Review of Computer Engineering Research
2024 Vol. 11, No. 4, pp. 140-154
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v11i4.4017
© 2024 Conscientia Beam. All Rights Reserved.

Exploring model-as-a-service for generative ai on cloud platforms

Harshad Pitkar[1+]
Sanjay Bauskar[2]
Devendra Singh Parmar[3]
Hemlatha Kaur Saran[4]

[1]*Luddy School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN, USA.*
Email: harshpitkar@gmail.com
[2]*Pharmavite LLC, Los Angeles, CA, USA.*
Email: sanjaybauskar@gmail.com
[3]*Discover Financial Service, Riverwoods, IL, USA.*
Email: davesingh081@gmail.com
[4]*Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, Maharashtra, India.*
Email: hemlathasaran@gmail.com

(+ Corresponding author)

Article History
Received: 21 October 2024
Revised: 29 November 2024
Accepted: 13 December 2024
Published: 27 December 2024

Keywords
Cloud computing
Compliance
Containerization
Deep learning
Edge computing
Generative AI
Kubernetes
Model-as-a-service
Security.

ABSTRACT

This study examines the exploration of Model-as-a-Service for generative AI on cloud platforms. Model-as-a-Service (MaaS) could revolutionize generative AI; thus, we examine its impact on sectors, implementation best practices, and future trends. Business usage of generative AI for content development, predictive modelling, and consumer engagement is flexible and scalable using Software as a Service (SaaS). We explore how MaaS lets companies access, train, and deploy complex generative models like Generative Adversarial Networks (GAN), Variational Autoencoders (VAE), and Transformers without expensive in-house AI infrastructure. Lifecycle management in MaaS simplifies model training, deployment, versioning, and continuous improvement for iterative development in dynamic business contexts. MaaS security and compliance are crucial in highly regulated areas, including healthcare, finance, and law. Encryption, network isolation, and access control protect data and models. Generative AI models handle sensitive data; hence, industry standards and data sovereignty must be followed. Ethical AI, edge computing, and low-code/no-code platforms will enable more people to use models in real time and follow responsible AI guidelines, making MaaS's future bright. Generative AI applications and real-world case studies in healthcare, banking, retail, and entertainment demonstrate how MaaS can create value and stimulate innovation. Our study finds that using MaaS for generative AI, businesses can immensely benefit and explains how developers can speed up development, improve customer experiences, and remain ahead in the ever-changing digital landscape.

Contribution/Originality: In this study, we explore Model as a Service and its potential in relation to Generative AI. The study focuses on leading cloud providers and their service offerings in this space. Unlike previous work, we have explored different use cases and demonstrated how various industries can take advantage of this model to improve scalability and affordability and drive innovation.

1. INTRODUCTION TO MODEL-AS-A-SERVICE (MAAS) IN GENERATIVE AI

1.1. Definition and Overview of MaaS in the Context of Generative AI

Cloud-based Model-as-a-Service (MaaS) enables customers to build, administer, and use machine learning models without infrastructure. MaaS works in generative AI because it simplifies the API interface and allows customers access to complex AI models like large-scale transformers, Generative Adversarial Networks (GANs), and Variational Autoencoders [1]. GANs may do computationally difficult tasks, including natural language processing, graphic creation, and creative content production. By coordinating resource-intensive models with cloud infrastructure, MaaS simplifies and speeds up generative AI adoption for enterprises. MaaS lets developers and organizations focus on innovation and real-world generative model use instead of training, deploying, and scaling models. By managing data pretreatment, training, deployment, and version control, MaaS simplifies generative AI model lifecycle. Generated AI is simple to integrate into apps since users may interact with these models via RESTful APIs or other interface protocols. Monitoring model performance, usage metrics, and security compliance is essential for generative AI systems in healthcare and banking.

MaaS's on-demand capabilities enable enterprises of any size to implement cutting-edge AI models without infrastructure or technical skills.

1.2. Importance and Benefits of Maas for Businesses and Developers

MaaS helps organizations and developers deploy complex AI models. MaaS simplifies generative AI for businesses by eliminating expensive hardware and technical skills. Subject matter experts (SMEs) may use generative AI without data centers or AI professionals. AI-driven solutions help them compete with larger companies and innovate in marketing, product development, and consumer interaction [2]. Automating content creation can save time and money, and generative AI can help e-commerce enterprises improve customer experience with personalized suggestions and synthetic media. MaaS's scalable and flexible platform simplifies generative AI model deployment and management. Developers may easily construct, test, and scale models using its experimental feature without deployment delays. MaaS lets developers deploy changes, track model performance, and troubleshoot. MaaS lets developers and organizations leverage generative AI to deliver AI-driven apps faster, cheaper, and more efficiently.

1.3. Overview of Maas Integrates with Cloud Platforms

For model lifecycle management and cloud platform integration, MaaS employs cloud-native tools and services. MaaS providers improve generative AI models' performance, scalability, and management using AWS, Google Cloud, and Microsoft Azure's huge infrastructure and services. MaaS generative model deployments used Kubernetes [3]. This simplifies scaling and improvements. Model data is safe and available with cloud data management, checkpoints, and output storage. MaaS solutions optimize model performance in real time by measuring, detecting anomalies, and cloud-based monitoring and recording. MaaS lets enterprises engage with data pipelines, databases, and Application Programming Interfaces (APIs) to programmatically receive generative model outputs or input real-time data. During peak demand or service failures, MaaS will remain accessible due to cloud infrastructure redundancy and recovery. Cloud providers' strong architecture may allow MaaS platforms to integrate generative AI into business and developer operations; it improves agility, productivity, and resilience.

2. BACKGROUND ON GENERATIVE AI AND CLOUD PLATFORMS

2.1. Explanation of Generative Ai, Including Deep Learning Models Like Gans, Vaes, and Transformers

Generative AI creates media files that look, sound, and feel like current ones, emulating human ingenuity. Unlike regression or classification AI models, generative models learn data distributions and produce comparable but unique data points. Due to their intricate neural network designs, models can recognize patterns, interpret complex structures, and provide coherent outputs. Generative AI uses several powerful deep learning models, depending on task difficulty and nature [4]. Generative Adversarial Networks (GANs) are popular models in this discipline.

GANs are trained by competing discriminator and generator neural networks, and this is antagonistic training. The generator generates new data instances, which the discriminator then compares to actual data. This iterative procedure fine-tunes the generator, making the outcomes virtually identical to data. GANs improve style transfer, video creation, and picture synthesis.

Virtual Analog Embeddings (VAEs) are another key generative AI paradigm. VAEs encode and reconstruct data into a compressed latent space to determine its distribution. VAEs prioritize interpretability and controlled data output over actual generated data, making them ideal for anomaly detection, data compression, and feature learning. VAEs generate structured data variations for healthcare and diagnostics, among other businesses. Transformers, originally used in natural language processing (NLP), have substantially enhanced generative AI, especially in text generation, translation, and image production [5]. Attention approaches in transformer architecture allow models to capture long-distance data relationships. Vision Transformers (ViTs) have expanded their application to images, while OpenAI models like Generative Pre-Trained Transformer (GPT) and Bidirectional Encoder Representations and Transformers (BERT) have created standards for coherent and contextually relevant text. These models can manage enormous datasets, remember details across long durations, and produce visual or textual content that looks and feels human-made.

2.2. Overview of Popular Cloud Platforms and Their AI Services

Many firms want to use AI, especially generative models, in their processes; thus, Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure have established AI services to satisfy their needs. These platforms offer data processing, model training, deployment, and scaling. Amazon Web Services (AWS) delivers AI and machine learning with Amazon SageMaker. SageMaker's machine learning model construction, training, and deployment toolbox includes pre-built algorithms and custom model deployment. Resource-intensive generative AI models run best on Graphical Processing Unit (GPU)-based AWS instances [6]. AWS AI services like Amazon Rekognition, Polly, and Comprehend improve text analytics, speech synthesis, and image recognition.

Google Cloud Platform (GCP) offers Google AI and Vertex AI, a platform for managing ML models from start to finish. Vertex AI lets you use pre-trained models, train, and deploy them. It also supports AutoML, so you may build models without coding. Because it has machine learning-friendly Tensor Processing Units (TPUs), GCP is popular for deep learning. Google's Natural Language API and pre-trained models, such as BERT, are cloud-accessible and demonstrate its NLP expertise for sentiment analysis and language translation.

Azure Machine Learning provides a complete solution for training, deploying, and merging models with MLOps. Azure Cognitive Services pre-builds language, voice, vision, and decision-making. Microsoft has integrated OpenAI's GPT-3 model into Azure's generative AI applications to give Azure users powerful NLP capabilities. Azure offers GPU-powered virtual machine instances for compute-intensive generative models [7]. Tools for data annotation, pipeline management, and model monitoring are also available.

Each cloud platform lets organizations choose the infrastructure and services they need by offering a selection of AI applications through MaaS.

2.3. Historical Perspective on the Evolution from on-Premises to Cloud-Based AI

Data volume, computational power, and AI model complexity have made on-premises AI less successful than cloud-based AI. AI systems were once only available to major companies with sophisticated computers. On-premises setups require specialized staff, expensive hardware, and ongoing maintenance. AI became a niche sector with limited access due to high infrastructure costs and scarce resources [8]. Cloud computing's dramatic rise in the early 2010s led artificial intelligence to abandon on-premises installations for more accessible, scalable, and cost-effective alternatives. Cloud platforms' pay-as-you-go pricing allowed companies to employ AI without huge initial expenses. GPUs and TPUs in cloud environments sped up AI development by training complex models in a tenth of the time needed by CPU-based systems. Figure 1 shows the adoption rate of MaaS between 2015 and 2023.

As you can see, more and more businesses are running their models in the Cloud; at the same time, there is a steady decline in utilizing on-premises hardware for these use cases.

MaaS has advanced AI development by enabling the deployment of sophisticated models as cloud services. Businesses may leverage innovative models without investing in the knowledge and labor needed to construct AI from scratch with MaaS. Cloud platforms' AI technologies let enterprises go from training to deployment quickly. This has democratized AI and expedited its adoption across industries.

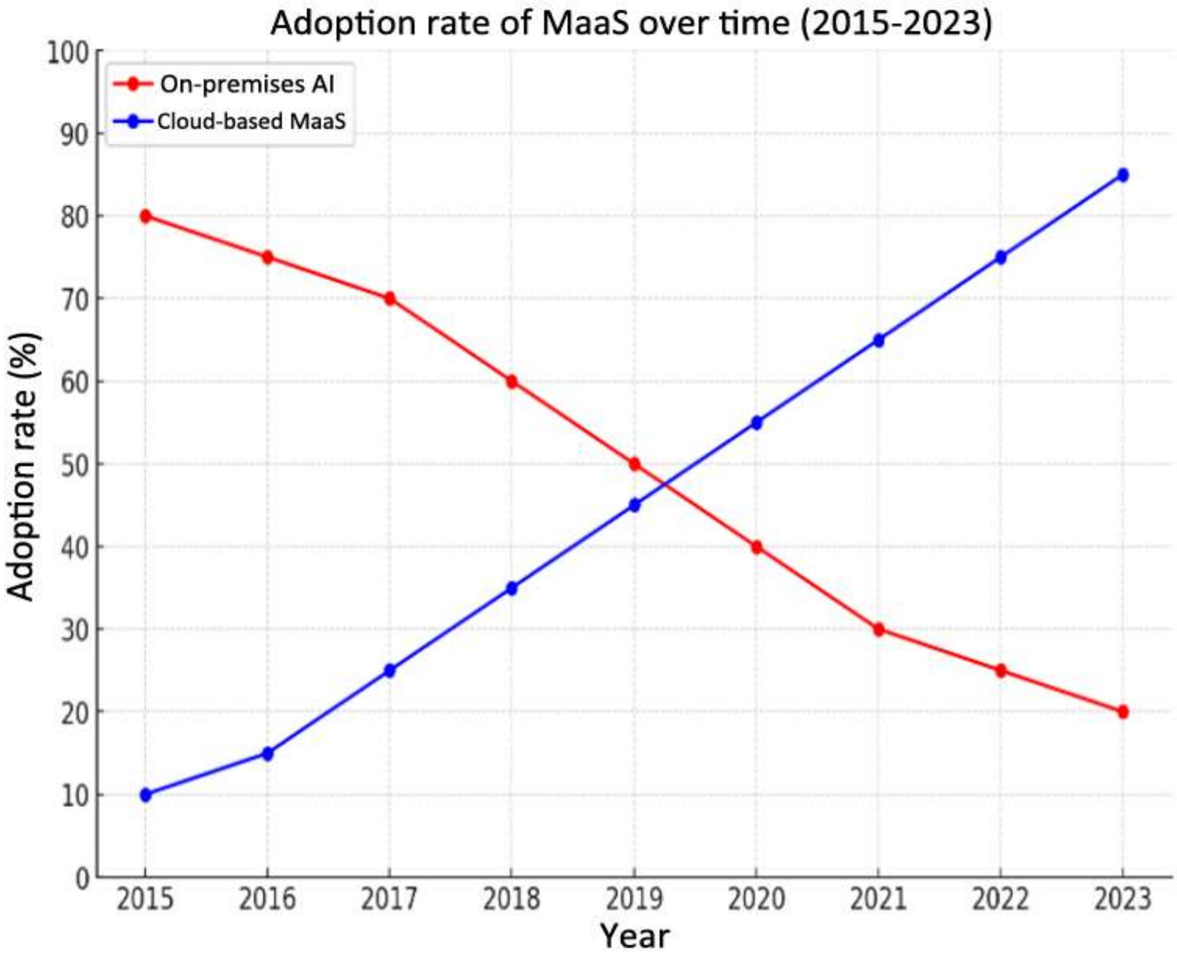

Figure 1. Adoption rate of MaaS over time.

Source: Cai, et al. [9].

2.4. Role of Maas as A Transformative Layer Within Cloud Platforms

The cloud platform transformational layer MaaS streamlines AI model deployment, scalability, and maintenance. Deploying AI models required MLOps, data engineering, and infrastructure management. Providing pre-trained or custom models as managed services streamlines this process for enterprises. Businesses may prioritize AI-driven solutions over infrastructural complexity. AI services from MaaS platforms' built-in model performance, versioning, and regulatory compliance tools are powerful and trustworthy [9]. Continuous monitoring and control make AI deployment a service that scales, optimizes, and updates models to meet needs. As cloud providers improve their MaaS offerings, businesses of all sizes will find it easier to install robust AI solutions quickly and efficiently, increasing AI adoption.

3. CORE COMPONENTS OF MODEL-AS-A-SERVICE (MAAS)

Hosting, scaling, and maintaining AI models on the cloud allows MaaS to simplify machine learning model deployment and management. MaaS allows organizations and developers to use, preserve, and deploy models as services without infrastructure or model management skills.

3.1. Key Components of Maas Architecture

API Endpoints: MaaS design includes an API layer that lets users communicate with deployed models.

Services, workflows, and applications on a variety of platforms and computer languages can integrate models thanks to APIs [10]. Web, mobile, and back-office apps may be easily linked with AI using API endpoints to feed data to models, get predictions, and more.

Model Storage: Model storage securely stores and performs versioning of trained models. In a MaaS arrangement, this storage solution permits several model versions, making A/B testing, upgrades, and rollbacks easy. Generative models typically require a large number of parameters to produce high-quality output. Thus, MaaS storage solutions use distributed or cloud-based solutions to reduce latency and increase speed.

Monitoring and Logging: Monitoring is the sole way to track model performance, utilization, and resource consumption. MaaS monitoring tools give real-time latency, API usage, response times, and prediction accuracy. These indicators help users understand the model's production performance, identify issues, and improve resource allocation. Auditing and monitoring are necessary to track outcomes, mistakes, and forecast requests.

Auto-scaling Mechanisms: Due to MaaS scaling, models can easily adjust to changing demand. By automatically adding computing capacity during peak hours, autoscaling reduces latency and response time [11]. when demand decreases, autoscaling reduces these resources to minimize costs. This versatility is crucial for computationally intensive generative models. Horizontal scaling and load balancing, where multiple models run simultaneously, let cloud providers disperse requests.

Model Management and Versioning: Version management becomes more important when models change, so testing, production, and development may use the same version. Users can deploy updates, trial different versions, and roll back if model management issues develop in MaaS. Updates improve training data and algorithms, essential for generative models.

Security: MaaS ensures the security of the model and data exchanged. Encryption, access control, and compliance checks protect sensitive or proprietary generative model data on MaaS platforms [12]. Security solutions like API key management, role-based access control (RBAC), and token-based authentication restrict access to authorized users. Data encryption at rest or in transit and Global Data Protection Regulation (GDPR) or Health Insurance Portability and Accountability Act (HIPAA) compliance assist MaaS providers in protecting sensitive data.

3.2. Model Hosting, Management, Scaling, and Security in MaaS

In MaaS, GPU- or TPU-enabled instances host AI-ready models. Memory stores the host models, making them ready to process requests in real time. Cloud-based model hosting gives enterprises on-demand access to high-performance resources, enabling faster model deployment than manual deployment and infrastructure provisioning. Managed hosting lets users focus on application integration rather than model configuration. Model management is essential for updating, managing dependencies, and maintaining models. MaaS clients control models via centralized dashboards. These dashboards show use, version control, and resource allocation [13]. MaaS makes replacing or upgrading models easy and minimizes downtime, even when new data requires retraining. Dependency monitoring in this management component keeps model versions current despite infrastructure and library changes.

MaaS cloud solutions use auto-scaling to meet variable demand. During peak hours, the system may add servers or GPUs to handle additional queries. Scaling down resources results in lower off-peak operational costs. Load balancing, which distributes requests over many model instances, is a typical scalability strategy. MaaS solutions guarantee consistent service quality and response times for enterprise-level apps and high-traffic scenarios due to their scalability.

Several approaches protect MaaS models and data [14]. Cloud service providers use firewalls, Virtual Private Networks (VPNs), and secret API endpoints to protect customer data. Token-based authentication, Internet Protocol (IP) address whitelisting, and (Advanced Encryption Standard) AES-256 data encryption protects API access and data at rest and in transit.

3.3. Lifecycle Management for Generative Models in Maas

Lifecycle services, an integral aspect of Model-as-a-Service, manage model training, deployment, versioning, and continuous improvement. This method improves model performance, update efficiency, and robustness, making generative model administration and scaling easier for businesses and developers [15]. Before training a model, it is necessary to collect data representing real-world scenarios. AWS SageMaker and Google Vertex AI offer managed deep learning environments. GPU or TPU instances enable these environments to process massive datasets and complex calculations. MaaS solutions automate deployment, allowing developers to deploy trained models with fewer configurations. To prepare the model for real-time queries and API endpoints, we ensure a smooth development-to-production transition.

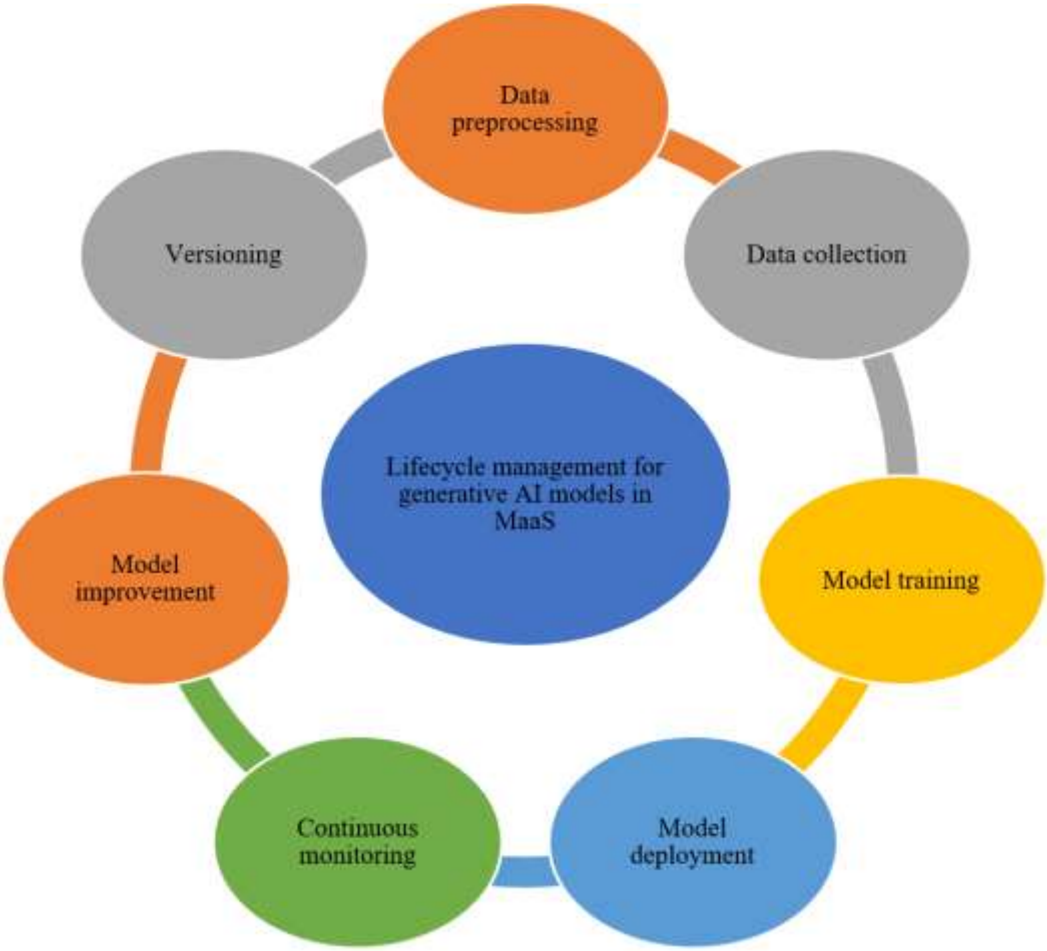

Figure 2. Lifecycle management for generative AI models in MaaS.

Versioning ensures that models remain current, and companies can update applications without disrupting them. Companies can perform A/B testing, track model iterations, and reverse with version management. The model's performance must be monitored continuously to discover faults early. The built-in analytics, usage statistics, and fault logs in MaaS systems let developers tailor models [16]. Generator models need regular adjustment, notably for transformer bias and GAN mode collapse. Looking at performance data can help businesses make generative AI solutions more accurate, scalable, and adaptable to new data and real-world demands. Figure 2 illustrates at a high level the various stages in the lifecycle of generative AI models in MaaS. This comprehensive lifecycle management method ensures correct and useful generative AI models for enterprises.

4. TECHNOLOGICAL ARCHITECTURE OF MAAS FOR GENERATIVE AI

MaaS for generative AI manages machine learning model lifecycles using a complicated cloud infrastructure, orchestration, containerization, and seamless data pipeline architecture. This architecture facilitates deploying AI models, especially generative models that use a lot of processing power, efficiently, scalably, and safely. This section discusses cloud-based generative AI service security, MaaS architecture, orchestration, containerization, and data storage and pipeline connections.

4.1. Overview of Cloud Infrastructure for Maas (Compute, Storage, Networking)

To host, manage, and deploy models at scale, MaaS for generative AI uses cloud infrastructure. This design often includes computing, storage, and networking.

For training and real-time predictions, GANs, VAEs, and transformers need lots of computer resources. Machine learning computer resources are available from AWS, Google Cloud, and Microsoft Azure [17]. GPU- or TPU-based VMs accelerate inference and training via parallel processing. Scaling up or down cloud compute instances according to demand provides flexibility and cost-effectiveness.

MaaS storage includes models and data. Trained AI models may require a few megabytes to several gigabytes or even terabytes of model repository storage, depending on their complexity. Amazon S3, Google Cloud Storage, and Azure Blob Storage allow businesses to securely store models. With these storage techniques, we may simply edit or revert a model and back it up at any moment. Data storage is crucial for managing huge generative model datasets. Cloud platforms offer Hadoop, Apache Hadoop Distributed File System (HDFS), and other distributed file systems, as well as Structured Query Language (SQL) and NoSQL managed databases for big datasets.

MaaS networking is essential for AI models and services to communicate across systems. Cloud systems leverage high-speed, low-latency networks to provide real-time model predictions [18]. Networking technologies like load balancers and virtual private clouds allow model hosting, data pipelines, and external APIs in the MaaS ecosystem to safely communicate. Content Delivery Networks (CDNs) typically send model predictions globally. This ensures low-latency model access worldwide.

4.2. Containerization (Docker, Kubernetes) and Orchestration in Maas

MaaS generative AI model deployment, scalability, and administration depend on containerization and orchestration.

Docker: Docker lets developers package software with all their dependencies, libraries, settings, and environments into small, portable containers. These containers run dependably across computing environments, simplifying model development, testing, and deployment. Docker makes it easy to deploy generative AI models to MaaS cloud settings without compatibility problems [19]. Encapsulating models in Docker containers ensures flawless operation across all cloud providers and infrastructures.

Kubernetes: Open-source Kubernetes automates containerized application deployment, scalability, and management. Kubernetes orchestrates AI model deployment and scaling in a decentralized, ever-changing cloud scenario as part of MaaS. Kubernetes distributes workloads over many containers for maximum availability and resource efficiency. To handle a generative model's high demand, Kubernetes can dynamically deploy extra containers. It can also self-heal and assess health to automatically replace failed containers. Kubernetes' rolling update functionality lets businesses deploy new model versions without disturbance. Kubernetes keeps previous model versions running during updates, ensuring model service users a smooth experience.

4.3. Integration of Maas with Data Storage, Preprocessing, and Data Pipelines

MaaS integrates data pipelines, storage, and preprocessing for efficient data management and model inference.

Data Storage: Data lakes and cloud databases let MaaS store data. These systems hold massive amounts of training data and model inference input and output data. Amazon Redshift, Google Cloud BigQuery, and Azure Data Lake store both structured and unstructured data. MaaS systems use various storage options to keep data safe, structured, and accessible and make model training and updating easier [20].

Data Preprocessing: Preprocessing raw data is essential for model training and inference. This may comprise data cleansing, normalization, feature extraction, and augmentation. Automation using cloud-based workflows and AI models is prevalent in MaaS. AWS Glue, Google Cloud Dataflow, and Azure Data Factory prepare massive datasets for efficient and scalable data pipelines. These services let organizations batch or real-time process data for generative AI model training or inference.

Data Pipelines: A MaaS "data pipeline" automates data flow from databases, IoT devices, and APIs to the model. MaaS workflows employ AWS SageMaker, Google AI Platform, and Azure ML for model training and inference [21]. These pipelines update models with new data. They simplify new-data model retraining and fine-tuning.

4.4. Security Considerations in Cloud-Based Maas for Generative AI Models

With generative AI models processing sensitive data, Model-as-a-Service security is even more important. Cloud data and models must be protected for availability, integrity, and privacy. Security techniques include RBAC to restrict model and resource access to authorized users. MaaS endpoints employ API keys, OAuth tokens, or multi-factor authentication (MFA) for user authentication. AES-256 encrypts sensitive data in transit and at rest. Hypertext Transfer Protocol Secure (HTTPS) and Transport Layer Security (TLS) protect API data over networks. Isolating resources and constraining MaaS workloads in VPCs improves network security. Infiltration can be stopped by firewalls, intrusion detection systems (IDS), and secure VPNs. MaaS platforms protect models during transfer with digital signatures and checksums [22]. Server availability and integrity are ensured by regular backups and replication across geographically dispersed server centers These security features make MaaS the best option for enterprises and developers concerned about generative AI models and sensitive data.

5. KEY ADVANTAGES OF MAAS FOR GENERATIVE AI ON CLOUD PLATFORMS

MaaS for generative AI on cloud platforms improves commercial, developer, and organizational capabilities. Using cloud infrastructure, MaaS simplifies sophisticated AI model deployment and management. Scalability, cost efficiency, cutting-edge hardware, and distributed creativity and cooperation are available.

5.1. Scalability and Flexibility for Developers and Enterprises

MaaS's scalability and flexibility help generative AI. Cloud MaaS platforms can scale to train models on massive datasets or provide real-time predictions for AI applications. AWS, Google Cloud, and Microsoft Azure offer on-demand storage, bandwidth, and computing. Scalability allows organizations to match cloud resources to workload [20]. Model training can strain the system computationally, whereas inference after deployment saves resources. MaaS systems support different cloud regions, model designs, and deployment situations. After testing GANs, VAEs, and transformers, developers can use TensorFlow or PyTorch.

5.2. Cost-Effectiveness And Operational Efficiency

SaaS platforms let companies pay for resources, saving money and enhancing efficiency.

This pay-as-you-go model enables any size company to use generative AI without pricey equipment. MaaS enables firms pay only for the computing, storage, and network resources they use, saving them a lot on on-premises infrastructure. Usage-based pricing from cloud service providers helps enterprises optimize spending. MaaS technologies automate model deployment and maintenance tasks like patching, scaling, and load balancing. Developers can focus on model optimization and application creation instead of infrastructure administration. MaaS is also compatible with Continuous Integration/Continuous Delivery (CI/CD) technologies, which speeds operations and permits rapid model changes and updates without production service interruptions.

5.3. Access to Advanced Hardware (GPUs, TPUs) and Support for Resource-Intensive Models

Generative AI models, especially deep learning ones, require GPUs or TPUs for efficient operation. MaaS systems on cloud services let developers and enterprises use modern hardware without investing in expensive on-premises infrastructure. AWS, Google Cloud, and Microsoft Azure offer powerful GPUs and TPUs for on-demand machine learning model execution. GPUs are great for deep neural network training due to their better parallel processing capabilities [23]. TPUs excel in AI tasks such as training large generative models like GANs and transformers, owing to their design for tensor-based computations.

With MaaS, developers can choose hardware type and size based on the workload. Allocating high-performance GPUs or TPUs speeds up resource-intensive model training, such as huge transformers. Real-time inference with trained models on less resource-intensive instances yields high performance and cost efficiency. GANs for large-scale image production and GPT-like transformers for natural language synthesis are computationally intensive, but enterprises can install and run them with on-demand high performance computing (HPC) resources. This access allows businesses without the means or knowledge to build their own AI infrastructure to participate.

5.4. Enhancements to Collaboration and Innovation in Distributed Environments

MaaS lets teams, regions, and organizations share and deploy generative AI models on one platform to collaborate better. Developers, data scientists, and business teams may collaborate better on cloud-based MaaS platforms since they can experiment and construct models together. Version control, shared repositories, and integrated workflows help teams interact easily across locations. MaaS gives organizations access to cloud providers' latest AI tools and services, helping them lead AI innovation. Cloud platforms may integrate with cutting-edge frameworks, pre-trained models, and APIs to let developers create new generative AI applications [8]. Deploying creative models to test in real-world situations lets developers quickly try new ideas.

6. CHALLENGES AND LIMITATIONS OF MAAS FOR GENERATIVE AI

MaaS has pros and cons for cloud-based generative AI. These concerns might hinder MaaS adoption, especially for firms and developers using AI in critical or regulated environments. Here are some of MaaS's biggest generative AI concerns.

6.1. Security and Privacy Concerns with Data and Model Storage

Cloud-based MaaS raises privacy and security concerns due to the sensitive data and models utilized in generative AI. Many AI applications train and infer sensitive data, like personal, financial, and organizational secrets. Without sufficient security, cloud storage exposes this data to breaches. Despite cloud platforms' strong encryption and security, businesses may be wary of entrusting critical data to third parties. Theft or unlawful access to AI models, especially generative ones, can threaten enterprises because of their intellectual property value [24]. In adversarial attacks, hostile users try to manipulate the generative model into biased or erroneous results or steal data. GDPR, HIPAA, and California Consumer Privacy Act (CCPA) may be challenging to comply with in MaaS. Businesses using MaaS must ensure their cloud providers comply with these criteria to avoid legal issues. These regulations severely limit data usage, storage, and transportation. Organizations should work with cloud providers to implement regulatory-compliant security solutions to address these risks.

6.2. Latency and Performance Issues in Large-Scale Applications

Latency and performance drastically impact real-time generative AI model performance. Artificial intelligence applications that analyze real-time data, such as live customer interactions, natural language processing, and photo or video production, require low latency. Even with high-performance cloud platforms, sending data between on-premises systems and cloud servers can cause network delays, especially if the user base is geographically dispersed. These delays can impair generative AI models' reactivity, resulting in poor user experiences. Despite strong cloud settings, pooled resources or unexpected traffic spikes might affect performance [25]. Due to cloud infrastructure stress, MaaS AI model performance may affect response times, especially during high demand.

These variances could pose challenges for organizations with large-scale applications that require consistent performance. MaaS providers may offer CDNs, edge computing, and multi-region installations to improve performance and latency. However, these systems may need more configuration and monitoring.

6.3. Limitations in Customization and Control Over the Models

Many customers benefit from MaaS solutions abstracting the difficulty of creating and operating AI models. However, this abstraction may limit organizations' model customization and governance. MaaS platforms' pre-built models and tools may not meet businesses' needs or performance goals. Pre-trained models' architecture, training data, and behavior may be harder to customize than with an on-premises solution, where the business controls model training and deployment. Business models and training pipelines may not be accessible on cloud systems [15]. Sectors with sensitive or proprietary use cases or heavily modified models may struggle with this lack of control. MaaS systems are general-purpose, making it difficult to create custom AI algorithms for specific applications.

6.4. Regulatory and Compliance Challenges

In regulated industries like healthcare, finance, and law, MaaS platforms make generative AI model deployment challenging. Cloud providers operating across multiple geographies may be in violation of data sovereignty laws, which restrict data storage in specific locations. Businesses must confirm their MaaS provider can handle data residency. Industry-specific compliance is another issue. To protect data privacy and model biases, many sectors have strong AI standards. These include HIPAA in healthcare and Basel III and MiFID II in finance. Many domains seek open, traceable AI systems; thus, humans must comprehend model decisions and training data [26]. MaaS companies may struggle to meet audit and regulatory standards owing to lack of control and transparency. These disciplines need thorough recordkeeping and transparency.

For generative AI, MaaS offers scalability, cost-effectiveness, and cutting-edge hardware, but there are downsides. The list includes security, privacy, performance, latency, model customization and control, and legal requirements. Consider the cloud platform's functionality, security, and industry or corporate demands to overcome these issues. By addressing these barriers, organizations can leverage MaaS for generative AI while maintaining security, performance, and compliance. Table 1 shows MaaS offerings by cloud providers, their key features to note, pricing model, strengths, and weaknesses.

Table 1. Overview of leading MaaS offerings for generative AI.

Cloud platform	MaaS offering	Key features	Unique tools	Pricing	Strengths	Weaknesses
AWS	Amazon SageMaker	Managed service for building and deploying AI models	Amazon Polly, deep learning AMIs	Pay-as-you-go, based on usage	Scalable, large model support	Complex pricing, requires AWS expertise
Google cloud	Vertex AI	Unified environment for building, deploying, scaling AI	TensorFlow extended, vertex AI workbench	Pay-as-you-go based on usage	Optimized for TensorFlow, strong AI tools	Limited model selection outside Google ecosystem
Microsoft Azure	Azure machine learning	Comprehensive platform for building and deploying AI	Azure cognitive services, AI gallery	Pay-as-you-go, subscription available	Integration with Microsoft products	Performance bottlenecks for large workloads
IBM cloud	Watson studio	AI development for building and deploying AI models	Watson machine learning, OpenScale	Custom pricing	Strong NLP and industry-specific tools	Less scalable for large AI workloads
Oracle cloud	Oracle cloud AI services	Managed services with Oracle integration	Oracle cloud infrastructure, digital assistant	Based on usage	Strong enterprise database integration	Limited pre-trained models, less AI advanced
Alibaba cloud	Alibaba cloud AI	AI training and inference services	Pai, Max Compute	Pay-as-you-go based on usage	High scalability, tailored for Asia pacific region	Limited global ecosystem and integration

Table 2. Comparative analysis of features, pricing, and performance.

Feature	Amazon SageMaker (AWS)	Vertex AI (Google cloud)	Azure machine learning (Microsoft)	IBM Watson studio
Model training	Fully managed, supports popular frameworks like TensorFlow, MXNet, PyTorch.	Easy integration with TensorFlow and Keras.	Supports popular frameworks and auto ML for easy model training.	Integrated AI tools with a focus on NLP and cognitive services.
Model deployment	Auto-scaling and multi-availability zone deployment.	Fully managed deployment with multi-cloud integration.	Managed deployment with Kubernetes, Docker support.	Managed deployment with scalability and monitoring.
Pre-trained Models	Wide range for NLP, image, and video tasks.	TensorFlow, custom models with Google pre-trained models.	Strong set of pre-built models for language and vision tasks.	Focus on NLP models, chatbots, and industry-specific AI solutions.
Integration with other services	Tight integration with AWS ecosystem (e.g., EC2, S3, Lambda).	Deep integration with Google's ecosystem (e.g., BigQuery, Dataflow).	Seamless integration with Microsoft products like power BI, Azure DevOps.	Focus on integration with IBM enterprise solutions.
Cost flexibility	Highly flexible, pay-per-use, can become expensive at scale.	Pay-per-use with flexible pricing for training and hosting.	Pay-as-you-go with extensive pricing calculators.	Custom pricing, but can be expensive for large workloads.
Security & compliance	High security with encryption and compliance certifications.	Strong security with Google's enterprise-grade encryption.	Security and compliance with industry standards like HIPAA, GDPR.	High security with an emphasis on data privacy.

6.5. Popular Maas Offerings and Comparison

MaaS services from AWS, Google Cloud, Microsoft Azure, and International Business Machines (IBM) Cloud offer several generative AI model installation functionalities. Each platform has strengths and cons based on use case, price, and infrastructure. AWS and Google Cloud offer scalability and AI tools, but Azure integrates better with Microsoft-focused enterprises. We have put together comparative analysis of popular MaaS service offerings on cloud platforms across a list of features supported in Table 2.

IBM Cloud performs well in NLP but poorly in scalability. The best MaaS supplier depends on business needs, budgets, and technological skills.

7. USE CASES OF MAAS FOR GENERATIVE AI ACROSS INDUSTRIES

Healthcare MaaS enables medical imaging, individualized treatment regimens, and generative AI drug discovery models. Researchers can utilize MaaS to train models with fake medical pictures to increase diagnostic accuracy without compromising patient privacy.AI models enabled by platforms like AWS SageMaker help doctors predict patient outcomes and uncover new treatment candidates, speeding up and lowering drug development costs [27]. MaaS automates market prediction, fraud detection, and financial reporting in banking. Generative Adversarial Networks can generate synthetic market data for financial system stress testing. Employing predictive models for risk assessment, financial institutions can use MaaS technologies like Azure's AI to extend compute power without managing complex infrastructure. Figure 3 depicts use cases of MaaS in different industries.

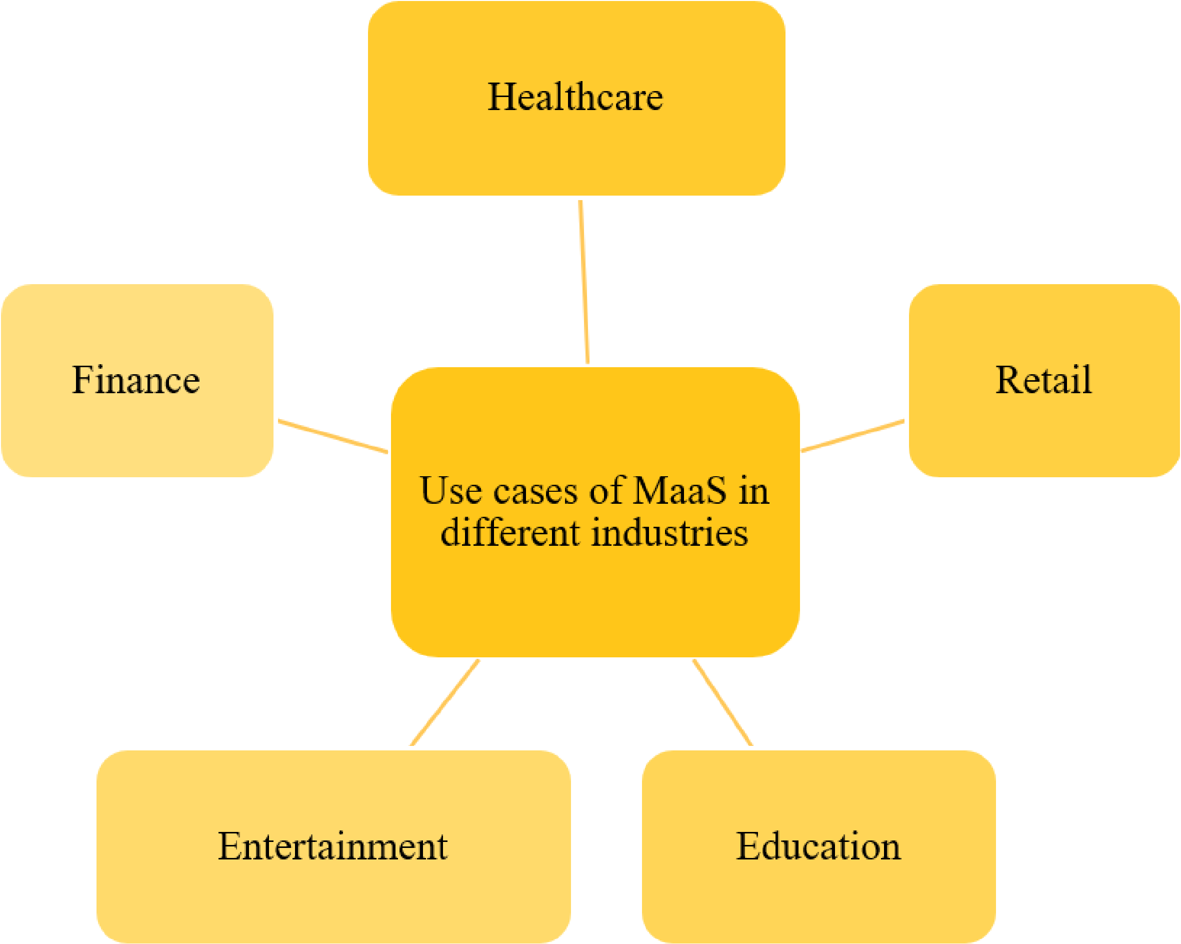

Figure 3. Use cases of MaaS in different industries.

MaaS is transforming retail by enabling more tailored consumer experiences. Retailers can develop virtual shopping experiences and personalized product suggestions using MaaS-enabled generative AI models. Marketers may utilize Google Cloud's Vertex AI to create real-time recommendation systems to boost user engagement and sales. The entertainment sector uses MaaS to generate scripts, songs, and videos. AI models can edit videos, produce audio tracks, and create convincing Computer Generated Imagery (CGI) figures. Generational models hosted on MaaS platforms like IBM Watson Studio have created dynamic screenplays and other digital media, simplifying and cost-effectively developing material [28].

MaaS in education can construct intelligent tutoring and adaptable learning models. For instance, generative AI may leverage student performance data to create unique learning routes, and AI-driven platforms like Google Cloud can automate grading and feedback output, freeing teachers to focus on other tasks. These apps help schools run efficiently and improve student learning. MaaS lets companies in various industries use cutting-edge AI models on a massive scale, enhancing innovation, efficiency, and savings.

8. FUTURE OF MAAS FOR GENERATIVE AI AND EMERGING TRENDS

Several new AI trends are expected to enhance the generative AI MaaS.

AI systems using generative models without coding are rising rapidly. More people are now using Generative AI for content creation, automation, and bespoke experiences, without the need for data scientists or technical teams. Future MaaS will stress customization and integration. Businesses can use improved training frameworks, model versioning, and seamless connection with proprietary data pipelines to develop generative AI models to their specifications. This enhanced flexibility will enable enterprises to create bespoke AI solutions with better efficacy and model generation and performance control. As these trends evolve, MaaS will establish its place in the AI ecosystem, allowing businesses to use generative AI to its fullest, most ethical potential.

9. CONCLUSION

We explored MaaS's revolutionary potential in generative AI, concentrating on its integration with AWS, Google Cloud, and Microsoft Azure.

We've seen how MaaS simplifies deploying and scaling complicated AI models, letting companies innovate without infrastructure management. GANs, VAEs, and transformers are available in the cloud for healthcare, finance, retail, and entertainment applications. We've highlighted MaaS's scalability, affordability, access to powerful hardware (GPUs, TPUs), and potential to inspire cooperation and creativity. Best practices include choosing the correct MaaS system, improving model performance, and satisfying security needs, as well as latency, model customization, and security. Ethical AI, edge computing, and low-code/no-code platforms shape MaaS. Growing use of generative AI makes it more accessible to developers and organizations worldwide. MaaS could change business adoption of cloud-based generative AI. MaaS's robust tools and scalable solutions will increase global industry productivity, innovation, and growth. Businesses and programmers should research, personalize, and monitor MaaS and generative AI news and trends.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable.
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] W. Gan, S. Wan, and S. Y. Philip, "Model-as-a-service (MaaS): A survey," presented at the In 2023 IEEE International Conference on Big Data (BigData) (pp. 4636-4645). IEEE, 2023.

[2] V. Liagkou, E. Filiopoulou, G. Fragiadakis, M. Nikolaidou, and C. Michalakelis, "The cost perspective of adopting large language model-as-a-service," presented at the In 2024 IEEE International Conference on Joint Cloud Computing (JCC) (pp. 80-83). IEEE, 2024.

[3] E. La Malfa *et al.*, "Language-models-as-a-service: Overview of a new paradigm and its challenges," *Journal of Artificial Intelligence Research*, vol. 80, pp. 1497-1523, 2024. https://doi.org/10.1613/jair.1.15865

[4] E. La Malfa *et al.*, "The ARRT of language-models-as-a-service: Overview of a new paradigm and its challenges," *arXiv preprint arXiv:2309.16573*, 2023.

[5] H. Wang, M. Liu, and W. Shen, "Industrial-generative pre-trained transformer for intelligent manufacturing systems," *IET Collaborative Intelligent Manufacturing*, vol. 5, no. 2, p. e12078, 2023. https://doi.org/10.1049/cim2.12078

[6] J. Duarte *et al.*, "FPGA-accelerated machine learning inference as a service for particle physics computing," *Computing and Software for Big Science*, vol. 3, pp. 1-15, 2019. https://doi.org/10.1007/s41781-019-0027-2

[7] C. Xue *et al.*, "OmniForce: On human-centered, large model empowered and cloud-edge collaborative AutoML system," *arXiv preprint arXiv:2303.00501*, 2023.

[8] S. Xue *et al.*, "Db-gpt: Empowering database interactions with private large language models," *arXiv preprint arXiv:2312.17449*, 2023.

[9] Z. Cai, R. Ma, Y. Fu, W. Zhang, R. Ma, and H. Guan, "LLMaaS: Serving large language models on trusted serverless computing platforms," *IEEE Transactions on Artificial Intelligence*, 2024. https://doi.org/10.1109/tai.2024.3429480

[10] L. Yu, Q. Chen, J. Lin, and L. He, "Black-box prompt tuning for vision-language model as a service," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence Aug*, 2023, pp. 1686-1694.

[11] Q. Yiting, K. Munir Hayet, Z. Shuqing, C. SiYuan, and C. Choonkit, "Enhancing sustainability in academic guidance: Develop an ai-driven agent for education 5.0," *Inti Journal*, vol. 2024, no. 40, pp. 1-10, 2024.

[12] Y. Qiu, S. Chen, and W. Y. Leong, "Creating a" ready-to-use" ai agent for navigating digital platform to enhance collaborative efficiency," *INTI Journal*, vol. 2024, 2024. https://doi.org/10.61453/intij.202422

[13] L. Yue and T. Chen, "AI large model and 6G network," in *In Proc. 2023 IEEE Globecom Workshops (GC Wkshps), Dec. 2023, pp. 2049-2054*, 2023.

[14] X. Shen *et al.*, "A split-and-privatize framework for large language model fine-tuning," *arXiv preprint arXiv:2312.15603*, 2023.

[15] A. Sabbatella, A. Ponti, I. Giordani, A. Candelieri, and F. Archetti, "Prompt optimization in large language models," *Mathematics*, vol. 12, no. 6, p. 929, 2024. https://doi.org/10.3390/math12060929

[16] A. Lekova, P. Tsvetkova, T. Tanev, P. Mitrouchev, and S. Kostova, "Making humanoid robots teaching assistants by using natural language processing (NLP) cloud-based services," *Journal of Mechatronics and Artificial Intelligence in Engineering*, vol. 3, no. 1, pp. 30-39, 2022. https://doi.org/10.21595/jmai.2022.22720

[17] M. Zhang, X. Pan, and M. Yang, "Jade: A linguistics-based safety evaluation platform for llm," *arXiv preprint arXiv:2311.00286*, 2023.

[18] S. Sheth, H. P. Baker, H. Prescher, and J. A. Strelzow, "Ethical considerations of artificial intelligence in health care: Examining the role of generative pretrained transformer-4," *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, p. 10.5435, 2022.

[19] G. De Seta, "Synthetic probes: A qualitative experiment in latent space exploration," *Sociologica*, vol. 18, no. 2, pp. 9-23, 2024.

[20] Q. Zheng, J. Wang, and Y. Shen, "Overview and trend of large-scale model deployment mode," in *In 2024 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) (pp. 1-6). IEEE*, 2024.

[21] H. Wang *et al.*, "A data and knowledge driven autonomous intelligent manufacturing system for intelligent factories," *Journal of Manufacturing Systems*, vol. 74, pp. 512-526, 2024. https://doi.org/10.1016/j.jmsy.2024.04.011

[22] J. Duan, S. Qian, D. Yang, H. Hu, J. Cao, and G. Xue, "MOPAR: A model partitioning framework for deep learning inference services on serverless platforms," *arXiv preprint arXiv:2404.02445*, 2024.

[23] S. Zhang, M. Xu, W. Y. B. Lim, and D. Niyato, "Sustainable AIGC workload scheduling of geo-distributed data centers: A multi-agent reinforcement learning approach," presented at the In GLOBECOM 2023-2023 IEEE Global Communications Conference (pp. 3500-3505). IEEE, 2023.

[24] J. Liang, A. Zhao, S. Hou, F. Jin, and H. Wu, "A GPT-enhanced framework on knowledge extraction and reuse for geographic analysis models in Google earth engine," *International Journal of Digital Earth*, vol. 17, no. 1, p. 2398063, 2024. https://doi.org/10.1080/17538947.2024.2398063

[25] J. Kim *et al.*, "Camp: Community agricultural model platform for multi-model ensemble simulations of crop growth and development," *Available at SSRN 4869988*, 2024. https://doi.org/10.2139/ssrn.4869988

[26] P. Balasubramanian, S. Nazari, D. K. Kholgh, A. Mahmoodi, J. Seby, and P. Kostakos, "TSTEM: A cognitive platform for collecting cyber threat intelligence in the wild," *arXiv preprint arXiv:2310.03541*, 2023.

[27] E. Yang *et al.*, "Continual learning from a stream of APIs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. https://doi.org/10.1109/tpami.2024.3460871

[28] N. Fip, "The new business of AI software vendors in the European manufacturing industry: An empirical study on business models of entrepreneurial AI software vendors," Ph.D. Dissertation, Vienna, 2021.

Views and opinions expressed in this article are the views and opinions of the author(s), Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.