State of AI in the Cloud 2024

Our research shows that AI is taking over the cloud: 70% of organizations are using managed AI services, making them already nearly as popular as managed Kubernetes (!).

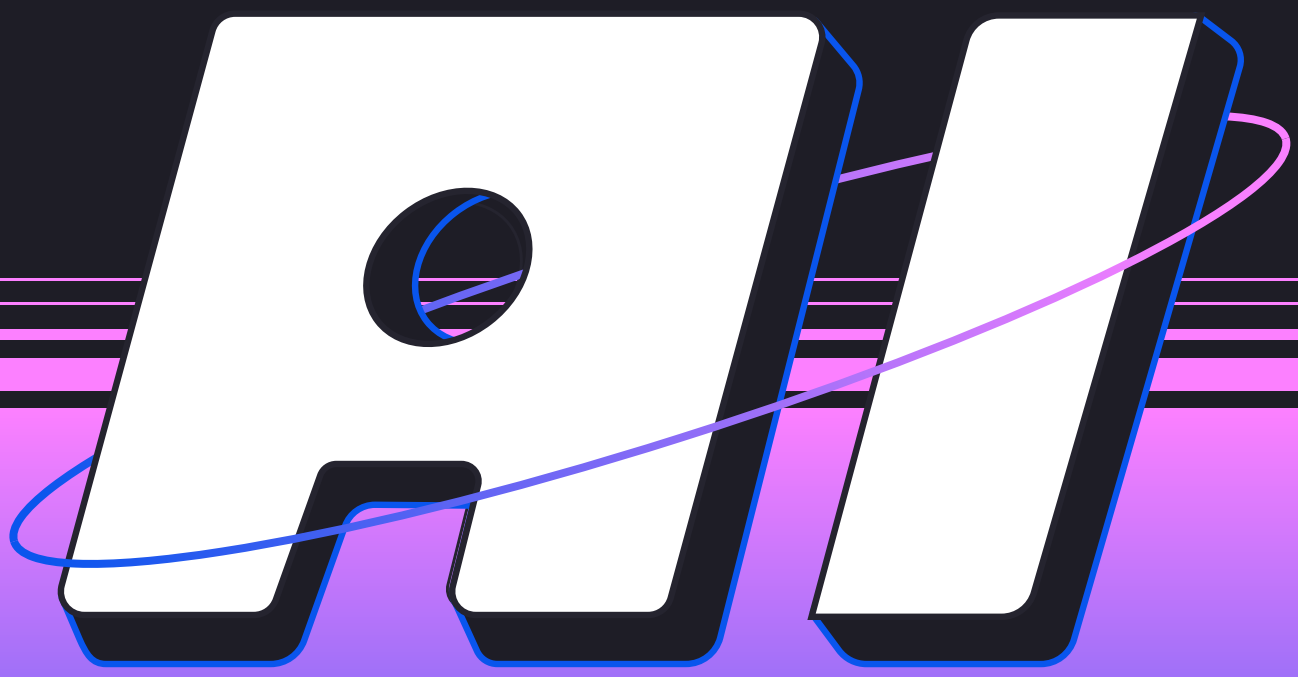


Table of Contents

Executive Summary

In this data-driven report, the Wiz research team explores the explosive adoption of managed AI services and self-hosted AI tools in cloud environments, based on a sample size of over 150,000 public cloud accounts. The report highlights the incredibly fast and widespread adoption of generative AI among cloud customers.

Over 70% of cloud environments now employ managed AI services. Microsoft, with its Azure AI Services (which includes Azure OpenAI), is leading in total AI deployment among the major Cloud Service Providers (CSPs). Moreover, Azure OpenAI usage grew by 228% in a 4-month period in 2023. It's interesting to note that while adoption of these services is high, a substantial portion of organizations (32%) still appear to be in the experimentation phase, deploying fewer than 10 instances of AI services in their cloud environments. This could indicate that most cloud customers aren't quite ready to deploy these services at scale in their production environments.

The recent rise of generative AI tools presents new challenges for cloud customers. As in the early days of cloud computing, development and adoption of AI technology often occurs without using fully established standards and governance. This is already raising concerns around the security and management of AI implementation. Some instances, such as Microsoft's accidental exposure of 38 terabytes of AI data, underscore the potential security risks associated with rapid adoption of AI technologies.

Looking forward, the cost of AI model training and inference, and the need for efficient and secure utilization of AI services, are becoming crucial considerations for businesses. Wiz recommends that organizations strive to improve visibility into their AI service usage and foster a culture of security ownership across teams, particularly to address the fast-emerging risk of "Shadow AI". Security teams should collaborate closely with developers, cloud engineers, data scientists and other AI practitioners to manage the attack surface introduced by AI tools into cloud environments.

Introduction

Over the past year and a half, generative AI has seen explosive growth among both end-users and businesses. While "traditional" AI and machine learning has been integrated into both scientific and commercial endeavors for many years now, generative AI and large language models (LLMs) in particular have made this technology a household name.

The trend began in July 2022 with the beta releases of image generation services including Midjourney and OpenAI's DALL-E 2, with other open-source models soon to follow, chiefly Stability.ai's Stable Diffusion in August 2022. The trend became undeniable with the release of text generation services, namely OpenAI's ChatGPT in November 2022.

Since then, both commercial and open-source generative AI models have seen widespread adoption by enthusiasts, established companies and new startups building their products around generative AI. This has also led to an emerging community of AI builders, taking part in the development effort of services and tools that enable training, finetuning, trimming, managing, and deploying AI models for everyday usage.

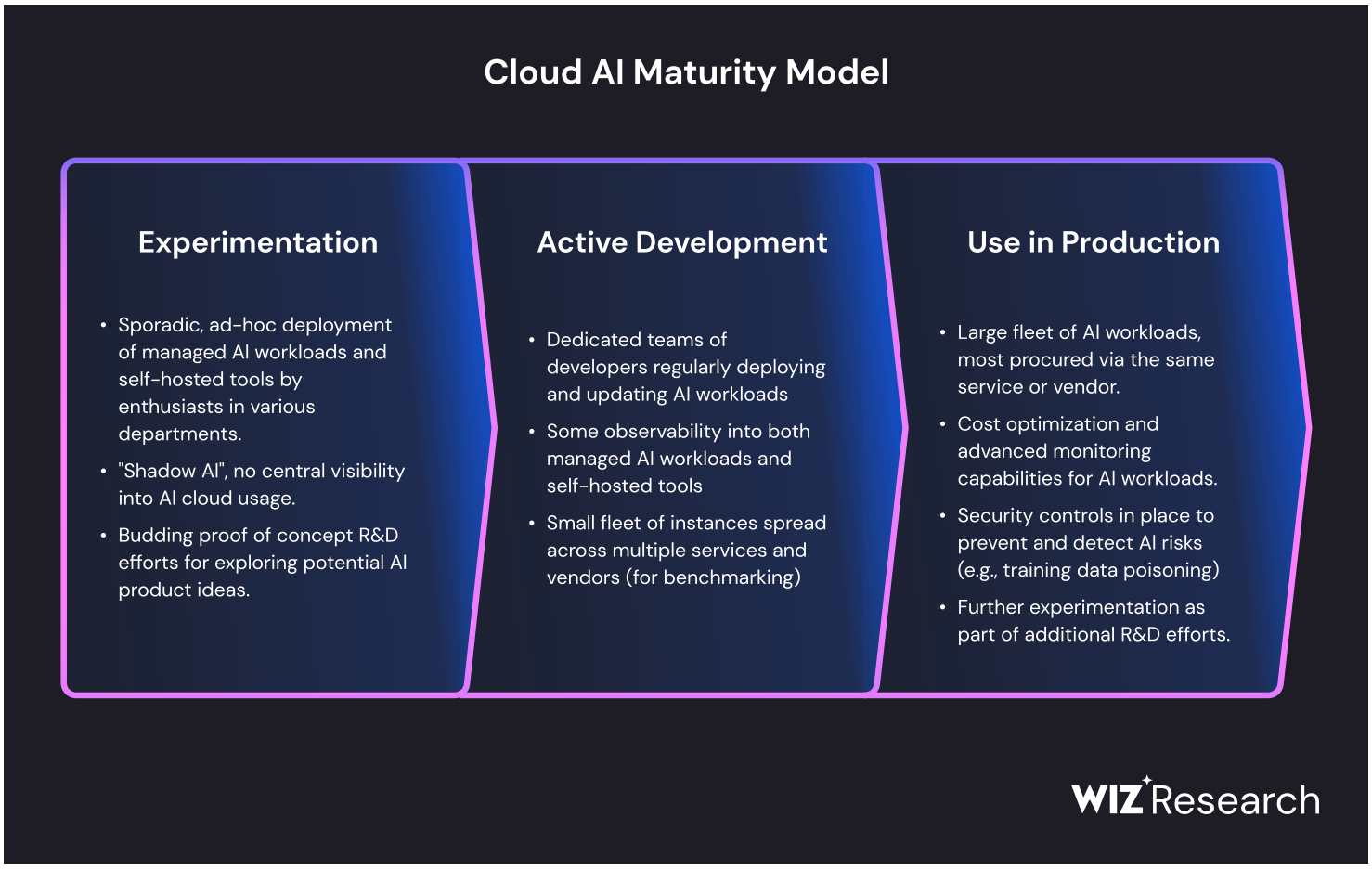
In the evolving technology landscape, generative AI appears to have become many things: a promising arena for innovation, a controversial subject inspiring a broad range of opinions, and a source of concern among consumers speculating about the changes it will bring to daily life. Regardless, organizations across industries are now leveraging both "traditional" and generative AI to enhance efficiency, automate processes, and gain a competitive edge.

But this surge in AI adoption recalls the challenges faced during the early stages of cloud computing — a rapid uptake of new technology without the necessary standards or governance in place. Generative AI is a truly cloud native technology, with training and inference remaining highly compute-intensive for most use-cases, and therefore encouraging organizations to take full advantage of the computation and storage scaling that the cloud has to offer.

Cloud service providers have stepped up to meet customer demand by offering both revamped versions of their existing AI and ML services as well as entirely new solutions for building with generative AI in the cloud. Azure recently released AI Studio in public preview and continues to invest in developing its AI Services (AKA Cognitive Services), a set of tools which among other things also houses Azure's integrated OpenAI offering, made generally available since January 2023; GCP added generative AI capabilities to Vertex AI in March 2023, while continuing to maintain its more ML-focused AI Platform as well; and AWS has been building new features into SageMaker while making Bedrock generally available in September 2023.

As with any technology that is so rapidly adopted, potential security drawbacks are beginning to show themselves, as recently demonstrated by the accidental exposure of 38 terabytes of AI data by Microsoft. Having said that, we're excited to see how cloud customers will leverage generative AI, and we're committed to helping them do so while introducing the correct security guardrails. Our research shows that as of the beginning of 2024, the adoption of both traditional and generative AI by cloud customers has reached unprecedented heights and shows no signs of slowing.

Organizations who choose to invest in AI as part of their technology portfolio will most likely go through several phases, as outlined in Gartner's AI Maturity Model or the COE's AI Guide. When examining how these developments might impact a company's cloud environment, we can generally define the following maturity model, showing how a company might go from sporadic experimentation by enthusiasts, through active product development, to deployment of production-ready AI features:



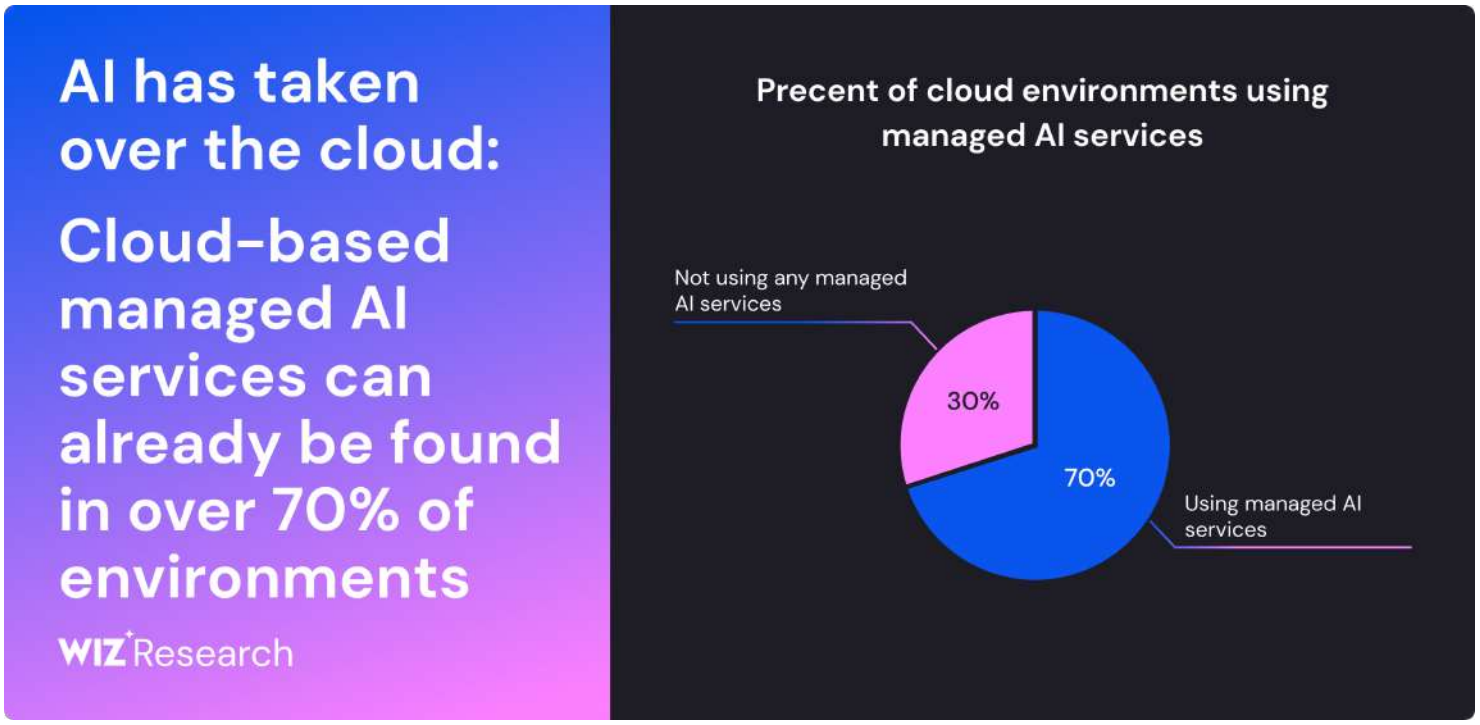
In this data-driven report, we'll examine the explosive adoption of AI services and tools by cloud customers across the three major cloud service providers (CSPs). We'll also consider what all this might mean for cloud customers as they face the dual prospects of rising costs and new security considerations associated with these new services.

This report includes insights from the Wiz Research team based on data collected throughout 2023. Over this period of time, a sample of over 150,000 cloud accounts were analyzed.

Key Findings:

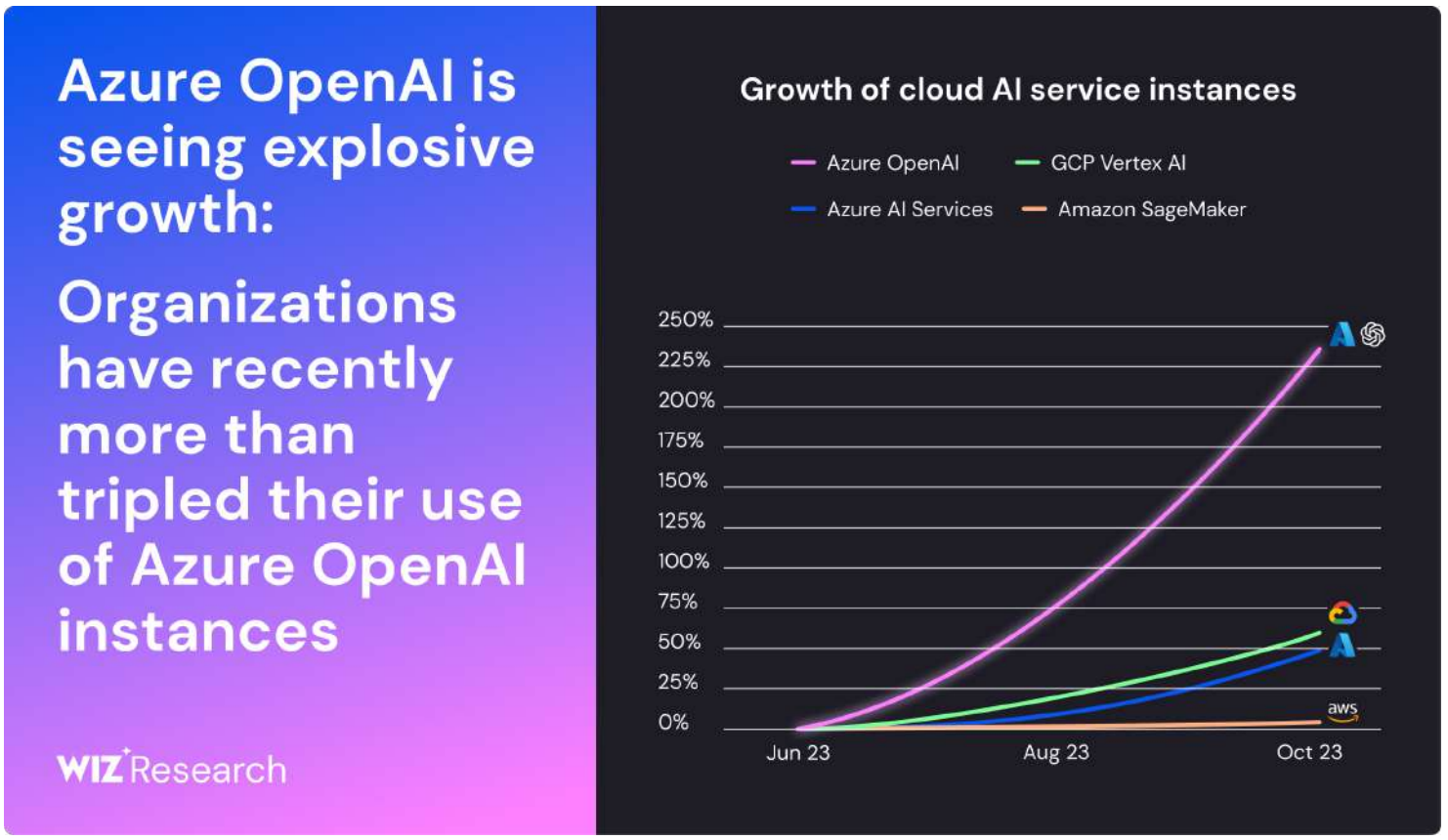
1 AI has taken over the cloud: Cloud-based managed AI services can already be found in over 70% of environments.

- Managed AI services offered by cloud service providers — such as Amazon SageMaker, Azure AI Services and GCP Vertex AI — can be found in over 70% of all cloud environments, meaning that products
  and services utilizing the most recent advances in AI are being adopted by organizations at an incredibly fast rate.
- For comparison, managed Kubernetes services offered by cloud providers (including EKS, AKS and GKE) are found on average in 81% of cloud environments. AI services are much newer than managed Kubernetes, but this data point shows that they are already almost as commonly used. To us, this indicates that cloud-based AI services are seeing an unusually high rate of adoption relative to other services.
- As mentioned earlier in this report, the spike in AI echoes the early days of cloud: massive adoption is happening, but processes and governance seem to be lagging behind.



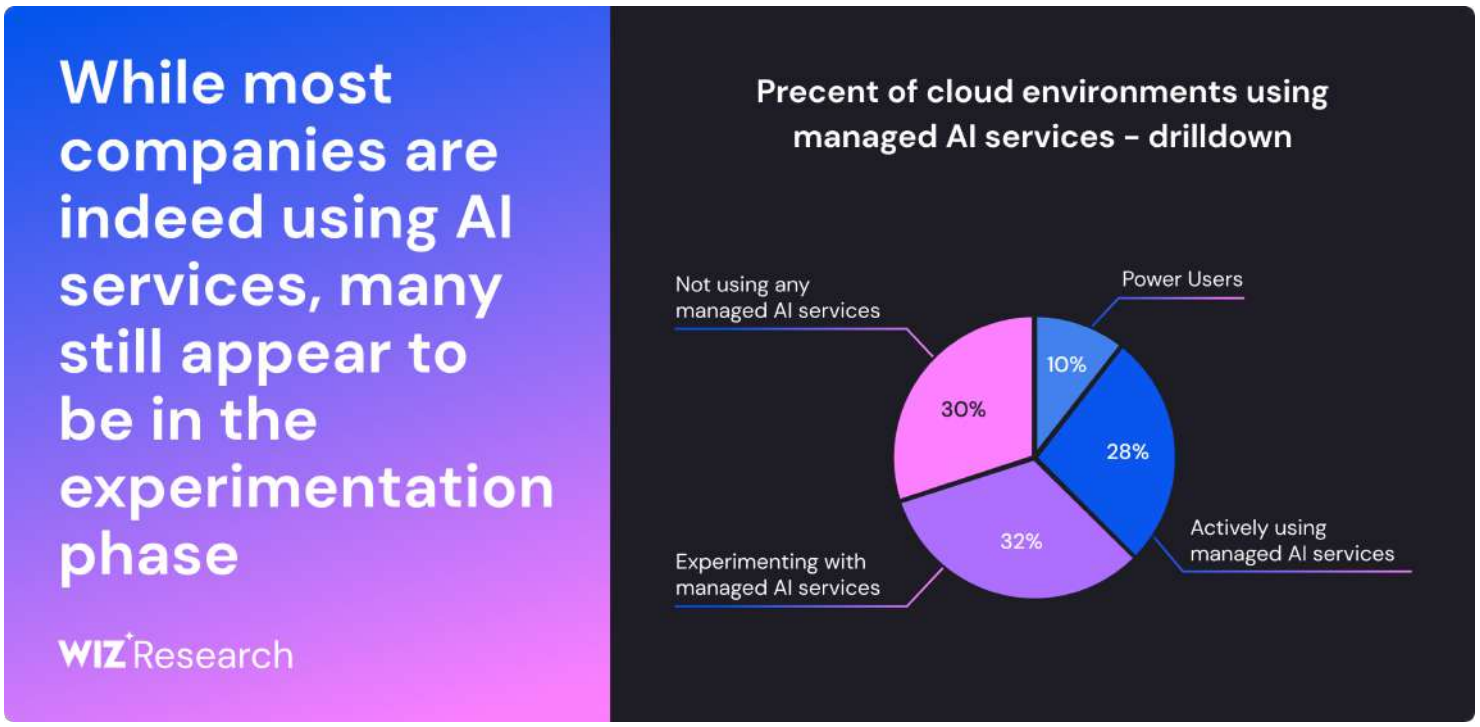
2 Azure OpenAI is seeing explosive growth; organizations have recently more than tripled their use of Azure OpenAI instances.

- Over a 4-month period between June and October 2023, the total number of Azure OpenAI instances observed across all cloud environments grew by a whopping 228% (with a ~40% month-over-month average). For comparison, the average instance growth in the same period for most other Azure AI Services (such as Text Analytics and Bing Custom Search) was only 13%.
- Considering that Azure OpenAI debuted in November 2021, was made generally available in January 2023, and likely garnered increased interest following the release of OpenAI's ChatGPT in November 2022, this rate of growth seems to indicate high customer attraction.
- While comparing between services offered by different CSPs is not an exact science, since each service might have slightly different and overlapping intended use-cases, our data shows that the number of Vertex AI instances grew by 45% (with a ~10% month-over-month average) during the same period.
- Since Amazon Bedrock was only made generally available in September 2023, we did not include it in this comparison.



3 While most companies are indeed using AI services, many still appear to be in the experimentation phase.

- As mentioned above, 70% of organizations are using managed AI services to some degree. However, around 32% have deployed less than 10 instances of these services in their environments, leading us to conclude that they are likely still testing the AI waters before they decide to invest more resources in their usage of this technology.
- While we don't know how many instances are devoted to experimentation as opposed to development or production purposes, or precisely what percent of cloud spend is devoted to managed AI services, we can state that at least 28% of organizations appear to be doing more than just experimenting — they are active in their deployment of AI instances, with 10 or more instances in their environment. In addition to those actively deploying AI services, another 10% of organizations could be considered "power users", with 50 or more instances in their environment.
- For this analysis, we're working under the assumption that the number of instances of a given service in any cloud environment positively correlates with actual usage of the service. Having more live instances is also likely to incur more costs, which is expected to drive organizations to minimize the number of instances while balancing the costs against the benefits of utilizing these services.
- One might wonder why the number of instances seems to be so low per customer, especially in comparison to other service types such as managed databases. We can identify two possible contributing factors here:
    1. First, considering the relatively high cost of these services (particularly training and finetuning, which can cost as much as 100$ per hour), we should probably expect the number of instances per cloud environment to remain relatively low in comparison to other comparable managed services.
    2. Second, at the moment, some CSPs are enforcing strict quotas and limiting the number of AI service instances deployable per customer (with some instance types also limited to specific regions).
- As efficiency rises and costs gradually go down, and as CSPs begin to remove limitations, we expect to see an eventual increase in the total number of managed AI service instances per cloud environment.

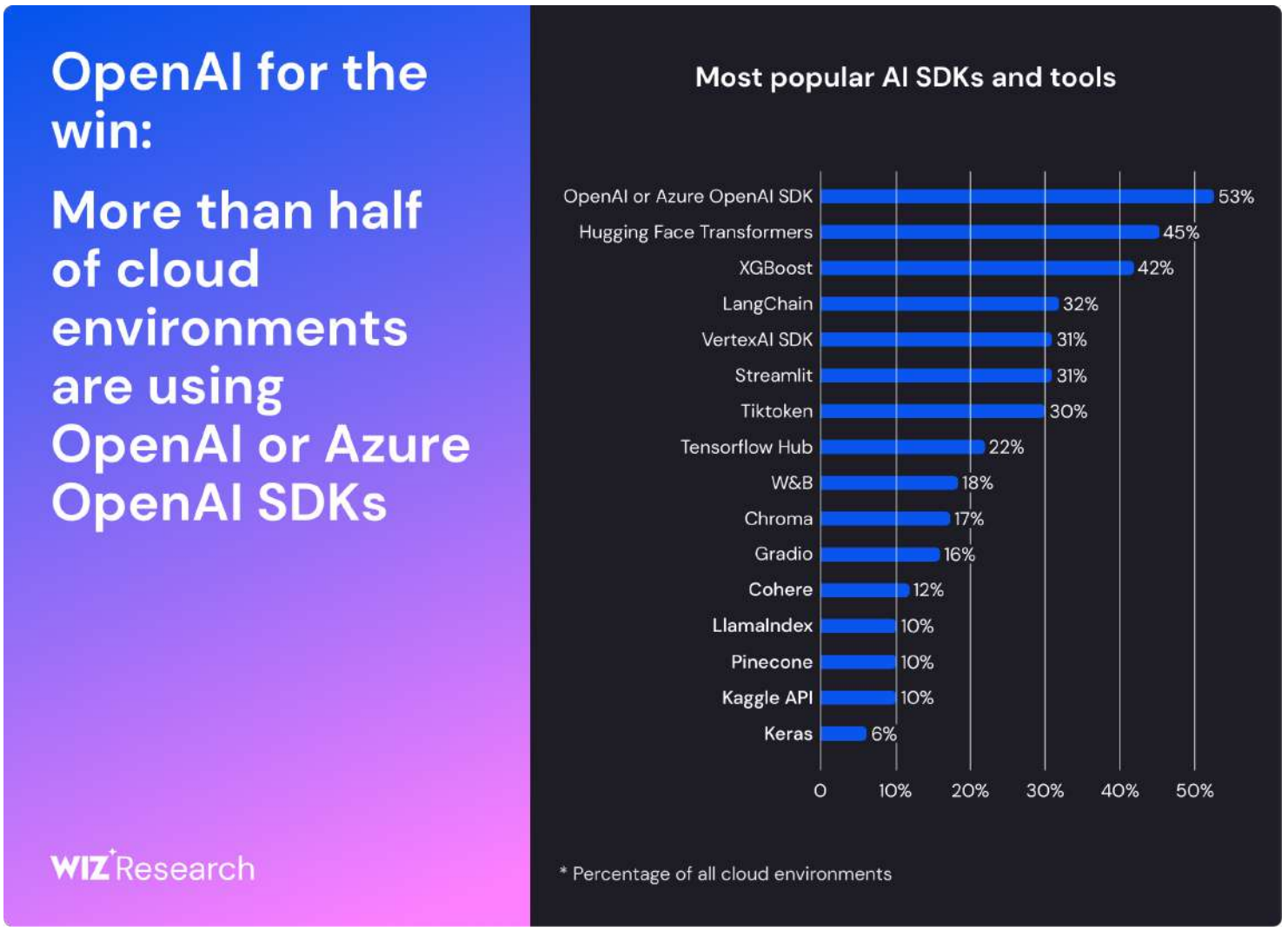


4 When it comes to the use of CSP-managed AI services in the cloud, Azure leads the pack.

- Microsoft appears to be leading the pack among CSP-managed services with use of AI in the cloud: 70% of Azure environments include Azure AI Service instances (equivalent to 39% of all cloud environments). Notably, 54% of Azure environments include managed Azure OpenAI instances in their environments (30% of all cloud environments). Note that this last data point doesn't account for organizations that might be implementing their own self-hosted or custom (non-managed) integrations with OpenAI.
- For comparison, 53% of AWS environments include SageMaker Notebook or SageMaker Domain instances (equivalent to 44% of all cloud environments), and 44% of GCP environments include Vertex AI or AI Platform instances (15% of all cloud environments).
- As mentioned above, since Amazon Bedrock was only made generally available in September 2023, we did not include it in this comparison. However, our latest data does show that at least 15% of organizations appear to be deploying instances of this service in their environment (based on activity analysis).

5 OpenAI for the win: More than half of cloud environments are using OpenAI or Azure OpenAI SDKs.

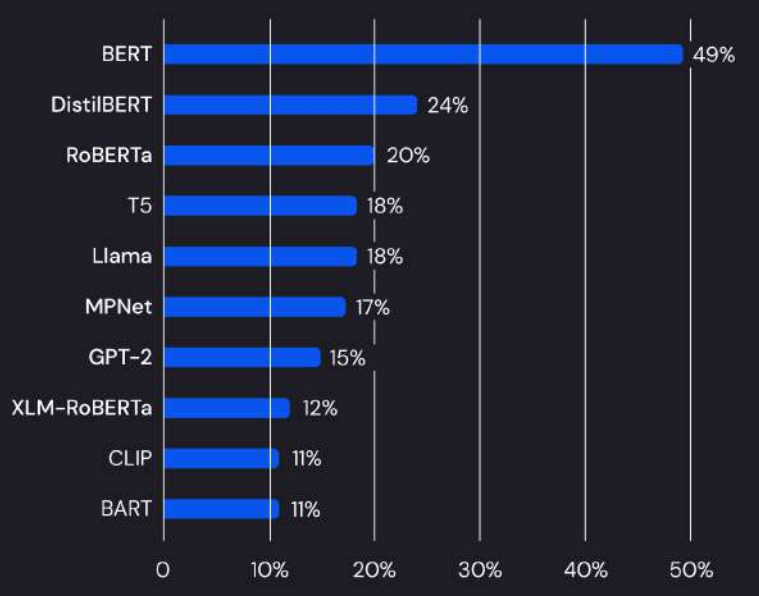
- About 53% of cloud environments are using OpenAI or Azure OpenAI SDKs (which allow integration with various OpenAI models such as GPT, DALL-E and Whisper), suggesting that generative AI tools are quickly becoming commonplace in cloud business models.
- In general, self-hosted AI and ML software is highly prevalent in the cloud and can be found in nearly 69% of environments. For example, Hugging Face Transformers can be found in 45% of environments, LangChain in 32%, and the Tensorflow Hub library in 22%.



- Based on our data, at least 42% of organizations have chosen to self-host AI models in their cloud environments (like those available from Hugging Face or trained internally). Among such organizations, the most popular identifiable model appears to be BERT, found in 49% of such environments (and 20% of all cloud environments).

While managed AI services are certainly popular, many organizations are also self-hosting AI models in the cloud

WIZ Research



Conclusions:

- Both managed AI services and self-hosted solutions are seeing fast and widespread adoption in cloud environments, with adoption rates now nearing those of managed Kubernetes clusters. We expect that even more cloud customers will be utilizing these services soon, and organizations will likely see continued growth in the number of instances deployed in their environment.
- Looking at these trends, we believe that the cost of training and inference, as well as the security aspects of utilizing AI services, are going to become critical business priorities for cloud customers in the coming year. This should include monitoring and optimizing the cost of developing and operating AI models as part of both internal and customer-facing applications, as well as building observability and robust security controls to manage what is essentially a brand-new attack surface, necessitating the adoption of new best practices.
- Security should certainly not be a blocker to AI innovation, but security teams may feel they have no other choice but to intervene, especially if the velocity of the business prevents them from being able to keep pace with the new risk factors that AI may introduce into their organizations' cloud environments. In other cases, security may not be a blocker at all, but possibly for the wrong reasons (e.g. security just isn't in the loop).
- Regardless, both managed and self-hosted AI services and tools appear to be on the rise among cloud customers. Therefore, 2024 will likely be the year in which many companies decide which experimentation paths are worth their investment, and what sort of AI-based products and features they're going to pursue. As many organizations are experimenting with generative AI in parallel, we expect the coming year to reveal precisely whether and how this technology can increase efficiency and enable never-before-seen features.

- The fast-paced adoption of this new technology is reminiscent of the cloud revolution itself, when almost every modern organization made the decision to shift many of their assets to the cloud, and the cloud also enabled an entire generation of new cloud-first companies. The current AI uptrend and the ongoing cloud revolution are intertwined, as generative AI is highly dependent on the scalability and capacity of the cloud. However, unlike the early days of cloud, the security considerations of AI must accompany the technological leap rather than emerge only years later.
- In conclusion, the future of AI in the cloud seems bright, and we're excited to see what the coming year of AI in the cloud has in store. Finally, we recommend the following strategies for success in safely incorporating all things AI into your cloud:

1. Check out our guide on choosing an AI-SPM solution in order to build visibility into AI service and product usage in your organization, and work to eliminate blind spots materializing as "Shadow AI". Observability will help you to understand which teams are using AI in your organization and what services and tools they're working with, so that you can ensure that both experimentation and development are conducted according to the most secure specifications. This will allow your organization to reap all the benefits of the latest and greatest in generative AI technology without introducing any significant risks or attack surface.
2. While you're at it, read our blogpost to learn how to securely build multi-tenant services that leverage generative AI models as part of their operation, so that your customers' data remains safe as well (and if you'd like to learn more about tenant isolation in general, be sure to take a look at the PEACH framework).
3. As in all things, foster a culture of security ownership towards AI. Almost every team in an organization can use AI (though time will tell whether every team should). In the new cloud operating model, security teams must work together with developers and cloud engineering teams, and now they must consider how best to bring data scientists and AI engineers into the fold.