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 destruction

Change the anatomy of work
augmenting the capabilities of knowledger workers by automating 60% to 70% of their individual activities

1 Boston Consulting Group, The CEO’s Guide to the Generative AI Revolution, March 2023
AI goes mainstream with ChatGPT

- **ChatGPT**: ~5 days
- **Instagram**: ~75 days
- **Spotify**: ~150 days

*Stylized Path To 1 Million Users* [# of days from launch]

Sources: Google, Subredditstats, Media Reports

*Path is stylized to the 1m milestone*
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.
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

*Source: Gartner*

Today

- **AI Systems**
  - Mostly Classify or Predict
  - Symbols
    - Lifetime Value Score, Intents, Risk Levels, “Turn Left, Image Category, Emotion Type …”
  - Artifacts
    - Video, Audio, Language, Images, Code, Synthetic Data, Design for Real World Objects

Generative

- **AI Systems**
  - Expanded to Generate
  - Symbols
  - Artifacts

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

Large Language Models – Multi-mode

Data

Text
Images
Speech
Structured Data
3D Signals

Training

Foundation Model

Adaption

Question Answering
Sentiment Analysis
Information Extraction
Image Captioning
Object Recognition
Instruction Following
Foundation Models – Contain Foundational Knowledge

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

5 Characteristics:
- Pretrained
- Generalized
- Adaptable
- Large
- Self-supervised

Foundation Models (FM)

Large Language Models (LLM)
ex: ChatGPT, Chinchilla, GPT-3

FM are models trained on broad data (using self-supervision at scale) that can be adapted to a wide range of downstream tasks. 
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

HTTPS://WWW.CATO.ORG/COMMENTARY/AI-COMING-OUR-JOBS-THATS-OKAY
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

<table>
<thead>
<tr>
<th>Model</th>
<th>Launch Date</th>
<th>Training Data</th>
<th>No. of Parameters</th>
<th>Max. Sequence Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-1</td>
<td>June 2018</td>
<td>Common Crawl, BookCorpus</td>
<td>117 million</td>
<td>1024</td>
</tr>
<tr>
<td>GPT-2</td>
<td>February 2019</td>
<td>Common Crawl, BookCorpus, WebText</td>
<td>1.5 billion</td>
<td>2048</td>
</tr>
<tr>
<td>GPT-3</td>
<td>June 2020</td>
<td>Common Crawl, BookCorpus, Wikipedia, Books, Articles, and more</td>
<td>175 billion</td>
<td>4096</td>
</tr>
<tr>
<td>GPT-4</td>
<td>March 2023</td>
<td>Unknown</td>
<td>Estimated to be in trillions</td>
<td>Unknown</td>
</tr>
</tbody>
</table>
LLM: Emergent Abilities at Scale

Source: OpenAI, 2023
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).

Source: OpenAI
Generative AI Application Landscape
### Large Language Model

- **Large in Size**
  - Billions of Parameters

- **Large Compute Resources**
  - HPC Scale clusters

---

The diagram illustrates the growth in model size and compute resources over the years.

- **2018**:
  - ELMO (94M)

- **2019**:
  - BERT-Large (340M)

- **2020**:
  - GPT-2 (1.5B)
  - T5 (11B)

- **2021**:
  - GPT-3 (175B)
  - Megatron-Turing NLG (17.2B)
  - PaLM (540B)

- **2022**:
  - OPT (175B)
  - YaLM (100B)

- **2023**:
  - Switch Transformer (1600B)
  - BLOOM (176B)
  - NLLB (54.5B)

---

- **Opt**: (175B)
- **BLOOM**: (176B)
- **NLLB**: (54.5B)
- **YaLM**: (100B)
- **PaLM**: (540B)
- **2023**:
- **2022**:
- **2021**
- **2020**
- **2019**
- **2018**

---

Model Size (parameters) vs. Year
Customizing LLMs for vertical use cases
Prompting vs Retrieval vs Fine-tuning

Simple
- Prompting
- Few-shot Prompting
- Retrieval + Prompting (LangChain, Vector DB)
- Iterative Refinement (Chain of Thought, Decomposition)

Complex
- Fine-tuning hosted model
- Fine-tuning OSS model
- Training OSS model from scratch
- Create custom model from scratch

Via Willem Pienaar: Emerging Patterns for LLMs in Production
LLM Customization

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 fine-tuning
$10M+ for infrastructure and resources
6+ months for development
# Numbers Every LLM Developer Should Know

Source: Waleed Kadous

<table>
<thead>
<tr>
<th>Prompts</th>
<th>Training and Fine Tuning</th>
</tr>
</thead>
<tbody>
<tr>
<td>40–90%</td>
<td>Cost to train a 13 billion parameter model on 1.4 trillion tokens</td>
</tr>
<tr>
<td>1.3</td>
<td>Cost ratio of fine tuning vs training from scratch</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Price</th>
<th>GPU Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>~50</td>
<td>V100</td>
</tr>
<tr>
<td></td>
<td>16GB</td>
</tr>
<tr>
<td>5</td>
<td>A100</td>
</tr>
<tr>
<td></td>
<td>24GB</td>
</tr>
<tr>
<td>10</td>
<td>A100</td>
</tr>
<tr>
<td></td>
<td>40/80GB</td>
</tr>
<tr>
<td>6</td>
<td>Typical GPU memory requirements of an LLM for serving</td>
</tr>
<tr>
<td>1</td>
<td>&gt;10x</td>
</tr>
<tr>
<td></td>
<td>Typical GPU memory requirements of an embedding model</td>
</tr>
<tr>
<td></td>
<td>~1GB</td>
</tr>
<tr>
<td></td>
<td>Throughput improvement from batching LLM requests</td>
</tr>
<tr>
<td></td>
<td>&gt;10x</td>
</tr>
<tr>
<td></td>
<td>1 MB</td>
</tr>
<tr>
<td></td>
<td>GPU Memory required for 1 token of output with a 13B parameter model</td>
</tr>
</tbody>
</table>

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

Presented by 🌍RAY & 💻anyscale with ❤️ Join the community ray.io or Request a Trial anyscale.com/signup today
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

![Total Compute Used During Training](chart.png)
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
Power use for AI GPUs purchased in the last year alone is similar to the output of a nuclear reactor.

Training a Large Language Model (LLM) can emit as much CO2 as 6 coast-to-coast passenger jet flights across the U.S.

Why be concerned with sustainability in AI?

## Datacenter Impact on Carbon Emissions

<table>
<thead>
<tr>
<th>Model name</th>
<th>Number of parameters</th>
<th>Datacenter PUE</th>
<th>Carbon intensity of grid used</th>
<th>Power consumption</th>
<th>CO₂eq emissions</th>
<th>CO₂eq emissions × PUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-3</td>
<td>175B</td>
<td>1.1</td>
<td>429 gCO₂eq/kWh</td>
<td>1,287 MWh</td>
<td>502 tonnes</td>
<td>552 tonnes</td>
</tr>
<tr>
<td>Gopher</td>
<td>280B</td>
<td>1.08</td>
<td>330 gCO₂eq/kWh</td>
<td>1,066 MWh</td>
<td>352 tonnes</td>
<td>380 tonnes</td>
</tr>
<tr>
<td>OPT</td>
<td>175B</td>
<td>1.09²</td>
<td>231 gCO₂eq/kWh</td>
<td>324 MWh</td>
<td>70 tonnes</td>
<td>76.3 tonnes</td>
</tr>
<tr>
<td>BLOOM</td>
<td>176B</td>
<td>1.2</td>
<td>57 gCO₂eq/kWh</td>
<td>433 MWh</td>
<td>25 tonnes</td>
<td>30 tonnes</td>
</tr>
</tbody>
</table>

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

1.1 exaflops of performance on the May 2022 Top500 list.

52.23 gigaflops/watt power efficiency.

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

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

Number of adopters

Amount of Customization

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

HPE AI Bundles
HPE AI Pods
# BETTER TOGETHER – NVIDIA H100 4X SXM AND 8X SXM

## HPE Cray Supercomputing XD665 vs HPE Cray XD670

<table>
<thead>
<tr>
<th>X665</th>
<th>X670</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPC Simulations</td>
<td>Discovery</td>
</tr>
<tr>
<td>AI Inferencing</td>
<td>Parallelization</td>
</tr>
<tr>
<td>Model Fine-Tuning</td>
<td>Speed</td>
</tr>
<tr>
<td>Transfer Learning</td>
<td>Industrial-Scale AI Training</td>
</tr>
</tbody>
</table>

### X665
- 2 AMD Genoa GPUs
- 4 NVIDIA H100 SXM GPUs
- 4U Height

### X670
- 2 Intel Saphire Rapids CPU
- 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

High Performance Networking
Photonics
Memory-Driven Computing
High Performance Storage & Data Management
Thank You!

Steve Heibein, HPE Public Sector AI Chief Technologist
Commonwealth Computer Summit
Practical tips for deploying GenAI and LLMs

Steve Heibein, HPE Public Sector AI Chief Technologist
October 16, 2023