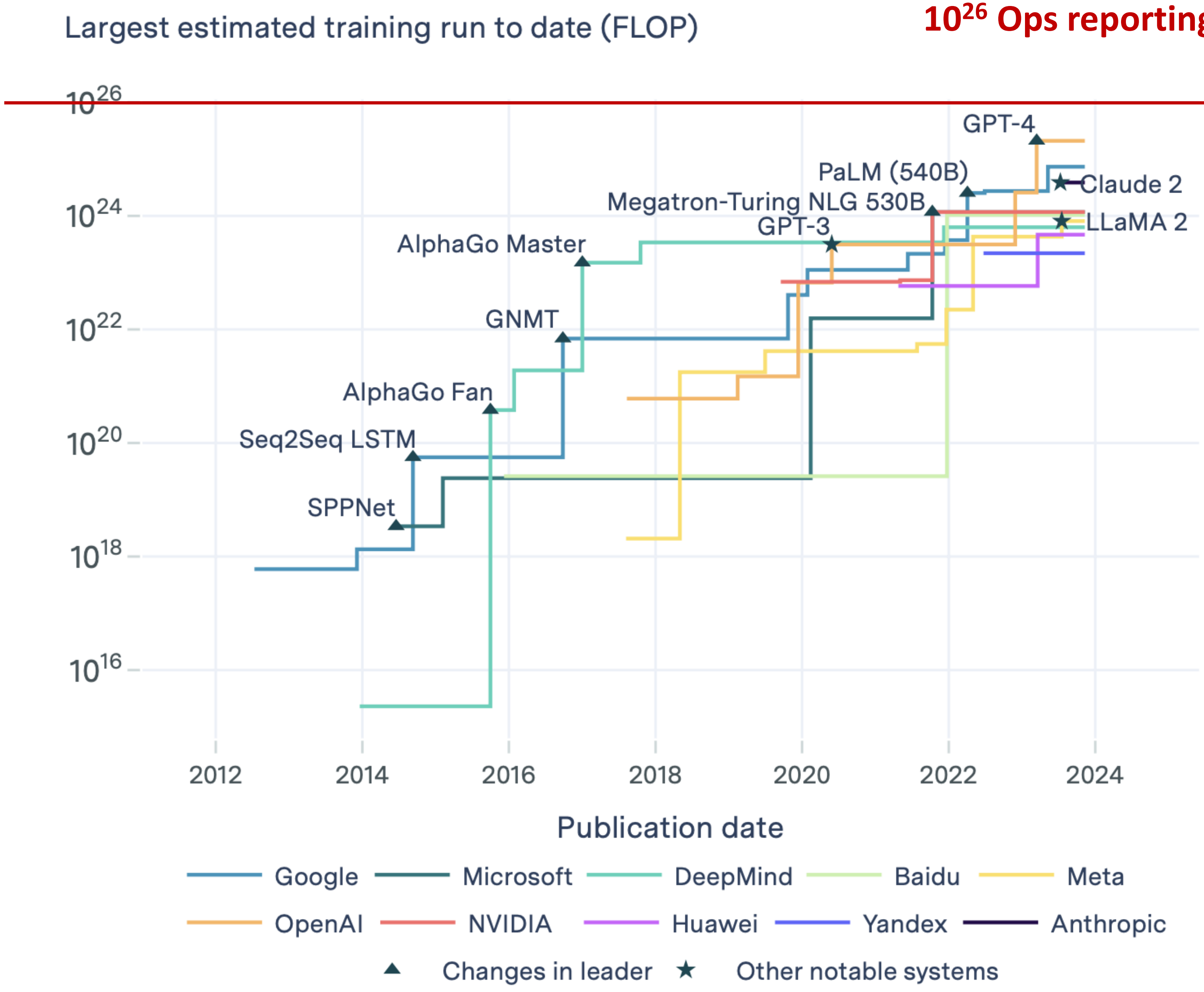
EF second

Figure 3: Trends in training compute of $n102$ milestone ML systems between 2010 and 2022. Notice the emergence of a possible new trend of large-scale models around 2016. The trend in the remaining models stays the same before and after 2016.

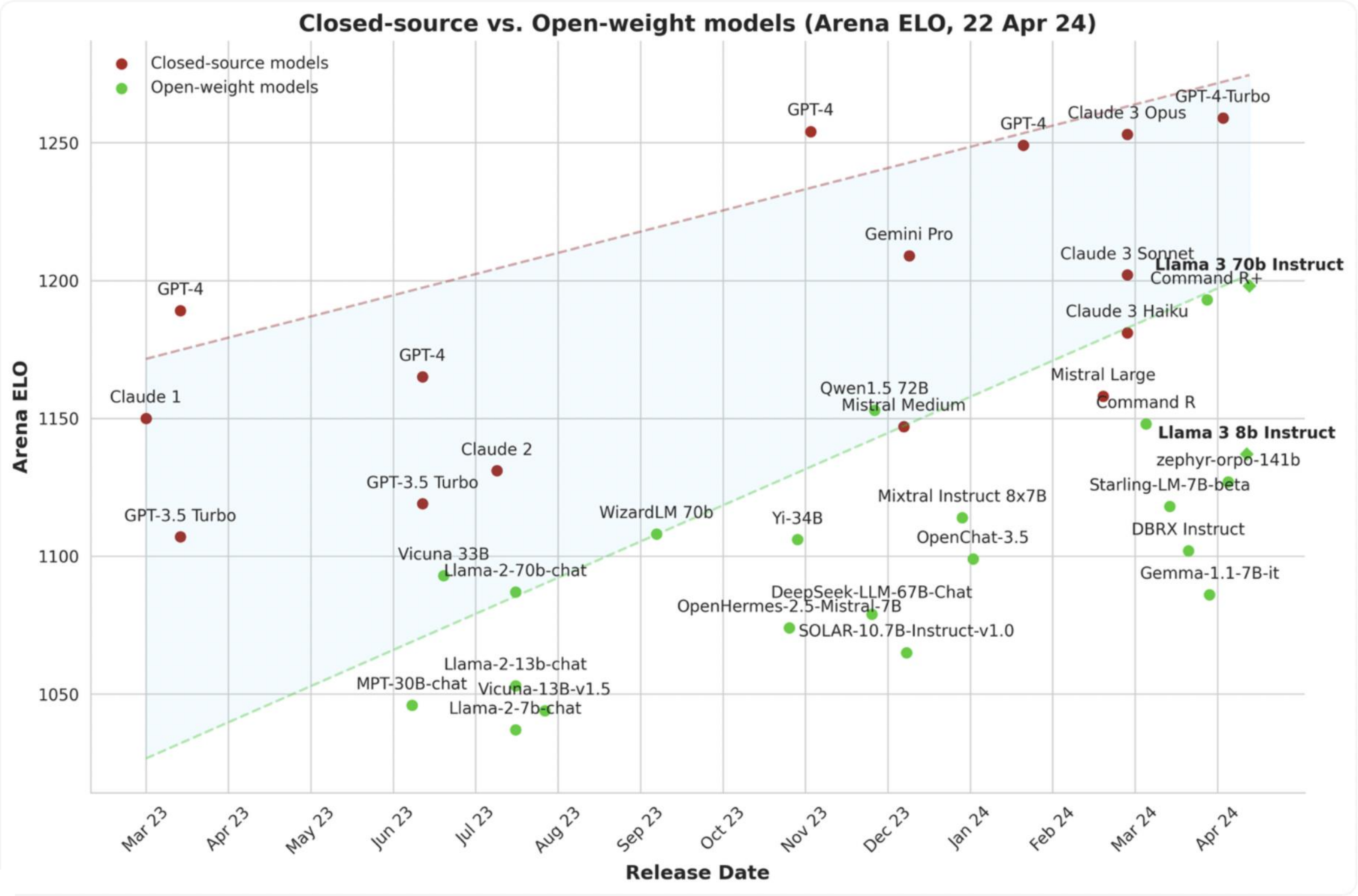
U.S. DEPARTMENT OF ENERGY

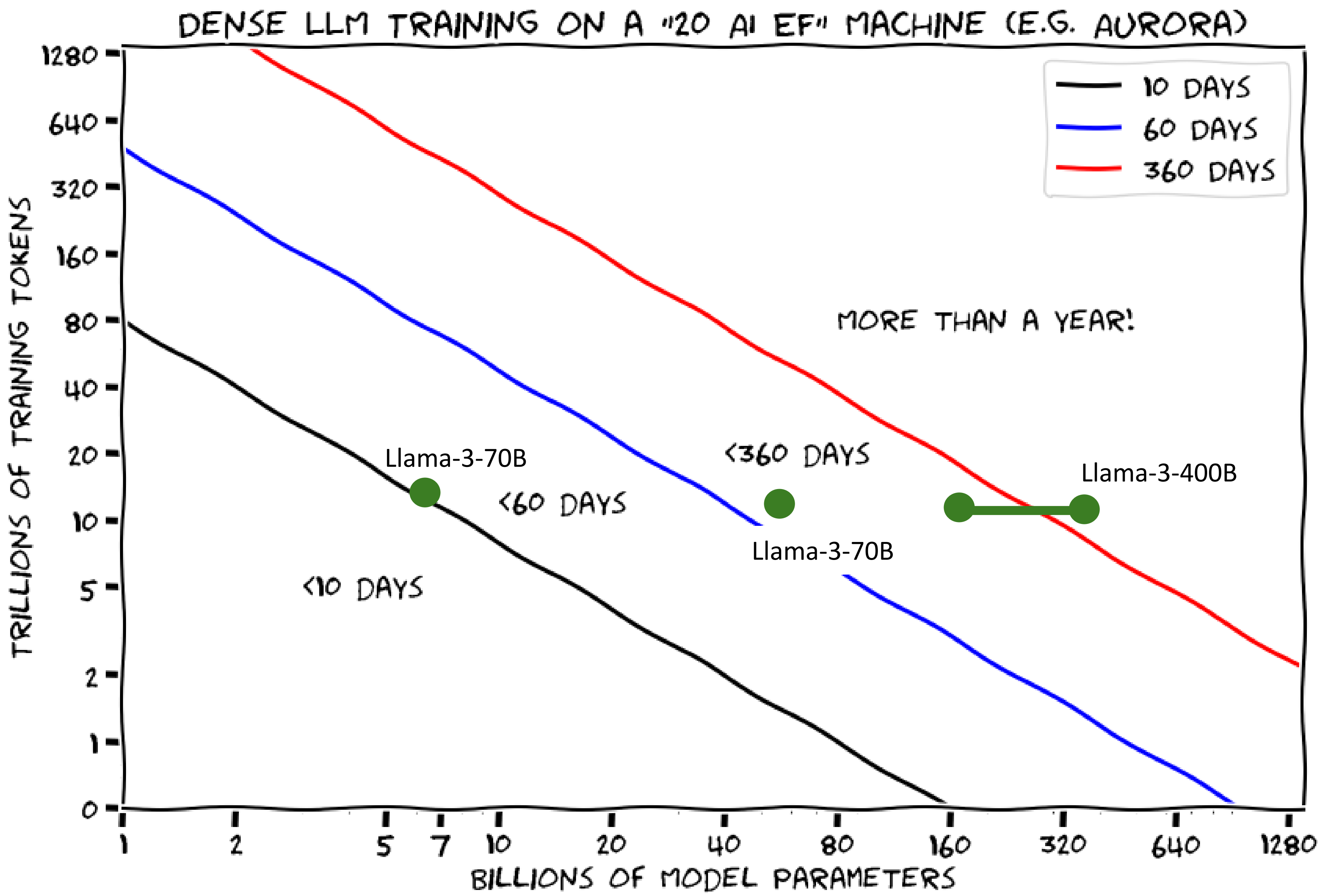
10^{26} **Ops reporting threshold from AI EO**

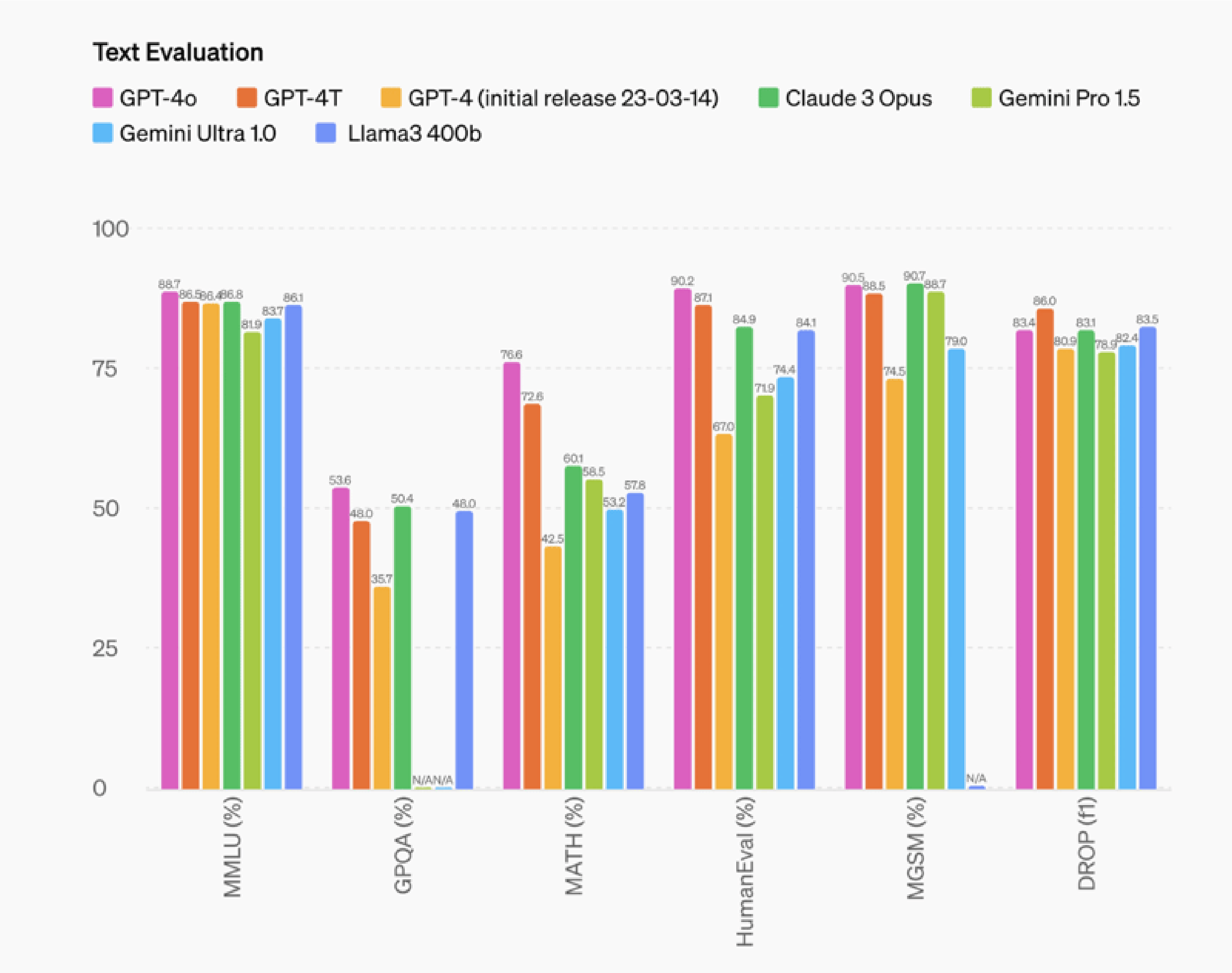
10 EF for 100 days



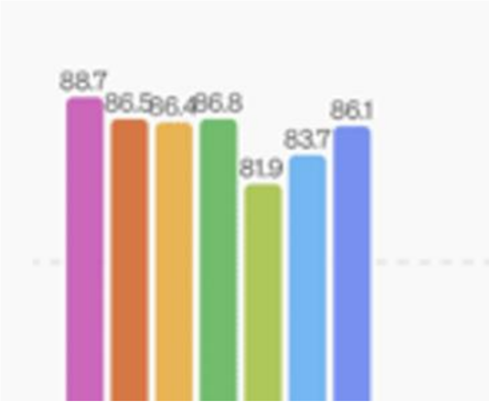
https://epochai.org/blog/who-is-leading-in-ai-an-analysis-of-industry-ai-research

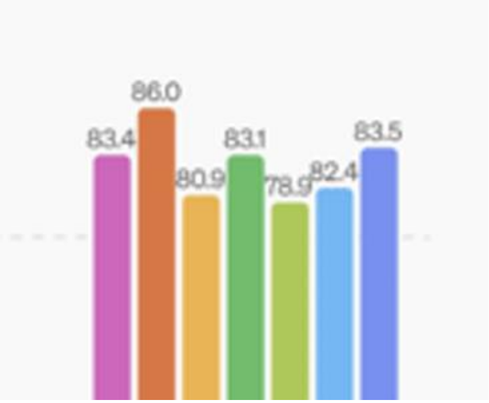

Closed-source vs. Open-weight models (Arena ELO, 22 Apr 24)
● Closed-source models
● Open-weight models
Arena ELO
1250
1200
1150
1100
1050
Release Date
Mar 23
Apr 23
May 23
Jun 23
Jul 23
Aug 23
Sep 23
Oct 23
Nov 23
Dec 23
Jan 24
Feb 24
Mar 24
Apr 24
GPT-4
GPT-4
GPT-4
Claude 3 Opus
GPT-4-Turbo
Gemini Pro
Claude 3 Sonnet
Llama 3 70b Instruct
Command R+
GPT-4
Claude 3 Haiku
GPT-4
Claude 1
Mistral Large
Qwen1.5 72B
Mistral Medium
Command R
Llama 3 8b Instruct
Claude 2
zephyr-orpo-141b
GPT-3.5 Turbo
Starling-LM-7B-beta
GPT-3.5 Turbo
WizardLM 70b
Yi-34B
Mixtral Instruct 8x7B
OpenChat-3.5
DBRX Instruct
Vicuna 33B
Llama-2-70b-chat
Gemma-1.1-7B-it
DeepSeek-LLM-67B-Chat
OpenHermes-2.5-Mistral-7B
SOLAR-10.7B-Instruct-v1.0
Llama-2-13b-chat
MPT-30B-chat
Vicuna-13B-v1.5
Llama-2-7b-chat


DENSE LLM TRAINING ON A "20 AI EF" MACHINE (E.G. AURORA)
TRILLIONS OF TRAINING TOKENS
1280
640
320
160
80
40
20
10
5
2
1
0
BILLIONS OF MODEL PARAMETERS
1
2
5
7
10
20
40
80
160
320
640
1280
10 DAYS
60 DAYS
360 DAYS
MORE THAN A YEAR!
<360 DAYS
<60 DAYS
<10 DAYS
Llama-3-70B
Llama-3-70B
Llama-3-400B



Improved Reasoning - GPT-4o sets a new high-score of 88.7% on 0-shot COT MMLU (general knowledge questions). All these evals were gathered with our new simple evals library. In... +





Llama 3 400B is close to GPT-4o

Keep training!!

https://www.reddit.com/r/LocalLLaMA/comments/1cr5yce/openai_gpt4o_eval_results_and_llama3400b/

Frontier AI
for Science Security and Technology

FASST Goals and Outcomes

- Ensure US (DOE) **leads the world in technical capability** for its missions in Science, Energy and National Security
- **Create, deploy and sustain world leading ”frontier” AI systems** and **applications** for DOE mission areas to provide advantage to US and partners
- Increase productivity and capabilities of the **DOE laboratories, academic, agency and international partners**
- Develop **an AI-forward workforce** for DOE

- **Discovery Science** – accelerate and improve effectiveness
- **Energy Transition** – accelerate, reduce risk, improve translation
- **National Security** – anticipate risk, mitigate risks, accelerate mission

Frontier AI for Science, Security and Technology

Data	Data aggregation, curation, representation, interfaces And infrastructure
Compute	Platforms for AI, next generation hardware , path forward for AI, cloud partnerships, Pathways to zettascale systems for AI
Models	Foundation Models for Science, Security and Energy Strategic multimodal FMs
Applications	Adapting Models for DOE Missions (many many targets)

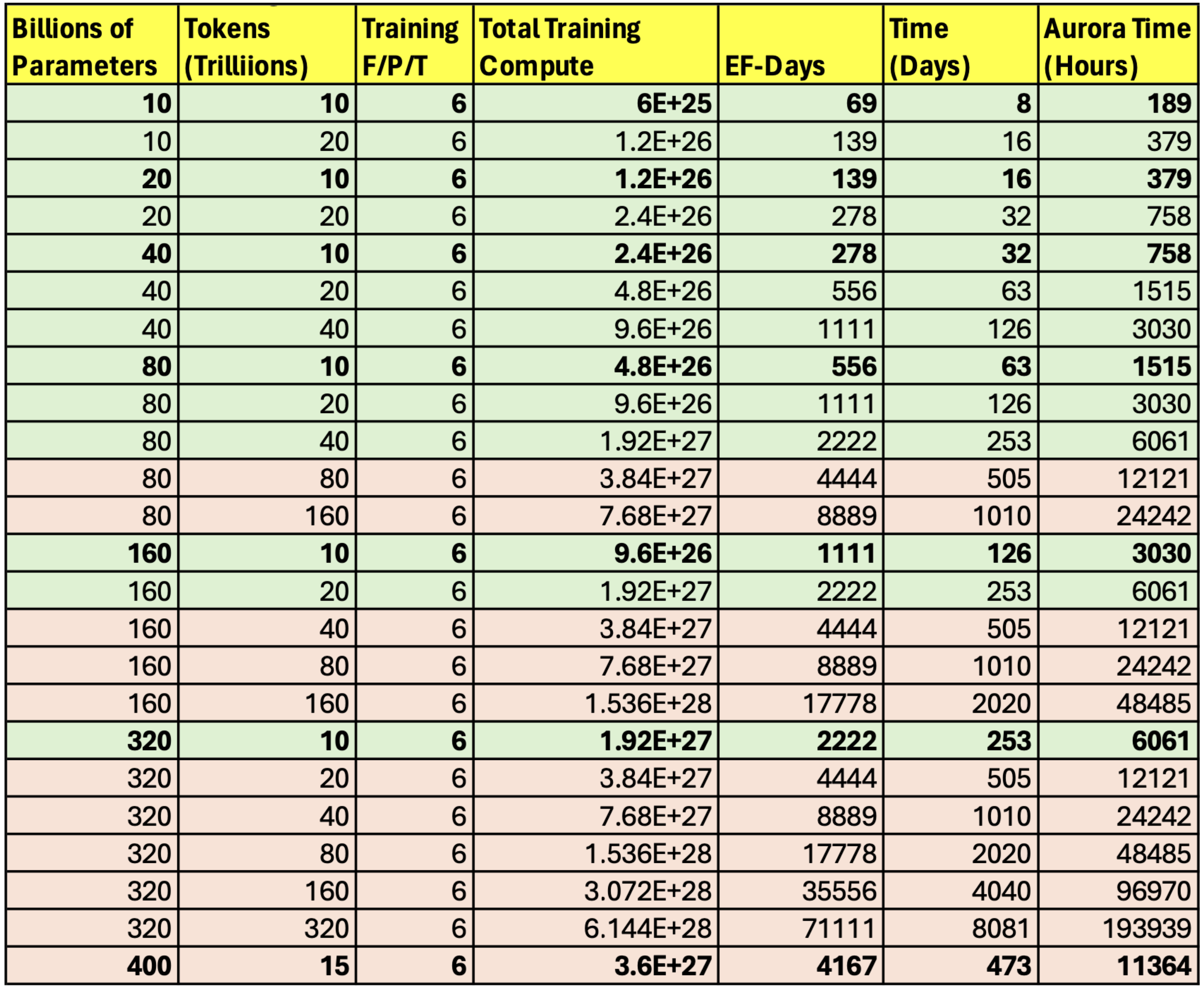
Key points on the FASST program

- **Data** – aggregating, cleaning, curating, transforming the many 100’s of petabytes of scientific data for AI training/testing, **target of 1000 Trillion tokens**
- **Platforms** – investing to create new platforms for training and inference, investments to push 1000EF @100MW, deployment of **multiple AI frontier training platforms** and **many AI inference optimized systems** – distinct from general HPC
- **Models** – ramping up to train order dozen domain oriented FMs each year to cover science, energy and national security
- **Applications** – downstream adaptation of models for 100’s-1000’s of DOE use cases

FASST Targets

- **Data effort must produce tokens on schedule** or whole effort will be rate limited on data preparation
  - Labor and inference intensive ⟹ **100 T tokens in first few years**
  - Common data APIs are needed, but not waste time on unneeded sw/standards
- **To train ~10 Frontier FMs per year will require building out of significant AI training resources to avoid cannibalizing LCFs**
  - Need 10x current Exascale AI flops in next few years ⟹ **200 AI EFs (a few sites)**
- **Inference hardware capacity is critical**
  - Need to serve models/apps for development and production
  - Will need thousands of inference servers ⟹ **200 AI EFs (deployed at ~10 sites)**
- **Large increase in staff are needed across the FASST program**
  - Much of the work in building and deploying FMs and applications is “engineering” and a project framework is needed to both manage to schedule and to integrate the hundreds of activities ⟹ **2000 FTEs**
- **Applications need to be deployed to get productivity boost**
  - Applications development should start now with open models and swap FMs as better ones become available ⟹ **100 frontier AI based applications**
  - Modular architecture with plug-in APIs are needed to avoid silos

How many models of what Scale can we train on our Current Exascale systems?



Billions of Parameters	Tokens (Trilliions)	Training F/P/T	Total Training Compute	EF-Days	Time (Days)	Aurora Time (Hours)
10	**10**	**6**	**6E+25**	**69**	**8**	**189**
10	20	6	1.2E+26	139	16	379
20	**10**	**6**	**1.2E+26**	**139**	**16**	**379**
20	20	6	2.4E+26	278	32	758
40	**10**	**6**	**2.4E+26**	**278**	**32**	**758**
40	20	6	4.8E+26	556	63	1515
40	40	6	9.6E+26	1111	126	3030
80	**10**	**6**	**4.8E+26**	**556**	**63**	**1515**
80	20	6	9.6E+26	1111	126	3030
80	40	6	1.92E+27	2222	253	6061
80	80	6	3.84E+27	4444	505	12121
80	160	6	7.68E+27	8889	1010	24242
160	**10**	**6**	**9.6E+26**	**1111**	**126**	**3030**
160	20	6	1.92E+27	2222	253	6061
160	40	6	3.84E+27	4444	505	12121
160	80	6	7.68E+27	8889	1010	24242
160	160	6	1.536E+28	17778	2020	48485
320	**10**	**6**	**1.92E+27**	**2222**	**253**	**6061**
320	20	6	3.84E+27	4444	505	12121
320	40	6	7.68E+27	8889	1010	24242
320	80	6	1.536E+28	17778	2020	48485
320	160	6	3.072E+28	35556	4040	96970
320	320	6	6.144E+28	71111	8081	193939
400	**15**	**6**	**3.6E+27**	**4167**	**473**	**11364**

GPT4 was trained on 15T tokens

GPT5 is training on X T tokens?

Llama 3 is trained on 15T tokens

Scientific FMs will need to be trained on more and multimodal data

Non LLMs have different scaling

LLM Scaling Models are Quite Good

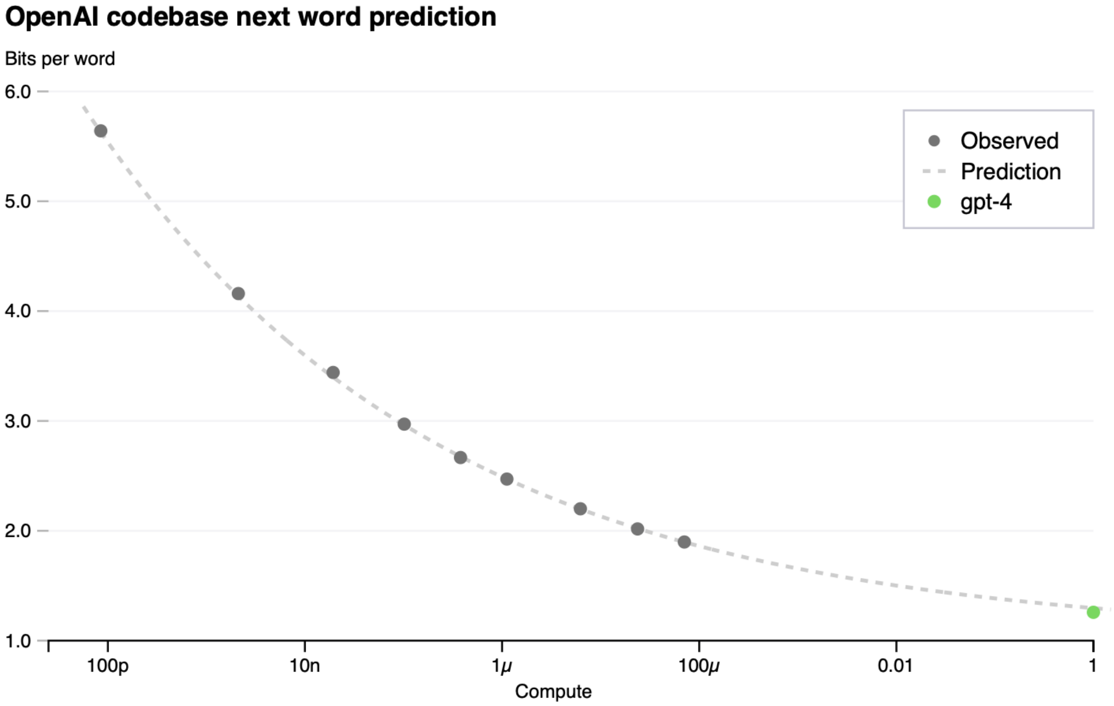
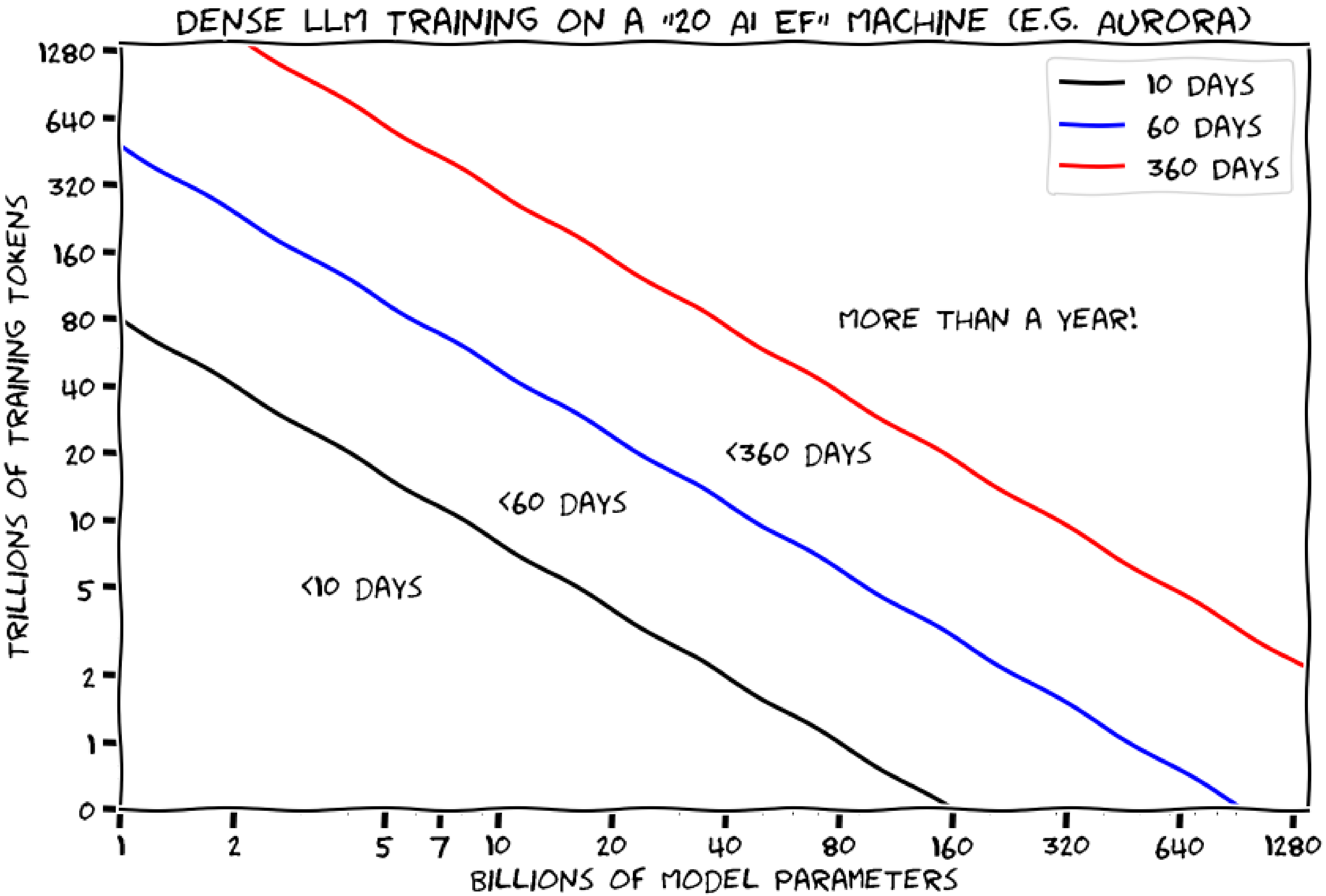
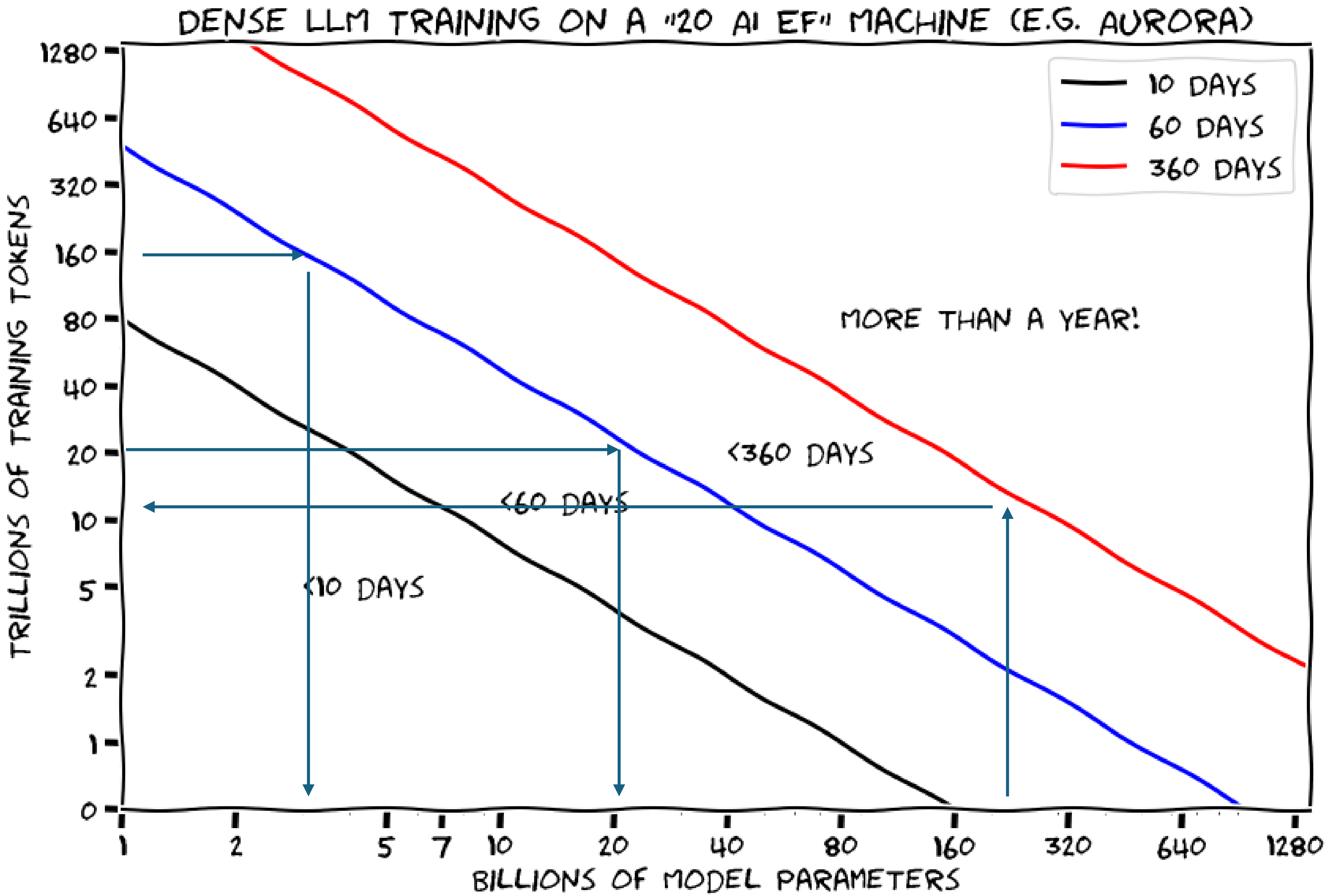


Figure 1. Performance of GPT-4 and smaller models. The metric is final loss on a dataset derived from our internal codebase. This is a convenient, large dataset of code tokens which is not contained in the training set. We chose to look at loss because it tends to be less noisy than other measures across different amounts of training compute. A power law fit to the smaller models (excluding GPT-4) is shown as the dotted line; this fit accurately predicts GPT-4's final loss. The x-axis is training compute normalized so that GPT-4 is 1.

https://arxiv.org/pdf/2303.08774


DENSE LLM TRAINING ON A "20 AI EF" MACHINE (E.G. AURORA)
10 DAYS
60 DAYS
360 DAYS
MORE THAN A YEAR!
<360 DAYS
<60 DAYS
<10 DAYS
TRILLIONS OF TRAINING TOKENS
1280
640
320
160
80
40
20
10
5
2
1
0
1
2
5
7
10
20
40
80
160
320
640
1280
BILLIONS OF MODEL PARAMETERS



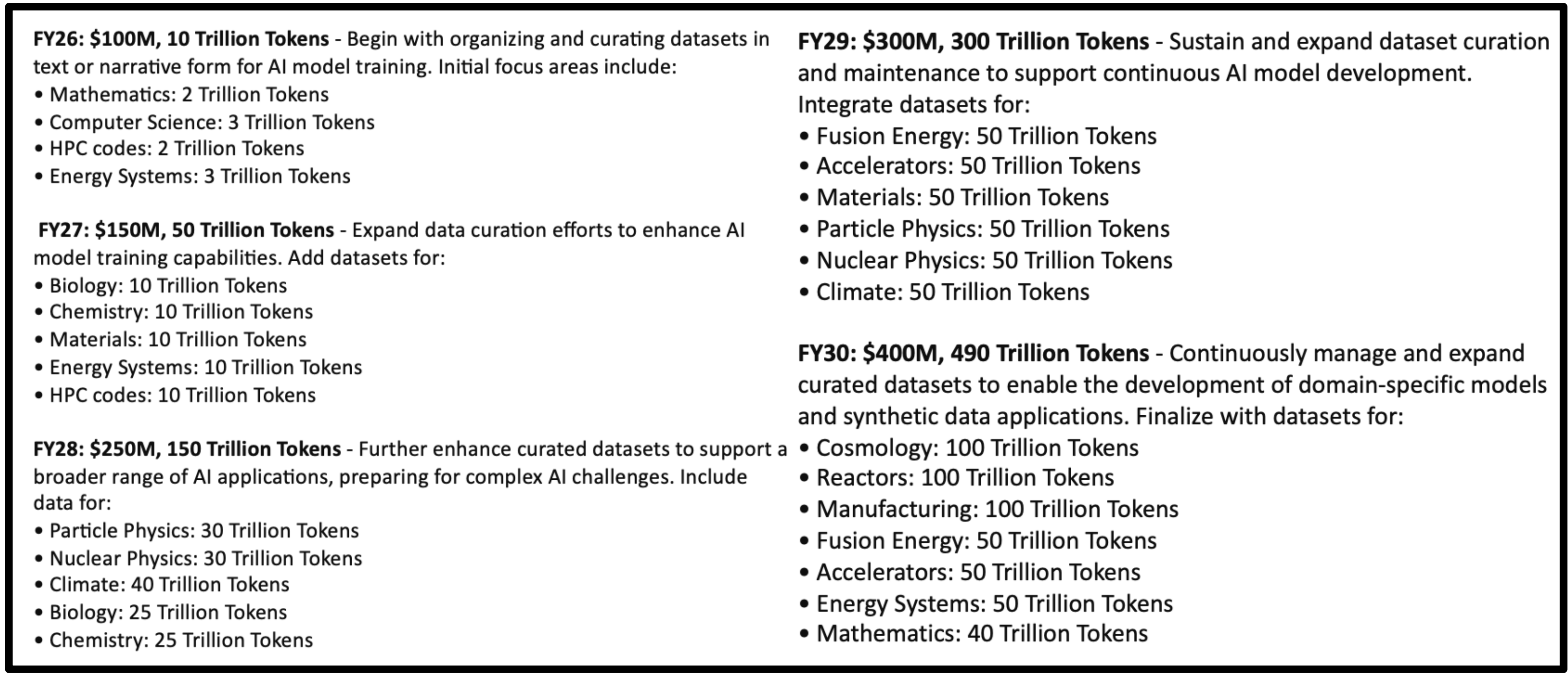
U.S. DEPARTMENT OF ENERGY

Five-year Sketch of Data Preparation for FMs

- Accelerators: 100 Trillion Tokens
- Biology: 35 Trillion Tokens
- Chemistry: 35 Trillion Tokens
- Climate: 90 Trillion Tokens
- Computer Science: 3 Trillion Tokens
- Cosmology: 100 Trillion Tokens
- Energy Systems: 63 Trillion Tokens
- Fusion Energy: 100 Trillion Tokens
- HPC codes: 12 Trillion Tokens
- Manufacturing: 100 Trillion Tokens
- Materials: 60 Trillion Tokens
- Mathematics: 42 Trillion Tokens
- Nuclear Physics: 80 Trillion Tokens
- Particle Physics: 80 Trillion Tokens
- Reactors: 100 Trillion Tokens

1000 Trillion Tokens over 5 years?

GPT4 trained on ~15T tokens
Llama3 trained on ~15T tokens



FY26: $100M, 10 Trillion Tokens - Begin with organizing and curating datasets in text or narrative form for AI model training. Initial focus areas include:

- Mathematics: 2 Trillion Tokens
- Computer Science: 3 Trillion Tokens
- HPC codes: 2 Trillion Tokens
- Energy Systems: 3 Trillion Tokens

FY27: $150M, 50 Trillion Tokens - Expand data curation efforts to enhance AI model training capabilities. Add datasets for:

- Biology: 10 Trillion Tokens
- Chemistry: 10 Trillion Tokens
- Materials: 10 Trillion Tokens
- Energy Systems: 10 Trillion Tokens
- HPC codes: 10 Trillion Tokens

FY28: $250M, 150 Trillion Tokens - Further enhance curated datasets to support a broader range of AI applications, preparing for complex AI challenges. Include data for:

- Particle Physics: 30 Trillion Tokens
- Nuclear Physics: 30 Trillion Tokens
- Climate: 40 Trillion Tokens
- Biology: 25 Trillion Tokens
- Chemistry: 25 Trillion Tokens

FY29: $300M, 300 Trillion Tokens - Sustain and expand dataset curation and maintenance to support continuous AI model development. Integrate datasets for:

- Fusion Energy: 50 Trillion Tokens
- Accelerators: 50 Trillion Tokens
- Materials: 50 Trillion Tokens
- Particle Physics: 50 Trillion Tokens
- Nuclear Physics: 50 Trillion Tokens
- Climate: 50 Trillion Tokens

FY30: $400M, 490 Trillion Tokens - Continuously manage and expand curated datasets to enable the development of domain-specific models and synthetic data applications. Finalize with datasets for:

- Cosmology: 100 Trillion Tokens
- Reactors: 100 Trillion Tokens
- Manufacturing: 100 Trillion Tokens
- Fusion Energy: 50 Trillion Tokens
- Accelerators: 50 Trillion Tokens
- Energy Systems: 50 Trillion Tokens
- Mathematics: 40 Trillion Tokens

Key FY26 Deliverables:

- 6 operational AI hubs
- 3 domain-specific foundation models trained on initial curated datasets
- 10 DOE AI FM applications developed and deployed
- Suite of curated datasets to enable further model development
- “20 AI EF” systems deployed
- Upgraded compute infrastructure to support model training
  - Supporting 1,000 DOE active scientific/engineering users
- Established partnerships to expand AI capabilities and workforce

Key FY27 Deliverables:

- 9 operational AI hubs (initial 6)
  - Deploying 3 FMs from FY26
- 6 domain-specific foundation models trained on initial curated datasets
- 20 DOE AI FM applications developed and deployed
- Suite of curated datasets to enable further model development
- “100 AI EF” systems deployed
- Upgraded compute infrastructure to support model training and inference
  - Supporting 2,000 DOE active scientific/engineering users
- Established partnerships to expand AI capabilities and workforce

Key FY28 Deliverables:

- 12 operational AI hubs
  - Deploying 6 FMs from FY27
- 8 domain-specific foundation models trained on initial curated datasets
- 30 DOE AI FM applications developed and deployed
- Suite of curated datasets to enable further model development
  - Add more science, energy and security topics
- “200 AI EF” systems deployed
- Upgraded compute infrastructure to support model training and inference
  - Supporting 5,000 DOE active scientific/engineering users
- Expanded partnerships to expand AI capabilities and workforce

Key FY29 Deliverables:

- 12 fully operational AI hubs
  - Deploying and supporting 10 world leading FMs from FY28
- 10 updated domain-specific foundation models trained on curated datasets and synthetic data
- 40 DOE AI FM applications developed and deployed
- Suite of curated datasets to enable further model development
  - Partnerships with industry on synthetic data augmentation
- “500 AI EF” systems deployed
- Upgraded compute infrastructure to support model training and inference
  - Supporting 10,000 DOE active scientific/engineering users
- Mature partnerships to sustain AI capabilities and workforce

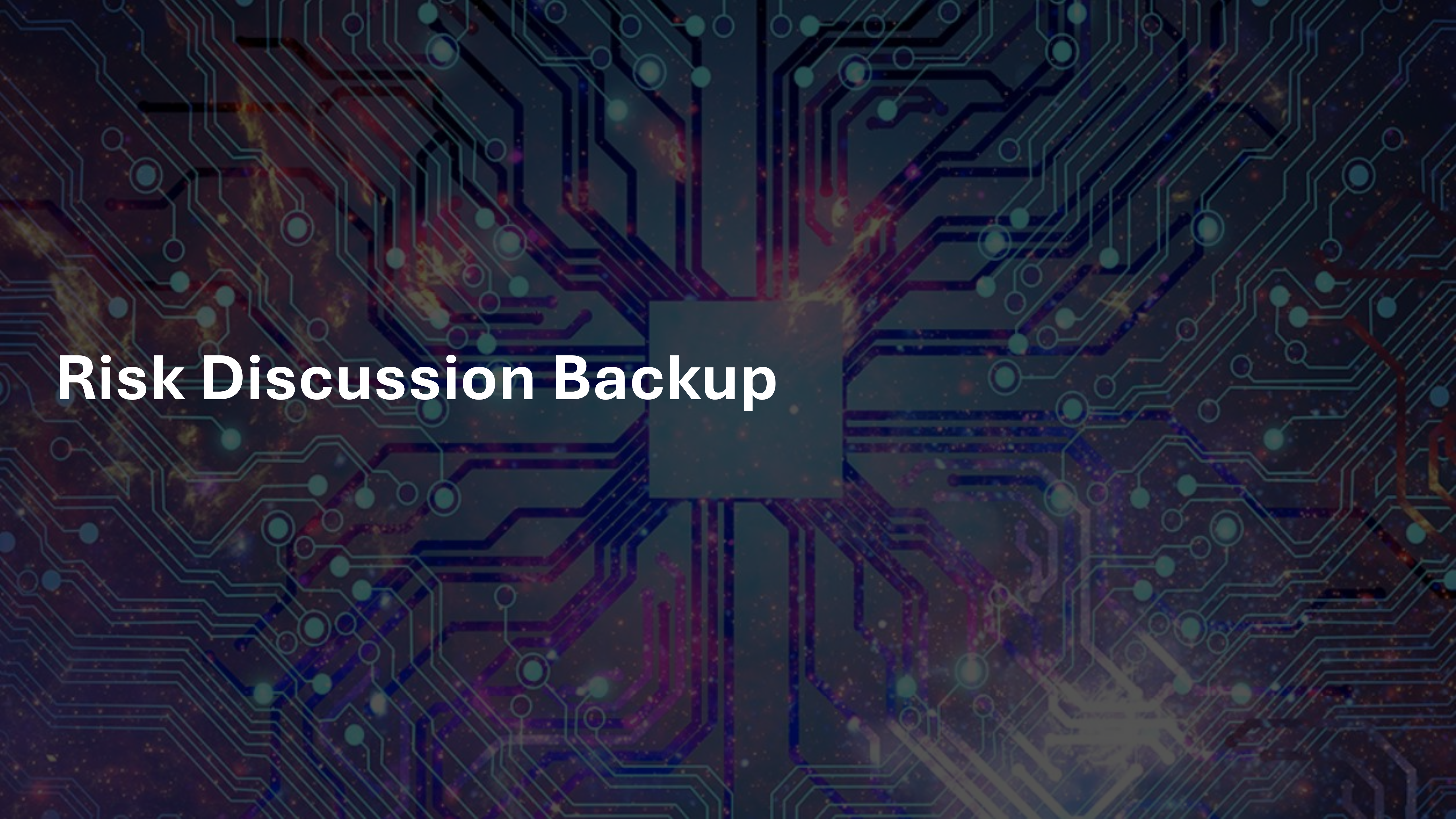
Key FY30 Deliverables:

- 12 fully operational AI hubs
  - Deploying and supporting 10 world leading FMs from FY29
- 12 updated domain-specific foundation models trained on curated datasets and synthetic data
- 60 DOE AI FM applications developed and deployed
- “1000 AI EF” systems deployed
- Suite of curated datasets to enable further model development
  - Partnerships with industry on synthetic data augmentation
- Upgraded compute infrastructure to support model training and inference
  - Supporting 20,000 DOE active scientific/engineering users
- Mature partnerships to sustain AI capabilities and workforce

The key elements are phased in incrementally each year, with 6 initial AI hubs and 3 foundation models in FY26, growing to 12 hubs and 12 mature FM models by FY30, and 60 FM based AI applications. Investments in computing, data curation, partnerships and other enabling capabilities also scale up year-over-year in proportion to the overall budget growth.

Frontier AI
for Science Security and Technology

Risk Discussion Backup



Frontier AI Systems and Risk

General Society AI Risks

- Disinformation and Deepfakes
- Surveillance and Privacy Violations
- Social and Behavioral Engineering
- Bias and Discrimination
- Market Manipulation

Global Security AI Risks

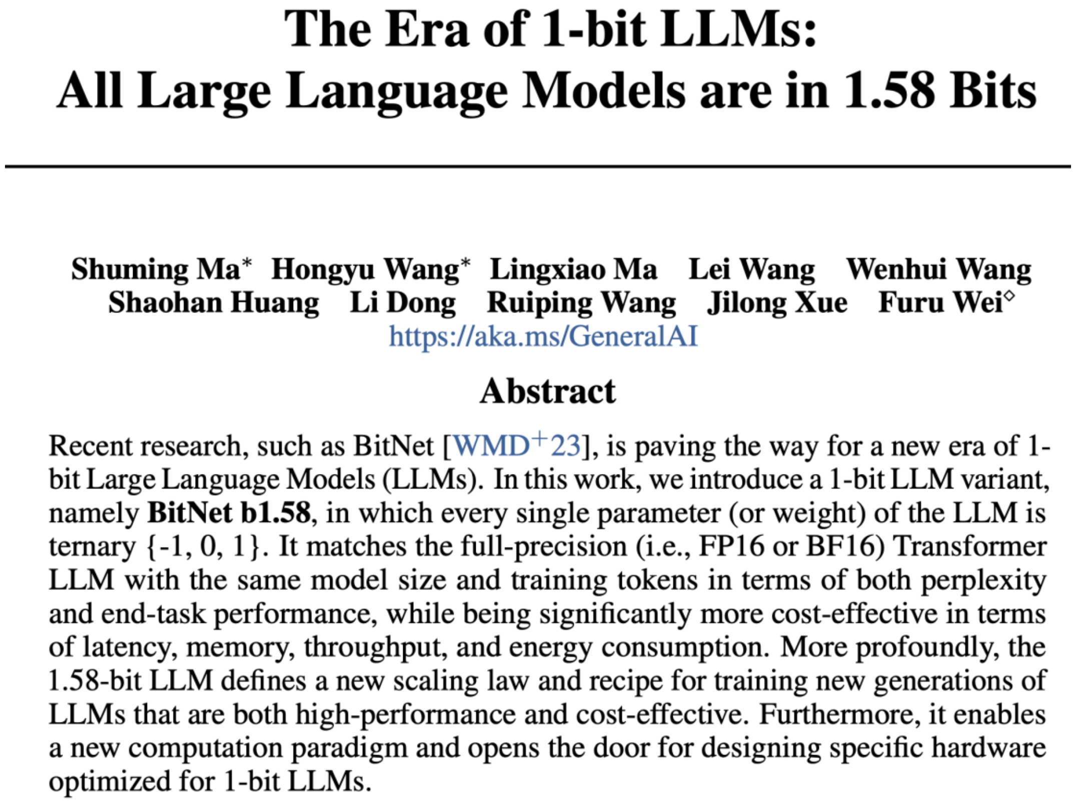
- Autonomous and Swarm Weapons
- Biosecurity and Novel Agents
- Nuclear Proliferation
- New Approaches to Chemical Weapons
- Accelerated Cyberwarfare

The AI risk landscape could change quickly

- There are > 600K LLM models in the wild handful of big models
- Barrier ~ 6FLOPS per Token per Parameter

$$\hat{L}(N, D) \triangleq E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$$

- Push towards small models HQ data
  - Improving quality of data (data efficiency)
  - Same capability with smaller models
- Push towards lower complexity
  - Subquadratic scaling of attention like things
  - Make big models cheaper to train



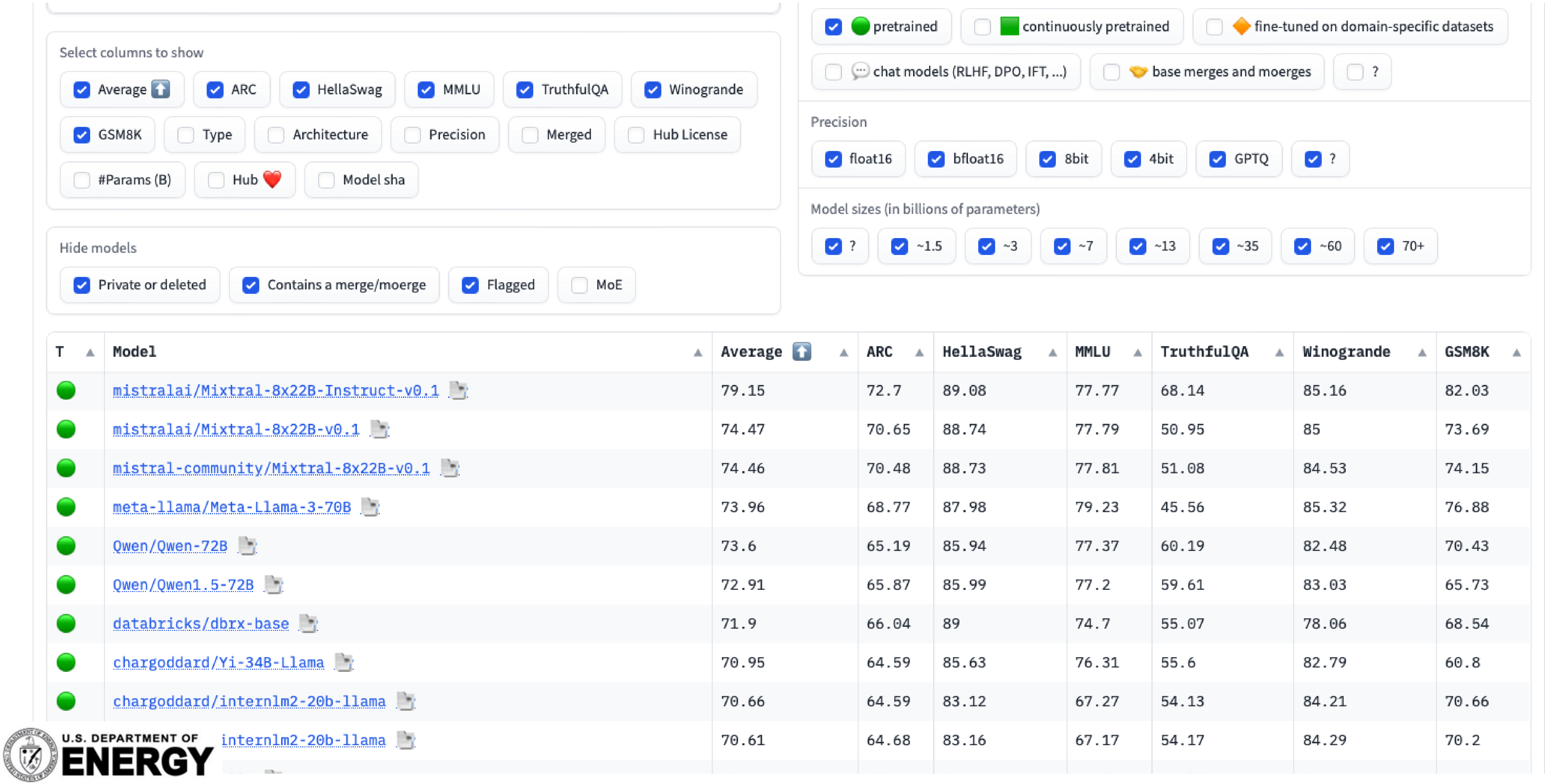
The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits

Shuming Ma* Hongyu Wang* Lingxiao Ma Lei Wang Wenhui Wang
Shaohan Huang Li Dong Ruiping Wang Jilong Xue Furu Wei◇
https://aka.ms/GeneralAI

Abstract

Recent research, such as BitNet [WMD+23], is paving the way for a new era of 1-bit Large Language Models (LLMs). In this work, we introduce a 1-bit LLM variant, namely **BitNet b1.58**, in which every single parameter (or weight) of the LLM is ternary {-1, 0, 1}. It matches the full-precision (i.e., FP16 or BF16) Transformer LLM with the same model size and training tokens in terms of both perplexity and end-task performance, while being significantly more cost-effective in terms of latency, memory, throughput, and energy consumption. More profoundly, the 1.58-bit LLM defines a new scaling law and recipe for training new generations of LLMs that are both high-performance and cost-effective. Furthermore, it enables a new computation paradigm and opens the door for designing specific hardware optimized for 1-bit LLMs.

Open Model Leaderboard



Select columns to show

☑ Average ⬆ ☑ ARC ☑ HellaSwag ☑ MMLU ☑ TruthfulQA ☑ Winogrande

☑ GSM8K ☐ Type ☐ Architecture ☐ Precision ☐ Merged ☐ Hub License

☐ #Params (B) ☐ Hub ❤ ☐ Model sha

Hide models

☑ Private or deleted ☑ Contains a merge/moerge ☑ Flagged ☐ MoE

☑ ● pretrained ☐ 🟩 continuously pretrained ☐ 🔶 fine-tuned on domain-specific datasets

☐ 💬 chat models (RLHF, DPO, IFT, ...) ☐ 🤝 base merges and moerges ☐ ?

Precision

☑ float16 ☑ bfloat16 ☑ 8bit ☑ 4bit ☑ GPTQ ☑ ?

Model sizes (in billions of parameters)

☑ ? ☑ ~1.5 ☑ ~3 ☑ ~7 ☑ ~13 ☑ ~35 ☑ ~60 ☑ 70+

T ▲	Model ▲	Average ⬆ ▲	ARC ▲	HellaSwag ▲	MMLU ▲	TruthfulQA ▲	Winogrande ▲	GSM8K ▲
●	mistralai/Mixtral-8x22B-Instruct-v0.1	79.15	72.7	89.08	77.77	68.14	85.16	82.03
●	mistralai/Mixtral-8x22B-v0.1	74.47	70.65	88.74	77.79	50.95	85	73.69
●	mistral-community/Mixtral-8x22B-v0.1	74.46	70.48	88.73	77.81	51.08	84.53	74.15
●	meta-llama/Meta-Llama-3-70B	73.96	68.77	87.98	79.23	45.56	85.32	76.88
●	Qwen/Qwen-72B	73.6	65.19	85.94	77.37	60.19	82.48	70.43
●	Qwen/Qwen1.5-72B	72.91	65.87	85.99	77.2	59.61	83.03	65.73
●	databricks/dbrx-base	71.9	66.04	89	74.7	55.07	78.06	68.54
●	chargoddard/Yi-34B-Llama	70.95	64.59	85.63	76.31	55.6	82.79	60.8
●	chargoddard/internlm2-20b-llama	70.66	64.59	83.12	67.27	54.13	84.21	70.66
	internlm2-20b-llama	70.61	64.68	83.16	67.17	54.17	84.29	70.2

U.S. DEPARTMENT OF ENERGY

Trustworthy and Responsible AI

Goals

- Alignment with human values and operational constraints
- **Compliance with known laws of physics and logic when required**
- **Exhibit reproducible behavior and results**
- **Robustness to noise and changes in operating environments**
- Respect privacy and are resistant to manipulation to reveal restricted info
- Compliance with regulatory or policy requirements
- **Can explain their reasoning and justify their conclusions**

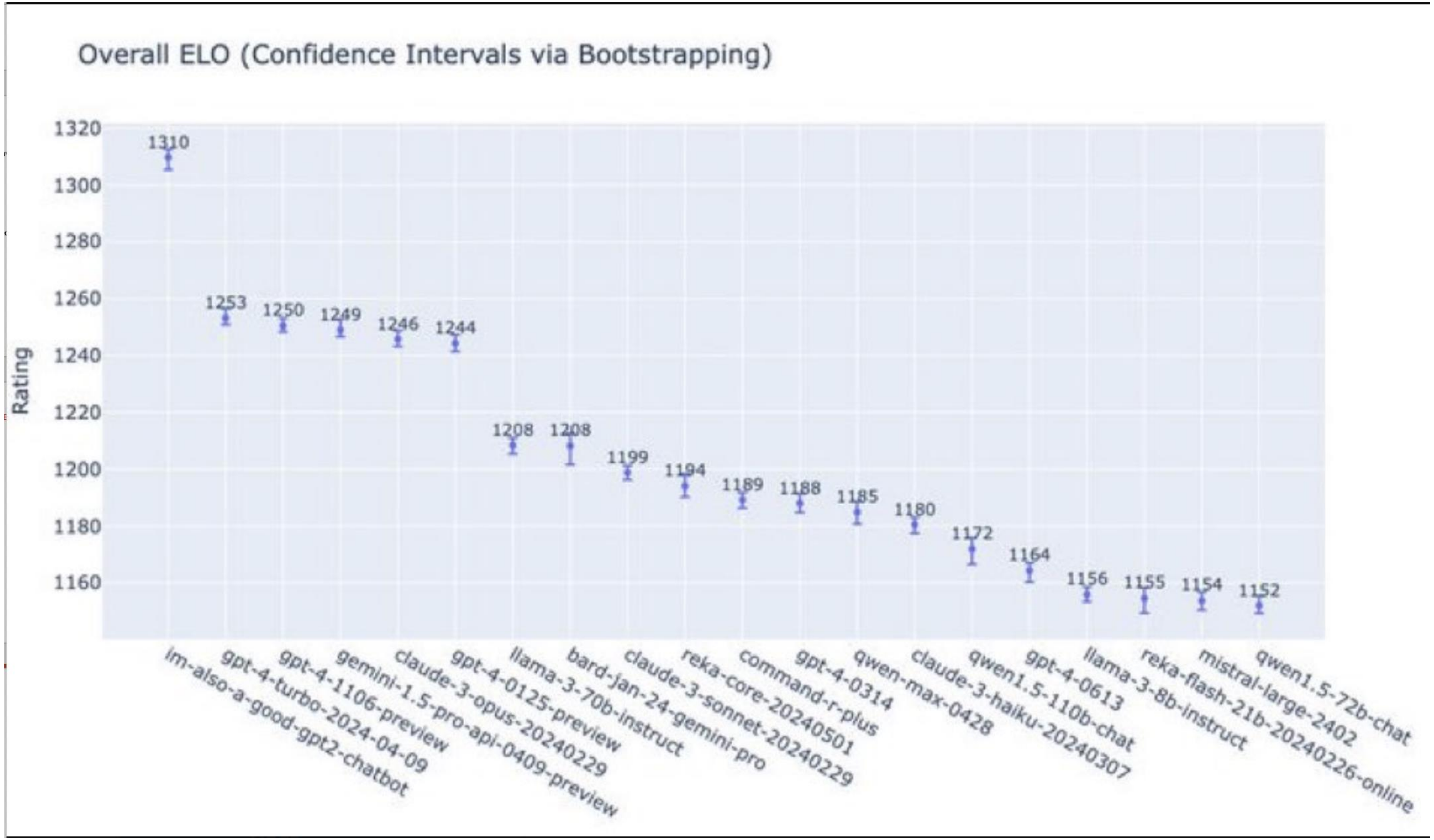
Open Questions

- How to systematically compare behaviors between models?
- How to comprehensively assess the domain knowledge of models?
- How to assess emergent behaviors or novel capabilities in science?
- How to assess scientific knowledge synthesis capabilities?

Thank You for Listening

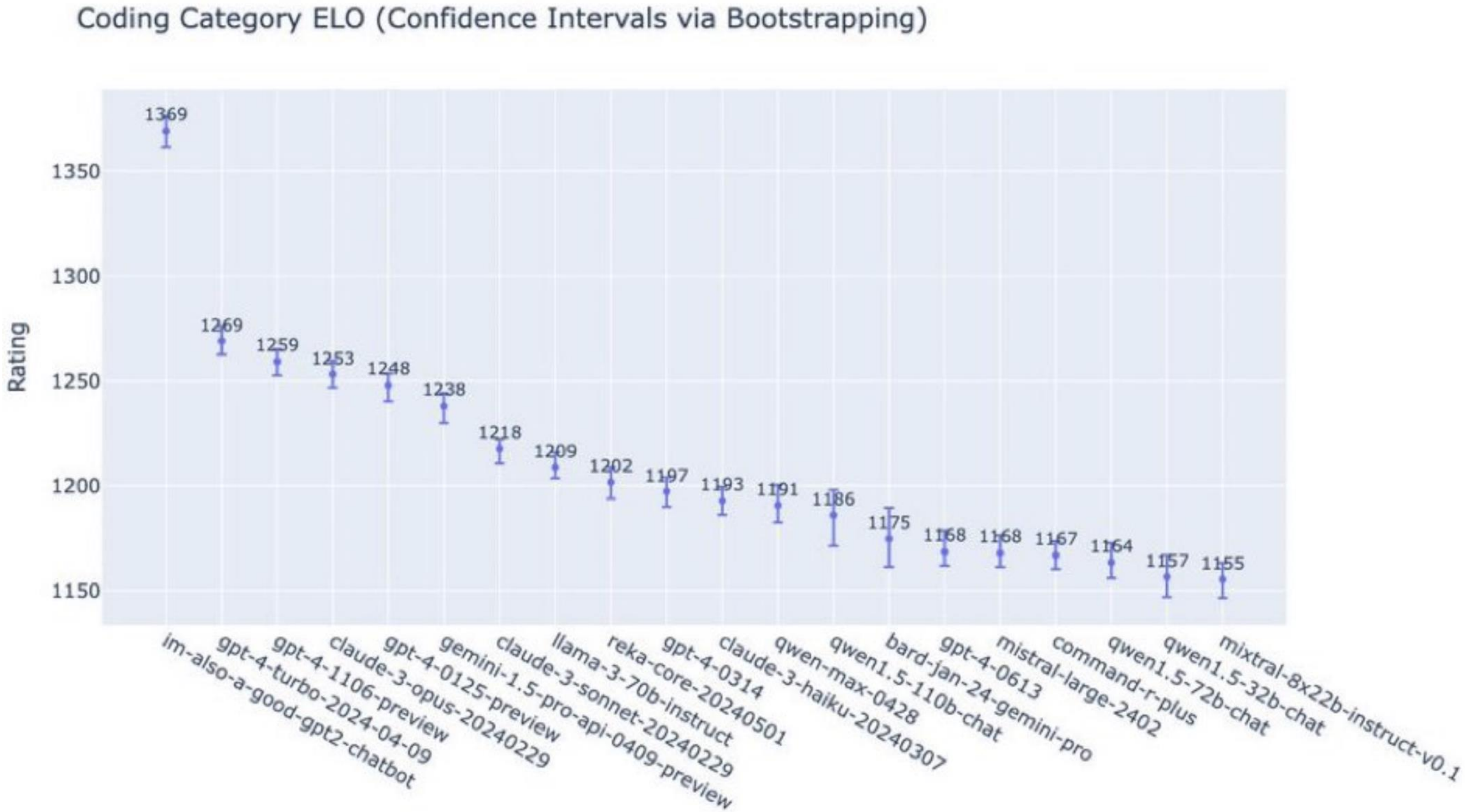
Benchmarking and Evaluation are at the heart of making progress in AI

Benchmarking and Evaluation is Needed for AI4SES



Sam Altman on x.com May 13th, 2024

Benchmarking and Evaluation is Needed for AI4SES



Sam Altman on x.com May 13th, 2024