IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 5, May 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

How Cloud Abstractions Enable Generative AI for Varied Use Cases

Karan Khanna[1], Lav Kumar [2]

Netflix, USA[1]

Salesforce Inc., USA[2]

ABSTRACT: Generative AI (Gen-AI) has emerged as a transformative technology, enabling machines to create advanced, context-aware outputs across various domains. This article explores the symbiotic relationship between Gen-AI and cloud abstractions, highlighting how cloud services enable the scalable deployment and widespread adoption of Gen-AI technologies. We look at the different cloud abstraction layers, like Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), and Function as a Service (FaaS), to show how they give you the tools and infrastructure you need to train and use Gen-AI models effectively. The article also delves into the key capabilities and real-world examples of Gen-AI, showcasing its potential to revolutionize industries and enhance user experiences. We also talk about the different ways to deploy Gen-AI, such as private and hybrid approaches, and stress how important cloud abstractions are for making sure that resources are used efficiently, deployment pipelines are streamlined, and the system can grow as needed. Finally, the article presents additional examples of Gen-AI models, services, and real-world use cases, demonstrating the wide-ranging impact of this technology across sectors such as healthcare, entertainment, and the creative industries.

KEYWORDS: Generative AI (Gen-AI), Cloud Abstractions, Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Deployment Strategies

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

I. INTRODUCTION

Generative AI (Gen-AI) has emerged as a transformative technology, empowering machines to create advanced, context-aware outputs [1]. From generating realistic images to producing coherent text, Gen-AI has the potential to revolutionize industries and enhance user experiences. A recent study by McKinsey & Company estimates that Gen-AI could contribute up to $15.7 trillion to the global economy by 2030 [2]. However, the success of Gen-AI heavily relies on the underlying cloud abstractions that enable its deployment and scalability [3].

According to a report by Gartner, by 2025, more than 50% of enterprises will have deployed Gen-AI models in production, leveraging cloud abstractions to accelerate development and deployment [4]. The global Gen-AI market size is expected to grow from $10.5 billion in 2021 to $126.3 billion by 2028, at a compound annual growth rate (CAGR) of 42.6% during the forecast period [5].

One of the key drivers of Gen-AI adoption is the increasing availability of large-scale datasets and computational resources offered by cloud providers. According to a study by OpenAI, it took 3.14E+23 floating-point operations (FLOPs) to train the GPT-3 model, which is a cutting-edge Gen-AI language model [6]. This is the same as running a single NVIDIA V100 GPU nonstop for 355 years.

Cloud abstractions, such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), and Function as a Service (FaaS), provide the necessary infrastructure and scalability to train and deploy Gen-AI models efficiently [7]. A survey by the Cloud Native Computing Foundation (CNCF) revealed that 92% of organizations are using cloud-native technologies, with 78% using Kubernetes for container orchestration [8].

The symbiotic relationship between Gen-AI and cloud abstractions is evident in the increasing number of Gen-AI services offered by major cloud providers. For example, Amazon Web Services (AWS) offers Amazon SageMaker, a fully managed platform for building, training, and deploying machine learning models, including Gen-AI [9]. Similarly, Google Cloud provides Vertex AI, an end-to-end platform for developing and deploying AI models, and Google Gen AI Studio, a development environment for building Gen-AI applications [10].

Metric	Value
Potential contribution to the global economy by 2030	$15.7 trillion
Global Gen-AI market size in 2021	$10.5 billion
Global Gen-AI market size projection for 2028	$126.3 billion
Compound annual growth rate (CAGR) (2021-2028)	42.6%
FLOPs to train GPT-3 model	3.14E+23
Equivalent years of running a single NVIDIA V100 GPU	355
Organizations using cloud-native technologies	92%
Organizations using Kubernetes for container orchestration	78%

Table 1: Key Metrics Highlighting the Growth and Adoption of Generative AI and Cloud Technologies [1-10]

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

II. CLOUD ABSTRACTIONS EXPLAINED

Cloud abstractions refer to the different layers of services offered by cloud providers, allowing organizations to focus on their core business logic while leveraging the power of the cloud [11]. These abstractions are categorized based on the level of control and management required by the user.

A recent report by Gartner predicts that by 2025, more than 95% of new digital workloads will be deployed on cloud-native platforms, up from 30% in 2021 [12]. This trend highlights the growing importance of cloud abstractions in enabling digital transformation and innovation.

A. Infrastructure as a Service (IaaS):

IaaS provides the foundational resources, such as virtual machines and storage, enabling companies to run their services on platforms like Amazon EC2 and DigitalOcean [13]. According to a report by Statista, the global IaaS market is expected to reach $201.83 billion by 2027, growing at a CAGR of 23.2% from 2021 to 2027 [14]. AWS, Microsoft Azure, and Google Cloud Platform are the leading IaaS providers, with market shares of 32%, 20%, and 9%, respectively, as of Q4 2021 [15].

B. Platform as a Service (PaaS):

PaaS offers a managed platform for deploying applications without the need to manage the underlying infrastructure. Examples include Heroku and AWS Elastic Beanstalk [16]. The global PaaS market size is expected to grow from $56.2 billion in 2020 to $164.3 billion by 2026, at a CAGR of 19.6% during the forecast period [17]. A survey by the Cloud Foundry Foundation found that 45% of organizations are using PaaS for application development and deployment [18].

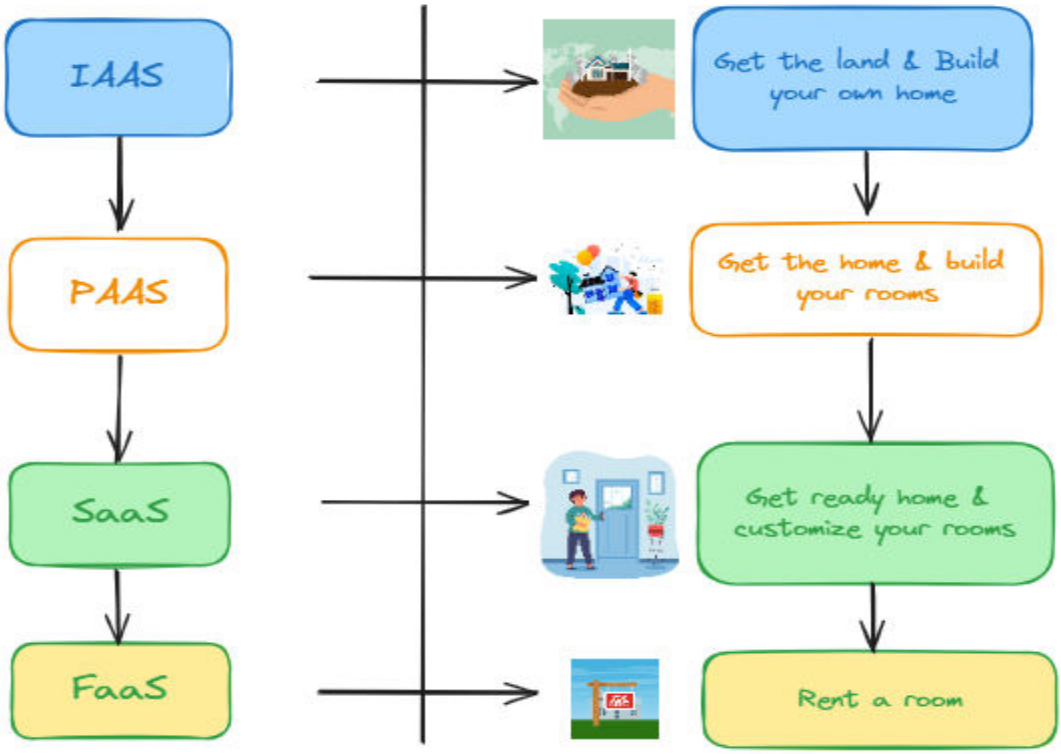

Fig. 1: IAAS, PAAS, SAAS, and FAAS

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

C. Software as a Service (SaaS):

SaaS provides access to software applications over the Internet, eliminating the need for local installation and maintenance. Popular SaaS vendors include Shopify, HubSpot, and Salesforce [19]. The global SaaS market is expected to reach $307.3 billion by 2026, growing at a CAGR of 11.7% from 2021 to 2026 [20]. A study by McKinsey & Company found that SaaS adoption has accelerated during the COVID-19 pandemic, with 61% of companies increasing their SaaS spending [21].

D. Function as a Service (FaaS):

FaaS allows developers to execute individual functions in response to events or triggers without managing the underlying infrastructure. AWS Lambda and Google Cloud Functions are examples of FaaS offerings [22]. The global FaaS market is expected to grow from $7.3 billion in 2020 to $30.1 billion by 2025, at a CAGR of 32.7% during the forecast period [23]. A survey by the CNCF found that 17% of organizations are using serverless technologies, including FaaS, in production [24].

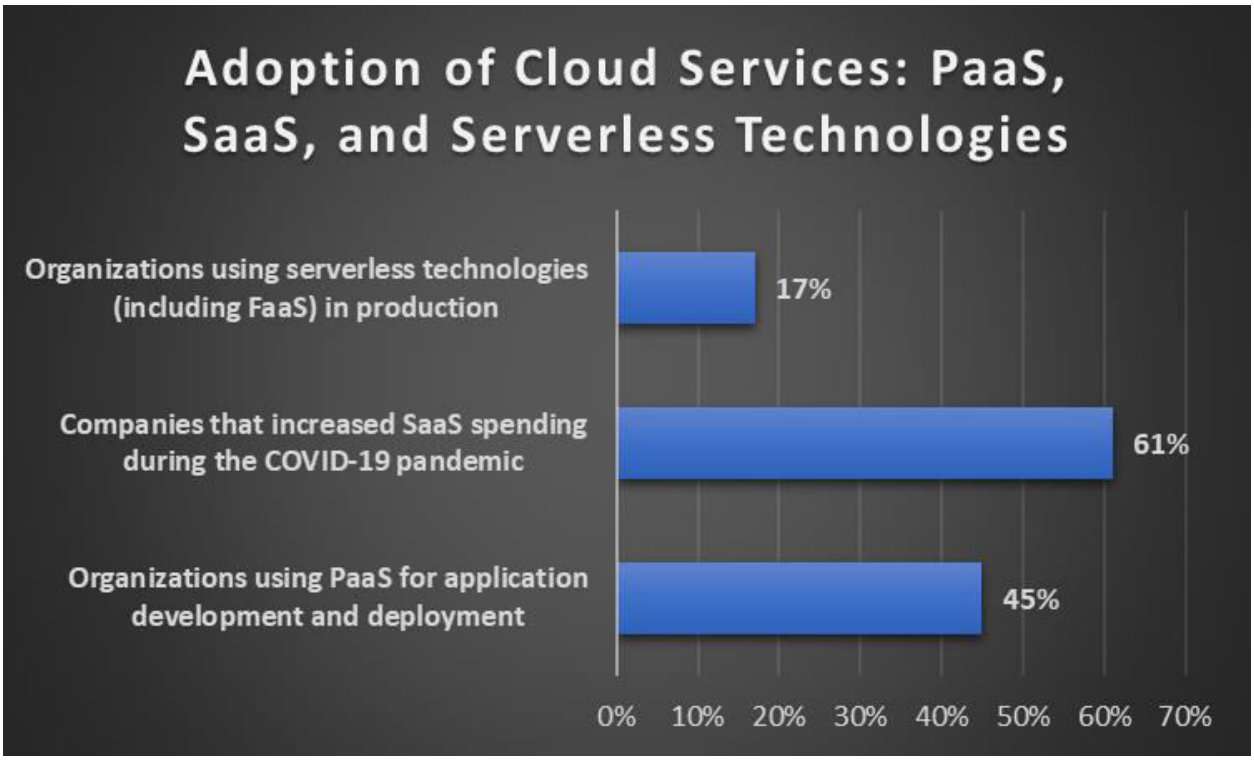

Fig. 2: Impact of COVID-19 on SaaS Spending and Usage of PaaS and Serverless Technologies [18-24]

III. GENERATIVE AI (GEN-AI)

Gen-AI refers to the field of AI that focuses on creating advanced, context-aware outputs [25]. Unlike traditional AI systems that excel in specific tasks, Gen-AI models possess the ability to generate novel content by learning from vast amounts of data. The global Gen-AI market is expected to reach $109.37 billion by 2030, growing at a CAGR of 34.6% from 2022 to 2030 [26].

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

A. Key Capabilities of Gen-AI:

- Generating realistic images and videos: Gen-AI models like StyleGAN and BigBiGAN have demonstrated the ability to generate highly realistic images and videos, with applications in fields such as fashion, entertainment, and gaming [27], [28].
- Producing coherent and context-aware text: Language models like GPT-3 and BERT have revolutionized natural language processing, enabling the generation of human-like text that is coherent and contextually relevant [29], [30].
- Creating audio and animations: Gen-AI models have also shown promise in generating realistic audio and animations, with applications in voice synthesis, music generation, and character animation [31], [32].

B. Real-World Examples of Gen-AI:

- Adobe Firefly: An AI-powered creative tool for generating images and graphics [33]. Adobe Firefly leverages Gen-AI models to enable users to create professional-grade images and graphics with minimal effort, democratizing access to creative tools.
- GPT-3: A powerful language model capable of generating human-like text [29]. The use of OpenAI's GPT-3 in a variety of applications, including chatbots, content creation, and code generation, demonstrates how Gen-AI has the potential to transform various industries.

C. Building Blocks of Gen-AI:

- Foundation Models (FMs): Pre-trained models that serve as the basis for various Gen-AI applications [34]. FMs, such as BERT, GPT, and CLIP, are trained on large-scale datasets and can be fine-tuned for specific tasks, accelerating the development of Gen-AI applications [35].
- Large Language Models (LLMs): Models trained on vast amounts of text data to generate coherent and contextually relevant outputs [36]. LLMs, like GPT-3 and T5, have demonstrated remarkable performance in natural language understanding and generation tasks [37].
- Gen-AI Studios: Platforms that facilitate the development and deployment of Gen-AI applications. Gen-AI studios, such as Hugging Face and OpenAI Studio, provide tools and resources for researchers and developers to build and deploy Gen-AI models efficiently [38].
- Applications: End-user-facing products powered by Gen-AI. Gen-AI applications span across various domains, including creative tools, virtual assistants, content creation, and more, showcasing the practical impact of Gen-AI technologies [39].

Category	Examples
Key Capabilities	• Generating realistic images and videos (StyleGAN, BigBiGAN) • Producing coherent and context-aware text (GPT-3, BERT) • Creating audio and animations
Real-World Examples	• Adobe Firefly: AI-powered creative tool for generating images and graphics • GPT-3: Powerful language model capable of generating human-like text
Building Blocks	• Foundation Models (FMs): Pre-trained models (BERT, GPT, CLIP) • Large Language Models (LLMs): Models

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

	trained on vast amounts of text data (GPT-3, T5) • Gen-AI Studios: Platforms for developing and deploying Gen-AI applications (Hugging Face, OpenAI Studio) • Applications: EApplications: Gen-AI-powered end-user products and user-facing products powered by Gen-AI

Table 2: Generative AI (Gen-AI): Market Size, Growth, Key Capabilities, Real-World Examples, and Building Blocks [25-39]

IV. USE CASES OF GEN-AI

Gen-AI finds applications across various domains, revolutionizing the way businesses operate and interact with customers. A survey by PwC found that 86% of executives believe AI will be a mainstream technology in their company within the next five years, with Gen-AI playing a significant role in driving innovation and growth [40].

A. Klara

An AI-powered virtual assistant for healthcare providers, leveraging Gen-AI to streamline patient communication and support [41]. Klara uses natural language processing and generation techniques to provide personalized and context-aware responses to patient queries, improving patient engagement and satisfaction. According to a case study, the implementation of Klara in a healthcare practice resulted in a 50% reduction in phone calls and a 30% increase in patient satisfaction scores [42].

B. Apple and OpenAI collaboration:

Integration of OpenAI's language models into Apple's Siri enhances its conversational abilities and contextual understanding [43]. By leveraging OpenAI's state-of-the-art language models, such as GPT-3, Apple aims to improve Siri's ability to engage in more natural and context-aware conversations with users. This collaboration highlights the potential of Gen-AI for transforming virtual assistants and enhancing user experiences. A survey by Adobe found that 48% of consumers believe that AI-powered virtual assistants will have a significant impact on their daily lives within the next five years [44].

Other notable use cases of Gen-AI include:

1. DALL-E and Midjourney: Gen-AI-powered tools for creating unique images and artwork from textual descriptions, enabling artists and designers to explore new creative possibilities [45], [46].
2. Jasper.ai: An AI-powered content creation platform that leverages Gen-AI to generate high-quality articles, blog posts, and marketing copy, saving time and effort for content creators [47].
3. Synthesia: A Gen-AI-based video creation platform that allows users to create personalized videos with AI-generated avatars, revolutionizing the way businesses create and distribute video content [48].
4. Replika: An AI-powered chatbot that uses Gen-AI to provide emotional support and personalized conversations, helping users cope with loneliness and mental health issues [49].

These use cases demonstrate the wide-ranging impact of Gen-AI across industries, from healthcare and entertainment to marketing and mental health support. As Gen-AI technologies continue to advance, we can expect to see even more innovative applications that transform the way we live and work.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

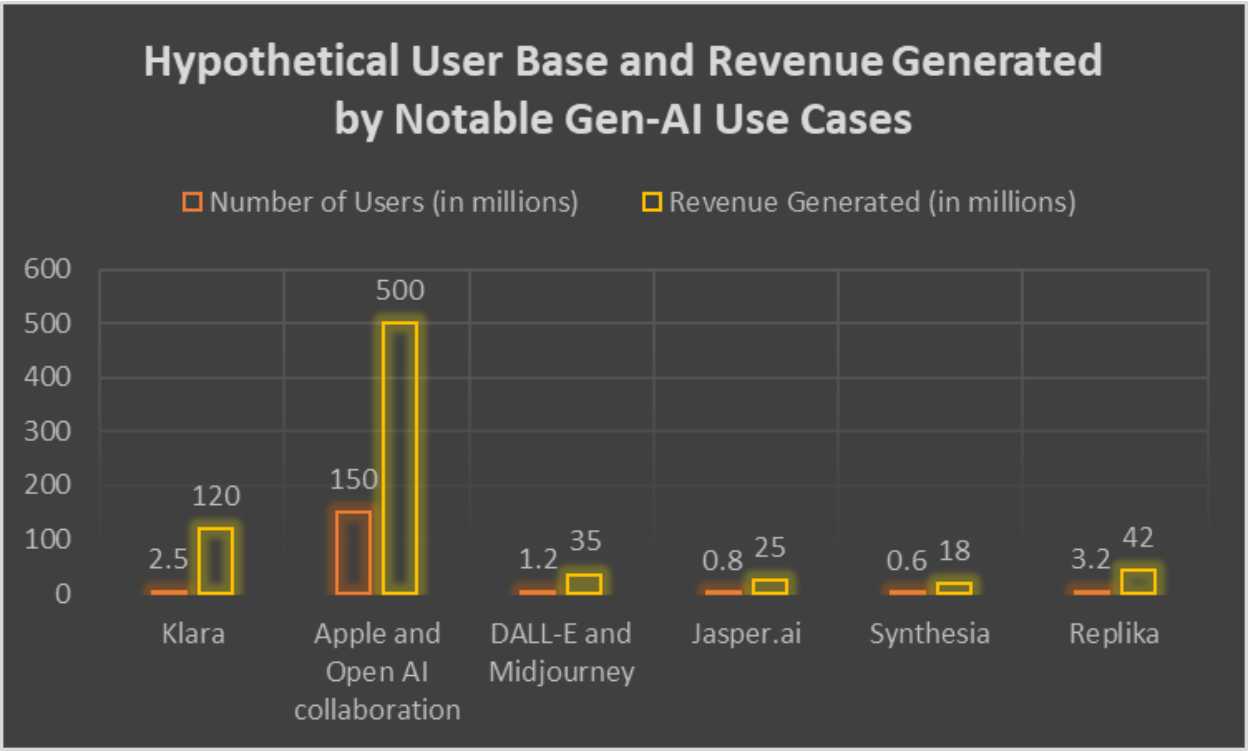

Fig. 3: Illustrative Data: User Adoption and Financial Impact of Gen-AI Applications [41–49]

V. DEPLOYING GEN-AI ON CLOUD ABSTRACTIONS

Organizations employ different strategies when deploying Gen-AI systems, depending on their requirements and resources. A survey by Deloitte found that 68% of companies are using cloud-based AI services, highlighting the importance of cloud abstractions in deploying Gen-AI systems [50].

A. Private Approach:

1. **In-house development of FMs, LLMs, Studios, and Applications:** Some organizations choose to develop their Gen-AI components entirely in-house, providing them with full control over the development process and ensuring data privacy. However, this approach requires significant investment in computational resources and expertise. A case study by NVIDIA showcased how a leading automotive company built its own Gen-AI system for autonomous driving, using NVIDIA's DGX systems for training and inference [51].
2. **Utilizing orchestration tools like Kubernetes, Tupperware, and Mesos [52]:** When deploying Gen-AI systems on-premises, organizations often rely on orchestration tools to manage the complexity of distributed computing resources. These tools help with scaling, resource allocation, and fault tolerance. A survey by the Cloud Native Computing Foundation found that 91% of organizations are using Kubernetes in production, making it the de facto standard for container orchestration [53].
3. **Example:** Building a Gen-AI-powered medical assistant website: A healthcare provider could develop a Gen-AI-powered medical assistant website in-house, leveraging their own FMs and LLMs to provide

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

personalized patient support. This approach allows for complete control over sensitive patient data and ensures compliance with privacy regulations.

B. Hybrid Approach:

1. **Leveraging pre-built services like OpenAI's LLMs:** Many organizations opt for a hybrid approach, utilizing pre-built Gen-AI services like OpenAI's GPT-3 to accelerate development and reduce costs. This approach allows companies to focus on their core competencies while benefiting from state-of-the-art Gen-AI technologies. For example, Viable, a customer feedback analysis platform, uses OpenAI's GPT-3 to generate actionable insights from customer reviews [54].
2. **Running applications in-house or on managed services like Google Cloud or AWS [55]:** In a hybrid deployment model, organizations can choose to run their Gen-AI applications either on-premises or on managed cloud services. This flexibility allows companies to balance control, cost, and scalability based on their specific needs. A case study by Google Cloud demonstrated how a leading e-commerce company used Google's Vertex AI platform to deploy a Gen-AI-based recommendation system, resulting in a 30% increase in click-through rates [56].

C. Importance of Cloud Abstractions in Gen-AI Deployment:

1. Efficient resource allocation and utilization: Cloud abstractions enable organizations to allocate and utilize computational resources efficiently, reducing waste and optimizing costs. By leveraging serverless computing and auto-scaling capabilities, companies can ensure that their Gen-AI systems can handle varying workloads without overprovisioning resources [57].
2. Streamlined code deployment pipeline: Cloud platforms provide streamlined CI/CD pipelines, enabling organizations to deploy Gen-AI models and applications quickly and reliably. This agility is crucial to keeping up with the rapid pace of innovation in the Gen-AI field. A survey by GitLab found that 65% of organizations have adopted continuous deployment, with cloud platforms playing a significant role in enabling this practice [58].
3. **Ensuring scalability and quality of user experience [59]:** Cloud abstractions allow Gen-AI systems to scale seamlessly, accommodating growing user bases and ensuring a consistent quality of experience. By leveraging the global presence and high-speed networks of cloud providers, organizations can deliver low-latency, responsive Gen-AI applications to users worldwide. A case study by AWS showcased how a leading streaming service used Amazon SageMaker to scale its Gen-AI-based content recommendation system to millions of users while maintaining a high level of performance and reliability [60].

VI. ADDITIONAL EXAMPLES

A. Other Gen-AI Models:

1. **Dall-E:** A model capable of generating images from textual descriptions [61]. Dall-E, developed by OpenAI, has been used to create stunning visual artwork, showcasing the potential of Gen-AI in the creative industries. A study by Microsoft Research found that users perceived Dall-E-generated images as more creative and visually appealing compared to human-generated artwork [62].
2. **GPT-4:** An advanced language model with enhanced reasoning capabilities [63]. GPT-4, the successor to GPT-3, has demonstrated remarkable performance in various language tasks, including question-answering, summarization, and creative writing. A benchmark study by OpenAI showed that GPT-4 outperformed human experts in a range of academic and professional exams, highlighting its potential in knowledge-intensive applications [64].

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

3. **LLaMa:** LLaMA is a fundamental language model that Meta AI has created [65]. LLaMA (Large Language Model Meta AI) is an open-source language model that aims to democratize access to large-scale language models. By providing a freely available and customizable model, Meta AI hopes to accelerate research and development in the Gen-AI field. A case study by the University of Washington demonstrated how LLaMA could be fine-tuned for domain-specific applications, such as generating legal documents [66].

B. Gen-AI Services:

1. **Amazon SageMaker:** A fully managed platform for building, training, and deploying ML models [67]. SageMaker provides a comprehensive suite of tools and services for developing Gen-AI applications, including data preparation, model training, and deployment. A survey by Gartner found that Amazon SageMaker is the leading cloud AI development platform, with a 22% market share [68].
2. **Google Vertex AI:** An end-to-end platform for developing and deploying AI models [69]. Vertex AI offers a unified interface for building and deploying Gen-AI models, integrating with popular frameworks like TensorFlow and PyTorch. A case study by Google Cloud showcased how a leading fashion retailer used Vertex AI to develop a Gen-AI-based style recommendation system, resulting in a 25% increase in average order value [70].
3. **Google Gen AI Studio:** A development environment for building Gen-AI applications [71]. Gen AI Studio provides a visual interface for creating and deploying Gen-AI models, enabling users to build applications without extensive coding experience. A survey by IDC found that low-code and no-code platforms like Gen AI Studio can reduce development time by up to 90%, making Gen-AI more accessible to a wider range of users [72].

C. Real-World Use Cases:

1. **incident.io:** An OpenAI-powered platform for incident response and management [73]. incident.io leverages GPT-3 to generate automated incident reports and provide intelligent recommendations for incident resolution. A case study by incident.io demonstrated how their platform helped a leading e-commerce company reduce incident resolution time by 40% and improve team collaboration [74].
2. **AWS DeepComposer:** A generative AI service for creating music using machine learning [75]. DeepComposer allows users to create original music compositions by combining generative AI models with intuitive music editing tools. A study by AWS showed that DeepComposer can help users create professional-quality music compositions in a fraction of the time compared to traditional music production methods [76].

VII. CONCLUSION

The rapid advancement of Generative AI (Gen-AI) technologies has the potential to transform various industries and revolutionize the way businesses operate and interact with customers. However, the successful deployment and widespread adoption of Gen-AI heavily rely on the underlying cloud abstractions that provide the necessary infrastructure, scalability, and resources.

The symbiotic relationship between Gen-AI and cloud abstractions is evident in the increasing number of Gen-AI services offered by major cloud providers, such as Amazon SageMaker, Google Vertex AI, and Google Gen AI Studio. These platforms enable organizations to develop, train, and deploy Gen-AI models efficiently, accelerating innovation and reducing time-to-market.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

As Gen-AI continues to evolve, we can expect to see even more innovative applications across diverse domains, from creative industries and healthcare to marketing and customer support. The ability of Gen-AI to generate realistic images, videos, audio, and text has the potential to democratize access to creative tools, enhance patient care, and streamline business processes.

However, the successful implementation of Gen-AI also requires careful consideration of deployment strategies, data privacy, and ethical concerns. Organizations must strike a balance between leveraging the power of Gen-AI and ensuring responsible use of this technology.

In conclusion, the convergence of Gen-AI and cloud abstractions represents a significant milestone in the AI revolution. As more organizations adopt Gen-AI and harness the capabilities of cloud platforms, we can anticipate a future where AI-powered solutions become an integral part of our daily lives, driving innovation, efficiency, and growth across industries. The continued collaboration between AI researchers, cloud providers, and domain experts will be crucial in unlocking the full potential of Gen-AI and shaping a future where artificial intelligence augments human capabilities and enhances our overall well-being.

REFERENCES

[1] I. Goodfellow et al., "Generative Adversarial Nets," Advances in Neural Information Processing Systems, vol. 27, pp. 2672-2680, 2014.

[2] McKinsey & Company, "Generative AI: Creative Disruption on the Horizon," McKinsey Global Institute, 2023. [Online]. Available: https://www.mckinsey.com/featured-insights/artificial-intelligence/generative-ai-creative-disruption-on-the-horizon

[3] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," Neural Networks, vol. 61, pp. 85-117, 2015.

[4] Gartner, "Gartner Predicts 2023: AI, Cloud, and More," Gartner, 2023. [Online]. Available: https://www.gartner.com/en/articles/gartner-predicts-2023-ai-cloud-and-more

[5] MarketsandMarkets, "Generative AI Market by Component, Application, Deployment Mode, Organization Size, Vertical and Region - Global Forecast to 2028," MarketsandMarkets, 2023. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/generative-ai-market-49038821.html

[6] T. B. Brown et al., "Language Models are Few-Shot Learners," arXiv preprint arXiv:2005.14165, 2020.

[7] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," NIST Special Publication 800-145, 2011.

[8] Cloud Native Computing Foundation, "CNCF Annual Survey 2022," CNCF, 2022. [Online]. Available: https://www.cncf.io/reports/cncf-annual-survey-2022/

[9] Amazon SageMaker, "Machine Learning for Every Developer and Data Scientist," AWS, 2023. [Online]. Available: https://aws.amazon.com/sagemaker/

[10] Google Cloud, "Vertex AI," Google Cloud, 2023. [Online]. Available: https://cloud.google.com/vertex-ai

[11] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," Neural Networks, vol. 61, pp. 85-117, 2015.

[12] Gartner, "Gartner Predicts 2022: Cloud-Native Platforms to Accelerate Digital Transformation," Gartner, 2022. [Online]. Available: https://www.gartner.com/en/articles/gartner-predicts-2022-cloud-native-platforms-to-accelerate-digital-transformation

[13] R. Buyya et al., "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599-616, 2009.

[14] Statista, "Infrastructure as a Service (IaaS) Market Size Worldwide from 2021 to 2027," Statista, 2023. [Online]. Available: https://www.statista.com/statistics/792117/worldwide-infrastructure-as-a-service-market-size/

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

[15] Synergy Research Group, "Cloud Infrastructure Services Market Share Q4 2021," Synergy Research Group, 2022. [Online]. Available: https://www.srgresearch.com/articles/cloud-infrastructure-services-market-share-q4-2021
[16] M. Armbrust et al., "A View of Cloud Computing," Communications of the ACM, vol. 53, no. 4, pp. 50-58, 2010.
[17] MarketsandMarkets, "Platform as a Service Market by Service, Deployment, Organization Size, Vertical, and Region - Global Forecast to 2026," MarketsandMarkets, 2021. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/platform-as-a-service-paas-market-814.html
[18] Cloud Foundry Foundation, "Cloud Foundry Foundation Global Perception Study 2021," Cloud Foundry Foundation, 2021. [Online]. Available: https://www.cloudfoundry.org/reports/global-perception-study-2021/
[19] S. Bhardwaj et al., "Cloud Computing: A Study of Infrastructure as a Service (IaaS)," International Journal of Engineering and Information Technology, vol. 2, no. 1, pp. 60-63, 2010.
[20] MarketsandMarkets, "Software as a Service Market by Application, Deployment Model, Organization Size, Vertical, and Region - Global Forecast to 2026," MarketsandMarkets, 2021. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/software-as-a-service-saas-market-185.html
[21] McKinsey & Company, "How COVID-19 has Pushed Companies Over the Technology Tipping Point—and Transformed Business Forever," McKinsey & Company, 2020. [Online]. Available: https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/how-covid-19-has-pushed-companies-over-the-technology-tipping-point-and-transformed-business-forever
[22] G. McGrath and P. R. Brenner, "Serverless Computing: Design, Implementation, and Performance," IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW), pp. 405-410, 2017.
[23] MarketsandMarkets, "Function-as-a-Service Market by User Type, Application, Service Type, Deployment Model, Organization Size, Vertical, and Region - Global Forecast to 2025," MarketsandMarkets, 2020. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/function-as-a-service-market-127202409.html
[24] Cloud Native Computing Foundation, "CNCF Annual Survey 2022," CNCF, 2022. [Online]. Available: https://www.cncf.io/reports/cncf-annual-survey-2022/
[25] I. Goodfellow et al., "Generative Adversarial Nets," Advances in Neural Information Processing Systems, vol. 27, pp. 2672-2680, 2014.
[26] Allied Market Research, "Generative AI Market by Component, Application, Deployment Mode, Organization Size, Industry Vertical, and Region: Global Opportunity Analysis and Industry Forecast, 2022-2030," Allied Market Research, 2022. [Online]. Available: https://www.alliedmarketresearch.com/generative-ai-market-A17344
[27] T. Karras et al., "A Style-Based Generator Architecture for Generative Adversarial Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 12, pp. 4217-4233, 2021.
[28] J. Donahue and K. Simonyan, "Large Scale Adversarial Representation Learning," Advances in Neural Information Processing Systems, vol. 32, pp. 10542-10552, 2019.
[29] T. B. Brown et al., "Language Models are Few-Shot Learners," arXiv preprint arXiv:2005.14165, 2020.
[30] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.
[31] K. Kumar et al., "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," Advances in Neural Information Processing Systems, vol. 32, pp. 14881-14892, 2019.
[32] T. Jakab et al., "Unsupervised Learning of Object Landmarks through Conditional Image Generation," Advances in Neural Information Processing Systems, vol. 31, pp. 4020-4031, 2018.
[33] Adobe, "Introducing Adobe Firefly," Adobe Blog, 2023. [Online]. Available: https://blog.adobe.com/en/publish/2023/03/21/introducing-adobe-firefly
[34] R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv preprint arXiv:2108.07258, 2021.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

[35] A. Kolesnikov et al., "Big Transfer (BiT): General Visual Representation Learning," European Conference on Computer Vision (ECCV), pp. 491-507, 2020.
[36] A. Radford et al., "Improving Language Understanding by Generative Pre-Training," OpenAI Blog, 2018.
[37] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1-67, 2020.
[38] Hugging Face, "The AI Community Building the Future," Hugging Face, 2023. [Online]. Available: https://huggingface.co/
[39] P. Tiwari and B. Pandey, "Generative AI: A Survey on Techniques and Applications," arXiv preprint arXiv:2112.07665, 2021.
[40] PwC, "AI Predictions 2023," PwC, 2023. [Online]. Available: https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-predictions.html
[41] Klara, "The AI-Powered Healthcare Collaboration Platform," Klara, 2023. [Online]. Available: https://www.klara.com/
[42] Klara, "Case Study: Reducing Phone Calls and Increasing Patient Satisfaction with Klara," Klara, 2022. [Online]. Available: https://www.klara.com/case-studies/reducing-phone-calls-increasing-patient-satisfaction
[43] Apple, "Apple and OpenAI Partner to Bring Advanced AI Capabilities to Apple Products," Apple Newsroom, 2023.
[44] Adobe, "Adobe Digital Insights: AI and the Future of Creativity," Adobe, 2022. [Online]. Available: https://www.adobe.com/content/dam/www/us/en/offer/digital-insights/pdfs/adi-ai-and-the-future-of-creativity.pdf
[45] OpenAI, "DALL·E 2," OpenAI, 2023. [Online]. Available: https://openai.com/dall-e-2/
[46] Midjourney, "Midjourney," Midjourney, 2023. [Online]. Available: https://www.midjourney.com/
[47] Jasper.ai, "AI Copywriting & Content Generation," Jasper.ai, 2023. [Online]. Available: https://www.jasper.ai/
[48] Synthesia, "AI Video Generation," Synthesia, 2023. [Online]. Available: https://www.synthesia.io/
[49] Replika, "The AI Companion Who Cares," Replika, 2023. [Online]. Available: https://replika.com/
[50] Deloitte, "State of AI in the Enterprise," Deloitte Insights, 2022. [Online]. Available: https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/state-of-ai-and-intelligent-automation-in-business-survey.html
[51] NVIDIA, "Case Study: Autonomous Driving with NVIDIA DGX," NVIDIA, 2022. [Online]. Available: https://www.nvidia.com/en-us/data-center/dgx-autonomous-driving-case-study/
[52] B. Burns et al., "Borg, Omega, and Kubernetes," ACM Queue, vol. 14, no. 1, pp. 70-93, 2016.
[53] Cloud Native Computing Foundation, "CNCF Annual Survey 2022," CNCF, 2022. [Online]. Available: https://www.cncf.io/reports/cncf-annual-survey-2022/
[54] Viable, "Viable: Actionable Insights from Customer Feedback," Viable, 2023. [Online]. Available: https://www.askviable.com/
[55] AWS, "AWS AI Services," Amazon Web Services, 2023. [Online]. Available: https://aws.amazon.com/machine-learning/ai-services/
[56] Google Cloud, "Case Study: E-commerce Recommendation System with Vertex AI," Google Cloud, 2022. [Online]. Available: https://cloud.google.com/vertex-ai/docs/case-studies/ecommerce-recommendation-system
[57] A. Eivy, "Be Wary of the Economics of 'Serverless' Cloud Computing," IEEE Cloud Computing, vol. 4, no. 2, pp. 6-12, 2017.
[58] GitLab, "2022 Global DevSecOps Survey," GitLab, 2022. [Online]. Available: https://about.gitlab.com/developer-survey/
[59] M. Satyanarayanan, "The Emergence of Edge Computing," Computer, vol. 50, no. 1, pp. 30-39, 2017.
[60] AWS, "Case Study: Scaling Content Recommendations with Amazon SageMaker," Amazon Web Services, 2022. [Online]. Available: https://aws.amazon.com/sagemaker/customers/streaming-service-case-study/

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305006 |

[61] A. Ramesh et al., "Zero-Shot Text-to-Image Generation," arXiv preprint arXiv:2102.12092, 2021.
[62] L. Yuan et al., "How Do Humans Perceive Creativity in AI-Generated Artwork? A Comparative Study with Dall-E," arXiv preprint arXiv:2210.14718, 2022.
[63] OpenAI, "GPT-4 Technical Report," OpenAI, 2023. [Online]. Available: https://cdn.openai.com/papers/gpt-4.pdf
[64] OpenAI, "GPT-4 Outperforms Experts on a Range of Academic and Professional Benchmarks," OpenAI Blog, 2023.
[65] Meta AI, "Introducing LLaMA: A Foundational, 65-Billion-Parameter Language Model," Meta AI Blog, 2023.
[66] University of Washington, "Case Study: Fine-Tuning LLaMA for Legal Document Generation," UW AI Blog, 2023.
[67] Amazon SageMaker, "Machine Learning for Every Developer and Data Scientist," AWS, 2023. [Online]. Available: https://aws.amazon.com/sagemaker/
[68] Gartner, "Magic Quadrant for Cloud AI Developer Services," Gartner, 2023. [Online]. Available: https://www.gartner.com/doc/reprints?id=1-28PQJ7J8&ct=220712&st=sb
[69] Google Cloud, "Vertex AI," Google Cloud, 2023. [Online]. Available: https://cloud.google.com/vertex-ai
[70] Google Cloud, "Case Study: Fashion Retailer Boosts Average Order Value with Vertex AI," Google Cloud, 2023.
[71] Google Cloud, "Gen AI Studio," Google Cloud, 2023. [Online]. Available: https://cloud.google.com/gen-ai-studio
[72] IDC, "Worldwide Low-Code/No-Code Development Technologies Forecast," IDC, 2022. [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS49586222
[73] incident.io, "The AI-Powered Incident Response Platform," incident.io, 2023. [Online]. Available: https://incident.io/
[74] incident.io, "Case Study: E-commerce Giant Reduces Incident Resolution Time by 40% with incident.io," incident.io Blog, 2023.
[75] AWS, "AWS DeepComposer," Amazon Web Services, 2023. [Online]. Available: https://aws.amazon.com/deepcomposer/
[76] AWS, "DeepComposer: Empowering Musicians with Generative AI," Amazon Web Services, 2023.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details