Viewpoint

Economics and Equity of Large Language Models: Health Care Perspective

Radha Nagarajan[1], PhD; Midori Kondo[2], MHA, PharmD; Franz Salas[3], MBA; Emre Sezgin[4], PhD; Yuan Yao[5], BSE; Vanessa Klotzman[6], MS; Sandip A Godambe[7], MD; Naqi Khan[8], MD, MS; Alfonso Limon[7], PhD; Graham Stephenson[9], MD; Sharief Taraman[10], MD; Nephi Walton[11], MD; Louis Ehwerhemuepha[7], PhD; Jay Pandit[12], MD; Deepti Pandita[9], MD; Michael Weiss[1], MD; Charles Golden[1], MD; Adam Gold[7], MBA; John Henderson[1], MBA; Angela Shippy[13], MD; Leo Anthony Celi[14], MD; William R Hogan[15], MD; Eric K Oermann[16], MD; Terence Sanger[7], MD, PhD; Steven Martel[7,17], MD

[1]Children's Hospital of Orange County, Orange, CA, United States
[2]Fred Hutch Patient Care, Seattle, WA, United States
[3]Amazon Web Services, Detroit, MI, United States
[4]Nationwide Children's Hospital, Columbus, OH, United States
[5]Amazon Web Services, San Francisco, CA, United States
[6]University of California Irvine, Irvine, CA, United States
[7]Children’s Hospital of Orange County, Orange, CA, United States
[8]Amazon Web Services, Seattle, WA, United States
[9]University of California Irvine Health, Irvine, CA, United States
[10]Cognoa LLC, Palo Alto, CA, United States
[11]National Institutes of Health, Bethesda, MD, United States
[12]Scripps Research Translational Institute, La Jolla, CA, United States
[13]Amazon Web Services, Houston, TX, United States
[14]Massachusetts Institute of Technology, Cambridge, MA, United States
[15]Medical College of Wisconsin, Milwaukee, WI, United States
[16]NYU Langone Medical Center, New York, NY, United States
[17]Physicians Specialty Faculty, Orange, CA, United States

Corresponding Author:
Radha Nagarajan, PhD
Children's Hospital of Orange County
1201 W. La Veta Ave
Orange, CA, 92868
United States
Phone: 1 714 997 3000
Email: Radha.Nagarajan@choc.org

Abstract

Large language models (LLMs) continue to exhibit noteworthy capabilities across a spectrum of areas, including emerging proficiencies across the health care continuum. Successful LLM implementation and adoption depend on digital readiness, modern infrastructure, a trained workforce, privacy, and an ethical regulatory landscape. These factors can vary significantly across health care ecosystems, dictating the choice of a particular LLM implementation pathway. This perspective discusses 3 LLM implementation pathways—training from scratch pathway (TSP), fine-tuned pathway (FTP), and out-of-the-box pathway (OBP)—as potential onboarding points for health systems while facilitating equitable adoption. The choice of a particular pathway is governed by needs as well as affordability. Therefore, the risks, benefits, and economics of these pathways across 4 major cloud service providers (Amazon, Microsoft, Google, and Oracle) are presented. While cost comparisons, such as on-demand and spot pricing across the cloud service providers for the 3 pathways, are presented for completeness, the usefulness of managed services and cloud enterprise tools is elucidated. Managed services can complement the traditional workforce and expertise, while enterprise tools, such as federated learning, can overcome sample size challenges when implementing LLMs using health care data. Of the

3 pathways, TSP is expected to be the most resource-intensive regarding infrastructure and workforce while providing maximum customization, enhanced transparency, and performance. Because TSP trains the LLM using enterprise health care data, it is expected to harness the digital signatures of the population served by the health care system with the potential to impact outcomes. The use of pretrained models in FTP is a limitation. It may impact its performance because the training data used in the pretrained model may have hidden bias and may not necessarily be health care–related. However, FTP provides a balance between customization, cost, and performance. While OBP can be rapidly deployed, it provides minimal customization and transparency without guaranteeing long-term availability. OBP may also present challenges in interfacing seamlessly with downstream applications in health care settings with variations in pricing and use over time. Lack of customization in OBP can significantly limit its ability to impact outcomes. Finally, potential applications of LLMs in health care, including conversational artificial intelligence, chatbots, summarization, and machine translation, are highlighted. While the 3 implementation pathways discussed in this perspective have the potential to facilitate equitable adoption and democratization of LLMs, transitions between them may be necessary as the needs of health systems evolve. Understanding the economics and trade-offs of these onboarding pathways can guide their strategic adoption and demonstrate value while impacting health care outcomes favorably.

(J Med Internet Res 2024;26:e64226) doi: 10.2196/64226

KEYWORDS

large language model; LLM; health care; economics; equity; cloud service providers; cloud; health outcome; implementation; democratization

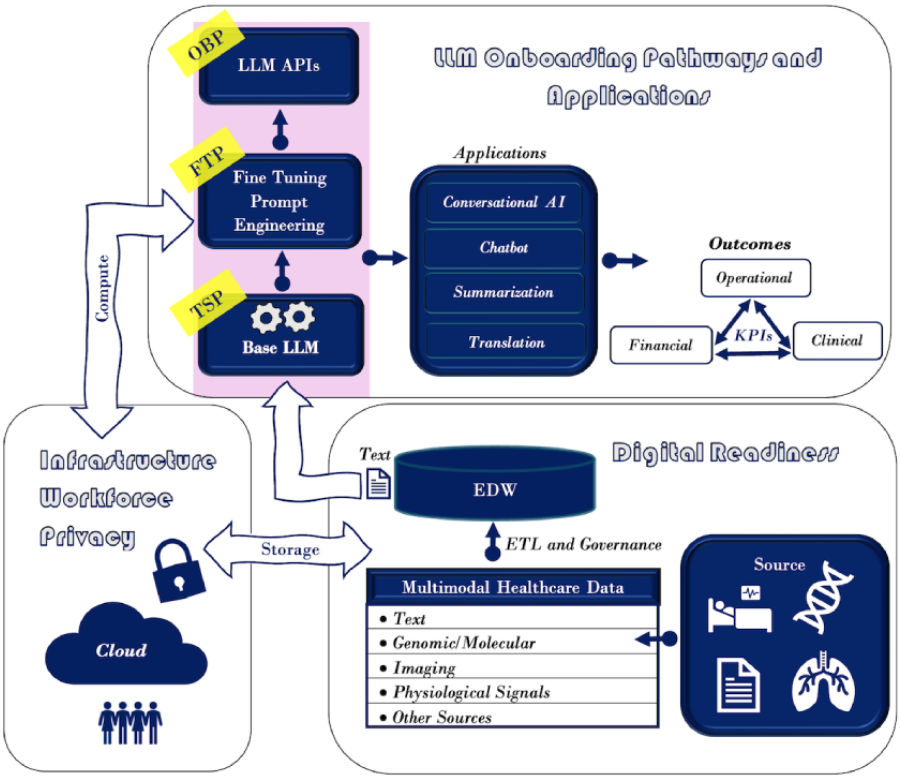
Introduction

Overview

The past decade has witnessed unprecedented growth and digitization of *multivariate* and *multimodal* health care data from diverse sources (eg, the electronic health record [EHR], claims, registries, Internet of Things, and molecular) [1,2]. While multivariate data represent data of a given type across a set of entities (eg, text), multimodal data represent distinct types of data (eg, text, image, and genomics) across entities of interest and play a critical role in generating comprehensive patient and population profiles. Multimodal health care data fall under 2 broad categories, namely *structured* (eg, diagnosis codes and the *International Classification of Disease*, ninth and tenth revisions) and *unstructured* data (eg, text and image). About 80% of health care data are unstructured, including text from clinical narratives [3,4]. The prevalence of unstructured textual clinical data is perhaps a primary motivating factor behind the continued evolution and adoption of natural language processing (NLP) [5] approaches for gaining novel insights from these datasets [6-8]. More recently, advanced machine learning techniques, such as deep learning (DL) [9-11], large language models (LLMs) [12,13], and foundation models, have accelerated these efforts with enhanced capabilities in deciphering patterns from unstructured data [14-16]. Multimodal health care data are usually extracted, transformed, and loaded (extract, transform, and load [ETL]) from diverse source systems into an enterprise data warehouse (EDW; Figure 1). Several variants (eg, Data Lake house) have also been proposed [17]. Subsequently, textual data from EDW are retrieved in a context-specific manner for downstream analytics and ingestion by LLMs (Figure 1). LLM implementations are governed by needs as well as affordability. This perspective discusses factors that impact LLM implementation and proposes 3 broad LLM onboarding pathways for its equitable distribution and adoption.

Multimodal digital footprints (Figure 1) capture unique characteristics of a given population with the potential to assist in decision-making in an evidence-based and data-driven manner, impacting outcomes and key performance indicators (KPIs). These outcomes typically fall under 3 broad categories (*clinical*, *operational*, and *financial*) that are not necessarily independent. For instance, data-driven approaches that can improve preventive care use can minimize aggressive disease-impacting clinical outcomes. Improved clinical outcomes can enable optimal resource allocation impacting operational outcome, reducing the economic burden on the patient, provider, as well as the payer impacting financial outcome. Therefore, the outcomes are represented by bidirectional arrows in Figure 1. While there is considerable excitement over the transformative potential of LLMs in health care [18], it is accompanied by significant *economic* challenges impacting their equitable distribution across health care organizations, especially those that serve economically disadvantaged communities. The choice of a particular LLM implementation pathway is dictated by the *needs* as well as *affordability*. In this perspective, 3 different LLM implementation pathways (training from scratch pathway [TSP], fine-tuned pathway [FTP], and out-of-the-box pathway [OBP]; Figure 1) across 4 major cloud service providers (CSPs; Amazon Web Services [AWS], Google Cloud Platform [GCP], Azure: Microsoft, and Oracle Cloud Infrastructure [OCI]) are presented as “onboarding points.” The risks, benefits, and economics of these pathways are presented and expected to assist in choosing a pathway and subsequent migration across pathways.

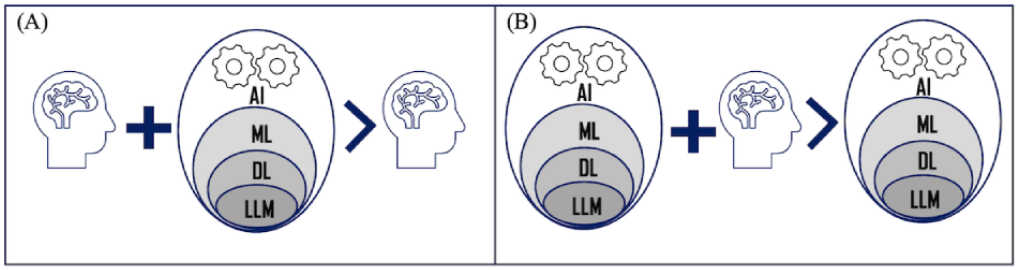
Figure 1. Essential ingredients for equitable distribution of large language models (LLMs) comprising 3 interconnected components: (1) digital readiness, (2) infrastructure workforce and privacy, and (3) LLM onboarding pathways (training from scratch pathway [TSP], fine-tuned pathway [FTP], and out-of-the-box pathway [OBP]) and applications to impact health care outcomes and key performance indicators (KPIs) in health care settings. AI: artificial intelligence; API: application programming interface; EDW: enterprise data warehouse; ETL: extract, transform, and load; MRI: magnetic resonance imaging.



Artificial Intelligence

Operational definition of artificial intelligence (AI) relies on the Turing test, which emphasizes the ability of computers to imitate humans in performing certain tasks [19]. These include automated reasoning, machine learning (ML), and NLP [19]. The hierarchical relationship of AI, ML, DL, and LLMs is shown in Figure 2. In contrast to classical statistical hypothesis testing, AI, ML, or DL assist in discovery, hypothesis generation, and validation in an evidence-based and data-driven manner, with LLMs exhibiting emergent abilities. ML is a branch of AI with a focus on the ability of machines to learn patterns from experience for a given task in an automated manner and draw inferences on previously unseen instances. Popular ML approaches with health care applications include supervised learning, unsupervised learning, reinforcement learning, and association mining [20,21]. DL [22-25] is a subfield of ML that specifically uses neural networks with multiple hidden layers to capture patterns of varying resolution in the given data. Unlike traditional ML, model parameters in deep neural networks (DNNs) can be considerably large with significant storage and computational demands. For example, DL models, such as transformers [26], an essential ingredient of LLMs, have billions of parameters [27]. The layered architecture of DNNs has also been shown to support "transfer learning," where an existing neural network trained on a large dataset can be reused as a base network for predictions on related datasets by retraining only a subset of the layers. This is in stark contrast to traditional, shallow ML models where all model parameters may be altered upon retraining. Recently, proposed foundation models [5,14], exploit the transfer learning ability of DNN, where a "base" model trained on large multimodal data is subsequently adapted for various downstream tasks. LLMs use transfer learning in conjunction with augmented intelligence for enhanced performance, as discussed in the subsequent sections.

Figure 2. Augmented intelligence along with hierarchical representation of artificial intelligence (AI), machine learning (ML), deep learning (DL), and large language model (LLM) is shown in panels (A) and (B), respectively. The primary driver is shown to the right of the inequality in each of the panels.



Augmented Intelligence

Augmented intelligence emphasizes the importance of AI and humans working in concert for enhanced performance and generalization ability (Figure 2). The fundamental theorem of biomedical informatics by Friedman [28] emphasized the importance of information resources assisting domain experts by complementing the knowledge of the domain expert. In the present context, it essentially addresses the question, "Can domain experts in partnership with an AI resource lead to better insights than those unassisted?" (Figure 2A). As noted earlier (Figure 1), health care data are typically large, high-dimensional, multimodal, ingested from diverse sources and evolving rapidly, challenging manual interpretation. Therefore, AI models can assist in gaining novel insights from these datasets in an evidence-based manner while validating what is already known. The primary driver in Figure 2A is the domain expert with AI assisting the knowledge discovery process. On a related note, AI models implicitly subscribe to optimizing a chosen objective function and may converge prematurely to a local optimum. Several factors dictate the convergence aspects of these models [29]. This raises the question, "Can AI models with feedback from domain experts perform and generalize better than those unassisted?" (Figure 2B). More specifically, the role of the domain expert can assist in narrowing the set of variables in a context-specific manner, significantly reducing the search space of potential solutions and improving the performance and generalization ability while imposing necessary guardrails for optimal performance. Unlike Figure 2A, the primary driver in Figure 2B is AI, with the domain expert assisting in the knowledge discovery process. LLMs incorporate human feedback (reinforcement learning with human feedback) in minimizing bias, toxicity, hallucinations with improved performance, and generalization ability [30].

LLMs in Health Care

Factors Impacting LLM Implementation

This section discusses three critical factors accompanying successful LLM implementations and deployments in health care settings: (1) *digital readiness*, (2) *infrastructure and workforce*, and (3) *privacy, ethics, and regulatory aspects* (Figure 1). These factors are closely related to the analytics maturity of a health care organization (eg, Healthcare Information and Management Systems Society Adoption Model for Analytics Maturity) [31], and their impact varies across the 3 LLM implementation pathways (TSP, FTP, and OBP).

Digital Readiness

Typically, LLMs (eg, LLaMA [Meta AI], GPT-4 [OpenAI], Med PALM-2, and Claude) are trained on massive amounts of data (eg, TB) integrated from multiple data sources, including those available publicly [32,33]. Large datasets used to train LLMs often lead to sizeable models with billions of parameters with a direct impact on their performance [27], while marking the transition from language models to LLMs with emergent abilities. So, access to large digital health care data (Figure 1) is critical in developing LLMs with superior performance. However, accessing sensitive health care data (eg, protected health information [PHI] and personally identifiable information) to train LLMs poses significant privacy and security challenges. Health care data, such as clinical narratives, are primarily governed by regulations, such as Health Insurance Portability and Accountability Act (HIPAA) [34] in the United States and the General Data Protection Regulation (GDPR) [35] in Europe. HIPAA establishes national standards for the protection of individually identifiable health information by covered entities and their business associates. Similarly, GDPR, although a European Union regulation, is designed to protect the privacy of European Union citizens and residents and applies to all organizations regardless of location. Both HIPAA and GDPR impose clear regulations on the release and sharing of health care data, with civil and criminal penalties for violations. Addressing privacy and security challenges associated with accessing sensitive health care data for training LLMs demands a robust data management strategy. Currently, there are 2 key strategies that stand out in this context: *data deidentification* and *confidential computing* [36,37]. Data deidentification involves removing personally identifiable information, both direct and indirect, from datasets to reduce the risk of patient reidentification. This allows the use of clinical data for model training purposes without compromising patient privacy. The deidentification process involves techniques such as removing names, addresses, social security numbers, and other direct identifiers, as well as managing quasi-identifiers such as dates of birth, gender, and medical diagnoses. However, these identifiers could potentially be used in combination with other information to reidentify individuals [38,39]. Therefore, deidentification singularly is not devoid of reidentification risks. The granularity of the data upon adequate deidentification can also significantly impact what could be effectively inferred. A recent study by de Kok et al [40] elucidated some of the challenges and best practices for sharing health care data compliant with GDPR. Subsequently, 4 approaches were proposed [40] for sharing open health care data (*consent*

pseudonymized, *no consent pseudonymized*, *no consent anonymized*, and *no consent cloud*). However, the sensitive nature of health care data sharing has a direct impact on the sample size for training and fine-tuning LLMs. In general, health care datasets are relatively smaller by several orders of magnitude compared with datasets used to train popular general-purpose LLMs. Recently, Jiang et al [12] pretrained their LLM (NYUTron) primarily on health care data comprising 4,112,249,482, nearly 4 billion words, resulting in a 109-million-parameter model. In a related yet independent study, Yang et al [41] pretrained their LLM (GATORTron) primarily on health care data comprising 82 billion deidentified clinical words, resulting in 3 different LLM models of varying sizes (baseline: 345 million parameters, medium: 3.9 billion parameters, and large: 8.9 billion parameters). It might not be surprising to note that the size of these LLMs was markedly smaller than general-purpose LLMs (eg, LLaMA: 65 billion, GPT-3: 175 billion, and Google Pathways Language Model: 540 billion) [27] by several orders of magnitude. Empirical studies on the scaling behavior of LLMs reported improved performance and emergent behavior with increasing number of parameters (ie, size of the model) [42,43], size of the data, and compute time [44]. However, sample size challenges are likely to persist for health care datasets. Recent efforts in boosting the sample size of the training data from LLMs have explored supplementing health care data from EDW with biomedical text from sources, such as PubMed [41]. Approaches, such as federated learning (FL) [45-48] can also assist in overcoming some of the sample size-related challenges. However, FL could be susceptible to data leakage, breach, and fair and equitable data representation [49].

Multimodal health care data (eg, text, images, and signals) are usually integrated from diverse enterprise source systems (eg, electronic medical record, registries, the Internet of Things, next-generation sequencing, and Picture Archiving and Communication System) and reside in a centralized EDW (Figure 1). EDWs support querying, reporting, and enterprise analytics. Data governance and ETL are essential ingredients of EDW implementation, dictating the quality and granularity of the data ingested by downstream analytics tools, such as LLMs, impacting their performance. Data governance and ETL can also exhibit marked variations across health care organizations attributed to several factors, including variations in source systems and business processes. More importantly, EDWs demand upfront investment, continued support from executive leadership, and an existing, evidence-based culture. EDW economics is governed by several factors, including (1) architecture of EDW (eg, Inmon [50] and Kimball and Ross [51]); (2) source systems; (3) type, velocity, and volume of data; (4) enterprise data governance and ETL processes; (5) infrastructure and workforce supporting secured storage and retrieval; and (6) existing culture of data-driven and evidence-based approaches in impacting outcomes and KPIs in the health care organization. It is important to note that ETL and EDW can exhibit marked variations across organizations, posing challenges in seamless sharing of health care data, deployment of AI and ML models, and demonstrating their generalization ability. Common data models, such as the Observational Medical Outcomes Partnership, have been helpful in data standardization, data sharing, and federated querying [52,53]. However, there are inherent limitations to standardization, including incompleteness in data models and terminologies resulting in data that cannot be mapped [54], errors in mapping [55], and the potential loss of information due to granularity mismatches between the source data and the standard. In addition, data standards require significant expertise and a steep learning curve in their absence. The challenges mentioned earlier can especially be accentuated across health care organizations that primarily serve the economically disadvantaged, underserved, and marginalized communities. Factors contributing to data inequities [56] are multidimensional and include ethnicity (eg, Hispanic), race (eg, African American), disease groups and treatment regimens (eg, rare disease), gender or gender identity (eg, lesbian, gay, bisexual, transgender, and queer [LGBTQ]), age (eg, pediatric population), geographic location (eg, rural areas), language barriers (eg, Spanish), digital divide, and patient literacy [57,58]. It is of interest to note that these dimensions are not mutually exclusive, and their combination can significantly impact the representation of a given population in the data and digital readiness of organizations that serve these communities. Inequity along these dimensions can lead to potential biases [59,60]. These biases broadly fall under *systemic*, *statistical*, *computational*, and *cognitive* bias [61]. Systemic bias can be further categorized into *measurement bias*, *missing validation bias*, *label bias*, and *modeling bias* [61]. Measurement bias [62,63] may be the result of variations in quality and representation of the entities of interest across subpopulations in the data. Missing validation bias may result due to a lack of adequate validation studies across certain populations. It is important to note that measurement bias may accentuate validation bias because validation at small sample sizes can be statistically challenging. Label bias may be a result of surrogate variables substituting the actual health care outcomes of interest. Modeling bias is attributed to the biased results generated by a specific model. Computational and statistical biases can be a result of inadequate representation of select groups and populations in the given data. Human cognitive bias is a bias due to human perception of AI and ML systems. Mitigating human cognitive bias may be critical for successful adoption of AI systems. From this discussion, digital readiness is critical for health care organizations to embark on their LLM journey and could dictate the choice of the LLM implementation pathway.

Infrastructure and Workforce Needs

Infrastructure Needs

LLMs primarily rely on the transformer architecture [26], with the ability to be trained in a massively parallel manner on sequential data. However, parallel processing in turn demands specialized hardware accelerators, such as graphics processing units (GPUs), and seamless interfaces between various hardware components in conjunction with significant network bandwidth for avoiding storage, data transfer, latencies, and bottlenecks (Figure 1). This is especially true when the training involves large amounts of data and pretrained LLMs have billions of parameters. Parallelism falls under 2 broad categories, namely (1) *data parallelism* [64] and (2) *model parallelism* [65]. Data

parallelism addresses challenges with large training datasets that cannot fit within a single GPU by partitioning and distributing them across multiple GPUs. This is especially helpful because optimization techniques, such as stochastic gradient descent, used by LLMs rely on small batches of data that can be distributed across GPUs. Model parallelism addresses challenges with the size of the LLMs (eg, billions of parameters) by distributing them (eg, weights and layers) across multiple GPUs and further subdivided into *pipeline parallelism* and *tensor parallelism*. While pipeline parallelism facilitates the distribution of the layers of a DNN [66], tensor parallelism [67] distributes the tensor computation across hardware accelerators. While partitioning the data and model is useful in overcoming some of the challenges with their size, fast interconnect between the GPUs is especially critical for enhanced communication between them in a cluster (eg, DGX A100 Data Center) and between clusters (eg, Super PODs) through specialized high-bandwidth (eg, 900 GB/s) communication links, such as NVLink/NVLink Switch. Training LLM models using multiple GPUs is also accompanied by significant carbon footprint and heating, demanding specialized cooling systems for optimal performance [68]. While LLM implementations are traditionally dependent on multiple software libraries [69], there has been recent interest in developing graphical user interfaces for LLMs to alleviate some of these challenges for the end users [70]. The size of the GPUs required is dependent on several factors, such as the size of the LLM models so that a model can be loaded into the GPU successfully while permitting necessary computation. For instance, LLM models with 7 billion parameters (13 GiB) may need a 16-GiB GPU, such as NVIDIA T4 Tensor Core for processing, while those with 13 billion (25 GiB) and 70 billion parameters (130 GiB) might require a 32 GiB (eg, NVIDIA-A100 GPU Server) and 160 GiB GPUs (eg, NVIDIA 2 × A100 multi-GPU Server), respectively.

Cloud-Computing Platforms

On-premise infrastructure had supported more traditional AI and ML implementations and analytics dashboards in the past. However, increasing size of the data and compute along with evolving needs of health care organizations, demand scalable infrastructure for storage (horizontal scaling) and computing (vertical scaling), with the latter playing a critical role for LLMs. For instance, data and model parallelism demand scalable infrastructure that can vary with training data size as well as the model size. Cloud-computing environments (Figure 1) provide infrastructure as a service and software as a service (SaaS) with pay-as-you-go payment models to address challenges from scalability, privacy, workforce, and economic standpoints. CSPs (AWS, GCP, Azure, and OCI) offer robust solutions for LLM deployment by leveraging their global infrastructures built around *high availability*, *enterprise security and compliance*, *low latency* access to computing resources, and managed services as needed. CSPs can play a critical role in bridging the chasm between needs, such as growth scale and security of information technology infrastructure, and affordability by leveraging existing technology investments on-premises or on other clouds. More importantly, CSPs can help facilitate enhanced democratization and equitable adoption of LLMs by the broader health care communities and not by a privileged few, a critical aspect for the equitable distribution, widespread adoption, and long-term success of LLMs in health care. CSPs also provide hybrid options, such as “cloud bursting,” that allow organizations to use their private cloud, on-premise infrastructure, for routine operations and burst into a CSP temporarily when additional computing resources are needed to handle peak loads and prevent queuing of computational workloads. A hybrid approach enables health care organizations to handle the variable and intensive computational demands of LLMs at scale. By cloud bursting, health care organizations can maintain a cost-effective private cloud for steady-state workloads and then burst into public clouds during periods of high demand without needing to overprovision their private cloud infrastructure while minimizing latencies. CSPs provide essential tools to help organizations manage security and compliance risks. AWS identity access management and Azure controls are examples of services that integrate into existing organizational security investments and aid in configuring fine-grained access controls and ensuring authorized access to sensitive data. Hybrid cloud offerings from AWS, GCP, and Azure also adhere to regulations and compliance (eg, HIPAA) for handling health care data containing PHI through data encryption and network security between on-premises and on-cloud workloads, creating a single flexible, cost-effective enterprise infrastructure technology solution. These environments can also assist in minimizing on-premises workforce needs and advanced skillset needs using readily available cloud-based tools and managed services. This aspect is especially critical for health care organizations that do not have sufficient up-front investment and a history of supporting analytics teams but would like to experiment with the utility of LLMs for impacting outcomes. In essence, CSPs can assist in equitable adoption by starting small with existing resources and scaling out as needed without large upfront investments. Utility services from CSPs can significantly reduce the operational overhead and technical challenges associated with LLM development, testing, and scale, allowing for health care organizations to focus more on outcome-based applications rather than infrastructure management [71]. Serverless options by CSPs for storage and computing are also expected to minimize carbon footprints [68].

Cloud-Based Hardware Accelerators

Hardware accelerators assist in overcoming data and computational bottlenecks working in concert with base processors (eg, central processing units) [72]. Several types of accelerators, such as application specific integrated circuits, field programmable gate arrays, GPUs, and dedicated chips for AI, have been explored [73-76]. Accelerators, such as GPUs, as noted earlier, have become an integral part of DL models and LLMs in overcoming computational bottlenecks. However, there is renewed interest in developing hardware accelerators for LLM training and inference that are affordable. Hardware accelerators offered by CSPs include Trainium and Inferentia (AWS), tensor processing units (TPUs; GCP), Maia (Azure), and MI300X (OCI). These chips are optimized for specific workloads, cost-effective relative to GPUs, and interface to popular open-source environments. However, multiple factors, such as the choice of open-source environment and

machine-learning libraries (eg, Tensorflow and PyTorch), can impact the benchmarking and performance of hardware accelerators [77].

- AWS—Trainium AI accelerator [78] supports training DL models with >100 billion parameters and up to 50% in cost-to-train savings over comparable elastic compute cloud instances. Inferentia [79] supports inference, delivering up to 2.3× higher throughput and up to 70% lower cost per inference than comparable elastic compute cloud instances. Both are supported by the AWS Neuro software development kit, which integrates natively with open-source DL environments such as PyTorch, TensorFlow, and HuggingFace. Inference on Meta's llama 8B would cost US $0.99/h on Inferentia, a savings of >65% on an NVIDIA A10G GPU instance.
- GCP—TPUs [80] are custom-designed AI accelerators optimized for training and inference of AI models [81]. They scale cost-efficiently for a wide range of AI workloads, spanning training, fine-tuning, and inference. TPUs provide the versatility to accelerate workloads on leading AI frameworks, including PyTorch, JAX, and TensorFlow.
- Azure—Maia 100 AI accelerator [82] is supported by the Maia software development kit and interfaces to open-source frameworks such as PyTorch, ONNX Runtime, and Triton from OpenAI. It can support services such as Microsoft Copilot and Azure OpenAI Service.
- OCI—MI300X accelerators are powered by AMD's CDNA 3 architecture, offering memory bandwidth and compute performance supporting a broad range of precision data that enable OCI to support larger and more complex computations for AI and ML workloads [83-85]. Compared with NVidia's H100GPUs, MI300X could provide a cost-effective alternative while still delivering competitive performance with its 304 compute units, 19,456 stream cores, and 1216 Matrix cores.

Cloud-Based Managed Services

Managed services from CSPs can assist in LLM implementation and management while minimizing on-premises workforce needs. Managed services features can be readily accessed through application programming interfaces (APIs), accelerating implementations with an enhanced focus on impacting health care outcomes. This may include access to multiple LLMs across vendors via a single API, enabling experimentation by end users without managing multiple end points, keys, and payloads. This permits experimentation with new LLM models and automates this process via CSP-managed services. These features are especially helpful in exploring the available LLMs that best suit the current needs of the health care organization. In contrast, the exploration phase could be fairly involved across on-premise implementations, accompanied by multiple tokens, API end points, and access controls. CSP-managed LLM services also facilitate centralized governance structures for access management, billing, auditing, and security scanning. Given the sensitive nature of health care data, using fewer end points allows health care organizations to set up the necessary access controls for data and APIs in a seamless manner.

Workforce Needs

LLM implementation and deployment in health care workflows typically demands a workforce with expertise across a spectrum of areas (Figure 1). However, the workforce needs will be dependent on several factors that include (1) LLM implementation pathway, (2) applications of LLM in health care workflows, (3) commercial or open-source platforms. and (4) on-premise or cloud-based implementations. Workforce needs are especially critical across LLM implementations that involve training and fine-tuning. This would ideally consist of a (1) core team comprising data scientists, architects, engineers, and CSPs with expertise in areas such as warehousing, NLP, LLMs, ML implementation, deployment, and assessment (eg, MLOps) [86] in concert with (2) an infrastructure team that addresses storage and computing needs on-premises and on-cloud, (3) subject matter experts and health care personnel who guide the implementation, reinforcement learning, and prompt engineering aspects of LLMs for optimal performance, (4) regulatory and compliance teams for ensuring ethical use of health care data and establishing guardrails, and (5) information technology that assists in ensuring privacy and security of health care data while deploying the LLM applications or API in enterprise workflows to impact outcomes and KPIs (Figure 1). Agile strategies (eg, DevOps or MLOps) may be critical for implementation, validation, and seamless deployment of LLMs in workflows. Given potential bias and toxicity that may accompany LLM implementations, an inclusive framework that incorporates critical feedback from stakeholders, patients, providers, and subject matter experts across diverse communities can minimize bias and assist in developing the necessary guardrails.

Privacy, Ethics, and Regulatory Aspects

Privacy and Security in the Cloud

Privacy and security are critical for storage, retrieval, and analysis of health care data (Figure 1). There has been an increasing shift in moving health care data and analytics from on-premises to the cloud [71]. CSPs provide confidential computing environments (CCEs) that facilitate computations in hardware-based trusted execution environments (TEEs). These ensure sensitive data (eg, PHI data) to remain encrypted in an isolated environment, preventing modification of data and applications by unauthorized parties, including CSPs, during processing and transmission [87,88]. They facilitate secure collaboration of first- and third-party data with the potential to assist in overcoming sample size constraints for LLM training as discussed in this section. CCE capabilities across the 4 major CSPs (AWS, GCP, Azure, and OCI) are discussed in the subsequent sections.

- AWS—Confidential computing capabilities through the processor agnostic AWS Nitro System and AWS Nitro Enclaves, enabling secure isolation of sensitive workloads [89].
- GCP—Confidential virtual machines (VMs) and Confidential Google Kubernetes Nodes allow customers to process sensitive data while keeping them encrypted in memory [90].

- Azure—Confidential computing VMs with AMD SEV-SNP and Intel SGX support ensuring VM-level confidentiality and protection from cloud operators [91].
- OCI—Confidential computing through Oracle's confidential instances leveraging AMD Secure Encrypted Virtualization for VMs and AMD secure memory encryption for bare metal instances protecting data and application processing the data [92].

Attestation

Attestation is a critical component of CCEs that ensures the trustworthiness of the computing environment. It allows the integrity and authenticity of the hardware, software, and configuration to be verified, effectively establishing trust between parties. Attestation offered by major CSPs (AWS, GCP, Azure, and OCI) is discussed in the subsequent sections.

- AWS—CCE attestation involves the use of an attestation document signed by the Nitro Hypervisor. This document is critical for providing the identity of the enclave to AWS Key Management Service (KMS), which validates the document against the KMS key policy. This allows the enclave to perform cryptographic operations with KMS keys [93].
- GCP—Provides attestation through the Binary Authorization and Certificate Authority Service. Through this service, the confidential workload collects measurements of itself and the TEE and sends an attestation request to the Binary Authorization service, which compares the measurements against an attestation policy. If they match, service returns a signed attestation [94].
- Azure—Provides CCE attestation via the Microsoft Azure Attestation service. The confidential workload includes an attestation client that collects measurements and evidence from the TEE. It then sends an attestation request with its evidence to the Microsoft Azure Attestation service that is verified against policy. If valid, it returns a signed attestation token.
- OCI—Provides attestation using a hardware-based trusted security module that generates an attestation report containing measurements of the hardware and firmware environment and verified by the customer to ensure confidential workload is running in a legitimate TEE.

FL Architecture

CSPs also provide FL architectures [47] for decentralized training of LLM addressing sample size challenges. An FL model is trained locally and refined through shared updates, resulting in an aggregated global model without explicit sharing of health care data. Using federated data for training may leverage collective knowledge, perhaps resulting in models with enhanced generalization ability. It is important to note that because only model updates are shared instead of the actual data, FL implicitly minimizes the amount of data transferred over the network. This can be particularly beneficial in scenarios where data transfer is costly or limited by bandwidth constraints. FL frameworks developed by CSPs include:

- AWS—FedML on AWS is an open-source library that supports several FL models. AWS provides tools, libraries, and algorithms to implement and experiment with FL algorithms in a private and secured environment [95].
- GCP—TensorFlow Federated is an open-source framework for ML on decentralized data. TensorFlow Federated is used to implement FL on Google Cloud, leveraging Google Kubernetes Engine for hosting and managing the FL process [96].
- Azure—the AzureML platform supports Azure FL frameworks NVFlare and Flower for running a FL pipeline. Azure's capabilities are leveraged for provisioning and orchestration of FL algorithms [97].
- OCI—Supports FL through various tools, frameworks, and services, such as Private Federated Learning with Domain Adaptation [98].

Confidential FL

While FL ensures compliance with data protection regulations, such as HIPAA and GDPR, it does have some limitations, as noted earlier [49]. A possible solution is to combine confidential computing and FL, resulting in confidential FL (CFL) [99]. This decentralized approach works well for a hybrid cloud environment that spans on-premise data centers, edge devices, and public clouds from different CSPs. Confidential computing TEEs secure the data during processing, while FL enables collaborative training without explicit sharing of health care data. Incorporating a deidentification process as a part of CFL workflow ensures access to large sensitive data for training LLMs without compromising privacy and security while maintaining the integrity of health care data management. Key characteristics of CFL across CSPs are discussed in the following sections.

- Enhanced privacy and security—while deidentification removes identifiable information from the data, CFL ensures that the data are processed in a secure and private manner.
- Compliance or regulations—Deidentification and CFL can help health care organizations comply with HIPAA and GDPR. CFL provides the necessary security measures to protect data in use. Both technologies address regulations regarding data security.
- Facilitation of data sharing—by combining these technologies, health care organizations can safely engage in collaborative data sharing initiatives and develop LLM models with enhanced performance and generalization ability.
- Intellectual property protection—CFL can protect intellectual property such as health care AI algorithms and research data during collaborative training.
- Building trust—secure handling of health care data builds trust among patients, providers, and payers.

While CFL has the potential to accelerate LLM implementation using sensitive federated health care data, some of the challenges listed here need to be addressed for its successful deployment and adoption.

- Data heterogeneity—CFL assumes that the data across participating organizations are independently and identically distributed. However, in practice, health care data may exhibit significant heterogeneity, which can impact the performance of the models. Techniques, such as transfer

learning and domain adaptation, can be used to address this challenge.
- Communication efficiency—the iterative nature of CFL involves frequent communication between the participating organizations and the central data server. Efficient communication protocols and compression techniques are necessary to minimize the communication overhead and ensure scalability.
- Model interpretability—CFL models may lack interpretability due to the distributed nature of the training process. Techniques, such as model distillation and explainable AI can be used to improve the interpretability of CFL models.
- Incentive mechanisms—encouraging health care organizations to participate in CFL initiatives may require appropriate incentive mechanisms. Developing fair and transparent incentive approaches that align with the interests of all stakeholders is an important consideration for success.
- Human in the loop—integrating human expertise and oversight as a part of the LLM training and decision-making process ensures that the models are accurate, reliable, and aligned with human values. Human in the loop also ensures that the models comply with legal and regulatory requirements, such as data protection laws and medical standards.

Ethics and Regulatory Aspects

The potential of AI tools, such as LLM, to transform health care outcomes does come with various ethical and regulatory challenges [100-102]. US president Joe Biden's October 2023 executive order [103] underscored the necessity of ensuring AI safety and security. It mandated AI-generated content to be clearly identified and called for substantial investments in AI-related education, training, and research. The order emphasized protecting intellectual property, supporting American workers, advancing equity, civil rights, while safeguarding privacy and civil liberties. It also directed the Department of Health and Human Services to establish safety parameters for AI, including frameworks for identifying and tracking clinical errors, generating improvement guidelines, and sharing these among health care organizations. A recent Health and Human Services 2024 ruling (Section 1557, Patient Protection and Affordable Care Act) also emphasized protection to patients against bias and discrimination from AI and ML decision support tools and the importance of mitigating such biases [104,105]. In addition, 29 countries attended the AI Safety Summit in November 2023 and signed the Bletchley Declaration [106] to "cooperate on AI to promote inclusive economic growth, sustainable development, and innovation, to protect human rights and fundamental freedoms, and to foster public trust and confidence in AI systems to completely realize their potential." The Institute for Healthcare Improvement's Lucian Leape Institute (LLI) [107] predicted increased use of AI in clinical documentation support, clinical decision support, and patient-supportive chatbots in the health care setting. They recommended prioritizing patient safety, engaging clinicians, ensuring AI efficacy and bias mitigation, establishing AI governance and oversight, and fostering collaborative learning across health systems. In addition, LLI emphasized the importance of AI systems in reporting confidence levels and rationale and the need for continuous human monitoring to maintain trust and accuracy in AI-generated outputs. LLI also suggested several considerations for regulators and policy makers: establishing clear guidelines for ethical and trustworthy AI use, promoting transparency and accountability, supporting AI literacy, incentivizing AI development that prioritizes safety, and facilitating localized decision-making. The European AI Act [108], the first legal framework on AI, categorized AI risks into 4 levels—unacceptable, high, limited, and minimal—and introduced transparency obligations for all AI models. Accreditation agencies, such as the Joint Commission, will need to advocate to create governance structures and processes for monitoring patient safety issues related to AI. The World Health Organization also recently commented on regulatory considerations for AI in health [109]. It outlined essential guidelines covering documentation and transparency, risk management, intended use validation, data quality, privacy protection, and stakeholder engagement. The ethical development of AI must adhere to principles, such as beneficence, nonmaleficence, autonomy, justice, data quality, transparency, fairness, responsibility, privacy, freedom, trust, sustainability, dignity, and solidarity. Trustworthiness of AI-based clinical decision support is often compromised by the lack of transparency in how AI tool's function and the basis of their decisions. There are concerns about the use of proprietary data, the absence of robust regulation, and the risk of bias from datasets that do not adequately represent marginalized populations. Overall, as AI continues to evolve and integrate into health care, maintaining a balance between innovation and ethical responsibility is crucial. Regulatory frameworks and ethical guidelines at health care organizations must evolve to ensure that AI enhances health care delivery while protecting the interests and rights of patients and providers alike.

LLM Guardrails for Responsible AI

There has been interest in developing guardrails and regulatory frameworks to facilitate responsible AI. These guardrails ensure the behavior of AI tools, such as LLM, falls within expected bounds while being resilient to adversarial attacks. These efforts include recent open-source initiatives, such as NeMo Guardrails [110] for improved trustworthiness [111] of LLM conversational systems. These guardrails assist in customizing the LLM interaction with users using *topical rails* and *execution rails* [110]. While topical rails prevent the LLM from veering off topic, execution rails assist in moderating the LLM output and ensure it is factual. A recent study by Meskó and Topol [102] on regulatory oversight of LLMs identified several associated challenges. The study pointed out the challenges in translating existing Food and Drug Administration's regulatory frameworks for medical devices [112] to contemporary AI-based technologies, such as LLMs, and the need for new regulatory frameworks for LLMs. More specifically, it highlighted 2 unique characteristics of LLMs: (1) the ability to adapt their performance to training data as well as tasks in contrast to traditional AI and ML approaches, and (2) the ability to learn in a self-supervised (autodidactic) manner without the need for explicit guidance and ground truth labels as in a more classical supervised ML paradigm. Subsequently, a list of LLM

regulatory challenges were identified (Table 1 in the study by Meskó and Topol [102]), including privacy, intellectual property, medical malpractice liability, quality control and standardization, informed consent, interpretability and transparency, fairness and bias, data ownership, overreliance, and need for continuous monitoring and validation. Some of these challenges have also been acknowledged in a more recent US Food and Drug Administration release with a broader focus on AI and medical products [113].

Table 1. Summary of risks and benefits of the 3 large language models (LLMs) onboarding pathways (training from scratch pathway [TSP], fine-tuned pathway [FTP], and out-of-the-box pathway [OBP]).

	TSP	FTP	OBP
Digital readiness			
Benefits	• LLMs are trained on health care data and capture characteristics of that target population with the potential to impact outcomes in that population. • Enhanced transparency of the data, implementation, and deployment. • Enhanced quality of training data through enterprise governance, minimizing bias.	• Digital readiness of FTP is much lesser than TSP because FTP focuses on fine-tuning as opposed to training.	• Digital readiness for OBP is minimum.
Risks	• Demands upfront investment in data warehousing, enterprise governance, and dedicated workforce.	• Susceptible to biases in training data used in the pretrained LLMs. • General purpose LLMs are often trained on nonhealth care data. • Prompt engineering demands can be significant.	• General purpose, out-of-the-box models pretrained on nonhealth care data may have limited utility, prone to bias, and hallucinations.
LLM			
Benefits	• Train LLM from scratch using either existing transformer architectures or novel architectures. • Long-term maintenance, customization, with evolving needs.	• Uses off-the-shelf pretrained LLMs without explicitly training from scratch. • The number of open-source pretrained LLMs continues to grow.	• Multiple choices of off-the-shelf LLMs accessed as APIs[a]. • Readily accessible with minimal training.
Risks	• Cost of training LLMs can be significant. • Sharing checkpointed LLMs trained on PHI[b] data is a risk. • Novel architectures demand considerable experimentation for optimal performance. May result in implementation and deployment delays.	• Susceptible to biases in the pretrained LLMs. • Pretrained LLMs on nonhealth care data may have performance limitations. • Dependency on pretrained LLMs is a risk. Limited transparency may be a security risk.	• Use is dependent on the features exposed by the vendors. • Generic nature of the output may have limited utility in addressing the unique needs of health systems. No explicit training of the LLM model.
Workforce			
Benefits	• On-premise workforce can assist in customizing LLMs with enhanced transparency and evolving needs.	• Workforce demand is significantly less than TSP.	• Minimal on-premise workforce needs and training. • Rapid implementation using managed services.
Risks	• Demands a skilled workforce with expertise across a spectrum of areas for implementation and deployment. • Demands recruitment, growth, and retention of skilled workforce. • Continued buy-in from leadership for sustaining workforce.	• Limited customization of the pretrained LLMs.	• Complete dependence on vendor models and services with minimal transparency
Infrastructure and security			
Benefits	• Options for training LLMs across cloud service providers with on-demand and spot instance pay-as-you-go pricing models. • Secured cloud environments are available	• Infrastructure needs for compute are significantly less than TSP.	• Infrastructure is needed primarily for inference.

	TSP	FTP	OBP
Risks	• Training LLMs on GPUs[c] is expensive. • Security and governance for training sensitive data in the cloud.	• Vendor pricing may vary based on adoption. • Prompt engineering on pretrained models can be involved.	• Infrastructure costs increase with the number of users. • Vendor pricing may vary with increasing adoption. Availability of services is not guaranteed. • Interfacing to downstream applications.

[a]API: application programming interface.

[b]PHI: protected health information.

[c]GPU: graphics processing unit.

LLM Implementation Triumvirate

Overview

As noted earlier, LLM implementation is dependent on several factors. In this section, 3 broad LLM implementation pathways (Triumvirate) are discussed along with associated risks, benefits, and economics (Figure 1). These pathways are not necessarily independent and expected to serve as onboarding points for equitable distribution and enhanced adoption of LLMs. A summary of the risk and benefits of these 3 pathways is also enclosed in Table 1 for convenience.

The TSP

In TSP, an LLM is trained from scratch using health care data and subsequently customized for specific needs and tasks of the health care organization. FTP and OBP may follow TSP.

Benefits

The TSP provides enhanced transparency of the data and code, complete ownership of the model parameters, implementation, and the ability to assess the quality of the training data at the most granular level. TSP is expected to facilitate long-term maintenance of the model and customization to the evolving needs of the health care organization, including seamless deployment in workflows and interface to enterprise dashboards. In addition to using an established LLM architecture, TSP may also implement novel architectures or modifications to existing general-purpose transformer architectures. Training LLMs using data from the EDW of a health care organization implicitly adheres to the data governance and ETL ensuring high-quality data and accommodating characteristics of the population served by the health care organization that may not necessarily be captured in generic datasets, such as those used to train general purpose LLMs. This in turn is expected to result in a model with enhanced performance [114] addressing the needs of the organization impacting clinical, operational, and financial outcomes and KPIs (Figure 1). Such a model is also expected to be better used while demonstrating value because the training data can significantly impact its behavior and performance [115]. As noted earlier, it is not uncommon for the organizational data to be supplemented by high-quality external data during training [41]. Unlike pretrained LLMs, access to the training data and enhanced transparency may assist in mitigating biases, minimizing perpetuation and amplification of biases, and reducing toxicity by the model as well as downstream applications and APIs that are dependent on the base model (Figure 1), leading to improved overall performance. This aspect is especially critical when deploying the LLM model in clinical workflows to assist in clinical decision-making.

Risks

TSP implicitly demands digital readiness, infrastructure and workforce, and regulatory compliance. Because sample-size challenges can impact TSP, approaches and FL techniques may be explored. The digital and analytics maturity of an organization, along with an existing culture of data-driven and evidence-based approaches to impacting outcomes, may be critical for successful TSP implementation and continued buy-in from the enterprise leadership. Unlike pretrained LLMs, TSP may demand experimentation to identify the optimal model size, parameters, and checkpointing the model before deployment. Agile implementation strategies across multiple teams, such as data science, information technology, clinical, and support from executive leadership may be critical for the timely progress of TSP. Therefore, timelines for TSP implementation and deployment are expected to be significantly larger than FTP and OBP. In addition to implementation and validation, timelines would also include seamless deployment in health care workflows and providing necessary training for the end users. Delays in demonstrating value are to be expected as with any new AI tool. TSP will also demand access to specialized infrastructure for storage and computing, including distributed frameworks and GPU Clusters or PODS. Enterprise CSPs can be critical partners in this regard. Unlike general purpose LLMs, TSPs using cloud infrastructure should follow strict compliance and security protocols that in turn may incur additional costs. Because TSP demands unique skill sets for implementation and critical evaluation, existing skillsets and investment in the workforce in areas such as data science, DL, NLP, LLM, and IT, as well as protected time for health care personnel for critical assessment of the models, would be important. The quality and performance of TSP will be dependent on the knowledge of the subject matter experts assisting with the reinforcement learning with human feedback process. FTP is likely to follow TSP as a part of customizing the LLMs to end users and downstream applications (Figure 1). Given the large number of parameters of LLM models, there is the possibility of the LLM models memorizing some of the information in the training data [116]. This in turn may discourage sharing checkpointed TSP models due to the risk of information leak.

Economics of TSP

Among the 3 pathways, TSP demands considerable up-front investment with regards to digital readiness, infrastructure, workforce, privacy, and regulatory aspects. TSP implementation will demand (1) an existing EDW for querying and retrieval of unstructured data for ingestion by LLMs; (2) a workforce with competencies across a spectrum of areas, including implementation, integration as well as technical aspects in multiple areas, including DL; (3) concerted working of multiple teams, including subject matter experts for validation and prompt engineering; and (4) regulatory oversight because the training phase in TSP would involve using health care data with the DL algorithm. Data-warehouse implementation and continued management could cost millions of dollars. The cost of hiring and retaining a workforce to support LLM implementation can be substantial, especially given the high demand for such specialized skillsets. For a midsized health care organization, the cost of a workforce capable of handling LLM development, implementation, deployment, and maintenance can range from US $2 million to US $5 million per year, including salaries, benefits, and training. Training and deploying LLMs require significant computational resources, including high-performance storage and computing infrastructure. For example, DGX A100 data centers (80 GB) were priced at approximately US $200,000 in 2020. LLaMA implementation [32] required 2048 A100 GPUs and 21 days for training their 65 billion parameter model, resulting in significant costs in millions of dollars. Therefore, the cost of compute, along with power consumption, physical space requirements, and dedicated personnel, could easily reach into the tens of millions of dollars for TSP. Working in partnership with cloud platforms can address several of these challenges. Major CSPs, such as AWS, Azure, GCP, and OCI, offer on-demand compute instances across GPU clusters and cost profiles based on the user needs and affordability within a secured framework (Table 2) [117-122]. TSP is usually followed by FTP to tune the response of the LLMs.

Because TSP demands building an LLM using the health care organizations data, it may require *hundreds of thousands of hours* of training [123]. Comparable pricing of 320 and 640 GiB of GPU memory using 8 × A100 GPUs in a single instance across the 4 major CSPs is presented in Table 2. On the basis of the comparison table and the time for training an LLM from scratch, it might be economical to purchase a long-term (3-year commitment), which may save around 60% when compared with on-demand costs. Another option would be to use a "spot instance" (Table 3) [124-127]. Spot instances are spare compute capacity offered by CSPs at a reduced cost compared with the on-demand pricing and serve as a suitable alternative. These pricing estimates vary with demand and can change throughout the day, week, or month. However, to use spot instances for LLM training, organizations need to implement strategies to handle instance reclamations and checkpoint management. This is especially critical for TSP, as it takes considerable time to train an LLM from scratch. CSP-managed services offer managed spot training or fine-tuning and resume jobs from the checkpoints. Due to spot interruptions, training or fine-tuning using spot instances may also take longer to complete compared with on-demand or reserved instances. With increasing adoption of LLMs by health care organizations in conjunction with the popularity of AI and ML availing spot instances in general can be challenging and could be prone to interruptions with marked variations in availability as well as pricing across the different geographic regions. Spot instances can also vary across CSPs, with some (eg, AWS, GCP, and OCI) providing more options and higher-memory GPU instances.

Table 2. Comparable per-hour pricing of graphics processing unit (GPU) clusters across cloud service providers (CSPs; Amazon Web Services [AWS], Azure, Google Cloud Platform [GCP], and Oracle Cloud Infrastructure [OCI]) for 320 or 640 GiB GPU memory for the training from scratch pathway (TSP). Representative data retrieved on June 2024.

CSP	Instance type	CPU[a] (cores)	Memory (GiB)	GPUs	Per GPU memory	Total GPU memory	On demand (US $)	On demand per GPU (US $)	1 year (US $)	3 years (US $)
AWS	p4d.24xlarge	96	1152	8	40	320	32.77	4.10	19.22	11.57
AWS	p4de.24xlarge	96	1152	8	80	640	40.96	5.12	24.01	14.46
Azure	ND96asr A100 v4	96	900	8	40	320	27.20	3.40	22.62	13.63
Azure	ND96amsr A100 v4	96	1900	8	80	640	32.77	4.10	20.97	14.42
GCP	a2-highgpu-8g	96	680	8	40	320	29.39	3.67	18.52	10.29
GCP	a2-ultragpu-8g	96	1360	8	80	640	40.22	5.03	—[b]	—
OCI	BM.GPU4.8	64	2048	8	40	320	24.40	3.05	—	—
OCI	BM.GPU.A100-v2.8	128	2048	8	80	640	32.00	4	—	—

[a]CPU: central processing unit.

[b]Not applicable.

Table 3. Comparison of per-hour spot-instance and on-demand pricing across cloud service providers (CSPs; Amazon Web Services [AWS], Azure, Google Cloud Platform [GCP], and Oracle Cloud Infrastructure [OCI]) large language models (eg, 320 GiB, 8 × A100 graphics processing units [GPUs] single instance) and smaller and medium large language models (eg, 64, 16 GiB V100 GPU). Representative data were retrieved in June 2024.

CSP	Large models			Medium and smaller models		
	Instance type	On-demand cost per hour (US $)	Spot cost per hour (US $)	Instance type	On-demand cost per hour (US $)	Spot cost per hour (US $)
AWS	p4d.24xlarge	32.77	8.37	p3.8xlarge	12.24	3.97
Azure	ND96asr A100 v4	27.20	8.19	NC24rs_v3	13.46	0.91
GCP	a2-highgpu-8g	29.39	11.75	n1-highmem-32	21.73	3.63
OCI	BM.GPU4.8	24.40	12.20	BM.GPU3.4	12.03	6.02

FTP Overview

While TSP focuses on pretraining LLMs, FTP focuses on adapting an existing pretrained LLM with a given architecture and parameters to tasks at hand in a domain-specific manner. This is usually accomplished by (1) adjusting the model parameters of the LLM using context-specific data that are much smaller than the training data and (2) adjusting the LLM performance and behavior by prompt engineering inputs and outputs of the LLM. The pretrained LLMs can be either open-source or proprietary LLMs, with those pretrained on health care data expected to perform better than general-purpose LLMs.

Benefits

The timeline for implementation, budgeting, infrastructure, and workforce needs for FTP is expected to be significantly lower than that of TSP because it does not involve training LLMs from scratch [114]. Typically, FTP uses readily available, pretrained proprietary or open-source LLM with open-source licenses for modifying the source code as per user needs. The number of open-source LLM offerings has continued to increase with time with communities, such as Hugging Face hosting leaderboards comparing their performance. Managed services by CSPs can assist in setting up multistep tasks across systems and data sources, generate knowledge bases from private data sources for FTP, and implement safeguards on inputs and outputs adhering to governance and responsible AI.

Risks

FTP will demand resources, access to quality data, relevant prompts (eg, input-output pairs), protected time for subject matter experts, and agile implementation strategies for adapting the pretrained LLMs for specific tasks. While automated approaches have been proposed for prompt engineering [128], prompt engineering risks for FTP may be relatively higher compared with fine-tuning on TSP because the training data of the pretrained models may not be domain-specific and can have potential biases. Implementation details of proprietary pretrained LLMs may not be readily accessible, limiting innovation and modification with evolving needs. Pricing of proprietary LLMs used by FTP may also increase with enhanced adoption across health care organizations, and their downtime may impact several dependent downstream applications in health care workflows. While several open-source LLMs are readily available from platforms, such as Hugging Face, these are primarily a result of crowdsourcing efforts and voluntary contributions posing challenges in translating them to enterprise tools. Dependency on an existing pretrained open-source or proprietary LLM can be a risk because these models are rapidly evolving and that could challenge active maintenance of legacy models. Open-source implementations traditionally do not support extensive documentation and training materials for onboarding. These in turn may demand workforce and digital capacity on-premises. This includes challenges in interfacing these tools with other systems and health care workflows. Open-source implementations may also be susceptible to vulnerabilities that may not be readily apparent; hence, they could be susceptible to security breaches, malicious content, malware, and ransomware attacks on models and downstream applications compromising patient privacy and leading to liabilities. Pretrained proprietary LLMs may have minimal flexibility, transparency, and interface options to downstream applications and dashboards in health care workflows. Because health care data can contain PHI, open-source and pretrained proprietary LLMs should be HIPAA compliant.

Economics of FTP

The economics of FTP (Table 4) are expected to be markedly lower compared with TSP as it does not involve the computationally intensive task of pretraining an LLM. More specifically, digital readiness and infrastructure and workforce costs are expected to be markedly lesser than those of TSP. In addition, FTP may require access to the annotated health care data that are several magnitudes less than the training data used in TSP. Unlike TSP, fine-tuning is typically accompanied by high GPU consumption for a short burst of time and ideal for availing pay-as-you-go models offered by CSPs in contrast to on-premise systems.

Table 4. Comparable pricing of graphics processing unit [GPU] clusters across cloud service providers (CSPs; Amazon Web Services [AWS], Azure, Google Cloud Platform [GCP], and Oracle Cloud Infrastructure [OCI]) for the fine-tuned pathway. Representative data were retrieved in June 2024.

CSP	Instance type	CPU[a] cores	Memory (GiB)	GPUs	Per GPU memory	Total GPU memory	On demand (US $)	1 year (US $)	3 year (US $)
AWS	p3.16xlarge	64	488	8	16	128	24.48	15.91	8.39
AWS	p3dn.24xlarge	96	768	8	32	256	31.22	18.39	9.64
Azure	NC24rs_v3	24	448	4	16	64	12.24	8.98	6.52
GCP	n1-highmem-64	64	416	8	16	128	23.63	14.88	10.63
OCI	BM.GPU3.8	52	768	8	16	128	23.60	—[b]	—

[a]CPU: central processing unit.

[b]Not applicable.

OBP Overview

OBP includes commercial off-the-shelf LLMs typically accessed by end users through Representational State Transfer APIs (SaaS) with minimal or no local customization.

Benefits

Unlike TSP and FTP, OBP is usually enterprise-ready. OBPs do not require digital readiness in access to integrated datasets and data warehouses. LLMs are accessed through a web interface; hence, no budgeting needs to be allocated for the storage and computing infrastructure needs of LLMs or the workforce to support LLM implementation and maintenance, as in TSP and FTP. Because OBP is provided by multiple vendors, there is an option to choose the best-performing LLM for a given price. Transitioning between OBP services can be done with ease because there is no explicit sharing of sensitive health care data or customized LLM architecture. User training needs in OBP are minimal compared with TSP and FTP. Managed services provided by CSPs can assist OBP implementation with minimal workforce needs on-premises. This includes CSP-managed API end points that offer access to 1 or more LLMs, whose cost is proportional to the number of input and output tokens. Services provided by the 4 major CSPs in this regard include (AWS: Claude, Azure: OpenAI Service, GCP: Gemini, and OCI: Cohere). Managed services by CSPs also support OBP LLM deployment on GPUs with minimal ease and a per-hour charge (AWS: Amazon SageMaker Canvas, Azure: AI Studio, and GCP: Vertex AI).

Risks

Unlike TSP and FTP, OBP is completely dependent on the SaaS or infrastructure as a service option provided by the vendor, with no control over training and fine-tuning and limited transparency on potential biases and the details of the LLM model and the training data. Prompt engineering risks of OBP may be larger than those of FTP and TSP because the OBP may use models that are not domain-specific, and the absence of domain-specific knowledge may result in OBP being sensitive to prompting. Output variability can be relatively higher for OBP, leading to inconsistent results and challenging deploying these models in clinical workflows. Copyright protection may limit the extent to which OBP can reveal the implementation details, and OBP services can be black boxes. Therefore, OBP services are expected to provide generic insights that may not necessarily accommodate the digital footprints and characteristics of the population served by a health care organization. This in turn may diminish the value and utility of OBP in addressing the needs that are specific to the organization. It might not be feasible to use PHI data across OBP services that are not HIPAA compliant. Posing questions to OBP services may unintentionally compromise patient privacy, especially when the sample size of the cohort being queried is small (eg, rare disease). OBP services may pose challenges in interfacing other applications, dashboards, and workflows without the explicit involvement of the vendor. Cost is usually incurred per inference and can aggregate over time with increasing dependency and number of users. These in turn may demand user authentication, quota allocation, and auditing. The reliability of the OBP service is dependent on the vendor and not necessarily guaranteed, with the potential for pricing options to increase with increasing adoption.

Economics of OBP

The cost of digital readiness for OBP is expected to be minimal as OBP services are provided as ready-to-use solutions without any need for training or fine-tuning as in TSP and FTP. Ideally, OBP would not require a dedicated team to support other than training materials for end users. However, hosting open-source LLM models, such as those from Hugging Face Hub, may demand infrastructure costs on-premises or in the cloud. Alternatively, LLM models in the OBP can be availed through APIs whose charges vary based on the input and output tokens and use patterns. Use patterns need to be carefully monitored as they could gradually increase to millions of tokens per month, significantly impacting the cost of OBP long-term. Pricing estimates of LLM APIs across major CSPs are shown in Table 5 [129-132]. Due to the per-unit pricing of certain CSPs, an average of 4 characters per token is assumed in generating these estimates.

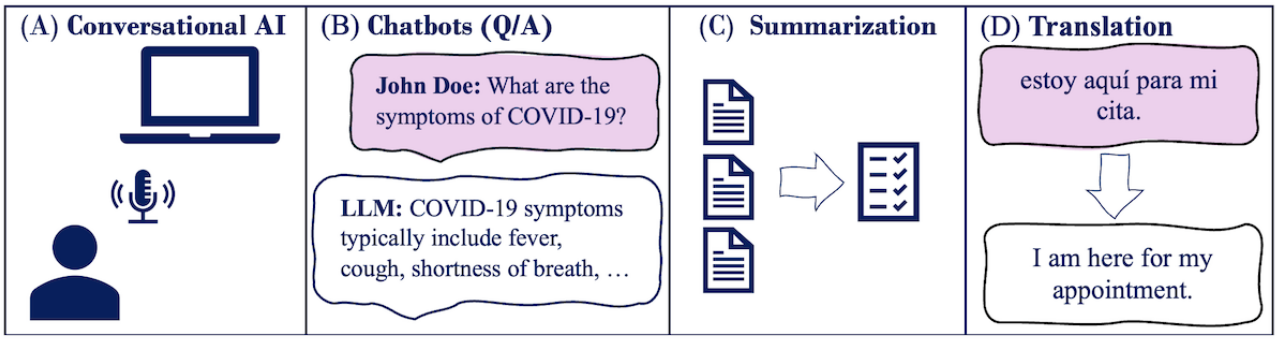
Table 5. Comparable pricing of large language models through application programming interface across cloud service providers (CSPs; Amazon Web Services [AWS], Azure, Google Cloud Platform [GCP], and Oracle Cloud Infrastructure [OCI]). Representative data were retrieved in June 2024.

CSP	Model	Input unit cost (US $)	Input parameters	Output unit cost (US $)	Output parameters	10,000 Input tokens (US $)	500 Output tokens (US $)	Total cost (US $)
AWS	Claude 3 Sonnet	0.003	Per 1000 tokens	0.015	Per 1000 tokens	0.030	0.008	0.038
Azure	GPT-4o global deployment	0.005	Per 1000 tokens	0.015	Per 1000 tokens	0.050	0.008	0.058
GCP	Gemini 1.5 pro	0.00125	Per 1000 characters (approximately 250 tokens)	0.00375	Per 1000 characters (approximately 250 tokens)	0.050	0.015	0.058
OCI	Cohere large	0.022	Per 10,000 transactions (approximately 2500 tokens)	0.022	Per 10,000 transactions (approximately 2500 tokens)	0.088	0.004	0.092

LLM Applications in Health Care

LLMs have the potential to offer health care providers new mechanisms for optimization and automation of documentation, clinical review, and direct patient communication. Its use is expected to reduce provider time in using systems, such as EHR, and improve clinical documentation while minimizing repetitive workflows. LLMs can also benefit patients by serving as an approachable tool to manage their health information assist in health literacy, appointment scheduling, and ambulatory encounters. Typical LLM applications in health care are shown in Figure 3. These applications broadly rely on the principles underlying dialogue systems [133]. They facilitate text-based and voice-based interactions between LLMs and the end user.

Figure 3. Typical health care applications of (A) large language models (LLMs) and conversational artificial intelligence (AI), (B) chatbots, (C) summarization, and (D) translation. Q/A: question and answer.



Conversational AI in Clinical Practice

Conversational AI (Figure 3A) focuses on the application of AI-based approaches, such as LLMs, for developing dialogue systems [133]. It is an omnibus term that includes chatbots and virtual assistants as well as text and voice-based interaction with the end user. The ability of these systems to understand user intent and generate relevant and accurate responses has continued to improve over time [134,135]. Conversational AI-enabled systems can significantly enhance information seeking and retrieval capabilities of health care information by patient, provider, and payer. Specifically in clinical practice, conversational AI integrated within EHRs [136] can facilitate increased patient-provider engagement and reduced documentation time, minimizing physician burnout. Moreover, these ambient listening, generative AI solutions have the opportunity to dramatically reduce documentation and cognitive burden, further leading to an improved clinician EHR experience. Depending on the services needed, the integration can be pursued through TSP for a bespoke solution, FTP for modifying existing models, and the OBP pathway with minimal customization. In addition to the risks and benefits of these pathways, their deployment in EHR workflows may demand expertise in user interface design for functional integration. The focus should be on ensuring that AI-supported interactions enhance rather than impede the patient-provider relationship. Operationally, such systems streamline data management processes, leading to increased efficiency and potential financial savings due to reduced time spent on repetitive tasks. Depending on digital and analytics maturity, there could be challenges related to integration and deployment challenges and enhanced adoption. In enhancing the conversational AI interfaces within EHR systems, a more detailed exploration of specific use cases and scenarios would reveal their impact on clinical workflow, such as patient portal communications or clinical decision support. Long-term adoption studies, focusing on the acceptance and resistance among health care professionals, can provide insights into the practical aspects of implementing these technologies. Moreover, a deeper discussion on how these AI interfaces evolve through ML and adaptation to user behaviors and preferences can illustrate the potential for increasingly efficient and user-friendly medical systems. The development of guardrails on AI-moderated operations and data security, particularly in the context of sensitive EHR data, is imperative to address concerns around privacy and compliance with health care regulations.

Conversational AI economics depends on the quantity, complexity, and scale of integration. For instance, conversational AI tools for EHR demand seamless and secured integration of the tool to the EHR either directly or through third-party interfaces, training the users for enhanced adoption, and customization of the tools through fine-tuning. Training conversational AI tools is critical to maintaining high accuracy on clinical tasks and dialogue understanding impacting its adoption. Successful deployment will demand active involvement of multiple teams, including analytics, information technology, the governance team, compliance, and security. Scalability considerations are crucial if deployed across multiple facilities, necessitating robust infrastructure investments. In addition, ongoing optimization costs are incurred as the AI must continuously update to align with new medical guidelines and treatment protocols, ensuring reliability and accuracy in patient data management.

Chatbots in Health Care

A chatbot is a conversational AI system that enables communication between computers and humans using natural language in completing specific tasks, such as question answering (Figure 3B). LLM-based chatbots are transforming the way patients, providers, and payers interact within health care settings [137]. They have the potential to improve health care outcomes by providing timely health information to patients, assisting in patient education and interventions, improving operational productivity by assisting decision-making processes, and reducing administrative tasks and overhead costs [138]. Chatbots can also assist in individualized services, including symptom assessment through virtual consults, appointment scheduling, and improving health literacy by making health care information accessible in a preferred language [139-141]. In developing chatbots, health care systems may opt for TSP for a completely customized solution, FTP for adapting existing models, or OBP for ready-to-use prebuilt templates. In addition to the risks and benefits of these pathways discussed earlier, chatbot implementations demand a good understanding of patient engagement strategies and health care communication norms. Successful deployment should ensure privacy and security of patients, adhere to regulations, and establish necessary guardrails [110]. Their design should also prioritize empathy and cultural sensitivity to ensure inclusive and respectful interactions with diverse patient populations in addition to minimizing bias and assumptions in conversations [137]. These, in turn, are expected to build trust and harmony with users, leading to enhanced adoption [142]. Successful chatbot design should also accommodate nuances of human interactions (eg, patient emotions and expectations). These in turn may demand well-articulated prompt engineering, fine-tuning, and optimization of the parameters in the underlying model [143]. Incorporating feedback from patients and health care providers will offer a deeper understanding of user experience and areas for improvement. Chatbots can also be integrated with other digital health technologies and workflows, such as telemedicine and electronic medical record. In addition, it can assist in personalization of services, tailoring interactions based on individual patient profiles and needs, significantly impacting patient engagement. Integrating chatbots with predictive analytics can also assist in assessing the usefulness of these tools by incorporating feedback.

The economics of LLM-powered chatbots in health care can be impacted by integration, compliance needs, feature complexity, and deployment scale. OBP is generally the least expensive, potentially costing a few thousand to tens of thousands of dollars annually, based on a subscription model. However, these figures can easily grow with the frequency of use and number of users. Fine-tuning existing models can cost tens to hundreds of thousands of dollars, depending on the extent of customization and licensing fees. Developing a highly customized chatbot from scratch is the most expensive option, with expenses running into hundreds of thousands to millions of dollars. Additional costs arise from ongoing maintenance, server costs, updates, and compliance with regulations such as HIPAA or GDPR, which necessitate robust security measures. The complexity of features, such as multilingual support advanced diagnostic capabilities, and the scale of deployment can also significantly impact the overall cost.

Summarization in Health Care

LLMs excel in summarization (Figure 3C), tasks critical for managing extensive medical documentation and improving clinical workflows [144]. These AI-driven systems can distill complex medical records into concise summaries for improved decision-making and patient management. Summarization also reduces task load while improving documentation quality, operational efficiency in processing documents, and financial savings by reducing the time and resources spent on administrative tasks. Summarization can also succinctly capture details of patient-provider conversations in team settings comprising a large number of clinicians, resulting in enhanced care continuity, coordination, and overall quality. While similar in integration complexity to conversational AI systems, the implementation of LLMs for summarization specifically requires tuning the models to capture critical medical insights accurately. Summarization leverages advanced natural language understanding capabilities, a step beyond general chatbot applications to ensure that summaries are not only succinct but also clinically relevant. These can be developed through TSP for high specificity, FTP for a balanced approach, or OBP for broader applications. The summarization process requires an augmented framework comprising a group of experts in the domain of AI, clinical knowledge and medical terminology, and data-processing infrastructure for critical validation. Successful implementation of LLM summarization is expected to ensure integrity of medical information, preventing any loss of critical details in the summarization process, minimizing the risk of misinterpretation of condensed information, and seamless integration of these tools into existing clinical workflows.

The economics of summarization while overlapping with those of broader AI integrations, such as conversational AI tools, are particularly influenced by the need for high-quality training data and the development of interfaces that clinicians can use effectively within existing digital health frameworks and workflows. The cost and investment may increase for summarization tools that meet high standards of accuracy and reliability in medical contexts while minimizing risks. This

includes rigorous testing and validation in real-world settings to adhere to the data handling and privacy regulations characteristic of the health care institution and industry. In addition, the ongoing maintenance to update the models with new medical information and guidelines further adds to the overall expenditure. These factors combined make the economics of summarization technologies substantial yet crucial for enhancing efficiency and decision-making in health care environments.

Machine Translation in Health Care

Machine translation (Figure 3D) serves diverse linguistic communities by the translation of text or speech from one language to another. Its role is especially helpful in overcoming language barriers in medical communication and documentation, especially across health care organizations that serve non–English-speaking communities. Machine translation can improve patient–health care provider communication, patient understanding of instructions, and discharge summaries, as well as operational benefits by facilitating multilingual documentation and financial advantages by potentially reducing the workload for human interpreters in low-resource settings [145]. It also has the potential to assist in the transmission of critical medical information in a culturally sensitive and empathetic manner, with the potential to minimize adverse events and impact health care outcomes favorable, especially across non–English-speaking communities. Machine translation can be developed through TSP for precise, context-specific translations using data from communities served by a specific health care organization, FTP for adapting existing models to medical language nuances, or OBP for immediate implementation with off-the-shelf translation tools. In addition to the applications mentioned here, machine translation requires collaboration with linguists and cultural sensitivity advisers to ensure translations are accurate and culturally appropriate. Ethical and regulatory considerations revolve around the accuracy and cultural appropriateness of translations. There is a strong emphasis on avoiding miscommunication in critical medical contexts while respecting linguistic diversity. Challenges include the risk of mistranslation, cultural insensitivity, and loss of nuanced medical context. The impact on target populations is usually diverse. It enables payers to offer multilingual services efficiently, aids providers in delivering equitable care to non–English-speaking patients and empowers patients by providing access to medical information in their preferred native languages. Here, underlining cultural competency alongside language translation is crucial. This includes not only translating text but also understanding and conveying cultural nuances, which is critical in medical contexts with potential favorable impact outcomes. Establishing specific metrics or standards to gauge the accuracy and reliability of translations can provide a benchmark for evaluating these tools. Discussing the legal implications and responsibilities in cases of mistranslation or miscommunication is also vital to understanding the potential liabilities involved. Extensive testing may be required before deployment in high-stakes areas, such as emergency medicine, where quick and accurate translation is vital.

The economics of machine translation in health care settings largely overlap with those of conversational AI technologies, as discussed previously. Costs can vary based on the development pathway chosen: OBP solutions may offer a lower upfront cost with general translation tools available for immediate use, while the FTP and the TSP require more substantial investments to adapt or develop models that handle medical language nuances and specific community dialects. However, OBP solutions could have severe limitations in high-stakes applications compared with FTP and TSP. Costs usually include customization, system integration with existing health care IT infrastructures such as EHRs, and ongoing expenses for maintenance and updates. Additional significant expenses are incurred in ensuring accuracy and cultural appropriateness, which involves collaboration with linguists and cultural experts. This collaboration is essential to mitigate risks of mistranslation and to comply with health care communication standards. Therefore, while the base technology may be like those used in conversational AI, the specificity and critical nature of medical translations can lead to higher costs, particularly when ensuring the system meets the stringent requirements of medical accuracy and regulatory compliance.

Conclusions

LLMs have the potential to meaningfully impact health care delivery and health outcomes. However, LLM implementations are impacted by the needs and affordability. This perspective provided 3 LLM implementation pathways (TSP, FTP, and OBP) and a road map for onboarding, enhanced democratization, and equitable adoption by the health care ecosystem. The economics, risks, and benefits of these pathways were also presented across 4 major CSPs (AWS, GCP, Azure, and OCI) to assist in choosing the best pathway for an organization. As LLMs continue to evolve [146], additional onboarding pathways are expected to join the repertoire.

The critical role of cloud-computing frameworks to support onboarding efforts from scalability, privacy, workforce, and economic standpoints was discussed. Pay-as-you-go models offered by CSPs alleviate the need for significant upfront investments while providing the ability to experiment with different pathways with the flexibility to scale and transition between pathways based on the usefulness of these tools in impacting outcomes and KPIs and use with evolving needs of health care organizations. Managed services provided by CSPs can assist in optimal management of resources and infrastructure while streamlining workflows and minimizing the need for considerable expertise across a variety of areas. These aspects are especially suited for organizations that do not have sufficient resources and upfront investment for LLM implementation. CSPs also provide privacy and security features, including confidential computing and TEEs for safeguarding sensitive health care data and maintaining regulatory compliance. The size of health care datasets used to train LLMs is often small compared with datasets used to train general-purpose models. CSPs can facilitate FL in conjunction with TEEs and deidentification in overcoming sample size constraints by enabling collaborative training strategies across health care organizations without explicit data sharing and ensuring privacy. FL approaches can result in models with enhanced

generalization ability in contrast to those trained using data from a single health care organization. Because LLMs may have the potential to memorize sensitive information from training data, hindering the sharing of checkpointed models due to privacy concerns. Techniques such as differential privacy and secure multiparty computation in CSPs can mitigate such risks. CSPs also provide access to specialized hardware accelerators, presenting an opportunity to improve the efficiency and cost-effectiveness of LLM training and inference. However, it is important to consider compatibility and performance trade-offs when integrating these accelerators into existing workflows.

As LLMs are used across health care, it is crucial to consider potential challenges and unintended consequences. Choice of an LLM implementation pathway can be significantly impacted by digital readiness, infrastructure, workforce, and ethical and regulatory landscape. Overreliance on LLMs could also diminish the critical thinking skills of health care professionals. As with all AI and ML tools, optimization is an essential ingredient of LLMs. Therefore, it is essential that LLMs be used within an augmented framework to support human decision-making rather than serving as a replacement. Establishing guardrails, ethical guidelines, and training programs for ensuring the responsible use of LLMs in clinical settings is important. Providing training and support for health care professionals and actively engaging them in the LLM implementation is critical for their successful deployment and long-term adoption. Workforce is a critical ingredient for LLM implementation, deployment, and maintenance. Prioritizing the inclusion of belonging, diversity, equity, and inclusion leaders as a part of LLM development is crucial to ensuring implementation that is inclusive and representative of diverse populations. Engaging policy makers and educating them about LLM limitations and adoption in health care is critical for realistic expectations from these tools and developing the necessary regulatory frameworks. Incentives could be introduced to encourage LLM adoption across health care organizations, and KPIs should be identified to assess its impact on health care outcomes. Short-term incentives can facilitate initial adoption of a particular onboarding pathway, while long-term incentives may assist in shifting across these pathways.

Identifying onboarding pathways for LLM implementation leveraging cloud computing along with metrics to demonstrate value while incorporating the necessary regulations and guardrails of responsible AI is critical for its equitable distribution and enhanced adoption in the health care ecosystem. Widespread adoption is also expected to facilitate feedback from diverse communities served by the health care ecosystem, improving patient outcomes and operational efficiency and addressing the unique challenges and considerations in health care.

Authors' Contributions

RN conceived the presented idea. All authors contributed to the content and development of the manuscript. All authors contributed to the editing, reviewing, and approval of the final manuscript.

Conflicts of Interest

ES serves on the editorial board of JMIR Publications.

References

1. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. Sep 15, 2022;28(9):1773-1784. [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
2. Topol EJ. As artificial intelligence goes multimodal, medical applications multiply. Science. Sep 15, 2023;381(6663):adk6139. [doi: 10.1126/science.adk6139] [Medline: 37708283]
3. Kong HJ. Managing unstructured big data in healthcare system. Healthc Inform Res. Jan 2019;25(1):1-2. [FREE Full text] [doi: 10.4258/hir.2019.25.1.1] [Medline: 30788175]
4. Sedlakova J, Daniore P, Horn Wintsch A, Wolf M, Stanikic M, Haag C, et al. Challenges and best practices for digital unstructured data enrichment in health research: a systematic narrative review. PLOS Digit Health. Oct 11, 2023;2(10):e0000347. [FREE Full text] [doi: 10.1371/journal.pdig.0000347] [Medline: 37819910]
5. Zhou C, Li Q, Li C, Yu J, Liu Y, Wang G, et al. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. arXiv. Preprint posted online on February 18, 2023. [FREE Full text]
6. Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. NPJ Digit Med. Dec 17, 2019;2(1):130. [FREE Full text] [doi: 10.1038/s41746-019-0208-8] [Medline: 31872069]
7. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform. Sep 2017;73:14-29. [FREE Full text] [doi: 10.1016/j.jbi.2017.07.012] [Medline: 28729030]
8. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc. Apr 01, 2019;26(4):364-379. [FREE Full text] [doi: 10.1093/jamia/ocy173] [Medline: 30726935]
9. Lee RY, Kross EK, Torrence J, Li KS, Sibley J, Cohen T, et al. Assessment of natural language processing of electronic health records to measure goals-of-care discussions as a clinical trial outcome. JAMA Netw Open. Mar 01, 2023;6(3):e231204. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.1204] [Medline: 36862411]

10. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc. Mar 01, 2020;27(3):457-470. [FREE Full text] [doi: 10.1093/jamia/ocz200] [Medline: 31794016]
11. Oleynik M, Kugic A, Kasáč Z, Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. J Am Med Inform Assoc. Nov 01, 2019;26(11):1247-1254. [FREE Full text] [doi: 10.1093/jamia/ocz149] [Medline: 31512729]
12. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. Jul 07, 2023;619(7969):357-362. [FREE Full text] [doi: 10.1038/s41586-023-06160-y] [Medline: 37286606]
13. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open. Oct 02, 2023;6(10):e2336997. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36997] [Medline: 37812419]
14. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv. Preprint posted online on August 16, 2021. [FREE Full text]
15. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. Aug 12, 2023;620(7972):172-180. [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
16. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv. Preprint posted online on May 16, 2023. [FREE Full text]
17. Armbrust M, Ghodsi A, Xin R, Zaharia M. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In: Proceedings of the 11th Annual Conference on Innovative Data Systems Research. 2021. Presented at: CIDR '21; January 11-15, 2021; Online. URL: https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf
18. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA. Sep 05, 2023;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
19. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. London, UK. Pearson; 2010.
20. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Cham, Switzerland. Springer International Publishing; 2009.
21. Aggarwal CC. Data Mining: The Textbook. Cham, Switzerland. Springer International Publishing; 2015.
22. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, MA. The MIT Press; 2016.
23. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. May 28, 2015;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
24. Sapoval N, Aghazadeh A, Nute MG, Antunes DA, Balaji A, Baraniuk R, et al. Current progress and open challenges for applying deep learning across the biosciences. Nat Commun. Apr 01, 2022;13(1):1728. [FREE Full text] [doi: 10.1038/s41467-022-29268-7] [Medline: 35365602]
25. Chollet F. Deep Learning with Python. Shelter Island, NY. Manning Publications; Nov 2017.
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. Preprint posted online on June 12, 2017. [FREE Full text]
27. Ananthaswamy A. In AI, is bigger always better? Nature. Mar 08, 2023;615(7951):202-205. [doi: 10.1038/d41586-023-00641-w] [Medline: 36890378]
28. Friedman CP. A "fundamental theorem" of biomedical informatics. J Am Med Inform Assoc. Mar 2009;16(2):169-170. [doi: 10.1197/jamia.m3092]
29. Allen-Zhu Z, Li Y, Song Z. A convergence theory for deep learning via over-parameterization. arXiv. Preprint posted online on November 9, 2018. [FREE Full text]
30. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv. Preprint posted online on March 4, 2022. [FREE Full text]
31. Adoption model for analytics maturity (AMAM). Healthcare Information and Management Systems Society. URL: https://www.himss.org/maturity-models/amam/ [accessed 2024-09-30]
32. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online on February 27, 2023. [FREE Full text]
33. Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, Passos A, et al. PaLM 2 technical report. arXiv. Preprint posted online on May 17, 2023. [doi: 10.48550/arXiv.2305.10403]
34. Health information privacy. U.S. Department of Health and Human Services. URL: https://www.hhs.gov/hipaa/for-professionals/security/index.html [accessed 2024-10-14]
35. General Data Protection Regulation (GDPR) compliance guidelines. General Data Protection Regulation. URL: https://gdpr.eu/ [accessed 2024-10-14]
36. Radhakrishnan L, Schenk G, Muenzen K, Oskotsky B, Ashouri Choshali H, Plunkett T, et al. A certified de-identification system for all clinical text documents for information extraction at scale. JAMIA Open. Oct 2023;6(3):ooad045. [FREE Full text] [doi: 10.1093/jamiaopen/ooad045] [Medline: 37416449]

37. Mulligan DP, Petri G, Spinale N, Stockwell G, Vincent HJ. Confidential computing—a brave new world. In: Proceedings of the International Symposium on Secure and Private Execution Environment Design. 2021. Presented at: SEED 2021; September 20-21, 2021; Washington, DC. [doi: 10.1109/seed51797.2021.00025]
38. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc. 2010;17(2):169-177. [FREE Full text] [doi: 10.1136/jamia.2009.000026] [Medline: 20190059]
39. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One. Dec 2, 2011;6(12):e28071. [FREE Full text] [doi: 10.1371/journal.pone.0028071] [Medline: 22164229]
40. de Kok JW, de la Hoz MÁ, de Jong Y, Brokke V, Elbers PW, Thoral P, Collaborator Group, et al. A guide to sharing open healthcare data under the General Data Protection Regulation. Sci Data. Jun 24, 2023;10(1):404. [FREE Full text] [doi: 10.1038/s41597-023-02256-2] [Medline: 37355751]
41. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med. Dec 26, 2022;5(1):194. [FREE Full text] [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
42. Schaeffer R, Miranda B, Koyejo S. Are emergent abilities of large language models a mirage? arXiv. Preprint posted online on April 28, 2023. [FREE Full text]
43. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv. Preprint posted online on March 31, 2023. [FREE Full text]
44. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. arXiv. Preprint posted online on January 20, 2020. [doi: 10.48550/arXiv.2001.08361]
45. Chen C, Feng X, Zhou J, Yin J, Zheng X. Federated large language model: a position paper. arXiv. Preprint posted online on July 18, 2023. [doi: 10.48550/arXiv.2307.08925]
46. Ye R, Wang W, Chai J, Li D, Li Z, Xu Y, et al. OpenFedLLM: training large language models on decentralized private data via federated learning. arXiv. Preprint posted online on February 10, 2024. [doi: 10.48550/arXiv.2402.06954]
47. Sadilek A, Liu L, Nguyen D, Kamruzzaman M, Serghiou S, Rader B, et al. Privacy-first health research with federated learning. NPJ Digit Med. Sep 07, 2021;4(1):132. [FREE Full text] [doi: 10.1038/s41746-021-00489-2] [Medline: 34493770]
48. McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. arXiv. Preprint posted online on February 17, 2016. [doi: 10.48550/arXiv.1602.05629]
49. Bak M, Madai VI, Celi LA, Kaissis GA, Cornet R, Maris M, et al. Federated learning is not a cure-all for data ethics. Nat Mach Intell. Mar 18, 2024;6:370-372. [doi: 10.1038/s42256-024-00813-x]
50. Inmon WH. Building the Data Warehouse. Hoboken, NJ. John Wiley & Sons; 2005.
51. Kimball R, Ross M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Hoboken, NJ. John Wiley & Sons; 2002.
52. Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, et al. The shared health research information network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. Sep 2009;16(5):624-630. [doi: 10.1197/jamia.m3191]
53. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124-130. [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
54. Saitwal H, Qing D, Jones S, Bernstam EV, Chute CG, Johnson TR. Cross-terminology mapping challenges: a demonstration using medication terminological systems. J Biomed Inform. Aug 2012;45(4):613-625. [FREE Full text] [doi: 10.1016/j.jbi.2012.06.005] [Medline: 22750536]
55. Matcho A, Ryan P, Fife D, Reich C. Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model. Drug Saf. Nov 2014;37(11):945-959. [FREE Full text] [doi: 10.1007/s40264-014-0214-3] [Medline: 25187016]
56. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med. Aug 18, 2022;5(1):119. [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
57. Chang BL, Bakken S, Brown SS, Houston TK, Kreps GL, Kukafka R, et al. Bridging the digital divide: reaching vulnerable populations. J Am Med Inform Assoc. 2004;11(6):448-457. [FREE Full text] [doi: 10.1197/jamia.M1535] [Medline: 15299002]
58. Wang M, Gago CM, Rodriguez K. Digital redlining-the invisible structural determinant of health. JAMA. Apr 16, 2024;331(15):1267-1268. [doi: 10.1001/jama.2024.1628] [Medline: 38497952]
59. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf. Mar 2019;28(3):231-237. [FREE Full text] [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
60. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. Oct 25, 2019;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
61. Blueprint for trustworthy AI: implementation guidance and assurance for healthcare coalition for health AI. Coalition for Health AI. Apr 4, 2023. URL: https://www.coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf [accessed 2024-10-14]
62. Charpignon ML, Byers J, Cabral S, Celi LA, Fernandes C, Gallifant J, et al. Critical bias in critical care devices. Crit Care Clin. Oct 2023;39(4):795-813. [doi: 10.1016/j.ccc.2023.02.005] [Medline: 37704341]

63. Teotia K, Jia Y, Link Woite N, Celi LA, Matos J, Struja T. Variation in monitoring: glucose measurement in the ICU as a case study to preempt spurious correlations. J Biomed Inform. May 2024;153:104643. [doi: 10.1016/j.jbi.2024.104643] [Medline: 38621640]
64. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. May 24, 2017;60(6):84-90. [doi: 10.1145/3065386]
65. Dean J, Corrado GS, Monga R, Chen K, Devin M, Le QV, et al. Large scale distributed deep networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. 2012. Presented at: NIPS'12; December 3-6, 2012; Lake Tahoe, Nevada.
66. Huang Y, Cheng Y, Bapna A, Firat O, Chen MX, Chen D, et al. GPipe: efficient training of giant neural networks using pipeline parallelism. arXiv. Preprint posted online on November 16, 2018. [doi: 10.48550/arXiv.1811.06965]
67. Shoeybi M, Patwary M, Puri R, LeGresley P, Casper J, Catanzaro B. Megatron-lm: training multi-billion parameter language models using model parallelism. arXiv. Preprint posted online on September 17, 2019. [doi: 10.48550/arXiv.1909.08053]
68. Faiz A, Kaneda S, Wang R, Osi R, Sharma P, Chen F, et al. LLMCarbon: modeling the end-to-end carbon footprint of large language models. arXiv. Preprint posted online on September 25, 2023. [doi: 10.48550/arXiv.2309.14393]
69. Bui ND, Le H, Wang Y, Li J, Gotmare AD, Hoi SC. CodeTF: one-stop transformer library for state-of-the-art code LLM. arXiv. Preprint posted online on May 31, 2023. [doi: 10.48550/arXiv.2306.00029]
70. Candel A, McKinney J, Singer P, Pfeiffer P, Jeblick M, Lee CM, et al. H2O open ecosystem for state-of-the-art large language models. arXiv. Preprint posted online on October 17, 2023. [FREE Full text] [doi: 10.18653/v1/2023.emnlp-demo.6]
71. Cresswell K, Domínguez Hernández A, Williams R, Sheikh A. Key challenges and opportunities for cloud technology in health care: semistructured interview study. JMIR Hum Factors. Jan 06, 2022;9(1):e31246. [FREE Full text] [doi: 10.2196/31246] [Medline: 34989688]
72. Peccerillo B, Mannino M, Mondelli A, Bartolini S. A survey on hardware accelerators: taxonomy, trends, challenges, and perspectives. J Syst Archit. Aug 2022;129:102561. [doi: 10.1016/j.sysarc.2022.102561]
73. Chen H, Zhang J, Du Y, Xiang S, Yue Z, Zhang N, et al. A comprehensive evaluation of FPGA-based spatial acceleration of LLMs. In: Proceedings of the 2024 ACM/SIGDA International Symposium on Field Programmable Gate Arrays. 2024. Presented at: FPGA '24; March 3-5, 2024; Monterey, CA. [doi: 10.1145/3626202.3637600]
74. Mohaidat T, Khalil K. A survey on neural network hardware accelerators. IEEE Trans Artif Intell. Aug 2024;5(8):3801-3822. [doi: 10.1109/tai.2024.3377147]
75. Machupalli R, Hossain M, Mandal M. Review of ASIC accelerators for deep neural network. Microprocess Microsyst. Mar 2022;89:104441. [doi: 10.1016/j.micpro.2022.104441]
76. Batra G, Jacobson Z, Madhav S, Queirolo A, Santhanam N. Artificial-intelligence hardware: new opportunities for semiconductor companies. McKinsey and Company. Dec 2018. URL: https://tinyurl.com/22a7vapn [accessed 2024-10-14]
77. Mince F, Dinh D, Kgomo J, Thompson N, Hooker S. The grand illusion: the myth of software portability and implications for ML progress. arXiv. Preprint posted online on September 12, 2023. [doi: 10.1016/0950-5849(89)90153-5]
78. Bshara N. AWS Trainium: the journey for designing and optimization full Stack ML hardware. In: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems. 2024. Presented at: ASPLOS '24; April 27-May 1, 2024; La Jolla, CA. [doi: 10.1145/3620666.3655592]
79. Romero F, Li Q, Yadwadkar NJ, Kozyrakis C. INFaaS: automated model-less inference serving. In: Proceedings of the 2021 USENIX Annual Technical Conference. 2021. Presented at: USENIX ATC 2021; July 14-16, 2021; Virtual Event. URL: https://www.usenix.org/conference/atc21/presentation/romero
80. Jouppi NP, Young C, Patil N, Patterson D, Agrawal G, Bajwa R, et al. In-datacenter performance analysis of a tensor processing unit. In: Proceedings of the 44th Annual International Symposium on Computer Architecture. 2017. Presented at: ISCA '17; June 24-28, 2017; Toronto, ON. [doi: 10.1145/3079856]
81. Wang YE, Wei GY, Brooks D. Benchmarking TPU, GPU, and CPU platforms for deep learning. arXiv. Preprint posted online on July 24, 2019. [FREE Full text] [doi: 10.1090/mbk/121/79]
82. Azure Maia for the era of AI: from silicon to software to systems. Azure. Apr 3, 2024. URL: https://azure.microsoft.com/en-us/blog/azure-maia-for-the-era-of-ai-from-silicon-to-software-to-systems/ [accessed 2024-10-14]
83. Smith A, Chapman E, Patel C, Swaminathan R, Wuu J, Huang T. 11.1 AMD InstinctTM MI300 series modular chiplet package – HPC and AI accelerator for exa-class systems. In: Proceedings of the IEEE International Solid-State Circuits Conference. 2024. Presented at: ISSCC 2024; February 18-22, 2024; San Francisco, CA. [doi: 10.1109/isscc49657.2024.10454441]
84. AMD CDNA™ 3 Architecture. Advanced Micro Devices. URL: https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/white-papers/amd-cdna-3-white-paper.pdf [accessed 2024-10-14]
85. AMD delivers leadership portfolio of data center AI solutions with AMD instinct MI300 series. AMD. Dec 6, 2023. URL: https://ir.amd.com/news-events/press-releases/detail/1173/amd-delivers-leadership-portfolio-of-data-center-ai [accessed 2024-10-14]
86. Kreuzberger D, Kühl N, Hirschl S. Machine learning operations (MLOps): overview, definition, and architecture. IEEE Access. 2023;11:31866-31879. [doi: 10.1109/access.2023.3262138]

87. A technical analysis of confidential computing. The Confidential Computing Consortium. Nov 2022. URL: https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/CCC-A-Technical-Analysis-of-Confidential-Computing-v1.3_unlocked.pdf [accessed 2024-10-14]
88. Confidential computing: hardware-based trusted execution for applications and data. The Confidential Computing Consortium. Nov 2022. URL: https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/CCC_outreach_whitepaper_updated_November_2022.pdf [accessed 2024-09-30]
89. Brown D. Confidential computing: an AWS perspective. Amazon Web Services. Aug 24, 2021. URL: https://aws.amazon.com/blogs/security/confidential-computing-an-aws-perspective/ [accessed 2024-09-30]
90. Confidential computing. Google. URL: https://cloud.google.com/security/products/confidential-computing [accessed 2024-10-14]
91. Azure offerings. Microsoft. May 21, 2024. URL: https://learn.microsoft.com/en-us/azure/confidential-computing/overview-azure-products [accessed 2024-10-14]
92. Confidential computing. Oracle. URL: https://docs.oracle.com/en-us/iaas/Content/Compute/References/confidential_compute.htm [accessed 2024-10-14]
93. Bean JD, Raghu A, Rudzitis A. Protect sensitive data in use with AWS confidential computing. Amazon Web Services. 2023. URL: https://d1.awsstatic.com/events/Summits/reinvent2023/CMP307_Protect-sensitive-data-in-use-with-AWS-confidential-computing.pdf [accessed 2024-10-14]
94. Confidential VM attestation. Google Cloud. URL: https://cloud.google.com/confidential-computing/confidential-vm/docs/attestation [accessed 2024-10-14]
95. Choudhury O, He C, Avestimehr S, Bhargavi D, Ravipati VS, Aziz W, et al. Federated learning on AWS with FedML: health analytics without sharing sensitive data – part 1. Amazon Web Services. Jan 13, 2023. URL: https://aws.amazon.com/blogs/machine-learning/part-1-federated-learning-on-aws-with-fedml-health-analytics-without-sharing-sensitive-data/ [accessed 2024-10-14]
96. Cross-silo and cross-device federated learning on Google Cloud. Google Cloud. 2024. URL: https://cloud.google.com/architecture/cross-silo-cross-device-federated-learning-google-cloud [accessed 2024-10-14]
97. Federated learning in Azure ML. GitHub. URL: https://github.com/Azure-Samples/azure-ml-federated-learning [accessed 2024-10-14]
98. Peterson D, Kanani P, Marathe VJ. Private federated learning with domain adaptation. arXiv. Preprint posted online on December 13, 2019. [doi: 10.48550/arXiv.1912.06733]
99. Eichner H, Ramage D, Bonawitz K, Huba D, Santoro T, McLarnon B, et al. Confidential federated computations. arXiv. Preprint posted online on April 16, 2024. [doi: 10.48550/arXiv.2404.10764]
100. Ong JC, Chang SY, William W, Butte AJ, Shah NH, Chew LS, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health. Jun 2024;6(6):e428-e432. [doi: 10.1016/s2589-7500(24)00061-x]
101. Goldberg CB, Adams L, Blumenthal D, Brennan PF, Brown N, Butte AJ, et al. To do no harm - and the most good - with AI in health care. Nat Med. Mar 2024;30(3):623-627. [doi: 10.1038/s41591-024-02853-7] [Medline: 38388841]
102. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. Jul 06, 2023;6(1):120. [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
103. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House. Oct 30, 2023. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ [accessed 2024-10-14]
104. Section 1557 of the Patient Protection and Affordable Care Act. U.S. Department of Health and Human Services. URL: https://www.hhs.gov/civil-rights/for-individuals/section-1557/index.html [accessed 2024-10-14]
105. LaRose C, Edwards E. 1557 final rule protects against bias in health care algorithms. National Health Law Program. URL: https://healthlaw.org/1557-final-rule-protects-against-bias-in-health-care-algorithms/ [accessed 2024-10-14]
106. The Bletchley Declaration by Countries attending the AI Safety Summit, 1-2 November 2023. United Kingdom Government. Nov 1, 2023. URL: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 [accessed 2024-10-14]
107. Artificial intelligence in health care: implications for patient and workforce safety. Institute for Healthcare Improvement. URL: https://www.ihi.org/resources/publications/artificial-intelligence-health-care-implications-patient-and-workforce#downloads [accessed 2024-10-14]
108. Shaping Europe's digital future: AI Act. European Commission. URL: https://tinyurl.com/455s5hb8 [accessed 2024-10-02]
109. Regulatory considerations on artificial intelligence for health. World Health Organization. 2023. URL: https://iris.who.int/bitstream/handle/10665/373421/9789240078871-eng.pdf [accessed 2024-10-14]
110. Rebedea T, Dinu R, Sreedhar M, Parisien C, Cohen J. NeMo guardrails: a toolkit for controllable and safe LLM applications with programmable rails. arXiv. Preprint posted online on October 16, 2023. [FREE Full text] [doi: 10.18653/v1/2023.emnlp-demo.40]
111. Sun L, Huang Y, Wang H, Wu S, Zhang Q, Li Y, et al. Trustllm: trustworthiness in large language models. arXiv. Preprint posted online on January 10, 2024. [doi: 10.48550/arXiv.2401.05561]

112. Software as a Medical Device (SAMD): clinical evaluation. U.S. Food & Drug Administration. Dec 8, 2017. URL: https://www.fda.gov/media/100714/download [accessed 2024-10-14]
113. Artificial intelligence and medical products: how CBER, CDER, CDRH, and OCP are working together. U.S. Food and Drug Administration. URL: https://www.fda.gov/media/177030/download?attachment [accessed 2024-10-14]
114. Lehman E, Hernandez E, Mahajan D, Wulff J, Smith MJ, Ziegler Z, et al. Do we still need clinical language models? arXiv. Preprint posted online on February 16, 2023. [doi: 10.48550/arXiv.2302.08091]
115. Yu KH, Healey E, Leong TY, Kohane IS, Manrai AK. Medical artificial intelligence and human values. N Engl J Med. May 30, 2024;390(20):1895-1904. [doi: 10.1056/NEJMra2214183] [Medline: 38810186]
116. Hartmann V, Suri A, Bindschaedler V, Evans D, Tople S, West R. SoK: memorization in general-purpose large language models. arXiv. Preprint posted online on October 24, 2023. [doi: 10.48550/arXiv.2310.18362]
117. Amazon EC2 P4 instances. Amazon Web Services. URL: https://aws.amazon.com/ec2/instance-types/p4/ [accessed 2024-10-14]
118. Azure machine learning pricing. Microsoft. URL: https://azure.microsoft.com/en-us/pricing/details/machine-learning/ [accessed 2024-10-14]
119. NDasrA100_v4 sizes series. Microsoft. Aug 22, 2024. URL: https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/gpu-accelerated/ndasra100v4-series?tabs=sizebasic [accessed 2024-10-14]
120. VM instance pricing. Google Cloud. URL: https://cloud.google.com/compute/vm-instance-pricing [accessed 2024-10-14]
121. Compute pricing. Oracle. URL: https://www.oracle.com/cloud/compute/pricing/#compute-gpu [accessed 2024-10-14]
122. Compute and EC2 instance savings plans. Amazon Web Services. URL: https://aws.amazon.com/savingsplans/compute-pricing/ [accessed 2024-10-14]
123. Gurney M. meta-llama / llama. GitHub. URL: https://github.com/meta-llama/llama/blob/main/MODEL_CARD.md [accessed 2024-10-14]
124. Amazon EC2 spot instances pricing. Amazon Web Services. URL: https://aws.amazon.com/ec2/spot/pricing/ [accessed 2024-10-14]
125. Linux virtual machines pricing. Microsoft. URL: https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/#pricing [accessed 2024-10-14]
126. Spot VMs pricing. Google Cloud. URL: https://cloud.google.com/spot-vms/pricing [accessed 2024-10-17]
127. Cloud cost estimator. Oracle. URL: https://www.oracle.com/cloud/costestimator.html [accessed 2024-10-17]
128. Hsieh CJ, Si S, Yu FX, Dhillon IS. Automatic engineering of long prompts. arXiv. Preprint posted online on November 16, 2023. [doi: 10.48550/arXiv.2311.10117]
129. AWS Bedrock pricing. Amazon Web Services. URL: https://aws.amazon.com/bedrock/pricing/ [accessed 2024-10-14]
130. Azure OpenAI Service pricing. Microsoft. URL: https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/ [accessed 2024-10-14]
131. Vertex AI pricing. Google Cloud. URL: https://cloud.google.com/vertex-ai/pricing [accessed 2024-10-14]
132. Generative AI service pricing. Oracle. URL: https://www.oracle.com/artificial-intelligence/generative-ai/generative-ai-service/pricing/ [accessed 2024-10-14]
133. McTear M. Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. Williston, VT. Morgan & Claypool Publishers; 2020.
134. Goodman RS, Patrinely JR, Stone CAJ, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. Oct 02, 2023;6(10):e2336483. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
135. Sezgin E. Redefining virtual assistants in health care: the future with large language models. J Med Internet Res. Jan 19, 2024;26:e53225. [FREE Full text] [doi: 10.2196/53225] [Medline: 38241074]
136. Haberle T, Cleveland C, Snow GL, Barber C, Stookey N, Thornock C, et al. The impact of nuance DAX ambient listening AI documentation: a cohort study. J Am Med Inform Assoc. Apr 03, 2024;31(4):975-979. [doi: 10.1093/jamia/ocae022] [Medline: 38345343]
137. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. Jun 01, 2023;183(6):589-596. [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
138. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. JMIR Med Inform. Feb 10, 2022;10(2):e32875. [FREE Full text] [doi: 10.2196/32875] [Medline: 35142635]
139. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc. Sep 01, 2018;25(9):1248-1258. [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
140. Kurniawan MH, Handiyani H, Nuraini T, Hariyati RT, Sutrisno S. A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. Ann Med. Dec 2024;56(1):2302980. [FREE Full text] [doi: 10.1080/07853890.2024.2302980] [Medline: 38466897]

141. Wollny S, Schneider J, Di Mitri D, Weidlich J, Rittberger M, Drachsler H. Are we there yet? - a systematic literature review on chatbots in education. Front Artif Intell. Jul 15, 2021;4:654924. [FREE Full text] [doi: 10.3389/frai.2021.654924] [Medline: 34337392]
142. Fan X, Chao D, Zhang Z, Wang D, Li X, Tian F. Utilization of self-diagnosis health chatbots in real-world settings: case study. J Med Internet Res. Jan 06, 2021;23(1):e19928. [FREE Full text] [doi: 10.2196/19928] [Medline: 33404508]
143. Schulhoff S, Ilie M, Balepur N, Kahadze K, Liu A, Si C, et al. The prompt report: a systematic survey of prompting techniques. arXiv. Preprint posted online on June 6, 2024. [doi: 10.48550/arXiv.2406.06608]
144. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Clinical text summarization: adapting large language models can outperform human experts. Research Square. Preprint posted online on October 30, 2023. [FREE Full text] [doi: 10.21203/rs.3.rs-3483777/v1] [Medline: 37961377]
145. Hudelson P, Chappuis F. Using voice-to-voice machine translation to overcome language barriers in clinical communication: an exploratory study. J Gen Intern Med. May 2024;39(7):1095-1102. [FREE Full text] [doi: 10.1007/s11606-024-08641-w] [Medline: 38347346]
146. Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, et al. Large language models: a survey. arXiv. Preprint posted online on February 9, 2024. [doi: 10.48550/arXiv.2402.06196]

Abbreviations

AI: artificial intelligence
API: application programming interface
AWS: Amazon Web Services
CCE: confidential computing environment
CFL: confidential federated learning
CSP: cloud service provider
DL: deep learning
DNN: deep neural network
EDW: enterprise data warehouse
EHR: electronic health record
ETL: extract, transform, and load
FL: federated learning
FTP: fine-tuned pathway
GCP: Google Cloud Platform
GDPR: General Data Protection Regulation
GPU: graphics processing unit
HIPAA: Health Insurance Portability and Accountability Act
KMS: Key Management Service
KPI: key performance indicator
LGBTQ: lesbian, gay, bisexual, transgender, and queer
LLI: Lucian Leape Institute
LLM: large language model
ML: machine learning
NLP: natural language processing
OBP: out-of-the-box pathway
OCI: Oracle Cloud Infrastructure
PHI: protected health information
SaaS: software as a service
TEE: trusted execution environment
TPU: tensor processing unit
TSP: training from scratch pathway
VM: virtual machine

Edited by A Coristine; submitted 11.07.24; peer-reviewed by K Srinivasan, A Abreu, Q Shi; comments to author 20.08.24; revised version received 28.08.24; accepted 16.09.24; published 14.11.24

Please cite as:
Nagarajan R, Kondo M, Salas F, Sezgin E, Yao Y, Klotzman V, Godambe SA, Khan N, Limon A, Stephenson G, Taraman S, Walton N, Ehwerhemuepha L, Pandit J, Pandita D, Weiss M, Golden C, Gold A, Henderson J, Shippy A, Celi LA, Hogan WR, Oermann EK, Sanger T, Martel S
Economics and Equity of Large Language Models: Health Care Perspective
J Med Internet Res 2024;26:e64226
URL: https://www.jmir.org/2024/1/e64226
doi: 10.2196/64226
PMID: