Special Report

Large Language Models: Finding Their Place in the HPC Ecosystem

Bob Sorensen and Tom Sorensen
September 2023

EXECUTIVE SUMMARY

The purpose of this study was to gain a better understanding of the capabilities of large language models (LLMs), an emerging class of AI algorithms, to benefit the overall HPC community. Key goals of this effort included describing the base of current and planned HPC-related activity that could incorporate LLMs, assessing the level of ongoing LLM activity within end user organizations, characterizing the interest in general-purpose LLM applications, exploring the prospects for LLM integration into traditional HPC algorithms, and highlighting the key challenges with integrating LLM capability into HPC-based workloads.

The survey, which was conducted in July 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%), representing a range of industry verticals led by computers and related electronics but that also included the financial sector, bioscience, advanced manufacturing, and geosciences. Organizations represented in this study consisted primarily of HPC sites in both research and production environments with some mixed HPC/enterprise sites.

Study highlights include:

- LLMs are considered to be an important emerging asset for both current and planned HPC-related activity.
- Respondent organizations are looking at a broad set of LLM-related end uses.
- There are numerous LLM applications currently being considered, and many individual organizations are looking at multiple options.
- There are some significant challenges ahead for organizations seeking to leverage LLM capability.
- The majority of surveyed organizations were willing to increase their computing budget to support LLM inclusion.

Key Findings

Key Finding #1: LLMs are considered to be an important emerging asset for both current and planned HPC-related activity.

When asked about the overall importance of LLMs to current or planned HPC-related activity, 78% of survey respondents indicated that LLMs are currently seen as being either very or somewhat important within their organization, rising to 90% within the next 12-18 months. Only 7% of survey respondents saw LLMs as either somewhat unimportant or very unimportant today, dropping to less than 1% in the next 12-18 months.

Key Finding #2: There is a wide range of ongoing LLM-related activity underway within the surveyed organizations.

The two most prominent activities identified were exploring the range of potential performance enhancements by integrating LLMs into existing HPC-based workloads and exploring in-house requirements for integrating LLMs into HPC-based workloads. However, a wide range of additional and likely parallel efforts were also underway such as LLM-relevant hardware and software procurements, holding discussions with LLM suppliers, and standing up LLM-related pilot programs.

There is a wide range of ongoing LLM-related activity underway within the surveyed organizations.

Key Finding #3: Respondent organizations are looking at a broad set of LLM-related end uses.

The most cited efforts were model, software, and data set development followed by mission critical scientific research, engineering, or production. A number of diverse LLM-related activities, such as writing and communications, data management and clean-up, and customer-facing applications, were also cited, illustrating the wide range of range of perceived use cases for LLM within the overall HPC community.

Key Finding #4: There are numerous LLM applications currently being considered, and many individual organizations are looking at multiple options.

The most selected general-purpose LLM application options were natural language processing (NLP) and questions answering systems. Additional options chosen by many included chatbots and virtual assistants, code generation and debugging, and content generation. Ultimately, the average respondent selected 3.8 different LLM applications currently under consideration.

Key Finding #5: Many different HPC-related scientific and engineer algorithms were seen as viable for LLM enhancements.

Data science/big data analysis was selected by 80% of the survey respondents as a top three choice and by 53% as the single most promising option for LLM enhancement. However, a number of other potential HPC workloads were seen as a top three option for LLM enhancement including Monte Carlo methods, dense linear algebra, and partial differential equations.

Key Finding #6: LLMs are viewed as having widespread benefit to surveyed organizations.

Some of the high-level organizational goals envisioned by the inclusion of LLMs into existing or planned HPC-related activities included enabling new HPC capabilities, driving faster times to solution on key HPC solutions, and enabling greater efficiency on HPC-based workloads.

Key Finding #7: **There are some significant challenges ahead for organizations seeking to leverage LLM capability.**

Top challenges cited were complexity with integrating LLM into existing HPC-based workloads, high/uncertain development costs, and concerns with the cost of LLM-specific hardware or software. Challenges considered less concerning, but still identified, included the notion that the technology is moving too fast for credible assessment of value, general confusion, uncertainty with LLM vendor selection, and uncertainty with demonstrated computational performance improvements.

Top challenges cited were complexity with integrating LLM into existing HPC-based workloads, high/uncertain development costs, and concerns with the cost of LLM-specific hardware or software.

Key Finding #8: **The majority of surveyed organizations were willing to increase their computing budget to support LLM inclusion.**

The most selected option was for surveyed organizations to increase their overall IT budget by 10% to less than 15% to support LLM inclusion. However, one-fifth of respondents indicated that their organization could commit at least 15% or more, while one in ten envisioned additional budget commitment exceeding 30% of current expenditures. Only 8% reported no plans to commit additional funds.

Key Finding #9: **Open source is currently the most preferred option for accessing LLM software.**

Overall, open source LLMs were the predominant source of LLM capability, followed to a lesser extent by commercial and in-house developed LLMs. The most popular open source providers of LLM software were Google, Microsoft, and Nvidia, but 23 different open source LLM providers were selected by respondents.

Key Finding #10: **Survey respondents looked to a wide range of LLM expertise to support the various stages of LLM development spanning foundation model construction, fine-tuning procedures, LLM integration into existing workloads, and supporting inference operations.**

- Respondents' organizations are looking to external LLM software developers at the foundational model stage, but with a shifting emphasis to in-house software development in the later stages of LLM roll out and use.
- In-house computational hardware needed to implement one or more stages of LLM development was the most prevalent hardware platform choice, with the sole exception of foundational model development, which showed a slight preference for external LLM hardware providers. External, most likely cloud based, LLM hardware was seen as a lesser imperative.
- There was no significant difference in the type of in-house hardware needed to support LLM development and use, generally being split between special purpose LLM hardware and general purpose counterparts, with the exception of model fine-tuning operations, which tended towards the use of special purpose LLM hardware.

Summary and Next Steps

The LLM user behaviors and expectations from this survey are reflective of broader market sentiments regarding generative AI: high confidence, eagerness to adopt, and a willingness to invest resources into the technology. This enthusiasm is also demonstrated by a healthy open-source ecosystem, pilot programs, and wide exploratory projects. For many users, the benefits of generative AI or LLM

integration is a foregone conclusion, the only obstacles being the cost and complexity of achieving this potential.

Those seeking to use LLM technology for natural language processing, chatbots, and virtual assistants are the most readily suited for benefits to their workloads, but there is also tremendous interest in other areas like code generation, fraud detection, and advertising. With LLM technology already offering improved capabilities in certain areas, many HPC users are rapidly exploring options with plans to adopt. In the fast changing and diversifying subject of LLM-enabled HPC, organizations that maintain AI literacy, manage realistic expectations, and make even-handed budget decisions will be best prepared to ride this wave without overextending.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

IN THIS STUDY

The intent of this study was to gain a better understanding of the capabilities of large language models (LLMs), an emerging class of AI algorithms, to benefit the overall HPC community. For the purpose of this particular study, HPC was broadly defined as any system that addressed traditional modeling and simulation workloads, data science/big data applications as well as the broad class of existing and emerging AI-based end uses that draw on either machine learning or deep learning paradigms. Such HPCs could be installed in a traditional on-premises data center, accessed through some form of cloud access model, either public or private, or some hybrid combination.

Research Approach

This study is based on an independent survey of organizations currently involved in LLMs or planning to employ LLMs within the next 12 to 18 months, to support current or planned HPC-based workloads.

Key study goals included:

- Describing the base of current and planned HPC-related activities that incorporate LLMs
- Assessing the current level of ongoing LLM activity spanning passive monitoring of LLM progress to running production LLM-enabled workloads
- Characterizing the interest in general-purpose LLM applications such as natural language processing, content generation, sentiment analysis, and question answering systems
- Exploring the prospects of LLM for integration into traditional HPC algorithms including differential equations, dense linear algebra, spectral methods, Monte Carlo methods, and data science/big data analysis
- Highlighting the key challenges with integrating LLM capability into existing HPC-based workloads
- Characterizing current and anticipated LLM-related budget commitments
- Identifying key sources of LLM software

Survey Background

The survey, which was conducted in July 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.

Ultimately 190 invitations were extended to potential participants in order to successfully gather 100 complete responses, a 53% success rate.

- Although not a definitive indicator as to the overall persuasiveness of LLMs across a wide range of sectors overall, the survey success rate supports the notion that there is wide interest in the technology.

Demographics of the respondents:

- A mix of major sectors: industry (66%), academic (23%), and government (14%)
- A range of industry verticals led by computers and related electronics but that also included the financial sector, bioscience, advanced manufacturing, and geosciences
- A regional combination of Norh America (21%), Europe (29%) and Asia/Pacific/ROW (12%)

- Job responsibilities that included scientific researchers, subject matter experts, HPC or IT staff, C-suite denizens, programmers, and data center staff
- A concentration on HPC sites in both research and production environments, but with some mixed HPC enterprise sites

Note that a more detailed description of respondent demographics can be found in the Appendix.

SURVEY RESULTS

The survey sought to consider a span of topics related to the use of LLM to address HPC-related workloads, summarized in this section, including:

- Scoping the overall interest and charactering the level or activity of LLMs within the surveyed organizations
- Exploring general-purpose LLM applications, such as natural language processing and question answering systems, as well as LLM applicability for inclusion in traditional HPC applications such as big data analysts or spectral methods
- Assessing LLM perspectives from an organization viewpoint, including organizational goals for LLMs, challenges with introducing LLMs onto existing workloads, and potential LLM budget commitments
- Delving into LLM software considerations including preferences for open source vs commercial LLM offerings as well as arraying the preferences for in-house, open source, and commercial LLM software suppliers

Scoping Overall LLM Interest and Activity

The survey results show strong indications that LLMs are considered to be an important emerging asset for both current and planned HPC-related activity.

- As seen below in Table 1, when asked about the overall importance of LLMs to current or planned HPC-related activity, 78% of survey respondents indicated that LLMs are currently seen as being either very or somewhat important, rising to 90% within the next 12-18 months.
- Conversely, only 7% of survey respondents saw LLMs as either somewhat unimportant or very unimportant today, dropping to less than 1% in the next 12-18 months.

Table 1

Importance of LLMs to HPC-Related Activity

	Currently	Next 12-18 months
Especially important	34%	48%
Somewhat important	44%	42%
Neither important nor unimportant	13%	8%
Somewhat unimportant	4%	1%

Table 1

Importance of LLMs to HPC-Related Activity

	Currently	Next 12-18 months
Very unimportant	3%	0%
Don't know/Not sure	2%	1%

N = 100

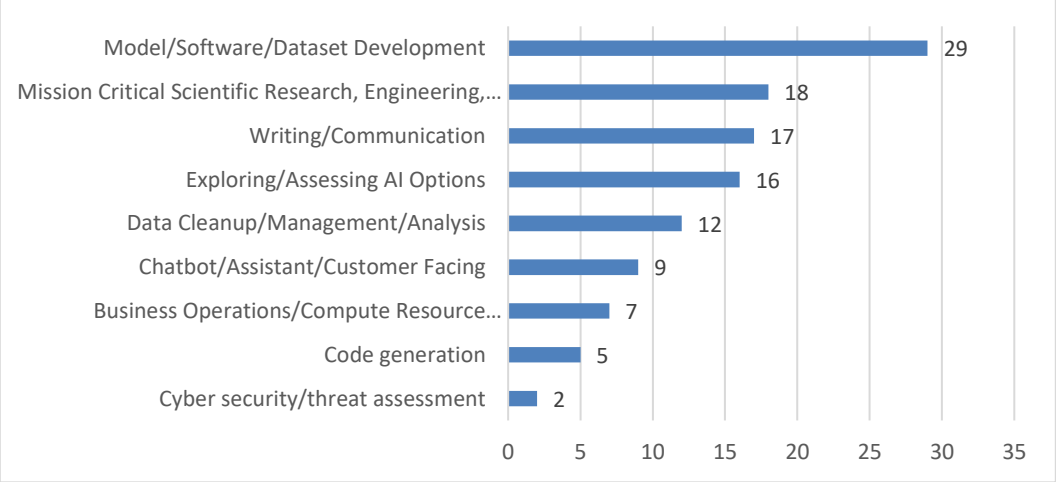
Source: Hyperion Research, 2023

Figure 1, seen below, lists the general classes of ongoing LLM-related activity in the surveyed organizations. The most cited activities centered on model, software, and data set development followed by mission critical scientific research, engineering, or production. A number of diverse LLM-related activities, such as writing and communications, data management and clean-up and customer facing applications, were also cited, illustrating the wide range of perceived use cases for HPC-related LLMs in the overall HPC community.

- For this particular question, respondents could select more than one answer. With a total response base of 115 from 99 respondents, the survey results reveal that most respondents are likely focusing their LLM work within a single HPC-related activity.

FIGURE 1

Current HPC-related LLM Activity Area



N = 99

Respondents could select more than one answer.

Source: Hyperion Research, 2023

Table 2, below, summarizes both the current and planned state of LLM-rated activity across the respondents' organizations. The two most selected options were exploring the range of potential performance enhancements by integrating LLMs into existing HPC-based workloads (58%) and exploring in-house requirements for integrating LLMs into HPC-based workloads (55%). However, a wide range of additional and likely parallel efforts were also under way including LLM-relevant hardware and software procurements, holding discussions with LLM suppliers, and standing up LLM-related pilot programs.

- Only about one in four respondents indicated that their organization was currently passively monitoring LLM-technology development.

Perhaps more important is the level of LLM activity acceleration anticipated within the next 12-18 months. For example, the number of organizations that plan to have production level LLMs running rises from 22% currently to 50% in the next 12-18 months, more than doubling organizational LLM participation. Likewise, the number of organizations that plan to stand up a fully funded LLM research effort will rise from 17% to 27% within the next 12-18 months, an almost 60% increase.

- Conversely, the number of organizations involved in passively monitoring LLM technology developments will decline from 27% today to 14% in the next 12-18 months, representing less than one in seven organizations in this study.

The number of organizations that plan to have production level LLMs running rises from 22% currently to 50% in the next 12-18 months.

Table 2

Summary of Current and Planned LLM-related Activities

	Currently	Next 12-18 months	Change Over Time
Exploring the range of potential performance enhancements by integrating LLMs into existing HPC-based workloads	58%	48%	-10%
Exploring in-house requirements for integrating LLMs into HPC-based workloads	55%	51%	-4%
Testing/assessing LLM-integrated workload performance	34%	45%	11%
Procuring access to necessary LLM software	31%	31%	0%
Reaching out to LLM hardware and software suppliers for information	30%	35%	5%
Passively monitoring LLM technology developments	27%	14%	-13%
Procuring access to necessary LLM hardware	26%	28%	2%
Standing up limited LLM-integrated pilot programs	26%	36%	10%

Table 2

Summary of Current and Planned LLM-related Activities

	Currently	Next 12-18 months	Change Over Time
Porting LLM capability into existing workloads	25%	34%	9%
Running production level LLM-enabled workloads	22%	50%	28%
Standing up a fully funded LLM research efforts	17%	27%	10%
No current activity	1%	0%	-1%
Other	1%	0%	-1%

N = 100

Respondents could select multiple options.

Source: Hyperion Research, 2023

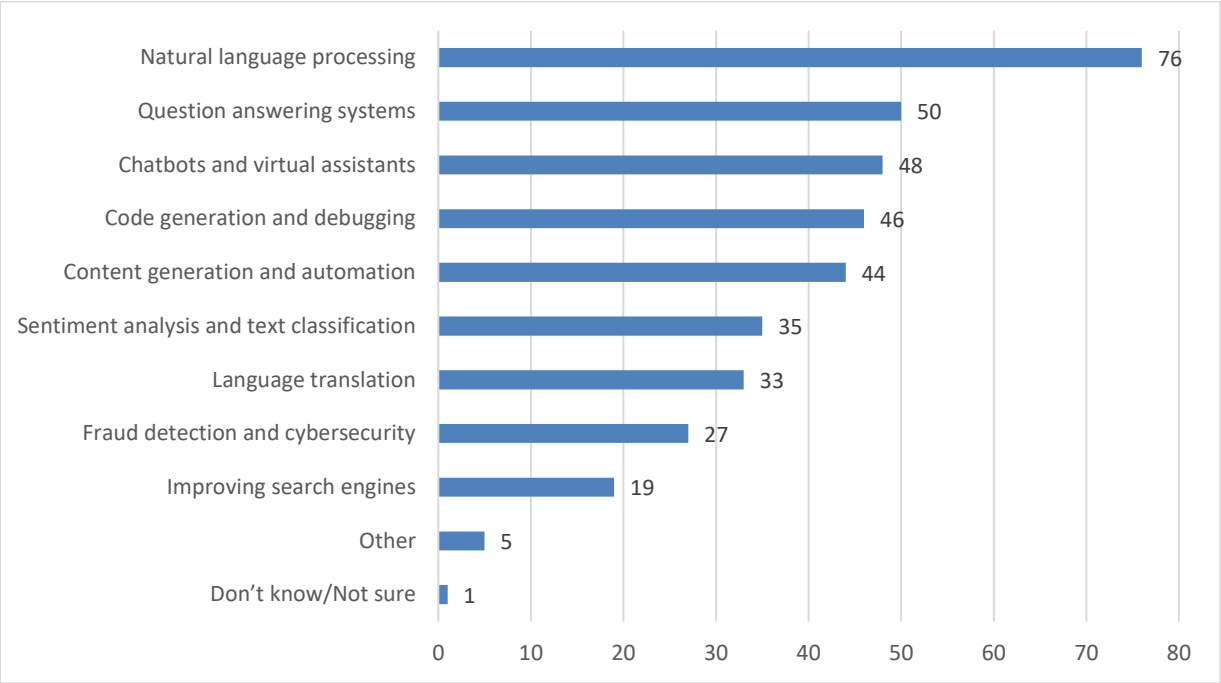
Exploring LLM Application Potential

Figure 2, below, arrays the current most important general-purpose LLM applications seen by survey respondents. The most selected options were natural language processing (NLP) (76 respondents) and question answering systems (50 respondents). Additional options chosen by many included chatbots and virtual assistant, code generation and debugging, and content generation. Respondents could select more than one answer and ultimately the average respondent selected 3.8 different options.

- It is too early to assess if this was due to a perceived broad range of LLM benefits appropriate to a respondent's organization or if in the early stage of LLM usage there had not emerged a clear primary preference.

FIGURE 2

Most Important General LLM Applications to Organization Today



N = 100

Respondents could select more than one answer.

Source: Hyperion Research, 2023

Table 3, below, arrays specific HPC workloads considered most promising for LLM enhancements, with data science/big data analysis selected by 80% of respondents for one of the top three choices and by 53% as the single most promising option. A number of other potential HPC workloads were seen as a top three option of LLM including Monte Carlo methods (37%), dense linear algebra (33%), and partial differential equations (31%).

Although data science-related activity was widely selected as the single most promising HPC workload (53%), almost half of the respondents selected other options as their top choice, with both dense linear algebra, algorithms, and libraries, as well as partial differential equations and boundary value problems each selected by 10% of respondents.

Table 3

HPC Workloads Considered Most Promising for LLM Enhancement

	Top Three Options	Single Most Promising Option
Data science/Big data analysis	80%	53%
Monte Carlo methods	37%	5%
Dense linear algebra, algorithms, and libraries	33%	10%
Partial differential equations and boundary value problems	31%	10%
Initial value problems and implicit methods	25%	6%
Sparse linear algebra	24%	6%
Spectral methods - Fast Fourier Transforms (FFTs) and applications	21%	4%
N-body / particle methods	20%	4%
Simple ordinary differential equations	15%	1%
Other	14%	1%

N = 100

For the top three options question, respondents could select the top three in order.

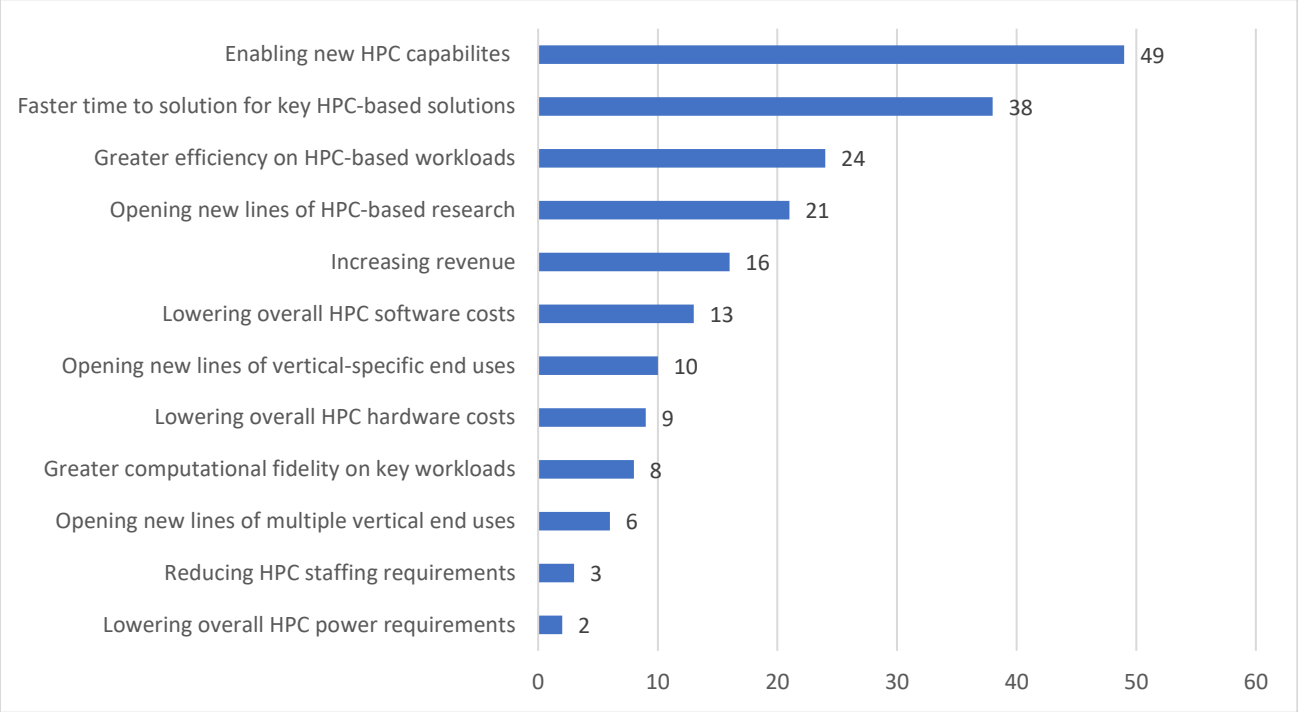
Source: Hyperion Research, 2023

LLMs from An Organization Perspective

Figure 3, below, arrays some of the higher-level larger organizational goals envisioned by the inclusion of new LLMs onto existing of planned HPC-related activities. Selected most often was enabling new HPC capabilities (49 of 100 respondents), followed by faster time to solution for key HPC-based solutions (38), and greater efficiency on HPC-based workloads.

FIGURE 3

Most Important Goal of New LLM capability



N = 100

Respondents could select more than one answer.

Source: Hyperion Research, 2023

Table 4, below, arrays the top challenges survey respondents' organizations face with introducing LLMs into their existing HPC-based workloads. Topping the list was complexity with integrating LLMs into existing HPC-based workloads (46%), followed by high/uncertain development costs (35%), and concerns with the cost of LLM-specific hardware or software (33%). In contrast, challenges that were of least concern included the notion that the technology is moving too fast for credible assessment of value (13%), confusion or uncertainty with LLM vendor selection (11%), and uncertainty of demonstrated computational performance improvements (10%).

- Of note is that concerns with integration complexity coupled with a lack of in-house LLM expertise points to a need for LLM suppliers to be able to assist end users to introduce and then properly operate an efficient tightly integrated HPC/LLM environment, at least for the short term.

Topping the list of challenges was complexity with integrating LLMs into existing HPC-based workloads (46%), followed by high/uncertain development costs (35%).

Table 4

Top Three Challenges with Introducing LLMs Into Existing HPC-based Workloads

	Option
Complexity with integrating LLMs into existing HPC-based workloads	46%
High/uncertain development costs	35%
Concerns with cost of LLM-specific hardware or software	33%
Lack of in-house expertise in LLMs	30%
Concerns with technical issues surrounding LLMs such as expandability and hallucinations	29%
Lack of demonstrated return on investment	25%
Long/uncertain implementation times	23%
Lack of credible data sources for LLM training	20%
High/uncertain operational costs	18%
The technology is moving too fast for credible assessment of value	13%
Confusion/uncertainty with LLM vendor selection	11%
Uncertainty of demonstrated computational performance improvements	10%
Other	6%

N = 100

Respondents could select up to three options.

Source: Hyperion Research, 2023

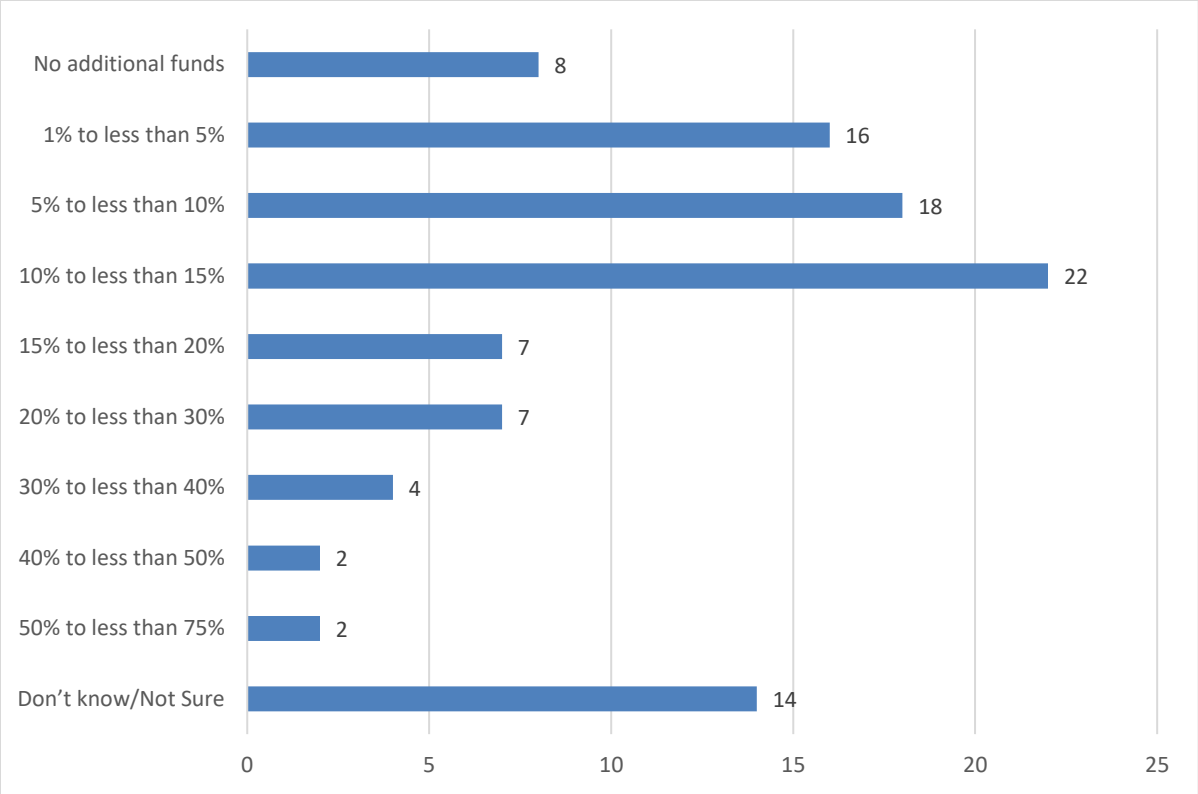
Figure 4, seen below, arrays the additional annual budget commitment the respondents believed their organization would be willing to commit to support LLM-enhanced HPC capabilities, with results widely dispersed, spanning little or no additional funds all the way to significant funding increases of 30% or more.

- The single largest response, 22 of the 100 respondents, was in the range of 10% to less than 15%.
- In total, 22 of respondents indicated that their organization would commit at least 15% or more, with one in ten envisioning additional budget commitment exceeding 30% of current expenditures.

- In contrast, 42 respondents indicated that their organization would likely commit to adding less than 10% of their current budget to support LLMs, including 8% seen as committing no additional funds.

FIGURE 4

Additional Annual Budget Commitment to Support New HPC-based LLM Enhancements



N = 100

Source: Hyperion Research, 2023

LLM Software Considerations

Table 5, below, lists by major sector the current sources of LLM software currently under consideration by respondents' organizations. Overall, open source LLMs are the predominant technology option (49%) for all respondents today, followed by commercial LLM sources (26%) and in-house developed LLMs (24%). There is some variation among the sectors surveyed, although the small sample size is not large enough to draw any significant nuanced conclusions.

- The academic sector favors open source at a rate more than 2 to 1 over commercial or in-house counterparts.
- The industry sector has the least bias towards any particular LLM software source and is the sector most open to exploring commercial LLM.

- The government sector has the lowest rate of commercial LLM sourcing, preferring open source by a wide margin.

Table 5

LLM Software Sources

	Total	Academic	Industry	Government
Average of % Open source:	49%	58%	46%	50%
Average of % Commercial LLM:	26%	22%	30%	18%
Average of % In-house development:	24%	20%	24%	32%

N = 100; 23; 63; 14

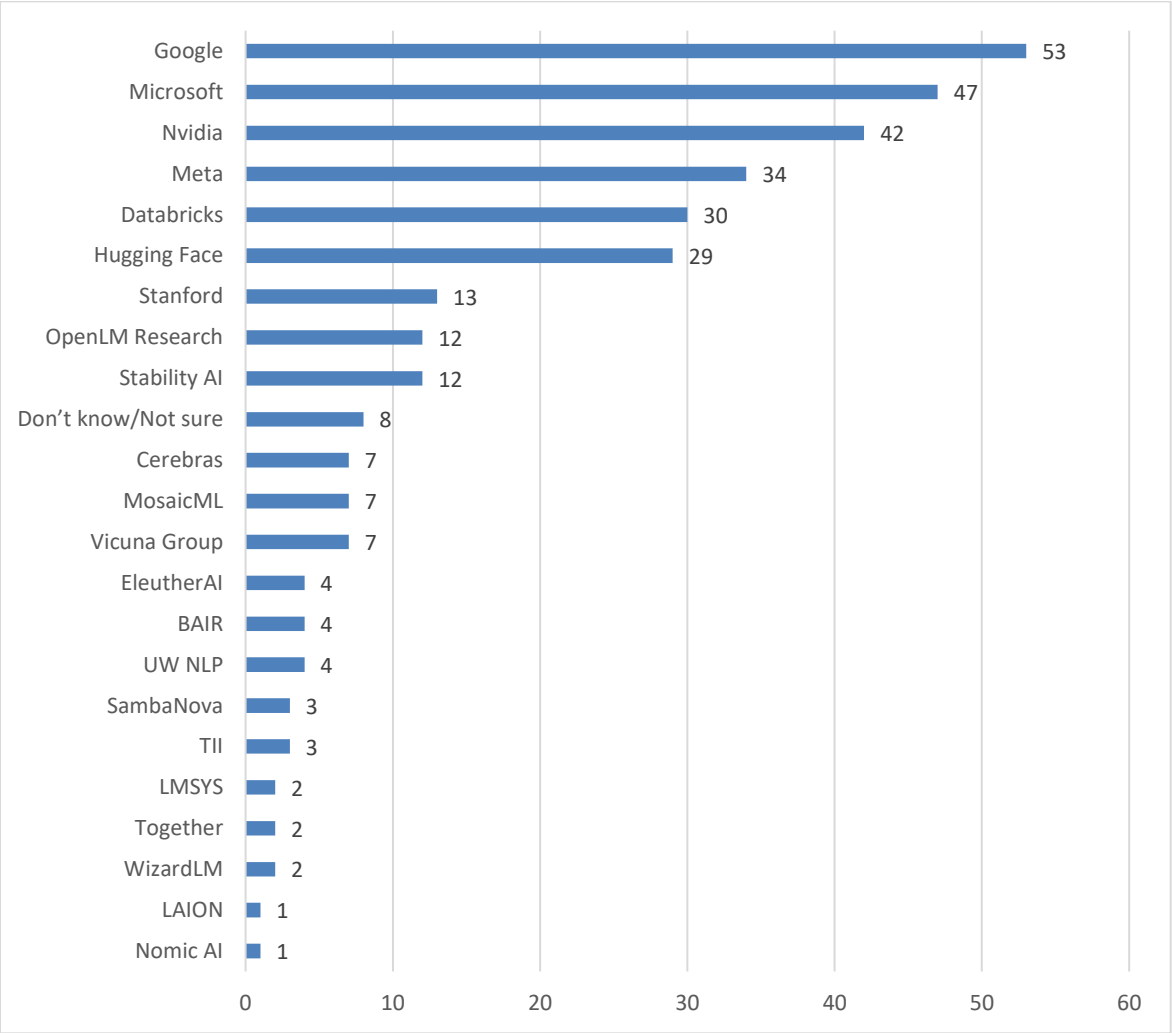
Source: Hyperion Research, 2023

Figure 7, below, lists the open source LLM developers used currently or within the next 12-18 months within respondents' organizations. The most selected options were Google (53 out of 100), Microsoft (47) and Nvidia (42).

- However, 23 different open source LLM providers were selected by the range of respondents.
- Ten different open source LLM providers were selected by 4 or less respondents, suggesting the range and variety of LLM sources under consideration by potential LLM end users.
- Respondents were invited to select all options that applied, and the respondent base in total selected 3.2 different open source providers per respondent.

FIGURE 7

Open Source LLM Developers Used Currently or Within the Next 12-18 Months



N = 100

Respondents could select all options that apply.

Source: Hyperion Research, 2023

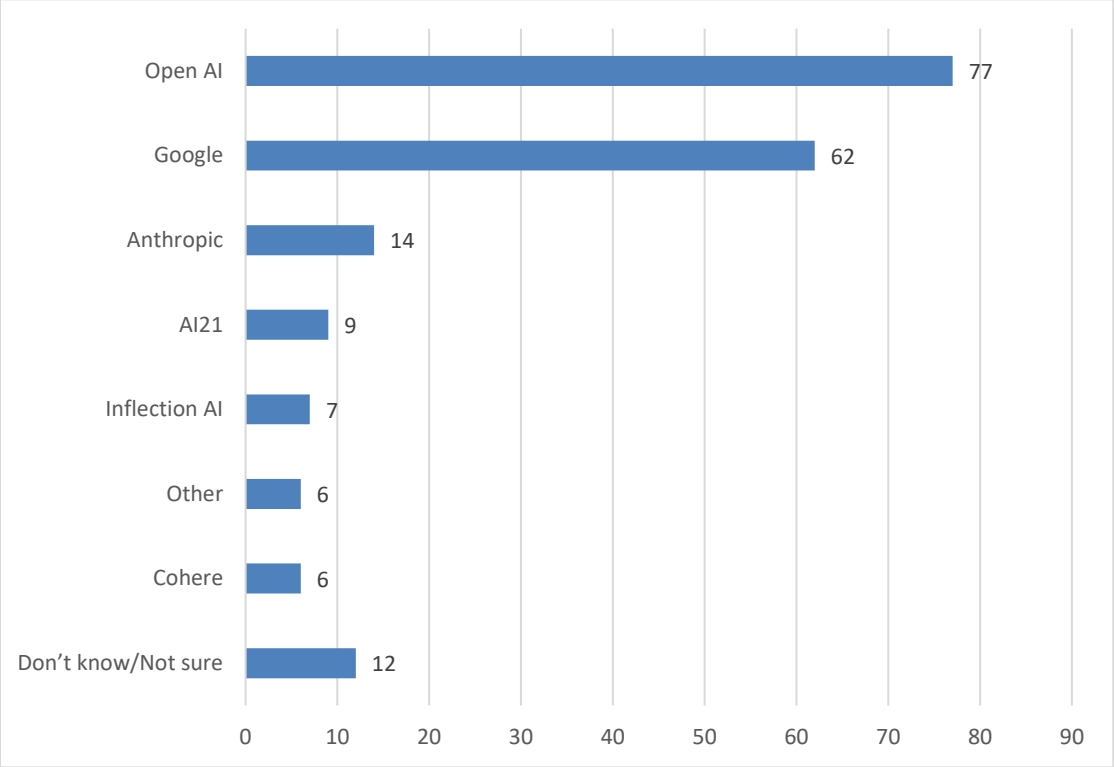
Figure 8, below, arrays the respondents' list of commercial LLM developers that their organization is either currently using or plans to use in the next 12-18 months. There were two major players identified in this sector: Open AI, which led the list (selected by 77 of the 100 respondents) followed by Google (62 of the 100).

- Smaller numbers went to Anthropic (14), AI21 (9) and Inflection AI (7). As was the case with the open source LLM providers, respondents could select all options that apply.

The average number of commercial developers cited per user was 1.9, compared with 3.2 LLM developers for the open source alternative. This difference could be due to either less sourcing options for the commercial space, more complexity with dealing with multiple commercial vendors, or the need to limit commercial variety due to a high cost of exploring multiple commercial LLM providers compared with open source counterparts.

FIGURE 8

Commercial Developers Used Currently or Within the next 12-18 Months



N = 100

Respondents could select all options that apply.

Source: Hyperion Research, 2023

Sources of LLM Support

Figure 9, below, arrays the various sources of LLM development that respondents' organizations could turn to at the various stages of LLM development, spanning foundation model construction, fine-tuning procedures, LLM integration into existing workloads, and supporting inference operations.

Notable overarching themes include:

- Respondents' organizations were looking to external LLM software developers at the foundational model stage but with a shifting emphasis to in-house software development in the later stages of LLM roll out and use.

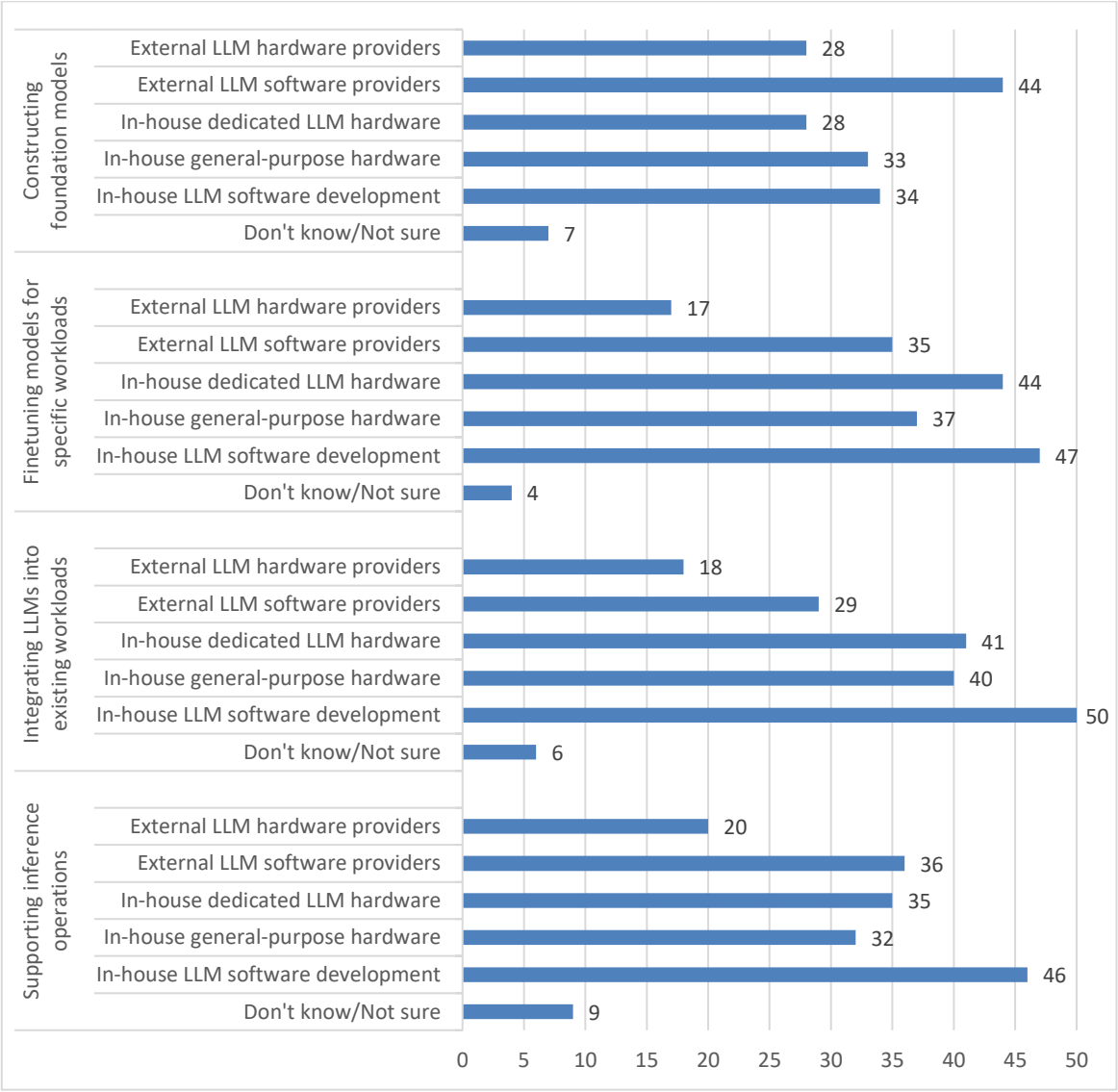
- In-house computational hardware needed to implement one or more stages of LLM development was the most prevalent hardware platform choice, with the sole exception of foundational model development that showed a slight preference for external LLM hardware providers. External, most likely cloud based, LLM hardware was seen as a lesser imperative.
- There was no significant difference in the type of in-house hardware considered appropriate to support LLM development and use, generally split between special purpose LLM hardware and general purpose counterparts, with the exception of model fine-tuning operation, which was seen as tending towards the use of special purpose LLM hardware.

Single LLM stage themes include:

- For foundational model construction, respondents preferred external LLM software providers, which was followed closely by in-house development. There was no clear preference for the use of either external or in-house hardware platforms to support such efforts. In addition, respondents were relatively split between using in-house dedicated LLM hardware versus in-house general purpose hardware for foundational model construction.
- For model fine-tuning procedures, respondents' organizations had a measurable emphasis on in-house software development using in-house, primarily dedicated, LLM computational hardware. For this stage, in-house hardware, either dedicated LLM or general-purpose, was selected over external hardware by a ratio of 4.5. Taken with the data above, it appears that many organizations may turn to external LLM suppliers for foundational model development but will rely more heavily on in-house resources for model fine-tuning.
- For LLM model integration, respondent organizations will most heavily rely on in-house resources, both in software and hardware. he emphasis on in-house hardware, either LLM dedicated or general-purpose, at a rate of 4.5 times that of external hardware, signals a strong preference and related need for the appropriate in-house skills to manage LLM integration into existing workloads.
- For supporting inference operations, respondents appear to favor in-house software support rather than turning to external LLM software providers. As with most of the other stages examined, with the exception of foundational model development, respondents' organizations were seen as preferring in-house hardware capabilities over external hardware providers.

FIGURE 9

Sources of LLM Development of Use at Major Stages: Model Building, Fine-tuning, Integration, and Inference Operations



N = 100

Respondents could select all options that apply.

Source: Hyperion Research, 2023

SUMMARY AND NEXT STEPS

HPC users leveraging or exploring LLM technology are a diverse group amidst an environment of rapid LLM advances and growth. Respondents' confidence in LLM technology within their organizations is high and this sentiment is represented in the anticipated short time to migrate workloads to an LLM-enabled system, their willingness to expand LLM-related budgets, and commitment to investing in long-term developments for LLM integration.

The study data suggests that this excitement is demonstrated by long-term expectations:

- Overall, LLM users expect a relatively speedy transition from an exploration phase to a production phase and have assumed a forward-looking position on AI-specific hardware and software procurement.
- This high expectation is perhaps most evident in the 50% of respondents who expect to be running production level LLM-enabled workloads within 12-18 months (more than doubling from the current 22%).

This optimism is further reflected in the anticipated diversity in LLM-enabled workloads. Data science and big data analysis are naturally the most prominent workload types, but traditional HPC workload types like Monte Carlo methods and dense linear algebra are also considered areas amenable to LLM integration. In the application area NLP is similarly the likely standout, but respondents indicated further exploration and expectations for applications from pure language processing like fraud detection, automation, and code generation.

LLM technology has already demonstrated much success and shows considerable promise in certain areas, as its applicability is explored in others. In such periods of rapid adoption, users must be able to identify how best to benefit from the new technology and, eventually, come to know its limits. With respondents indicating high levels of exploratory engagement and willingness to expand budgets for LLMs, the consensus suggests that the limits of LLM integration are far beyond its current use. With this in mind, the next 1.5-2.5 years will be a critical time for transforming this exciting new technology into a reliable, cost-effective tool and determining which applications can realistically benefit from LLM integration.

The next 1.5-2.5 years will be a critical time for transforming this exciting new technology into a reliable, cost-effective tool.

Though the potential is recognized, the path to effectively integrating an LLM into an HPC workflow has considerable obstacles. Simply put, as identified by respondents, it is expensive and complicated. While methods like tapping into the healthy open-source ecosystem or designing robust procurement strategies can help address the cost concerns, the worries surrounding complexity and sufficient expertise will take more time to soften. Not only are specialty experts rare, especially for organizations currently exploring different options, but they are also inherently costly in their own right.

- The ability to field a wide base of LLM experts is moving slower than the overall generative AI boom, and while there are experts in the field, the demand for their skills is growing considerably. Currently, user outlook is hopeful regarding these potential roadblocks.
- Notably, concerns surrounding the reliability of the technology as well as its potential to produce unintelligible or useless 'hallucination' outputs have taken a back seat.

Data collected in this survey is reflective of widespread user sentiments within the current generative AI boom. Confidence is high and many are already n the path of exploring the effectiveness of integrating LLMs to their HPC workflows. The most critical question that users and organizations must ask themselves is not if LLM technology could somehow be effective in their industry, but if they can identify where, and how they can implement a strategy to achieve those benefits in a cost-friendly way.

With high confidence comes high expectations. Organizations that can maintain realistic expectations can avoid over-costly procurements, integrate new technology at a safe pace, and more accurately identify when LLM integration is not a viable option.

The next 1.5-2.5 years are extremely important in the development of LLM technology and its role in advanced computing. With a significant number of users on the precipice of mission critical LLM integration and exploring new applications at a rapid pace, the fulfillment of these expectations remains a question. LLM's have already reached usefulness beyond what could be considered their most logical application area, but how much further could it go? Organizations that manage to maintain relevant AI literacy, make decisions based realistic expectations, and an openness to cost-friendly innovation stand a good chance of coming out ahead in this rush to adopt.

APPENDIX: SURVEY DEMOGRAPHICS: RESPONDENTS, ORGANIZATIONS, AND BUDGETS

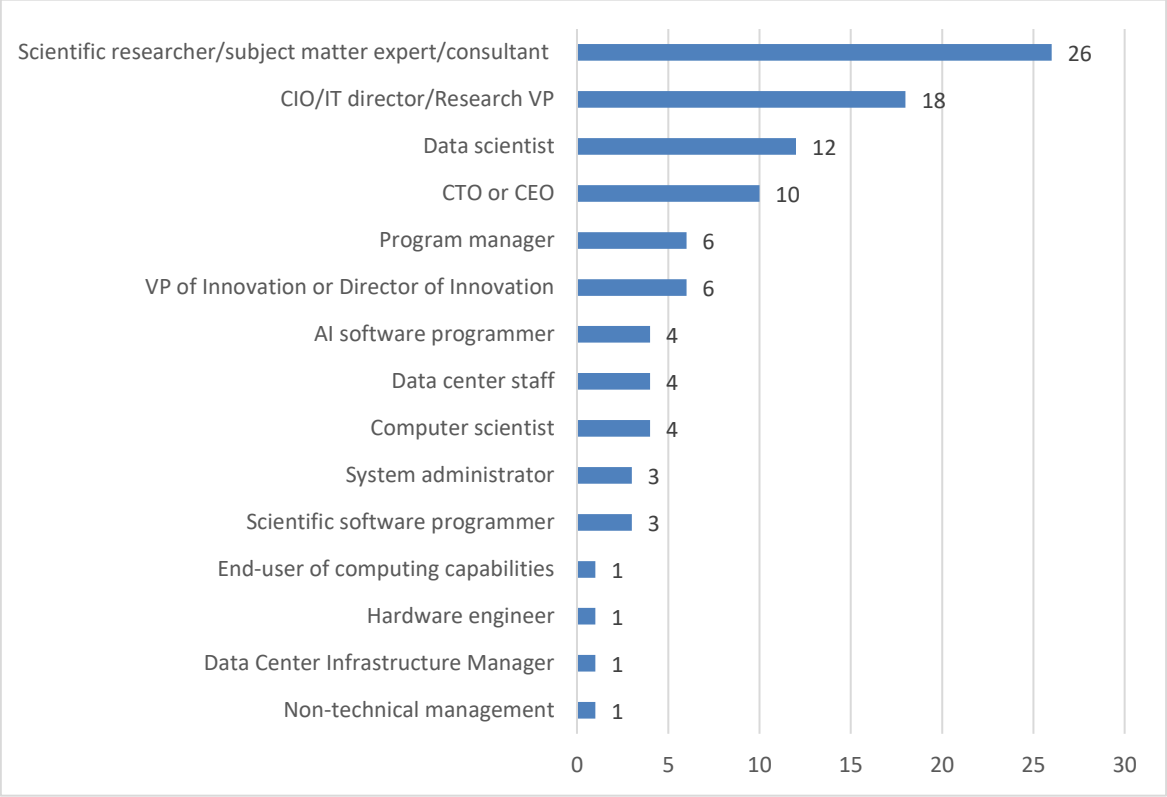
This appendix provides additional details on the demographics of the survey respondents and the organizations they represented. It consists of three main segments: individual survey respondents' background information, general demographics about the organization they represent, and select data on the budgetary levels of those organizations.

Respondent Demographics

Figure 10, seen below, lists the various self-identified job titles of survey respondents. Although scientific research/subject matter expert was the most selected option, the broad range of job titles suggests that the impetus, or at least the interest, to consider an LLM activity can draw attention from a number of sources within any one organization.

FIGURE 10

Respondent's Job Title



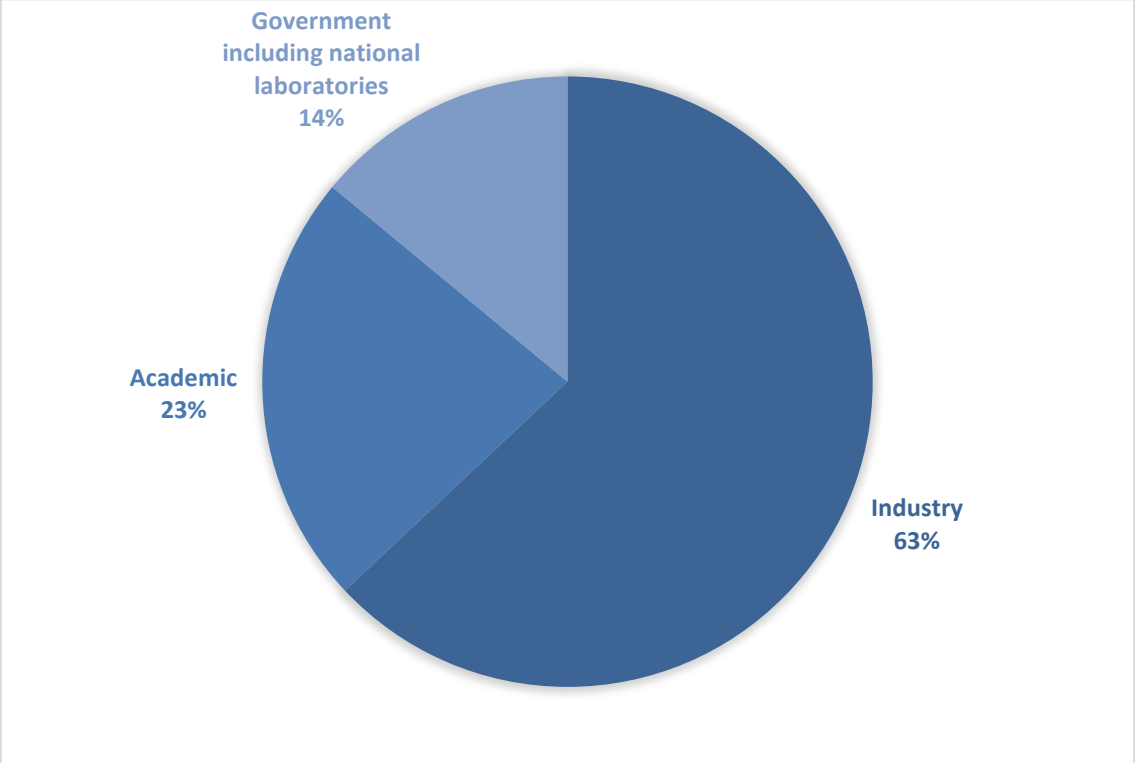
N = 100

Source: Hyperion Research, 2023

Figure 11, seen below, shows the breakdown of respondents' organizational sectors. The majority were from the industry sector (63%), followed by academia (23%), and government (14%).

FIGURE 11

Respondent's Sector Affiliation



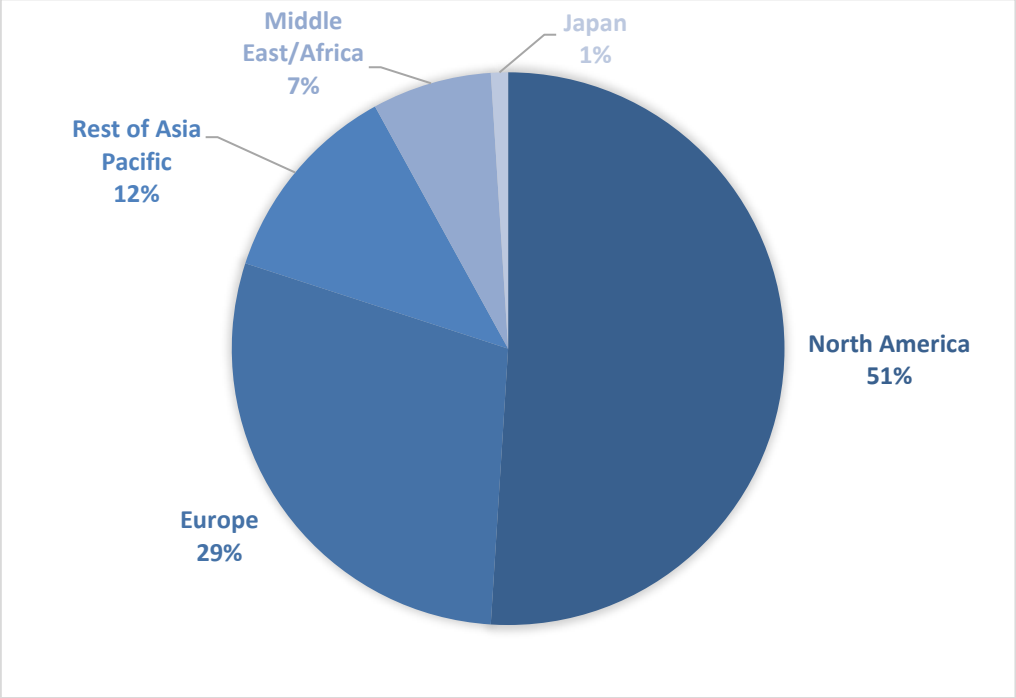
N= 100

Source: Hyperion Research, 2023

Figure 12, seen below, outlines the respondents' headquarters location. About half were sited in North Amecia with the remainder divided among Europe, the rest of Asia Pacific (without Japan and China), and Japan. However, this is but one measure of locating a particular center of activity for HPC-related LLM work. Increasingly, respondents are operating in countries or regions different from their organization headquarter location, and many organizations may have such efforts spread across a number of different countries.

FIGURE 12

Respondent's Organization Headquarters Location



N=100

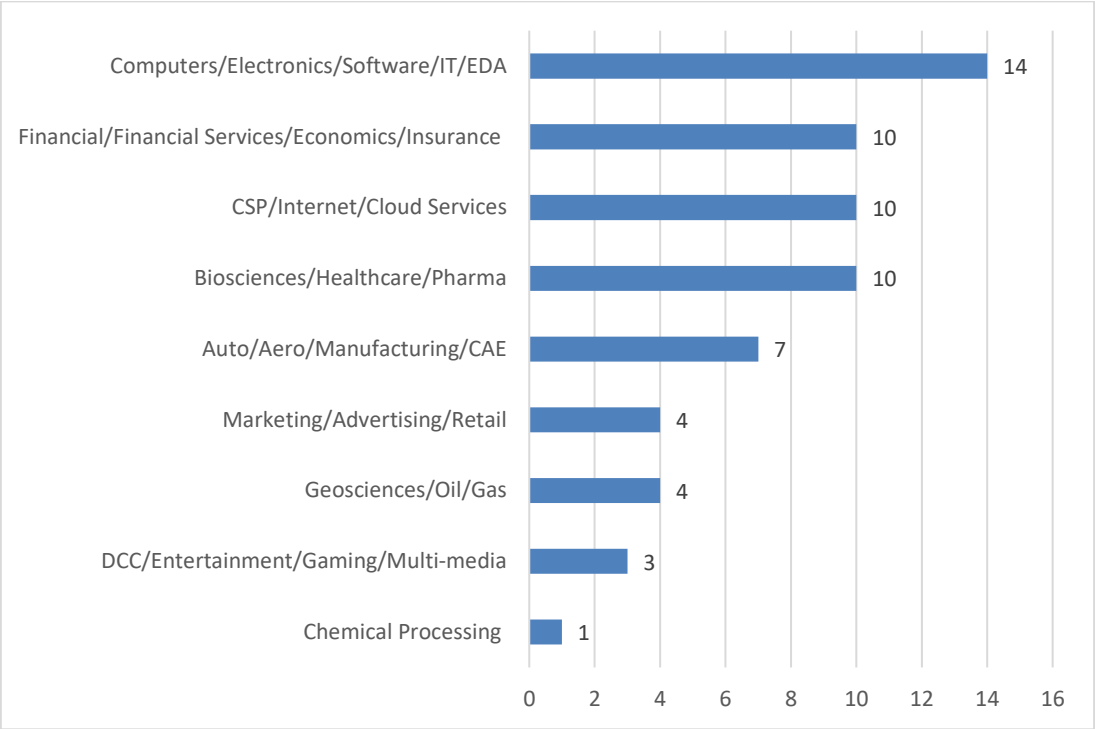
Source: Hyperion Research, 2023

Organizational Demographics

Figure 13, seen below, outlines the respondents' main area of activity for those that identified as coming from the industry sector. Computers/electronics/Software/IT/EDA was the most selected industrial category, followed by finance, CSPs, and the bio-life sciences.

FIGURE 13

Industry Respondent's Organizational Main Area of Activity



N= 63

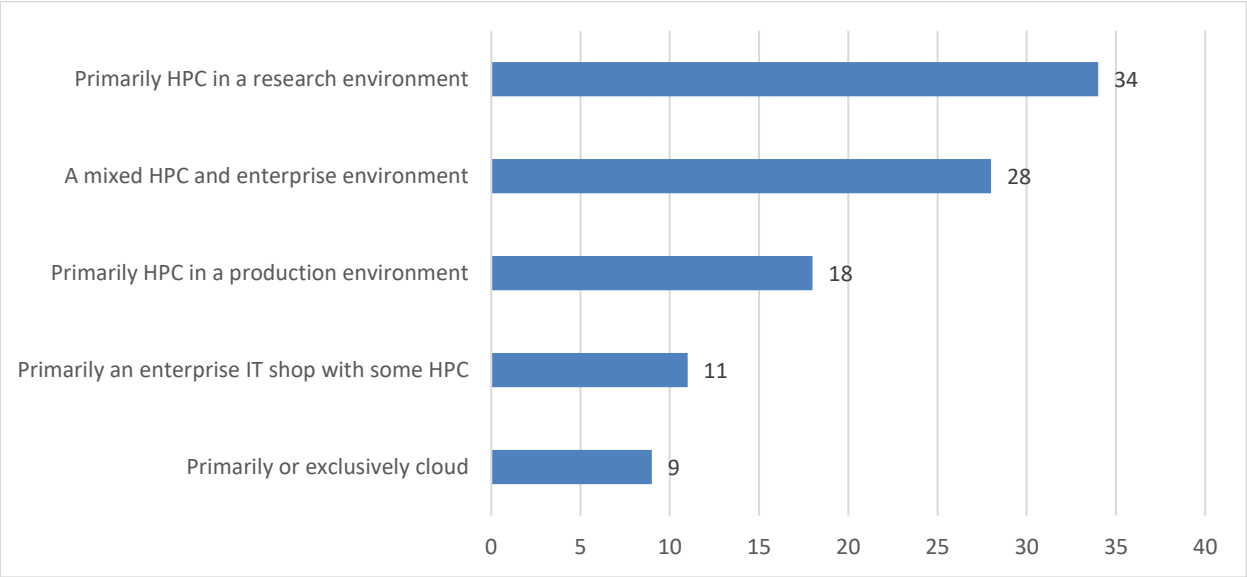
Source: Hyperion Research, 2023

Figure 14, shown below, arrays the self-identified computing environments for the set of survey respondents. The major selection was primarily HPC in a research environment, but many respondents indicated that their site was a mixture of HPC and enterprise class systems.

- Additional information about the overall budgetary commitments to the three major hardware platforms (on-premises, public cloud, and private cloud) are shown in Tables 15-17 below.

FIGURE 14

Organziation's Overall Compute Environment



N = 100

Source: Hyperion Research, 2023

Survey Organizations' Budget Specifics

Table 5, seen below, lists the overall current annual budget of respondents' organizations to support their combined HPC-based requirements, including both on-premises and cloud-based resource commitments. Results here show that the survey reached a relatively broad base of HPC sites including those under US$100,000 (7%) to those exceeding US$100 million (5%).

- For a sectoral perspective, academia had the most participation at the lower levels of annual budget (29%) between US$100,00 to US$500,000, zero for government sites.
- Meanwhile, government sites tended towards higher value annual budgets, more so than the other sectors: only 16% academic sites had budgets between US$$20 million and US$100 million, and only 15% of industry sites and 30% of government sites fell within that budget range.

For the subsequent table, there are differing rates of sector participation: academic had 21 respondents, industry 47, and government 13.

Table 5

Organization's Current Annual Budget (in US dollars) to Support HPC-based Requirements (on-premises and cloud-based)

	Total	Academic	Industry	Government
Under $100,000	7%	19%	4%	0%
$100,000 to less than $250,000	5%	10%	4%	0%
$250,000 to less than $500,000	17%	14%	15%	31%
$500,000 to less than $1 million	16%	10%	19%	15%
$1 million to less than $5 million	16%	14%	17%	15%
$5 million to less than $10 million	9%	10%	11%	0%
$10 million to less than $20 million	9%	14%	6%	8%
$20 million to less than $50 million	12%	10%	13%	15%
$50 million to less than $100 million	4%	0%	2%	15%
$100 million to less than $500 million	4%	0%	6%	0%
More than $500 million	1%	0%	2%	0%

N = 81; 21; 47; 13

Source: Hyperion Research, 2023

Tables 15-17, seen below, show the distribution of respondents' organization budget commitments to three major hardware platforms: on-premises, public cloud, and private cloud, respectively.

For this section, the sample size was reduced to 85 due to inconsistencies in some responses.

Organizational budget commitment to on-premises capabilities is the most widely dispersed option of the three platform options. 26 out of the 85 respondents (30%) commit at least 90% of their overall IT budget to on-premises capabilities, while 14 out of 85 (16%) have zero on-premises budget commitment.

- Likewise, almost one half of the respondents' organizations commit between 20% and 80% of their budget to on-premises, with the remainder split across some form of cloud services.

Organizational budget commitment skews towards less dependence on public clouds than the on-premises alternatives. Indeed, almost 60% of survey respondents reported committing less than 10% of their overall budget commitment to public cloud-based platforms.

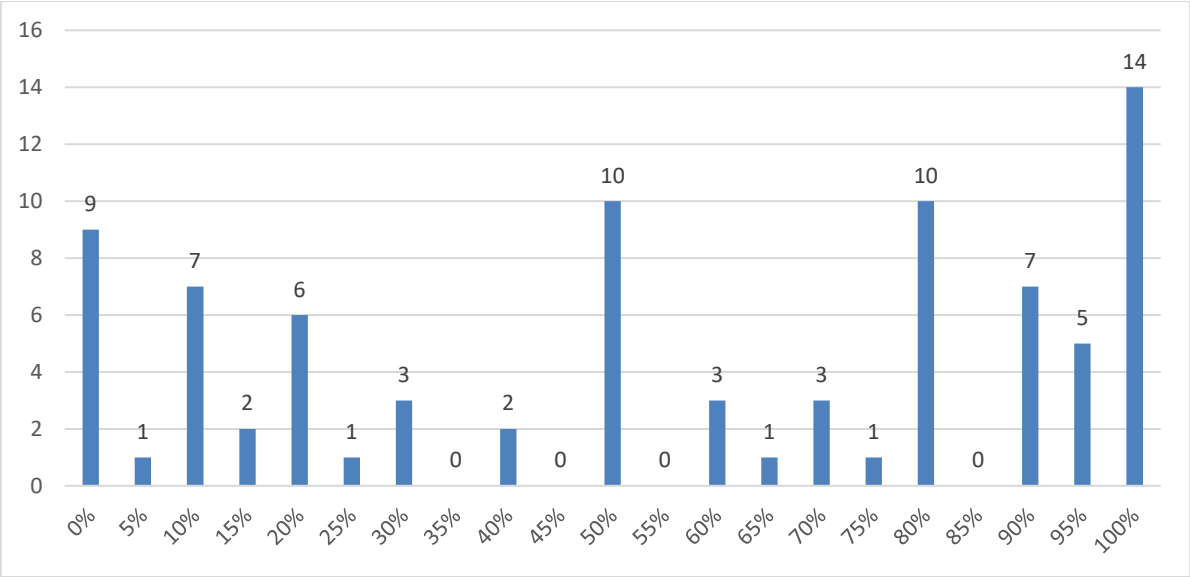
- Only about 10% committed more than 80% of the budget to public cloud-based resources, and only one respondent organization was wholly committed to a cloud-based infrastructure.

Organizational budget commitment to private cloud was low compared to public cloud, 43% of survey residents reported zero budget commitment to private clouds, and only one in ten committed more than 50%.

- Budget commitment to both cloud modes, public and private, were reported by a number of respondents.

FIGURE 15

Organization's Annual Budget Commitment by Platform (% of Total Budget): On-Premises

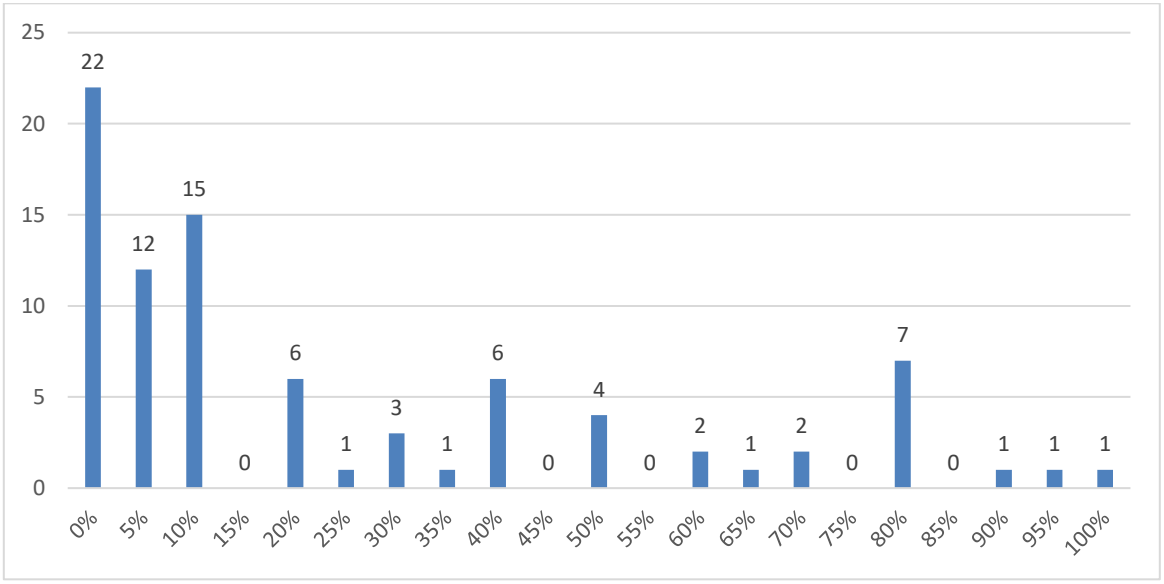


N = 85

Source: Hyperion Research, 2023

FIGURE 16

Organization's Annual Budget Commitment by Platform (% of Total Budget): Public Cloud

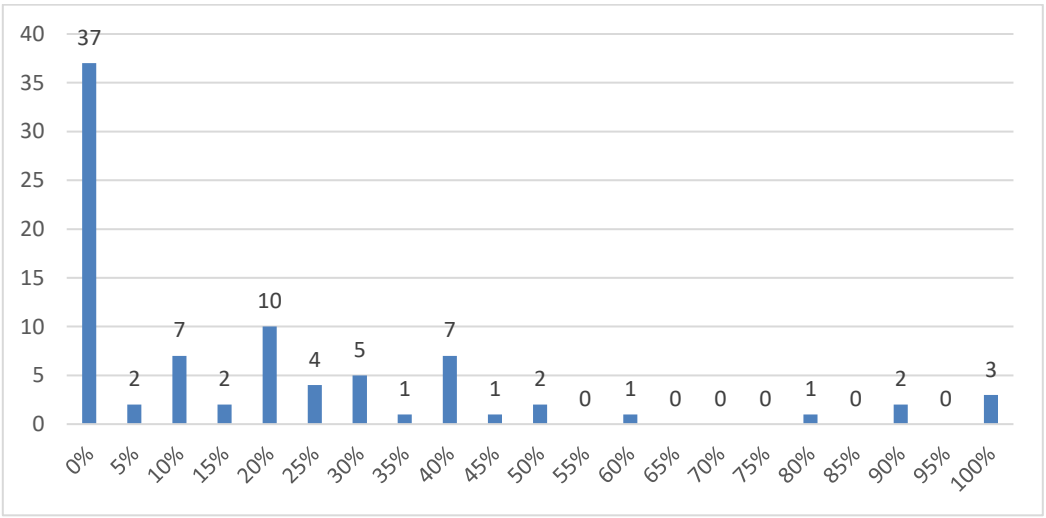


N = 85

Source: Hyperion Research, 2023

FIGURE 17

Organization's Annual Budget Commitment by Platform (% of Total Budget): Private Cloud



N = 85

Source: Hyperion Research, 2023

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). We provide 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
