# Trending Questions About Generative AI

20 March 2024 - ID G00806282 - 16 min read

By Ben Yan, Frances Karamouzis, and 5 more

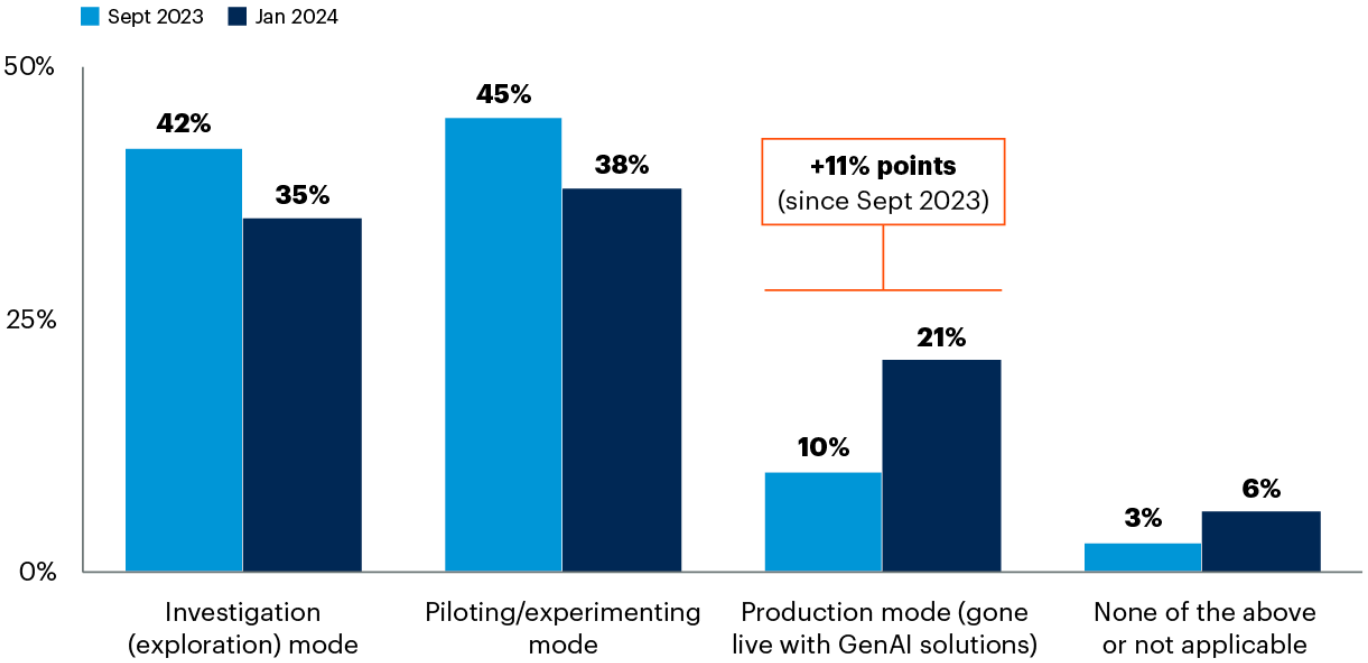
Generative AI continues to be one of the most highly discussed topics in business and on the top of the agenda of the C-suite across all industries and geographies. This research collection can help you better understand trends in GenAI and large language models and choose suitable solutions.

## Analysis

Gartner has conducted a series of webinars and polled its clients during the webinars. [1] The most recent poll reveals that one in five organizations now have generative AI (GenAI) solutions in production. This is up from 10% in September 2023 to 21% in January 2024 (see Figure 1). [2]

**Figure 1: Enterprise Investments for Generative AI Initiatives**

**Enterprise Investments for Generative AI Initiatives**
Percentage of respondents



n = 1,419 (September 2023); 1,299 (January 2024)

Q. How would you characterize your organization's generative AI investments (time, money, resources)?

Sources: Generative AI Realities: Proactive Approaches for Quantifiable Business Results, Webinar Polling, September 2023;
Generative AI Realities: Measuring and Quantifying Business Results, Webinar Polling, January 2024

806282_C

Gartner

Additional polling questions in the same webinar series revealed that nearly two-thirds of organizations are using GenAI across multiple business units. This represents a 19 percentage point jump since September 2023, and 40% of respondents say that GenAI has been deployed in more than three business units.

The primary business functions that have adopted or intended to invest in some form of GenAI solution are customer service (16%), marketing (14%) and sales (12%). The IT function primarily focuses GenAI deployment on the software development life cycle, and infrastructure and operations. The percentage of functions such as finance, HR, procurement and legal that have deployed GenAI is still below 10%. [2]

Here, we have rounded up relevant and up-to-date Gartner research to give answers to the trending questions about GenAI and large language models (LLMs) in five areas: business value and use cases, impact on workforce and organizations, technology, vendors and ecosystem, and security and risk management.

# Research Highlights

*Some recommended content may not be available as part of your current Gartner subscription.*

## Business Value and Use Cases

### How Do Some Organizations Identify, Vet and Fund Suitable GenAI Use Cases?

The savviest organizations are creating an ongoing self-service educational curriculum to build awareness, increase knowledge and foster creation among their staff. This approach feeds a dynamic, iterative process of collecting ideas and use cases in a methodical manner. Targeted multidisciplinary teams then utilize frameworks (e.g., Gartner's AI radars or use-case prisms) to vet and juxtapose the ideas based on business value and feasibility (see **Gartner AI Opportunity Radar: Set Your Enterprise's AI Ambition**). At this stage, the C-suite and technology leaders (responsible for AI, analytics, data, applications, integration or infrastructure) can work together to determine how to vet and fund the various AI initiatives. In order to do this, they look at cost, value and risk. And this is a complex undertaking.

### What Are the Common Use Cases Applicable to My Industry?

Gartner's GenAI prisms define use cases by industry and assess their business value and feasibility to foster strategic planning (see **Use-Case Prisms for Generative AI: A Guide to Emerging Opportunities in Industries**).

Behind the scenes, in order to generate use cases, organizations need to have multidisciplinary teams that do the heavy lifting of defining (and scoring) business value and feasibility. Business value is often captured through the lens of efficiency, effectiveness or business agility for either competitive parity or competitive advantage (see **Identify the Strategic Benefit, Cost and Risk of Generative AI Use-Case Types**). Feasibility may involve a myriad of technology factors, AI-ready data analysis, skills availability, process debt and trust levels.

### What Are GenAI Case Studies in the Real World?

Gartner analysts have organized dozens of case study snapshots that can be filtered by industry, business value, risk, strategy alignment category, results and use-case category. These are short, focused looks at real-life examples to enable executives and their teams with strategic planning (see **Generative AI Case Study Snapshots**). Business and technology leaders looking for ways to take full advantage of the potential of AI can leverage these real-world examples to expand their thinking on the art of the possible and the impact organizations are having with AI.

## How Are Organizations Aligning GenAI Initiatives to Business Goals and Assessing Business Value?

Gartner's framework for GenAI investments falls into three competitive impact categories:

- **Defend** — Incrementalism, marginal gains and microinnovations
- **Extend** — Growth in either market size, reach, revenue or profitability
- **Upend** — Creation of new markets and products

Heads of AI and executive leadership must assess the potential benefits and cost of new GenAI investments. Experimentation can be done inexpensively for most use cases. **How to Calculate Business Value and Cost for Generative AI Use Cases** provides a decision framework for assessing costs and realizing value of enterprise GenAI initiatives.

## How Are Government Policies and Regulations Impacting Business Value?

Across the major regions of the world such as the U.S., Europe and Asia, various government entities have introduced or enacted policies and regulations that will impact organizations and business value.

Gartner has published research to explore possible impacts (see **The Impact of the 'U.S. Executive Order on AI'** and **Quick Answer: The EU AI Act and Its Anticipated Impact**). The common theme, regardless of geographical region, is that business leaders need to either step up or step aside. The overall message is that if you are in an executive leadership role, you have responsibilities related to AI. You must proactively take measures to be compliant and prevent harm. In other words, even if you are not ready for AI, you need to be AI-ready. Executives should adjust leadership priorities, reconcile AI investment with redistributed risk and prepare today for a more regulated tomorrow.

# Impact on Workforce and Organizations

## How Will the Workforce Be Impacted?

Gartner's position is that the near-term impact of GenAI will be to augment targeted activities or tasks, rather than the entire job. Gartner's strategic planning assumption is that by 2026, over 100 million humans will engage robocolleagues (synthetic virtual colleagues) to contribute to enterprise work. In a large majority of cases, job reduction or elimination will be limited for the next two to three years. The primary focus of many organizations is the profile of "productivity pursuers" using everyday AI. A limited number of organizations have embarked on "game-changing AI." The impact that GenAI will make on the workforce will manifest itself on a case-by-case basis. It will vary by industry, geography and organizational complexity. The extent of that impact depends on strategy, execution, risk management, governance, technology choices and the ability to engender trust.

The role that is most frequently inquired about is the leader of the AI strategy and execution, namely, the head of AI (see **Quick Answer: Do Enterprises Need a Chief AI Officer?**). Gartner's position is that most organizations do not need a chief AI officer. However, they do need a leader to orchestrate a holistic or integrated approach to AI and GenAI with multidisciplinary governance. The focus must be on a business strategy that is infused with AI rather than an AI technology roadmap masquerading as a strategy.

## How Will AI Governance Be Impacted?

The application of GenAI is becoming increasingly pervasive, usually with initiatives launched in multiple business units. To coordinate and govern a growing number of AI/GenAI projects and resulting AI/GenAI systems, a best practice is to create an AI center of excellence (COE). The AI COE should enable and coordinate local initiatives, ensuring business alignment, trust, and a balance between scale and governance. It should avoid an overly centralized approach. The AI COE should also work closely together with IT and data and analytics (D&A) teams, both for data access and the embedding of AI in applications.

With AI's business criticality and strategic importance becoming higher, an AI organization should further evolve toward a federated or hybrid structure, composed of central and decentralized teams. To scale AI and increase its impact, a portfolio of AI initiatives should be proactively orchestrated by an AI lead, supported by the execution of an AI strategy and roadmap to further mature organizationwide AI capabilities. See **Become an AI-First Organization: 5 Critical AI Adoption Phases**.

### How Should Organizations Develop a GenAI Governance Model to Manage GenAI Solutions?

Additional policies and guidelines are required to use GenAI responsibly and to manage limitations and risks relating to areas such as trust, fairness, intellectual property and security. The governance of GenAI should be complementary and aligned with existing governance of AI, data, IT and other areas. In addition, it should be compliant with emerging regulations, such as the EU's AI Act, as well as with regional, cultural and ethical values. To be effective, GenAI governance should be implemented through clearly defined roles and responsibilities, procedures, communication, awareness sessions and training. It should be further operationalized through practical guidelines and tool support for the development, deployment and monitoring of AI systems. Leading and coordinating AI governance is typically the responsibility of an AI COE, which is owned by senior or C-level leadership and often supported by an advisory board. See **Tool: Generative AI Policy Template**.

## Technology

### How Are Organizations Combining Enterprise Data With GenAI Solutions?

Most organizations that purchased Bing Chat, ChatGPT, Gemini, Poe, Quora or any other comparable tool and applied it were inevitably asked within days or weeks how to incorporate enterprise data into the tool and manage data access. Gartner research provides multiple examples of approaches for combining enterprise data with GenAI solutions and examines the pros and cons of each approach (see **Quick Answer: How to Supplement Large Language Models With Internal Data**).

### Which Way Can Better Incorporate Enterprise Data Into LLMs, Prompt Engineering or Fine-Tuning?

Prompt engineering (inclusive of retrieval-augmented generation [RAG]) is often the initial option that most organizations try when incorporating internal knowledge/data into LLMs, especially if the data is dynamic or needs access control. Organizations then quickly turn to ask about fine-tuning. The rationale behind fine-tuning is either to add new knowledge to a model or to tweak the model for better alignment, while prompt engineering with RAG architecture can do both with additional token consumption/inference costs and longer execution (retrieval and inference) time.

RAG has its limitations. It will take weeks to implement a prototype, which seems easy. But it's difficult to scale it to a high production level, which will take months. Every step (such as trunking, tokenization, embedding, semantic/keyword search fusion, reranking and the inference of LLM itself) in the RAG architecture pipeline could hurt the accuracy of final answers generated by the LLM. And you can apply RAG on top of fine-tuned models to achieve better results.

See **Prompt Engineering With Enterprise Information for LLMs and GenAI** and **Quick Answer: When to Fine-Tune Large Language Models**.

## How Do I Choose Between Open-Source Models and Proprietary Models?

The key benefits of open-source models include customizability, better control over deployment options, enhanced privacy and security, the ability to leverage collaborative development, model transparency, and the potential to reduce vendor lock-in. Besides general-purpose open-source models, there will be lots of open-source task-specific LLMs that enterprises can choose from.

Some enterprises can leverage cloud infrastructures (infrastructure as a service [IaaS] or via APIs) for open-source model fine-tuning and inference. Other enterprises can choose smaller open-source models, perform lightweight fine-tuning (instruction tuning) and then host them on-premises. In addition, enterprises must consider other factors (see **Quick Answer: What Are the Pros and Cons of Open-Source Generative AI Models?**).

## How Should Enterprises Evaluate Different GenAI Deployment Options?

Enterprises may feel overwhelmed, having numerous GenAI deployment models to choose from. Gartner research explores five different approaches for deploying GenAI and examines the pros and cons of each option (see **How to Choose an Approach for Deploying Generative AI**).

IT leaders responsible for AI solutions should:

- Understand and document the technical differences between different approaches, so that they will not be locked into the approaches prescribed by their vendors. This involves acknowledging that the deployment and ongoing maintenance of a solution is a responsibility shared with the vendor.
- Account for all critical decision factors (costs, the control of model outputs, security and implementation complexity) and make objective decisions on a use-case-by-use-case basis.

The five approaches that Gartner research explores for deploying GenAI are not mutually exclusive. Most organizations may need to combine them and adopt a hybrid version.

## How Should I Control AI Models' Outputs?

AI observability is the ability to monitor and assess the behavior of an AI model to better understand and control its outputs. LLM adoption requires organizations to use guardrail tools to monitor and control AI models. Such tools monitor requests, token usage, toxicity, the readability of prompts, leakages of personally identifiable information (PII), the citation of responses and the evaluation of responses. They can even leverage other LLMs to evaluate existing models. Guardrails are the primary way of implementing content anomaly detection at runtime in LLM applications. Organizations must establish policy and technical mechanisms to improve AI observability. Please note that most control around AI models and observability is limited to internal working mechanisms of LLMs. See **Introduce AI Observability to Supervise Generative AI**.

## What Are the Major GenAI Trends and the Maturity Level of GenAI Technologies?

- In addition to massive GenAI models (such as GPT-4), the future might see more use-case-specific or context-specific models of smaller size.
- Open-source models are rising in prominence and aggressively competing against closed-source ones. With AI-related regulations increasing, customers may favor open-source models which have better deployment flexibility and customizability, and enable better control over security and privacy.
- We will see the advent of more vertical-domain-specific models, particularly in the areas of healthcare, life sciences, financial services and legal services. Most of these models will be

built on top of AI foundation models but with domain-specific data.

- Composite AI will become popular. More solutions will combine GenAI models with other AI techniques to address various business problems that GenAI models cannot solve effectively or efficiently alone.

- AI foundation models will become increasingly multimodal. They will be trained with different types of data (modalities) and be capable of handling more than one modality in their inputs and outputs.

- LLMs/GenAI gained much traction in the AI field in 2023. But most of them are far from mature for enterprises to adopt. How to incorporate enterprise knowledge into LLMs effectively, efficiently and safely remains a challenge. We saw rare cases of large-scale production deployments and deep business adoption by 2023. We can foresee a very fast evolution of technologies such as GenAI marketplaces, multimodal models and autonomous GenAI agents. But it is still too early to combine these technologies or weave them into systems.

See **Hype Cycle for Generative AI, 2023** for more information on GenAI innovations, trends and maturity.

## Vendors and Ecosystem

### How Can I Search for a Suitable GenAI Vendor Solution?

Gartner has developed a vendor identification toolkit that enables organizations to search for suitable vendors by functional areas (such as finance, HR, sales and marketing), industries and use cases (see **Tool: Vendor Identification for Generative AI Technologies**). The purpose of the toolkit is to help organizations search the market in a fast manner for potential off-the-shelf or out-of-the-box options.

### What Are the Key Vendors of Base LLMs and Infrastructure?

Base LLMs are often the backbone of GenAI applications. They can be open-source models or provided by proprietary vendors. Examples of base LLM vendors include Amazon (Titan), Anthropic (Claude), Google (PaLM, Gemini) and OpenAI (GPT-3.5, GPT-4). It should be noted that these vendors are often in partnership with one another, which allows users to access a model from different platforms.

One strategic move that hyperscalers such as Amazon Web Services, Google Cloud Platform and Microsoft Azure have taken is to develop their own proprietary AI-optimized chips to meet developers' growing demand for AI computing and to run their SaaS applications. Typical infrastructure vendors include Amazon, Google, IBM, Microsoft, NVIDIA and Oracle.

See **A Comparison of Generative AI Platform Offerings**.

### What Tools Should Organizations Use to Extend LLMs and Customize LLM Solutions?

Typical tools include:

- Model orchestration tools that help build complex LLM-based solutions by connecting various technical components.
- Vector databases which currently play an important role in the GenAI application architecture by providing retrieval steps of RAG architecture.
- Embedding models that can translate organizational knowledge and user queries into embeddings which vector databases use for semantic search.
- Model observability tools.
- Model cache tools.

See **A CTO's Guide to the Generative AI Technology Landscape**.

## Security and Risk Management

### What Are the Attack Surfaces of LLM Solutions?

The major security concerns about LLMs remain data leakages and prompt injections. Attack surfaces vary depending on how LLMs are consumed. As for applications like ChatGPT, their main attack surfaces are prompts, which can be susceptible to business logic abuse and injections. LLMs' outputs can pose a risk too because they may include malicious links and content. When integrating a third-party LLM by building its orchestration layer (e.g., prompts and RAG), you will see the attack surface expanding, especially regarding the security of API calls. In addition, the Top 10 security risks in deploying and managing LLMs, released by the

Open Worldwide Application Security Project (OWASP), [3] can be used to assess LLM attack surfaces and frame a security discussion.

See **Generative AI Adoption: Top Security Threats, Risks and Mitigations**.

## What Are the Regulatory Risks Concerning LLM Usage?

Depending on their locations and/or jurisdictions of operation, organizations face many potentially different constraints related to their use of LLMs (see **9 Trust and Ethical Implications of Generative AI**). It’s critical to engage with legal specialists before the design, deployment or use of any LLM. Concerns vary widely across jurisdictions, and the effects of forthcoming legislation are yet to be understood. As for general-purpose LLMs provisioned by third parties, end users will find it impossible to control risks concerning where data is processed or sent to, the legitimacy of training data/methods, the reliability and desirability of outputs (e.g., harmful and false information), and transparency of the design, training and functions of a model. Compliance concerns may come from intellectual property (see **Impact of Generative AI on Intellectual Property**), privacy, data protection, and AI-specific technology-focused laws. There may exist a need to ensure that requirements for privacy and confidentiality extend beyond prompt content and training/pretraining data, and should cover logs of user queries, enterprise context data for prompt engineering and training data for fine-tuning.

## How Should I Develop a Company Policy on LLMs?

Define rules based on the user team, the use case and the type of data. Most organizations would formulate a policy with rules for inputs (e.g., no client or private data in ChatGPT), and rules on outputs, especially for the generation of client-facing content (e.g., manual review processes for all generated content expected to go public). See **Tool: Generative AI Policy Kit**.

## Do I Need Additional Security Controls for LLMs?

Yes. Many organizations leverage their secure web gateways or security service edge providers to implement policy controls, security monitoring and sometimes data loss prevention (DLP) features. A splash page, intercepting the traffic before users access the application, is frequently used to show a link to the company policy, and a contact for approving access to ChatGPT. Additionally, organizations track the use of other LLM applications to determine access restriction for unmanaged LLM applications.

See **Innovation Guide for Generative AI in Trust, Risk and Security Management**.

## ⊕ Evidence