Innovation Guide for Generative AI Technologies

Published 28 July 2023 - ID G00793932 - 28 min read

By Analyst(s): Radu Miclaus, Anthony Mullen, Arun Chandrasekaran

Initiatives: Artificial Intelligence; Evolve Technology and Process Capabilities to Support D&A; Software Engineering Technologies

Generative AI technologies are emerging rapidly and promise substantial value in the enterprise. Like the broader category of AI, GenAI permeates the entire technology stack and the majority of industry verticals. IT, Data and Analytics and software engineering leaders can use this high-level guide to ground themselves in the vendor landscape for GenAI.

Overview

Key Findings

- While generative AI (GenAI) has had several niche applications over the last years, 2023 was a breakout moment with the GPT-based ChatGPT chatbot from Open AI gaining huge adoption and mind share from both enterprise buyers and vendors looking to use this technology in their solutions.
- Vendors in the GenAI space span all layers of the enterprise stack, from underlying compute to development tooling and end applications. We see both incumbent platforms (e.g., DSML, CRM, ERP) adding GenAI to what they do as well as net new generative platforms and services coming to market.
- While generative technologies and applications are diverse in what they generate (images, text, videos, code, designs, 3D models), we see a large portion of market activity and investment driven by foundation models, and in particular, large language models (LLMs) and their surrounding ecosystem.
- The first wave of vendors in the market has centered on the rapid production of content and experiences, aided by enterprise information and knowledge bases. The second wave of disruption, and resulting market offerings, will look at dynamic process/workflow and generative orchestration using approaches such as multiagent systems, plug-ins and simulation.

Recommendations

- **Plan ahead to reduce the technical debt of GenAI pilots.** While today, ChatGPT is the most popular generative model application, other models are joining the market to provide both general-purpose capabilities along with industry- and task-centric solutions. Design solutions to be loosely coupled to generative models to allow for flexible selection and combinations of models. Ensure you can flexibly use enterprise knowledge assets (e.g., content, data, rules/heuristics, corpora, digital twin models, knowledge graphs) to prompt and ground the behavior of GenAI models to a wide selection of generative services. If you haven't developed the semantic data layer for your business, you must begin now.

- **Consider ethical and responsible AI practices of vendors.** Check content training provenance of solutions to appraise the risk you are exposed to. Many of the popular generative models are embroiled in class action lawsuits due to copyright violations. While legal penalties don't exist today, the landscape will tighten as legislation comes into force over the coming years. While markets are evolving to support responsible GenAI, they are immature, and initiatives to watermark generative output will take time to gain a networking effect of adoption. Ensure procurement and legal teams are included in vendor rationales and selection.

- **Check your existing application portfolio for their GenAI roadmaps.** While it might be tempting to just pay to enable new GenAI features in applications you already own, doing so will have major risk and cost implications that must be appraised. The vendor landscape in the first half of 2023 was very tool-focused. However, as vendors begin to adapt core GenAI technologies to domains, expect the market in the second half of 2023 to be complemented by a rich set of solutions specializing by role, business unit and industry.

- **Evaluate vendor solutions thoroughly, and defer large AI architecture decisions until 2024 when solutions stabilize.** While the GenAI paradigm offers much promise and an overhaul of the technology marketplace and ecosystem, there are many unknowns in using this technology. Along with technical considerations of repeatability and unintended consequences is the issue of price and business model. While GenAI may be applied in many locations across the enterprise, the monetary cost of doing so is not clear. Develop and refine a cost/value model to compare the as-is versus GenAI-enabled versions of your business.

Strategic Planning Assumption(s)

- By 2025, the top five vendors across all categories of enterprise software will use GenAI in their pipeline.
- By 2026, the number of companies using open-source AI directly (not indirectly via other vendors) will increase tenfold.
- By 2026, GenAI will facilitate an increased use of other AI technologies (aside from GenAI) by 400%.

The following research is part of a new initiative Gartner is piloting to provide updates at a greater frequency. It is a work in progress that does not represent our final position. While we continue to monitor this topic, we invite you to provide constructive feedback. All relevant updates and feedback will be incorporated into the final research, which will undergo our standard review process.

Table of Contents

Use these jump links to navigate the document:

Market Definition

Back to top

GenAI is not a market per se; it permeates across the entire technology stack and the majority of verticals. The new way to interface with technology is bringing disruption to the technology usage patterns for both consumers and workers.

Gartner defines generative AI as technologies that "can generate new derived versions of content, strategies, designs and methods by learning from large repositories of original source content. GenAI has profound impacts on the business including content discovery, creation, authenticity and regulations; automation of human work; and the customer and employee experience."

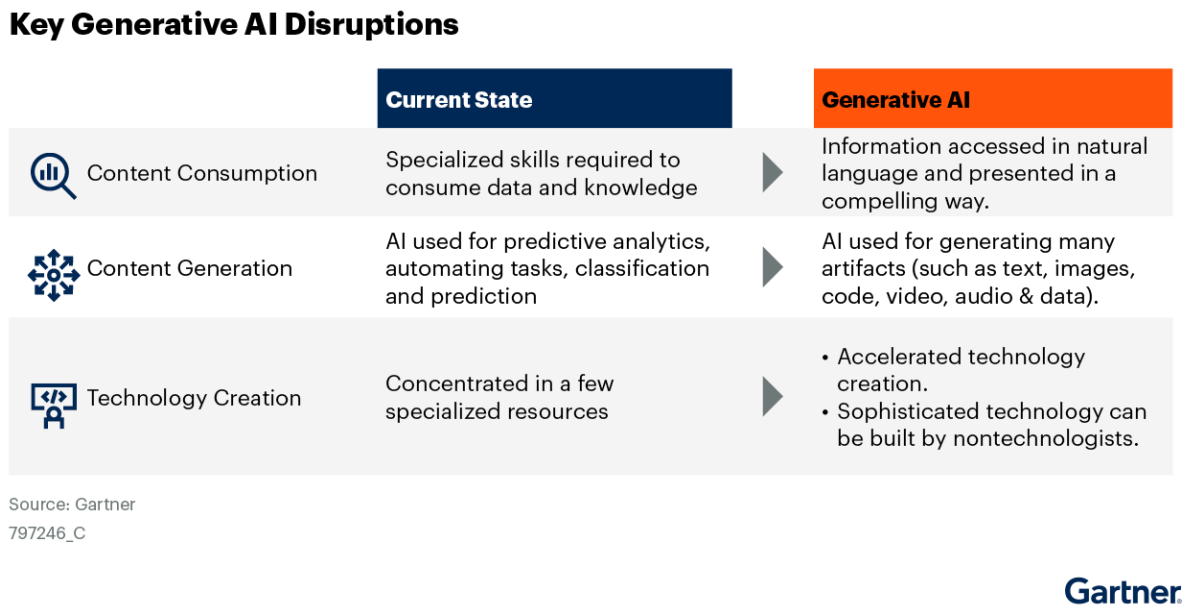
The GenAI market is composed of the following segments, with vendors often present in multiple segments:

- **Infrastructure providers** — This segment comprises the vendors of infrastructure (both hardware and IaaS and PaaS offerings) to support the GenAI needs for compute, storage and network. It also includes software providers offering capabilities for infrastructure orchestration, tuning and scaling for GenAI development and applications in production.
- **Model providers** — This layer of vendors offers access to commercial or open-source foundation models such as LLMs and other types of generative algorithms (such as GANs, genetic/evolutionary algorithms or simulations). These models can be provided for developers to embed into their applications or be used as base models for fine-tuning customized models for their software offerings or internal enterprise use cases.
- **AI engineering** — The vendors in this segment are made up by incumbent and startup vendors covering full-model life cycle management, specifically adjusted to and catering to development, refinement and deployment of generative models (e.g., LLMs) and other GenAI artifacts in production applications.
- **Generative AI apps** — GenAI applications use GenAI capabilities for user experience and task augmentation to accelerate and assist the completion of a user's desired outcomes. When embedded in the experience, GenAI offers richer contextualization for singular tasks such as generating and editing text, code, images and other multimodal output. As an emerging capability, process-aware GenAI agents can be prompted by users to accelerate workflows that tie multiple tasks together.

GenAI has had several niche applications in the last few years, especially targeting use cases such as simulation, synthetic data generation, conversational AI, advanced intelligent document processing and search. The research in transformer-based models and LLMs has been progressing rapidly over the last three years with major breakthroughs in 2022, culminating with the release of the GPT-based ChatGPT chatbot from OpenAI. The capabilities of ChatGPT illustrate the massive opportunity for LLMs being used in reinventing the interface with technology and the way we use data analysis and synthesis (structured, semistructured, unstructured).

For the enterprise specifically, three major areas of GenAI disruption are related to usage patterns, as described in How to Pilot Generative AI (see Figure 1).

Figure 1: Key Generative AI Disruptions



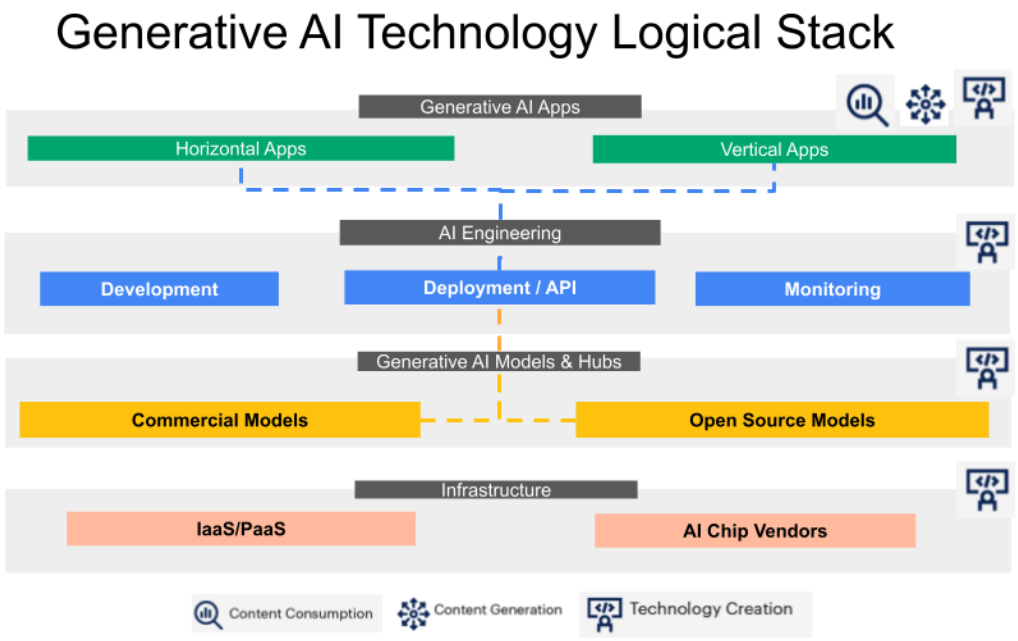
These areas of disruption are important when buyer organizations explore their investments and adoption routes for GenAI. The technology decisions need to align with the business use cases and the organization's AI maturity. The technology stack is evolving rapidly and can meet a wide range of needs for organizations, whether they are looking for off-the-shelf productivity tools augmented with GenAI or looking to build their own GenAI applications with models refined on their proprietary data.

Market Map

Back to top

In Figure 2, we map the logical representation of the technology stack with the GenAI disruptions in the enterprise. The stack flows from the bottom up, from infrastructure to model providers, AI engineering and finally GenAI applications.

Figure 2: Generative AI Technology Logical Stack



Market Dynamics

Back to top

While generative technologies and applications are diverse today, we see a large portion of market activity and investment driven by foundation models, and in particular, LLMs and their surrounding ecosystem.

The market dynamics for vendors are evolving rapidly. Each section of the logical stack above has different competitive aspects, vendor focuses, and buyer profiles and needs met. Below is a conversation on the player types in each layer and their dynamic in the market.

Infrastructure

At a high level, the infrastructure layer is shared by the AI chip vendors who have seen an increasing demand in their products to sustain both training and inference workloads as well as the IaaS and PaaS offerings from hyperscaler CSPs. The buyers of infrastructure are the technology vendors themselves, service providers and enterprises who are interested in technology creation by building models using AI engineering and integrating them into applications either for internal use cases or to monetize externally.

IaaS and PaaS Vendors

The majority of training and inference is happening in cloud providers due to the scalable, elastic and cost-efficient options for compute, storage and network. Hyperscalers are interested in supporting the activity for the entire technology stack all the way to applications. The GenAI development services offered tend to be anchored by the availability of LLMs (either developed in-house, acquired through partnerships or OSS). The ancillary services related to infrastructure orchestration overlap both IaaS and PaaS capabilities (distributed computing, cluster management, memory management, storage and network optimization as well as robust observability) and support the build of applications by both technology vendors and enterprises.

Most of the demand for compute, network and storage hardware is currently with the CSPs. The occasional technology vendor may stand up its own architecture to build a custom model. However, until the demand for support of GenAI materializes more substantially in enterprise data centers, CSPs will continue to meet the consumption of infrastructure for enterprise use cases.

AI Chip Vendors

Generative processes are adding to the workload profile of traditional AI methods. The computational profiles for both LLM training and inference are more demanding than traditional AI. This has caused a boom in the demand for chip manufacturers that specialize in high-memory, accelerated computing. Both GPU and special CPU configurations augmented by software layers are the computational engines driving the innovation in and deployment of GenAI models in production.

Some AI chip manufacturers are creating deeper differentiation by integrating vertically in the software layers that support model development and implementation, getting them closer to the value creation in the applications at the top of the tech stack.

Model Providers

The model war is in full swing as the opportunities for monetizing generative models are presently peaking. The leaders in model quality and versatility are companies that have been aggressive in internet data collection to support the massive task of training and fine-tuning LLMs. As the commercial options started appearing and the opportunity became clearer for the technology community, the open-source movement started focusing toward LLM options that attempt to level the playing field for developers and organizations looking to build on top of generative models. In addition to the size of models for certain use cases starting to shrink, the amount of data and human input to refine and fine-tune the models is getting streamlined. The optionality for developers is increasing between the commercial and open-source options.

Commercial Models

The commercial/closed models such as GPT-3, GPT-4, LaMDA, Amazon Titan, ERNIE and PaLM are powering the economic engines of the companies who invested in developing them. The owning vendors use commercial models to build applications on top of them to monetize these applications as well as offer them to developers via APIs for embedding into applications. Special partnership agreements allow partners to deploy instances of the models in their cloud infrastructure and build applications on top while wrapping enterprise security and privacy around the deployments.

Open-Source Models

Open-source LLMs such as BLOOM, GPT-J, Llama 2, Dolly and OpenLLaMA are the result of community efforts to offer options for the developer communities that allow them to innovate and commercialize on top with value-add applications. The open-source communities have always been proactive in building capabilities for developers (and other personas), and GenAI is a massive opportunity. Open-source options are very attractive for companies piloting and proving use cases, as well as for enterprises who have mature and experienced engineering and operations teams that can take on the operational maintenance of the models (including testing) and even contribute to the projects.

The models can be accessed through APIs from the commercial vendors or through community-maintained model hubs (especially popular for open-source models). Enterprises using commercial model APIs or even deploying models on their infrastructure engage in building applications using easy to medium AI design patterns. Enterprises using foundation models to refine and fine-tune them for domain-specific applications are more advanced and are engaging in the more difficult AI design patterns for LLMs, including advanced AI engineering. For a more detailed conversation on difficulty levels of embedding LLMs into enterprise applications, see AI Design Patterns for Large Language Models.

AI Engineering

To date, the most evolved AI engineering discipline for GenAI was in the simulation market. Simulations create a model of the world ("grounding data" in LLM-speak), which can be used along with AI to generate artifacts and events (synthetic data) as well as use multiagent systems ("plug-ins" in LLM-speak) and reinforcement learning to generate processes, learning methods and strategies (see Predicts 2023: Simulation Combined With Advanced AI Techniques Will Drive Future AI Investments).

Today, however, the leap for many enterprises to simulation as an overarching paradigm of design and development is a step too far. However, leveraging foundation models easily with existing content and knowledge to deliver a variety of use cases is attractive to the majority of organizations. Currently, the most popular generative model to bring into AI engineering workflows are LLMs. Buyers of technologies targeting AI engineering for LLMs have a high AI maturity and deep knowledge and discipline around ModelOps. AI engineering capabilities will help technology vendors, service providers and enterprises engage in creating custom applications powered by a mix of general-purpose, domain-specific and task-centric generative models.

AI engineering for GenAI will bring learning curves to the enterprise with regard to processes for model development, model deployment and monitoring. We observe the following in the market.

Development

Development options in the GenAI market range from using, training or building individual generative models to composite AI assemblies through to broader generative systems development. In LLM development, enterprises have options from both the incumbent DSML engineering platforms as well as new entrant startups specializing in LLM development (see Market Guide for DSML Engineering Platforms).

The AI engineering workflow (train, design, build, tune) for GenAI is different from traditional machine learning development, including the handling of artifacts such as corpora, content and semantic assets like knowledge graphs. AI systems and projects have a mix of models, code and artifacts that challenge advanced-analytics-centric, MLOps-only approaches. Expect a greater intersection between the MLOps and DevOps markets (XOps). This will produce a learning curve for enterprises and will require iteration. Complementary development markets, such as data labeling and synthetic data, support generative development by providing capabilities such as corpora labeling and synthetic training and test data.

One emerging area in model development is the use of composite AI via chains of LLMs, which allows for the design of sequences of generative tasks to enable support of more complex use cases. The learning curves for developers working with frameworks such as LangChain (GitHub) or Transformers Agent (Hugging Face) will entail a mentality switch toward the design of agent-like behavior that adapts to the prompt inputs from applications users to complete more complex tasks.

Deployment

Model deployment will encompass architecting the generative service in a scalable and cost-efficient manner, managing the integration endpoints, performance testing and CI/CD capabilities. Besides the vendors that will extend XOps capabilities for LLM processes, data store vendors like vector database providers will be an important consideration on how the back-end services are configured and designed.

Monitoring

When it comes to monitoring objectives, generative models are different from traditional machine learning models. Since generative models respond to prompting or seed conditions in different ways, it is very important to monitor the interaction between the prompt and completion, observing how the model interprets the inbound prompt and how it responds to it. Elements such as loss of context, factual accuracy drift, hallucination or tone alteration (abusive/rude) are part of an automated monitoring capability. Vendors providing these capabilities will also add observability and reporting for owners of generative applications to understand how the user population is using the applications. If alerts are brought up, the monitoring functions will flag the model for more fine-tuning by the developers.

As enterprises either implement existing models or build their own, they will increasingly demand monitoring for ethical use, abuse prevention as well as IP compliance assurance based on the provenance of the model and/or the data used to train it.

Generative AI Apps

GenAI-enabled applications will primarily target technology users interested in content consumption and content generation/creation. The delivery mechanism for GenAI apps can be:

- Brand-new GenAI applications
- Existing applications that have added GenAI capabilities

The buyer organizations will focus on two (sometimes complementary) needs:

- The needs of users (internal and external) for content consumption in the context of their tasks
- The needs of users for content generation/creation in the context of their tasks

Horizontal Applications

Horizontal applications are GenAI apps that cut across multiple verticals. The horizontal nature can be task-oriented — such as communications, creative design, business process and workflow as well as low-code or no-code generation — but also function-oriented, including marketing, sales, customer support, HR, IT, software engineering, knowledge management, general productivity and collaboration tools. The vendor landscape for horizontal applications will include incumbent vendors that are adding GenAI capabilities for content consumption and content generation via assistants and/or chatbot plug-ins, as well as startups that may choose to offer new processes and redesigned experiences for users.

Vertical Applications

Vertical applications will have similar elements as the horizontal applications but will focus more on the vertical/industrial domain. These applications will use fine-tuned LLMs with domain refinement and vertical-specific workflows and tasks that enable productivity for specialized users working directly with the respective domains. Examples of vertical GenAI applications include drug discovery and research for life sciences fields, compound and material sciences applications, generative wealth management tools for the financial sector, and assistants for legal and compliance research.

Market Evolution

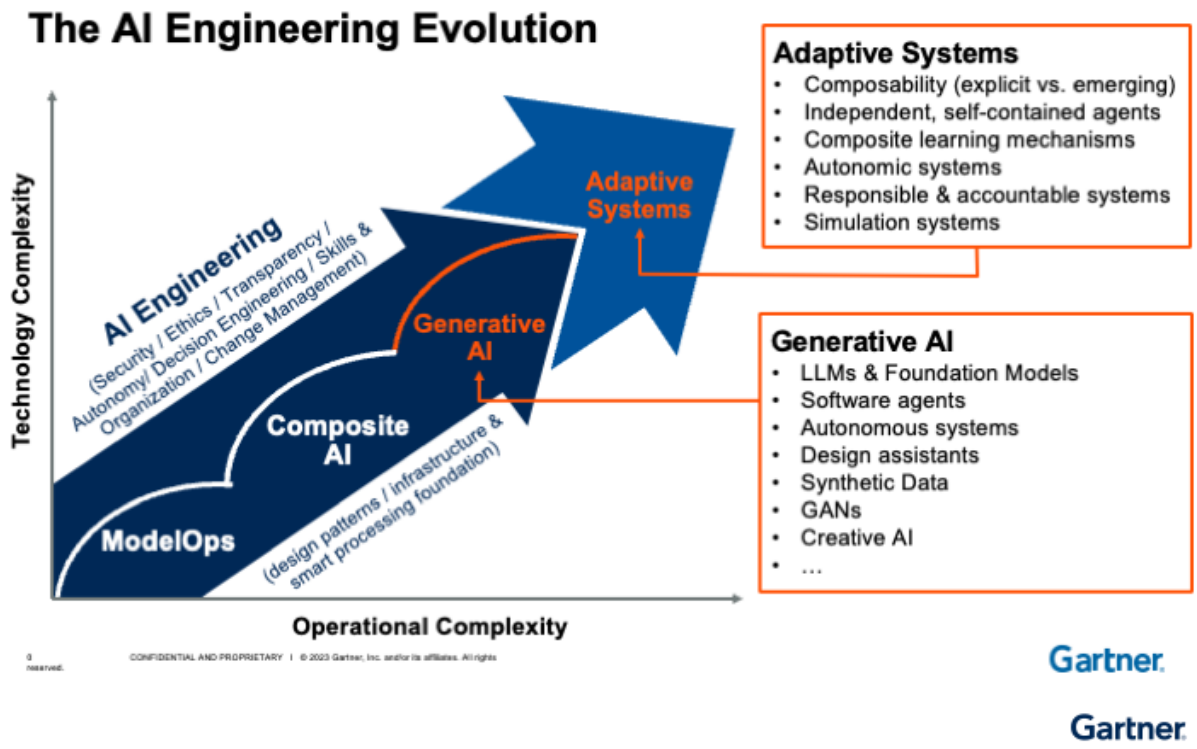
Back to top

While generative capabilities will be present in many software categories and create new markets, GenAI also has a systemic impact on AI overall and unlocks the next phase of AI — namely, adaptive AI (see Figure 3).

"Adaptive AI systems allow for model behavior change post deployment by learning behavioral patterns from past human and machine experience, and within runtime environments to adapt more quickly to changing, real-world circumstances."

See Top Strategic Technology Trends for 2023: Adaptive AI.

Figure 3: The AI Engineering Evolution



GenAI provides the following key capabilities to support the next era of AI development — adaptive AI. The ability to:

- **Create not just artifacts (text, code, imagery, design components) but processes and subtasks** — The ability to generate processes and subtasks using technologies like LLMs gives workflows much more flexibility and ultimately the potential for process and code to be dynamically generated from high-level tasks.
- **Support for developing and coordinating multiple AI and software components** — Historically, data scientists and experts would have to create composite AI systems that brought together different techniques, which took time and money. GenAI dramatically democratizes the ability to create complex systems from input prompts and seed conditions.

GenAI alone cannot provide a complete infrastructure for the much broader field of adaptive AI. It will both complement and improve it. The markets it will intersect with deeply are:

- **Multiagent systems** — Today, these systems are uncommon and most often found in academia. However, the recent release of ChatGPT plug-ins (a network of experts) and tools such as Hugging Face's Transformers Agent will quickly shift focus from research to commercialization. GenAI can turn external and internal marketplaces into composable AI assets by dynamically commissioning them at run time.
- **Composite AI** — Composite AI is an approach that uses different AI techniques in combination to solve problems. Related terms are hybrid AI, neuro-symbolic AI and causal AI. GenAI will accelerate the development of composite AI systems, and we expect the following intersections:
  - GenAI tools with access to AI marketplaces and assets can create composite AI experiences, bringing together different types of AI such as machine learning, logic and rules, and optimization techniques.
  - GenAI will drive the use of non-GenAI AI technology, where models are selected dynamically from a pool of resources. It will accelerate the use of other AI.
  - GenAI will use composite techniques to tackle the challenges present in GenAI models such as bias, fact checking and hallucination.
- **Simulation technologies** — These technologies can be used to develop reusable environmental models (geographic, physical, conceptual) and provide a "stage" for multiagent systems (see Innovation Insight: AI Simulation). Simulation technologies and GenAI have multiple intersections:
  - Using generative techniques within simulations to generate events (without the need for training data)
  - Using simulation environments to support techniques such as reinforcement learning and multiagent learning
  - Using generative technologies to develop assets (2D, 3D, audio) for simulations

- **Semantic AI technologies** — These technologies can be used to model human heuristics, relationships and rules to steer the behavior of systems — generative and otherwise. They also provide a cause and effect analysis and make for a foundation of a shared language between AI and humans to empower explainable AI.
  - Semantic assets can be used as data payloads for generative models, act as a "grounding" of facts for generation and support collaboration in multiagent systems.
  - LLMS can be used to reverse-engineer semantic assets such as taxonomies, ontologies and graphs.

- **Decision intelligence platforms** — These technologies support the formulation and expansion of the discipline used to improve decision making by explicitly understanding and engineering how decisions are made and how outcomes are evaluated, managed and improved via feedback (see Innovation Insight for Decision Intelligence Platforms). We see the following:
  - Use of LLMs to mine unstructured and semistructured text to extract division intelligence assets (e.g., decision trees, eligibility criteria).

- **Insight engines** — Insight engines combine search, composite AI and GenAI to enable context-enriched analysis and synthesis, as well as the delivery of actionable information, derived from all types of content and data within and outside of an organization. They will evolve to both deliver generative experience and services within the platform as well as act as a companion for GenAI systems, where they will deliver high-speed, high-volume content and artifacts for GenAI workloads (see Magic Quadrant for Insight Engines).

Business Benefits (Use Cases)

Back to top

Gartner research is continuously focusing on both helping enterprises define the business benefits of adopting GenAI technologies as well as mapping the most impactful use cases to achieve the business outcomes. In our Board Brief on Generative AI, we describe the following areas of revenue and cost and productivity opportunities.

Revenue Opportunities

- **Product development:** AI will enable many enterprises to create new products more quickly. Pharma, healthcare and manufacturing (CPG, food and beverages, chemicals and materials science) will become AI-first industries to create new drugs, less-toxic household cleaners, novel flavors and fragrances, new alloys, and faster and better diagnoses.

- **New revenue channels:** Enterprises with greater levels of AI maturity will gain greater financial benefits associated with revenue, according to our recent survey. Top AI use cases among mature AI organizations are leveraging AI more for creating new revenue channels (34%). [1]

Cost and Productivity Opportunities

- **Worker augmentation:** Use cases demonstrate howGenAI can augment workers' ability to draft and edit text, images and other media. It can summarize, simplify and classify content. It can also generate software code, translate and verify, and improve chatbot performance. At this stage, GenAI is highly proficient at creating a wide range of artifacts that users can describe or imagine quickly and at scale.

- **Long-term talent optimization:** Employees will be distinguished by their ability to conceive and execute ideas, projects, processes, services and relationships in partnership with AI. This symbiotic relationship will accelerate workers' time to proficiency and greatly extend their range and competency. Going forward, the impact on job roles and staff skills will be profound.

- **Process improvement:** GenAI can derive real, in-context value from vast stores of content, such as documents, correspondence and transcripts. Until now, a wealth of data has gone largely unexploited. Now, with GenAI, it will generate more value for enterprises that leverage it. Most content, data and workflow jobs will change.

The use cases for GenAI in business domains and industrial verticals will continue to be overlaid in our evolving research series around applying AI: Applying AI — A Framework for the Enterprise. The familiar concept of use-case prisms can guide enterprise customers in their ideation and prioritization of use cases and investment priorities. See the range of use-case prisms in Uncovering Artificial Intelligence Business Opportunities in Over 20 Industries and Business Domains. An example of an industry-focused use-case prism is available in Use-Case Prism: Generative AI for Manufacturing.

Piloting and Evaluating Vendors

Back to top

Enterprises are approaching the adoption curve for GenAI at different paces and entry points, and exploration and piloting are a crucial step. The comprehensive research on exploration and piloting efforts for enterprises (How to Pilot Generative AI) provides details on how to run a GenAI pilot.

When it comes to selecting technologies and using out-of-the-box productivity features, technology leaders can use the table below. It describes examples (not exhaustive) of buyer roles, technology use cases and high-level capabilities for the technology stack layers discussed in the Market Definition and Market Dynamics sections.

Table 1: Buyer Roles, Use Cases and Capabilities for the Different Technology Stack Layers

(Enlarged table in Appendix)

	Infrastructure	Model Providers	AI Engineering	Generative AI Apps
Buyer Roles	CTOs, CIOs, cloud architecture leaders	CTOs, digital apps teams, CDAOs	CDAOs, operations leaders, chief data scientists, software engineers	LOB leaders, IT leaders, application leaders
Key Use Cases	- LLM development infrastructure orchestration - GenAI apps infrastructure optimization	- Build applications on top of GenAI models - Monetize access to model APIs	- LLM-focused ModelOps - Generative agent development and deployment - Hand-off to GenAI apps	- Horizontal productivity - Domain-specific analysis and research - Generative automation
Core Capabilities (Buyer-Oriented)	- Infrastructure as code - Compute/storage staging - Network services and/or configuration - Monitoring/FinOps - No-code admin control plane - Infrastructure services marketplace	- Robust developer experience (documentation and support) - Community support - Freemiums/trial periods - Pricing transparency and calculators - Performance benchmarking, testing and SLAs - Model hub access and navigation (optional)	- Infrastructure staging - Vector Management (DBs) - Options for base model selection - Data preparation/labeling/generation for training and fine-tuning (Q&A pairs, labeled data) - Prompt engineering and management - Evaluation, bias detection and responsible AI reporting - Process for using reinforcement learning with human feedback - Ability to build LLM chains or pipelines - Documentation options - Deployment packaging and handoff	- Conversation AI interface (assistant or chatbot) - Prompt engineering experience - Knowledge activation for grounding the generative output - Documentation and process summarization - Personalization driven by usage history and metadata - Collaboration features for sharing/annotations, comments - Decision support/augmentation - User activity/activation observability - FinOps and cost controls - Transparent pricing and calculators

Source: Gartner

Managing Risks

Back to top

The GenAI technologies market is evolving rapidly and, as mentioned above, it will affect the entire technology stack across all verticals. GenAI has the opportunity to enhance existing technology usage patterns and also introduce new ones. As such, technology buyers must both understand their choices for technology and map them to their use cases, learning curves, costs and risks of adoption. The list below represents the areas of consideration for the enterprise when making technology choices, which in turn will affect the way the market will evolve:

- **Learning curves** — The enterprise learning curves will vary depending on the use cases tackled and their current state of maturity in knowledge management, AI engineering, and digital applications building and operations. Pilots and deployment in production may experience delays due to friction in learning curves. Organizations need to plan for extended timelines for the learning curves as well as plan for investment in upskilling proactively.
- **Cost unpredictability** — Enterprise buyers for both centralized procurement and LOB departmental purchases will need to be ready for experiencing change in how they interact and transact with vendors as they add GenAI features. The use cases for GenAI investment, how they interpret/measure productivity and the other investments needed for readiness in the enterprise are all important factors. For a deeper conversation on the value and cost of GenAI, see Assess the Value and Cost of Generative AI With New Investment Criteria.

Buyers of GenAI technologies are exposed to risks and cannot rely solely on policies, controls and assurances provided by the vendors. Covered in detail in our Board Brief on Generative AI, enterprises adopting GenAI capabilities will be opened to the following risks:

- **Lack of transparency risks** — GenAI models are not explainable nor predictable. For boards whose auditors and regulators require attestations regarding the data used by the enterprise, this will limit enterprise use or create risks.
- **Accuracy risks** — GenAI systems consistently produce inaccurate and fabricated answers. Outputs generated by GenAI should be assessed for accuracy, appropriateness and actual usefulness before being accepted.

- **Bias risks** — Enterprises must have policies or controls in place to detect biased outputs and deal with them in a manner consistent with company policy and any relevant legal requirements.
- **Data privacy, intellectual property and copyright risks** — There are currently no verifiable data governance and protection assurances regarding confidential enterprise information.
- **Cyber and fraud risks** — Enterprises must prepare for malicious actors' use of GenAI systems for cyberattacks and fraud, such as deepfakes and those that use deepfakes for social engineering of personnel, and ensure the executive team has mitigating controls in place.
- **Sustainability risks** — GenAI uses significant amounts of electricity. Enterprises that invest in GenAI should encourage the executive team to choose vendors that reduce power consumption and leverage high-quality renewable energy to mitigate the impact on sustainability goals.

For more information on the trust, risk and security management (TRiSM) implications of GenAI, see 4 Ways Generative AI Will Impact CISOs and Their Teams.

The dynamic between buyers and vendors of technology and services will be an important one in covering the risk exposure for enterprises, especially in industries where regulatory scrutiny is high. Buyer organizations should prioritize vendors who focus on and are transparent in delivering enterprise-grade capabilities such as security, privacy, auditability and observability, factual grounding, responsible and ethical AI practices, training content as well as cost visibility and controls.

Vendor Profiles

Back to top

Table 2: Vendor Profiles

(Enlarged table in Appendix)

Vendor	Product (Models)	Infrastructure	Model Providers/APIs	AI Engineering	Generative AI Apps
Adobe	Firefly, Sensei				X
Algolia	Algolia				X
Amazon	AWS, Bedrock, (Titan)	X	X	X	
Anthropic	Claude		X		X
Arize AI	Arize			X	
Bloomberg	BloombergGPT		X		X
Cohere	Embed, Semantic Search, Generate, Command Model, Classify		X		
Databricks	Lakehouse AI, Databricks Marketplace		X	X	
GitHub	Copilot				X
Glean	Glean				X
Google	GCP, Duet AI, Vertex AI, (PaLM, LamDA)	X	X	X	X
Grammarly	Grammarly				X
Hugging Face	Hugging Face		X		
IBM	watsonx	X	X	X	
Jasper	Jasper Everywhere, Jasper App		X		X
Meta	Make-A-Video, Voicebox (Llama 2)		X		
Microsoft	Azure, Office, Dynamics 365, Power Platform	X	X	X	X
MOSTLY AI	Mostly			X	
Nvidia	DGX Cloud, (NeMo, BioNeMo, Picasso)	X	X	X	
OpenAI	ChatGPT (GPT, DALL-E)		X		X
Otter.ai	Otter.ai				X
Salesforce	Einstein GPT, AI Cloud, Service GPT, Sales GPT		X		X
Stability AI	DreamStudio		X		X

Source: Gartner

For a larger set of vendors and classifications, use our Tool: Vendor Identification for Generative AI Technologies to identify software vendors offering development support and out-of-the-box applications.

This is beta research and will be updated frequently in the content and variables presented. Gartner recognizes the vibrant and innovative GenAI community we are a part of and invites vendors to propose their inclusion in this tool by emailing us with relevant details.

Evidence

[1] **2022 Gartner AI Use-Case ROI Survey:** This survey sought to understand where organizations have been most successful in deploying AI use cases and figure out the most efficient indicators that they have established to measure those successes.The research was conducted online from 31 October through 19 December 2022 among 622 respondents from organizations in the U.S. (n = 304), France (n = 113), the U.K. (n = 106) and Germany (n = 99). Quotas were established for company sizes and for industries to ensure a good representation across the sample. Organizations were required to have developed AI to participate. Respondents were required to be in a manager role or above and have a high level of involvement with the measuring stage and at least one stage of the life cycle from ideating to testing AI use cases.

Disclaimer: The results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

Table 1: Buyer Roles, Use Cases and Capabilities for the Different Technology Stack Layers

	Infrastructure	Model Providers	AI Engineering	Generative AI Apps
Buyer Roles	CTOs, CIOs, cloud architecture leaders	CTOs, digital apps teams, CDAOs	CDAOs, operations leaders, chief data scientists, software engineers	LOB leaders, IT leaders, application leaders
Key Use Cases	- LLM development infrastructure orchestration - GenAI apps infrastructure optimization	- Build applications on top of GenAI models - Monetize access to model APIs	- LLM-focused ModelOps - Generative agent development and deployment - Hand-off to GenAI apps	- Horizontal productivity - Domain-specific analysis and research - Generative automation
Core Capabilities (Buyer-Oriented)	- Infrastructure as code - Compute/storage staging - Network services and/or configuration - Monitoring/FinOps - No-code admin control plane - Infrastructure services marketplace	- Robust developer experience (documentation and support) - Community support - Freemiums/trial periods - Pricing transparency and calculators - Performance benchmarking, testing and SLAs - Model hub access and navigation (optional)	- Infrastructure staging - Vector Management (DBs) - Options for base model selection - Data preparation/labeling/generation for training and fine-tuning (Q&A pairs, labeled data) - Prompt engineering and management - Evaluation, bias detection and responsible AI reporting - Process for using reinforcement learning with human feedback	- Conversation AI interface (assistant or chatbot) - Prompt engineering experience - Knowledge activation for grounding the generative output - Documentation and process summarization - Personalization driven by usage history and metadata - Collaboration features for sharing/annotations, comments

		- Ability to build LLM chains or pipelines - Documentation options - Deployment packaging and handoff	- Decision support/augmentation - User activity/activation observability - FinOps and cost controls - Transparent pricing and calculators

Source: Gartner

Table 2: Vendor Profiles

Vendor	Product (Models)	Infrastructure	Model Providers/APIs	AI Engineering	Generative AI Apps
Adobe	Firefly, Sensei				X
Algolia	Algolia				X
Amazon	AWS, Bedrock, (Titan)	X	X	X	
Anthropic	Claude		X		X
Arize AI	Arize			X	
Bloomberg	BloombergGPT		X		X
Cohere	Embed, Semantic Search, Generate, Command Model, Classify		X		
Databricks	Lakehouse AI, Databricks Marketplace		X	X	
GitHub	Copilot				X
Glean	Glean				X
Google	GCP, Duet AI, Vertex AI, (PaLM, LamDA)	X	X	X	X
Grammarly	Grammarly				X

Hugging Face	Hugging Face		X			
IBM	watsonx	X	X	X		
Jasper	Jasper Everywhere, Jasper App		X			X
Meta	Make-A-Video, Voicebox (Llama 2)		X			
Microsoft	Azure, Office, Dynamics 365, Power Platform	X	X	X		X
MOSTLY AI	Mostly			X		
Nvidia	DGX Cloud, (NeMo, BioNeMo, Picasso)	X	X	X		
OpenAI	ChatGPT (GPT, DALL-E)		X			X
Otter.ai	Otter.ai					X
Salesforce	Einstein GPT, AI Cloud, Service GPT, Sales GPT		X			X
Stability AI	DreamStudio		X			X

Source: Gartner