The Hidden Costs, Challenges, and Total Cost of Ownership of Generative AI Adoption in the Enterprise

Revealing the Profound Gap Between Hyper-inflated Expectations and Business Reality

Insights into critical cost considerations, investment benchmarks, key use cases, and pricing metrics

Presented by CLEAR|ML in partnership with AI INFRASTRUCTURE ALLIANCE

This global survey report is the second conducted by ClearML and the AI Infrastructure Alliance (AIIA) in 2023 on the business adoption of Generative AI. Like our first report[1], we surveyed 1,000 AI Leaders and C-level executives in charge of spearheading Generative AI initiatives within their organizations. Both survey reports shed light on the adoption, economic impact, and significant challenges these professionals face in unleashing Generative AI's potential at scale.

In our last report, while the majority of respondents said they need to scale Generative AI, they also said they lacked the budget, resources, talent, time, and technology to do so. Given AI's force-multiplier effect on revenue, new product ideas, and functional optimization, it's clear that critical resource allocation is needed for companies to effectively invest in AI to transform their organization. Key challenges in Generative AI adoption within the enterprise include:

- Managing overall running and variable costs at scale
- Having complete oversight and understanding of LLM performance
- Hiring human capital and the lack of availability of specialized talent
- Improving efficiency and productivity while managing costs and TCO
- Increasing governance and visibility

In this report, we put our finger on the various considerations of the hidden costs and unknowns of Generative AI business adoption. We tried to unpack how AI leaders are navigating the uncharted territory of hidden operating costs related to Generative AI, which are often described as unfamiliar and unpredictable. We also explored how global organizations plan to balance Gen AI investments with expected outcomes and their overall running and variable costs.

Our findings show that it is essential for organizations and AI leadership to develop an effective, strategic approach to calculating, forecasting, and containing these costs tailored to their own organization and its unique business use cases. Hence, we chose to **identify** how confident AI leaders and C-level executives feel about accurately **predicting** and **forecasting** the TCO and ROI for Gen AI in their organizations while considering key factors and cost drivers such as setup, training, maintenance, running costs, specific use cases, and variable costs such as compute.

[1] - Our initial survey report, "Enterprise Generative AI Adoption: C-Level Key Considerations, Challenges, and Strategies for Unleashing AI at Scale," can be downloaded here: https://go.clear.ml/new-research-report-on-enterprise-generative-ai-adoption

KEY FINDINGS

1 Blind spot

Respondents are not considering the total costs of Generative AI

Based on survey answers, we found that most respondents believe their Gen AI costs are centered around model development, training, and systems infrastructure. For example, the costs associated with how a model works – human capital, the tools and systems to run it, and the app/UI for users. Unfortunately, the reality is quite different. We believe respondents are underestimating how messy data can be and the heavy lifting needed for data prep. It's worth noting that this is even more challenging if their company is using AI as a Service (i.e. using an API to connect to a LLM such as OpenAI ChatGPT, Google Bard, or Cohere™ Generate). Similarly, respondents are underestimating the time required by SMEs to work with the engineering team to ensure the model is accurate and "good enough" to roll out. Most importantly, a shockingly low 8% of respondents said they would attempt to control their budget by limiting models and/or access to Gen AI to better manage their budgets, which means they are not thinking about running costs, which we expect is going to be a huge surprise for them given their pivotal impact on TCO as a pricey cost driver.

2 Unrealistic expectations

Most companies want to implement and run Gen AI themselves

Nearly every respondent (91%) plans to resource or staff in-house to support future Gen AI efforts. That's bad news for consultants who all seem to be building Gen AI capabilities into their talent pool, but it does lead us to believe that organizations are considering scaling Gen AI for the long haul. However, that requires some serious cost considerations for how they are budgeting going forward and how to be most efficient using their budgets year-over-year. They may well be overestimating how much they can do with their budget, particularly in light of the findings above. It's interesting that 21% of respondents want to use their existing team, which means finding more ways to scale themselves efficiently to do more with less -- or just produce fewer models.

3 Critical prioritization

The key to selecting use cases for implementation within budget

While 82% of respondents are considering 4-9 use cases for their organization with end users ranging from 501-5,000, an alarmingly low 20% of respondents have allocated an annual budget of more than $2 million. That is worrying, as according to ClearML's TCO calculator the first year of training, fine-tuning, and serving a model for 3,000 employees hovers around $1 million (depending on data corpus and use case) using an in-house team, with future economies of scale possible through shared compute usage.

Meanwhile, 32% of respondents reported they are currently using ChatGPT, and these respondents are likely to find scaling across their business quite expensive, as costs grow linearly with token usage. We're hard-pressed to understand how this usage will ultimately fit within estimated future budgets, another gap between vision and reality.

4 One size does not fit all

Organizations will need a wide variety of Gen AI tools and models

We found that 37% of respondents plan to use Gen AI for content generation, which can be accomplished with popular out-of-the-box single use case applications such as Jasper™ and Copy.ai or available LLM APIs such as Cohere™ Generate. This use case is the least expensive to address, as organizations can simply purchase off-the-shelf apps with a low-cost subscription per user. Best of all, there is no need for organizations to share proprietary data with the app developer, so it is also a low-risk activity, one that is easy to outsource.

Having said that, three of the most commonly requested use cases require significant access to internal documents and internal organizational data in order to produce accurate and helpful results.

These are:

- Content recommendation engine for supporting internal teams
- Assistant for strategy, corporate planning, and finance; and
- Gen AI as a product feature

The models for these use cases will likely need to be in-house and most likely on-prem in order to protect company data and IP, which means businesses will need to make the investment to build their in-house teams to support multiple models for multiple use cases.

5 Reality gap

Limiting access to Gen AI is not seen as an effective way to stay on budget

92% of respondents are committed to growing budget inline with users and do not want to stay under budget by limiting access. 42% of respondents said they would grow budgets to accommodate more users and 50% will try to find savings through economies of scale.

However, achieving economies of scale through AI as a Service is virtually impossible because the price to use the service increases linearly with usage. Not only that, prompt engineering efforts are typically customized for each use case, so for businesses running multiple use cases, there are no time/energy savings. For enterprises running multiple models for various business units and use cases, the easiest way to attain economies of scale is through resource pooling: leveraging human capital that can build, maintain, and monitor the models across the business, as well as sharing compute power for serving.

Another concern that highlights the gap between hyped vision and reality is the willingness to give employees access to Gen AI (while that's great) will lead to spiraling costs that may catch organizations unawares. Underestimating costs as usage goes up seems to be a common theme in the results of this global survey. This is likely to leave organizations in a very difficult spot in the future, one that might cause a reshuffling of resources and the need to supplement budgets mid-year to bridge the usage gap.

6 Caveat emptor

It's astoundingly difficult for AI Leaders to predict the future hidden costs of Gen AI for their business

As we mentioned before, only a mere 9% of respondents are thinking about running costs; any organization not considering the total cost of ownership for Gen AI is in for a huge surprise when the bill comes due.

Meanwhile a third of respondents acknowledge that OpenAI's APIs are slow/unresponsive/unstable, and the costs of the LLM models' APIs are high and/or growing too fast, although 64% of respondents accept that it may cost more than $200/year/user. But compare this to the 50% of respondents who believe that 11-25% of all employees will be using Gen AI in year 1 of rollout, escalating to 26-50% of employees in year 2, and you can see how quickly the margins grow and how total costs will accumulate.

The bottom line? Organizations of all sizes seem ill-prepared to scale Generative AI. While they recognize running costs are high, they are not accounting for them in their forecasts and estimations of cost drivers. We believe organizations need to better align their Generative AI strategy with their business goals and operating budgets and allocate the necessary resources and governance in order to bridge the gap between their vision and reality.

We all know that this technology has the potential to unleash huge revenue opportunities, but it doesn't seem possible when companies indicate they aren't considering all the various cost factors. With that warning shot across the bow, let's dive into who answered our survey.

DEMOGRAPHY

To start with, let's look at the demographics of who we spoke with in our survey.

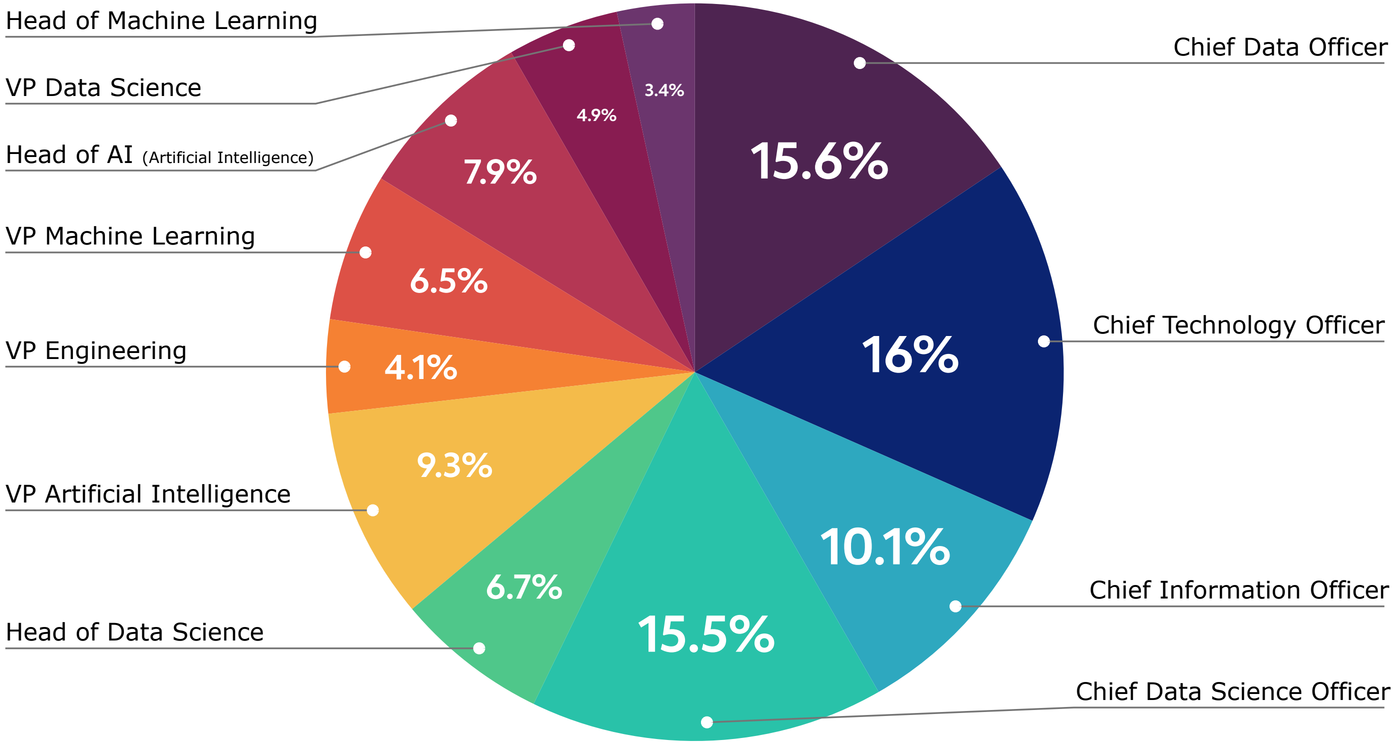
Essentially, we surveyed 1,000 respondents from companies with more than 500 employees, with the bulk of organizations (60%) employing 1,000-9,999 employees and classified as enterprises. Most of the survey respondents (89%) were between the ages of 35 and 54. The global survey primarily focused on 3 major markets: North America, EMEA and Asia-Pacific.

We also talked with AI/ML leadership and the heads of teams, with job titles such as Chief Data Officer, Chief Information Officer, Chief Technology Officer, and Chief Data Science Officer. That means the results primarily represent the C-suite as well as VP-level and heads of AI departments.

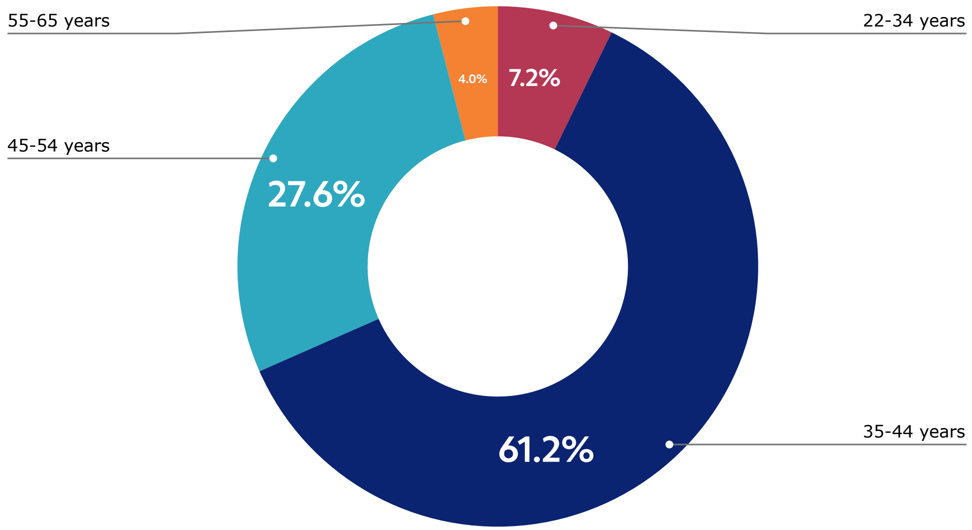
Lastly, we spoke to people across an impressive array of industries – everything from Financial Services, Retail, Manufacturing, and CPG to Telecommunications, Energy, Technology, Healthcare, and more. The largest representations came from Automotive, followed by Retail/ Wholesale Trade, Computer Software, and Energy/Utilities/Oil & Gas, but no vertical we surveyed represented more than 6% of the total respondents, so we had a wide range of views across a wide variety of verticals.

Here's a visual look at the demographics of our surveyed respondents:

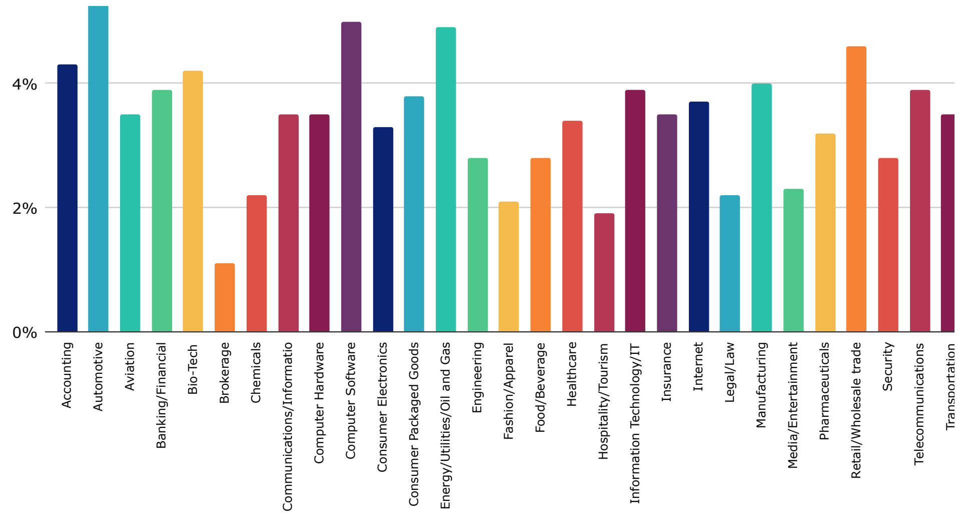
Title/Position of Surveyed Respondents



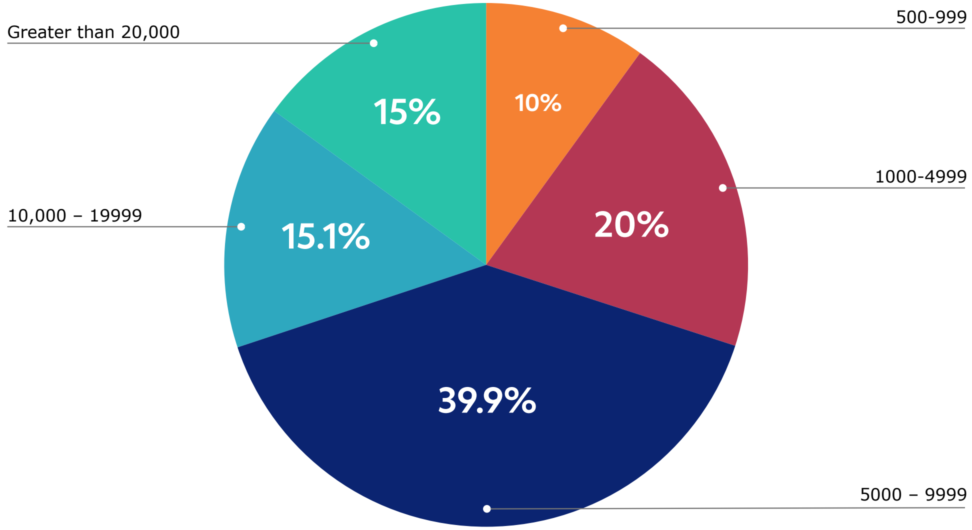
Age of Surveyed Respondents



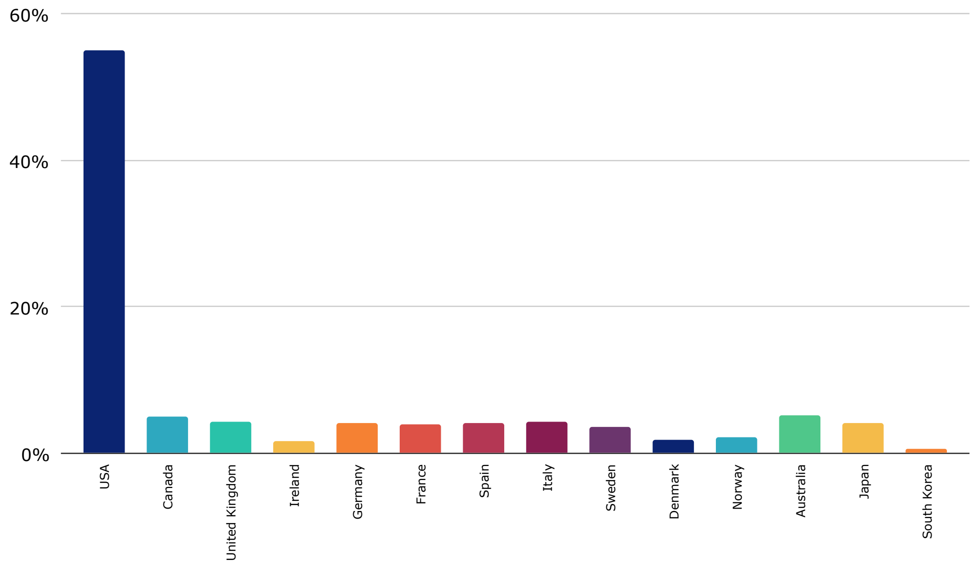
Industries Represented by Respondents



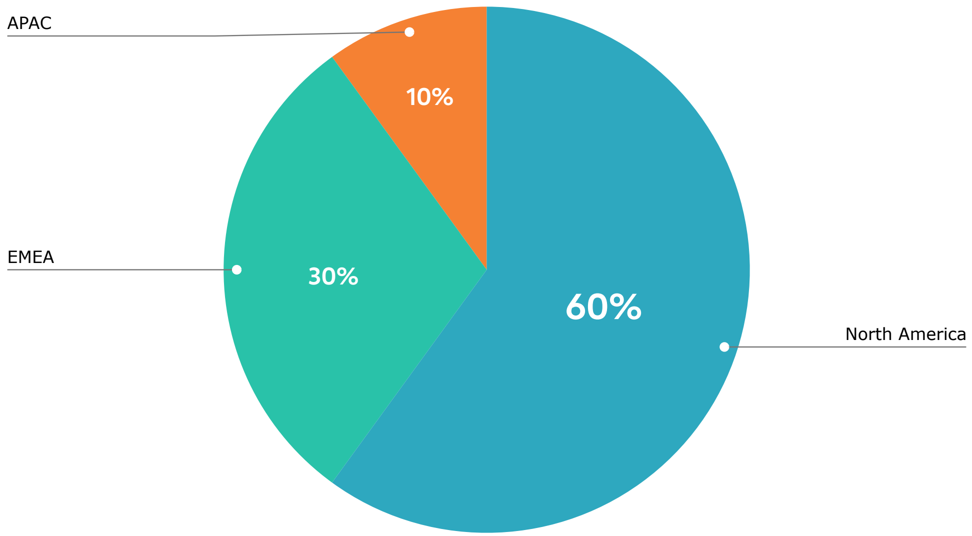
Company Size



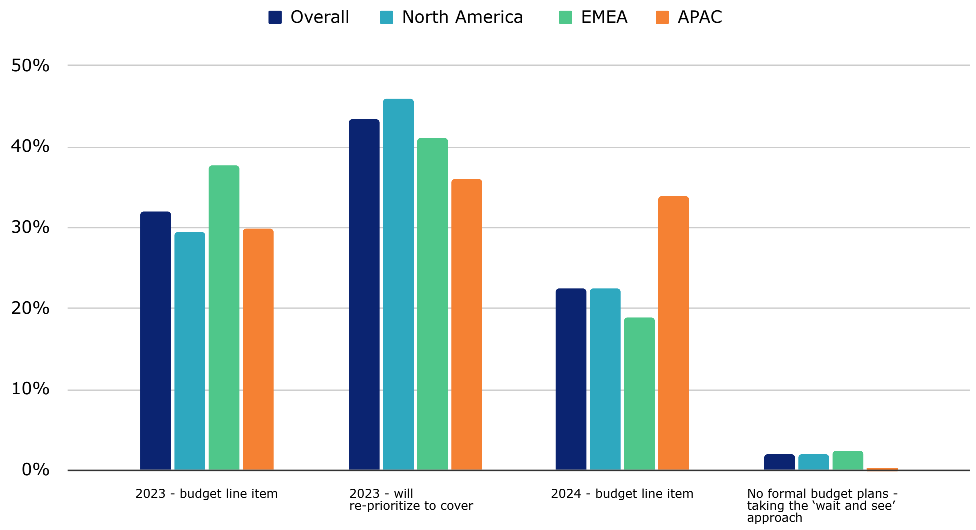
Headquarters of Responding Businesses



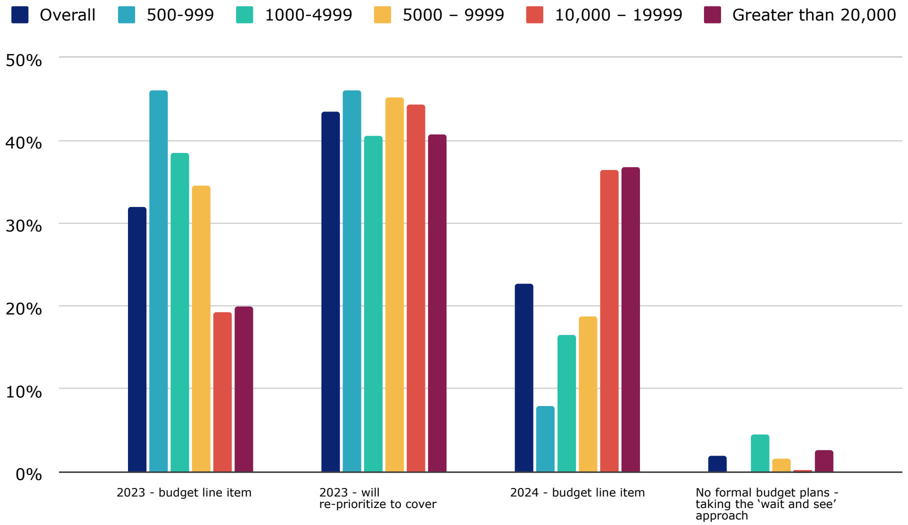
Geographic Regions of Headquarters



Generative AI as a Budget Line Item by Geography



Generative AI as a Budget Line Item by Company Size

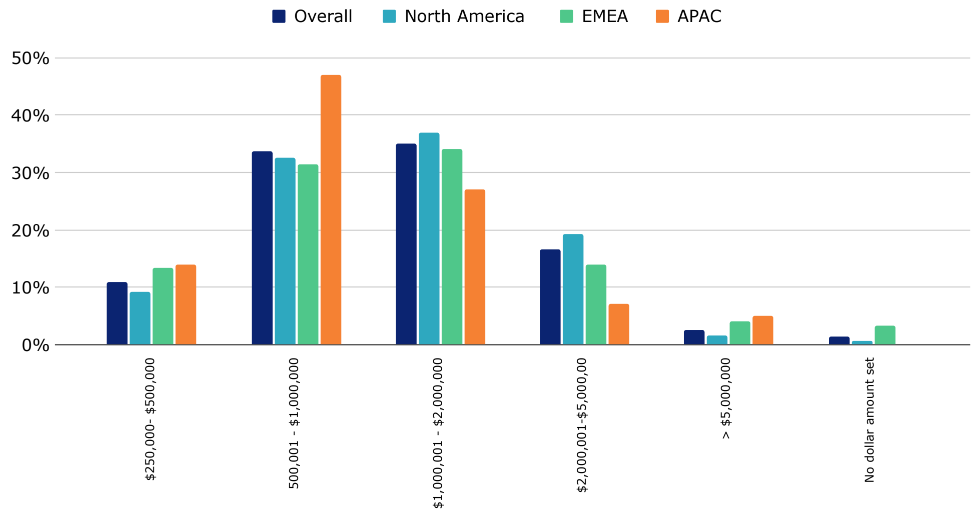


QUESTIONS

Now let's turn to the actual questions at the heart of the survey. As with most technology, we started with the foundation of planning, which is budget. From the respondents' answers, it's clear that organizations are investing in Generative AI now and are voting with their dollars. In fact, 76% of respondents indicated they have a GenAI budget line item in 2023, with an additional 23% saying it's a 2024 line item.

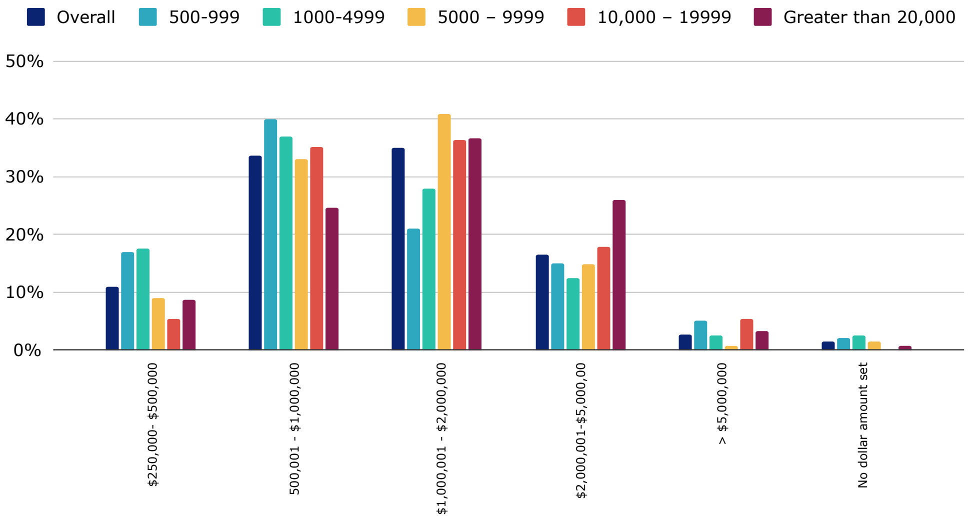
From our first survey, “Enterprise Generative AI Adoption: C-Level Key Considerations, Challenges, and Strategies for Unleashing AI at Scale,” we know that they are primarily investing to harness the power of Gen AI to drive internal and external product innovation (44.4% of respondents ranked this benefit of Gen AI as first or second), supercharge knowledge workers’ efficiency (36.2% of respondents ranked this benefit of GenAI as first or second), and drive revenue (45.9% of respondents ranked this benefit of Gen AI as first or second).

Generative AI Budget Allocation by Geography



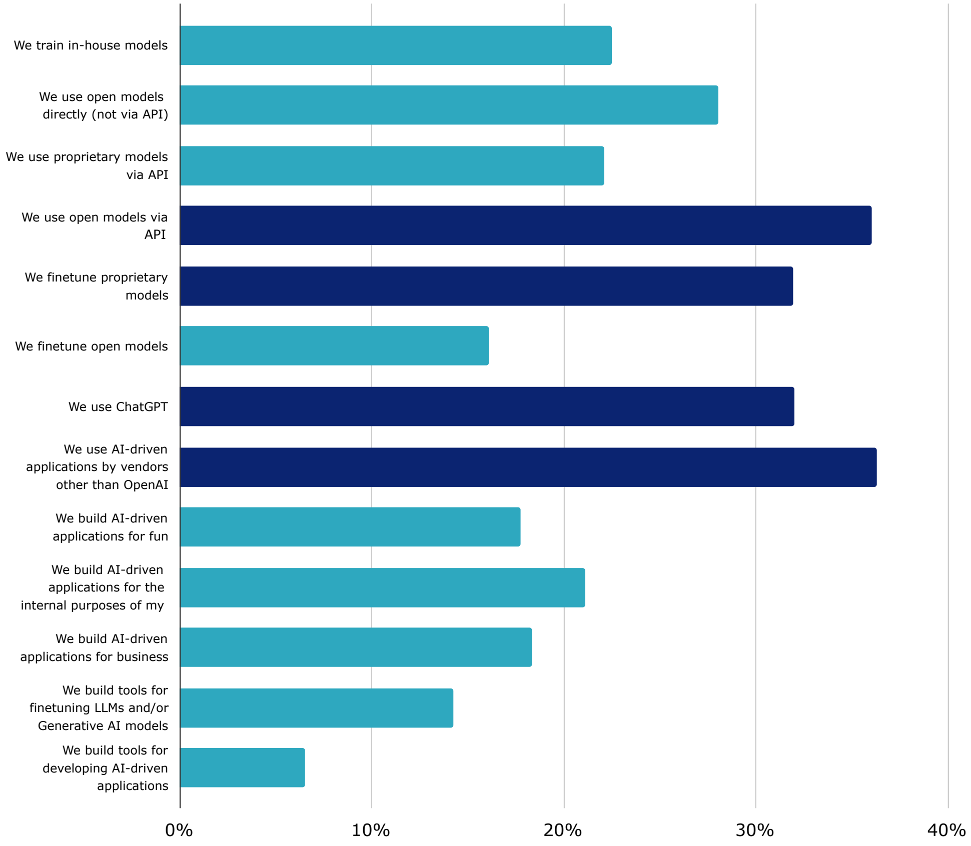
Given that most organizations have already budgeted for Generative AI, we wanted to know the size of their budgets. We found that 56% of organizations plan to invest $1M -$5M in Generative AI adoption. An additional 34% of large start-ups and mid-market organizations with 500-1000 employees respondents plan to spend $500K-$1M in Generative AI.

Generative AI Budget Allocation by Company Size



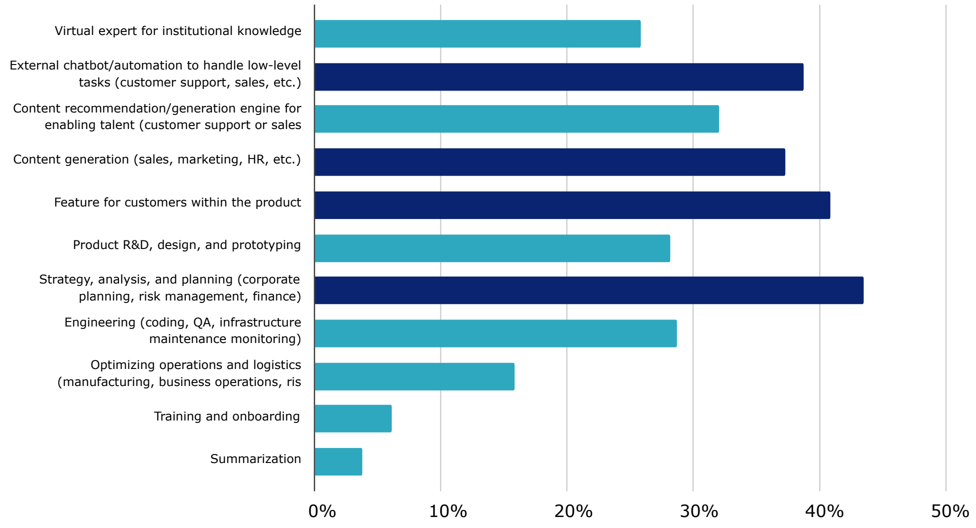
Next, we asked how respondents' organizations plan to work with LLMs, Gen AI models, and AI-driven applications. 36% of respondents said they use open models via an API or indicated the use of AI-driven applications by vendors other than OpenAI, while 32% of respondents reported they plan to fine-tune proprietary models themselves or use ChatGPT. In the chart below, you can see all of the various ways companies plan to work with Generative AI:

How Companies Work with Gen AI



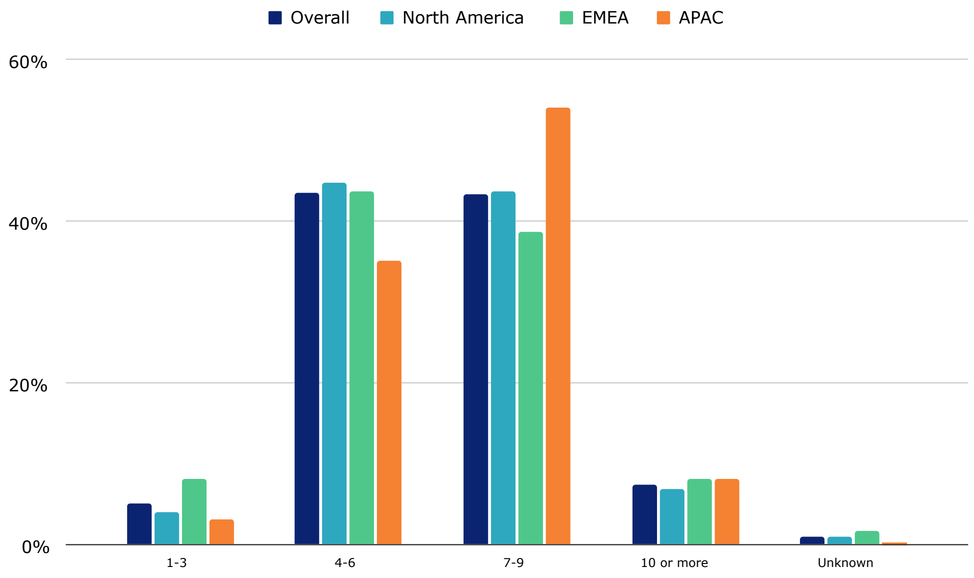
When it comes to business use cases for Gen AI, the majority of respondents highlighted five critical use cases, with 43% highlighting "Strategy, analysis, and planning (corporate planning, risk management, finance)" as their leading use case, followed closely with 40% choosing "Feature for customers within the product" as their leading use case. 38% of respondents chose "External chatbot/automation to handle low-level tasks (customer support, sales, etc.)" as a key use case with 37% flagging "Content generation (sales, marketing, HR, etc.)" as a critical use case. Closing the top-five use case list was "Content recommendation/generation engine for enabling talent (customer support or sales representatives)" with 32% of AI leaders ranking it as top priority:

Top Generative AI Business Use Cases

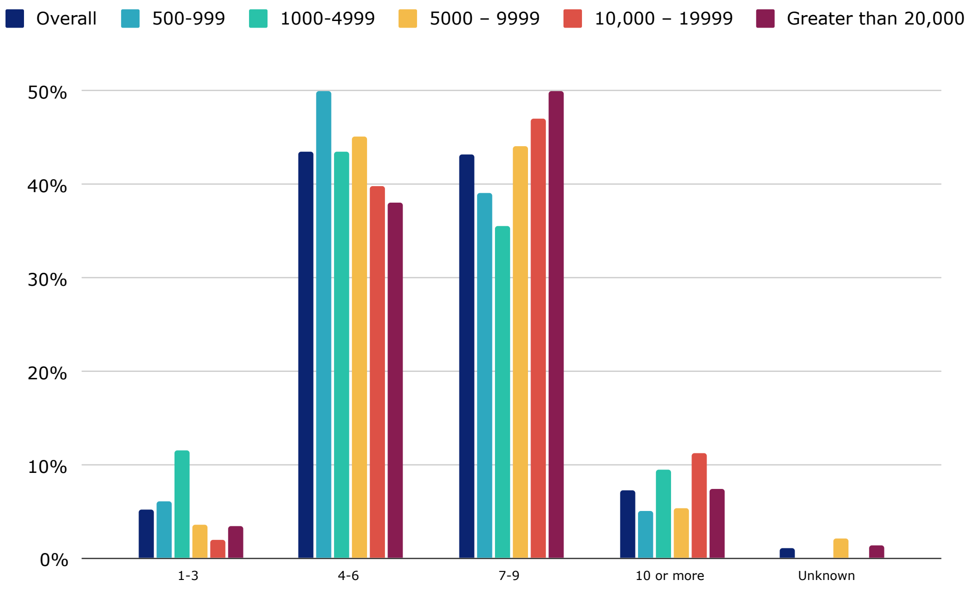


With the top use cases identified, we wanted to get a sense of just how many Gen AI use cases respondents had identified and planned to address in the next 18 months. The vast majority of mid-market, enterprise, and large enterprises indicated they'll be investing in multiple use cases within their organization, with 93% of organizations investing in 4-10 (or more) different use cases across the business.

Total Number of Use Cases by Geography

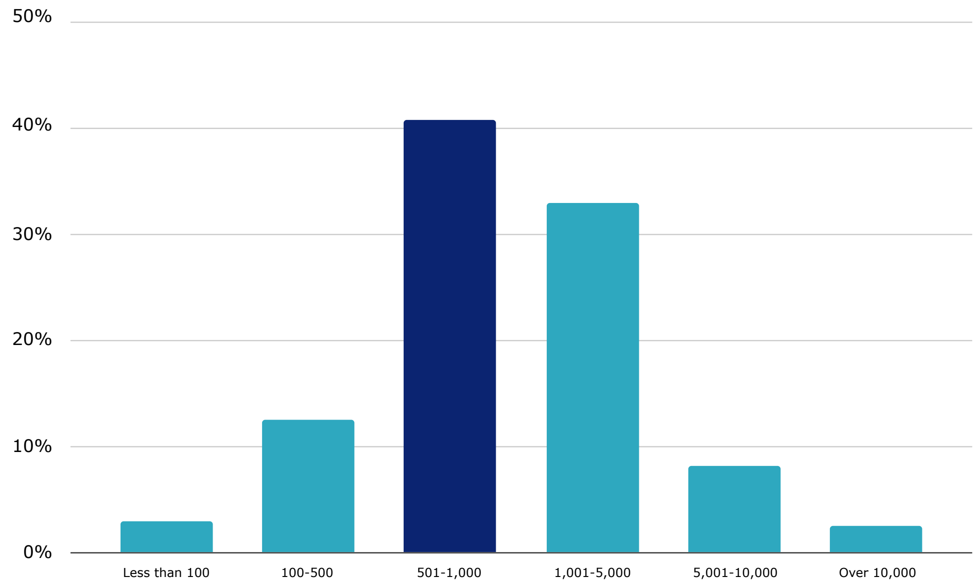


Total Number of Use Cases by Company Size



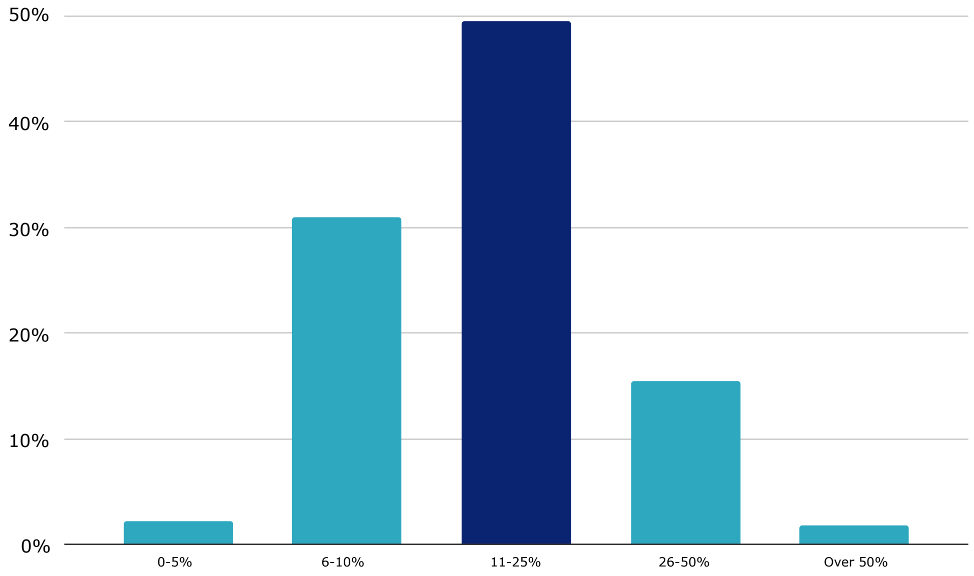
That being said, we asked how many users will need access to the Generative AI model per business use case, and an astounding 85% reported they expect between 501 and 10,000+ users will need access to a Generative AI model within their respective use case or multiple use cases.

Number of Users Needing Access to Gen AI Models



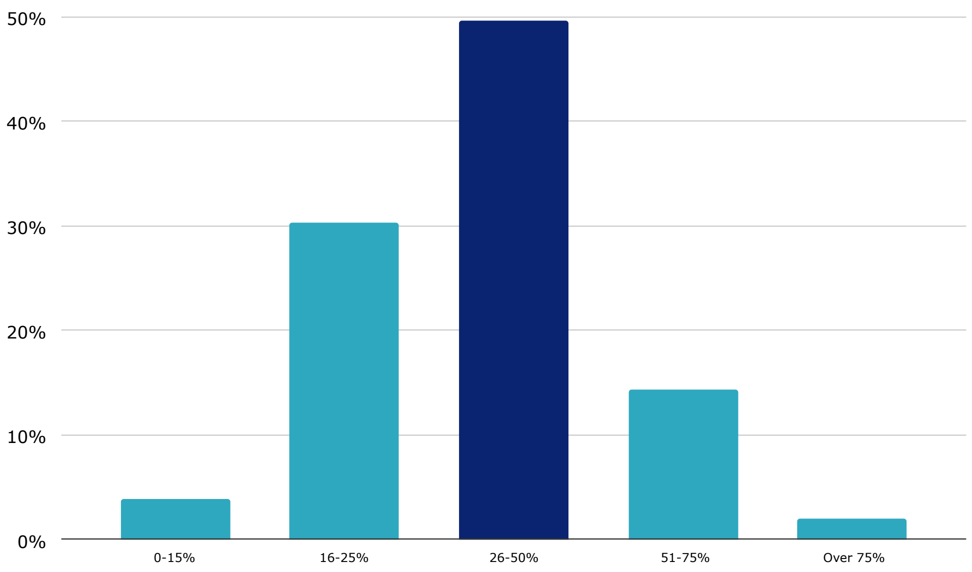
Given the extensive number of users needing access to Gen AI models, it was imperative to understand what percentage of employees were expected to use Gen AI in the first year of testing preliminary business use cases and rollout. It's clear that AI business leaders have an optimistic outlook of Generative AI adoption across their workforce, with 50% of leaders indicating 11-25% of employees, and an additional 18% of leaders reporting that 26-50% or more of employees, will use Generative AI in the first year of testing or rollout.

Percentage of Employees Using Gen AI - First Year



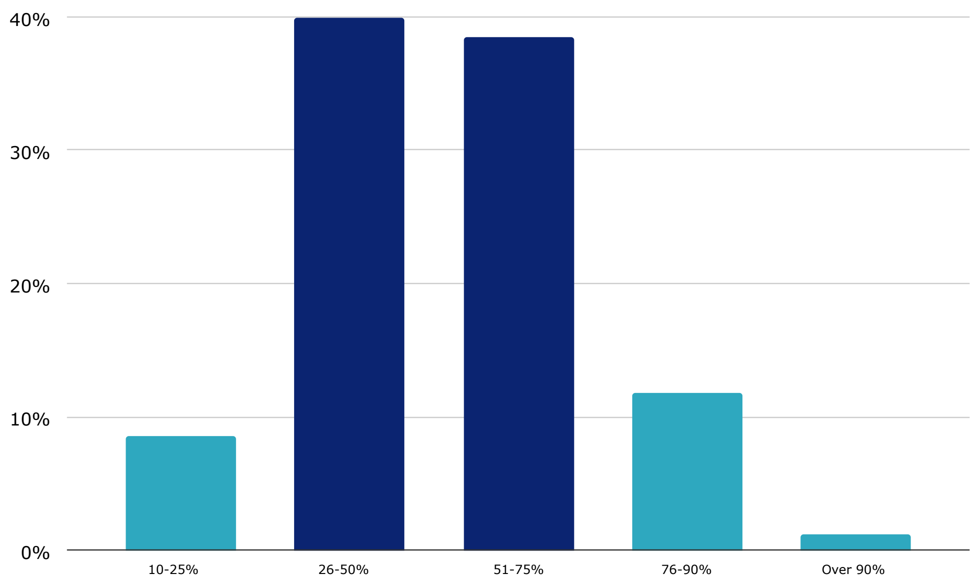
Now that we know what percentage of employees are expected to use Gen AI in its first year of testing and rollout, we wanted to know the percentage in year two after rollout. AI and ML leaders see additional cross-organizational adoption in the second year after initial adoption, with 30% or respondents indicating that 16-25% of employees are expected to use Gen AI, with 50% of respondents stating 26-50% of employees will use GenAI – and an additional 14% reporting they expect an astounding 51-75% of employees to use Gen AI as part of their day-to-day in the second year.

Percentage of Employees Using Gen AI - Second Year



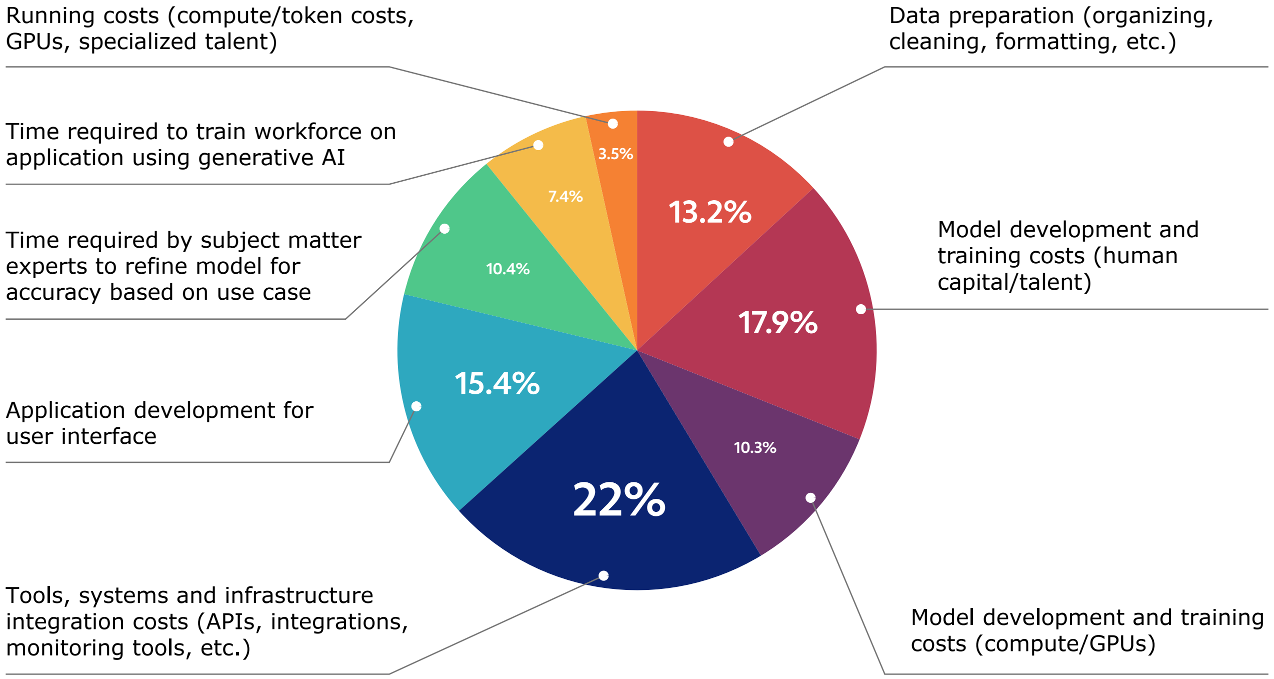
As a follow up to AI leaders' forecast on internal employee adoption of Gen AI, we asked them what percentage of their employees they expected would eventually use this technology. It's plain that respondents foresee vast Generative AI adoption across use cases, departments, and business units over time, with 40% of respondents reporting that they anticipate 26-50% of their workforce eventually using Generative AI in their discipline and an additional 39% sharing that they expect 51-75% of their entire workforce using Generative AI, and lastly 12% indicating a staggering 76%-90% of their employees will be using Gen AI eventually.

Ultimate Employee Adoption of Gen AI



Next, we asked which Generative AI cost drivers have respondents considered for developing, deploying, maintaining, and running generative AI in their enterprise as part of total cost of ownership (TCO). Based on survey answers, we found that most respondents believe their Gen AI costs are centered around model development, training and systems infrastructure. For example, the costs associated with how a model works – human capital, the tools and systems to run it, and the app/UI for users. 59% of executives overseeing AI reported that tools, systems, and infrastructure integration costs (APIs, integrations, monitoring tools, etc.) are the top AI cost drivers. 48% reported that model development and training costs (human capital/ talent) are a top cost driver, and 42% indicated application development for user interface as a top cost driver.

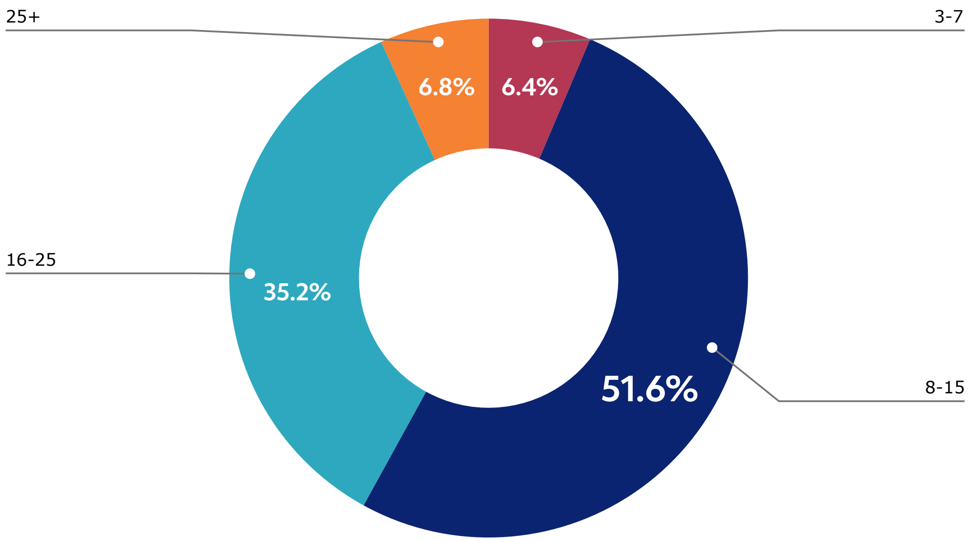
Key Gen AI Cost Drivers



This is an excellent example of the gap between a company's vision and reality. We believe respondents are underestimating how messy data can be, and the heavy lifting needed for data prep as well as underestimating the cost of vast usage by users at a company-wide scale. It's worth noting that this is even more challenging if their company is using AI as a Service. Similarly, respondents are underestimating the time required by SMEs to work with the team to ensure the model is accurate and "good enough" to roll out either internally or externally. Most importantly, a shockingly low 8% of respondents said they would attempt to control their budget by limiting models and/or access to Gen AI to better manage their budgets, which means they are not thinking about running costs, which we expect is going to be a huge surprise for them. Moreover, our previous survey found that although AI and ML adoption is now a key revenue and ingenuity engine within the enterprise, an astonishing 59% of C-level leaders are inadequately resourced to deliver on business leadership's expectations of Generative AI innovation. They lack the budget and resources needed to drive adoption successfully across the enterprise and create value. Clearly, something's got to give.

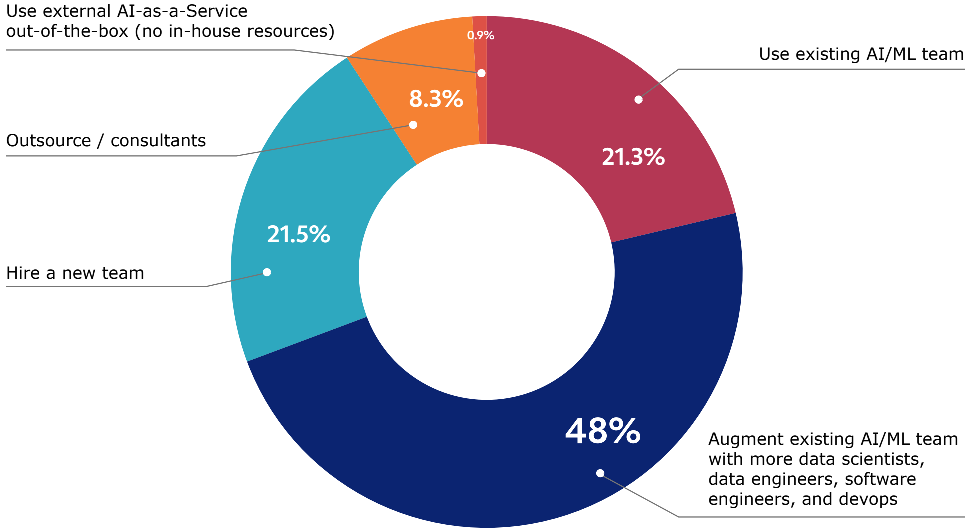
When asked about the size of their current team overseeing Generative AI (data engineers, data scientists, AI engineers, etc.), the results tell an interesting story. The majority of respondents (52%) indicated that their AI/ML team size overseeing Generative AI ranges from 7 to 15 team members, with an additional 42% indicating their teams are larger than 16. This data certainly correlates with the company size breakdown, where mid-market & SMB organizations with 500-1000 employees report team size to be up to 15 employees and larger enterprises and F1000 companies reporting a much larger team size, with more than 16 team members.

Team Size Overseeing Gen AI



When asked how they plan to resource and staff future Gen AI efforts, a clear preference emerged amongst AI leaders to keep Generative AI as an internal effort. Almost half of respondents (48%) are choosing to augment their existing AI/ML team with more data scientists, data engineers, software engineers, and devops to support their Generative AI efforts. 21% of respondents indicated they'll use their existing AI/ML team (which means finding more ways to scale themselves efficiently to do more with less -- or just produce fewer models) and an almost even number (22%) said they will hire a new team to resource and staff their future Generative AI efforts. Only 8% mentioned they will consider outsourcing it to consulting firms or consultants and just 1% selected using external AI-as-a-Service out-of-the-box (no in-house resources) as part of their plan to staff future Gen AI efforts.

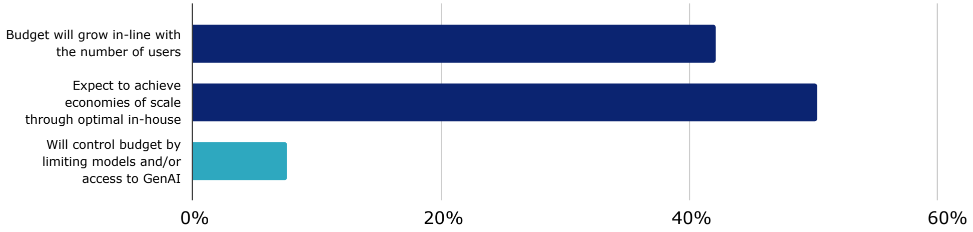
Plans to Resource and Staff Future Gen AI Efforts



Given that nearly every respondent (91%) plans to resource or staff in-house to support future Gen AI efforts, that's bad news for consultants who are preparing for such projects to constitute a large part of their business, but it does lead us to believe that organizations are considering scaling Gen AI for the long haul. However, that requires some serious cost considerations for how they are budgeting going forward and how to be most efficient using their budgets year-over-year. They may well be overestimating how much they can do with their budget.

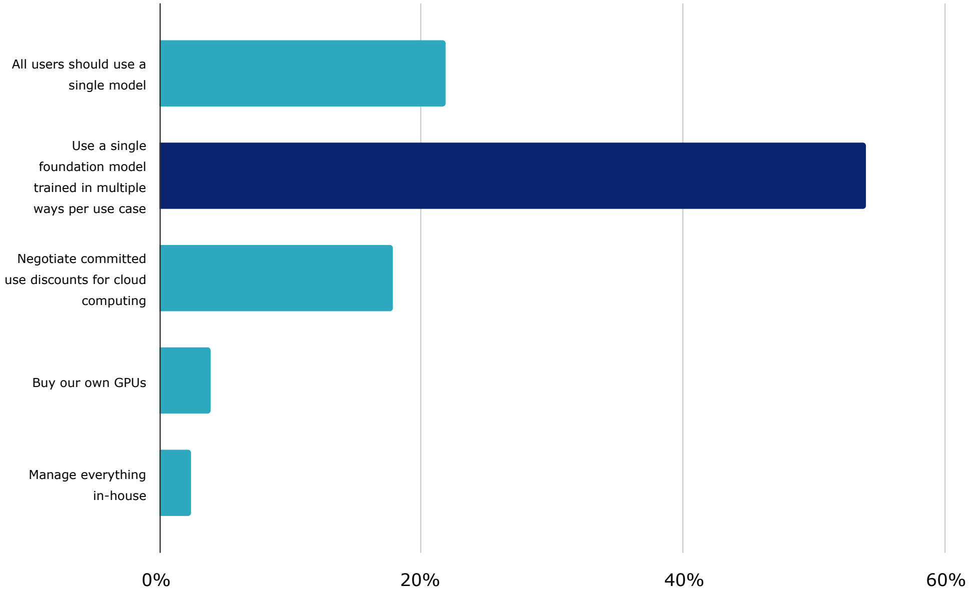
When respondents were asked how they anticipate managing their Gen AI budget to scale across use cases, business units, and departments, the results were almost evenly split between two of the four strategies, with 50% highlighting they expect to achieve economies of scale through optimal in-house or cloud GPU usage, and an additional 42% indicating their Generative AI budget will grow in-line with the number of users as they scale over time or add additional use cases to the mix. Only 8% said they will attempt to control their budget by limiting models and/or access to Gen AI to better manage their budgets.

Strategies to Manage Budget



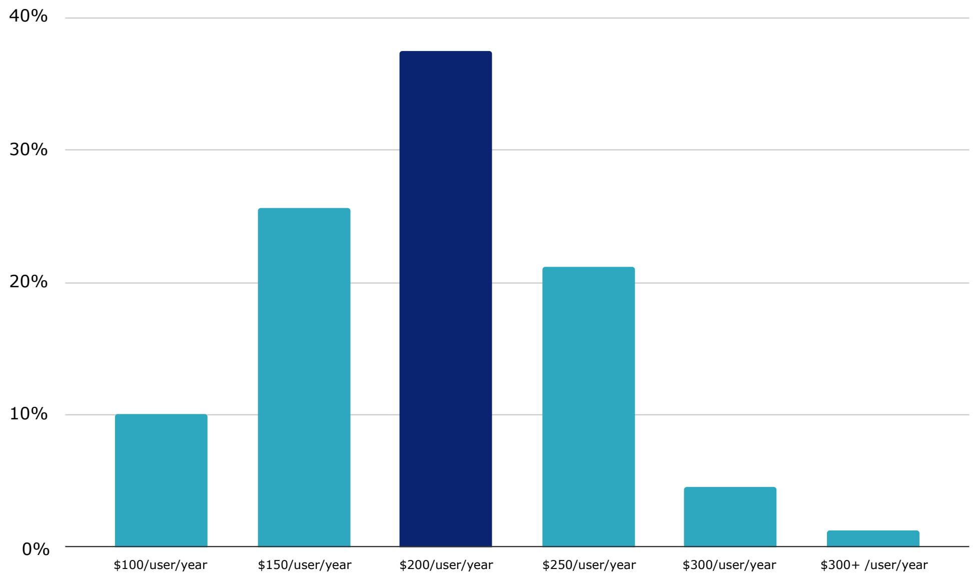
AI leaders were asked to identify where they would expect to see economies of scale when adopting Generative AI in their organizations. The majority of respondents (54%) said that using a single foundation model trained in multiple ways per use case would be the most successful factor driving economies of scale.

Factors Driving Economies of Scale



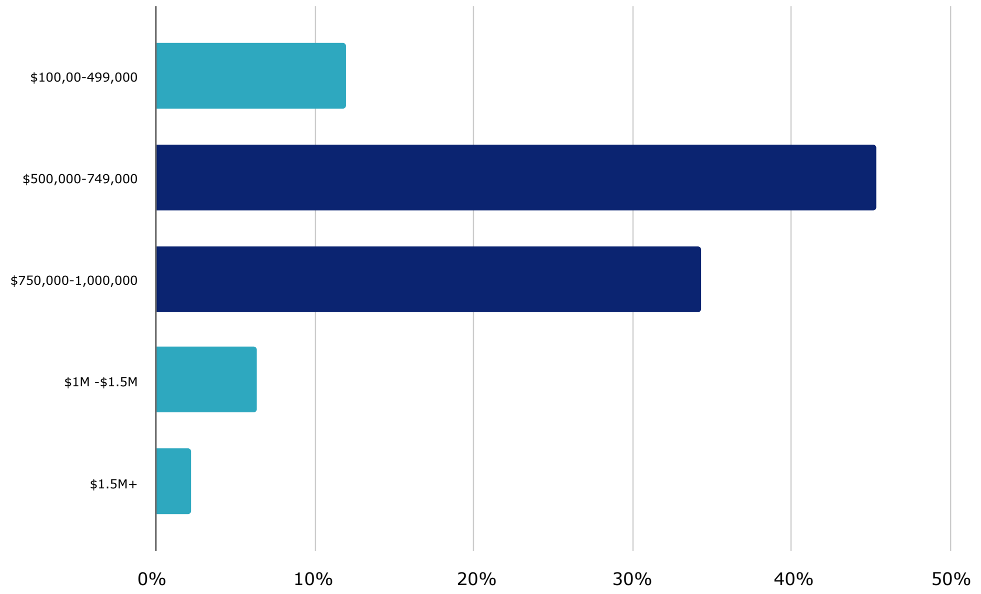
We wanted to dive deep into the key considerations and hidden costs of adopting Gen AI at scale across businesses and enterprises, and asked survey participants how much they are willing to pay per user per year to develop, deploy, and maintain Gen AI in the organization. We wanted to better understand organizations' appetite to fund Gen AI initiatives per user in a typical SaaS model. The majority of respondents (65%) said they were willing to pay $200/user/year or more – with 27% willing to pay $250+ per user per year.

Price Per User



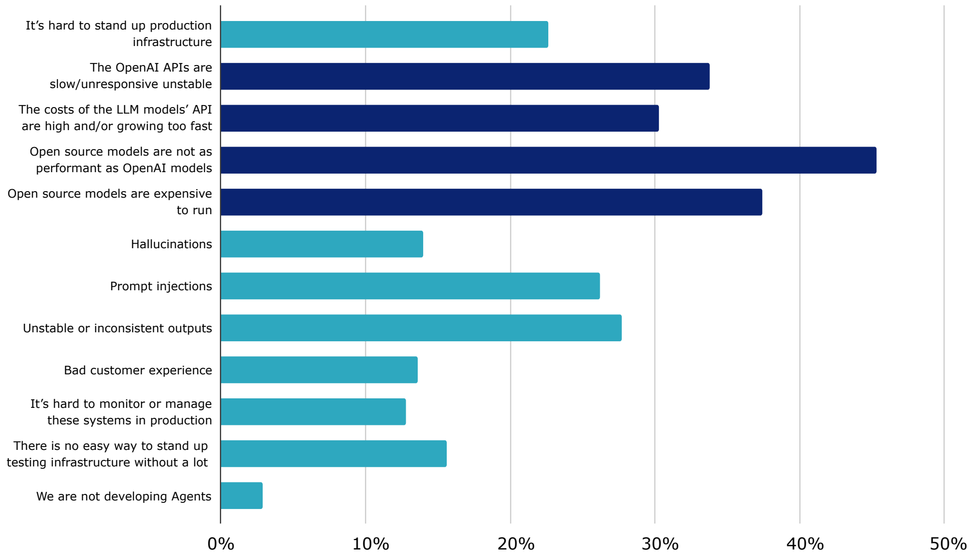
Survey participants were also asked about their willingness to pay for Generative AI platforms. When asked how much they would pay per year for a platform that allows their AI team to manage all of the organization's Generative AI models company-wide while optimizing GPU compute costs, the vast majority of respondents said they were willing to pay between $500,000-$1,000,000 per year for such a platform. In our previous survey, "Enterprise Generative AI Adoption: C-Level Key Considerations, Challenges, and Strategies for Unleashing AI at Scale," 68% of C-level, Fortune 1000 respondents overseeing Gen AI enterprise adoption indicated that given the latest advancements and release of Generative AI and LLM platforms, they believe the importance of creating value from their AI investments is greater compared to last year. 57% of respondents reported that their board expected a double-digit increase in revenue from AI/ML investments in the coming fiscal year and an additional 37% reporting the expectation of a single-digit increase.

Gen AI Platform Investment



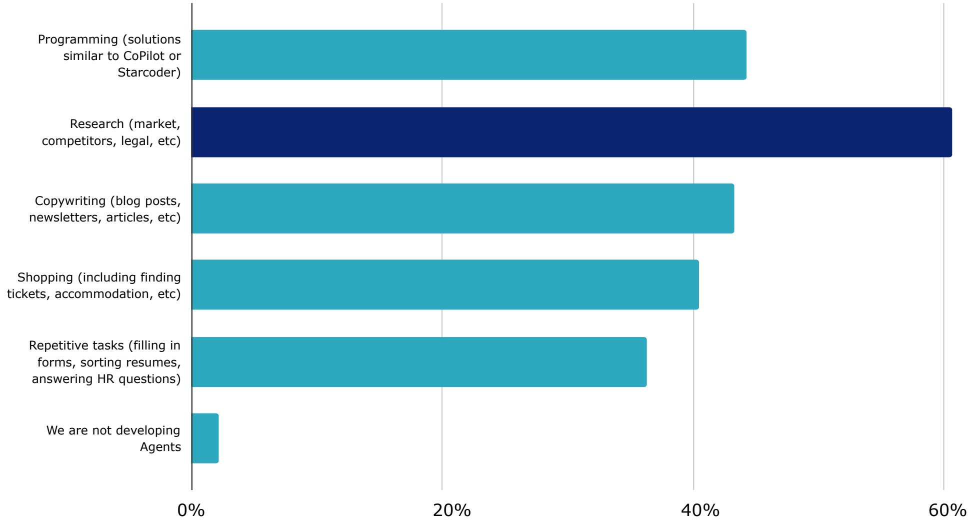
When asked about the challenges faced by respondents developing Generative AI applications (or agents), responses were split, with a clear top-4 responses. 45% of respondents indicated that they will face a challenge with Open Source models that are not as performant as OpenAI models. 37% of respondents said Open Source models are expensive to run, while 34% of AI leaders pointed to the fact that OpenAI APIs are slow/unresponsive/unstable, and 30% flagging that the costs of the LLM models' API are high and/or growing too fast. Respondents understand these critical issues are not going away, but rather believe they can be solved with agents and automation. While those may help contain costs initially, it will be challenging for businesses to maintain them for the ever-changing AI as a Service models. The cost of maintaining their agents and automation across multiple use cases as time goes on will also increase.

Challenges in Developing Gen AI Apps/Agents



Finally, when asked if they were developing a Gen AI application (or agent), which areas will respondents apply them to, the majority of survey respondents voiced that research (market, competitors, legal, etc.) is the top area to target Gen AI applications, with 44% highlighting programming (solutions similar to Microsoft 365™ Copilot or StarCoder from Hugging Face™ and ServiceNow™ Research) as the leading factor. 43% indicated copywriting (blog posts, newsletters, articles, etc) as their leading consideration.

Areas for Gen AI Apps/Agents



CONCLUSION

As we've seen from these global survey results , while most organizations are planning and budgeting for Gen AI, their vision may not be aligned with the amount of money that is actually needed for success. We believe that will result in collisions between expectations and reality when the bills come due, and that many organizations may well be shocked at the actual total cost of ownership of successfully deploying Generative AI at their organizations at scale.

While it's clear that companies are onboard with the potential of Gen AI and are planning multiple use cases as well as looking to roll out the technology to employees and customers alike, what remains to be seen is how successful that adoption will be and what price tag they are actually forecasting, and then willing and able to pay to drive Generative AI ingenuity.

Right now, as the hype cycle around Gen AI continues to escalate, C-level executives and AI leaders are faced with a vast array of choices and are weighing their options. It seems like every single vendor is claiming to have Gen AI baked into their platform, leading to well-deserved skepticism amongst the hype when claims are put to the test against true ROI and TCO as well as productivity, time to market (TTM) and efficiency multipliers.

AI leaders are right to be cautious, but we urge readers not to be discouraged. While it might seem like the Wild West right now, this hyper-inflated, over-the-top phase will subside. By considering and taking into account the various cost drivers of Generative AI business adoption, leaders can be confident in accurately predicting and forecasting the TCO for Gen AI in their organization. When budgeting and planning for Gen AI business adoption, leaders should consider the multiple factors that impact their total cost of ownership, such as: setup, training, maintenance, running costs, use cases and variable costs such as compute.

What's the right level of customization, and what's needed from a data or knowledge protection standpoint? What are the right use cases and the right model(s) to support them? Organizations must decide that for themselves, and we are hopeful the data, metrics, and insights contained in this report has illuminated the thinking behind many of the decisions that leaders must make today, and in the future.

NEXT STEPS

If you are a commercial leader trying to unlock value and drive sustainable growth in today's analytical, data-driven business environment, consider the use of an open source LLM platform like ClearML's ClearGPT – the only secure, enterprise-grade generative AI platform that removes the blockers and risks of using LLMs to fuel innovation. To learn more and request a demo, please visit https://cleargpt.ai.

If you're a Gen AI company looking to connect with enterprises and you want to support the release of clear, unbiased information about AI and the AI market, then please contact the AIIA at infra@ai-infrastructure.org to become a partner.

About AIIA

The AI Infrastructure Alliance is dedicated to bringing together the essential building blocks for the Artificial Intelligence applications of today and tomorrow. The Alliance and its members bring striking clarity to this quickly developing field by highlighting the strongest platforms and showing how different components of a complete enterprise machine-learning stack can and should interoperate. They deliver essential reports and research, virtual events packed with fantastic speakers, and visual graphics that make sense of an ever-changing landscape. To learn more, visit https://ai-infrastructure.org/.

About ClearML

ClearML is used by more than 1,300 enterprise customers to develop a highly repeatable process for their end-to-end AI model lifecycle, from product feature exploration to model deployment and monitoring in production. Use all of our modules for a complete ecosystem or plug in and play with the tools you have. ClearML is trusted by more than 150,000 forward-thinking Data Scientists, Data Engineers, ML Engineers, DevOps, Product Managers and business unit decision makers at leading Fortune 500 companies, enterprises, academia, and innovative start-ups worldwide. To learn more, visit the company's website at https://clear.ml.

Unleashing AI
in the Enterprise