2025 Multiclient Study Highlights

Key Findings of the 2025 HPC/AI End User Perspectives on HPC AI, ML, and Big Data

Tom Sorensen and Bob Sorensen
February 2025

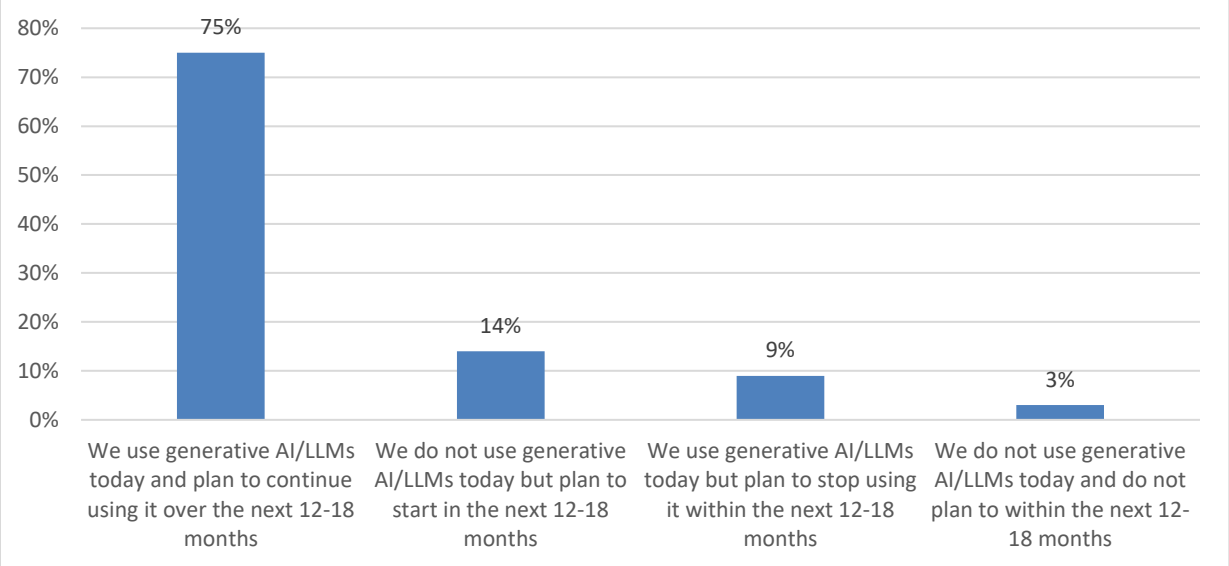
OVERVIEW

Results from the 2025 iteration of Hyperion Research's global site survey of 107 HPC/AI sites representing over 2,500 systems, from industry, government, and academia offer a number of insights about significant developments in the sector:

- The most prominent barriers to furthering AI capabilities are considered to be the quality and scale of available data.
- AI budgets are continuing to increase at unprecedented rates, with industry organizations leading the way.
- The majority of AI workloads are implemented to support or enhance existing traditional HPC workloads among surveyed organizations.
- Most HPC Users Are Now Using AI Capabilities in One Way or Another.

FIGURE 1

Expected and Current use of Generative AI/LLMs for Current or New Workloads



N = 107

Source: Hyperion Research, 2025

DEMOGRAPHICS

Key demographics of the 2025 Hyperion Research HPC/AI End User Perspective study include:

- Inputs were gathered from 107 HPC/AI sites with over 2,500 systems from industry (72%), government (15%), and academia (13%).
- Of the 77 industrial sites surveyed, the most prevalent respondent sectors were EDA/IT/ISV (26%), economics/financial (22%) and bio-sciences (16%). Other sectors represented included CAE, geosciences and digital content and creation (DCC).
- Roughly half the sites surveyed had peak performance of their largest HPC at 25 Pflops or less, while four sites had exascale-class systems.
- Roughly half of all surveyed sites indicated that the price of their largest technical computing/HPC/AI server was less than $3 million, while one in nine reported prices tags for their most powerful system at $150 million or more.

HIGHLIGHTS

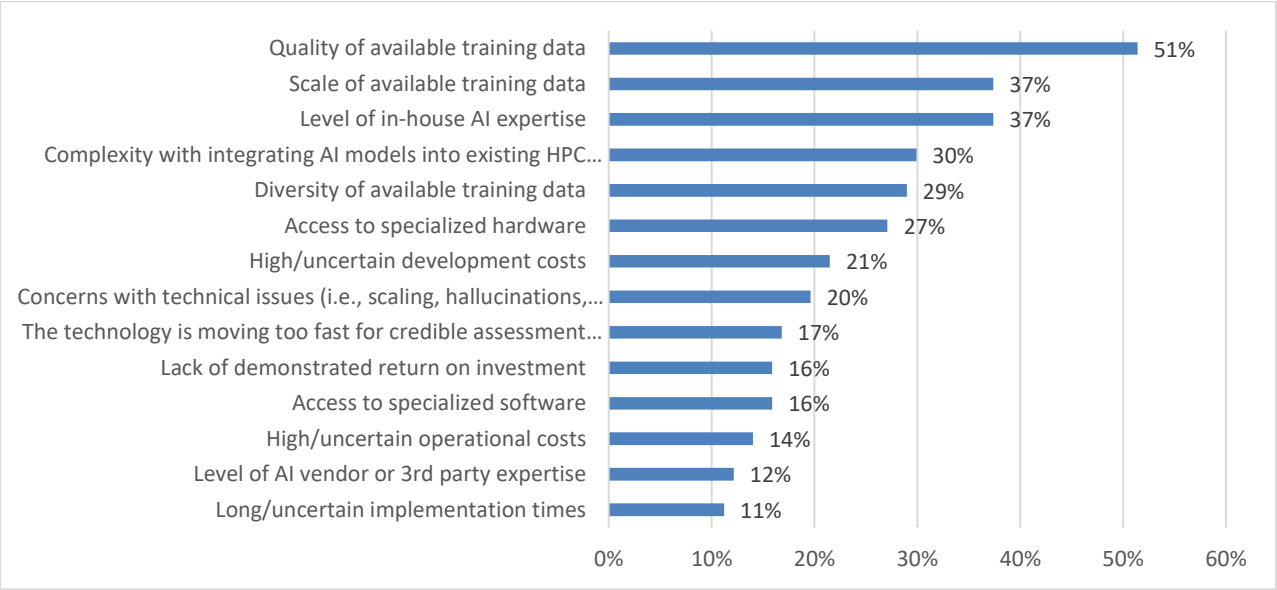
Barriers to Expanding the Use of AI

When asked about barriers to furthering AI capabilities, the first and second most selected answers were quality of available training data and scale of available training data;

- Of the top five perceived barriers, data related concerns were cited three times: quality, scale, and diversity (see Figure 2).
- Also high on the list were level of in-house expertise and the complexity of integrating AI models into existing models.
- As training continues, quality data to ameliorate models is harder to come by; models require more data, better data, and more application-specific data as its potential applications diversify. Data acquisition, management, and synthesis are expected to become considerably more important activities.

FIGURE 2

Major Barriers to AI Development



N = 107

Source: Hyperion Research, 2025

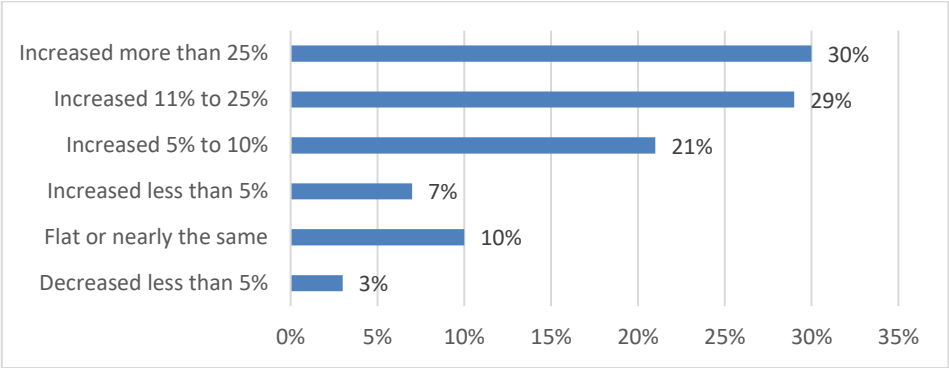
Budgets are Increasing at a Rapid Rate

Budgets are increasing at a rapid rate for AI/ML/DL/LLM integration and adoption, with industry respondents leading the way:

- As shown in Figure 3, 59% of respondents indicated that they increased their AI budgets by more than 11% in the last year, with 30% of respondents indicated an increase of more than 25%.
- For industry respondents: 62% reported an increase in budget of over 11% and 32% reported an increase in AI budget of over 25%. Assets contributing significantly to these budget increases include AI expertise, application appropriate hardware procurement, and cloud usage.

FIGURE 3

AI/ML/DL/LLM Budget Changes Over the Last Year



N = 107

Source: Hyperion Research, 2025

The Majority of AI Workloads Are Implemented to Support or Enhance Traditional HPC Applications

Despite developments of novel applications in the consumer space of production-tier AI workloads, 72% of HPC users are primarily using AI leveraged to support existing, traditional HPC workloads.

- AI/ML/DL/LLM capabilities are viewed as powerful accelerants to introduce efficiency to existing workflows.

Most HPC Users Are Now Using AI Capabilities in One Way or Another

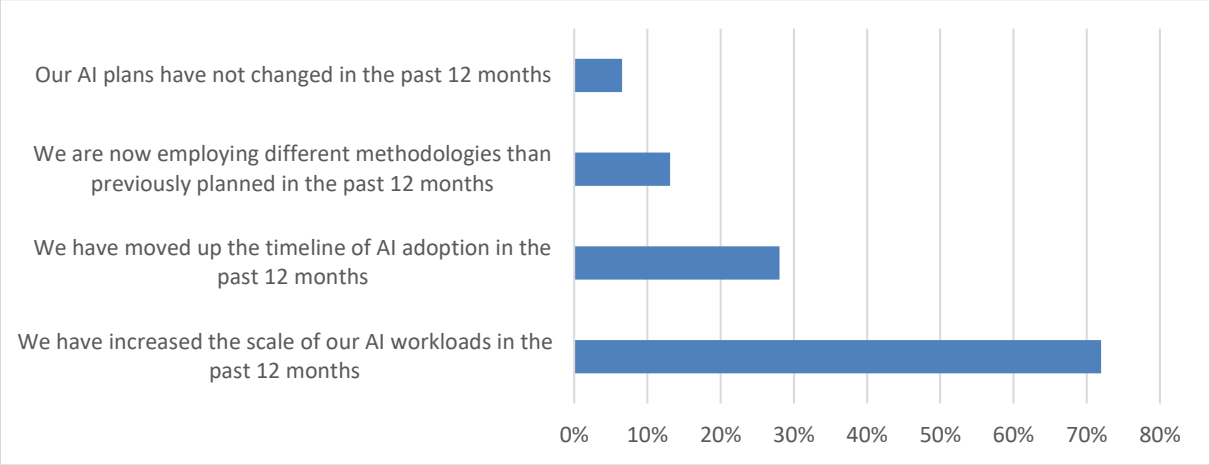
96% reported that their organization is currently or planning to run in the near future AI/ML/DL/LLM workloads. This marks an important stage in the maturation of AI integration into the HPC user community as the end of mass adoption and the beginning of a stage that will be marked by scaling, tuning, and provisioning.

Furthermore, as seen in Figure 1 above, these respondents are responding quickly to shifting needs and technological development, with over 70% reporting an increased scale of workloads in the last 12 months and almost 30% reporting a moving up in the timeline within the same time period (Figure 4). While this already represents a sea change over the past several years, it does not represent the level of activity or integration within each representative organizations, which still have much room to grow and further adopt.

The increased scale represented in Figure 4 will be reflected in a marked increase in AI expertise, application specific hardware, and software license procurement. Furthermore, the protracted timeline of top-of-the-line hardware like GPUs is expected to translate to a continued rise in public cloud usage among HPC users.

FIGURE 4

Changes in Scale of AI Workload Adoption



N = 107

Source: Hyperion Research, 2025

CONCLUSION

Hyperion Research closely tracks the ongoing intersection and integration of AI methodologies into advanced computing and HPC. Over the last several years, adoption has exploded into a state of use by almost all HPC sites. These efforts are usually focused on supporting existing HPC workloads that commonly leverage GPU hardware and consist of a training-heavy mix of both training and inference.

For most users, access to large amounts of quality data and necessary expertise are the greatest perceived barriers to furthering their AI efforts. Most representative groups preferred a mix of on-premises and public cloud resources to meet their compute needs. When asked to break down the ratio of these two sources, users most commonly reported more than half of their workload demands being met on-premises, with a minority of workloads being handled in the cloud.

In response to the expected positive impact on traditional research, science, and engineering workloads, user organizations have considerably ramped up their budgets and integration efforts related to AI/ML/DL. While this saturation of adoptive organizations is nearing entirety, it does not represent the extent to which integration within an organization is complete, efficient, and well-maintained, areas in which there is still considerable work to be done. Currently, excitement and buy-in are still at very high levels, and models continue to become more well-suited for their specific application requirements.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology, and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). Hyperion Research provides thought leadership and practical guidance for users, vendors, and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com