A CTO's Guide to the Generative AI Technology Landscape

Published 18 September 2023 - ID G00793970 - 15 min read

By Analyst(s): Arun Chandrasekaran, Radu Miclaus, Eric Goodness

The viral adoption of ChatGPT has created immense interest in GenAI, but the technology landscape is confusing to discern and rapidly evolving. CTOs can use this research to assess this landscape of technologies and key vendors and create robust strategies to accelerate adoption.

Overview

Key Findings

- The generative AI (GenAI) technology landscape, which consists of infrastructure, models, engineering tools and applications, is rapidly evolving, creating a new ecosystem of vendors and products that technology innovation leaders find difficult to navigate.
- AI foundation models, which are at the center of the GenAI revolution, are rapidly evolving due to the advent of more closed source models, burgeoning innovation in open-source communities as well as popularity of model hubs, which are democratizing their access to developers.
- New engineering tools that help enterprises operationalize their GenAI use cases such as vector databases; API orchestration tools and AI trust; and risk and security management tools are all experiencing early interest.
- Both incumbent vendors and new startups are embedding GenAI models to build a differentiated SaaS portfolio aimed at enabling business-function-specific and verticalized use cases for enterprises. The intense competitive jostling in this space and unclear competitive moats between various vendors are exacerbating supplier selection for CIOs and CTOs.

Recommendations

CTOs responsible for the digital future and understanding GenAI should:

- Take an objective view of the adequate balance between accuracy, costs, security and privacy principles and time to value when deploying GenAI models to determine the appropriate model needed. Not all use cases require the largest or the most customized models.
- Adopt a platform approach to GenAI by investing in centralized AI engineering tools that can provide automation, governance and use-case enablement across a broad set of AI models and providers.
- Prioritize security and privacy practices, model training process, model IP ownership, workflow integration and track record of delivering innovation when selecting GenAI application vendors.
- Work with I&O leaders to assess current infrastructure fit and to explore reliable and efficient infrastructure utilization practices.

Strategic Planning Assumptions

By 2026, more than 80% of enterprises will have used GenAI APIs, models and/or deployed GenAI-enabled applications in production environments, which is a significant increase from less than 5% today.

By 2026, more than 70% of independent software vendors (ISVs) will have embedded GenAI capabilities in their enterprise applications, which is a major increase from less than 1% today.

By 2028, the number of independent vendors offering GenAI tools will reduce by 30% due to industry consolidation and startup failures due to poor product market fit.

Introduction

The GenAI landscape is rapidly evolving with accelerated product launches from incumbent vendors and the emergence of thousands of startups, which are vying for enterprise wallet share. The thousands of startups in this space will lead to inevitable consolidation in the future. Although most innovation in the IT space has come from startup companies, most enterprises are risk averse and reluctant to do business with them. Startups do carry unique risks, but many of these risks can be minimized. See The CIO's Guide to Working With Startups for more guidance.

Navigating the GenAI ecosystem is often overwhelming for enterprise CTOs due to a chaotic and fast-moving ecosystem of technologies and vendors. In this research, we explain this technology landscape by identifying the key technology segments and exploring why they are important.

The categories will evolve, new categories will be created in the future and the vendors mentioned here are only samples (i.e., there are no more than 10 vendors per category), meaning this is not to be construed as an exhaustive list. Large technology companies such as Microsoft (and OpenAI), Google, Amazon Web Services (AWS), IBM, Salesforce and NVIDIA are building end-to-end GenAI platforms to address many technology segments listed in this research. See A Comparison of Generative AI Platform Offerings for more details.

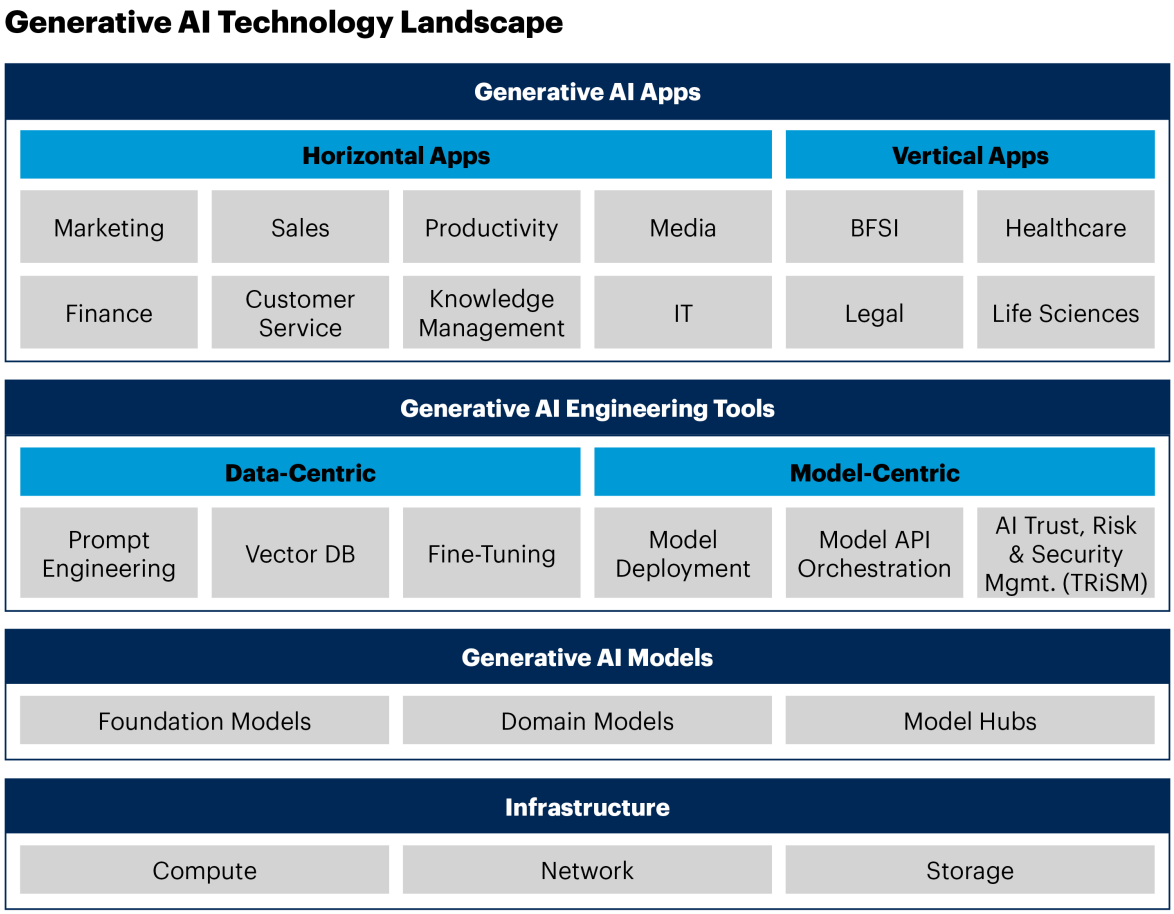
Analysis

The GenAI landscape consists of four critical layers — infrastructure, models, engineering tools and applications:

- AI foundation models continue to be an important driver of innovation and progress in this ecosystem.
- As organizations mature in their adoption, they are seeking engineering tools to customize, automate and govern the applications.
- GenAI applications are becoming more common as incumbent vendors and startups embed GenAI models to automate workflows for business users.
- The enormous demand for custom infrastructure to train the GenAI models has created a huge mismatch between supply and demand (particularly for GPUs).

Across each of these layers are specific technology segments, some of which are quite novel, while others predate the existence of GenAI. Figure 1 illustrates the various layers of the GenAI technology landscape.

Figure 1: Generative AI Technology Landscape



Source: Gartner
793970_C

Gartner

GenAI Models — Assess Their Fit for Your Case

GenAI models are at the core of hype and disruption in this space. These models have evolved significantly from early models such as BERT and GPT-2 in terms of size, capabilities and performance. The GenAI model landscape consists of foundation models, domain models and model hubs. Assess a model's fit for your case based on accuracy, costs, time to value and security and privacy principles.

Foundation Models

What Are They?

Foundation models are massively pretrained models on a huge corpus of unlabeled internet data. They can either be transformer-architecture-based models (like most large language models) or diffusion-based models (like most computer vision models).

Why Are They Important?

It isn't an exaggeration to say that the real revolution in the GenAI space was heralded by the launch of foundation models. Their versatility and ability to perform a wide variety of tasks has made them very popular with enterprise users. A huge ecosystem of APIs (such as ChatGPT APIs) and applications have been built on top of these foundation models.

Who Are Some Sample Vendors?

Alibaba, Anthropic, Baidu, Cohere, Google (PaLM 2, LaMDA), Meta, MidJourney, NVIDIA, OpenAI and Microsoft (GPT-4, DALL·E 2) and Stability AI

Domain Models

What Are They?

While foundation models are a significant technical advance, they may be impractical for many use cases due to their general purpose nature and cost, complexity and hallucination risks. The proliferation of open-source foundation models such as Bloom, GPT-J and Llama 2 are giving rise to use-case-specific domain models. These fine-tuned models may or may not be open source, but are generally smaller than the foundation models, often optimized for specific use cases or the needs of vertical industries. A recent example is the release of Med-PaLM 2 by Google and BloombergGPT. Several companies building foundation models are also releasing fine-tuned models.

Why Are They Important?

Domain models can improve the use-case alignment within the enterprise while delivering optimal accuracy at a lower cost. Through more targeted training datasets, these models have the potential to lower hallucination risks associated with large models but at the risk of sacrificing versatility.

Who Are Some Sample Vendors?

AI21 Labs, AWS, Bloomberg, Databricks, Google, IBM, Microsoft, NVIDIA, OpenAI and Salesforce

Model Hubs

What Are They?

Model hubs are an equivalent of an "app store" — they serve as a marketplace or hubs for models — and are often open source. Model hubs such as Hugging Face are also offering automation and governance tools and curated datasets, as well as model APIs and GenAI applications, targeting enterprise needs.

Why Are They Important?

Model hubs help developers find GenAI models and datasets to embed in their custom applications or workflow. Thousands of models are available across a broad spectrum of use cases, and model hubs remove a lot of friction for developers looking to rapidly iterate and build GenAI applications. Model hubs serve an important role in democratizing access to the vibrant ecosystem of open-source models.

Who Are Some Sample Vendors?

Hugging Face, Replicate

GenAI Engineering Tools — Adopt a Platform Approach for Enabling Automation, Standardization and Trustworthiness

GenAI engineering tools enable enterprises to operationalize models faster, balancing both governance and time to market. AI engineering tools can be subdivided into model-centric and data-centric tools. Terms such as DataOps, LLMOps, LangOps or FMOps, or more broader terms such as ModelOps or MLOps, are used frequently, but we believe they are a subset of AI engineering.

Some of the prominent market categories are outlined in this research.

Prompt Engineering Tools

What Are They?

Prompt engineering tools enable user organizations to provide adequate context, examples and data retrieval to a GenAI model to steer it to provide the outcomes that the user desires. In prompt engineering, the GenAI models are frozen but steered via simple prompt engineering techniques, in context prompting or prompt augmentation.

Why Are They Important?

Prompt engineering tools help users by automating prompt creation, making available popular and precreated prompts (prompt marketplaces), providing prompt versioning (for reuse), enabling governance in prompt engineering (such as blocking unsafe prompts or theft of intellectual property) and enabling data retrieval to deploy retrieval augmented generation (RAG) architectures. Prompt engineering is an effective way to steer the model without incurring significant model retraining costs, meaning prompt engineering tools will be critical for IT leaders.

Who Are Some Sample Vendors?

Promptable, PromptBase, PromptLayer and Microsoft

Vector Databases

What Are They?

A vector database is a type of database that stores vectors or spatial representations of enterprise data. By storing, indexing and retrieving vector embeddings, the vector database performs the function of similarity search that finds the best match between the user's prompt (the question) and the particular vector embedding.

Why Are They Important?

Vector databases enable applications to respond with low latency to high concurrency requests (prompts), thereby ensuring better user experience for a variety of GenAI use cases. The usage of vector databases also enables superior prompt augmentation and reuse of organization data, which can be searched and retrieved easily, thereby boosting model performance and accuracy.

Who Are Some Sample Vendors?

Activeloop, AWS, Chroma, Elastic, Google, Microsoft, Pinecone, Qdrant, Weaviate and Zilliz

Model Fine-Tuning Tools

What Are They?

Fine-tuning tools allow enterprise data to be labeled, annotated and used for model fine-tuning to create task-specific models. Through prebuilt connectors, these tools enable enterprises to harness data from a variety of popular data sources.

Why Are They Important?

Fine-tuning tools enable the creation of high-quality enterprise data that can be used to fine-tune base models for task specificity, higher model performance and lower hallucinations. As smaller and open-source models gain traction, and as proprietary models enable fine-tuning via their APIs, more enterprises may seek to use fine-tuning tools to customize the models.

Who Are Some Sample Vendors?

Cleanlab, Scale AI and Snorkel AI

Model Deployment Tools

What Are They?

Model deployment tools aid in experiment tracking; automate infrastructure configuration; visualize datasets; enable versioning of APIs and applications for optimized production deployments; and perform automated unit tests to measure reliability and performance.

Why Are They Important?

Model deployment tools automate the life cycle from build, experimentation and to actual product launch, and they provide flexibility for enterprises to deploy the models anywhere with appropriate checks and balances.

Who Are Some Sample Vendors?

AirOps, Anyscale (Ray), Cerebrum, HoneyHive, Humanloop, Neptune.ai, OctoML, Seldon Technologies and Weights & Biases

Application Framework Tools

What Are They?

GenAI application frameworks provide an abstraction layer to enable prompt chaining; model chaining, interfacing with external APIs; retrieving contextual data from data sources; and maintaining statefulness (or memory) across various model requests. Some of these tools also provide templates for developing GenAI applications.

Why Are They Important?

GenAI application frameworks orchestrate workflows by chaining prompts or models together to achieve intended outcomes. They enable effective prompting through prompt templates, input prompt optimization and output parsing. In addition, they help with model accuracy by providing an intuitive and efficient way to search and summarize corporate data and documents, enabling organizations to leverage AI models and by augmenting their value through innovative indexing techniques.

Who Are Some Sample Vendors (or Projects)?

deepset, Dust AI, LangChain, LlamaIndex and Microsoft

AI Trust, Risk and Security Management (TRiSM) Tools

What Are They?

AI TRiSM tools are necessary to manage new risks and threats introduced by GenAI and consist of tooling and frameworks that help implement AI governance. This includes solutions and techniques for model explainability, data protection, model monitoring, fairness, privacy, content moderation and application security.

Why Are They Important?

The field of GenAI has been progressing quickly; however, if ungoverned, GenAI will exacerbate the internal and external risks for an organization. These risks include loss of intellectual property, hallucinations, lack of model explainability, bias and toxicity in model output and misinformation. AI TRiSM tools can help organizations mitigate, if not eliminate, the risks.

Who Are Some Sample Vendors?

Arize AI, Arthur, CalypsoAI, Credo AI, Fiddler AI, Gretel, MOSTLY AI, Protopia AI, Spectrum Labs and TruEra

GenAI Applications — Assess Security and Privacy Practices, Workflow Integration and Track Record of Delivering Innovation

What Are They?

GenAI-enabled applications use GenAI for UX and task augmentation to accelerate and assist users through the completion of their desired outcomes. When embedded in the experience, GenAI offers richer contextualization for singular tasks like generating and editing text, code, images and other multimodal output.

Why Are They Important?

The use cases will permeate across a wide spectrum of domains and skill sets within the knowledge workforce, reimagining how enterprises, institutions and the public sector think of scale and productivity. Enterprises will onboard gradually, likely embracing the GenAI features that will be delivered custom-made into their productivity and business processes tools.

The enterprises that already have established maturity in AI engineering and app development will likely access the midlevel to lower level of the stack to build/refine their own models and applications to differentiate further from the competition.

Who Are Some Sample Vendors?

The ISVs targeting the GenAI applications will be both incumbents and new entrants. The incumbent vendors that already have domain-based applications will move to enhance the user experience within the workflows and tasks of the applications with GenAI capabilities. New entrants will likely be designed specifically around GenAI and will focus heavily on knowledge management, retrieval and generative experience around specific tasks (like productivity) and vertical domains (healthcare, legal, finance, other).

Table 1: Horizontal Apps (Business Function Focused) Sample Vendors

(Enlarged table in Appendix)

Business Function	Sample Vendors
Marketing, Communications and Advertising	Jasper AI, Tavus, Tome, Rytr
Sales	Lavender, Microsoft, Outplay, Regie.ai, Salesforce
Enterprise Search and Knowledge Management	Algolia, Glean AI, Google, Microsoft, Sana Labs
Design and Creativity	Adobe, Character.AI, Diagram, Dubverse, Rephrase.ai, Synthesia
Customer Service	Cresta, Landbot, Quickchat AI
Workforce Productivity	Cogram, Google, Mem, Microsoft, Notion, OpenAI, Supernormal
Finance and Accounting	AlphaWatch AI, Truewind
IT	AWS, Codeium, GitHub (Microsoft), Google, Red Hat (IBM), Tabnine

Source: Gartner (September 2023)

Table 2: Vertical Apps — Sample Vendors

Industry	Sample Vendors
Financial Services	Bloomberg, Charli AI
Healthcare	Google, Microsoft, Replikr
Legal	Casetext (Thomson Reuters), Darrow, Harvey, PatentPal
Life Sciences	Exscientia, Huma.ai, Insilico, Zephyr AI

Source: Gartner (September 2023)

Infrastructure — Work With I&O Leaders to Assess Infrastructure Fit and Encourage Efficient, Reliable Utilization Practices

GenAI models and applications require significant investments in specialized training infrastructure due to their massive size, which increases the amount of infrastructure required for initial training, retraining and inferencing. Infrastructure for GenAI models can be classified into three major subcategories — computing, networking and storage.

Computing

What Are They?

When building and using GenAI tools, compute resources are required during two distinct phases: training the model and then during runtime (or inference). GPUs are the workhorse powering the model training process. While the model training is an extremely compute intensive process, more compute cycles will be spent in inference during the life cycle of the model, enabling specialized computation both during training and inference of these models.

Why Are They Important?

GPUs and other specialized compute resources enable model training and inference at scale with high performance. Beyond the capabilities of AI chips, functionalities such as distributed training, parallel computing, memory management, performance profiling and optimization will become important. While most of model training and inference for GenAI applications happen in the public cloud, the increase in open-source models may herald investments in data center compute infrastructure, particularly in regulated industries, although most model deployments will continue to be cloud-centric.

Who Are Some Sample Vendors?

AMD, AWS, Cerebras, CoreWeave, Google, IBM, Intel, Microsoft, NVIDIA and SambaNova

Networking

What Are They?

Networking hardware and software as well as connectivity services are essential in the collaborative development and operationalization of GenAI models that span resource interconnectivity, data collection and ongoing model monitoring, optimization and management. Switching infrastructures will need upgrades to keep up with the high scale and throughput requirements of dense GPU requirements associated with GenAI models.

Why Are They Important?

Networking and connectivity services are essential for processes such as data acquisition, data preprocessing, distributed training, model deployment, model monitoring and security in developing models and applications for GenAI. Packet loss and low throughput can often be the core issues throttling performance of AI models as they have struggled to keep pace with compute and storage performance improvements.

Who Are Some Sample Vendors?

Arista Networks, AWS, Cisco, DriveNets, Google, Huawei, Juniper Networks, Microsoft and NVIDIA

Storage

What Are They?

Foundation models access and process significant amounts of data storage due to the model size. Storage products capable of scaling in capacity and performance are necessary to store large volumes of training data and model checkpoints as well as to process them in real time. Storage for GenAI applications must be able to feed data at large scale and high speeds to keep up with the performance of specialized GenAI-centric compute resources typically based on GPUs instead of CPUs.

Why Are They Important?

Storage solutions differ based on the specific requirements of the foundation models, including data size, access patterns, the decentralization of training, scalability and cost considerations. Depending on the scale of the project, a combination of local storage (memory bound), parallel file systems and object storage can support the high bandwidth, low latency and high IOPS needs of GenAI apps. The most effective storage platform can accommodate multiple access protocols and scale performance and capacity independently, allowing on-demand performance flexibility and long-term cost-effectiveness.

Who Are Some Sample Vendors?

AWS, DDN, Dell, Google, IBM, Microsoft, Pure, VAST and WEKA

Evidence

Generative AI: A Creative New World, Sequoia.

This research is based on detailed vendor briefings with many vendors profiled in this research.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Generative AI: The Basics (Shareable Slides)

How to Pilot Generative AI

How to Choose an Approach for Deploying Generative AI

Glossary of Terms for Generative AI and Large Language Models

Quick Answer: What Are the Pros and Cons of Open-Source Generative AI Models?

Hype Cycle for Generative AI, 2023

Table 1: Horizontal Apps (Business Function Focused) Sample Vendors

Business Function	Sample Vendors
Marketing, Communications and Advertising	Jasper AI, Tavus, Tome, Rytr
Sales	Lavender, Microsoft, Outplay, Regie.ai, Salesforce
Enterprise Search and Knowledge Management	Algolia, Glean AI, Google, Microsoft, Sana Labs
Design and Creativity	Adobe, Character.AI, Diagram, Dubverse, Rephrase.ai, Synthesia
Customer Service	Cresta, Landbot, Quickchat AI
Workforce Productivity	Cogram, Google, Mem, Microsoft, Notion, OpenAI, Supernormal
Finance and Accounting	AlphaWatch AI, Truewind
IT	AWS, Codeium, GitHub (Microsoft), Google, Red Hat (IBM), Tabnine

Source: Gartner (September 2023)

Table 2: Vertical Apps — Sample Vendors

Industry	Sample Vendors
Financial Services	Bloomberg, Charli AI
Healthcare	Google, Microsoft, Replikr
Legal	Casetext (Thomson Reuters), Darrow, Harvey, PatentPal
Life Sciences	Exscientia, Huma.ai, Insilico, Zephyr AI

Source: Gartner (September 2023)