AI-at-HYPERION: A Special Analysis

AI Investments for HPC and Advanced Computing Continue to Expand and Begin to Show ROI Results

Tom Sorensen, Bob Sorensen, and Earl Joseph
November 2025

EXECUTIVE SUMMARY

The purpose of this study was to gain a better understanding of the return on investments made by AI/HPC across critical science and engineering verticals. Key goals included creating a picture of investment expectations, snapshots of current integration projects in progress, budget allocations, provisioning methods, future plans, and challenges involving achieving desired goals in these areas.

The report reveals a strategic shift among high-performance computing (HPC) and scientific users in their approach to generative AI integration. While continuing a pace of relatively rapid expansion, organizations are now focused on refining existing implementations and responding more deliberately to success metrics, whether by scaling, maintaining, or redirecting their efforts. AI is primarily used today in supportive roles such as data preprocessing and workflow management, reflecting the sector's need for precision and reproducibility. Despite costs often exceeding expectations and monetary returns projected to be years away, users remain committed to AI as a long-term innovation tool, prioritizing research advancement over immediate financial gain.

Some highlights from the study:

- 75.8% of the sites felt that their AI projects met or exceeded expectations.
- 74.9% of respondents indicated plans to moderately or significantly expand generative AI to support HPC workloads, roughly 28% of those characterize this expansion as significant.
- Less than 3% expect to contract their use of AI, none of which would characterize that contraction as significant.
- Roughly 40% of respondents are already using agentic AI models.

However, technical challenges continue to bring hesitation when it comes to broad adoption. Hallucinations, lack of explainability, and integration complexity are persistent concerns, especially in scientific and engineering contexts where accuracy is paramount. These issues, combined with the high computing demands of generative models, have led to a projected slowdown in the high rate of growth. This signals a transition from reactive adoption to more measured, application-specific onboarding. To sustain momentum, future progress will depend on innovations that reduce technical friction and elevate AI from peripheral support to more central roles in HPC workflows.