# Model-as-a-Service (MaaS): A Survey

Wensheng Gan<sup>1,2</sup>, Shicheng Wan<sup>3\*</sup>, Philip S. Yu<sup>4</sup>

<sup>1</sup>Jinan University, Guangzhou 510632, China

<sup>2</sup>Pazhou Lab, Guangzhou 510330, China

<sup>3</sup>South China University of Technology, Guangzhou 510641, China.

<sup>4</sup>University of Illinois Chicago, Chicago, IL 60637, USA

Email: {wsgan001, scwan1998}@gmail.com, psyu@uic.edu

Abstract—Due to the increased number of parameters and data in the pre-trained model exceeding a certain level, a foundation model (e.g., a large language model) can significantly improve downstream task performance and emerge with some novel special abilities (e.g., deep learning, complex reasoning, and human alignment) that were not present before. Foundation models are a form of generative artificial intelligence (GenAI), and Model-as-a-Service (MaaS) has emerged as a groundbreaking paradigm that revolutionizes the deployment and utilization of GenAI models. MaaS represents a paradigm shift in how we use AI technologies and provides a scalable and accessible solution for developers and users to leverage pre-trained AI models without the need for extensive infrastructure or expertise in model training. In this paper, the introduction aims to provide a comprehensive overview of MaaS, its significance, and its implications for various industries. We provide a brief review of the development history of "X-as-a-Service" based on cloud computing and present the key technologies involved in MaaS. The development of GenAI models will become more democratized and flourish. We also review recent application studies of MaaS. Finally, we highlight several challenges and future issues in this promising area. MaaS is a new deployment and service paradigm for different AI-based models. We hope this review will inspire future research in the

Index Terms—foundation models, artificial intelligence, generative AI, Model-as-a-Service, review

# I. INTRODUCTION

We currently live in an era of fast-developing within big data [1], [2], artificial intelligence (AI) [3], and Web 3.0 [4], [5], organizations within various industries are increasingly leveraging the power of machine learning (ML) models [6] to gain insights, automate processes, and make data-driven decisions. However, developing, training, and deploying ML models are challenging and resource-intensive. Especially in the digital age, generative artificial intelligence, e.g., large language model (LLM) [7], has recently not only become a public-familiar word on social media but also a hot spot in the academic field. Generative artificial intelligence (GenAI) [8] refers to the ability of a machine learning model to generate new data that is similar to the training data. This is in contrast to the traditional AI approach, which is designed to recognize differences between distinct classes or categories of data. GenAI models are often adopted in tasks such as image generation, text generation [9], and music composition [10]. LLM is an advanced model within generative AI that aims to simulate human language ability and intelligence (i.e., generation capability). Developers and engineers train LLM using massive amounts of textual data and then utilize deep learning algorithms [11], [12], [13] to understand as many languages as possible. Indeed, the training process is the core of LLM [14], [15]. It involves training the model on huge amounts of data to learn the statistical patterns and semantic relationships of language. During the training process [15], LLM predicts the next word or sentence to optimize its parameters and improve its predictive abilities. Through iterative training, LLM gradually enhances its language generation and understanding quality. Benefiting from generating human language text capability, LLM has become an important service in various domains. It provides people with a lot of useful unexpected functions and outstanding applications, such as data analytic [16], [17], natural language processing (NLP) [18], intelligent conversation system [19], machine translation [20], information retrieval [21], and text-to-multimodal [22].

Since the foundation model is still in its exploration stage, there is no clear definition and consensus in academia and industry [23], [24]. This is where Model-as-a-Service (MaaS) comes into play. MaaS is a cloud computing-based service framework that offers AI and ML models and related infrastructure as a service to developers and businesses. It provides a convenient and cost-effective way to access and utilize large models without the need for extensive knowledge or infrastructure. MaaS allows users to leverage pre-trained ML models and algorithms through simple interfaces, application programming interfaces (APIs), or software development kits (SDKs). MaaS allows users to access functions of the large model through calling API without the need to train and maintain complex models themselves.

In simple terms, MaaS is a new business model. As the foundation infrastructure of the AI era, MaaS provides secure, efficient, and cost-effective model usage and development support for downstream applications. Looking at the entire industry structure of MaaS, the core idea follows the path of "Model–Single Point Tool-Application Scenarios". Users can directly call, develop, and deploy models in the cloud without the need to invest in building and maintaining the infrastructure, hardware, and specialized knowledge required for their own models. Large models, as a critical component

<sup>\*</sup>Corresponding author.

of MaaS, are a product of the combination of "high computing power and strong algorithms" and will be a development trend for the future of artificial intelligence. Large models are rapidly evolving, transitioning AI from a "craft workshop" to a "factory model", achieving greater versatility and intelligence, and enabling AI to empower applications across various industries more widely. MaaS finds wide applications across multiple domains. In intelligent conversational systems, MaaS can be used to build chatbots, voice assistants, and other interactive systems that engage in smooth and natural conversations with users. In information retrieval, MaaS can provide intelligent search and recommendation functionalities, helping users quickly find the desired information. However, the development of MaaS also faces some challenges. Firstly, there are performance and stability considerations. Large models require significant training data and computational resources to achieve optimal performance, and they need to address limitations in tasks such as long-text generation and logical reasoning. Secondly, there are privacy and security concerns as MaaS may involve users' personal information and sensitive data, requiring appropriate security measures in data transmission and privacy protection.

In the growing landscape of AI and ML, MaaS is emerging as a groundbreaking paradigm that revolutionizes the deployment and utilization of AI models. MaaS provides a scalable and accessible solution for organizations and developers to leverage pre-trained AI models without the need for extensive infrastructure or expertise in model training. However, there has not yet been a comprehensive overview of MaaS proposed in the academic. This paper aims to introduce MaaS by highlighting key components of the MaaS model, their underlying technologies, the advantages of MaaS, and the differences from previous cloud computing-based service models.

**Contributions**: To fill this research gap, this paper tries to provide a comprehensive literature survey of MaaS, and the contributions are as follows:

- This is the first review to introduce the characteristics of the interaction and related technologies of MaaS, and we also review the history of cloud computing services.
- We analyze differences between traditional and GenAI model-based technology stacks and conclude the latter stack presents many new functions and features.
- We also highlight some promising application fields within MaaS that are currently developing or will be popular in the near future.
- We analyze the challenges and issues encountered by MaaS in more detail, including utilization constraints, model interpretability, technological ethics, and morality.

**Roadmap:** We first introduce the history of previous cloud-based service platforms in Section II. We analyzed the relationship between MaaS and previous XaaS in Section III, and listed the benefits of MaaS in Section V. We then introduce some key technologies related to MaaS in Section IV and present several case applications in Section VI. We also discuss several challenges and issues faced by MaaS in Section VII.

Finally, Section IX concludes this survey with a discussion of potential future research. Note that Table I presents some basic symbols in this survey.

TABLE I: Summary of symbols and their explanations

Symbol	Definition
AI	Artificial intelligence
GenAI	Generative artificial intelligence
ML	Machine learning
LLM	Large language model
NLP	Natural language processing
API	Application programming interface
SaaS	Software-as-a-Service
PaaS	Platform-as-a-Service
IaaS	Infrastructure-as-a-Service
MaaS	Model-as-a-Service

# II. HISTORY OF XAAS

When reviewing the development history of Web 2.0, the "X-as-a-Service" model paradigm (XaaS) is typical of the second wave [25]. Indeed, since the complex and resource-intensive disadvantages of developing environment construction, the core target of XaaS is to provide users with infrastructure and then let them focus on solution design.

Software-as-a-Service (SaaS) [26] is one of the earliest cloud computing service models, which originally emerged between the late 1990s and early 2000s. SaaS delivers internetbased applications and then allows users to use these applications according to the online subscription method. This new kind of Internet service lets users no longer need to install local software, such as BaiduNetdisk, Dropbox, and Gmail. In general, SaaS applications are capable of holding abundant users because of their good scalability. The SaaS provider handles almost all maintenance tasks, including updates, bug fixes, security, and data backups, while users are only permitted to do human-software interaction. However, since the number of users involved in the Internet world is continuously rising, the rise of personalized customization for different users (i.e., more software products in various fields) has led to an explosion of requirements and then formed an urgent issue.

Platform-as-a-Service (PaaS) [27] emerged after SaaS (around the mid-2000s). It provides program developers with a stable cloud computing platform for building, testing, and deploying applications. In PaaS solutions, a provider hosts both the hardware and software on their dedicated infrastructure and delivers this platform to users as an all-in-one solution, software stack, or service accessible through the Internet. Developers can primarily concentrate on coding, relieving them of the concerns related to maintaining and managing the foundational infrastructure or platform traditionally linked to the development process. This paves the path for continued development and innovation with minimal interruptions, simultaneously diminishing the necessity for extensive infrastructure setup and coding efforts. Google's App Engine application platform [28], [29] and Microsoft's Azure app service [30] are two classic examples of PaaS. PaaS providers are responsible for maintaining the underlying

infrastructure, while users only focus on application development and management. Nevertheless, because of operational costs, PaaS may impose limitations on the different resources required for applications and workloads (e.g., computational resources, storage resources, and network resources). In some cases, specific application requirements might not be compatible with the PaaS environment, which could result in issues when deploying and running applications.

To provide users with greater flexibility, control, and adaptability to meet various application requirements and business scenarios, the latest model in cloud computing services, i.e., Infrastructure-as-a-Service (IaaS) [31], emerged around 2010. IaaS provides virtualized computing resources through web services, which allow users to rent infrastructure services such as servers, networking, and storage on-demand. Generally, users can utilize the infrastructure through application programming interfaces (APIs). Users obtain greater control over the configuration of fundamental infrastructure (e.g., virtual servers, storage, and networking) and take over responsibilities previously managed and maintained by PaaS providers within the IaaS model. The IaaS providers only focus on managing hardware modules and dealing with hardware issues. Engineers have developed many mature IaaS platforms (e.g., Google Compute Engine [32], DigitalOcean [33], and Amazon Elastic Compute Cloud [34]). These platforms have been deeply integrated into our daily lives. Yet, SaaS offers broader applications to users and various non-technical departments, whereas PaaS and IaaS predominantly cater to development teams. Fig. 1 depicts the change path of XaaS.

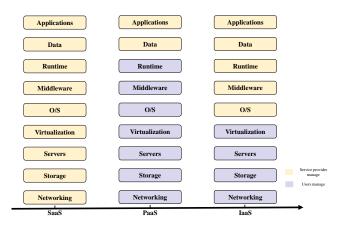


Fig. 1: A comparison between the infrastructures involved by the user and service provider. (Source: https://www.redhat.com/en/topics/cloud-computing/what-is-paas)

## III. RELATIONSHIP BETWEEN OTHER XAAS

Similar to the previous XaaS models, in the forthcoming Web 3.0 era, MaaS is a new service model in the field of AI that involves providing access to pre-trained machine learning models as a service on the Internet. It involves hosting, managing, and giving developers access to models that have already been trained through APIs. This lets developers add AI functions to their own systems and apps. The target behind

MaaS is to provide a platform where developers can access pre-trained machine learning models that have been trained on large datasets and optimized for specific tasks. Thus, MaaS abstracts away the complexities of model training and deployment and makes it easier for developers to leverage AI technologies. Besides, MaaS enables developers to focus on their specific use cases and application logic, without the need to delve into the intricacies of model training, hyperparameter tuning, and infrastructure management. By using MaaS, developers can save time and resources by utilizing existing models rather than building and training models from scratch. Additionally, in order to create sophisticated models and AI capabilities, it democratizes AI for a wider user base, including developers with no prior AI knowledge.

During the Web 2.0 age, SaaS, PaaS, and IaaS are all cloud computing-based solutions [35]. They offer services over the Internet to reduce the need for on-premises deployment. In terms of cost consideration, this kind of subscription-based flexible payment model benefits both customers and providers significantly [36]: 1) Customers do not consider how to invest in and maintain fundamental infrastructure, such as software and hardware, anymore. 2) Pieces of equipment procured by providers can be efficiently reused across customers, which results in substantial savings in acquisition expenses. 3) The high scalability of these models empowers customers to adapt resource usage and payment obligations according to their specific requirements. However, compared to traditional cloud service models, MaaS has some characteristics and advantages.

- MaaS takes further consideration in the service model by focusing on the training, deployment, and invocation of AI and ML models. In contrast, IaaS provides basic computing resources; PaaS offers application development and deployment platforms; and SaaS provides complete application software. In particular, MaaS abstracts a higher level of service that enables users to directly utilize the capabilities of AI and ML models.
- MaaS aims to offer specialized services and customized functions for users to complete different tasks (such as conversation, video production, and artistic image creation). It is supposed to be a more professional tool for users in the field of development or application. Hence, the core components of MaaS are the training, deployment, and invocation of various AI models. Traditional cloud service models, however, focus more on providing infrastructure, platforms, or application software and offer lower levels of model selection and customization.
- MaaS provides a lot of simplified tools that make it easier for users to utilize and deploy AI models. In contrast, traditional cloud service models often require users to have more technical knowledge and experience to fully utilize the related services, and they also limit the widespread adoption of AI technologies. These can be a barrier for non-experts or most developers. MaaS lets AI usage be more democratic by providing user-friendly operation interfaces and low entry barriers.

In conclusion, MaaS refers to the delivery of machine learning models as a cloud-based service. As a service model specifically targeting machine learning models, MaaS differs from traditional cloud service models (i.e., IaaS, PaaS, and SaaS) in terms of abstraction level, service objects, model selection and customization, and usability. The advantages of MaaS are that it offers more advanced services and abstractions tailored to machine learning models. It is convenient and efficient for users to use and deploy machine learning models.

## IV. RELATED TECHNOLOGIES OF MAAS

During the whole of Web 2.0, most application development technology stacks can be summarized as a three-tier architecture (Fig. 2(a)). Both desktop and mobile applications depend on their execution environment (i.e., operating systems). As a popular social network and communication platform, there are some obvious differences between the WeChat app [37] on Android and iOS systems. For example, on the Android system, the mini-programs can be set as a floating window. Users can quickly open them by sliding to the right on WeChat. Furthermore, users can add mini-programs to the home screen to avoid repeated operations. However, the iOS system does not support the floating window function. If users accidentally close a mini program, they have to open WeChat again and then use the closed mini program. The reason for the above case is that the differences between the two operating systems themselves lead to distinct development environments and design principles. Essentially, this is due to the significant differences in chips used by Android and iOS systems.

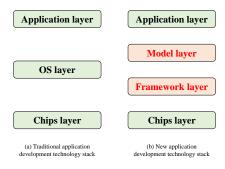


Fig. 2: Comparison between traditional and model-based technology stacks.

However, these applications will be unified within the MaaS platform. The new application development technology stack is a four-tier architecture (Fig. 2(b)). Compared with the traditional technology stack, the training framework and model layers replace the original operating system layer. Almost all large models are conducted by similar training frameworks (e.g., DeepSpeed [38], MindSpore [39], Megatron [40] Fairscale [41], PaddlePaddle [42]), which reduces the effect of operating systems. Besides, the AI-based application is a new kind of product that was developed based on the understanding, generation, and memory capabilities of large models. Indeed, MaaS has gradually become an integral part of the modern technological landscape. In this section, we will

explore the related technologies of MaaS. Several technologies and components are involved in the implementation of MaaS. Here are some important elements we suppose:

Cloud computing: MaaS relies on cloud computing infrastructure [43], [44] to host and deploy machine learning models. Cloud services grant access to scalable and reliable platforms that effectively meet the computational demands of serving models to numerous users. The reliance on cloud computing infrastructure marks a significant shift in how machine learning is utilized. Traditionally, setting up and maintaining the necessary computational resources for machine learning tasks was a cumbersome and expensive process.

Model training and optimization: MaaS providers train machine learning models on large datasets to achieve high accuracy and performance. This involves using techniques like deep learning [38], [45], transfer learning [46], [47], and ensemble learning [48], [49] to develop robust models. Besides, trained models are optimized for deployment on MaaS platforms. This includes techniques like model compression, quantization, and pruning [50], which reduce the model size and improve its efficiency without sacrificing performance.

API and development tools: API and development tools are key technologies that make sure MaaS services are easy to use. MaaS platforms provide a set of APIs for developers to dumb down the use and integration of models. These APIs provide endpoints for making requests and receiving predictions or insights from the models. Correspondingly, development tools such as software development kits (SDKs), command-line interfaces (CLIs), RESTful APIs, and GraphQL are commonly used in MaaS implementations [51], [52], [53].

Monitoring and analytics: MaaS providers implement monitoring and analytics capabilities to track model performance, usage records, and user feedback. This helps in optimizing models, identifying potential issues, and continuously improving the service quality. Analytic tools allow MaaS providers to recognize abnormal decisions, user preferences, and context while running models. The collected information is invaluable for adjusting models to meet the specific needs of users. For instance, if analytic results reveal that users predominantly employ a chatbot for customer support inquiries during peak hours, providers can allocate additional resources to ensure seamless service during these times.

Scalability and load balancing: The essence of scalability in MaaS platforms is all about flexibility and adaptability. In general, the platforms have to handle numerous concurrent requests from users [54], [55]. When the demand for services surges, a scalable system can timely manage its resources to meet the increased load. This means the system can swiftly allocate more computing power, storage, and bandwidth, ensuring that user experiences remain smooth and uninterrupted. Scalability and load-balancing technologies ensure that the system can handle the increasing demand and distribute the workload efficiently across multiple servers or instances.

**Security and privacy**: MaaS platforms are confronted with the imperative task of mitigating security and privacy risks inherent to the management of sensitive data and model accessibility. To enhance the integrity of these platforms, a comprehensive arsenal of security measures is deployed, including encompassing data encryption, access control mechanisms, and the utilization of secure communication protocols [56], [57]. Therefore, these strategies collectively serve the dual purpose of safeguarding user data and preserving the sanctity of confidential information.

# V. ADVANTAGES OF MAAS

Because of its benefits, MaaS is very useful for developers and businesses that are working on making and using GenAI models. MaaS speeds up the entire model development and deployment lifecycle, effectively reducing technical barriers and delivering a series of benefits. These benefits encompass robust performance, flexible payment options, and continuous optimization. Some significant advantages of MaaS are listed as follows:

Lower technical barriers: MaaS effectively lowers the technical barriers associated with the utilization of GenAI models. Developers do not need to become experts in machine learning technologies or master complex algorithms to effectively use advanced models for their development and application needs. They can pay more attention to the creative and practical works. In addition, this kind of user-friendly approach also will encourage innovation across different industries.

Simplified model development: MaaS offers pre-trained AI models as accessible services. Although the pre-train paradigm lets developers leverage AI capabilities without the need for extensive model construction and training, traditional AI model development is still time-consuming and resource-intensive work. Without constructing and training model steps, MaaS empowers developers to seamlessly integrate open-source models into their workflows. This not only optimizes time and resource allocation but also significantly diminishes the initial learning curve and implementation barriers.

High performance and scalability: Since cloud computing services are the basic infrastructure of MaaS, the powerful capability of high-performance computing resources meets numerous data and complex computational requisites. Scalability is an important characteristic of MaaS, which ensures MaaS automatically adjusts resource allocation in response to changing computational requirements. In particular, adaptability is crucial in scenarios where workloads vary dramatically, such as in video production applications. For instance, a model used for real-time image recognition may experience significant spikes in demand during peak usage hours. The cloud infrastructure supports MaaS to seamlessly allocate additional computing resources to handle these spikes and thus ensures fast and consistent response times.

Shared knowledge and collaboration: Since MaaS is built upon a wealth of data and expert experience, the models usually represent broad knowledge and generalization abilities. Developers benefit from this feature without considering gathering or processing large amounts of data independently. The collective knowledge and experience underpinning MaaS models serve as a valuable resource that transcends the confines

of individual capabilities. Developers can leverage this shared knowledge to finish their projects as soon as possible. Moreover, MaaS serves as a conducive platform for sharing models and exchanging experiences, which will promote collaboration and knowledge sharing among developers. This accelerates innovation and problem-solving and drives the development of the entire machine-learning community.

Intelligence decision support: MaaS can offer businesses intelligent decision support by leveraging machine learning models to predict future business trends and financial conditions, aiding enterprises in making wiser decisions. MaaS translates analytical results into user-friendly reports and visualizations and provides customized business solutions. Enterprises can select different machine learning models and data processing algorithms based on their specific business requirements and data characteristics, thus achieving more intelligent and personalized business processes [58].

Flexible payment models: In fact, MaaS conventionally utilizes flexible payment models. The subscription-based model allows users to pay only for the actual usage of MaaS services. This kind of payment model is a cost-effective solution and beneficial for small and medium-sized enterprises as well as individuals. Besides, it can convert customers into subscribers, which reduces the cost of attracting new customers and increases customers' loyalty. This is particularly valuable in competitive markets where attracting new customers is expensive and hard [59]. Additionally, the subscription model not only ensures a steady income stream for enterprises but also enables providers to deliver consistent updates, improvements, and support services to subscribers.

Wide scope of applications: The GenAI model exhibits compatibility across diverse domains (e.g., natural language processing, image creation, recommendation systems, and video production). Large language models, with their extensive training on diverse data sources, show the ability to perform well in multimodal downstream tasks [60], [61]. The versatility of GenAI models outperforms the constraints of specific industries. Developers no longer need to find fine-tuned models for different cases. Instead, they can draw from a unified model service that adapts to the unique demands of various application scenarios.

## VI. APPLICATIONS WITHIN MAAS

MaaS has a wide range of applications in various domains, including but not limited to the following. Firstly, we present an overview of applications within MaaS in Fig. 3.

# A. MaaS in Healthcare

MaaS will have multifaceted implications for the healthcare sector in the future. Its capacity to amalgamate a wealth of diverse healthcare data sources, encompassing electronic health records (EHRs), medical literature, genomics data, and real-time patient monitoring data, is the cornerstone of its transformative potential [62]. MaaS's analytical prowess transcends data integration, enabling it to unveil intricate patterns, correlations, and insights that underpin treatment

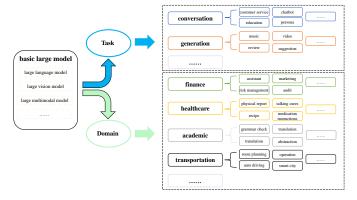


Fig. 3: The applications of various industries within MaaS.

recommendations and prognosis assessments. The integration of clinical guidelines and principles of evidence-based medicine further distinguishes MaaS as a pivotal healthcare ally. Its analytical engines, fueled by extensive medical literature and clinical trial data, distill the most up-to-date and efficacious treatment modalities for specific medical conditions and patient profiles, thereby furnishing healthcare professionals with an invaluable resource for decision-making. MaaS's capabilities extend to patient profiling and risk assessment, where it leverages machine learning algorithms to scrutinize a series of patient attributes. By juxtaposing this data against expansive patient cohorts and research findings, MaaS crafts personalized risk assessments, forecasting disease progression, treatment responses, and potential adverse events with remarkable precision. As a decision support system, MaaS offers physicians real-time, evidence-based guidance. Through the amalgamation of patient-specific data and clinical knowledge, MaaS optimizes treatment plans, suggesting tailored medication options, dosage adjustments, or alternative therapies based on individual patient characteristics. In addition, predictive modeling is another advantage of MaaS. A potent tool for forecasting treatment outcomes and prognostic assessments. By delving into historical patient data and treatment responses, MaaS identifies predictive factors and forges models that estimate the likelihood of treatment success, disease recurrence, or patient survival. Continuously absorbing fresh patient data, treatment outcomes, and clinical research findings, MaaS updates its algorithms and models to adapt the novel emerging evidence. Finally, the result is a system that continually refines its personalized treatment recommendation tasks and prognostic assessment metrics. This ensures the healthcare system remains adaptable for serving.

# B. MaaS in Academic

MaaS has emerged as a transformative force to profoundly influence research fields and community development through a series of profound effects. The most influential is the acceleration of academic research. MaaS empowers researchers with access to pre-trained models, datasets, and infrastructure, thus streamlining the research and development process. This acceleration is pivotal in allowing researchers to channel their

energies directly into their research objectives, without the need for extensive investments of time and resources in model creation and training [63]. As a consequence, research cycles become shorter, fostering rapid advancements and breakthroughs in a myriad of research domains. MaaS also carries the torch of democratization in research. By offering readily available models, tools, and resources, it removes barriers for researchers and offers state-of-the-art models, techniques, and tools that require limited resources or expertise in machine learning. This inclusivity opens the doors for a wide range of researchers to benefit from the progress in their research fields. Facilitating collaboration and knowledge sharing represents another pivotal role of MaaS platforms. These platforms serve as a nexus for researchers to exchange models, datasets, and findings, fostering transparency and nurturing a culture of collaboration. This mutual exchange of knowledge and expertise catalyzes the cross-pollination of ideas, expedites research, and promotes interdisciplinary approaches to problem-solving. MaaS platforms also play a substantial role in reinforcing the reproducibility of research results. By providing access to pretrained models and standardized datasets, we can rigorously validate and benchmark the results against established standards. This transparency elevates the credibility of research findings and permits equitable comparisons between various approaches, thereby encouraging healthy competition and propelling innovation. Perhaps one of the most consequential effects of MaaS is the swift deployment of research outcomes. MaaS empowers us to seamlessly transition their models and solutions into real-world applications. This capacity serves as a conduit for translating research findings into practical use cases, conferring benefits upon industries, organizations, and society as a whole. We can effectively employ MaaS platforms to bridge the gap between academia and industry.

# C. MaaS in Blockchain

The intersection of MaaS and blockchain technology brings forth a series of transformative effects, profoundly influencing data privacy, model sharing, training, and decentralized model inference within the blockchain ecosystem. The fusion of MaaS and blockchain augments data privacy and security. The inherent attributes of blockchain, including decentralization, immutability, and traceability, play a pivotal role in safeguarding data integrity and privacy. By executing model training and inference on distributed nodes, sensitive data is shielded from the risk of transmission to centralized servers [64]. This integration not only protects users' data privacy but also improves the overarching system's security. Furthermore, MaaS in a blockchain context can actualize model sharing and marketization [65]. The blockchain ledger records the ownership and usage rights of models, thereby establishing a framework for model providers to vend or license their models to fellow users. This ingenious marketization mechanism empowers model developers to enhance their creations while simultaneously affording users a wider array of choices and greater flexibility in model selection. In terms of model training, blockchain's verifiability attribute finds its application

by validating the credibility of model training processes. This is accomplished by documenting the entire training process and its outcomes on the blockchain, guaranteeing transparency, auditability, and resistance to tampering. This feature is of paramount significance in applications where the trustworthiness and traceability of the model training process are crucial, such as in the domains of finance and healthcare. The decentralized nature of blockchain is further leveraged for decentralized model inference. Models are deployed across multiple nodes within the blockchain, enabling parallel inference and consensus verification of results. This decentralized model inference approach augments both the efficiency and reliability of model inference. It is particularly pertinent in applications requiring high reliability and resilience against tampering, such as in data encryption, smart contract execution [66], and identity validation and authentication [67].

# D. MaaS in Web 3.0

Web 3.0 will be the next evolution of the Internet, characterized by decentralization, user control, intelligence, and data privacy [4]. Intelligence is one of the significant features of Web 3.0, as well as of MaaS [68]. Consequently, MaaS serves as an intelligent solution that smoothly aligns with the ideals of Web 3.0. Besides, Web 3.0 also emphasizes personal data autonomization within user control. Since the individuals' data is a cash cow, in the context of MaaS, users transmit their data to the cloud computing platform for processing while retaining full control over the utilization and dissemination of their data. The flexibility to selectively share data in exchange for model predictions safeguards personal privacy, offering users the autonomy to manage and leverage their data in harmony with the principles of Web 3.0. Data transmission and storage are inherent to the process of sending data to the cloud for model inference. To uphold data privacy and security, MaaS incorporates measures such as encryption, authentication, and access control (as illustrated in the last subsection). These safeguards protect user data from unauthorized access and misuse, in alignment with the data privacy and security objectives of Web 3.0. Web 3.0 also contains smart contract technology, which ensures automated business logic in decentralized applications. By uniting smart contracts with MaaS, a higher level of automated decision-making and intelligent business logic can be achieved. Web 3.0 encourages data interoperability and cross-platform interaction. MaaS complements this objective by offering standardized interfaces and protocols, enabling diverse platforms and applications to effortlessly invoke machine learning models.

## VII. CHALLENGES AND ISSUES

Model-as-a-Service (MaaS), as a new deployment and service paradigm for different AI-based models, will face several challenges and potential issues for future development. In this section, we highlight the key challenges and issues associated with MaaS. Details are described below.

First and foremost, one significant challenge that MaaS encounters is ensuring security and privacy. In the architecture

of MaaS, models and rich data often need to be transmitted and processed across different environments. This can potentially lead to the exposure of sensitive data and the misappropriation of confidential models. Therefore, ensuring the security and privacy of models and data becomes one of the key issues that MaaS needs to address. Future directions could include the application of encryption and secure computing technologies, along with the establishment of more rigorous access control and permission management mechanisms.

Secondly, MaaS also grapples with challenges related to model management and version control. With models undergoing continuous iterations and updates, effective management and control of model versions have become a critical concern. Additionally, the deployment and service aspects of models must consider scenarios where multiple models coexist. Issues such as model selection, composition, management, and scheduling of different model versions need to be addressed. Future directions might involve the creation of comprehensive model repositories and version control systems, as well as automated model selection and composition algorithms.

Furthermore, MaaS needs to address challenges related to computational power and resource allocation. As machine learning models grow in complexity and scale, the demand for computational resources also increases. Within the MaaS architecture, achieving elastic scalability and load balancing to meet the varying scales and demands of model services becomes a significant challenge. Future directions could include leveraging containerization and cluster management technologies to achieve flexible deployment and scheduling of models [69], along with optimizing the computational and inference processes of models to enhance resource utilization.

In practice, the future various development of MaaS also focuses on multi-platforms and cross-organizational integration. Deploying and utilizing models often involve collaboration among multiple platforms and organizations. Challenges include achieving model interaction and communication between different platforms and ensuring data sharing and cooperation among various organizations. Future directions might encompass the formulation of standardized model interfaces and communication protocols to facilitate interoperability and collaboration between different platforms.

There are several respects in which MaaS may have trouble:

- Latency: Latency is the time between a request and a response. Depending on the complexity of each GenAI model and the volume of requests, a latency issue is inevitable. Real-time applications may require low-latency responses, which may be difficult to achieve, particularly for foundation models. In fact, we need low-latency, real-time inference services for applications.
- Interpretability of model and results: A "black box"
  model has poor interpretability, and the users cannot
  accurately fill in the decision-making results. Almost all
  GenAI models exhibit a high degree of complexity and
  perform poorly in their interpretability. This often poses
  a significant constraint when it comes to elucidating the
  rationale behind the models' decisions to users.

- Governance, risk, and compliance (GRC): The main compliance refers to data compliance, content compliance, platform operation compliance, platform management compliance, etc. Organizations often have to comply with various rules and regulations, including but not limited to General data protection regulation (GDPR) [70] and industry-specific standards. These guideline steps will meet some significant challenges when we deploy MaaS.
- Data quality and bias: High-quality big data is the cornerstone of foundation models. If the training data used to build models is biased or of poor quality, it can lead to biased predictions and inaccurate results when deployed as a service. Pre-trained large models may inherit biases present in the training data, leading to biased or unfair outcomes [71], [72]. Careful evaluation and mitigation of biases are necessary to ensure the ethical and fair application of the models.
- Lack of transparency: Ensuring that models can work seamlessly with a variety of platforms and technologies used by different clients can be challenging. Compatibility issues may arise. Pre-trained models used in MaaS may be complex and difficult to interpret. Understanding the inner workings and making decisions based on the model's output can be challenging [73], particularly for critical applications requiring explainable AI.
- Dependency on service availability: Using MaaS from a particular provider can lead to vendor lock-in, making it difficult to switch to a different service or bring the model in-house. MaaS relies on the availability and reliability of the service provider's infrastructure. If the provider experiences downtime or interruptions, it can disrupt the functioning of applications relying on the models.
- Limited offline capabilities: Ensuring the availability
  and reliability of the MaaS can be challenging. Downtime or service interruptions can disrupt applications that
  depend on the model. MaaS often requires an active
  Internet connection to access and utilize the models.
  In scenarios where Internet connectivity is limited or
  unreliable, offline functionality may be compromised.
- Limited customization: This may seem counterintuitive, but it is indeed the case. Sometimes, the vast majority of ordinary users lack the capability to modify these parameters. Some MaaS providers might limit the level of customization available, which can be a limitation for applications with specific requirements. Pre-trained models provided by MaaS may lack the flexibility to be extensively customized according to specific requirements. Users may have limited control over the underlying model architecture and parameters.

With MaaS, how to achieve "one model to serve all"? This is still an open question. Addressing these challenges and limitations often requires careful planning, architecture design, and ongoing maintenance. Organizations should evaluate their specific needs and constraints before deciding to implement MaaS to ensure it aligns with their goals and resources.

## VIII. FUTURE DIRECTIONS

MaaS holds great potential in the fields of machine learning and artificial intelligence. There are several noteworthy future directions that can take further research. These include, but are not limited to, addressing security and privacy concerns, exploring encryption and secure computing techniques, and developing access control mechanisms. Enhancing model management, version control systems, and optimizing resource allocation for scalability are also the keys to MaaS research. Additionally, cross-organizational integration through standardized interfaces and communication protocols is promising for future work. These directions are pivotal for unlocking the full potential of MaaS. We discuss some of them as follows:

- Data privacy and security: With the increasing adoption of MaaS, the growing demand for models and data will raise people's concerns about security and privacy [70], [74]. In the future, there will be a necessity for more advanced encryption schemes, access control mechanisms, and data protection methods to ensure the security of data transmission and processing in MaaS.
- Algorithm optimization and model efficiency: Future research can focus on improving algorithms and optimizing model efficiency to enhance the performance and responsiveness of MaaS. The efficient utilization of limited resources is an urgent concern within MaaS. We should explore ways to optimize the computation and inference processes of large models to reduce resource consumption [75]. This optimization work is vital for maximizing the potential of MaaS in the future.
- Model lifecycle management: Model lifecycle management, which covers development, training, deployment, updates, and retirement, is a critical area for research and development. Scholars and engineers can focus on establishing more sophisticated model management theories and developing corresponding version control systems to ensure effective model governance and maintenance.
- Explainable AI and ethical considerations: How to improve the explainability of MaaS models and address ethical and moral issues related to AI to ensure fairness and transparency in their applications? In MaaS research, a crucial challenge is the complexity of model decision processes. We must devise novel paradigms to enhance the interpretability of these models and then let users understand the rationale behind the models' decisions.
- Environmental friendliness and energy efficiency: Future research can explore ways to improve the environmental friendliness and energy efficiency of MaaS by optimizing the use of computational and communication resources, reducing its impact on the environment.
- Multimodal learning and cross-device collaboration:
  We can explore multimodal learning [76] (involving multiple modalities) and cross-device collaboration (e.g., federated learning) to enable the sharing and training of models in distributed environments, improving model performance while preserving data privacy.

Social impact and policy issues: We must carefully
consider the societal implications of MaaS and actively
contribute to the establishment of regulations and policies
that will guide its utilization and advancement. It is imperative that these regulations align with ethical principles
and the broader interests of society.

## IX. CONCLUSION

This is the first survey that concentrates on the topic of MaaS based on GenAI models. We first discuss the history between XaaS models, which we suppose is a continuous conversion. Then, we review the relationship between cloud platforms (i.e., SaaS, PaaS, and IaaS) and draw out differences from the perspective of Web eras. Several important technologies of the MaaS platform are introduced in detail, as well as their benefits. Subsequently, we present some application scenarios, since MaaS has a wide range of applications in various domains. Finally, we identify several key problems and challenges that MaaS will face and point out some future directions for studying MaaS. MaaS has emerged as a transformative paradigm in the era of AI, enabling organizations to leverage pre-trained models seamlessly. Its benefits, including accelerated adoption, low cost, and easy operation advantages, have paved the way for widespread AI integration in industries. As MaaS grows in the future, data privacy, customization, and ethical considerations will be key troubles and challenges to taking full advantage of its potential advantages. With the evolution of MaaS for social good, ubiquitous AI technologies will be deeply engaged in our daily lives.

# ACKNOWLEDGMENT

This research was supported in part by the National Natural Science Foundation of China (Nos. 62272196 and 62002136), the Natural Science Foundation of Guangdong Province (No. 2022A1515011861), and the Young Scholar Program of Pazhou Lab (No. PZL2021KF0023).

## REFERENCES

- [1] W. Gan, J. C.-W. Lin, H.-C. Chao, and J. Zhan, "Data mining in distributed environment: a survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 6, p. e1216, 2017.
- [2] J. Sun, W. Gan, Z. Chen, J. Li, and P. S. Yu, "Big data meets metaverse: A survey," arXiv preprint arXiv:2210.16282, 2022.
- [3] M. Minsky, "Steps toward artificial intelligence," *The Institute of Radio Engineers*, vol. 49, no. 1, pp. 8–30, 1961.
- [4] W. Gan, Z. Ye, S. Wan, and P. S. Yu, "Web 3.0: The future of internet," in *Companion Proceedings of the Web Conference*. ACM, 2023, pp. 1266–1275.
- [5] S. Wan, H. Lin, W. Gan, J. Chen, and P. S. Yu, "Web3: The next internet revolution," arXiv preprint arXiv:2304.06111, 2023.
- [6] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
- [7] L. Hu, Z. Liu, Z. Zhao, L. Hou, L. Nie, and J. Li, "A survey of knowledge enhanced pre-trained language models," *IEEE Transactions* on Knowledge and Data Engineering, pp. 1–19, 2023.
- [8] R. Gozalo-Brizuela and E. C. Garrido-Merchán, "A survey of generative AI applications," arXiv preprint arXiv:2306.02781, 2023.
- [9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

- [10] Z. Epstein, A. Hertzmann, I. of Human Creativity, M. Akten, H. Farid, J. Fjeld, M. R. Frank, M. Groh, L. Herman, N. Leach *et al.*, "Art and the science of generative AI," *Science*, vol. 380, no. 6650, pp. 1110–1111, 2023
- [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, pp. 1–11, 2017.
- [12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in NAACL-HLT, 2019, pp. 4171–4186.
- [13] A. Radford, J. Wu, D. Amodei, D. Amodei, J. Clark, M. Brundage, and I. Sutskever, "Better language models and their implications," *OpenAI blog*, vol. 1, no. 2, 2019.
- [14] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson et al., "Extracting training data from large language models," in *The Conference on 30th USENIX Security Symposium*, 2021, pp. 2633–2650.
- [15] F. Zeng, W. Gan, Y. Wang, and P. S. Yu, "Distributed training of large language models," in *IEEE 29th International Conference on Parallel* and Distributed Systems. IEEE, 2023, pp. 1–8.
- [16] W. Gan, J. C. W. Lin, P. Fournier-Viger, H. C. Chao, V. S. Tseng, and P. S. Yu, "A survey of utility-oriented pattern mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1306–1327, 2021.
- [17] J. Sun, W. Gan, H.-C. Chao, P. S. Yu, and W. Ding, "Internet of behaviors: A survey," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11117–11134, 2023.
- [18] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.
- [19] J. Yu, X. Zhang, Y. Xu, X. Lei, X. Guan, J. Zhang, L. Hou, J. Li, and J. Tang, "XDAI: A tuning-free framework for exploiting pre-trained language models in knowledge grounded dialogue generation," in *The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 4422–4432.
- [20] X. Garcia, Y. Bansal, C. Cherry, G. Foster, M. Krikun, M. Johnson, and O. Firat, "The unreasonable effectiveness of few-shot learning for machine translation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 10867–10878.
- [21] L. Bonifacio, H. Abonizio, M. Fadaee, and R. Nogueira, "Inpars: Unsupervised dataset generation for information retrieval," in *The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 2387–2392.
- [22] L. Yu, J. Chen, A. Sinha, M. Wang, Y. Chen, T. L. Berg, and N. Zhang, "CommerceMM: Large-scale commerce multimodal representation learning with omni retrieval," in *The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 4433– 4442.
- [23] H. Huang, O. Zheng, D. Wang, J. Yin, Z. Wang, S. Ding, H. Yin, C. Xu, R. Yang, Q. Zheng et al., "ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model," *International Journal of Oral Science*, vol. 15, no. 1, p. 29, 2023.
- [24] Y. Shen, L. Heacock, J. Elias, K. D. Hentel, B. Reig, G. Shih, and L. Moy, "ChatGPT and other large language models are double-edged swords," p. e230163, 2023.
- [25] C. M. Mohammed and S. R. Zeebaree, "Sufficient comparison among cloud computing services: IaaS, PaaS, and SaaS: A review," *Interna*tional Journal of Science and Business, vol. 5, no. 2, pp. 17–30, 2021.
- [26] M. Cusumano, "Cloud computing and SaaS as new computing platforms," Communications of the ACM, vol. 53, no. 4, pp. 27–29, 2010.
- [27] E. Van Eyk, L. Toader, S. Talluri, L. Versluis, A. Uţă, and A. Iosup, "Serverless is more: From PaaS to present cloud computing," *IEEE Internet Computing*, vol. 22, no. 5, pp. 8–17, 2018.
- [28] N. Chohan, C. Bunch, S. Pang, C. Krintz, N. Mostafa, S. Soman, and R. Wolski, "AppScale: Scalable and open appengine application development and deployment," in *The 1st International Conference on Cloud Computing*. Springer, 2010, pp. 57–70.
- [29] D. M. Laxmaiah, D. Y. K. Sharma et al., "A comparative study on google app engine amazon web services and microsoft windows azure," *International Journal of Computer Engineering & Technology*, vol. 10, no. 1, pp. 54–60, 2019.
- [30] D. Chappell, "Introducing the azure services platform," White paper, vol. 1364, no. 11, pp. 1–32, 2008.

- [31] S. S. Manvi and G. K. Shyam, "Resource management for infrastructure as a service (IaaS) in cloud computing: A survey," *Journal of Network* and Computer Applications, vol. 41, pp. 424–440, 2014.
- [32] S. Krishnan, J. L. U. Gonzalez, S. Krishnan, and J. L. U. Gonzalez, "Google compute engine," 2015.
- [33] V. Pikkuhookana and J. Soini, "Software development using cloud services," 2023.
- [34] A. E. C. Cloud, "Amazon elastic compute cloud," 2009
- [35] J. Peng, X. Zhang, Z. Lei, B. Zhang, W. Zhang, and Q. Li, "Comparison of several cloud computing platforms," in *International Symposium on Information Science and Engineering*. IEEE, 2009, pp. 23–27.
- [36] C. L. Wang, Y. Zhang, L. R. Ye, and D.-D. Nguyen, "Subscription to fee-based online services: What makes consumer pay for online content?" *Journal of Electronic Commerce Research*, vol. 6, no. 4, pp. 304–311, 2005.
- [37] C. Montag, B. Becker, and C. Gan, "The multipurpose application WeChat: a review on recent research," *Frontiers in Psychology*, vol. 9, p. 2247, 2018.
- [38] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, "DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters," in *The 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020, pp. 3505–3506.
- [39] L. Huawei Technologies Co., "Huawei MindSpore AI development framework," in *Artificial Intelligence Technology*. Springer, 2022, pp. 137–162.
- [40] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training multi-billion parameter language models using model parallelism," arXiv preprint arXiv:1909.08053, 2019.
- [41] M. Baines, S. Bhosale, V. Caggiano, N. Goyal, S. Goyal, M. Ott, B. Lefaudeux, V. Liptchinsky, M. Rabbat, S. Sheiffer *et al.*, "Fairscale: A general purpose modular pytorch library for high performance and large scale training," 2021.
- [42] S. Hu, C. He, C. Zhang, Z. Tan, B. Ge, and X. Zhou, "Efficient scene text recognition model built with PaddlePaddle framework," in *The 7th International Conference on Big Data and Information Analytics*. IEEE, 2021, pp. 139–142.
- [43] D. Hilley, "Cloud computing: A taxonomy of platform and infrastructure-level offerings," Georgia Institute of Technology, pp. 44– 45, 2009.
- [44] Y. Khmelevsky and V. Voytenko, "Cloud computing infrastructure prototype for university education and research," in *The 15th Western Canadian Conference on Computing Education*, 2010, pp. 1–5.
- [45] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, "ZeRO: Memory optimizations toward training trillion parameter models," in SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2020, pp. 1–16.
- [46] T. T. Joy, S. Rana, S. Gupta, and S. Venkatesh, "A flexible transfer learning framework for bayesian optimization with convergence guarantee," *Expert Systems with Applications*, vol. 115, pp. 656–672, 2019.
- [47] M. Jiang, Z. Wang, S. Guo, X. Gao, and K. C. Tan, "Individual-based transfer learning for dynamic multiobjective optimization," *IEEE Transactions on Cybernetics*, vol. 51, no. 10, pp. 4968–4981, 2020.
- [48] V. J. Kadam and S. M. Jadhav, "Performance analysis of hyperparameter optimization methods for ensemble learning with small and medium sized medical datasets," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no. 1, pp. 115–123, 2020.
- [49] F. Wang, Y. Li, F. Liao, and H. Yan, "An ensemble learning based prediction strategy for dynamic multi-objective optimization," *Applied Soft Computing*, vol. 96, p. 106592, 2020.
- [50] J. Kim, S. Chang, and N. Kwak, "PQK: model compression via pruning, quantization, and knowledge distillation," arXiv preprint arXiv:2106.14681, 2021.
- [51] T. D. Nguyen, A. T. Nguyen, H. D. Phan, and T. N. Nguyen, "Exploring API embedding for API usages and applications," in *IEEE/ACM 39th International Conference on Software Engineering*. IEEE, 2017, pp. 438–449.
- [52] S. I. Ross, F. Martinez, S. Houde, M. Muller, and J. D. Weisz, "The programmer's assistant: Conversational interaction with a large language model for software development," in *The 28th International Conference* on Intelligent User Interfaces, 2023, pp. 491–514.
- [53] N. Li, B. Kang, and T. De Bie, "SkillGPT: a RESTful API service for skill extraction and standardization using a large language model," arXiv preprint arXiv:2304.11060, 2023.

- [54] Y. Xu, L. Wu, L. Guo, Z. Chen, L. Yang, and Z. Shi, "An intelligent load balancing algorithm towards efficient cloud computing," in Workshops at the 25th AAAI Conference on Artificial Intelligence, 2011, pp. 27–32.
- [55] M. Mesbahi and A. M. Rahmani, "Load balancing in cloud computing: a state of the art survey," *International Journal of Modern Education and Computer Science*, vol. 8, no. 3, pp. 64–78, 2016.
- [56] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in The 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015, pp. 1310–1321.
- [57] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "Sok: Security and privacy in machine learning," in *IEEE European Symposium on Security and Privacy*. IEEE, 2018, pp. 399–414.
- [58] A. Papa, L. Dezi, G. L. Gregori, J. Mueller, and N. Miglietta, "Improving innovation performance through knowledge acquisition: the moderating role of employee retention and human resource management practices," *Journal of Knowledge Management*, vol. 24, no. 3, pp. 589–605, 2020.
- [59] M. Labus and M. Stone, "The CRM behaviour theory-managing corporate customer relationships in service industries," *Journal of Database Marketing & Customer Strategy Management*, vol. 17, pp. 155–173, 2010.
- [60] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu et al., "PaLM-E: An embodied multimodal language model," arXiv preprint arXiv:2303.03378, 2023.
- [61] Y. Zang, W. Li, J. Han, K. Zhou, and C. C. Loy, "Contextual object detection with multimodal large language models," arXiv preprint arXiv:2305.18279, 2023.
- [62] R. Kosklin, J. Lammintakanen, and T. Kivinen, "Knowledge management effects and performance in health care: a systematic literature review," *Knowledge Management Research & Practice*, vol. 21, no. 4, pp. 738–748, 2023.
- [63] D. Feser, "Innovation intermediaries revised: a systematic literature review on innovation intermediaries' role for knowledge sharing," *Review of Managerial Science*, vol. 17, no. 5, pp. 1827–1862, 2023.
- [64] G. Zyskind, O. Nathan et al., "Decentralizing privacy: Using blockchain to protect personal data," in *IEEE Security and Privacy Workshops*. IEEE, 2015, pp. 180–184.
- [65] J. Chiu and T. V. Koeppl, "Blockchain-based settlement for asset trading," *The Review of Financial Studies*, vol. 32, no. 5, pp. 1716– 1753, 2019.
- [66] N. Atzei, M. Bartoletti, and T. Cimoli, "A survey of attacks on ethereum smart contracts (sok)," in *The 6th International Conference on Principles* of Security and Trust. Springer, 2017, pp. 164–186.
- [67] A. Norta, R. Matulevičius, and B. Leiding, "Safeguarding a formalized blockchain-enabled identity-authentication protocol by applying security risk-oriented patterns," *Computers & Security*, vol. 86, pp. 253–269, 2019.
- [68] F. A. Alabdulwahhab, "Web 3.0: the decentralized web blockchain networks and protocol innovation," in *The 1st International Conference* on Computer Applications & Information Security. IEEE, 2018, pp. 1–4.
- [69] O. Bentaleb, A. S. Belloum, A. Sebaa, and A. El-Maouhab, "Containerization technologies: Taxonomies, applications and challenges," The Journal of Supercomputing, vol. 78, no. 1, pp. 1144–1181, 2022.
- [70] G. D. P. Regulation, "General data protection regulation (GDPR)," 2018.
- [71] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training deep neural networks with biased data," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9012–9020.
- [72] J. Su, T. Y. Zhuo, J. Mansurov, D. Wang, and P. Nakov, "Fake news detectors are biased against texts generated by large language models," arXiv preprint arXiv:2309.08674, 2023.
- [73] M. H. Jarrahi, D. Askay, A. Eshraghi, and P. Smith, "Artificial intelligence and knowledge management: A partnership between human and ai," *Business Horizons*, vol. 66, no. 1, pp. 87–99, 2023.
- [74] Y. Chen, Y. Gui, H. Lin, W. Gan, and Y. Wu, "Federated learning attacks and defenses: A survey," in *IEEE International Conference on Big Data*. IEEE, 2022, pp. 4256–4265.
- [75] C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen, "Large language models as optimizers," arXiv preprint arXiv:2309.03409, 2023.
- [76] J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu, "Multimodal large language models: A survey," in *IEEE International Conference on Big Data*. IEEE, 2023, pp. 1–10.