



**Hewlett Packard
Enterprise**

Commonwealth Computer Summit

Practical tips for deploying GenAI and LLMs

Steve Heibein, HPE Public Sector AI Chief Technologist

October 16, 2023

LARGE LANGUAGE MODELS AS THE FOUNDATION FOR GENERATIVE AI HAVE THE POTENTIAL TO:

Disrupt nearly every industry

promising both competitive advantage and creative destruction¹

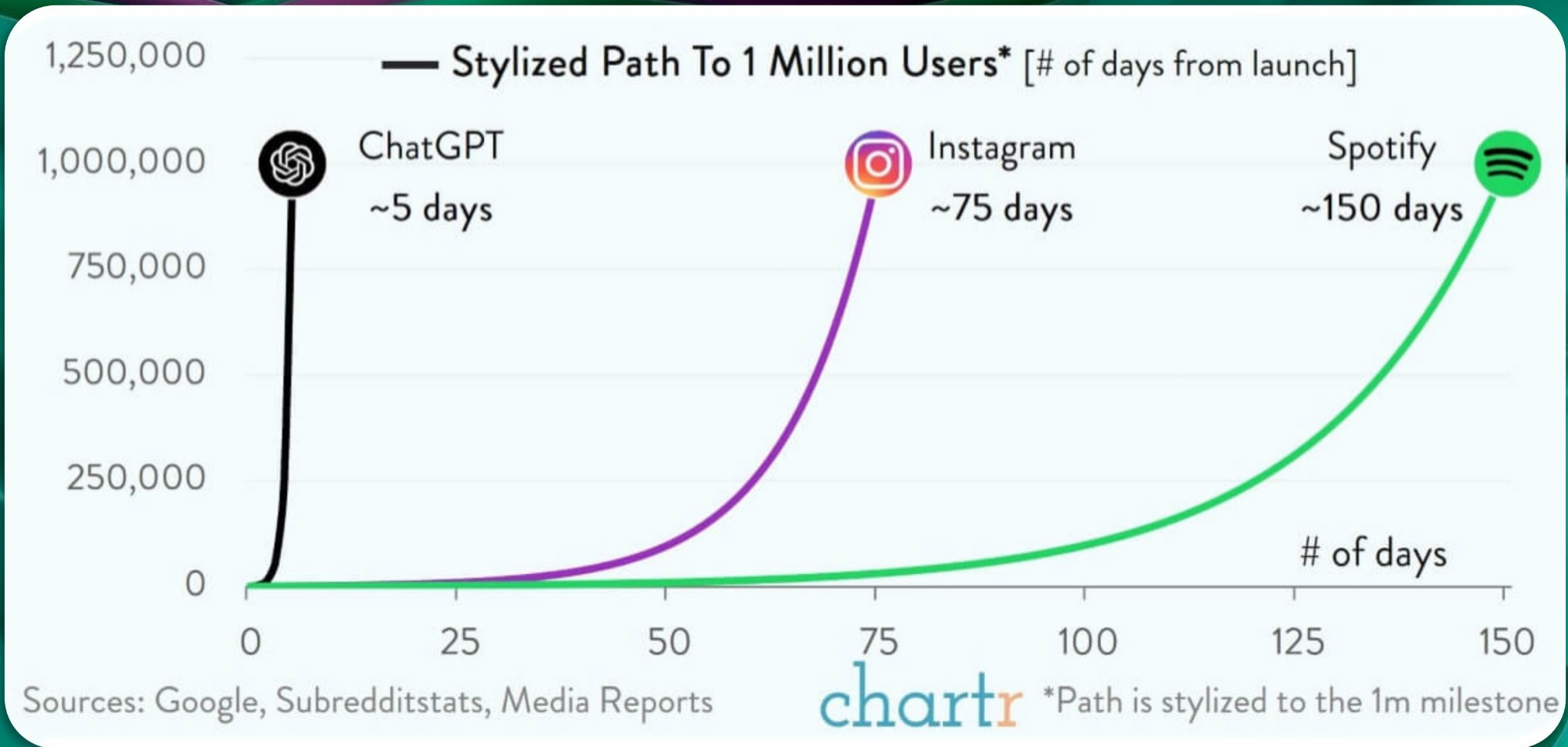
Change the anatomy of work

augmenting the capabilities of knowledge workers by automating 60% to 70% of their individual activities²

¹ Boston Consulting Group, [The CEO's Guide to the Generative AI Revolution](#), March 2023

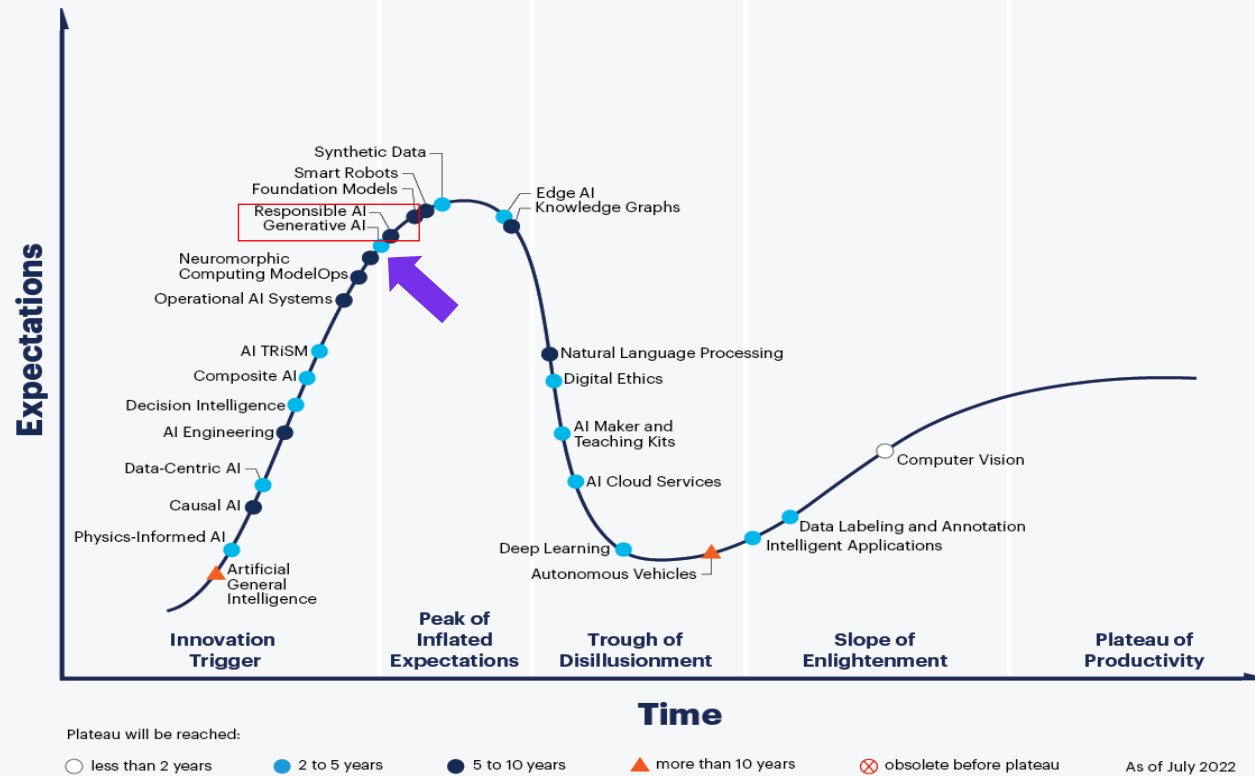
² McKinsey & Company, [The Economic Potential of Generative AI – The next productivity frontier](#), June 2023

AI goes mainstream with ChatGPT



Generative AI : Time is Now

Hype Cycle for Artificial Intelligence, 2022



Gartner Generative AI predictions:

- By 2025, we expect more than 30% — up from zero today — of new drugs and materials to be systematically discovered using generative AI techniques
- By 2025, 30% of outbound marketing messages from large organizations will be synthetically generated, up from less than 2% in 2022.
- By 2030, a major blockbuster film will be released with 90% of the film generated by AI (from text to video), from 0% of such in 2022.

gartner.com

Source: Gartner
© 2022 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1957302

Gartner®

Why all the Hype about Generative AI?

Compared with traditional AI methods, Generative AI has the potential to ...

- Exhibit emergent capabilities
- Less reliance on labeled data
- Have better predictive accuracy
- More seamlessly handle multimodal data
- Novel interfaces for human-AI interaction (e.g., prompting)



Large Language Models Boost Productivity

Pros

- Like having an infinite number of Assistants
- Knows a lot about a lot of topics

Cons

- Needs guidance
- Confidently Incorrect
- “Hallucinates”



Generative AI: Expanding the Output of AI Systems

Today

AI Systems

Mostly Classify
or Predict

Symbols

Lifetime Value Score, Intents, Risk Levels, “Turn Left, Image Category, Emotion Type ...”

Generative

AI Systems

Expanded to
Generate

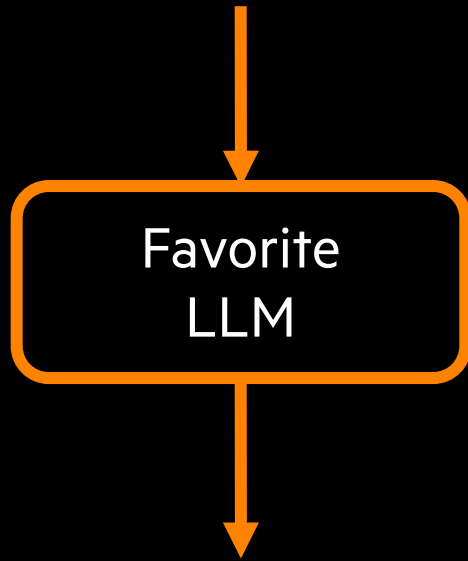
Artifacts

Video, Audio, Language, Images, Code, Synthetic Data, Design for Real World Objects



GenAI Examples

Input Prompt: Write a term paper with citations about the Cuban Missile Crisis explaining President Kennedy's options.



Output: 4-pages of very confident text written like a high schooler.



Example of GPT-4 visual input:

User What is funny about this image? Describe it panel by panel.



Source: <https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/>

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

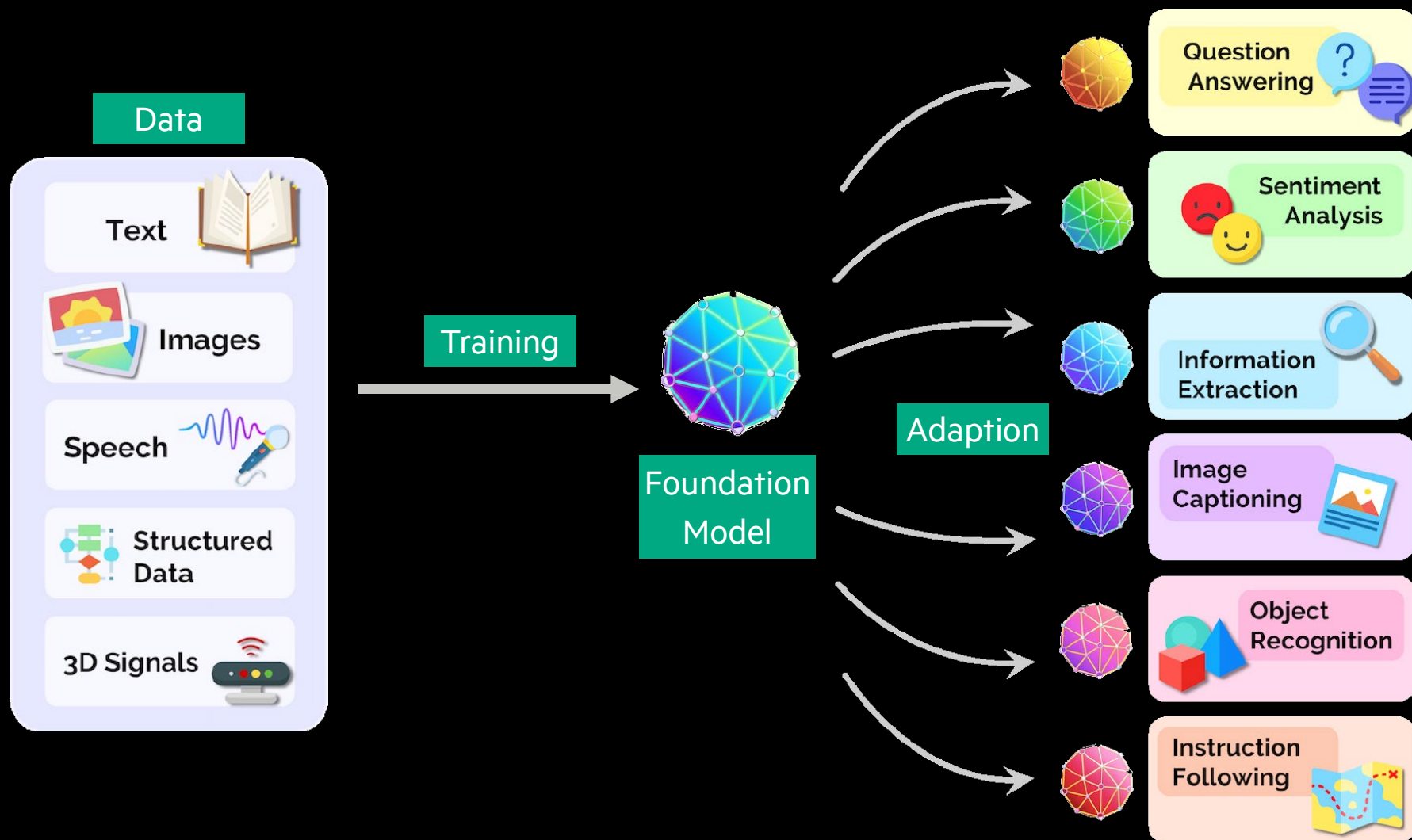
The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

DALL-E 2 Prompt: An astronaut riding a horse in photorealistic style.



Source: OpenAI : <https://openai.com/product/dall-e-2>

Large Language Models – Multi-mode

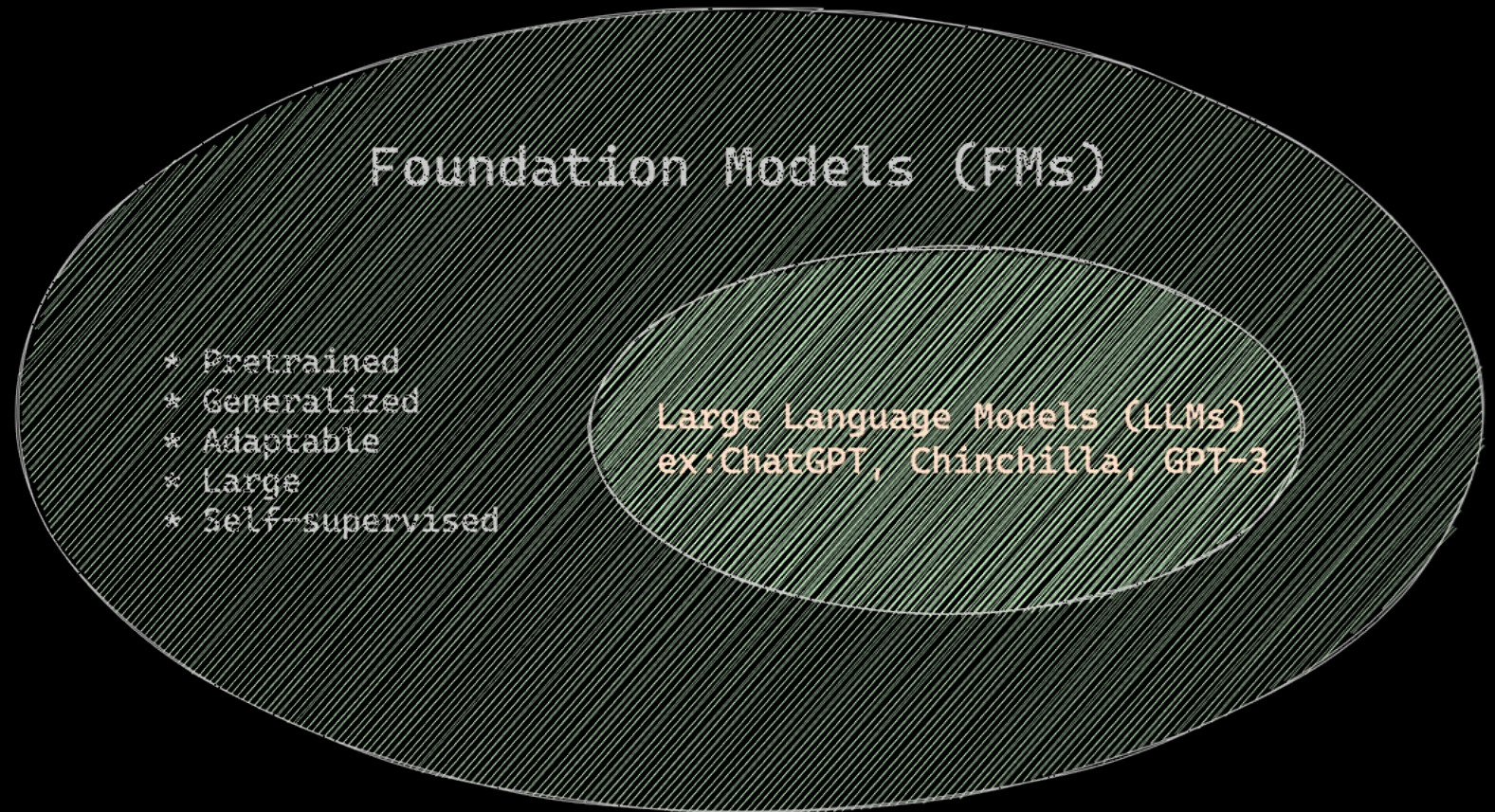


Foundation Models – Contain Foundational Knowledge

Contain basic understanding of words, sentence structure, patterns, context

5 Characteristics:

- Pretrained
- Generalized
- Adaptable
- Large
- Self-supervised



FMs are models trained on broad data (using self-supervision at scale) that can be adapted to a wide range of downstream tasks.
<https://hai.stanford.edu/news/reflections-foundation-models>

Generative AI Application Areas

Application Areas

- Question Answering
- Conversation
- Code completion
- Creative Generation
- Search
- Translation
- Classification

Extended App by Chaining

- Multi-modal data sources (public, proprietary, etc.)
- System & user inputs
- Prompt Templates
- Vector databases
- Links to real world (plug-ins...)

Common Use Cases

- Summarizing documents
 - Legal
 - Financial
- Detecting fraud in claim forms
- Performing NER and semantic search in audio transcriptions
- Answering complex questions at a pharmacy about prescriptions
- Helping physicians write patient post-visit reports

Will AI Take my Job?

- Maybe, but probably not
 - Low skill information workers are at risk
 - Customer Support / Customer Service
 - Data Entry & Analysis
 - Accountants + HR
- AI will augment (disrupt) many jobs
 - Creatives (content, “influencers”, marketing)
 - Software
 - Law (Legal research, Contract analysis)
 - Medicine (medical imaging)
 - Education & Learning



ChatGPT – Application Built on Generative AI Large Language Model



- Chatbot version based on OpenAI's LLM, Generative Pre-trained Transformer 3.5 (GPT-3.5) model
- Launched on Nov. 30, 2022
- 175 billion parameters
- Text-based tool that can produce human-like responses to user requests
 - Poetry in the style of William Shakespeare
 - Advice on workout plan
 - Dinner suggestion given contents of your kitchen
 - Book, contract, article summary
- ChatGPT performance is a substantial step forward from using Google search or online symptom checker.

Evolution of GPT – Generative Pretrained Transformer

| Model | Launch Date | Training Data | No. of Parameters | Max. Sequence Length |
|-------|---------------|--|------------------------------|----------------------|
| GPT-1 | June 2018 | Common Crawl, BookCorpus | 117 million | 1024 |
| GPT-2 | February 2019 | Common Crawl, BookCorpus, WebText | 1.5 billion | 2048 |
| GPT-3 | June 2020 | Common Crawl, BookCorpus, Wikipedia, Books, Articles, and more | 175 billion | 4096 |
| GPT-4 | March 2023 | Unknown | Estimated to be in trillions | Unknown |

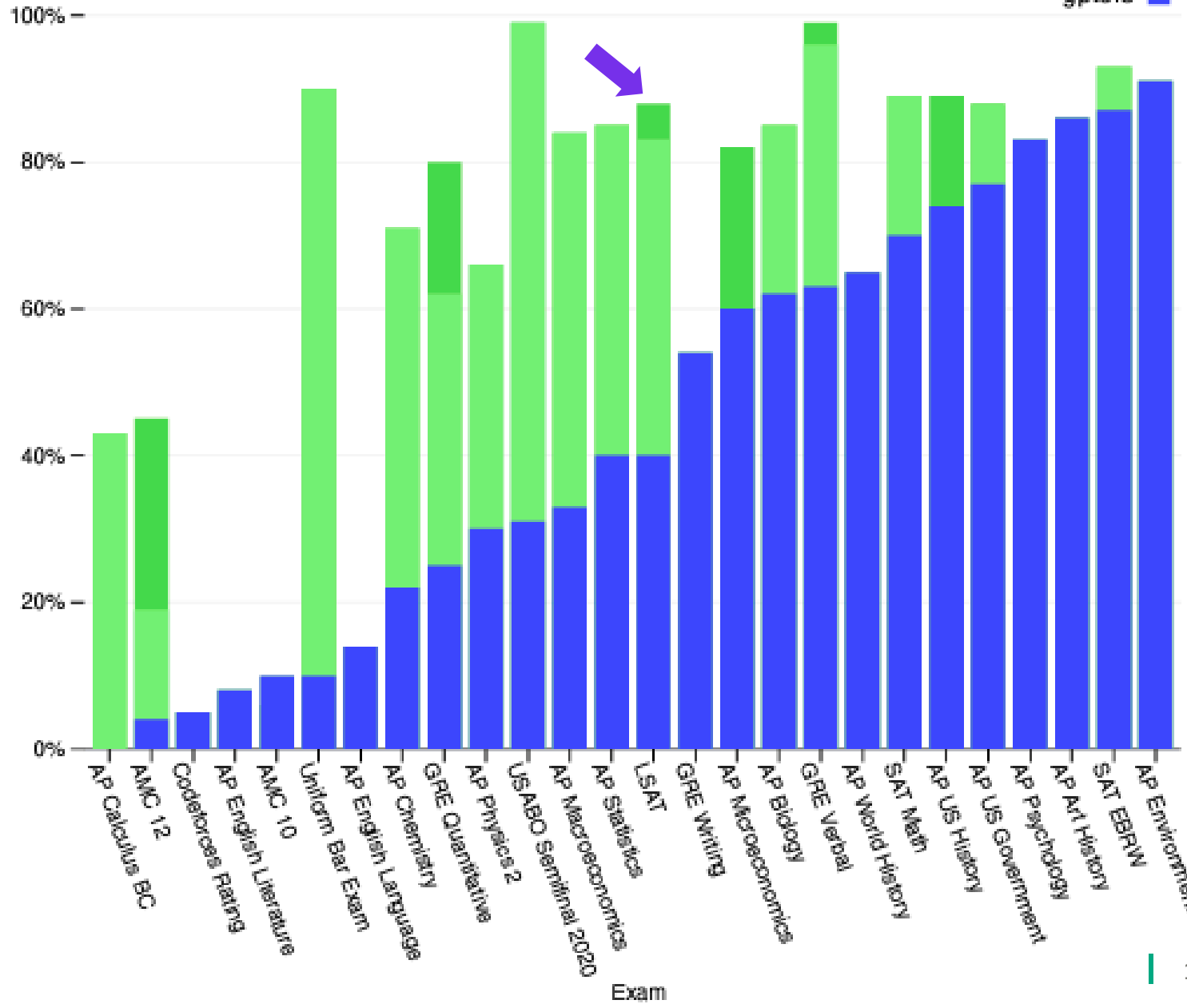
LLM: Emergent Abilities at Scale

Source: [OpenAI, 2023](#)

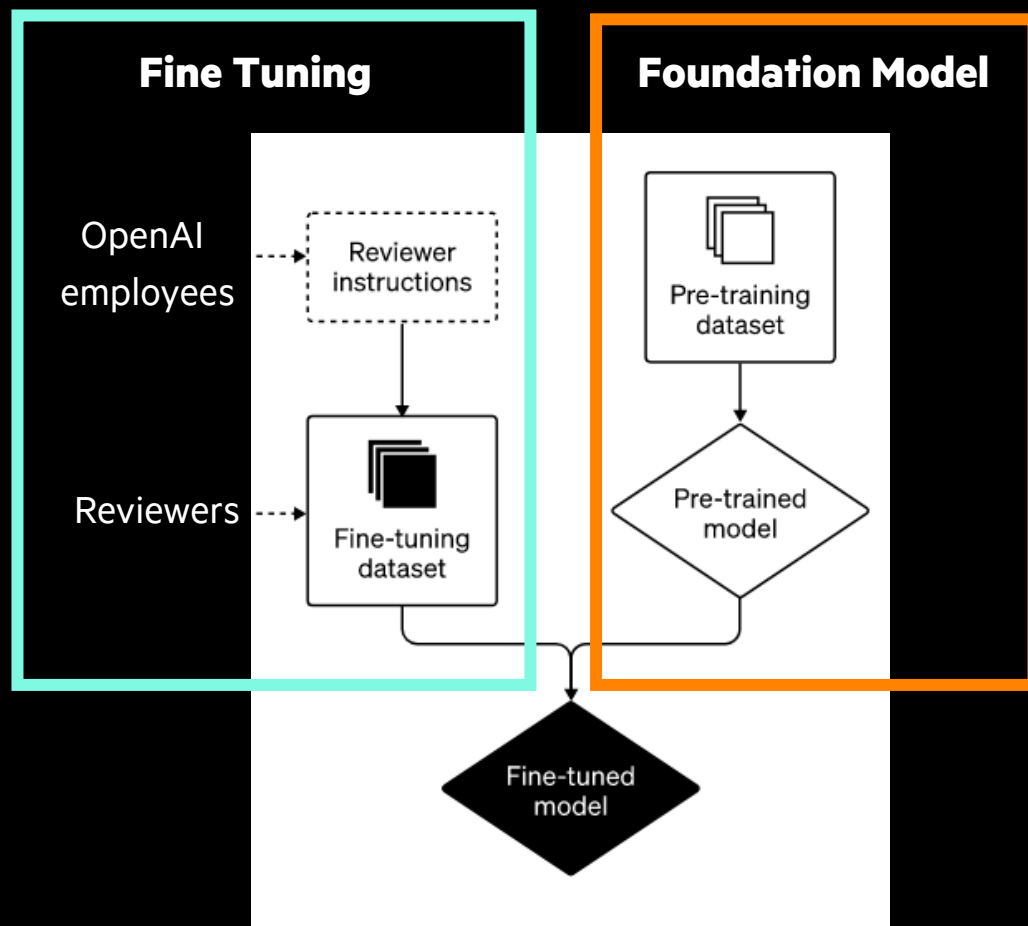
Exam results (ordered by GPT-3.5 performance)

Estimated percentile lower bound (among test takers)

gpt-4
gpt-4 (no vision)
gpt3.5



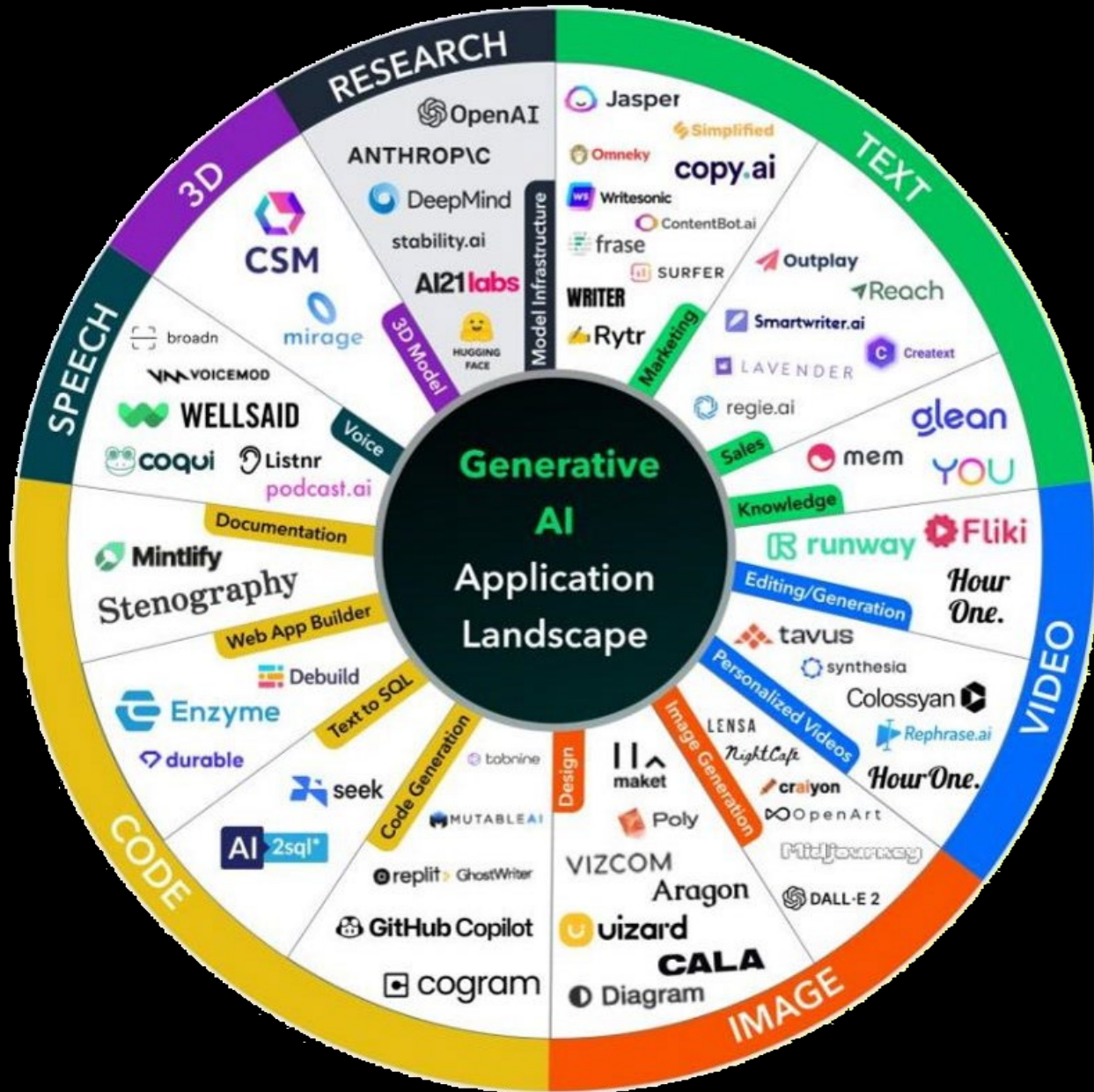
How ChatGPT Was Trained



An initial “**pre-training**” phase in which the model learns to predict the next word in a sentence, informed by its exposure to lots of Internet text (and to a vast array of perspectives). GPT-3 is an example of such pre-trained model.

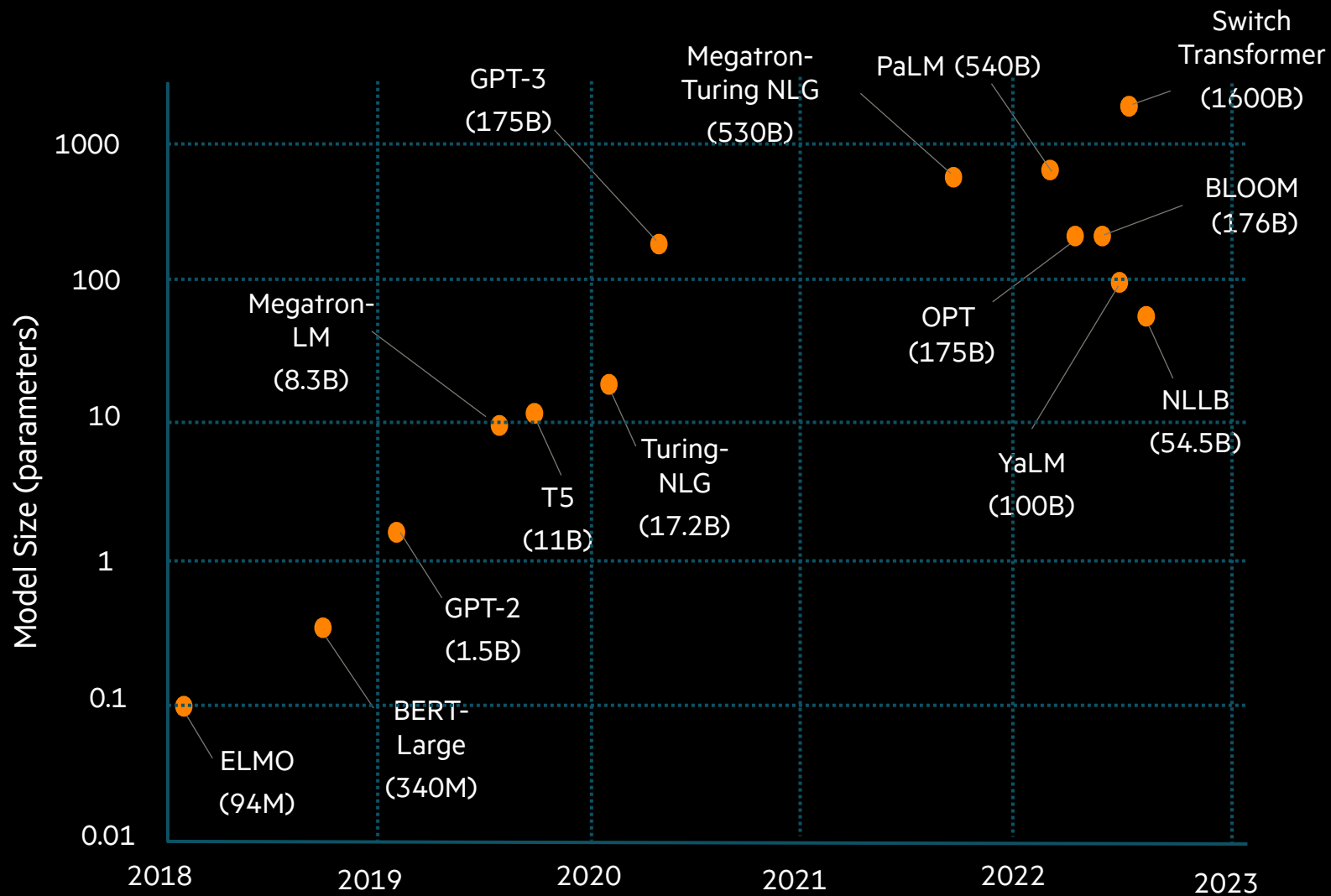
This is followed by a second “**Fine-tuning**” phase in which models were fine-tuned to narrow down system behavior. Fine-tuning process leveraged both supervised learning as well as reinforcement learning in a process called reinforcement learning from human feedback (RLHF).

Generative AI Application Landscape



Large Language Model

- **Large in Size**
 - **Billions of Parameters**
- **Large Compute Resources**
 - **HPC Scale clusters**



Customizing LLMs for vertical use cases

Prompting vs Retrieval vs Fine-tuning



LLM Customization

Number of adopters

Less Customization

Generative AI as a Service - ChatGPT, Google Bard, Amazon Bedrock, Existing Services
Consumption model, \$ per inference
Fastest time to market



Moderate Customization

P-tuning and fine tuning of pre-trained model
\$M+ for infrastructure and resources
Weeks to months for development



Extensive Customization

Custom foundation models or extensive finetuning
\$10M+ for infrastructure and resources
6+ months for development



Amount of Customization

Numbers Every LLM Developer Should Know

Source: Waleed Kadous
bit.ly/llm-dev-numbers

🔮 Prompts

40–90% Amount saved by appending “Be Concise” to your prompt

1.3 Average tokens per word

💰 Price

~50 Cost Ratio of GPT-4 to GPT-3.5 Turbo

5 Cost Ratio of generation of text using GPT-3.5-Turbo vs OpenAI embedding

10 Cost Ratio of OpenAI embedding to Self-Hosted embedding

6 Cost Ratio of OpenAI base vs fine tuned model queries

1 Cost Ratio of Self-Hosted base vs fine-tuned model queries

💡 Training and Fine Tuning

~\$1 million Cost to train a 13 billion parameter model on 1.4 trillion tokens

<0.001 Cost ratio of fine tuning vs training from scratch

🧠 GPU Memory

16GB V100 GRAM capacity
24GB A10G GRAM capacity
40/80GB A100 GRAM capacity

2x number of parameters Typical GPU memory requirements of an LLM for serving

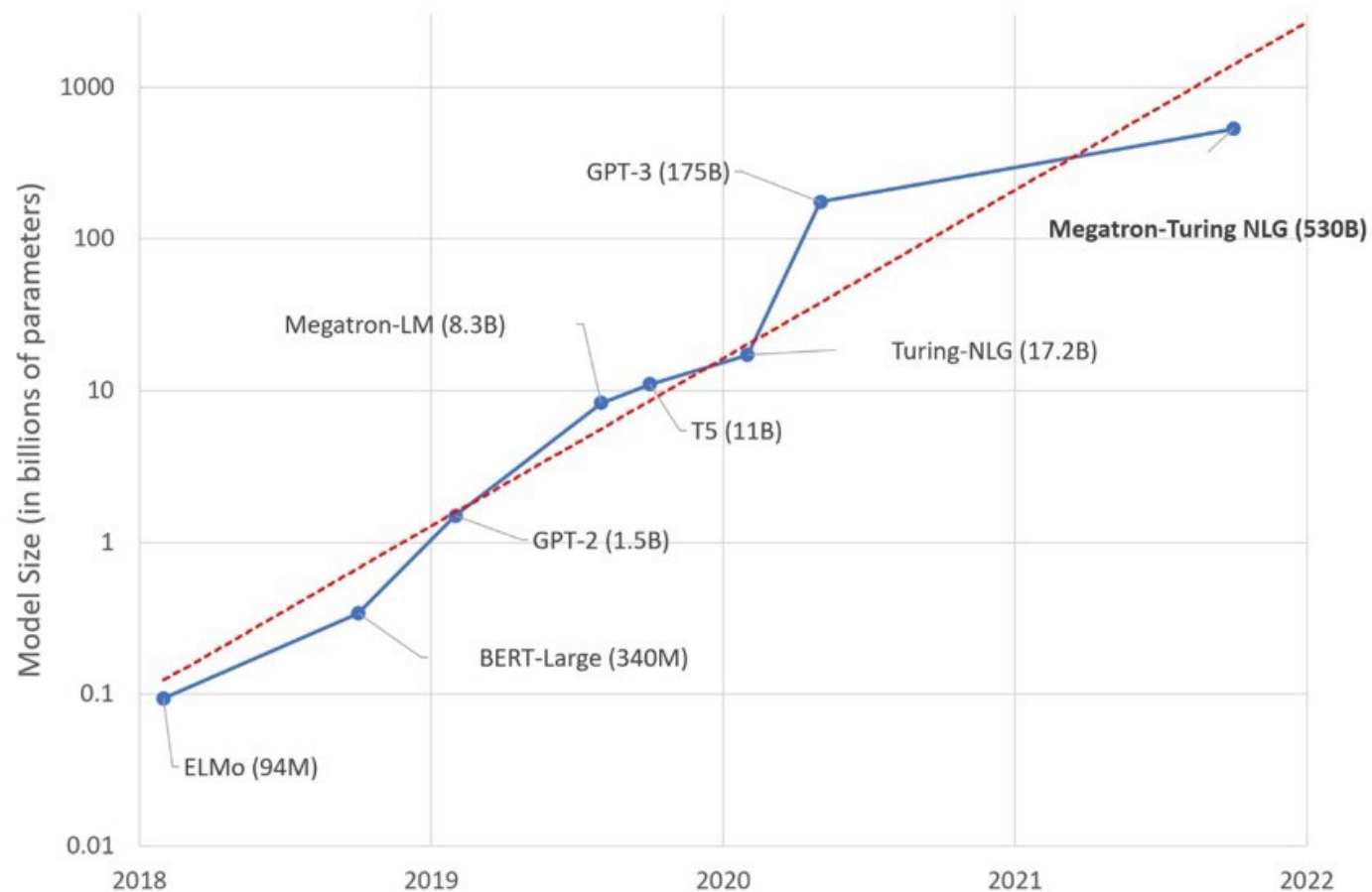
~1GB Typical GPU memory requirements of an embedding model

>10x Throughput improvement from batching LLM requests

1 MB GPU Memory required for 1 token of output with a 13B parameter model

* Check out bit.ly/llm-dev-numbers for how we calculated the numbers

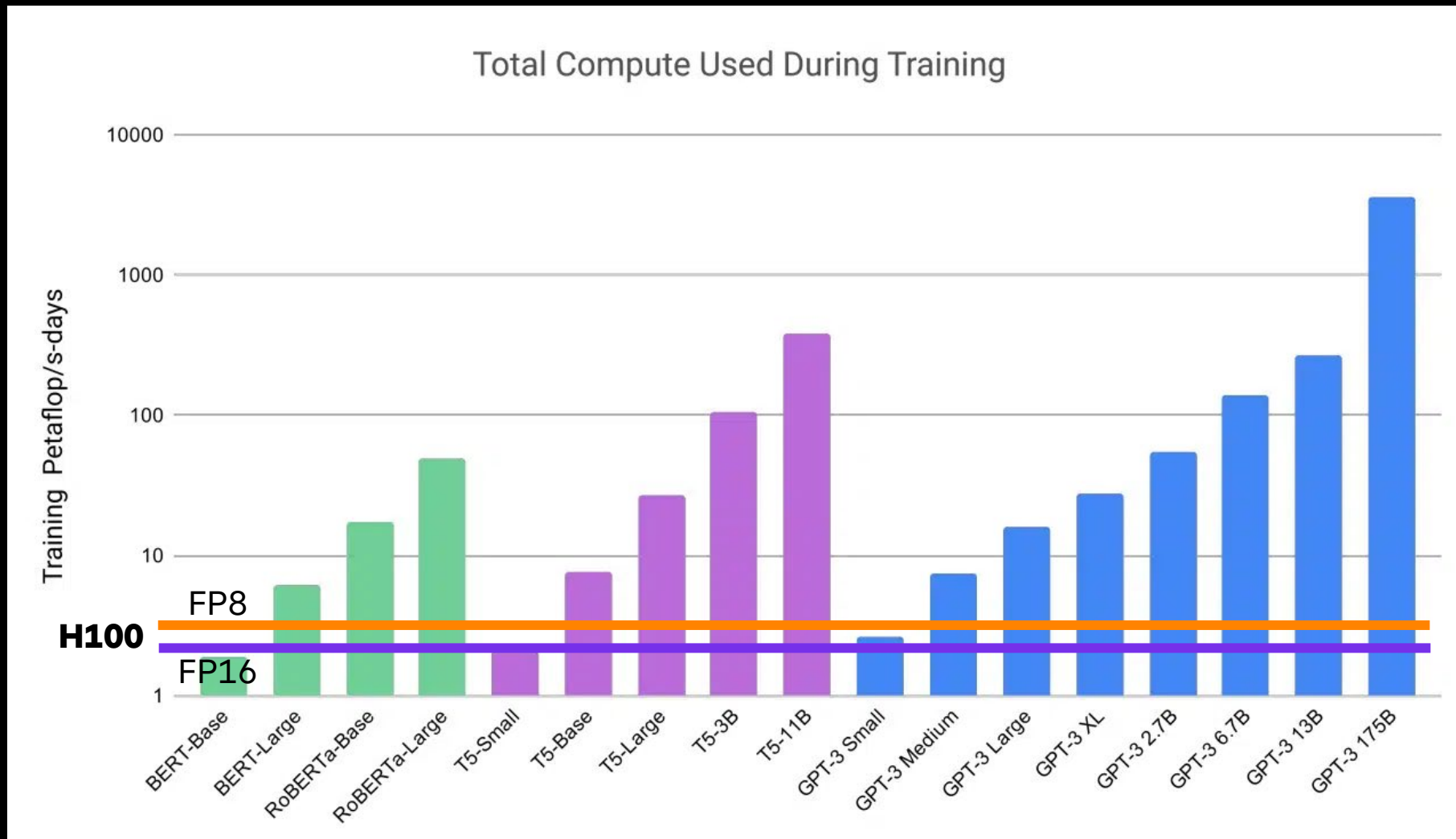
Cost and Power



What's the problem?

- Training GPT-3 **once** costs ~**\$3 million** on the public cloud
- Estimate:
 - 1 month x 1024 A100 x \$4.09/hour (AWS on-demand for p4d.24xlarge)
- It also uses **extreme** amounts of energy:
1200 megawatt hours
(Patterson et. al. 2021)

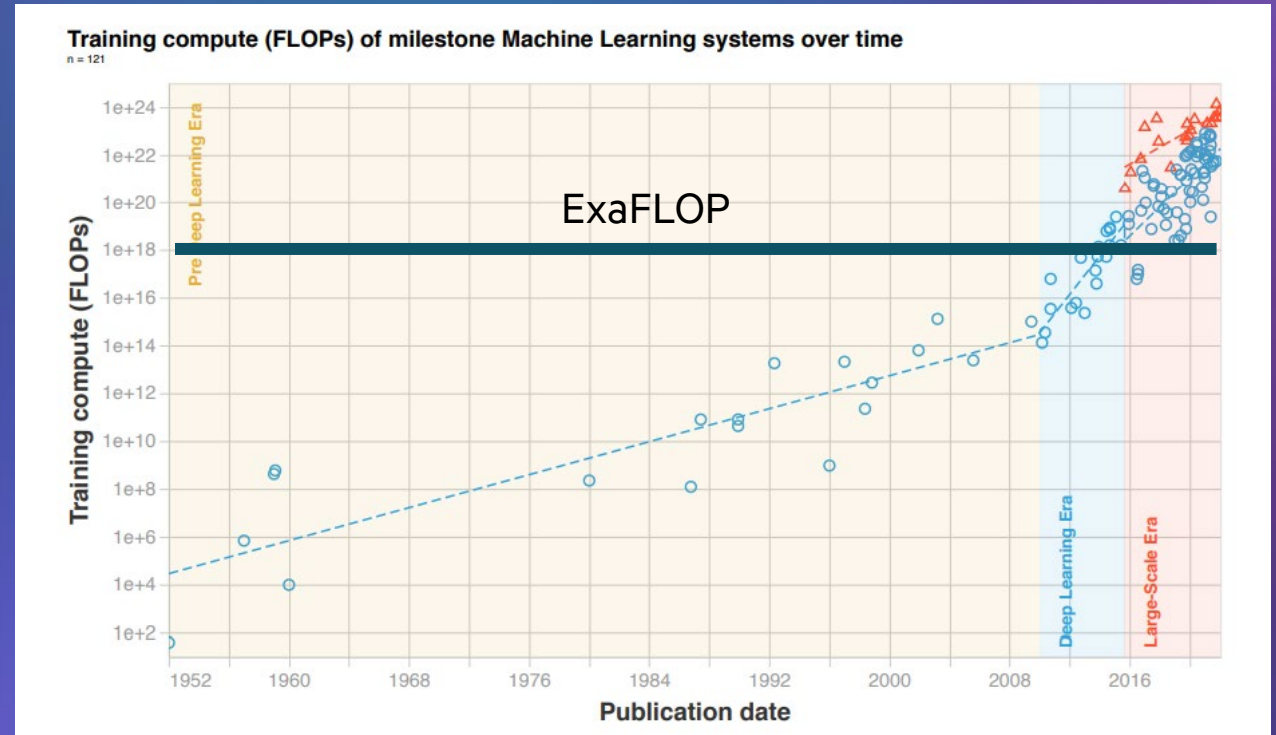
Generative AI has become an HPC problem



Large language models are a supercomputing problem

Development, training, tuning and deployment are very compute-intensive

“The computing requirements for large-scale AI models doubled every 10.7 months from 2016 to 2022.”

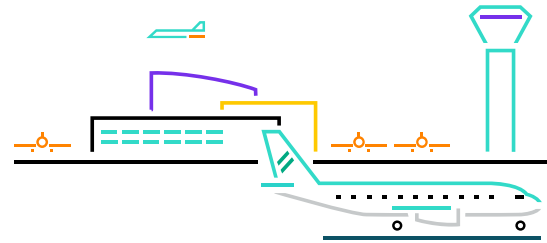


Source: [Compute trends across three eras of machine learning](#), University of Aberdeen, Centre for the Governance of AI, University of St. Andrews, MIT, University of Tübingen, Complutense University of Madrid, March 2022

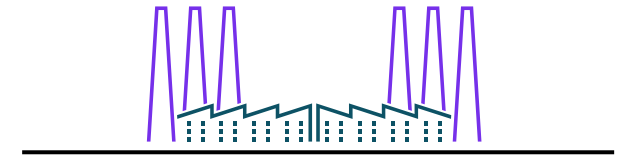
Why be concerned with sustainability in AI?

Training a Large Language Model (LLM) can **emit as much CO2** as

6 coast-to-coast passenger jet flights across the U.S.¹



Power use for AI GPUs purchased in the last year alone is similar to the output of a **nuclear reactor**²



1 Source: Carbon Emissions and Large Neural Network Training (<https://arxiv.org/pdf/2104.10350.pdf>)
2 Source: US Energy Information Administration (<https://www.eia.gov/tools/faqs/faq.php?id=104&t=3>)

Datacenter Impact on Carbon Emissions

| Model name | Number of parameters | Datacenter PUE | Carbon intensity of grid used | Power consumption | CO ₂ eq emissions | CO ₂ eq emissions × PUE |
|------------|----------------------|--------------------------|----------------------------------|-------------------|------------------------------|------------------------------------|
| GPT-3 | 175B | 1.1 | 429 gCO ₂ eq/kWh | 1,287 MWh | <i>502 tonnes</i> | 552 tonnes |
| Gopher | 280B | 1.08 | 330 gCO ₂ eq/kWh | <i>1,066 MWh</i> | <i>352 tonnes</i> | 380 tonnes |
| OPT | 175B | <i>1.09</i> ² | <i>231 gCO₂eq/kWh</i> | <i>324 MWh</i> | 70 tonnes | <i>76.3 tonnes</i> ³ |
| BLOOM | 176B | 1.2 | 57 gCO ₂ eq/kWh | 433 MWh | 25 tonnes | 30 tonnes |

Table 4: Comparison of carbon emissions between BLOOM and similar LLMs. Numbers in *italics* have been inferred based on data provided in the papers describing the models.

HPE delivers the world's first Exascale Supercomputer for US DOE*




- 74 HPE Cray EX cabinets
- 9,408 AMD EPYC CPUs, 37,632 AMD GPUs
- HPE Slingshot 11 interconnect
- 700 petabytes of storage capacity, peak write speeds of 5 terabytes per second using Cray ClusterStor Storage System

TOP500

#1

ORNL's Frontier supercomputer is #1 on the TOP500.

1.1 exaflops of performance on the May 2022 Top500 list.




GREEN500

#1

ORNL's Frontier supercomputer is #1 on the GREEN500.

52.23 gigaflops/watt power efficiency.




HPL-AI

#1

ORNL's Frontier supercomputer is #1 on the HPL-AI list.

6.88 exaflops on the HPL-AI benchmark.



* Source: May 30, 2022, Top500 release

HPE Cray Supercomputers—Customer Choice



ANL “Aurora”

- >1.5 EF Peak performance
- **Intel** Xeon CPU & Xe GPU
- Slingshot interconnect
- Mixed AI and HPC workload



ORNL “Frontier”

- >1.1 EF Peak performance
- **AMD** EPYC CPU & MI250 GPU
- Slingshot interconnect
- Mixed AI & HPC Workload



LANL “Venado”

- Near EF Peak performance
- **NVIDIA** Grace Hopper SoC
- Slingshot interconnect
- Mixed AI & HPC Workload

“Anyone can build a fast CPU. The trick is to build a fast system.” Seymour Cray

<https://www.alcf.anl.gov/aurora>

<https://www.olcf.ornl.gov/frontier/>

<https://discover.lanl.gov/news/0530-venado>

Why HPE for AI sustainability?

Economic and carbon savings

Train LLMs with

20%

fewer compute resources

ASHA is

10x

faster than standard approaches



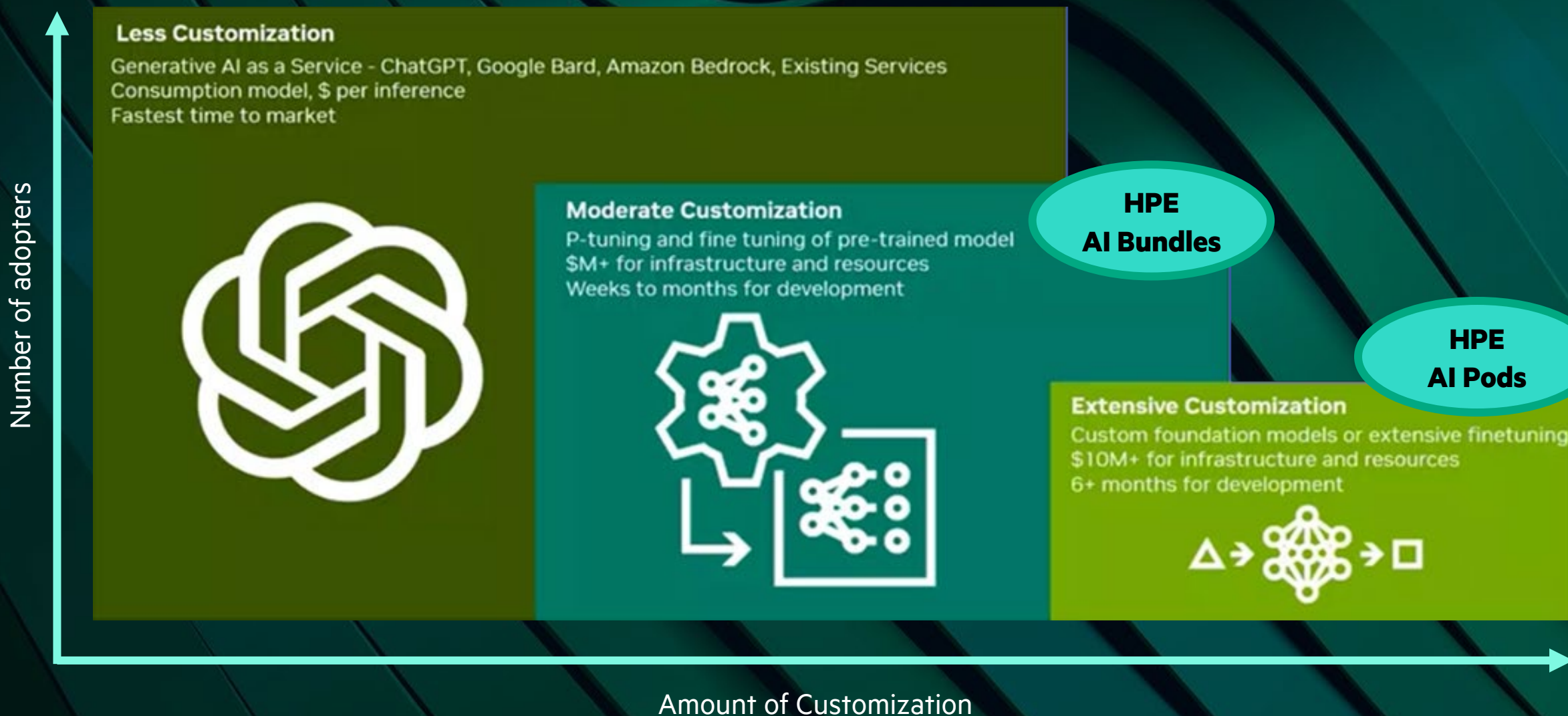
Determined AI

HPE ethical commitment

HPE seeks to use and develop responsible AI with beneficial outcomes for people and businesses and public services guided by ethical principles

HPE applies AI ethical principles through an ethical review process, partner risk assessment and cross BU enablement and engagement

AI Bundle Target Markets



BETTER TOGETHER – NVIDIA H100 4X SXM AND 8X SXM

HPE Cray Supercomputing XD665 vs HPE Cray XD670

HPE Cray XD Supercomputing

XD665

HPC Simulations | AI Inferencing
Model Fine-Tuning | Transfer Learning



2 AMD Genoa CPUs
4 NVIDIA H100 SXM GPUs
4U Height

XD670

Discovery | Parallelization | Speed
Industrial-Scale AI Training



2 Intel Sapphire Rapids CPUs
8 NVIDIA H100 SXM GPUs
5U Height

Accelerate your AI Inference Initiatives

Computer Vision AI at the Edge

Purpose-built for AI at the edge

Loss prevention
Smart spaces

HPE ProLiant DL320 Gen11

Up to Four NVIDIA L4 GPUs



NVIDIA Metropolis ecosystem

Generative Visual AI

Optimized for visual apps

3D animation
Image/video generation

HPE ProLiant DL380a Gen11

Up to Four NVIDIA L40S GPUs



NVIDIA AI Enterprise suite

Natural Language Processing AI

Powering large language models

Speech AI
Fraud detection

HPE ProLiant DL380a Gen11

Up to Four NVIDIA H100 GPUs
supporting NVLink



NVIDIA AI Enterprise suite

HPE GreenLake for HPC/AI

HPC/AI aaS

To define, deliver and integrate the right solution, reliably

T-Shirt Sized or Custom
Purpose-built for HPC/AI Workloads
On-prem or Co-lo



GreenLake for LLM

Industry leading technology developed to solve the world's biggest problems

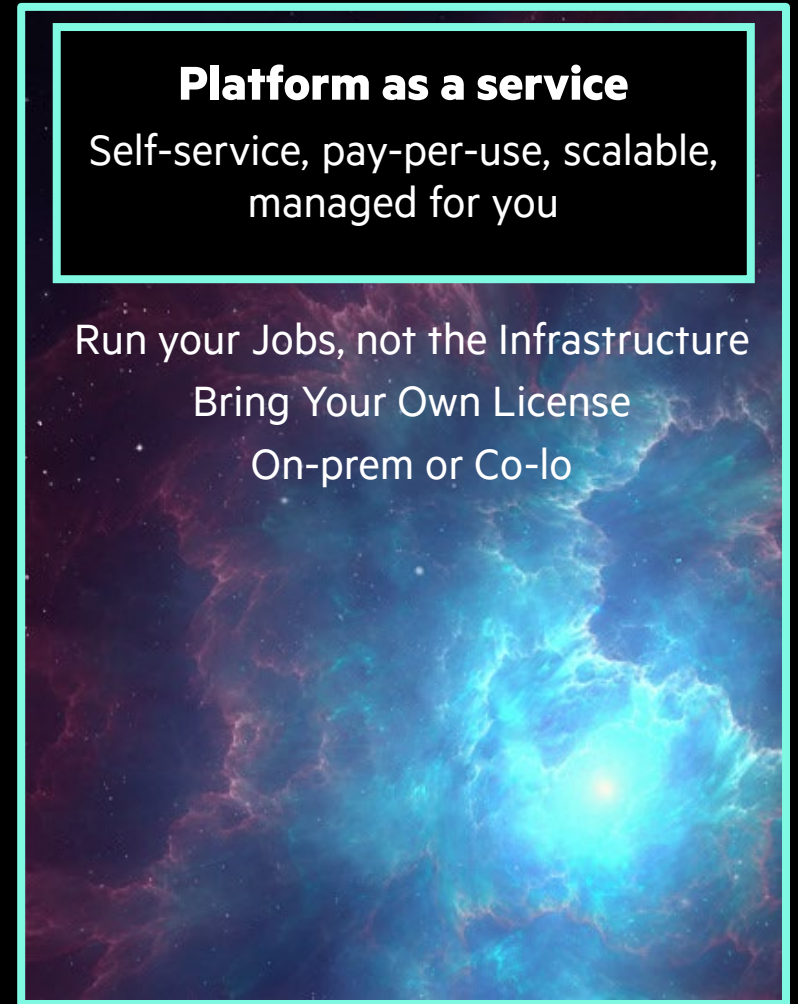
LLM Hardware/Software Stack
Hotel Pricing Model
100% Renewable Energy
>80% Natural Cooling
Reclaimed Heat



Platform as a service

Self-service, pay-per-use, scalable, managed for you

Run your Jobs, not the Infrastructure
Bring Your Own License
On-prem or Co-lo



HPE Leadership Computing in the Age of Insight

Unrivaled expertise
in
HPC / AI

Largest applications
and performance team
in the industry

HPE trusted
supply chain

Hewlett Packard
Labs

Differentiated
IP & systems
capabilities

High Performance
Networking

Photonics

Memory-Driven
Computing

High Performance
Storage & Data
Management

At Scale S/W
& Full Dev
Ecosystem

HPE Cray Programming
Environment

Machine Learning Dev
and Data Management
Environments

Converged workloads

Cluster Mgt

Meet the
customer where
they are

On-premises
Colo
Public Cloud

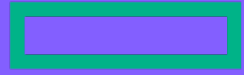
Optimized HPC Cloud
Instances

CAPEX
HPE GreenLake

Thank You!

Steve Heibein, HPE Public Sector AI Chief Technologist





**Hewlett Packard
Enterprise**

Commonwealth Computer Summit

Practical tips for deploying GenAI and LLMs

Steve Heibein, HPE Public Sector AI Chief Technologist

October 16, 2023