Hype Cycle for Generative AI, 2023

Published 11 September 2023 - ID G00795274 - 97 min read

By Analyst(s): Arun Chandrasekaran, Leinar Ramos

Initiatives: Digital Future; Artificial Intelligence; Evolve Technology and Process Capabilities to Support D&A; Generative AI Resource Center

ChatGPT has made generative AI a top priority for the C-suite and has sparked tremendous innovation in new tools beyond foundation models. This inaugural GenAI Hype Cycle guides technology innovation leaders through these fast-moving technologies and markets.

Additional Perspectives

- Invest Implications: Hype Cycle for Generative AI, 2023
  (12 September2023)
- Summary Translation: Hype Cycle for Generative AI, 2023
  (12 October 2023)

Strategic Planning Assumption

By 2026, more than 80% of enterprises will have used generative AI APIs or models, and/or deployed GenAI-enabled applications in production environments, up from less than 5% in 2023.

Analysis

What You Need to Know

We've created this inaugural generative AI (GenAI) Hype Cycle to demystify the core technologies underpinning this transformative trend. The technologies on our Hype Cycle fall into four key areas:

- **GenAI models:** Pretrained AI models are at the core of the GenAI trend. They're incorporating modalities beyond text and becoming tuned to handle a variety of use cases.
- **AI engineering tools:** A growing ecosystem of GenAI tools and techniques enables organizations to build, govern and customize GenAI applications.
- **Applications and use cases:** The huge number of GenAI applications and use cases provides great potential for rapid adoption and business value. But it also forces organizations to rethink GenAI trust, risk and security issues.
- **Enablement techniques and infrastructure:** GenAI exists and will progress because of both novel techniques and several preexisting AI techniques. Meanwhile, specialized infrastructure will accelerate model training and the inference process.

Technology innovation leaders, including CTOs, should use this Hype Cycle to identify GenAI innovations that can be leveraged according to their appetite for risk versus potential rewards. Doing so will help leaders understand this trend's potential for their organizations.

The Hype Cycle

The GenAI technology landscape consists of four key technology areas.

GenAI models: These models are rapidly evolving. Pretrained foundation models, the most popular of which are large language models (LLMs), are at the core of the GenAI wave. They are evolving to become more multimodal and instruction trained to be conversational.

Use-case-specific models are emerging across business functions, and domain-specific GenAI models are appearing in sectors such as healthcare, life sciences, legal, financial services and the public sector.

Model hubs are becoming an important part of the value chain by helping developers navigate the wide range of GenAI models.

Open-source models are becoming a credible alternative to dominant proprietary LLMs due to their customizability, better control over privacy and security, and the ability to leverage their collaborative development (see Quick Answer: What Are the Pros and Cons of Open-Source Generative AI Models? for more details). More energy- and resource-efficient LLMs will appear that can run in local "edge" devices.

However, despite tremendous progress since 2017, we're still a long way from artificial general intelligence (or human-level intelligence).

Assess:

- Artificial general intelligence
- Domain-specific GenAI models
- Edge LLMs
- Foundation models
- Large language models
- Model hubs
- Multimodal GenAI
- Open-source LLMs

AI engineering tools: A growing ecosystem of GenAI tools and techniques helps organizations build, govern and customize GenAI-powered applications. The growing ecosystem of tools will become increasingly important as most enterprises are gearing up to deploy multiple models (from multiple providers) within their environments.

Prompt engineering, retrieval augmented generation and reinforcement learning from human feedback are key techniques for training and using GenAI models. Other emerging technologies, such as GenAI application orchestration frameworks and LangOps, are becoming more important for embedding GenAI into broader systems.

The underlying data infrastructure for GenAI is also adapting with the emergence of specialized databases such as vector databases.

Explore:

- GenAI application orchestration frameworks
- LangOps
- Prompt engineering
- Reinforcement learning from human feedback
- Retrieval augmented generation
- Vector databases

Applications and use cases: GenAI-enabled virtual assistants, such as ChatGPT, have attracted much attention, but a huge number of GenAI applications and use cases go even further. GenAI models and APIs are increasingly embedded into many enterprise applications (generative AI-enabled applications) and organizations in many domains are using them successfully. Their strongest growth is in software engineering (AI-augmented software engineering).

The massive growth in GenAI use cases and applications is forcing organizations to rethink AI trust, risk and security management (AI TRiSM) for GenAI.

Examine:

- AI-augmented software engineering
- AI TRiSM
- Autonomous agents
- GenAI-enabled virtual assistants
- Generative AI-enabled applications
- Synthetic data

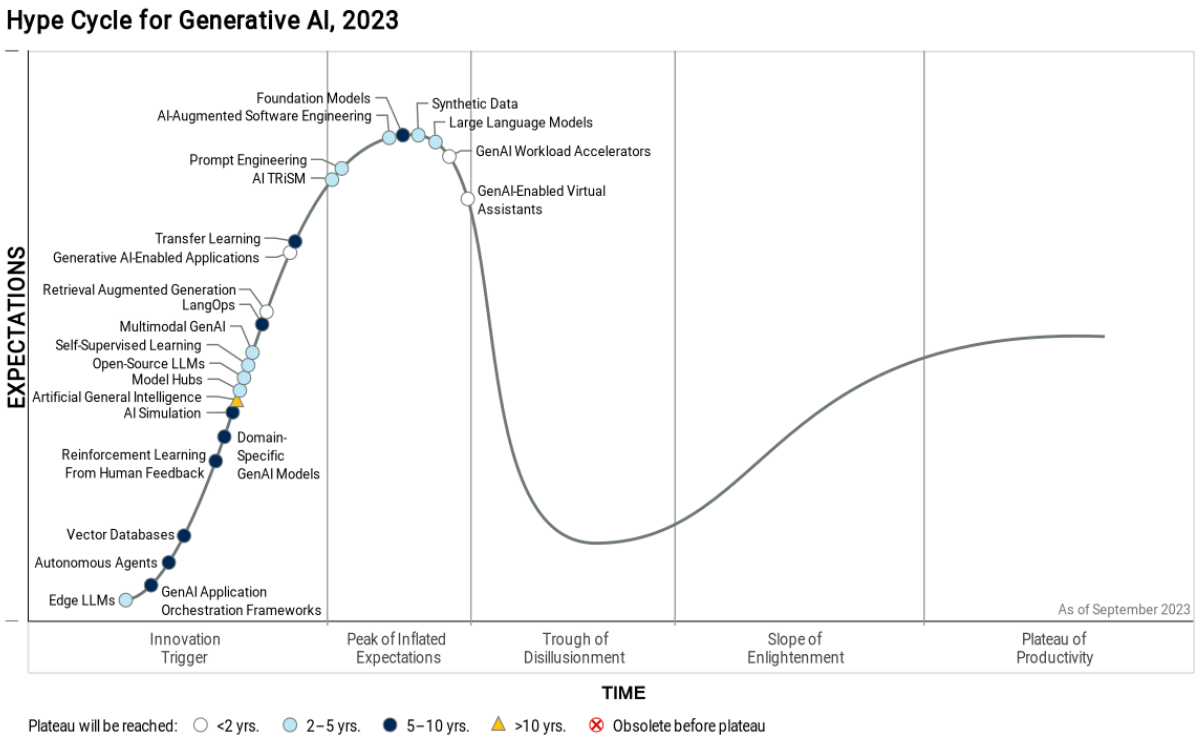
Enablement techniques and infrastructure: GenAI builds on several preexisting AI techniques. AI simulation is an increasingly important technique for generating complex-world models and environments that can be used to train AI agents and generate synthetic data. Similarly, most AI foundation models are trained using self-supervised learning techniques on vast amounts of internet data, while transfer learning is an extremely valuable technique for using foundation models as an advanced starting point for new tasks and use cases. These underlying techniques will continue to enable GenAI progress.

Specialized infrastructure — such as generative AI workload accelerators — will play an important role in model training and the inference process. Most models will be trained and run in the cloud, but as more open-source models become available, clients may choose to train or run these models outside cloud environments. Demand will continue to grow for specialized AI chips and tools to run them efficiently and at lower cost.

Evaluate:

- AI simulation
- Generative AI workload accelerators
- Self-supervised learning
- Transfer learning

Figure 1: Hype Cycle for Generative AI, 2023



The Priority Matrix

When compared to other Hype Cycles, the Generative AI Hype Cycle has almost all innovations in the Innovation Trigger or Peak of Inflated Expectations, which is characteristic of an early-stage market. In addition, most technologies tend to have transformational or high benefit ratings as well as a faster path to the Trough of Disillusionment and beyond.

The innovations that deserve particular attention within the two- to five-year period to mainstream adoption include multimodal GenAI, open-source LLMs and edge LLMs. Early adoption of these innovations can lead to significant competitive advantage and time-to-market benefits.

Several innovations have a five- to 10-year period to mainstream adoption, and from these innovations, domain-specific GenAI models, reinforcement learning from human feedback and autonomous agents offer the highest potential.

Table 1: Priority Matrix for Generative AI, 2023

(Enlarged table in Appendix)

Benefit ↓	Years to Mainstream Adoption: Less Than 2 Years ↓	2 - 5 Years ↓	5 - 10 Years ↓	More Than 10 Years ↓
Transformational		AI-Augmented Software Engineering Large Language Models Multimodal GenAI Self-Supervised Learning	Autonomous Agents Foundation Models	Artificial General Intelligence
High	GenAI-Enabled Virtual Assistants GenAI Workload Accelerators Generative AI-Enabled Applications Retrieval Augmented Generation	AI TRiSM Edge LLMs Model Hubs Open-Source LLMs Prompt Engineering Synthetic Data	AI Simulation Domain-Specific GenAI Models GenAI Application Orchestration Frameworks LangOps Reinforcement Learning From Human Feedback Transfer Learning Vector Databases	
Moderate				
Low				

Source: Gartner (September 2023)

On the Rise

Edge LLMs

Analysis By: Ray Valdes, Leinar Ramos, Tyler Bray

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

An edge LLM is a large language model (LLM) designed to be deployed at the edge of a network. Edge LLMs strive to bring the capabilities of full LLMs to the edge, balancing trade-offs of size, computation and performance to fit the resource constraints of edge deployments. Target devices include smartphones, IoT devices, edge gateways and robots. Edge LLMs are more likely to be based on open-source models, because of the need for modifications for resource-constrained environments.

Why This Is Important

Edge LLMs are still emerging, but potentially can reshape our interactions with computers by adding intelligence at the edge, in smartphones, mobile devices, field robots and IoT devices in buildings, vehicles, factories and other locales. They can reduce latency, save bandwidth and add resiliency in case of connectivity failure. Moreover, they can address enterprise data privacy concerns and boost scalability through controlled local deployments.

Business Impact

Because edge LLMs are still embryonic, their business impact is yet to be proven. As they mature, they will have an impact in many areas. They have the potential to stimulate market growth, deliver productivity improvements, save costs, offer a competitive edge, enhance user interactions, ensure data security and unlock novel device capabilities. They can enable natural interactions on devices such as cars, smartphones, factory machinery, thermostats, field robotics and construction equipment.

Drivers

- **Privacy concerns about cloud LLMs:** Many enterprises are concerned about data privacy compromise in cloud LLMs. This concern fuels interest in running LLMs locally (on-premises), which, in terms of privacy, can be viewed in the same category as edge LLMs. This is not to say that edge LLMs and local LLMs are immune from data leakage. But current perception is that noncloud LLMs provide greater control and lockdown over high-value confidential data.
- **Open-source LLM wave:** The availability of powerful open-source LLMs, such as Llama 2, has fueled activity among vendors, entrepreneurs and investors in new deployment scenarios (edge and local LLMs). The massive investment required to build and deploy general-purpose cloud LLMs has sparked exploration of alternative deployment approaches. Open-source LLMs provide a flexible technology foundation for these efforts.
- **Inference cost reduction:** LLMs can be run on a smartphone or laptop for little added cost, unlike cloud LLMs, where every word fragment (token) is metered.
- **Know-how about resource minimization:** In parallel with open-source LLMs, there is increased know-how regarding resource minimization and cost-efficiency. Different techniques continue to be discovered for efficient training and tuning, reduced memory consumption, faster model loading, and so on.
- **AI computational resources at the edge:** In 2017, Apple added a Neural Engine computational capability to the iPhone. Huawei added an NPU, and Google defined a neural net API for Android devices. These capabilities are improving over time.
- **Need for disconnected LLM:** There are still locations where internet connectivity is not available, such as in remote locations or deep inside a building. Edge LLMs can work in disconnected mode.
- **Need for reduced latency:** Edge LLMs can obviate the task of sending data to the cloud for processing and waiting for a token-by-token response. With sufficient local hardware, responsiveness can be better than cloud LLMs.

Obstacles

- **Resource-constrained devices:** Many edge devices have minimal computing resources. Evolving techniques can address these constraints, but they have trade-offs. Also, specialized skills for low-resource computing remain rare in the workforce.

- **Reduced accuracy:** To run LLMs on edge devices, models must be downsized in various ways. Often, the outcome is reduced model performance and low quality of results. Today, an edge LLM on a Raspberry Pi makes for a good conversation piece but is not useful in real-world use cases.
- **Data privacy and security:** Running LLMs on edge devices means the entire dataset is exposed: the model, weights, prompts and context. This situation is similar to a company laptop. Standard techniques to secure the company laptop can be layered on top of the edge LLM, but these are not yet widely adopted.
- **Edge heterogeneity:** At the edge can be a proliferation of device types, platforms and communications protocols. This can make manageability and reliability challenging.

User Recommendations

- To use edge LLMs effectively, product leaders must address key areas such as data management, model optimization, testing, security, monitoring and logging, and governance and compliance. Product leaders must balance trade-offs of size, computational efficiency and performance to fit the constraints of edge deployments.
- Product leaders must track the evolving landscape and be prepared to change strategies as events unfold. The space is rapidly evolving, with major recent developments in enabling technology, such as Llama 2. The pace of change will continue through the near term.
- Product leaders should consider open-source LLMs as a flexible and customizable foundation for edge LLMs, but not rule out the hyperscale vendors of cloud LLMs that have ample resources to address a diverse set of use cases, possibly including edge LLMs.

Sample Vendors

GPT4All; Meta; OpenOrca; Qualcomm; Replicate

Gartner Recommended Reading

Innovation Insight for Edge AI

Emerging Tech: Aligning Benefits to Use Cases and Industry Sectors Is Key to Selling the Value of Edge AI

Deploy Leaner AI at the Edge: Comparing Three Architecture Patterns to Enable Edge AI

What Are the Pros and Cons of Open-Source Generative Models?

GenAI Application Orchestration Frameworks

Analysis By: Arun Chandrasekaran

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Generative AI (GenAI) application orchestration 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. Also, these frameworks provide templates for developing new GenAI applications.

Why This Is Important

GenAI application orchestration frameworks provide functionalities that expand AI foundation models' capabilities, making them more adaptable, interactive, context-aware and efficient in various applications. GenAI application orchestration frameworks manage workflows by chaining prompts and models together to achieve business outcomes. Also, these frameworks enable effective prompting through prompt templates, input prompt optimization and output parsing.

Business Impact

GenAI application orchestration frameworks can enable composable and extensible applications. Machine learning (ML) developers can easily swap out components and customize chains to meet their specific needs, thus enhancing overall flexibility for workloads. IT leaders can reduce tooling fragmentation by creating a centralized ML platform to help enterprises integrate data sources with models and chain prompts. This GenAI framework can also help standardize AI engineering tools and scale applications across the organization.

Drivers

- **Use-case maturity:** As use cases evolve, enterprises increasingly deploy multiple models across various use cases. These models may have to be chained together, or the prompts need to be chained — which these frameworks orchestrate.
- **Need for automation:** While AI foundation models are a great advance, they still require extensive human intervention (in terms of prompts) to achieve business outcomes. GenAI application orchestration frameworks enable developers to build agents to reason about problems and break them into smaller, executable subtasks. For example, with LangChain, developers can introduce context and memory into completions by creating intermediate steps and chaining commands.
- **Generative AI application development:** Developers are keen to capitalize on GenAI models to build applications. GenAI application orchestration frameworks provide developers with a new way to build user interfaces and automate application builds.
- **Model customization:** GenAI models can be combined with enterprise data through advanced prompt engineering techniques or model fine-tuning. GenAI application orchestration frameworks provide an open approach to integrating data sources with AI models.

Obstacles

- **Lack of awareness:** These tools are new, and there is a lack of understanding of what these tools do, which one to use and how to safely deploy them.
- **Immaturity:** Most tools are open source, with limited commercial support. Therefore, enterprise ML teams must cautiously navigate software bugs and security issues, characteristic of most early-stage open-source software (OSS) projects.
- **Lack of clear winners:** While these orchestration frameworks strive to bridge the GenAI application development, their longevity and ability to innovate iteratively are currently unknown.

User Recommendations

- Encourage experimentation on what these tools do and their potential fit in your technical architecture.
- Identify the use cases where you are implementing data retrieval or fine-tuning, which can benefit from using these tools.
- Explore the autonomous agent capabilities of these tools cautiously, given the black-box nature of machine-to-machine interaction.
- Take a centralized platform approach to achieve standardization and automation across the GenAI applications you are building.

Sample Vendors

deepset; Dust; LangChain; LlamaIndex; Microsoft

Gartner Recommended Reading

Innovation Guide for Generative AI Technologies

Autonomous Agents

Analysis By: Christian Stephan

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Embryonic

Definition:

Autonomous agents are combined systems that achieve defined goals without human intervention. They use a variety of AI techniques to identify patterns in their environment, make decisions, invoke a sequence of actions and generate outputs. These agents have the potential to learn from their environment and improve over time, enabling them to handle complex tasks.

Why This Is Important

Autonomous agents represent a significant shift in AI capabilities. By their independent operation and decision capabilities, they can improve business operations, enhance customer experiences and enable new products and services. On the other hand, tech executives need to manage new challenges in transparency, ethics and workforce adoption. The early stage is inflating expectations, despite most agents providing limited use, but the fast development in this area demands proper observation.

Business Impact

Business operations will be significantly enhanced with autonomous agents by boosting efficiency through automating complex tasks, improving organizational productivity. They will also enhance customer experience with 24/7 personalized service via conversational agents empowered with robotic process automation (RPA) capabilities to trigger organization processes. This will probably come with cost savings, granting a competitive edge. It also poses an organizational shift of workforce from delivery into supervision.

Drivers

- **Investment and progress in AI development:** Rapid advancements in AI, particularly in the reasoning ability of large language models (LLMs) and reinforcement learning, are also favoring autonomous agent development.
- **Transfer learning and adaptive AI:** Applying learned knowledge to new tasks or adapt to changing environments will enhance the versatility and effectiveness of autonomous agents.
- **Multimodality:** The ability to process and integrate multiple types of data (text, images, audio) enables autonomous agents to handle more complex tasks and provide richer interactions. Decision-level fusion of modality is simpler to integrate in the action sequence of autonomous agents.
- **Digitalization of organizations:** By digitizing assets and processes for effective operations, companies have generated data and automation that autonomous agents can connect to in order to perform complex tasks in the environment.
- **Demand for personalization:** In an era of increasing customer expectations on personalized experiences and high availability, autonomous agents can deliver such extended services at scale.
- **Promise on cost savings:** Setting up an autonomous infrastructure always bears opportunity for savings on human labor that could probably be replaced by IT infrastructure.

Obstacles

- **AI trust, risk and security management (TRiSM) challenges:** Autonomous agents most likely will have access to sensitive information and critical infrastructure and need effective protection. Transparent and explainable decision making is mandatory.
- **Lack of human supervision:** Without a human in the loop, the potential to mitigate AI errors is reduced.
- **Upcoming regulations:** Agency, in particular, is a much discussed and debated policy issue. Early regulatory proposals point to strict regulations and liabilities for autonomous actors.
- **Organizational resistance:** The fear of replacement will expose autonomous agents to massive resistance within the workforce.
- **Multimodal understanding:** While late fusion of modality seems easier to integrate, it misses the complex interactions between different types of data compared to an early feature-level fusion.

User Recommendations

- **Observe and experiment:** The technology is immature and needs to evolve. In some niches, like knowledge management, document processing and actionable chatbots, autonomous agents are more likely than in others. Search for signals that point toward your preferred use cases. Be open about experimentation, but don't expect ROI on this.
- **Develop a data strategy:** Autonomous agents will require high-quality data to function effectively. Develop and implement strategies for data collection, cleaning, management and privacy for your organization.
- **Invest in AI and data literacy:** Invest in training your workforce. This will help to understand the potential and limitations of autonomous agents and how to work with them effectively.

Sample Vendors

Adept; AgentGPT; AutoGPT; Amazon Web Services; BabyAGI; Cognosys; Generally Intelligent; Inflection AI; LangChain; MULTI ON

Gartner Recommended Reading

Autonomous Things: Technology Use Cases for R&D

Emerging Tech: Tech Innovators in Tabular Synthetic Data — Domain-Focused

2023 Utility Trend: Establish Decision Intelligence Before Chasing Autonomous Business

Vector Databases

Analysis By: Arun Chandrasekaran, Radu Miclaus

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Vector databases store numerical representations of data. In such databases, each point is represented by a vector with a fixed number of dimensions, which can be compared via mathematical operations, such as distance measures. Vector databases are commonly used in machine learning (ML) solutions, where vectors represent data features/attributes, such as text embeddings. Storing these vectors in a database enables users to search for similar data points with low latency.

Why This Is Important

Vector databases serve such use cases as similarity search and product recommendation. Rapid innovation in generative AI and adoption of AI foundation models have spawned interest in vector databases. When customers adopt generative AI models, vector databases store the embeddings that result from the model training. Storing vector embeddings representing the model training, the database can do a similarity search, which matches a prompt (the question) with specific or similar vector embedding.

Business Impact

Businesses thrive by delivering differentiated customer experience (CX). Generative AI is increasingly embedded in applications to empower the human-machine symbiosis, and organizations need scalable and accelerated ways to build and support these applications long term. Vector databases are an important back-end service that allows businesses to future-proof and scale their generative-AI-enabled applications. These drive business value through customer engagement and adoption.

Drivers

- **Popularity of vector embeddings:** With the rise of AI foundational models, embeddings have become the cornerstone for semantic search. Hence, they are the working inputs for training large foundation models.
- **Performance and scalability needs:** The applications looking to embed generative methods that use embeddings-based models need back-end services that can respond with low latency to high concurrency requests (prompts) and responses (completions) for generative AI use cases.
- **Service architecture:** Because most applications are built on service-based architectures, vector databases are ideally presented to applications as services that communicate with the interface via APIs.
- **Hybrid implementation of retrieval and generative models:** Vector databases are optimal for both semantic search (retrieval based on vector similarity) and generative inference through foundation models. This hybrid combination of models drives the need for optimized vector databases, because both generative and retrieval are used together for grounding of facts.
- **Developer focus:** Developers of new applications are driving the demand for vector databases by presenting use cases that cannot scale without the ability for embeddings to be stored in an optimized structure for high-throughput production applications.

Obstacles

- Enterprises lack an understanding of what vector databases do and the unique use cases they enable.
- Vector databases are superspecialized databases that may cause challenges around data migration and integration and limited extensibility across use cases.
- Most vector databases are delivered as cloud-managed service — the complexity of deploying, configuring and operating them outside cloud environments requires deep technical skills and know-how.
- Vector databases can be expensive to implement, given the newness of the technology and lack of industry skills to deploy and manage it.
- The vector database market is nascent and populated mostly by startups, which may not have extensive experience working with enterprise clients, as well as unproven product market fit.

User Recommendations

- Determine whether your functional requirements can be satisfied by incumbent vendors that can support the storage and retrieval of vector embeddings — you may not always need a purpose-built vector database.
- Prioritize developer experience, ecosystem integration, use case fit, reliability and performance as important selection criteria, and validate them thoroughly via a POC process.
- Select managed, cloud-based vector databases as deployment modes, unless you have stringent requirements and deep technical skills for an on-premises, self-managed deployment mode.
- Conduct internal training and education on the appropriate use cases for vector databases, how to leverage their true potential, and effective ways to optimize their deployment and maximize their value.

Sample Vendors

Couchbase; Croma; Elastic; Google; Pinecone Systems; Qdrant; Redis; Weaviate; Zilliz

Gartner Recommended Reading

Innovation Insight for Artificial Intelligence Foundation Models

Quick Answer: What Is GPT-4?

Executive Pulse: AI Investment Gets a Boost From ChatGPT Hype

How Large Language Models and Knowledge Graphs Can Transform Enterprise Search

Reinforcement Learning From Human Feedback

Analysis By: Jasleen Kaur Sindhu, Wellington Holbrook

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Reinforcement learning from human feedback (RLHF) is a machine learning technique that combines automated and human guidance to train an AI agent or a model. It includes training a reward model based on human feedback (positive and negative feedback), then using this model to optimize the AI agent to better understand natural language, learn from human experience and better align with human preference.

Why This Is Important

RLHF is an important technique to steer a self-supervised learning model with human preferences to increase the safety and usefulness of the foundation model. Human language preferences are often dynamic and vary across segments, regions and generations. RLHF ensures foundation models are trained and aligned with these dynamic human preferences and, as a result, are more accurate and capable to execute specific tasks, such as humanlike language generation and text classification.

Business Impact

RLHF allows LLM-based models to learn from human behavior and respond with greater adaptability, enhanced logical reasoning and more reliable decision-making. It reinforces trust and credibility by ensuring responses are accurate, appropriate and nonoffensive. RLHF-trained LLMs are finding widespread adoption in industries such as education, healthcare, gaming, entertainment and financial services, and in use cases that allow nontechnical employees and consumers to directly interact with AI.

Drivers

- Increased attention and rapid growth of LLM-based models and generative AI (GenAI) tools such as ChatGPT will create environments where humans and AI agents interact in real time, driving the need for humans to guide the learning process.
- Increased need to improve content appropriateness and reduce discriminatory or offensive content to support widespread adoption of GenAI-based tools.
- Growing realization of a possibility to create human-like intelligence by bridging the gap between AI-based machine intelligence and advanced human cognitive intelligence.
- Greater appetite to democratize access to AI-embedded tools and solutions that enable nontechnical employees and consumers to directly interact with these interfaces with ease.
- Continued interest from the data scientist community and GenAI solution providers to use RLHF to improve performance of the foundation models, improve explainability and reduce bias present in initial training data.
- Growing available RLHF services from commercial vendors that allow firms to adopt and utilize RLHF for training their foundation models.

Obstacles

- Gathering and processing high-quality human feedback can be very time-consuming and expensive for many firms, and can limit the scalability and wider application of RLHF.
- While human feedback allows the LLM models to better understand natural language and perform in environments where it is difficult to specify the reward, it also makes the model susceptible to human biases and errors. The LLM-based model may learn to generate information that may be offensive or discriminatory to certain demographics.
- It can be difficult to integrate and reconcile feedback from multiple sources. Human feedback can vary depending on the task, cultural background and individual preferences. Especially conflicting human feedback may cause significant learning problems.
- LLM models may continue to demonstrate undesirable behaviors that might not be captured by human feedback, or may exploit loopholes in the feedback process to achieve higher rewards.

User Recommendations

- Apply RLHF-trained models where business outcomes and constraints are hard to quantify or the ability to sound human is particularly important — for example, personal assistants in healthcare.
- Apply RLHF-trained models where there is a greater risk of incorrect or inappropriate responses that may negatively impact your business (for example, a customer-facing application), or use cases with a high risk of biases toward specific customer demographics.
- Consider RLHF for verification and accuracy of information in use cases that require understanding and generation of natural language, such as conversational agents and text summarization.
- Leverage off-the-shelf RLHF services from vendors in the market, but assess suitability and the level of domain and technical skills of the human talent that may be critical for the industry.
- Conduct regular assessment of the output generated to ensure human biases and errors have not negatively impacted the underlying model behavior.

Sample Vendors

Appen; Argilla; Clickworker; Invisible AI; Prolific; Scale AI; Surge AI; Toloka AI

Gartner Recommended Reading

How to Pilot Generative AI

Assess the Value and Cost of Generative AI With New Investment Criteria

Design and Implement Human-in-the-Loop Interfaces for Control, Performance and Transparency of AI

Domain-Specific GenAI Models

Analysis By: Jim Hare, Leinar Ramos

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

Domain-specific models are generative AI (GenAI) models that have been optimized for the needs of specific industries, business functions, or tasks. They are aimed at improving performance and reducing the need for advanced prompt engineering, compared with general-purpose models, for a narrower set of use cases. They can be built from scratch or fine-tuned from existing general-purpose models.

Why This Is Important

While general-purpose models perform well across a broad set of applications, they may be impractical for many enterprise use cases that require domain-specific data. Domain-specific GenAI models can improve use case alignment within the enterprise while delivering improved accuracy and better contextualized answers, reducing the need for advanced prompt engineering. Through more targeted training, these models have the potential to lower hallucination risks associated with large models.

Business Impact

Domain-specific GenAI models can achieve:

- Faster time to value for AI projects by providing a more advanced starting point for industry-specific or use-case-specific tasks.
- Improved performance by a reduction in inaccuracies and hallucinations, compared with general-purpose models, as they are trained with more relevant domain-specific data and for more targeted domain tasks.
- Broader applicability of GenAI models, as organizations will be able to apply them to enterprise use cases where general-purpose models built using consumer data are not performant enough.

Drivers

- **The proliferation of open-source foundation models.** The increased availability of high-performing and commercially usable open-source large language models makes it easier to build domain-specific models. Open-source models as a foundation can be fine-tuned with domain-specific data or further trained for domain-specific tasks.
- **Increased specialization in GenAI model use.** As GenAI adoption in organizations moves from general use cases to more specific ones, targeted models that have been trained with industry-specific data and for domain-specific use cases will become more important.
- **Growing industry use cases.** Demand is increasing for GenAI in many industries, such as healthcare, life sciences, legal, financial services and the public sector. Technology service providers are looking to meet this increased demand by offering products and services tailored to the needs of these industries, including domain-specific models.
- **Limitations of general-purpose models.** General-purpose models often require significant prompt engineering and fine-tuning effort to optimize the output for enterprise use cases since the models haven't been trained using industry or business function data. Domain-specific models can be more easily adapted to specific requirements, utilize available resources more efficiently, and deliver powerful language processing capabilities that address specific challenges and tasks.

Obstacles

- **Reduced model versatility.** By optimizing for a narrow set of use cases, domain-specific models sacrifice versatility, losing the ability to perform well across a wider set of tasks and domains, compared with general-purpose models.
- **Model proliferation.** An organization might need different domain-specific models for different use cases, and thus have an increased portfolio of models to manage. Many organizations might not be ready to operationalize many different models at scale.
- **Persistent inaccuracies and hallucinations.** Even though inaccuracies might reduce in the target domain, compared with general-purpose models, domain-specific models can still make mistakes and hallucinate.

User Recommendations

- Consider domain-specific models for situations that require accurate and contextually appropriate outputs within a specific industry or business function.
- Look for off-the-shelf domain-specific models for your specific use case before deciding to build your own custom domain model. The development of custom domain models poses resource-related issues, predominantly in terms of computational resources and specialized skills.
- Ask domain-specific model providers about the specific use cases for their model along with the pedigree, amount and quality of the data used to train the model.

Sample Vendors

Amazon Web Services; AI4Finance; Bloomberg; Google; Meta; Microsoft; Salesforce

Gartner Recommended Reading

Innovation Insight for Artificial Intelligence Foundation Models

AI Design Patterns for Large Language Models

Accelerate Adoption of Generative AI by Offering an FMOps- or a Domain-Specific Partner Ecosystem

Prompt Engineering With Enterprise Information for LLMs and GenAI

AI Simulation

Analysis By: Leinar Ramos, Anthony Mullen, Pieter den Hamer, Jim Hare

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

AI simulation is the combined application of AI and simulation technologies to jointly develop AI agents and the simulated environments in which they can be trained, tested and sometimes deployed. It includes both the use of AI to make simulations more efficient and useful, and the use of a wide range of simulation models to develop more versatile and adaptive AI systems.

Why This Is Important

Increased complexity in decision making is driving demand for both AI and simulation. However, current AI faces challenges, as it is brittle to change and requires a lot of data. Conversely, realistic simulations can be expensive and difficult to build and run. To resolve these challenges, a growing approach is to combine AI and simulation: Simulation is used to make AI more robust and compensate for a lack of training data, and AI is used to make simulations more efficient and realistic.

Business Impact

AI simulation can bring:

- Increased AI value by broadening its use to cases where data is scarce, using simulation to generate synthetic data (for example, robotics and self-driving cars)
- Greater efficiency by leveraging AI to decrease the time and cost to create and use complex and realistic simulations
- Greater robustness and adaptability by using simulation to generate diverse scenarios to increase AI performance in uncertain environments
- Decreased technical debt by reusing simulation environments to train future AI models

Drivers

- **Limited availability of AI training data is increasing the need for synthetic data techniques, such as simulation.** Simulation is uniquely positioned among synthetic data alternatives in its ability to generate diverse datasets that are not constrained by a fixed "seed" dataset to generate synthetic data from.
- **Advances in capabilities are making simulation increasingly useful for AI.** Simulation capabilities have been rapidly improving, driven both by increased computing performance and more efficient techniques. This has made simulation environments a key part of the training pipeline of some of the most advanced real-world AI use cases, such as robotics and self-driving cars.
- **The growing complexity of decision making is increasing the interest in AI simulation.** Simulation is able to generate diverse "corner case" scenarios that do not appear frequently in real-world data, but that are still crucial to train and test AI to perform well on uncertain environments. As the complexity of the environments and decision making goes up, the ability to build AI systems that are robust becomes more important.
- **Increased technical debt in AI is driving the need for the reusable environments that simulation provides.** Current AI focuses on building short-lived AI models with limited reuse, accumulating technical debt. Organizations will increasingly deploy hundreds of AI models, which requires a shift in focus toward building persistent, reusable environments where many AI models can be trained, customized and validated. Simulation environments are ideal since they are reusable, scalable, and enable the training of many AI models at once.
- **The growing sophistication of simulation drives the use of AI to make it more efficient.** Modern simulations are resource-intensive. This is driving the use of AI to accelerate simulation, typically by employing AI models that can replace parts of the simulation without running resource-intensive step-by-step numerical computations.

Obstacles

- **Gap between simulation and reality:** Simulations can only emulate — not fully replicate — real-world systems. This gap will reduce as simulation capabilities improve, but it will remain a key factor. Given this gap, AI models trained in simulation might not have the same performance once they are deployed: differences in the simulation training dataset versus real-world data can impact models' accuracy.
- **Complexity of AI simulation pipelines:** The combination of AI and simulation techniques can result in more complex pipelines that are harder to test, validate, maintain and troubleshoot.
- **Limited readiness to adopt AI simulation:** A lack of awareness among AI practitioners about leveraging simulation capabilities can prevent organizations from implementing an AI simulation approach.
- **Fragmented vendor market:** The AI and simulation markets are fragmented, with few vendors offering combined AI simulation solutions, potentially slowing down the deployment of this capability.

User Recommendations

- Complement AI with simulation to optimize business decision making or to overcome a lack of real-world data by offering a simulated environment for synthetic data generation or reinforcement learning.
- Complement simulation with AI by applying deep learning to accelerate simulation and generative AI to augment simulation.
- Create synergies between AI and simulation teams, projects and solutions to enable a next generation of more adaptive solutions for ever-more complex use cases. Incrementally build a common foundation of more generalized and complementary models that are reused across different use cases, business circumstances and ecosystems.
- Prepare for the combined use of AI, simulation and other relevant techniques, such as graphs, natural language processing or geospatial analytics, by prioritizing vendors that offer platforms that integrate different AI techniques (composite AI), as well as simulation.

Sample Vendors

Altair; Ansys; Cosmo Tech; Epic Games; MathWorks; Microsoft; NVIDIA; Rockwell Automation; The AnyLogic Company; Unity

Gartner Recommended Reading

Innovation Insight: AI Simulation

Predicts 2023: Simulation Combined With Advanced AI Techniques Will Drive Future AI Investments

Cool Vendors in Simulation for AI

Artificial General Intelligence

Analysis By: Pieter den Hamer

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Artificial general intelligence (AGI) is the (currently hypothetical) intelligence of a machine that can accomplish any intellectual task that a human can perform. AGI is a trait attributed to future autonomous AI agents that can achieve goals in a wide range of real or virtual environments at least as effectively as humans can. AGI is also called “strong AI.”

Why This Is Important

As AI becomes more sophisticated and powerful, with recent great advances in generative AI in particular, a growing group of people see AGI as no longer purely hypothetical. Improving our understanding of at least the concept of AGI is critical for steering and regulating AI’s further evolution. It is also important to manage realistic expectations and to avoid prematurely anthropomorphizing AI. However, should AGI become real, its impact on the economy, (geo)politics, culture and society cannot be underestimated.

Business Impact

In the short term, organizations must know that hype about AGI exists today among many stakeholders, stoking fears and unrealistic expectations about current AI's true capabilities. This AGI anticipation is already accelerating the emergence of more AI regulations and affects people's trust and willingness to apply AI today. In the long term, AI continues to grow in power and, with or without AGI, will increasingly impact organizations, including the advent of machine customers and autonomous business.

Drivers

- Recent great advances in applications of generative AI and the use of foundation models and large language or multimodal models drive considerable hype about AGI. These advances have been enabled largely by the massive scaling of deep learning, as well as by the availability of huge amounts of data and compute power. To further evolve AI toward AGI, however, current AI will need to be complemented by other (partially new) approaches, such as knowledge graphs, multiagent systems, simulations, evolutionary algorithms, causal AI, composite AI and likely other innovations yet unknown.
- Vendors such as Google, IBM, NNAISENSE, OpenAI and Vicarious are actively researching the field of AGI.
- Humans' innate desire to set lofty goals is also a major driver for AGI. At one point in history, humans wanted to fly by mimicking bird flight. Today, airplane travel is a reality. The inquisitiveness of the human mind, taking inspiration from nature and from itself, is not going to fizzle out.
- People's tendency to anthropomorphize nonliving entities also applies to AI-powered machines. This has been fueled by the humanlike responses of ChatGPT and similar AI, as well as AI being able to pass several higher-level education exams. In addition, more complex AI systems display behavior that has not been explicitly programmed. Among other reasons, this results from the dynamic interactions between many system components. As a result, AI is increasingly attributed with humanlike characteristics, such as understanding. Although many philosophers, neuropsychologists and other scientists consider this attribution as going too far or being highly uncertain, it has created a sense that AGI is within reach or at least is getting closer. In turn, this has triggered massive media attention, several calls for regulation to manage the risks of AGI and a great appetite to invest in AI for economic, societal and geopolitical reasons.

Obstacles

- The current issues regarding unreliability, hallucinations, lack of transparency and lack of reasoning or logic capabilities in generative AI-powered chatbots (one possible direction toward AGI), are not easy to overcome with the intrinsically probabilistic approach of deep learning. More data or more compute power for ever bigger models are unlikely to resolve these issues. Better or curated training data, improved prompt interpretation and engineering or more domain-specific foundation models may help to improve reliability, but not sufficiently.
- There is little scientific consensus about what “intelligence” and related terminology like “understanding” actually mean, let alone how AGI should be exactly defined and interpreted. Flamboyant representations of AGI in science fiction create a disconnect from reality. Scientific understanding about human intelligence is still challenged by the enormous complexity of the human brain and mind. Several breakthrough discoveries are still needed before human intelligence is properly understood at last. This in turn is foundational to the “design” or at least validation of AGI, even when AGI will emerge in a nonhuman, nonbrainlike form. Moreover, once AGI is understood and designed, further technological innovations will likely be needed to actually implement AGI. For these reasons, strong AI is unlikely to emerge in the near future. This may be sooner if one would settle for a more narrow, watered-down version of AGI in which AI is able to perform not all but only a few tasks at the same level as humans. This would no longer really be AGI as defined here.
- If AGI materializes, it is likely to lead to the emergence of autonomous actors that, in time, will be attributed with full self-learning, agency, identity and perhaps even morality. This will open up a bevy of legal rights of AI and trigger profound ethical and even religious discussions. Moreover, the (anticipated) emergence of AGI and the risk of human life being negatively impacted by AGI, from job losses to a new, AI-triggered arms race and more, may lead to a serious backlash and possibly regulatory bans on the development of AGI.
- The anticipated possible emergence of AGI urges governments to take measures before its risks can no longer be mitigated. Regulations to ban or control AGI are likely to emerge in the near future.

User Recommendations

- Today, people may be either overly concerned about future AI replacing humanity or overly excited about current AI's capabilities and impact on business. Both cases will hamper a realistic and effective approach toward using AI today. To mitigate this risk, engage with stakeholders to address their concerns and create or maintain realistic expectations.
- Stay apprised of scientific and innovative breakthroughs that may indicate the possible emergence of AGI. Meanwhile, keep applying current AI to learn, reap its benefits and develop practices for its responsible use.
- Although AGI is not a reality now, current AI already poses significant risks regarding bias, reliability and other areas. Adopt emerging AI regulations and promote internal AI governance to manage current and emerging future risks of AI.

Sample Vendors

AGI Innovations; Google; IBM; Kimera Systems; Microsoft; New Sapience; NNAISENSE; OpenAI; Vicarious

Gartner Recommended Reading

The Future of AI: Reshaping Society

Innovation Insight for Generative AI

Innovation Insight: AI Simulation

Applying AI — Key Trends and Futures

Innovation Insight for Artificial Intelligence Foundation Models

Model Hubs

Analysis By: Eric Goodness

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Model hubs are repositories that host pretrained and readily available machine learning (ML) models, including generative models. Model hubs can serve as for-fee marketplaces for models available for commercial use or repositories of open-source generative models which are entitled as part of a service such as MLOps platforms. Model hubs increasingly offer automation and governance tools, curated datasets, model APIs and generative AI-enabled applications targeting specific enterprise needs.

Why This Is Important

Model hubs make it easier for developers to find generative AI models and open datasets to embed in their applications or workflows. Model hubs provide access to thousands of models — both closed source and open source — across a broad spectrum of use cases. Many model hubs curate generative models based on capabilities and specific use cases, and remove a lot of friction for developers looking to rapidly iterate and build generative AI applications from pilot to production faster.

Business Impact

By leveraging model hubs, enterprises can execute a range of design patterns to use generative model capabilities based on business needs. Ultimately, model hubs can greatly reduce friction for developers to support their company's need to create value for both internal operations and customers including proofs of concept for securing investment.

Drivers

- Enterprises increasingly view model hubs as an approach to reducing the complexity of integrating generative AI-enabled capabilities and a way to access state-of-the-art generative models more easily and faster. Without model hubs, organizations would require significant resources to develop and train their models, which is time-consuming and cost-intensive.
- Enterprises seek shorter time to value, and model hubs foster collaboration and knowledge sharing within the AI community. Enterprises can contribute to improving existing models, share experience and expertise, and collaborate to develop novel and creative solutions.
- There is growing demand for ease of use relating to model evaluation and benchmarking capabilities to allow enterprises to assess the quality and performance of different models and select the best fit for their use cases.
- There is a growing need for faster iterations to experiment and innovate with generative AI models, such as proofs of concept for retrieval augmented generation, and domain-specific models including BioBERT. Such features enable exploration and companies to develop unique and novel AI-based solutions.

Obstacles

- It is unclear how independent software vendors (ISVs) will charge for model hub capabilities. Today, most model hubs feature various MLOps software with many ISVs stating they plan to create charging schema based on utilization.
- Model hubs are new and emerging approaches to democratizing generative models. Growing pains have surfaced in this emergence relating to issues faced by developers such as software bugs, the introduction of breaking changes and a lack of acceptable documentation.
- While many pursue model hubs based on the variety of available generative models in the repository, users should perform due diligence to understand the ease of use and efficacy of tools and other services model hub providers offer as part of their platform. These tools and services may lack functionality found in stand-alone tools or form the catalogs of other providers, such as hyperscale providers.
- Smaller, emerging software companies lack the tools and input of large communities to add new, state-of-the-art models.

User Recommendations

- Perform broad due diligence by comparing the available repository of closed- and open- source models, tools available across the large language model operations (LLMOps) life cycle, and a review of the model hub community to reveal issues about technical code, support, general model and hub usability.
- Improve success when seeking access to state-of-the-art models by engaging larger providers such as the cloud hyperscalers that partner with smaller providers, to create broad and diverse hubs that provide both closed- and open-source models.
- Increase the knowledge and engineering skills for using, refining and fine-tuning off-the-shelf models found in model hubs.
- Reduce bias and ethical challenges by engaging providers of model hubs and ensuring the availability of tools and services to enable responsible AI.

Sample Vendors

Amazon Web Services; Databricks; Google; Hugging Face; IBM; Microsoft; Replicate; Snowflake

Gartner Recommended Reading

Applying AI — Governance and Risk Management

Toolkit: Delivery Metrics for DataOps, Self-Service Analytics, ModelOps and MLOps

Competitive Landscape: Cloud Providers Artificial Intelligence Services

Magic Quadrant for Cloud AI Developer Services

Critical Capabilities for Cloud AI Developer Services

Open-Source LLMs

Analysis By: Eric Goodness

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Open-source large language models (LLMs) are deep-learning foundation models distinguished by the terms of use, distribution granted to developers, and the developers' access to source code and the model architecture. Open-source LLMs are made available to the public through a license that enables anyone to access, use, modify and distribute the model source code without restriction.

Why This Is Important

Open-source LLMs, frameworks, libraries, tools and datasets accelerate enterprise value from the implementation of generative AI by democratizing access, reducing complexity and removing impediment for developers. In addition, open-source LLMs provide access to developer communities in enterprises, academia and other research roles that are working toward common goals to improve and make the models more valuable.

Business Impact

The key benefits of open-source models include increased customizability, better control over privacy and security, the ability to leverage collaborative development, model transparency, and the potential to reduce vendor lock-in. Ultimately, open-source LLMs offer enterprises smaller models that are easier and less costly to train, and enable business applications and core business processes.

Drivers

- Increased interest in customizing LLMs is driving open-source adoption. Open-source LLMs are flexible to customize, because engineers can access the model parameters and source code. Such access enables developers to customize these models (if they decide to do so) and have more control over costs, output and alignment for their use cases. Product ownership based on open-source LLMs enables enterprises to continuously develop them, based on internal and customer demands, and makes their applications harder to imitate by competitors.
- Growing development in AI open-source communities drives interest in and adoption of open-source LLMs. Open-source LLMs are generally supported and enriched by the collaborative power of development communities that continuously refine the models. This driver is predicated on the vibrancy of the community. Some developer communities build on top of these models and fine-tune them for specific use cases and runtime scenarios.
- The need for LLM transparency as proprietary LLMs is notoriously opaque. For example, closed-source LLMs do not provide transparency relating to LLM architectures, training methodologies, datasets or access to information relating to model weights and checkpoints. Open-source LLMs offer more transparency, enabling developer inspection and analysis.
- There is a growing need to avoid vendor lock-in for LLMs, even in scenarios in which the models are consumed with commercial support from a vendor. The generative AI model landscape is rapidly evolving; hence, open-source can provide flexibility for users to swap models or model providers in the future, with fewer exit barriers.

Obstacles

- The investments in data engineering, tooling integration and infrastructure to train and run these models can be high. This represents a significant fixed cost, compared with proprietary alternatives, as well as longer time-to-value and an impediment to implementation.
- Upgrading and managing the life cycle of open-source models, particularly if a lot of customization is built on top of a given version, will be difficult.
- The complexity of the variety of licensing models in open source impedes implementation. Open source can impose restrictions on the consumer and require rigorous review from legal teams before adoption. For example, not all open-source models are certified for commercial use.
- As measured by various benchmarks (e.g., ARC, BIG-Bench, HELM, HellaSwag and TruthfulQA), the gap in accuracy between proprietary LLMs and open-source LLMs reduces demand. This gap might not matter, depending on the accuracy required for your use case.

User Recommendations

- Perform due diligence to understand the legal exposure related to training data and the potential biases in the models, such as verifying the open-source license and checking for restrictions on its commercial use.
- Investigate to ensure that data privacy and security measures are in place when using open-source LLMs to process sensitive information.
- Proactively engage with open-source LLM communities as part of the due-diligence process to discern the positives and drawbacks associated with various models.
- Evaluate different open-source LLMs, based on such factors as performance, resource requirements, compatibility and documentation. Test the models on sample data, and compare their outputs against your defined objectives.
- Consider experimenting with fine-tuning on domain-specific data to assess the LLM's adaptability to your specific use cases.

Sample Vendors

BigCode; BigScience; Cerebras; EleutherAI; Google; H2O.ai; Meta; NVIDIA; Replit; StabilityAI

Gartner Recommended Reading

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

Quick Answer: How Do I Compare LLMs?

How to Create and Enforce a Governance Policy for Open-Source Software

A CTO's Guide to Open-Source Software: Answering the Top 10 FAQs

Tool: OSS Governance Policy Template

Hype Cycle for Open-Source Software, 2023

Market Guide for AI Trust, Risk and Security Management

Self-Supervised Learning

Analysis By: Pieter den Hamer, Erick Brethenoux

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Self-supervised learning is an approach to machine learning in which labels or supervisory signals are created from the data itself, without having to rely on historical outcome data or external (human) supervisors that provide labels or feedback. It is inspired by the way humans learn through observation, gradually building up general knowledge about concepts, events and their relations, or spatiotemporal associations in the real world.

Why This Is Important

Self-supervised learning aims to overcome one of the biggest drawbacks of supervised learning: the need to have access to typically large amounts of labeled training data. This is not only a practical problem in many organizations with limited relevant data, or where manual labeling is prohibitively expensive, but also a more fundamental problem with current AI, limiting its learning versatility and broader applicability.

Business Impact

Self-supervised learning enables the extended applicability of machine learning to use cases where labeled training datasets are not available. It may also shorten development time and improve the robustness and accuracy of models. Its relevance is most prominent in computer vision (CV), natural language processing (NLP) — including large language models (LLMs) such as GPT-4, Internet of Things (IoT) analytics/continuous intelligence, robotics, or other AI applications that rely on unstructured data or typically unlabeled sensor data.

Drivers

- **Making ML feasible in the absence of labeled training data:** In self-supervised learning, labels can be generated automatically from the data itself, without the need for human annotation. In essence, this is done by masking elements in the available data (e.g., a part of an image, a sensor reading in a time series, a frame in a video or a word in a sentence) and then training a model to "predict" the missing element. Thus, the model learns how one part relates to another, how one situation (captured through video and/or other sensors) typically precedes or follows another, and which words often go together, for example. In other words, the model increasingly represents the concepts and their spatial, temporal or other relations in a particular domain. This model can then be used as a foundation to further fine-tune the model (e.g., using "transfer learning") for one or more specific tasks with practical relevance.

- **Helping derive more value from the growing availability of IoT sensor data and other diverse, possibly external, sources of data:** Taken alone, these data sources (e.g., visual, sound, pressure, temperature or textual data) may be of limited value. More value can be derived from data by identifying associations between data sources, in essence, using the elements or events in one source to label elements or events in another source.

- **Stepping toward broader AI with more efficient learning:** Self-supervised learning has the potential to bring AI closer to the way humans learn. This occurs mainly via observation and association, building up general knowledge about the world through abstractions and then using this knowledge as a foundation for new learning tasks, thus incrementally building up ever-more knowledge that in future AI scenarios may serve as common sense. For example, ChatGPT and other generative AI rely heavily on this use of self-supervised learning.

Obstacles

- **Skills and experience are still very scarce:** Self-supervised learning is currently only practiced by a limited number of innovative AI companies. This includes its use by large tech firms in the context of foundation models for natural language processing and computer vision.
- **Tool support is still limited:** Although open-source ML frameworks, such as TensorFlow and PyTorch, have started to support self-supervised learning, broader tool support is lacking, which makes implementation a knowledge-intensive and low-level coding exercise.

User Recommendations

- **Apply self-supervised learning only when the value of such application justifies the risks of a still experimental approach.** Scarce, highly experienced ML experts are needed to carefully design a self-supervised learning task, based on masking of available data, which allows a model to build up knowledge and representations that are meaningful to the business problem at hand.
- **Apply self-supervised learning when manual labeling or annotating of data is too expensive or infeasible —** but only after comparing alternative approaches, such as the use of (external) data labeling and annotations services, synthetic data, reinforcement learning, active learning or federated learning.
- **Track the developments in self-supervised learning,** once more mature, self-supervised learning has the potential of becoming a pervasively used foundation for a next generation of applications with AI and machine learning, not limited to foundation models.
- **Complement self-supervised learning with other machine learning approaches.** Self-supervised learning can be used to create a baseline model, which can then be further improved or fine-tuned by using a (smaller) labeled dataset for supervised learning, or by applying reinforcement learning.

Sample Vendors

Amazon; craftworks; Google; Helm.ai; Microsoft; OpenAI; Speechmatics; V7

Gartner Recommended Reading

Three Steps to Boost Data for AI

Innovation Insight for Artificial Intelligence Foundation Models

Innovation Insight for Generative AI

Quick Answer: What Is GPT-4?

Multimodal GenAI

Analysis By: Danielle Casey, Roberta Cozza

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

Multimodal GenAI is the ability to combine multiple types of data inputs and outputs in generative models, such as images, videos, audio, text and numerical data. Multimodality augments the usability of generative AI by allowing models to interact with and create outputs across various modalities.

Why This Is Important

Multimodal GenAI is important because data in the real world is typically multimodal. Multimodality helps capture the relationships between different data streams and scales the benefits of GenAI across potentially all data types and applications. This allows AI to support humans in performing more tasks, regardless of the environment. Multimodal functionality will increasingly be present in tech offerings, as users demand associated performance.

Business Impact

Multimodal GenAI will have a transformational impact on enterprise applications by enabling the addition of new features and functionality otherwise unachievable. The impact of multimodality is not limited to specific industries or use cases, and can be applied at any touchpoint between AI and humans. Today, many multimodal models are limited to two or three modalities, though this will increase over the next few years to include more modalities. The future of AI is multimodal.

Drivers

Multimodal GenAI is being driven by:

- **Demand for reduced data silos:** Multimodality removes traditional data barriers by allowing users to interact with, manipulate, and create outcomes from numerous data types. Most data in reality is multimodal. The presence of large datasets that combine text, video, audio and other modalities will be a driver of multimodal model training.
- **Extensibility of automation:** Multimodality supports new tasks, such as data extraction, converting one data type to another, and creating new data outcomes. Applications that support multimodality will have a higher automation potential.
- **Improved user experience (UX):** Multimodality improves UX by enabling richer experiences, by meeting users where they are with the data that is available.
- **Customer and employee experience (CX/EX):** ChatGPT has already raised the bar on interactive experiences for users. As users demand richer experiences that incorporate text, voice, video and other modalities, generative AI models and applications need to rise up to that challenge.
- **Increased multimodal research:** AI labs are focusing their resources on creating models that can work across many different modalities. There is an increased appetite to use GenAI models across many modalities.

Obstacles

Multimodal GenAI models can be inhibited by:

- **Training challenges:** Multimodal GenAI models use deep learning, data alignment and fusion techniques to train, integrate and analyze data sourced from the multiple modalities. Multimodal data has varying degrees of quality and formats compared to unimodal data.
- **Lack of data:** Data availability may be limited in some modalities. For example, availability of large-scale audio datasets is more limited compared to other modalities like images and text. This impacts model training and accuracy.
- **Data exposure:** Multimodal GenAI increases the exposure to a wider range of sensitive data. Examples of particularly sensitive data types include maps or geolocation data, biometric data or health data.

- **Other risks:** Bias and inaccurate or fabricated outputs are amplified by multimodal data sources. Also, regulations and standards are a work in progress and are lagging GenAI's capability advancement.

User Recommendations

- Identify the two or three modalities that are most important to your organization when choosing which multimodal GenAI model to use to account for current limited multimodality support.
- Determine the need for a multimodal GenAI by assessing the level of operational inefficiencies and data silos within the organization.
- Assess the technical complexities of processing and integrating data inputs and outputs from diverse multimodal sources. Then, validate early on how these can best be integrated with key legacy or more current workflows.
- Create specific cross-modal data policies and tools aimed at protecting privacy, detecting bias, and ensuring compliance to emerging AI regulations.

Sample Vendors

Google; Meta; Microsoft; NVIDIA; OpenAI; Stability AI; Twelve Labs

LangOps

Analysis By: Bern Elliot, Soyeb Barot

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Enterprise natural language operations (LangOps) are practices that support the management of the full life cycle of the language models and solutions implemented in enterprise architectures. This includes the development and curation of training data, artificial intelligence (AI) models and semantic data (e.g., knowledge graphs), as well as the continuous delivery of retrained models and functionality integrated with relevant business processes and applications.

Why This Is Important

A growing number of natural language-based solutions are being used broadly across enterprises. Users want to customize and adapt capabilities based on their business requirements. To be successful, enterprise leaders must operationalize their approaches with the implementation of natural language technologies (NLTs) and techniques, to enable repeatability and reduce technical debt. This practice is called LangOps.

Business Impact

Enterprise LangOps operate natural language technologies — enabling enterprises to integrate these technologies across enterprise applications, improving efficiency, reusability of assets and scalability of deployments. LangOps may initially be limited to single NLT areas, such as text analytics, conversational platforms, large language models or translation. However, synergies across multiple NLT areas will drive language centers of excellence (COEs) and cross-functional LangOps streams.

Drivers

- Rapid increases in NLT by business areas. The business requirements are becoming more complex, and the underlying use of NLT is proliferating across increasingly broad business domain areas.
- The increasing need for sophisticated data and modeling practices, along with the volume of data. There is also demand for sharing training data and data-handling tools for semistructured and unstructured content.
- AI ModelOps and XOps practices that are not always applicable or adaptable to language areas. This drives a need for specialized streams in some language areas.
- Natural language solutions' need to draw from diverse technology and business skills. LangOps enables these diverse participant efforts to be organized and focused.
- The desire to share best practices for ingesting, managing, storing, governing, and monetizing large sets of unstructured data — for example, transcribed call recordings, specialized terminology libraries, and the use of vector data bases, as well as ontologies and knowledge graphs.
- Cross-functional uses for language technology, including model customization or fine-tuned language models. These might be used across multiple areas, including marketing, customer service, and websites.

Obstacles

- **Immature practices and methods:** The overall maturity of many emerging NLTs is accompanied by immaturity of the best practices for managing the full solution life cycle. LangOps will need to fit with ModelOps, XOps and emerging GenAI practices. Implementing human in the loop (HITL) practices for training may be nonstandard to Ops practices.
- **Language applications span different NLT platforms:** This causes inconsistencies for LangOps across platforms. LangOps approaches are sometimes focused on specific use cases for NLTs, such as translation and localization.
- **Organizational issues:** The cross-functional nature of LangOps requires people with different skills and reporting responsibilities to work together. This poses challenges that can be complicated by the potential for conflicting "turf" and budget control issues. For example, a single language solution may be used by different areas of the business. Previously, separate solutions were handled by each business area.

User Recommendations

- Develop an informal community of interest that includes business users of language technology as well as AI, generative AI, data science, machine learning (ML) and language experts. For many organizations, thinking about language technology as an interconnected area is a new concept. This community will assist in socializing this concept.
- Define a strategic enterprise NLT roadmap. This will start by viewing language initiatives as part of a broader portfolio, not as discrete projects. Allow projects to advance at their own pace, but look for how and where synergies across the organization will be useful. Provide guidelines for operationalizing NLT projects and the management of metadata and semantic data.
- Enlist the support of CxO-level sponsors as part of the planning. This will allow diverse business areas and groups to work together to reduce "turf" conflicts.
- Focus initially on language areas in which technical support to improve operations is already in use.

Sample Vendors

IBM; Microsoft, Unbabel; Veritone

Gartner Recommended Reading

Market Guide for AI-Enabled Translation Services

Tool: Vendor Identification for Natural Language Technologies

ChatGPT Research Highlights

Innovation Insight for Generative AI

Best Practices for the Responsible Use of Natural Language Technologies

Retrieval Augmented Generation

Analysis By: Radu Miclaus

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Retrieval augmented generation (RAG) is a design pattern that uses search functionality to retrieve relevant data and add it to the prompt of a generative AI model in order to ground the generative output with factual and new information. RAG can be used for both retrieving public internet data as well as for retrieving data from private knowledge bases.

Why This Is Important

The majority of large language models (LLMs) are trained on statistical language patterns in internet data and therefore can be susceptible to producing inaccurate outputs. While this can be overlooked for consumer-focused tools, enterprises need factual backing, privacy and access control to information without the need to shoulder the build of custom LLMs.

Business Impact

Traditionally, workers spend 20% to 30% of their time looking for the information needed to complete business tasks. Retrieval augmented generation, or RAG has the potential to enhance how information is synthesized which, when paired with RAG, has the potential to contextualize content and accelerate productivity. The applications for self-service content consumption for employees, as well as customer facing applications, will have a significant impact on employee and customer satisfaction, brand equity and workforce productivity.

Drivers

- The current worker experience of searching and finding information and knowledge documented in enterprises can be frustrating and requires a better information retrieval and synthesis approach.
- The advances of generative AI applied toward content synthesis has so far been the missing functionality in search applications for achieving a robust content consumption experience.
- The popularity of generative tools like ChatGPT and Bard have extended to shadow use (without IT oversight) of consumer tools in the enterprise for business activities. This has exposed risks around privacy and accuracy of output which in turn drives the need for the RAG pattern implementation on private knowledge bases.
- Generative AI service vendors, either via a hyperscale marketplace or via specialty model APIs, are making it easier for organizations to configure the RAG pattern on their knowledge bases.
- Enterprises are showing increased interest in using LLMs on their private knowledge bases, with the enterprise privacy and security guardrails in place.
- Competitive pressure to adopt the latest innovations in order to increase productivity and maintain competitive edge is a driver for organizations to activate their knowledge bases to support the RAG pattern for internal and external-facing applications.

Obstacles

- The need for making the knowledge bases available for retrieval is not new, it is a legacy need from enterprise search and has traditionally been under-invested in. Companies that do not have the discipline and skill sets for building robust retrieval and search to support the RAG pattern will experience friction during adoption.
- Knowledge engineering talent is specialized and different from traditional data engineering. The ramp-up in knowledge engineering professionals that can ingest, process, enrich knowledge bases and unstructured data repos as well as configure search pipelines and optimize search infrastructures will add to the adoption timeline.
- Configuration of knowledge base access controls for the RAG pattern is not a trivial task and can add obstacles in widespread adoption across all the organizational knowledge.
- Concerns about IP protection and responsible AI in the use of LLMs will remain obstacles in the adoption of RAG-based applications.

User Recommendations

IT and data and analytics leaders looking at adopting generative AI capabilities on top of private corporate data should:

- Assess their maturity and readiness in relation to knowledge management and information retrieval for employees and customers in order to inform how investments need to be prioritized.
- Pilot applications using the RAG pattern on a well-known knowledge base in order to assess the lift in the content consumption experience and gain buy-in for further investment.
- Plan for investment in filling any skill gaps that exist in their knowledge engineering capabilities, either through upskilling or external hiring.
- Engage with technology vendors or services providers that combine both technology and services to accelerate their adoption of RAG architecture.

Sample Vendors

Amazon; Charli AI; Google; Microsoft; Perplexity AI

Gartner Recommended Reading

AI Design Patterns for Large Language Models

AI Design Patterns for Knowledge Graphs and Generative AI

Prompt Engineering With Enterprise Information for LLMs and GenAI

How to Pilot Generative AI

Generative AI-Enabled Applications

Analysis By: Radu Miclaus, Arun Chandrasekaran

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Generative AI-enabled applications use generative AI for user experience (UX) and task augmentation to accelerate and assist the completion of a user's desired outcomes. When embedded in the experience, generative AI offers richer contextualization for singular tasks like generating and editing text, code, images and other multimodal output. As an emerging capability, process-aware generative AI agents can be prompted by users to accelerate workflows that tie multiple tasks together.

Why This Is Important

Fast-moving advances in foundation models drive generative AI-enabled applications, which have the potential to democratize the workforce. Since applications can now be enabled with generative AI capabilities that process and provide output in human consumable modalities (text, images, sound, etc.), the use cases will permeate a wide spectrum of domains and skill sets within the knowledge workforce, reimagining how enterprises think of scale and productivity.

Business Impact

Generative AI chatbots/agents/co-pilots within applications will target time-consuming, manual-prone and repetitive tasks, such as knowledge discovery, summarization and contextualization, software engineering and coding, graphic and video design, and workflow design and execution. With these tools at their disposal, knowledge workers and creatives will sustain new learning curves toward innovative ways to scale businesses. Businesses not making use of these tools will struggle to compete.

Drivers

- **Fast advancement of foundation models:** Foundation models like GPT are advancing at an accelerated rate. There is a movement toward democratizing foundation models via open-sourcing variations, such as Meta AI (Large Language Model Meta AI [LLaMa]) or BigScience Large Open-science Open-access Multilingual Language Model (BLOOM).

- **Wider range of applications:** Among others, the most common pattern for generative AI-embedded capabilities today is text-to-X, which democratizes the access for knowledge workers to what used to be specialized tasks via prompt engineering using natural language. For example, **text-to-text** supports knowledge discovery, summarization and contextualization in communication applications across the enterprise. **Text-to-code** is emerging as developer processes get augmented through "pair programming" with AI co-pilots directly into the coding experience, with use cases ranging across the software development life cycle. **Text-to-image/video (image-to-image)** applies when applications from graphics design to video editing and full video generation see generative capabilities added both by traditional technology players as well as new startups. **Text-to-process/workflow** is emerging as generative AI agents enable users to use text and voice to generate workflows and generative tasks together in cohesive domain-specific applications. **Text-to-multimodal** supports the building of high-fidelity avatars, or digital objects that have image, sound and narrative/text modalities, as an example of multimodal application in metaverse and gaming.

- **Domain specialization:** Specialization on top of foundation models is extending into domain-specific refinement, as well as refinement based on internal/private/licensed knowledge bases and process definitions for enterprises.

- **Acceptance into professional life:** Consumers are pulling the generative AI-enabled applications into their professional life.

- **Computation cost optimization:** The computational innovations for training and inference are focusing on optimizing and refining the cost structures across the entire software stack (infrastructure, methodologies and integrations).

Obstacles

- **Security, consumer privacy and enterprise intellectual property (IP) protection concerns:** A large number of inquiries from potential buyers of generative capabilities are concerned with the wide umbrella of trust and security. While large hyperscale vendors and startups are racing to make generative AI services enterprise-ready, in the short- to midterm, there will still be a lack of regulation and appropriate adaptable oversight.
- **Accuracy and veracity of outputs:** Hallucinations and inaccuracy will continue to be a concern for generative AI.
- **Fear around automation and job replacement:** Human nature brings a blend of excitement and fear around widespread adoption.
- **Learning curves and uncertainty:** As generative AI technology evolves, there is confusion about the implementation that is right for enterprises, how quickly the market is evolving and the lack of skills on transformers available in the market.
- **Regulation:** While currently lagging, regulations will follow and may increase the friction in innovation speed and adoption.

User Recommendations

- **Seek technology providers that can offer vertical specialization:** Vendors who will accelerate the refinement and adoption of generative AI capabilities in the context of vertical and business processes of the enterprise should be prioritized in evaluation for existing and future needs.
- **Use enterprise-ready technologies:** For enabling and embedding generative AI in applications, (a hybrid build-and-buy approach), prioritize research into the roadmaps of enterprise-ready generative AI services with a focus on addressing the privacy, security and IP protection needs of the enterprise.
- **Encourage steady growth:** Challenge knowledge workers to engage in new learning curves, and improve or redesign business processes to respond to this disruption.

Sample Vendors

Adobe; AgentGPT; Amazon; Anthropic; Google; Hugging Face; Inflection; Microsoft; OpenAI; Salesforce

Gartner Recommended Reading

Innovation Insight for Generative AI

Emerging Tech: Generative AI Needs Focus on Accuracy and Veracity to Ensure Widespread B2B Adoption

Quick Answer: How Can You Manage Trust, Risk and Security for ChatGPT Usage in Your Enterprise?

Innovation Insight for ML-Powered Coding Assistants

Quick Answer: Will Machine-Learning-Generated Code Replace Developers?

Transfer Learning

Analysis By: Ben Yan, Shubhangi Vashisth, Radu Miclaus, Wilco van Ginkel

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

Transfer learning reuses previously trained machine learning models as an advanced starting point for new purposes, in order to reduce the learning time and data required to attain acceptable performance.

Why This Is Important

Transfer learning is attractive. It enables rapid training, reduces the amount of data needed and may provide better predictive performance than models trained from scratch. A starting point is a repository of models, and these models can be customized based on internal or external data. Transfer learning advances the broader field of AI, as it allows AI to generalize — i.e., use what is learned in one task to more quickly learn another, related task. Transfer learning can also be used to further refine existing models with smaller amounts of data.

Business Impact

The business impact of transfer learning can be summarized as follows:

- It will impact how organizations apply machine learning (ML), especially for natural language processing (NLP) scenarios.
- It promises to attain acceptable performance with less training data, significantly less computational overhead (green IT) and faster development speed.
- It utilizes the model from the source (data-rich) domain, opening ML use cases that were previously infeasible due to lack of data.

Drivers

- **Broader application in foundation models:** The popularity of ChatGPT and image generation models is driving attention to the fine-tuning (a form of transfer learning) of foundation models. Compared with prompt engineering, fine-tuning techniques can incorporate more data into the models, and build customized models for organizations.
- **Increase in model marketplaces and community:** The availability of repositories, such as Hugging Face, allows developers to find and reuse pretrained models with easy community collaboration. The open-source repositories also enable organizations to build models with fewer barriers to entry.
- **Applicability to multiple use cases and verticals:** We see transfer learning being used across a number of model types (language, computer vision, predictive and multimodal) in many domains, such as finance, healthcare, gaming, autonomous driving and e-commerce.
- **Proliferation of AI models within organizations:** Many AI models can be reused. As more models are created, opportunities for transfer learning increase. These models and their datasets can be reused between departments, or even in external organizations.

Obstacles

- The adoption of transfer learning highly depends on the availability of existing models and the relevance/similarity between domains. It is hard to determine upfront whether transfer learning works.
- Transfer learning today is a capability embedded into existing platforms or a method applied by systems integrators and analytics consultancies.
- Transfer learning remains a technical challenge. Fine-tuning foundation models is even harder, and requires proper data, computing resources and talents. The ROI of fine-tuning customized models needs to be measured case by case.
- As the AI field becomes more regulated, documentation of source data and model lineage may be required to support explainability and trust. Not all model providers could provide sufficient information.
- In the quest for more explainable, fair and transparent AI, transfer learning can be seen as a further complication to the AI development process.

User Recommendations

- **Maintain repositories of AI models and datasets:** Work with your data and analytics leaders to utilize metadata management initiatives to identify AI models and their datasets. Document the successful transfer learning examples. AI centers of excellence (COEs) or similar should facilitate.
- **Explore useful internal and external models:** Seek transfer learning opportunities to reuse AI models. Organizations with a more mature level of AI adoption should additionally assess how their current models might be reused in related domains and/or similar tasks.
- **Check the ML tools you use to create and train models, and determine their support for transfer learning:** ML tools should include capabilities that facilitate transfer learning, such as fine tuning.
- **Loop in CSOs, legal teams and business stakeholders:** Teach them about transfer learning to develop your initial position on AI risk. For example, educate them on the lineage of the original data for the base model and the security risks of open base models you may use.

Sample Vendors

4Paradigm; Alibaba Group; Amazon; Google; H2O.ai; Hugging Face; IBM; Microsoft; NVIDIA; OpenAI

Gartner Recommended Reading

Innovation Insight: Transfer Learning

Transfer Learning in China: Increase the Value of Your Data Every Time You Use It

Three Steps to Boost Data for AI

AI Design Patterns with ChatGPT

At the Peak

AI TRiSM

Analysis By: Avivah Litan, Jeremy D'Hoinne, Bart Willemsen

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

AI trust, risk and security management (AI TRiSM) ensures AI model governance, trustworthiness, fairness, reliability, robustness, efficacy and data protection. AI TRiSM includes solutions and techniques for model interpretability and explainability, data and content anomaly detection, AI data protection, model operations and adversarial attack resistance.

Why This Is Important

AI models and applications deployed in production should be subject to protection mechanisms. Doing so ensures sustained value generation and acceptable use based on predetermined intentions. Accordingly, AI TRiSM is a framework that comprises a set of risk and security controls and trust enablers that helps enterprises govern and manage AI models and applications' life cycle — and accomplish business goals. The collateral benefit is enhanced compliance with forthcoming regulations, like the EU AI Act.

Business Impact

Organizations that do not consistently manage AI risks are exponentially inclined to experience adverse outcomes, such as project failures and breaches. Inaccurate, unethical or unintended AI outcomes, process errors and interference from malicious actors can result in security failures, financial and reputational loss or liability, and social harm. AI misperformance can also lead organizations to make suboptimal business decisions.

Drivers

- ChatGPT democratized third-party-provisioned generative AI and transformed how enterprises compete and do work. Accordingly, the risks associated with hosted, cloud-based generative AI applications are significant and rapidly evolving.

- Democratized, third-party-provisioned AI often poses considerable data confidentiality risks. This is because large, sensitive datasets used to train AI models are shared across organizations. Confidential data access must be carefully controlled to avoid adverse regulatory, commercial and reputational consequences.
- AI risk and security management imposes new operational requirements that are not fully understood and cannot be addressed by existing systems. New vendors are filling this gap.
- AI models and applications must be constantly monitored to ensure that implementations are compliant, fair and ethical. Risk management tools can identify and eliminate bias from training data and AI algorithms.
- AI model explainability must be constantly tested through model observations. Doing so ensures original explanations and interpretations of AI models remain active during model operations. If they don't, corrective actions must be taken.
- Detecting and stopping adversarial attacks on AI requires new methods that most enterprise security systems do not offer.
- Regulations for AI risk management — such as the EU AI Act and other regulatory frameworks in North America, China and India — are driving businesses to institute measures for managing AI model application risk. Such regulations define new compliance requirements organizations will have to meet on top of existing ones, like those pertaining to privacy protection.

Obstacles

- AI TRiSM is often an afterthought. Organizations generally don't consider it until models or applications are in production.
- Enterprises interfacing with hosted, large language models (LLMs) are missing native capabilities to automatically filter inputs and outputs — for example, confidential data policy violations or inaccurate information used for decision making. Also, enterprises must rely on vendor licensing agreements to ensure their confidential data remains private in the host environment.
- Once models and applications are in production, AI TRiSM becomes more challenging to retrofit to the AI workflow, thus creating inefficiencies and opening the process to potential risks.
- Most AI threats are not fully understood and not effectively addressed.

- AI TRiSM requires a cross-functional team, including legal, compliance, security, IT and data analytics staff, to establish common goals and use common frameworks — which is difficult to achieve.
- Although challenging, the integration of life cycle controls can be done with AI TRiSM.

User Recommendations

- Set up an organizational task force or dedicated unit to manage your AI TRiSM efforts. Include members who have a vested interest in your organization's AI projects.
- Work across your organization to effectively manage best-of-breed toolsets for enterprise-managed AI and applications that use hosted AI as part of a comprehensive AI TRiSM program.
- Avoid, to the extent possible, black-box models that stakeholders do not understand.
- Implement solutions that protect data used by AI models. Prepare to use different methods for different use cases and components.
- Establish data protection and privacy assurances in license agreements with vendors hosting LLM models — for example, Microsoft or OpenAI.
- Use enterprise-policy-driven content filtering for inputs and outputs to and from hosted models, such as LLMs.
- Incorporate risk management mechanisms into AI models and applications' design and operations. Constantly validate reliable and acceptable use cases.

Sample Vendors

AIShield; Arize AI; Arthur; Fiddler; ModelOp; Modzy; MOSTLY AI; Protopia AI; SolasAI; TrojAI

Gartner Recommended Reading

Use Gartner's MOST Framework for AI Trust and Risk Management

Top 5 Priorities for Managing AI Risk Within Gartner's MOST Framework

Prompt Engineering

Analysis By: Frances Karamouzis, Afraz Jaffri, Jim Hare, Arun Chandrasekaran, Van Baker

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Prompt engineering is the discipline of providing inputs, in the form of text or images, to generative AI models to specify and confine the set of responses the model can produce. The inputs prompt a set that produces a desired outcome without updating the actual weights of the model (as done with fine-tuning). Prompt engineering is also referred to as "in-context learning," where examples are provided to further guide the model.

Why This Is Important

Prompt engineering is the linchpin to business alignment for desired outcomes. Prompt engineering is important because large language models (LLMs) and generative AI models in general are extremely sensitive to nuances and small variations in input. A slight tweak can change an incorrect answer to one that is usable as an output. Each model has its own sensitivity level, and the discipline of prompt engineering is to uncover the sensitivity through iterative testing and evaluation.

Business Impact

Prompt engineering has the following business impacts:

- **Performance:** It helps improve model performance and reduce hallucinations.
- **Business alignment:** It allows subject data scientists, subject matter experts and software engineers to steer foundation models, which are general-purpose in nature, to align to the business, domain and industry.
- **Efficiency and effectiveness:** Alternative options, such as building a model from scratch or fine-tuning, can be much more complex, drive longer time to market and be more expensive.

Drivers

- **Balance and efficiency:** The fundamental driver for prompt engineering is it allows organizations to strike a balance between consuming an "as is" offering versus pursuing a more expensive and time-consuming approach of fine-tuning. Generative AI models, and in particular LLMs, are pretrained, so the data that enterprises want to use with these models cannot be added to the training set. Instead, prompts can be used to feed content to the model with an instruction to carry out a function.
- **Process or task-specific customizations or new use cases:** The insertion of context and patterns that a model uses to influence the output generated allows for customizations for a particular enterprise or domain, or regulatory items. Prompts are created to help improve the quality for different use cases — such as domain-specific question answering, summarization, categorization, and so on — with or without the need for fine-tuning a model, which can be expensive or impractical. This would also apply to creating and designing new use cases that utilize the model's capability for image and text generation.
- **Validation and verification:** It is important to test, understand and document the limits and weaknesses of the models to ensure a reduced risk of hallucination and unwanted outputs.

Obstacles

- **Embryonic nature of the discipline:** Prompt engineering processes and roles are either unknown or enterprises have a low level of understanding and experience. Gartner webinar polling data (over 2,500 responses; see Executive Pulse: AI Investment Gets a Boost From ChatGPT Hype) revealed that approximately 60% of respondents self-reported that they had not heard of prompt engineering. And 90% of those same respondents revealed that their organization did not currently have prompt engineers.

- **Role alignment:** Data scientists are critical to understanding the capabilities and limits of models, and to determine whether to pursue a purely prompt-based or fine-tuning-based approach (or combination of approaches) for customization. The ultimate goal is to use machine learning itself to generate the best prompts and achieve automated prompt optimization. This is in contrast to an end user of an LLM who concentrates on prompt design to manually alter prompts to give better responses.

- **Lack of business alignment:** There is often a lack of consensus on prompt engineering's business approach, as well as agreed-upon standards, methodology and approaches. This has led to fierce debates on the value of prompt engineering and how to establish governance.

- **Risk:** Beyond the early stages of awareness and understanding, the biggest obstacle may be that prompt engineering is focused on verification, validation, improvement and refinement; however, it's not without risk. Prompt engineering is not the panacea to all of the challenges. It helps to manage risk, not remove it completely. Errors may still occur, and potential liability is at stake.

User Recommendations

- Rapidly build awareness and understanding of prompt engineering in order to quickly start the journey of shape-shifting the appropriate prompt engineering discipline and teams.
- Build critical skills across a number of different team members that will synergistically contribute critical elements. For example, there are important roles for data scientists, business users, domain experts, software engineers and citizen developers.
- Communicate and cascade the message that prompt engineering is not foolproof. Rigor and diligence need to permeate and work across all the enterprise teams to ensure successful solutions.

Sample Vendors

FlowGPT; HoneyHive; LangChain; PromptBase; Prompt Flow; PromptLayer

Gartner Recommended Reading

Quick Answer: How Will Prompt Engineering Impact the Work of Data Scientists?

Quick Answer: What Impact Will Generative AI Have on Search?

Accelerate Adoption of Generative AI by Offering an FMOps- or a Domain-Specific Partner Ecosystem

Glossary of Terms for Generative AI and Large Language Models

AI-Augmented Software Engineering

Analysis By: Arun Batchu, Hema Nair, Oleksandr Matvitskyy

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

The use of artificial intelligence (AI) technologies (e.g., machine learning [ML] and natural language processing [NLP]) to help software engineers create, deliver and maintain applications is designated AI-augmented software engineering (AIASE). This is integrated with engineers' existing tools to provide real-time, intelligent feedback and suggestions.

Why This Is Important

Today's software development life cycle includes such routine and repetitive tasks as boilerplate functional and unit-test code and docstrings, which AIASE tools automate. AI-powered automation enables software engineers to focus their time, energy and creativity on such high-value activities as feature development. Emerging AI tools discover the configurations that meet operational goals. Software builders who use these tools remain productive and engaged, and they stay longer in their jobs.

Business Impact

AIASE accelerates application delivery and allocates software engineering capacity to business initiatives with high priority, complexity and uncertainty, helping quality teams develop self-healing tests and nonobvious code paths. These tools automatically generate test scenarios previously created manually by testers, and detect test scenarios often missed by test teams. AIASE tools detect issues with code security, consistency or maintainability and offer fixes.

Drivers

Demand drivers include:

- The increasing complexity of software systems to be engineered
- Increasing demand for developers to deliver high-quality code faster
- Increasing numbers of application development security attacks
- Optimizing operational costs

Technology solution drivers include:

- The application of AI models to prevent application vulnerabilities by detecting static code and runtime attack patterns
- The increasing impact of software development on business models

- The application of large language models to software code
- The application of deep-learning models to software operations

Obstacles

- Hype about the innovation has caused misunderstandings and unrealistic expectations about the benefits of AIASE.
- There is a lack of deep comprehension of generated artifacts.
- There is limited awareness about production-ready tools.
- Software engineers who fear job obsolescence have shown resistance.
- There is a lack of transparency and provenance of data used for model training.
- Uneven, fragmented solutions that automate only some of the tasks in the software development life cycle (SDLC).
- AI skills such as prompt engineering, training, tuning, maintaining and troubleshooting models.
- High model training and inference costs at scale.
- Intellectual property risks stemming from models trained on nonpermissive licensed code.
- Privacy concerns stemming from code, and associated proprietary data leaking as training data for AI models.
- Technical employees' fear of jobs being automated by AI.

User Recommendations

- Pilot, measure and roll out tools only if there are clear gains.
- Verify the maintainability of AI-generated artifacts, including executable requirements, code, tests and scripts.
- Track this rapidly evolving and highly impactful market to identify new products that minimize development toil and improve the experience of software engineers, such as those that ease security and site operations burden.
- Reassure software engineers that AIASE is an augmentation toolset for human engineers, not a replacement.
- Pick providers (including open-source vendors) that supply visibility to training data and transparency on how the model was trained.
- Establish the correct set of metrics, such as new release frequency and ROI, to measure the success of AIASE.

Sample Vendors

Akamas; Amazon Web Services; Diffblue; Google; IBM; Microsoft; OpenAI; SeaLights; Sedai; Snyk

Gartner Recommended Reading

Innovation Insight for ML-Powered Coding Assistants

Infographic: Artificial Intelligence Use-Case Prism for Software Development and Testing

Market Guide for AI-Augmented Software Testing Tools

Foundation Models

Analysis By: Arun Chandrasekaran

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Foundation models are large-parameter models that are trained on a broad gamut of datasets in a self-supervised manner. They are mostly based on transformer or diffusion deep neural network architectures and will potentially be multimodal in the near future. They are called foundation models because of their critical importance and applicability to a wide variety of downstream use cases. This broad applicability is due to the pretraining and versatility of the models.

Why This Is Important

Foundation models are an important step forward for AI due to their massive pretraining and wide use-case applicability. They can deliver state-of-the-art capabilities with higher efficacy than their predecessors. They've become the go-to architecture for NLP, and have also been applied to computer vision, audio and video processing, software engineering, chemistry, finance, and legal use cases. Primarily text-based, large language models (LLMs) are a popular subset of foundation models. ChatGPT is based on one (GPT-4).

Business Impact

With their potential to enhance applications across a broad range of natural language use cases, foundation models will have a wide impact across vertical industries and business functions. Their impact has accelerated, with a growing ecosystem of startups building enterprise applications on top of them. Foundation models will advance digital transformation within the enterprise by improving workforce productivity, automating and enhancing CX, and enabling rapid, cost-effective creation of new products and services.

Drivers

Foundation models:

- **Require only limited model customization to deliver effective results.** Foundation models can effectively deliver value through prebuilt APIs, prompt engineering or further fine-tuning. While fine-tuning may deliver the best value because of customization possibilities, the other two options are less complex.
- **Deliver superior natural language processing.** The difference between these models and prior neural network solutions is stark. The large pretrained models can produce coherent text, code, images, speech and video at a scale and accuracy not possible before.

- **Enable low-friction experimentation.** The past year has seen an influx of foundation models, along with smaller, pretrained domain-specific models built from them. Most of these are available as cloud APIs or open-source projects, further reducing the time and cost to experiment.
- **Have accelerated AI innovation with massive model sizes.** Examples include OpenAI's GPT-4; Google's AI's PaLM; Google DeepMind's Gopher and Chinchilla; Meta AI's LLaMA; and Alibaba's M6. In addition, companies such as Hugging Face, Stability AI and EleutherAI have open-sourced their models.

Obstacles

Foundation models:

- **Do not deliver perfect results.** Although a significant advance, foundation models still require careful training and guardrails. Because of their training methods and black-box nature, they can deliver unacceptable results or hallucinations. They also can propagate downstream any bias or copyright issues in the datasets.
- **Require appropriate skills and talent.** As with all AI solutions, the end result depends on the skills, knowledge and talent of the trainers, particularly for prompt engineering and fine-tuning.
- **Expand to impractical sizes.** Large models are up to billions or trillions of parameters. They are impractically large to train for most organizations because of the necessary compute resources, which can make them expensive and ecologically unfriendly.
- **Concentrate power.** These models have been mostly built by the largest technology companies with huge R&D investments and significant AI talent, resulting in a concentration of power among a few large, deep-pocketed entities. This situation may create a significant imbalance in the future.

User Recommendations

- **Create a strategy document** that outlines the benefits, risks, opportunities and execution plans for these models in a collaborative effort.
- **Plan to introduce foundation models into existing speech, text or coding programs.** If you have any older language processing systems, moving to a transformer-based model could significantly improve performance. One example might be a text interpretation, where transformers can interpret multiple ideas in a single utterance. This shift in approach can significantly advance language interfaces by reducing the number of interactions.
- **Start with models that have superior ecosystem support,** have adequate enterprise guardrails around security and privacy, and are more widely deployed.
- **Explore new use cases,** such as natural language inference, sentiment analysis or natural-language-based enterprise search, where the models can significantly improve both accuracy and time to market.
- **Designate an incubation team** to monitor industry developments, communicate the art of the possible, experiment with BUs and share valuable lessons learned companywide.

Sample Vendors

Alibaba Group; Amazon; Baidu; Cohere; Google; Hugging Face; IBM; Microsoft; OpenAI; Stability AI

Synthetic Data

Analysis By: Arun Chandrasekaran, Anthony Mullen, Alys Woodward

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Synthetic data is a class of data that is artificially generated rather than obtained from direct observations of the real world. Synthetic data is used as a proxy for real data in a wide variety of use cases including data anonymization, AI and machine learning development, data sharing and data monetization.

Why This Is Important

A major problem with AI development today is the burden involved in obtaining real-world data and labeling it. This time-consuming and expensive task can be remedied with synthetic data. Additionally, for specific use-cases like training models for autonomous vehicles, collecting real data for 100% coverage of edge cases is practically impossible. Furthermore, synthetic data can be generated without personally identifiable information (PII) or protected health information (PHI), making it a valuable technology for privacy preservation.

Business Impact

Adoption is increasing across various industries. Gartner predicts a massive increase in adoption as synthetic data:

- Avoids using PII when training machine learning (ML) models via synthetic variations of original data or synthetic replacement of parts of data.
- Reduces cost and saves time in ML development.
- Improves ML performance as more training data leads to better outcomes.
- Enables organizations to pursue new use cases for which very little real data is available.
- Is capable of addressing fairness issues more efficiently.

Drivers

- In healthcare and finance, buyer interest is growing as synthetic tabular data can be used to preserve privacy in AI training data.
- To meet increasing demand for synthetic data for natural language automation training, especially for chatbots and speech applications, new and existing vendors are bringing offerings to market. This is expanding the vendor landscape and driving synthetic data adoption.
- Synthetic data applications have expanded beyond automotive and computer vision use cases to include data monetization, external analytics support, platform evaluation and the development of test data.
- Increasing adoption of AI simulation techniques is accelerating synthetic data.
- There is an expansion to other data types. While tabular, image, video, text and speech applications are common, R&D labs are expanding the concept of synthetic data to graphs. Synthetically generated graphs will resemble, but not overlap the original. As organizations begin to use graph technology more, we expect this method to mature and drive adoption.
- The explosion of innovation in AI foundation models is boosting synthetic data creation. These models are becoming more accessible and more accurate.

Obstacles

- Synthetic data can have bias problems, miss natural anomalies, be complicated to develop, or not contribute any new information to existing, real-world data.
- Data quality is tied to the model that develops the data.
- Synthetic data generation methodologies lack standardization.
- Completeness and realism are highly subjective with synthetic data.
- Buyers are still confused over when and how to use the technology due to lack of skills.
- Synthetic data can still reveal a lot of sensitive details about an organization, so security is a concern. An ML model could be reverse-engineered via active learning. With active learning, a learning algorithm can interactively query a user (or other information sources) to label new data points with the desired outputs, meaning learning algorithms can actively query the user or teacher for labels.
- If fringe or edge cases are not part of the seed dataset, they will not be synthetized. This means the handling of such borderline cases must be carefully accommodated.
- There may be a level of user skepticism as data may be perceived to be “inferior” or “fake.”

User Recommendations

- Identify areas in your organization where data is missing, incomplete or expensive to obtain, and is thus currently blocking AI initiatives. In regulated industries, such as healthcare or finance, exercise caution and adhere to rules.
- Use synthetic variations of the original data, or synthetic replacement of parts of data, when personal data is required but data privacy is a requirement.
- Educate internal stakeholders through training programs on the benefits and limitations of synthetic data and institute guardrails to mitigate challenges such as user skepticism and inadequate data validation.
- Measure and communicate the business value, success and failure stories of synthetic data initiatives.

Sample Vendors

Anonos (Statice); Datagen; Diveplane; Gretel; Hazy; MOSTLY AI; Neuromation; Rendered.ai; Tonic.ai; YData

Gartner Recommended Reading

Innovation Insight for Synthetic Data

Innovation Insight for Generative AI

Data Science and Machine Learning Trends You Can't Ignore

Cool Vendors in Data-Centric AI

Case Study: Enable Business-Led Innovation with Synthetic Data (Fidelity International)

Large Language Models

Analysis By: Leinar Ramos, Bern Elliot

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Large language models (LLMs) are AI foundational models that have been trained on vast amounts of unlabeled textual data. Applications can use LLMs to accomplish a wide range of tasks, including question answering, content generation, content summarization, retrieval-augmented generation (RAG), code generation, language translation and conversational chat.

Why This Is Important

LLMs are one of the core technologies powering the generative AI revolution. During the past few years, LLMs based on transformer architectures have demonstrated surprising and significant capabilities across a wide range of domains and industries. They have served as general-purpose foundation models that tackle many different use cases, and bring state-of-the-art AI capabilities to organizations.

Business Impact

As a result of their general-purpose nature, LLMs are being adopted across all industries and business functions. They are being used to generate text, build question-answering systems, summarize and classify documents, translate and edit text, and generate and explain programming code, as well as many other use cases. As LLMs are incorporated into enterprise applications, including integration with chatbot and conversational front ends (e.g., ChatGPT), their impact has been increasing.

Drivers

- **Access to large volumes of unlabeled text:** The public availability of large volumes of textual data, particularly on the web, makes the creation of high-performing LLMs possible.
- **Vendor adoption:** Large technology companies are investing heavily in the development of new LLMs, and applying them to new use cases as they race to achieve a strong position in the LLM space. Similarly, LLMs are being incorporated into many enterprise applications, expanding adoption.
- **Open-source LLMs:** The improved performance and increased availability of commercially licensed open-source LLMs is expanding the potential adoption of LLMs. This has lowered the barrier to entry for startups and other vendors, as well as increased the feasibility of a private deployment approach for enterprises that cannot use cloud-based, proprietary LLMs.
- **Growing ecosystem around LLMs:** LLMs can be augmented by additional capabilities that create more-powerful systems. There is a growing ecosystem of tools, plugins, extensions, orchestration and data retrieval layers that build more-complex LLM-based applications.
- **Ability to customize LLMs:** LLMs can be customized with domain-specific data for specific industries, functions and use cases. This flexibility will drive broader application of LLMs in the enterprise. This LLM customization is achievable through prompt engineering/RAG patterns or via LLM fine tuning.
- **Increased compute power:** The exponential increase in computing power and performance during the past decade has enabled the training and use of ever-larger LLMs, improving their performance as they scale.

Obstacles

- **Unreliability:** LLMs produce inaccuracies and hallucinations, which makes them less suitable for a broad range of use cases. LLM use requires customization, governance and, often, human supervision.
- **Explainability:** LLMs are not currently explainable. LLM use case adoption will be constrained by the need for decision-making transparency.
- **LLM proliferation**: Organizations will struggle to navigate the wide range of LLMs, validate vendor claims and select the right models for their enterprise use cases.
- **Uncertainty of regulatory environment:** Intellectual property in LLM training datasets, privacy and confidentiality of enterprise data, and legal liability are some of the issues that will limit adoption.
- **Cybersecurity:** LLMs are susceptible to adversarial attacks and prompt injections that could result in data leakage and malicious LLM use.
- **Energy consumption**: LLM training and use are highly energy-intensive. As LLMs become larger and more widely used, their environmental impact will increase.

User Recommendations

- Identify AI use cases in which LLMs can be deployed to deliver a more immediate business impact. Experiment with use cases in which the cost of errors is acceptable, and transparency and explainability are not required.
- Follow prompt engineering best practices to optimize LLM performance. Combine LLMs with other applications and tools for more complex design patterns, such as combining LLMs with RAG for use cases that require bringing your own private data into LLMs.
- Validate vendor LLM claims. Consider a wide range of criteria when comparing LLMs, This should include the type of model required, the model's multidomain benchmark performance, the ability to fine-tune or customize the model, as well as the broader ecosystem of tools that can support the LLM.
- Define a policy for the acceptable use of LLMs and LLM-based applications that clearly identifies the risks of using these tools, as well as the expected user actions to mitigate them.

Sample Vendors

A121; AWS; Baidu; Cohere; Google; IBM; Meta; Microsoft; NVIDIA; OpenAI

Gartner Recommended Reading

AI Design Patterns for Large Language Models

Balance the Environmental Perils and Promises of Generative AI

Glossary of Terms for Generative AI and Large Language Models

GenAI Workload Accelerators

Analysis By: Alan Priestley, Arun Chandrasekaran

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

Generative AI workload accelerators, either GPUs or custom ASICs, are chips designed to operate alongside a CPU and support the highly parallel processing operations necessary to develop (train) and deploy (inference) applications based on large generative AI models.

Why This Is Important

The rapid growth in development of large-scale generative AI models demands significantly more processing performance than can be delivered by standard CPUs. Leveraging workload accelerator chips optimized for these tasks can significantly reduce training times and enable cost-effective deployment of models to support high volumes of user transactions.

Business Impact

Leveraging generative AI workload accelerators enables:

- Reduction in time taken to train or fine tune generative AI models for a specific use case or dataset.

- Cost-effective deployment of generative AI-based applications for use in a wide spectrum of business use cases.

Drivers

- Development of generative AI models requires the use of processing techniques that leverage high-performance systems with workload accelerators.
- Executing generative AI-based applications typically requires the use of computer systems that can execute high volumes of highly parallel math operations.
- Generative AI models require training with large sets of data. Workload accelerators such as data center GPUs can be used for this task, but high-performance custom ASICs are being developed that can deliver a more cost-effective solution to this problem.
- Many generative AI models are being trained by using large cloud-based clusters of servers that integrate workload accelerators, but many enterprise organizations will need to deploy workload accelerator-based infrastructure within their on-premises data centers to deploy these models for use.

Obstacles

- The most commonly utilized workload accelerator is a high-performance data center GPU. These are expensive and have very high power demands, impacting both server design and the data center infrastructure required to support them.
- Deployments may be limited where a data center has insufficient power per rack or inadequate cooling solutions to support high-power workload accelerators.
- New server with higher specification CPUs, more memory and storage along with high bandwidth networking equipment may be required to deploy generative AI applications in on-premises data centers.
- A limited number of vendors offer ASIC based workload accelerators that can be used by enterprises as an alternative to GPUs, creating access and support challenges.
- While ASIC-based workload accelerators can offer significantly better performance (at lower power) than GPU-based solutions, GPUs may be more readily available from a wider range of vendors.

User Recommendations

- Use CPUs (on-premises or cloud infrastructure) when generative AI workload demand is light enough to fit in conventional CPU-based infrastructure.
- Use GPUs or dedicated workload accelerators when generative AI workloads would otherwise consume excessive server resources or when performance needs aren't met with CPUs.
- Select workload accelerators and vendors that offer or support the broadest set of generative AI models and toolsets.
- Use compute optimization tools to boost the efficiency of your compute accelerators for better ROI.

Sample Vendors

Amazon Web Services; AMD; Google; Intel; NVIDIA; SambaNova Systems

Gartner Recommended Reading

Forecast: AI Semiconductors, Worldwide, 2021-2027, 2Q23 Update

Forecast Analysis: AI Semiconductors, Worldwide

Quick Answer: How Will GenAI Impact the Semiconductor Industry?

Emerging Tech Impact Radar: Artificial Intelligence

GenAI-Enabled Virtual Assistants

Analysis By: Danielle Casey, Bern Elliot

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Generative AI (GenAI)-enabled virtual assistants (VAs) represent a new generation of VAs that leverage large language models (LLMs) to achieve functionality that cannot be obtained with previous VA methods. GenAI is being used to improve VA performance, add new functionality, extend task automation and support new value outcomes.

Why This Is Important

LLMs will materially augment VAs, and providers are either currently piloting, already using, or planning on adding GenAI capabilities to their R&D roadmaps. There are different approaches that can be taken when incorporating an LLM into a VA. They include chaining multiple LLM API integrations, embedding an out-of-box model into an offering and retraining an LLM to create a customized model. Each option must be evaluated against time and costs.

Business Impact

GenAI-enabled VAs will:

- Operate with improved accuracy in conversational dialogue and content discovery
- Improve GenAI-enabled capabilities, such as text summarization, content generation and content visualization
- Likely be multimodal, due to LLM flexibility around data ingestion and supported outputs
- Extend automation
- Improve operational efficiency
- Save resources
- Enable new service offerings

Drivers

GenAI-enabled virtual assistants are driven by:

- **Customer demand for LLMs.** The hype about LLMs, primarily ChatGPT, has created customer demand for LLMs in product offerings and heightened expectations about VA performance. By 2025, GenAI will be embedded in 80% of conversational AI offerings, up from 20% in 2023.

- **Vendors pivoting rapidly.** Many VA providers have preexisting knowledge graphs, indexed vector databases and technical in-house expertise. Coupled with existing, high-performing out-of-box models, repositioning VAs as GenAI enabled is relatively achievable for most vendors. Notably, some vendors have been using LLMs (such as GPT-2) for several years.
- **Revenue opportunities.** GenAI-enabled VAs will be able to perform additional tasks, support more complex use cases and deliver higher levels of operational efficiency. An eager customer base is willing to pay for these performance improvements and associated value outcomes. To not use LLMs is to miss out on significant and immediate revenue opportunities.

Obstacles

Obstacles inhibiting adoption include:

- **Accuracy issues.** LLMs are not entirely accurate and have the unique problem of hallucinating (i.e., making up facts). Techniques for improving accuracy and reducing hallucinations include prompt engineering, policy injection, knowledge graphs, indexed vector databases and model fine-tuning.
- **The lack of explainability.** LLMs differ from traditional, non-GenAI models in that they lack transparency to provide explainable outcomes.
- **Regulatory uncertainty.** LLMs have received particular scrutiny from regulators across geographies. Moreover, advancements in responsible AI trail GenAI adoption.
- **Cost of customization.** Creating industry-specific LLMs by retraining open-source models on an industry dataset is costly and time-consuming. Yet, custom LLMs will augment domain-specific VAs.
- **Data privacy and security concerns.** There are outstanding data privacy and security concerns about LLM data usage.

User Recommendations

- Support innovation while mitigating risk by developing clear and comprehensive company guidance, policies, tools and evaluative frameworks concerning the use of GenAI in VAs or other applications.
- Create a tiered use case strategy that prioritizes internal-facing data- and text-heavy use cases in the near term, and pushes decision intelligence and creative, external-facing use cases to the middle to long term.
- Prepare for future regulations and ensure compliance by investing in responsible AI.
- Expect the distinction between GenAI-enabled VAs and VAs to diminish rapidly as GenAI methods become a basic requirement and feature of all virtual assistants.

Sample Vendors

Amelia; Anthropic; Avaamo; Baidu; Google; Microsoft; Moveworks; OpenAI; Openstream.ai

Gartner Recommended Reading

Emerging Tech: Use Generative AI to Transform Conversational AI Solutions

Emerging Tech Roundup: ChatGPT Hype Fuels Urgency for Advancing Conversational AI and Generative AI

Emerging Tech: Top Use Cases for Generative AI

Appendixes

Hype Cycle Phases, Benefit Ratings and Maturity Levels

Table 2: Hype Cycle Phases

(Enlarged table in Appendix)

Phase	Definition
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant media and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the innovation is pushed to its limits. The only enterprises making money are conference organizers and content publishers.
Trough of Disillusionment	Because the innovation does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the innovation's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the innovation are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology's target audience has adopted or is adopting the technology as it enters this phase.
Years to Mainstream Adoption	The time required for the innovation to reach the Plateau of Productivity.

Source: Gartner (September 2023)

Table 3: Benefit Ratings

Benefit Rating ↓	*Definition* ↓
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (September 2023)

Table 4: Maturity Levels

(Enlarged table in Appendix)

Maturity Levels	Status	Products/Vendors
Embryonic	In labs	None
Emerging	Commercialization by vendors Pilots and deployments by industry leaders	First generation High price Much customization
Adolescent	Maturing technology capabilities and process understanding Uptake beyond early adopters	Second generation Less customization
Early mainstream	Proven technology Vendors, technology and adoption rapidly evolving	Third generation More out-of-box methodologies
Mature mainstream	Robust technology Not much evolution in vendors or technology	Several dominant vendors
Legacy	Not appropriate for new developments Cost of migration constrains replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source: Gartner (September 2023)

Recommended by the Authors

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

Understanding Gartner's Hype Cycles

Tool: Create Your Own Hype Cycle With Gartner's Hype Cycle Builder

Innovation Guide for Generative AI Technologies

How to Pilot Generative AI

How to Choose an Approach for Deploying Generative AI

A Generative AI Playbook for CDAOs

Table 1: Priority Matrix for Generative AI, 2023

Benefit ↓	*Years to Mainstream Adoption*			
	Less Than 2 Years ↓	***2 - 5 Years*** ↓	***5 - 10 Years*** ↓	***More Than 10 Years*** ↓
Transformational		AI-Augmented Software Engineering Large Language Models Multimodal GenAI Self-Supervised Learning	Autonomous Agents Foundation Models	Artificial General Intelligence
High	GenAI-Enabled Virtual Assistants GenAI Workload Accelerators Generative AI-Enabled Applications Retrieval Augmented Generation	AI TRiSM Edge LLMs Model Hubs Open-Source LLMs Prompt Engineering Synthetic Data	AI Simulation Domain-Specific GenAI Models GenAI Application Orchestration Frameworks LangOps Reinforcement Learning From Human Feedback Transfer Learning Vector Databases	
Moderate				
Low				

Source: Gartner (September 2023)

Table 2: Hype Cycle Phases

Phase	*Definition*
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant media and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the innovation is pushed to its limits. The only enterprises making money are conference organizers and content publishers.
Trough of Disillusionment	Because the innovation does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the innovation's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the innovation are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology's target audience has adopted or is adopting the technology as it enters this phase.
Years to Mainstream Adoption	The time required for the innovation to reach the Plateau of Productivity.

Phase ↓	Definition ↓

Source: Gartner (September 2023)

Table 3: Benefit Ratings

Benefit Rating ↓	Definition ↓
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (September 2023)

Table 4: Maturity Levels

Maturity Levels ↓	*Status* ↓	*Products/Vendors* ↓
Embryonic	In labs	None
Emerging	Commercialization by vendors Pilots and deployments by industry leaders	First generation High price Much customization
Adolescent	Maturing technology capabilities and process understanding Uptake beyond early adopters	Second generation Less customization
Early mainstream	Proven technology Vendors, technology and adoption rapidly evolving	Third generation More out-of-box methodologies
Mature mainstream	Robust technology Not much evolution in vendors or technology	Several dominant vendors
Legacy	Not appropriate for new developments Cost of migration constrains replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source: Gartner (September 2023)