HYPERION RESEARCH

# Hyperion Research Special Report Overview
## *Large Language Models: Finding Their Place in the HPC Ecosystem*

Interest in LLMs exploded in 2023, driven by their unique capabilities to answer queries, generate concise summaries, and even produce unique works of fiction. Within the HPC community, many organizations are eager to explore the opportunities that LLMs can bring to the advanced computing space by offering tangible gains in computational capability for key workloads. Based on the results of a Hyperion Research study performed in mid-2023, LLMs are considered to be an important emerging technology for both current and planned HPC-related activities. Additional highlights from this study include:

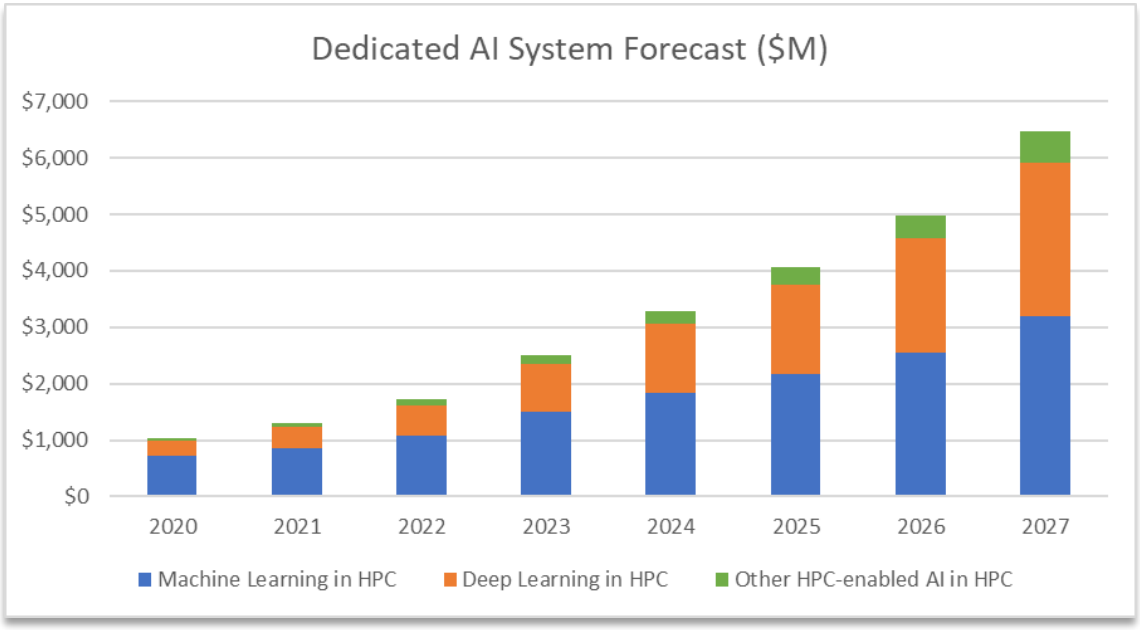
- LLMs are viewed as having widespread benefits to organizations.
- Numerous LLM applications are currently being considered, and many organizations are looking at multiple options.
- Organizations seeking to leverage LLM capabilities do, however, see challenges with the complexity of integrating LLMs into existing HPC-based workloads as well as have concerns with the cost of LLM-specific hardware or software.
- The majority of surveyed organizations are willing to increase their computing budget to support LLM inclusion.
- Survey respondents are looking for a broad range of LLM expertise to support the various stages of LLM development.

*Note:*
*The survey, which was conducted in July/August 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: industry (63%), academic (23%), and government (14%).*

"LLMs show promise in being able to bring a spate of new capabilities to the overall AI space in key areas including content creation, language translation, and even code development," said Bob Sorensen, Senior Vice President of Research. "For their part, HPC end users are interested in LLM's potential to accelerate some of their most demanding computational problems in fields as diverse as finance, bioscience, and advanced manufacturing. Exploration of LLM capability has begun in earnest across a wide swath of HPC sites. Keeping apace of the rapid development in this field will be complicated but necessary for any organization seeking to make the optimal decisions about the how's, when's, and why's of including LLM as part of their computational arsenal."

But LLM's are only one part of the overall AI spectrum, albeit one of the most recent and attention grabbing, that depends heavily on advanced AI-centric computer systems. Hyperion Research forecasts that the purchase of systems dedicated for overall AI use will grow at over 30% a year at least out to 2027, as shown in the figure below.

## Growth in Dedicated HPC-related AI Systems




*Source: Hyperion Research, September 2024*

For more information about the LLM Special Study and purchase options, please contact Jean Sorensen at jeansorensen@hyperionres.com

For more about Hyperion Research's new AI Beacon CIS program and the AI Advisory Committee, go to https://hyperionresearch.com/ai-beacon/.