Hyperion Research

HPC and Generative AI: A Game Change in the Making?

September 2023

Bob Sorensen
Tom Sorensen

www.HyperionResearch.com
www.hpcuserforum.com

Large Language Models and HPC

Focus on the most recent, and computational demanding,, AI space

- **AI has developed yet another programming paradigm (logic programming, expert systems, ML, DL) with the rise of large language models (LLMs)**
  - Trained on broad data (generally using self-supervision at scale), LLMs can be adapted—or focused through fine-tuning--to a wide range of downstream tasks
  - Applications can include natural language processing, questions answering systems, chatbots and virtual assistants, code generation and debugging, and content generation.
  - LLMs are based on standard deep leaning and transfer techniques (knowledge learned in one realm that transfers to another) <u>but their scale results in new emergent capabilities</u>
  - Current <u>popular</u> exemplars include BERT, DALL-E, GPT-3.5

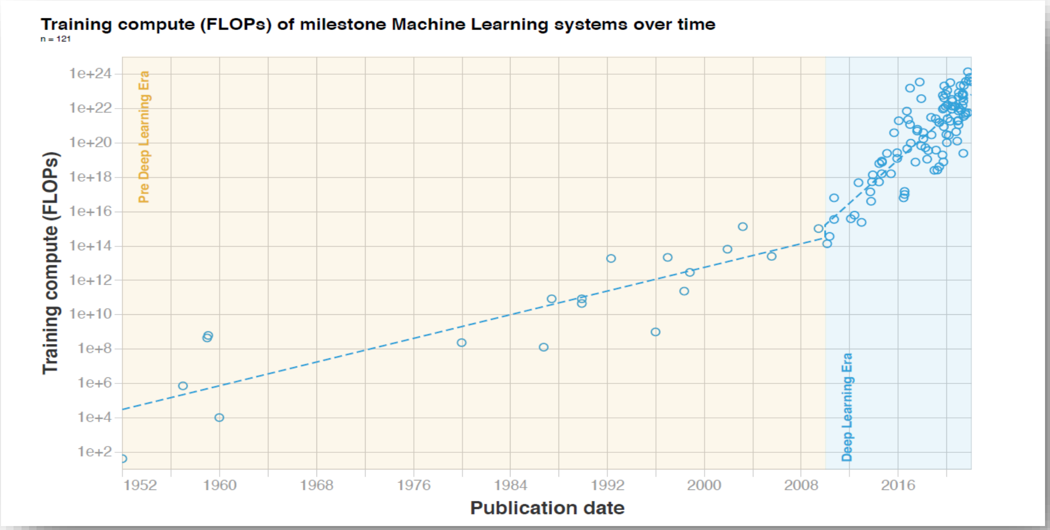
Framing LLM/HPC Requirements

Three elements dominate scaling of LLMs on HPCs

- **<u>Compute</u>: the absolute number of floating-point operations needed to train a LLM to a desired degree of accuracy**
- **<u>Dataset size</u>: input data set used for training the LLM**
- **<u>Model size</u>: number of tokens or parameters**
  - The larger the number of parameters, the more nuance in the model's understanding of each word's meaning and context
- **This scaling heuristic been called the ideal gas law of machine learning**
    - PV= nRT encompasses a range of complex action
    - Scaling moves here as a f(C,D,M)
- **LLMs requirements ultimately define necessary HPC specifications**

LLMs Consume Significant Flops

LLM sum flops growth eclipses Top 500 growth

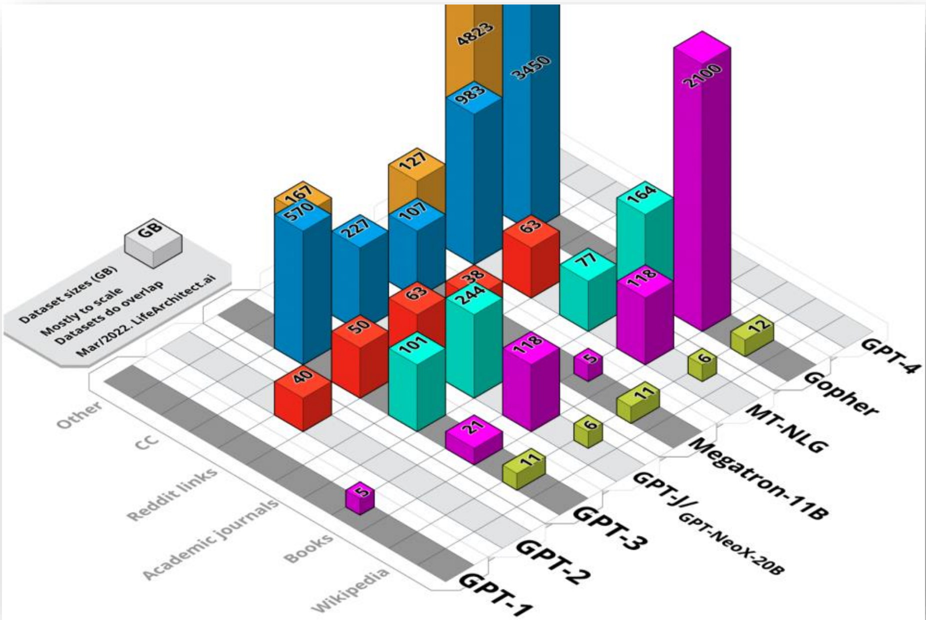


- Pre 2010: Typical training compute flops:
  - On the order of 2 X 10^{12} (2 Tflops)
  - Flops requirements doubling every 21.3 month
  - Post 2010 to Current:
- Currently on the order of 6 X 10^{22} flops (60 Zettaflops)
  - Flops requirements doubling every 5.6 months
  - Roughly 11X faster than HPC Top 1 Linpack performance growth rate

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

Putting the HPC in HPC-AI: Data Set Size

LLM validity related directly to data sources



	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GPT-1		*4.6*					4.6
GPT-2				**40**			40
GPT-3	*11.4*	*21*	*101*	*50*	**570**		753
The Pile v1	**6**	**118**	**244**	**63**	**227**	**167**	825
Megatron-11B	**11.4**	*4.6*		**38**	**107**		161
MT-NLG	*6.4*	*118*	*77*	*63*	*983*	*127*	1374
Gopher	**12.5**	**2100**	*164.4*		*3450*	*4823*	10550

Table 1. Summary of Major Dataset Sizes. Shown in GB. Disclosed in **bold**. Determined in *italics*. Raw training dataset sizes only.

- **There are natural limits here**
  - Is more data better?
  - Is more data even available? Is targeted data available?
  - To what extent will ‘good’ data availability limit LLM progress?

Alan D. Thompson. 2022. What’s in my AI? A Comprehensive Analysis of Datasets used to Train GPT…
https://LifeArchitect.ai/whats-in-my-ai 52021

LLM Data Set: Bigger != Better

Data set size important, but quality matters more

- **Input data set used for training LLMs**
  - Most LLM are trained using a mix of preexisting data sets
  - Some examples of widely-used data sets
    - Common Crawl: Contains billions of web pages and is updated monthly
    - Wikipedia: The online encyclopedia
    - Project Gutenberg: A large collection of free e-books
    - OpenWebText: A collection of over 40GB of text from the web, pre-processed to remove low-quality text
    - Reddit: A popular social news site that contains a wealth of information on a wide range of topics
    - Cornell Movie Dialogs Corpus: A dataset of movie scripts and conversations, a useful source of conversational training data
  - Recent Philippines government study concludes two million domestic crowdworkers currently editing images and text large data sets to be LLM-friendly

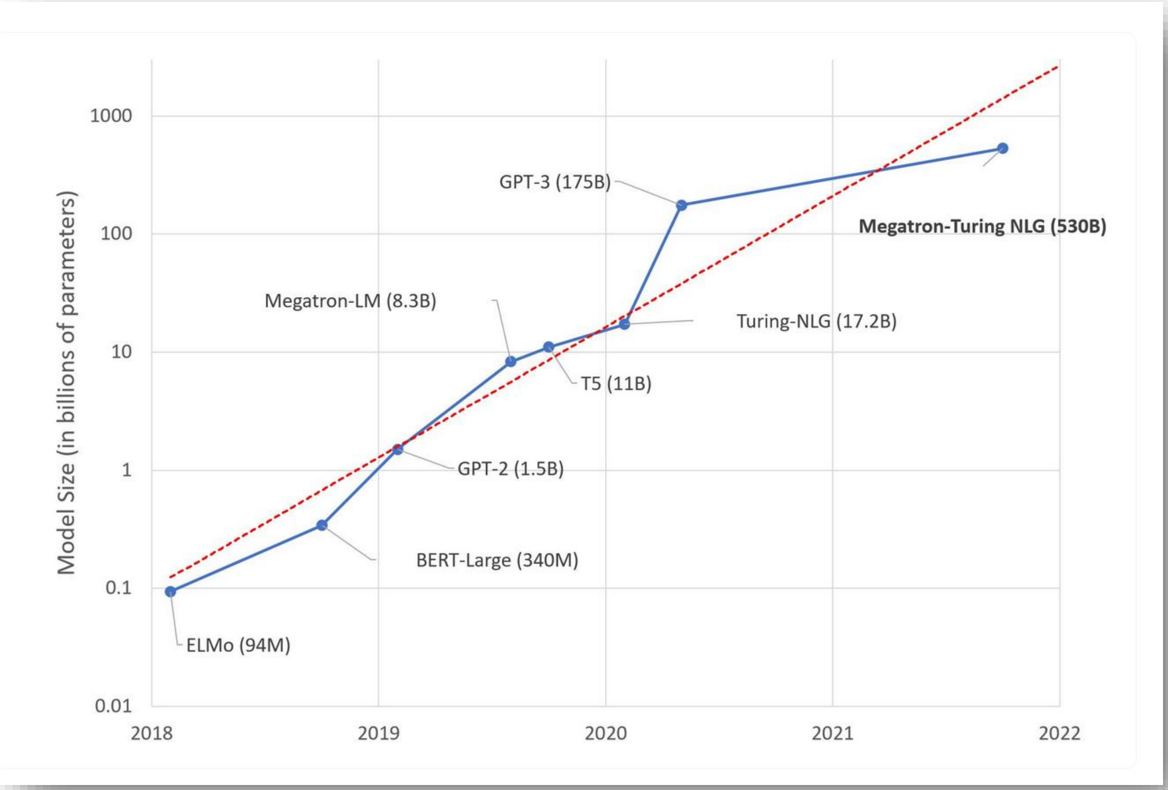
Bigger Models Enable Better LLMs

Tokens and related storage requirements

- **Model size: measure in tokens (or parameters)**
  - Tokens are the basic units of text or code that an LLM uses to process and generate language
  - Can be characters, words, subwords, or other segments of text or code
  - Stored using Byte-Pair Encoding (BPE) scheme
    - 16 bits per token
    - Introduced in 1994 by Phillip Gage as an algorithm for data compression in the C User Journal
- **The larger the number of parameters or tokens, the more nuance in the model's understanding of each word's meaning and context**
- **100-200 billion tokens used in current large scale language models…and that's likely to increase**

Putting HPC in HPC-AI: Model Size

*Language model size on a steep upward trend as well**



- **Megatron Turing used a model size of 530 billion tokens**
- **Training a 530 billion parameter model requires over 10 terabytes of aggregate memory for the model weights, gradients, and optimizer states**

- M-T NLG: 530 billion tokens– three OOMs in four years?
  - Trained on NVIDIA DGX SuperPOD-based Selene HPC

* See Blog at Hugging Face blog, https://huggingface.co/blog/large-language-models

Putting This All Together

Is this (another) new HPC architectural paradigm in the works?

- **Based on a recent LLM analysis by Riken**
- **GPT variant flops requirements**
  - GPT-3.5 (ChatGPT): $3X10^{24}$ flops (estimated)
  - GPT-4.0: $3X10^{25}$ flops (estimated)
- **OpenAI System: Microsoft/Open AI collaboration**
  - Top 5 system when stood up
  - GPU-based BF16 312 Tflop/s x 25,000 = 7.8 Eflop/s TPP
  - GPT-3.5 (ChatGPT): 4.5 days X 2
  - GPT-4.0 45 days X 2
- **Fugaku:**
  - FP32 6.76 Tflop/s X 158,976 = 1.07 Eflop/s (TPP)
  - GPT3.5 (ChatGPT): 32 days X 10
  - GPT-4.0 45 days X 2: 328 days X 10 ~= 8.9 years

Distributed Training of Large Language Models on Fugaku, https://t.co/idofa7Tjyu

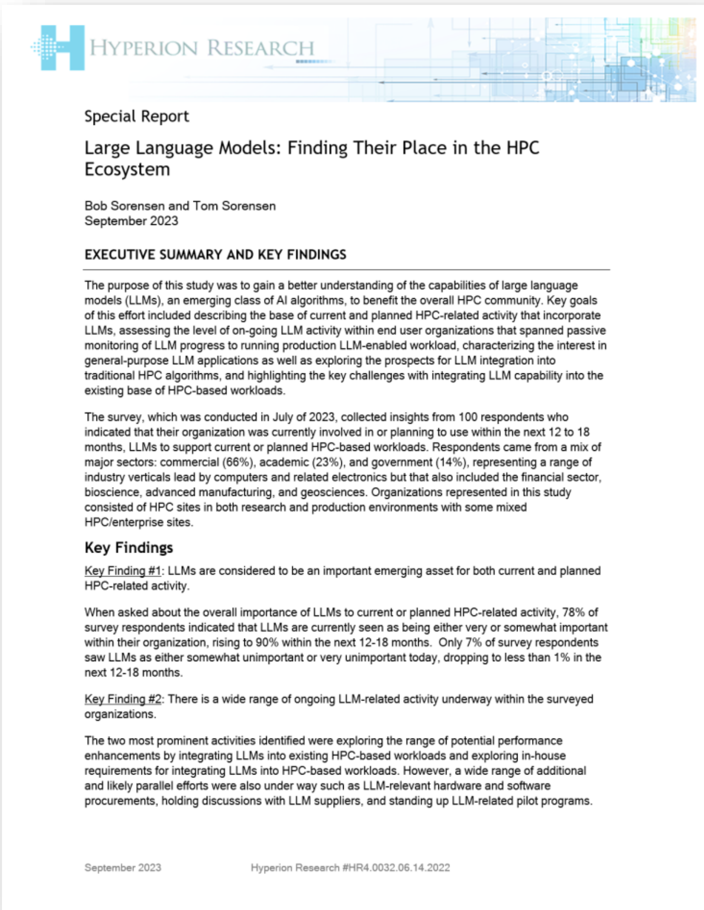
LLM HPCs Of Note

New Machines and New Suppliers

- **Google's AI-focused A3 VM HPC**
  - 26,000 Nvidia H100 Hopper GPUs in a single cluster
  - 26 Eflop/s of "AI performance"
  - GPU-to-GPU data interface supporting CPU bypass
- **Microsoft/OpenAI HPC**
  - Announced 08/2020, 100's of million of dollars
  - AMD Epyc Rome CPUs: 285,000 cores total
  - 10,000 Nvidia A100 Ampere GPUs
  - Computational engine for GPT-3
- **Meta Research SuperCluster (Phase 2)**
  - 16,000 A100s
  - One of the largest known flat InfiniBand fabrics in the world, with 48,000 links and 2,000 switches.
- **Nvidia DGX Cloud**
  - Nvidia H100 or A100s, 640 GB memory instances
  - $36,999/ per month per instance
  - Cloud instances through Oracle, Azure, Google

Large Language Models: Finding Their Place in the HPC Ecosystem

Soon to be available HR Study



Hyperion Research

Special Report

Large Language Models: Finding Their Place in the HPC Ecosystem

Bob Sorensen and Tom Sorensen
September 2023

EXECUTIVE SUMMARY AND KEY FINDINGS

The purpose of this study was to gain a better understanding of the capabilities of large language models (LLMs), an emerging class of AI algorithms, to benefit the overall HPC community. Key goals of this effort included describing the base of current and planned HPC-related activity that incorporate LLMs, assessing the level of on-going LLM activity within end user organizations that spanned passive monitoring of LLM progress to running production LLM-enabled workload, characterizing the interest in general-purpose LLM applications as well as exploring the prospects for LLM integration into traditional HPC algorithms, and highlighting the key challenges with integrating LLM capability into the existing base of HPC-based workloads.

The survey, which was conducted in July of 2023, collected insights from 100 respondents who indicated that their organization was currently involved in or planning to use within the next 12 to 18 months, LLMs to support current or planned HPC-based workloads. Respondents came from a mix of major sectors: commercial (66%), academic (23%), and government (14%), representing a range of industry verticals lead by computers and related electronics but that also included the financial sector, bioscience, advanced manufacturing, and geosciences. Organizations represented in this study consisted of HPC sites in both research and production environments with some mixed HPC/enterprise sites.

Key Findings

Key Finding #1: LLMs are considered to be an important emerging asset for both current and planned HPC-related activity.

When asked about the overall importance of LLMs to current or planned HPC-related activity, 78% of survey respondents indicated that LLMs are currently seen as being either very or somewhat important within their organization, rising to 90% within the next 12-18 months. Only 7% of survey respondents saw LLMs as either somewhat unimportant or very unimportant today, dropping to less than 1% in the next 12-18 months.

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

The two most prominent activities identified were exploring the range of potential performance enhancements by integrating LLMs into existing HPC-based workloads and exploring in-house requirements for integrating LLMs into HPC-based workloads. However, a wide range of additional and likely parallel efforts were also under way such as LLM-relevant hardware and software procurements, holding discussions with LLM suppliers, and standing up LLM-related pilot programs.

September 2023 Hyperion Research #HR4.0032.06.14.2022

- **190 invitations to gather 100 complete responses**
- **Commercial (66%), academic (23%), and government (14%)**
- **Verticals: computers and related electronics but that also included the financial sector, bioscience, advanced manufacturing, and geosciences**
- **Regional variety US and non US resident headquarters locations**

Arraying Current LLM Activities

	Currently	Next 12-18 months	Change Over Time
Exploring LLM potential for existing HPC-based workloads	58%	48%	-10%
Exploring LLM integration requirements for 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%

QUESTIONS?

bsorensen@hyperionres.com
tsorensen@hyperionres.com