AI at Scale with a Focus on HPC

September 2024
HPC User Forum @ ANL

Bob Sorensen
Tom Sorensen

www.HyperionResearch.com
www.hpcuserforum.com

Hyperion Research: AI at Scale and HPCs

Key AI-focused Studies and Related Coverage

- LLM Study: *currently available*
  - Conducted across industry verticals, governmental organizations, and academic institutions to capture the current activity, intent, and details regarding experimental and production LLM applications
- AI in the Cloud Study: *Just completed August 2024*
  - Captures insights at the intersection of advanced AI usage and cloud computing resources
- End User Inferencing: *available Fall 2024*
  - Targeted towards the inferencing side of production and near-production integration of advanced AI/LLM
- HPC End-User Multi-Client Study 2024 *available RSN*
  - The seventh edition of a comprehensive study that surveys many HPC customer sites worldwide to create a detailed profile of HPC workloads, plans, and spending on various aspects of HPC, including AI
- AI Subscription Service
  - Hyperion Research's rolling 12-month AI subscription service consisting of targeted documents summarizing and providing analysis on recent studies, news, recent technology developments, products, and HPC User Forum AI-related presentations

Large Language Models (LLMs) and HPC

Focus on the most demanding, and recent, AI space

- AI writ large is undergoing a paradigm shift with the rise of LLM models (e.g., BERT, DALL-E, GPT-4++) trained on broad data sets (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks
- Language–centric class of foundation models, so called to underscore their critically central yet incomplete character
- LLMs have demonstrated uses in language, vision, robotic manipulation, reasoning, human interaction
- LLM are based on standard deep learning and transfer techniques (knowledge learned in one realm that transfers to another), but their scale results in new emergent capabilities

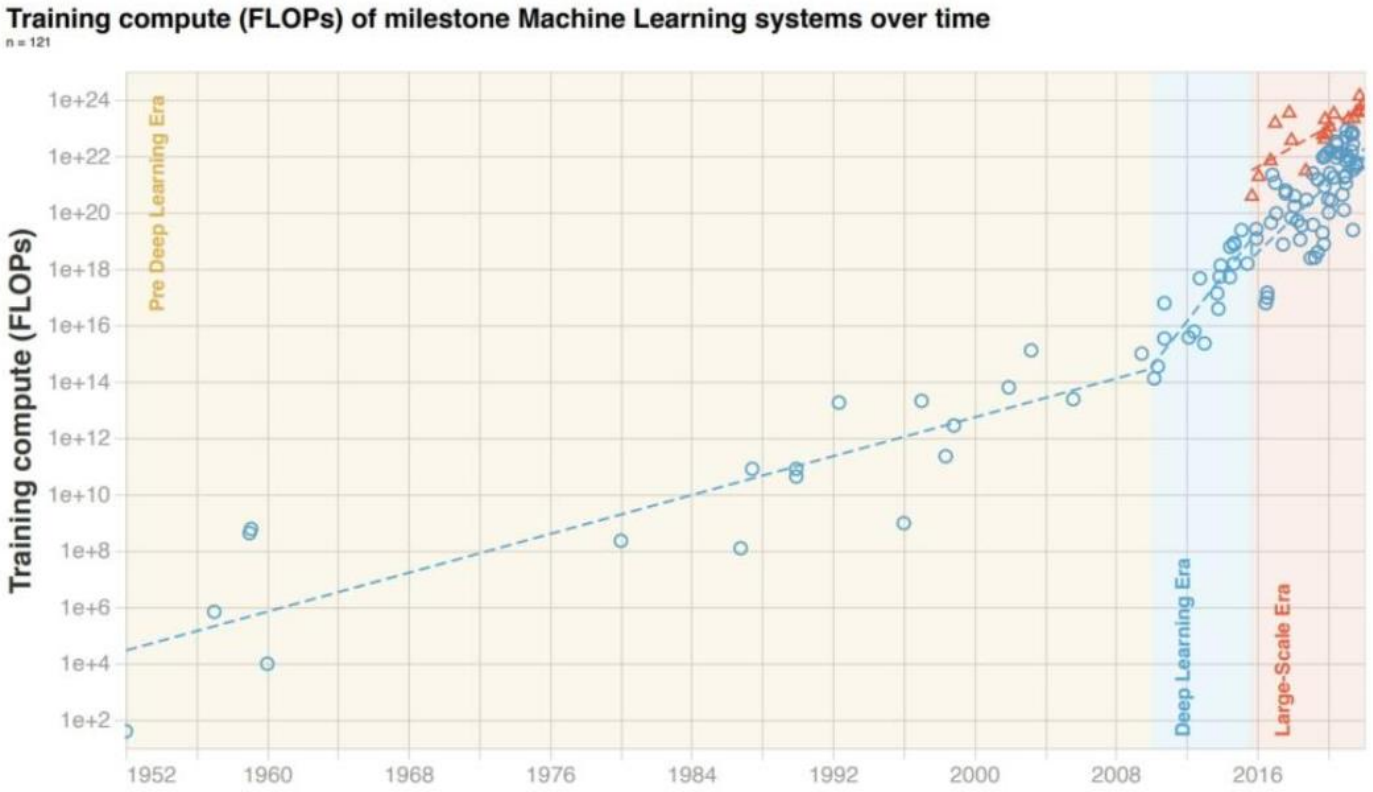
Framing LLM/HPC Requirements

Three elements dominate scaling of LLMs on HPCs

- Compute: the absolute number of floating-point operations needed to train a LLM to a desired degree of accuracy
- Dataset size: input data set used for training the LLM
- Model size: number of parameters/tokens
  - The larger the number of parameters and tokens, the more nuance in the model's understanding of each word's meaning and context
    - Parameters: refer to the trainable weights and biases within the model. They represent the neural connections that are learned during the training process.
    - Tokens: the discrete units into which text/data is divided. For example, the sentence "Large Language Models are impressive" can be tokenized into six tokens: ["Large", "Language", "Models", "are", "impressive"]
- These LLMs fundamentals ultimately define necessary HPC specifications

LLMs Consume Significant FLOPs

LLM flops growth eclipses Top 500 growth

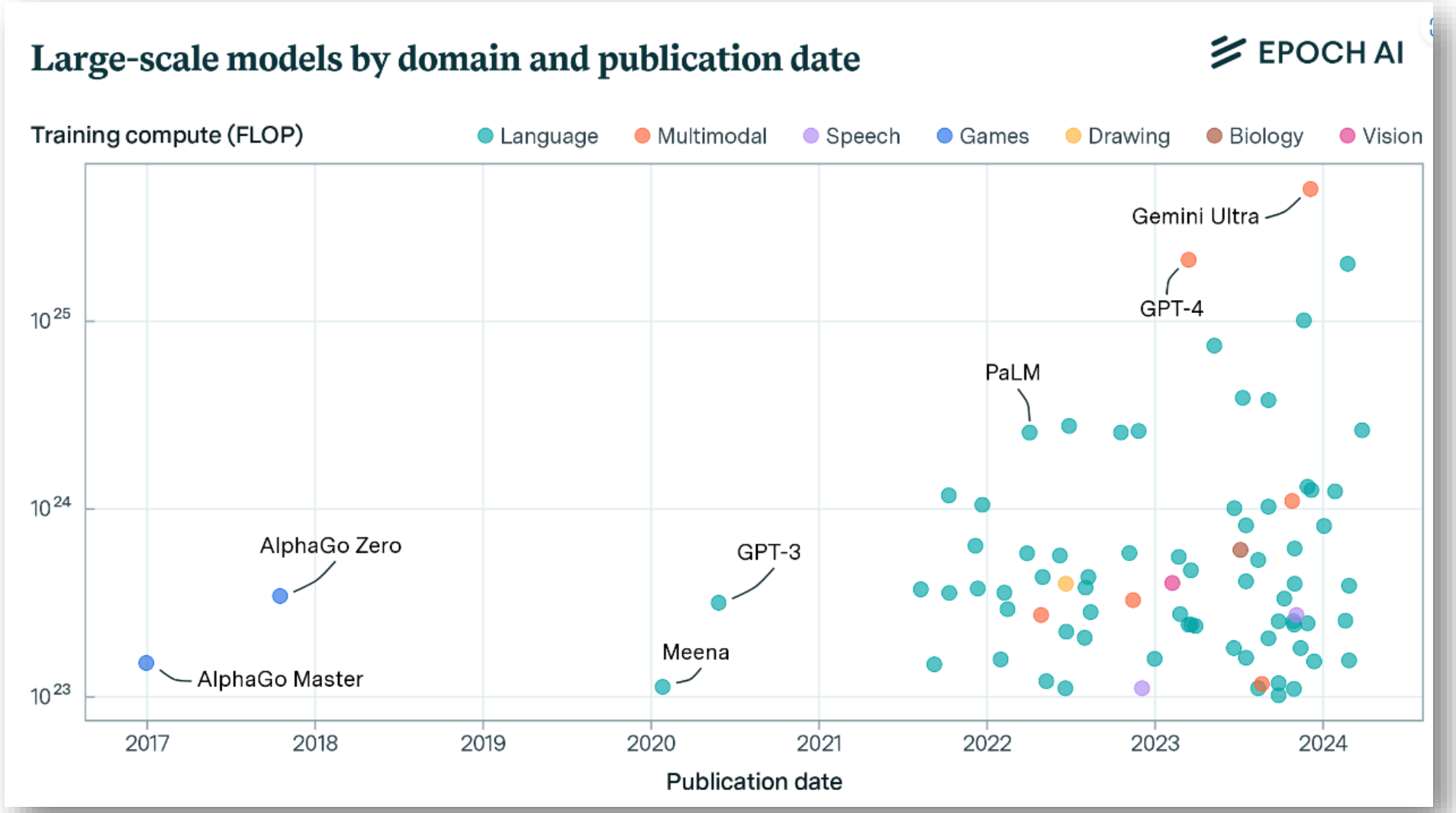


- Currently, highest end training session require on the order of 10^{25} FLOPs (Tens of YottaFLOPs)
- FLOPs requirements doubling every 5.6 months
- Roughly 11X faster than HPC Top One Linpack performance growth rate

- Fastest HPC in 2024, Frontier at Oakridge National Lab, DOE peaks at 10^{18} floating point operations *per second* -> 10^{25} /10^{18} = 10^{7} or ~ten million seconds or ~115 days
  - This assumes 100% efficiency: reality is much less

See *Compute Trends Across Three Eras of Machine Learning*, arXiv:2202.05924

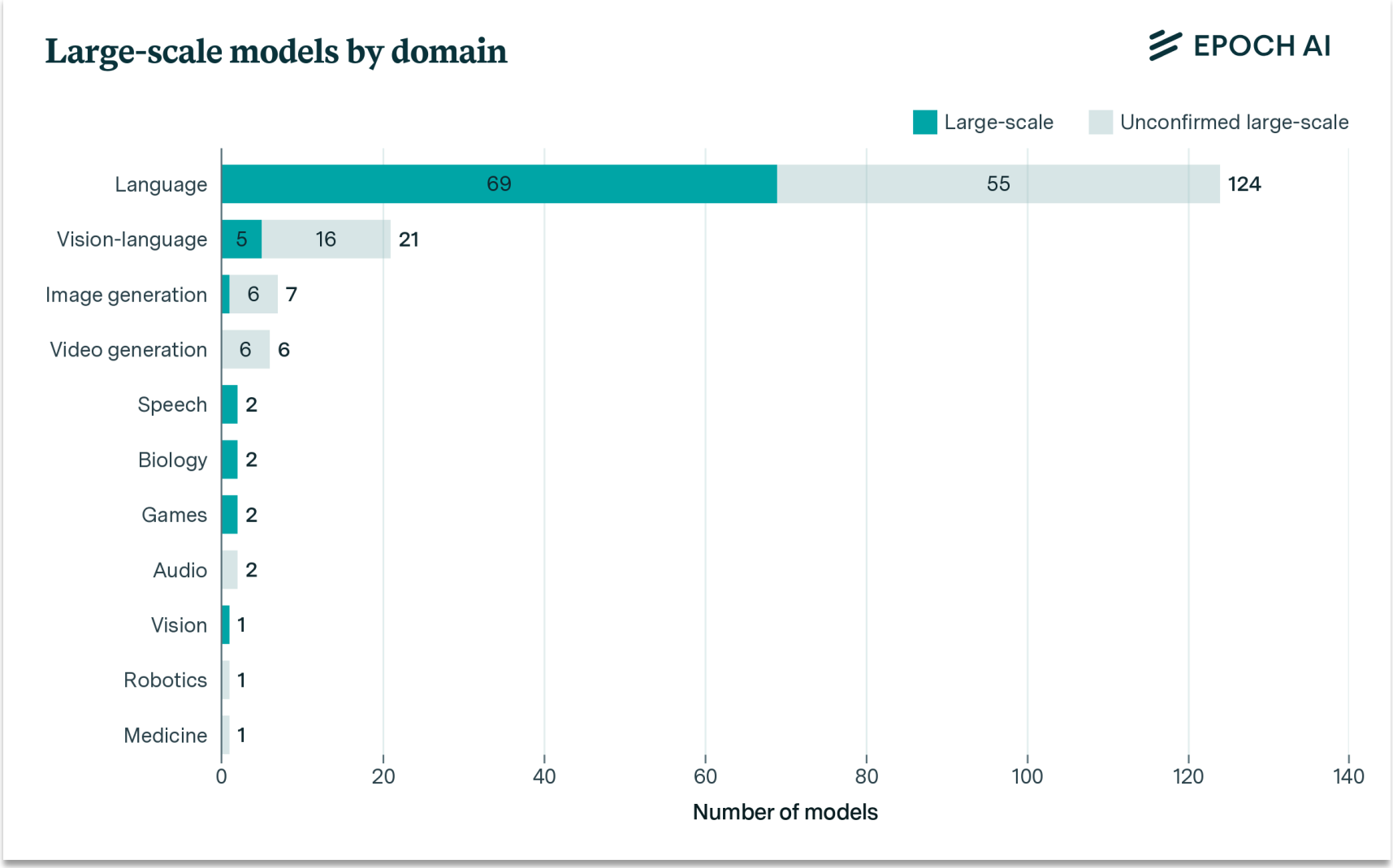
Foundation Training FLOPs Requirements



Tracking foundation models over 10^{23} FLOPs

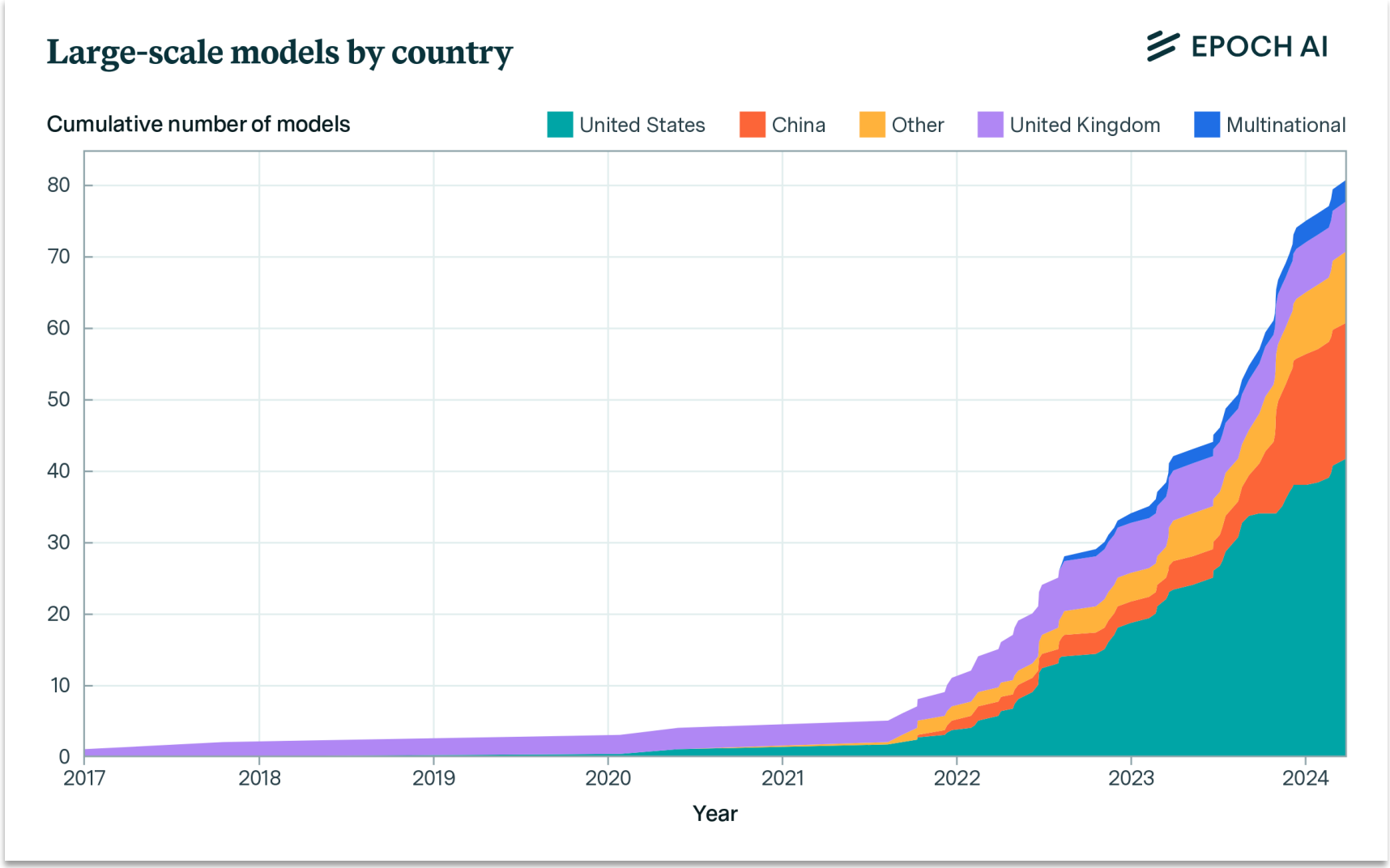
- 81 confirmed models
- 86 unconfirmed
- This is the tip of the model training iceberg
  - See Hugging Face

Foundation Domains



- Language models dominate
- Majority rely on text or graphics data
- Encoding/decoding processes well understood
- Clouds on the Horizon?
  - Encoding/decoding
    - Complexity
    - Computational requirements
  - Time series
  - Multiple data input formats
  - Multiple inputs to a single event

Foundation Developers by Country



Leading large-scale model developers

- Google
- Meta,
- DeepMind4
- Hugging Face
- OpenAI
- Anthropic

China-based:

- Alibaba, Tsinghua

Developers

- Industry (71)
- Academia (2)
- Academia/industry collaborations (6)
- Government (2)

HR Study LLMs: Wide Range of Activities

Key Finding #2: There is a wide range of ongoing LLM-related activity underway within the surveyed organizations

Table 2

Summary of Current and Planned LLM-related Activities

	Currently	Next 12-18 months	Change Over Time
Exploring the range of potential performance enhancements by integrating LLMs into existing HPC-based workloads	58%	48%	-10%
Exploring in-house requirements for integrating LLMs into HPC-based workloads	55%	51%	-4%
Testing/assessing LLM-integrated workload performance	34%	45%	11%
Procuring access to necessary LLM software	31%	31%	0%
Reaching out to LLM hardware and software suppliers for information	30%	35%	5%
Passively monitoring LLM technology developments	27%	14%	-13%
Procuring access to necessary LLM hardware	26%	28%	2%
Standing up limited LLM-integrated pilot programs	26%	36%	10%
Porting LLM capability into existing workloads	25%	34%	9%
Running production level LLM-enabled workloads	22%	50%	28%
Standing up a fully funded LLM research efforts	17%	27%	10%
No current activity	1%	0%	-1%
Other	1%	0%	-1%

N = 100

Respondents could select multiple options.

Source: Hyperion Research, 2023

- The two most prominent activities identified:
  - Exploring the range of potential performance enhancements by integrating LLMs into existing HPC-based workloads
  - Exploring in-house requirements for integrating LLMs into HPC-based workloads

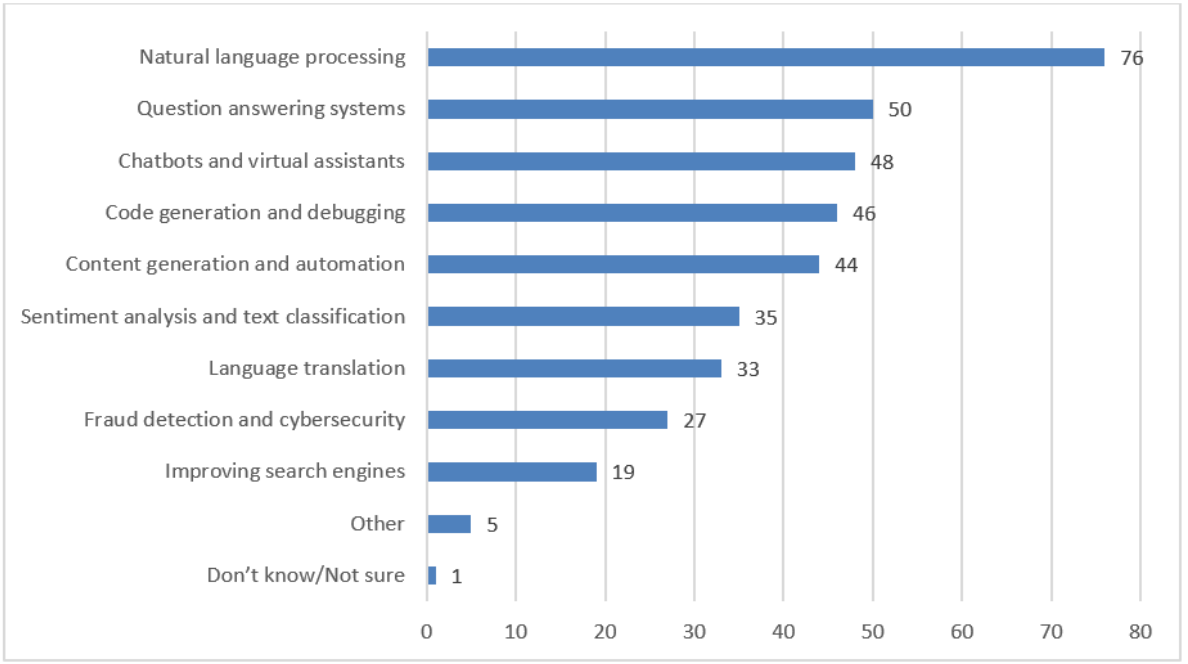
However, a wide range of additional and likely parallel efforts were also underway such as LLM-relevant hardware and software procurements, holding discussions with LLM suppliers, and standing up LLM-related pilot programs

HR Study LLMs: Exploring Multiple Options

Key Finding #4: Numerous LLM applications currently being considered, many looking at multiple options

FIGURE 2

Most Important General LLM Applications to Organization Today



N = 100

Respondents could select more than one answer.

Source: Hyperion Research, 2023

- The most selected general-purpose LLM application options were natural language processing (NLP) and questions answering systems
- Additional options chosen by many included chatbots and virtual assistants, code generation and debugging, and content generation
- Ultimately, the average respondent selected 3.8 different LLM applications currently under consideration

HR Study LLMs: Potential HPC Applicability

Key Finding #5: Many different HPC-related scientific and engineer algorithms were seen as viable for LLM enhancements

Table 3

HPC Workloads Considered Most Promising for LLM Enhancement

	Top Three Options	Single Most Promising Option
Data science/Big data analysis	80%	53%
Monte Carlo methods	37%	5%
Dense linear algebra, algorithms, and libraries	33%	10%
Partial differential equations and boundary value problems	31%	10%
Initial value problems and implicit methods	25%	6%
Sparse linear algebra	24%	6%
Spectral methods - Fast Fourier Transforms (FFTs) and applications	21%	4%
N-body / particle methods	20%	4%
Simple ordinary differential equations	15%	1%
Other	14%	1%

N = 100

For the top three options question, respondents could select the top three in order.

Source: Hyperion Research, 2023

- Data science/big data analysis was selected by 80% of the survey respondents as a top three choice
  - 53% as the single most promising option for LLM enhancement
- Second-tier selections led by Monte Carlo methods, dense linear algebra, and partial differential equations

HR Study LLMs: Challenges Loom

Key Finding #7: There are some significant challenges ahead for organizations seeking to leverage LLM capability

Table 4

Top Three Challenges with Introducing LLMs Into Existing HPC-based Workloads

	Option
Complexity with integrating LLMs into existing HPC-based workloads	46%
High/uncertain development costs	35%
Concerns with cost of LLM-specific hardware or software	33%
Lack of in-house expertise in LLMs	30%
Concerns with technical issues surrounding LLMs such as expandability and hallucinations	29%
Lack of demonstrated return on investment	25%
Long/uncertain implementation times	23%
Lack of credible data sources for LLM training	20%
High/uncertain operational costs	18%
The technology is moving too fast for credible assessment of value	13%
Confusion/uncertainty with LLM vendor selection	11%
Uncertainty of demonstrated computational performance improvements	10%
Other	6%

N = 100

Respondents could select up to three options.

Source: Hyperion Research, 2023

- Top challenges cited were
  - Complexity with integrating LLM into existing HPC-based workloads
  - High/uncertain development costs
  - Concerns with the cost of LLM-specific hardware or software
- Challenges considered less concerning, but still identified, included the technology is moving too fast for credible assessment of value, general confusion, uncertainty with LLM vendor selection, and uncertainty with demonstrated computational performance improvements

HR Study Cloud/On-Premises AI Activities

AI activity in the cloud more prevalent than on-premises in every major category

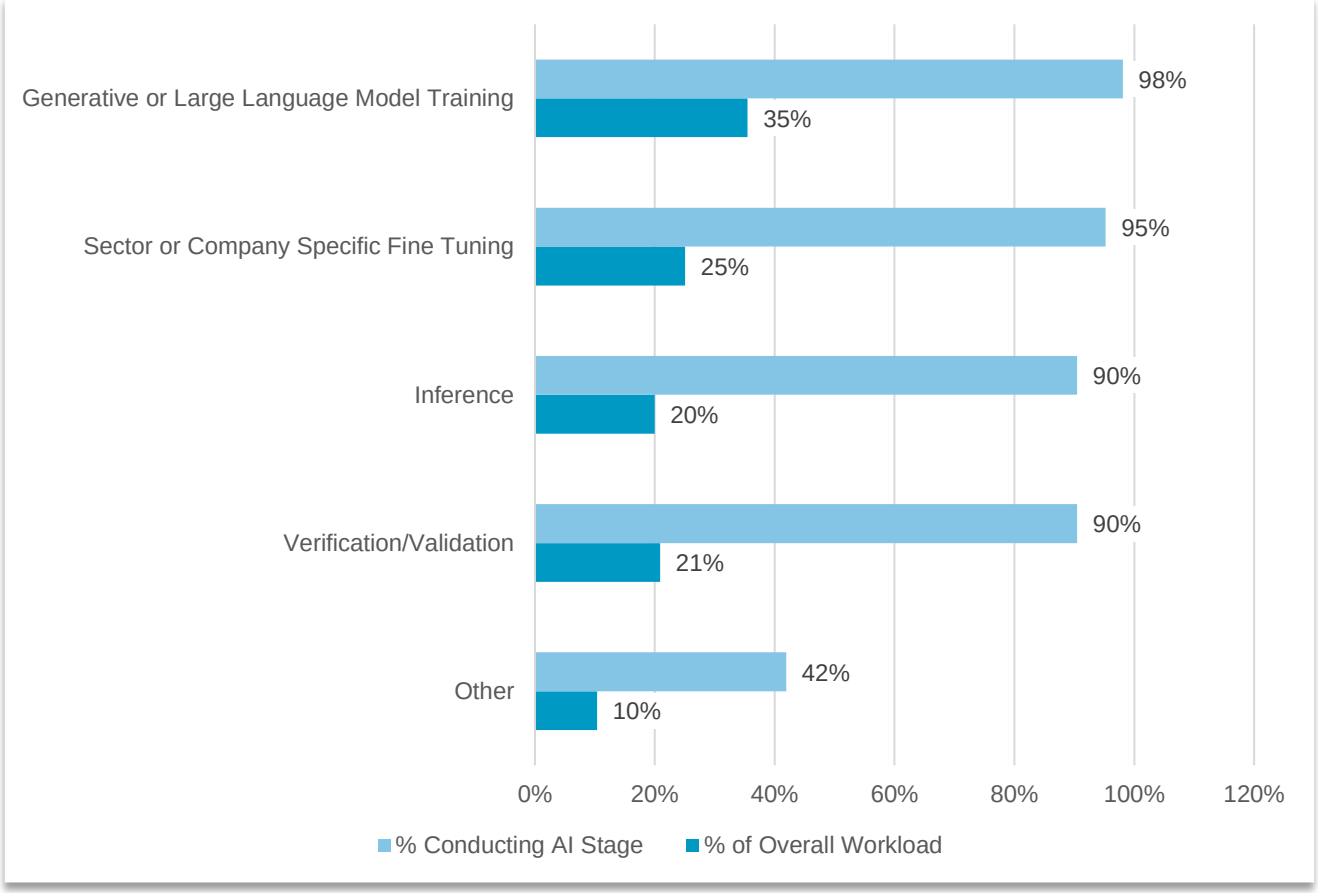
On premises: exploring the range of potential AI performance enhancements	50%
Cloud: exploring the range of potential performance enhancements	64%
On premises: reaching out to AI hardware and software suppliers for information	30%
Cloud: Reaching out to cloud service providers for hardware and software information	35%
On-premises hardware procurement for AI activities	25%
Cloud-based hardware procurement for AI activities	30%
On-premises software procurement for AI activities	20%
Cloud-based software procurement for AI activities	34%
On-premises: standing up limited AI-integrated pilot programs	22%
Cloud: standing up limited cloud-based AI-integrated pilot programs	31%
On premises: testing/assessing AI-integrated workload performance	25%
Cloud: testing/assessing cloud-based AI-integrated workload performance	39%
On premise: running production level AI-enabled workloads on-premises	30%
Cloud: running production level AI-enabled workloads in the cloud	43%

N= 103, Respondents could select all options that apply

Source: Hyperion Research 2024

HR Study Cloud-based AI Workload Stages

LLM training and fine tuning most prevalent AI types and highest workload commitment



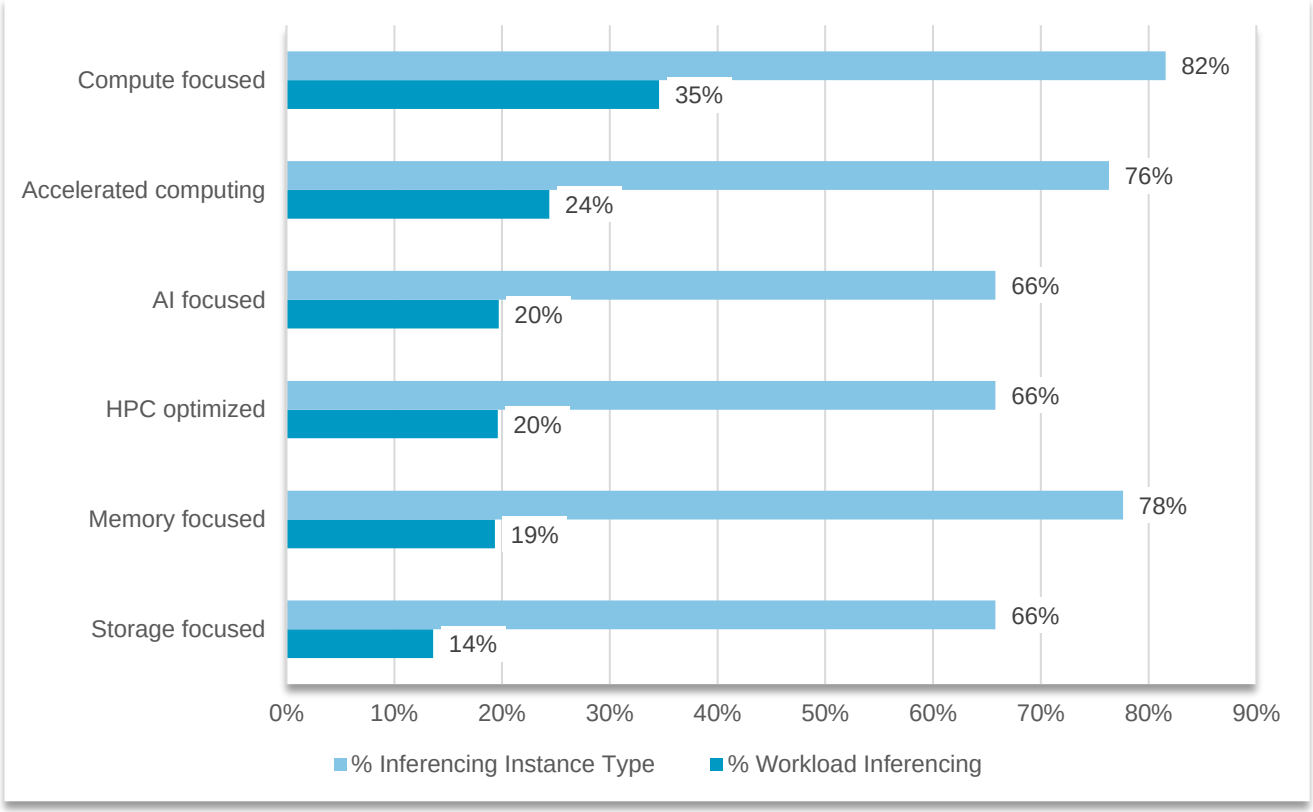
- Generative or LLM training is conducted at almost all respondent sites (98%)
  - For those 98% of respondents, they indicated that 35% of their overall workload was committed to LLM training
- Fine tuning was conducted at 95% of respondent organizations, consuming one quarter of total AI runtime
- 90% of respondents are conducting inference operations, that consumes on average 20% of their overall AI workload

N= 105. Workloads based on runtime. Other not specified.

Source: Hyperion Research 2024

HR Study Cloud-based Inferencing Instances

Compute and accelerated instances support almost 60% of all cloud-based inferencing



- 82% do compute focused cloud-based inferencing for an average of 35% of their cloud-based inferencing workload
- 76% do accelerated computing focused cloud-based inferencing for an average of 24% of their cloud-based inferencing workload
- 66% of respondents do AI-focused inferencing to address an average of 20% of their on-premises inferencing workload

N= 76: Workloads based on runtime.

Source: Hyperion Research 2024

Next Up: Inferencing Study

Building an inferencing capability: critical decision Points for AI end users

- End User Inferencing: available Fall 2024
  - Targeted towards the inferencing side of production and near-production integration of advanced AI/LLM
  - The survey will delve into the hardware and software requirements of user groups and organizations managing high inferencing demands, as well as related budgetary and infrastructure requirements
  - Survey respondents will provide insights on their specific inference types, level of integration and experimentation, and other details of their advanced AI usage including plans and methods of scaling
  - Additional questions welcome

HYPERION RESEARCH

QUESTIONS?

bsorensen@hyperionres.com
tsorensen@hyperionres.com

Too great a burden of knowledge can clog the wheels of imagination. When a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong.

-Arthur C. Clark