Hype Cycle for Artificial Intelligence, 2025

11 June 2025 - ID G00828523 - 123 min read

By: Haritha Khandabattu, Birgi Tamersoy

Initiatives: Drive AI Transformation for Sustainable Competitive Advantage; Strategy, Risks and Opportunities; Technologies and Markets

AI investment remains strong, but focus is shifting from GenAI hype to foundational innovations like AI-ready data, AI agents, AI engineering and ModelOps. This research helps leaders prioritize high-impact, emerging AI techniques while navigating regulatory complexity and operational scaling.

Analysis

What You Need to Know

As AI adoption matures, the focus is shifting from experimentation to scale. Generative AI (GenAI) is now at the Trough of Disillusionment, signaling a maturing understanding of its potential and limits. Enterprises are directing investment toward AI-enabling capabilities — from high-quality, contextualized data to responsible AI governance — to ensure consistent, scalable delivery. As concerns around AI safety, sovereignty and ethics intensify, responsible AI is becoming not just a compliance requirement, but a design principle and a differentiator.

At the same time, newer innovations like multimodal AI, AI trust risk and security management (TRiSM) adoption, and AI agents (LLM-based) dominate the Peak of Inflated Expectations. Composite AI has retained its relevance, serving as a foundational strategy for combining diverse AI techniques. Innovations in simulation, embodied AI and world models continue to gain traction, offering powerful capabilities that are setting the stage for future widespread adoption.

Organizations must bring AI closer to the point of decision, using fit-for-purpose AI solutions tailored to context and workload. Success will depend on tightly business-aligned pilots, proactive infrastructure benchmarking, and coordination between AI and business teams to optimize value delivery across environments.

The Hype Cycle

The two biggest movers in this year's Hype Cycle, AI-ready data and AI agents, are experiencing heightened interest. This trend, accompanied by ambitious projections and speculative promises, places them at the Peak of Inflated Expectations. As businesses invest in AI-ready data strategies and deploy AI agents, they must navigate the complexities of implementation and integration and the risks associated with deployment. It is essential to manage expectations realistically and plan strategically to get the full potential of these innovations, while also preparing for the inevitable challenges that follow.

AI governance platforms and FinOps for AI are two new entries on the Hype Cycle that indicate a focus on establishing robust frameworks for managing AI's risks and expanding its role in enterprises. Their arrival marks a shift toward not only developing cutting-edge AI technologies, but also implementing structured oversight and financial efficiency measures to support sustainable growth.

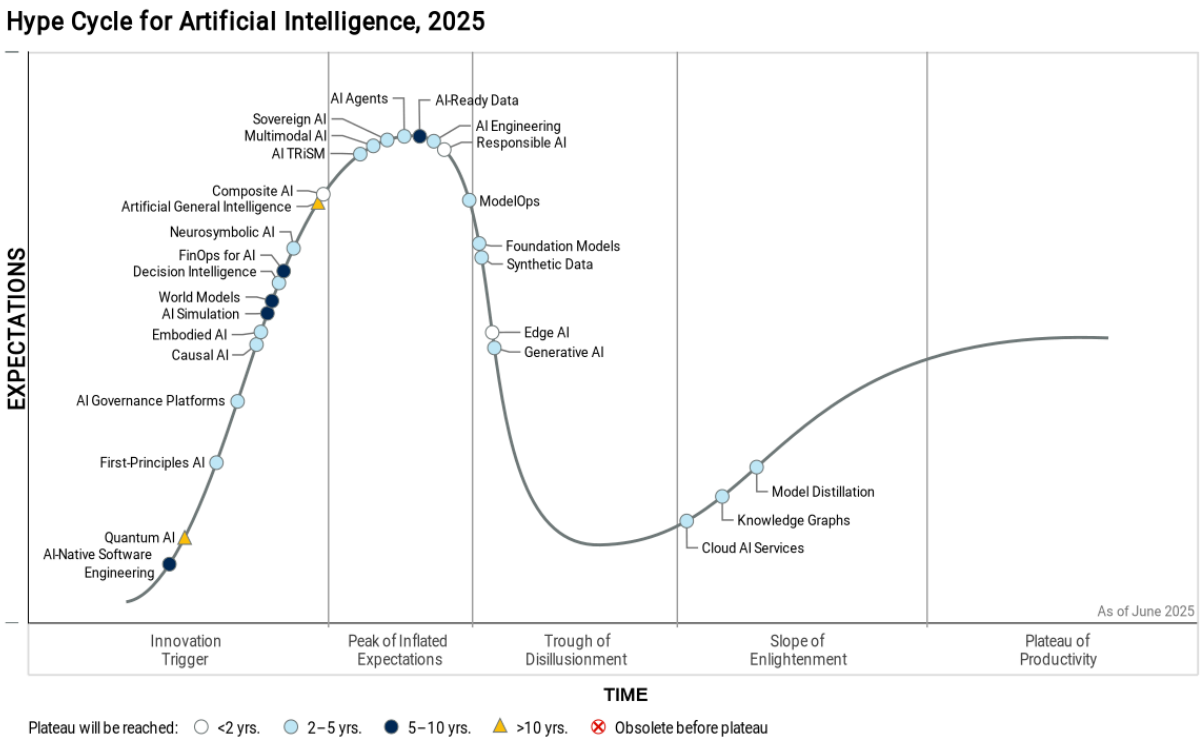
The entry of AI-native software engineering into the Hype Cycle highlights how rapid changes in the field are profoundly reshaping roles and tasks. Similar to other evolving roles, AI-native software engineers are redefining their way of working — orchestrating agentic tools and workflows to design, build and scale intelligent systems with greater autonomy and creative intent.

Advancements like model distillation, already present in later stages of the Hype Cycle, emphasize the maturing of capabilities that enhance efficiency and enable more versatile and comprehensive AI applications.

Meanwhile, innovations such as quantum AI, first-principles AI and causal AI have experienced slower progression compared to last year's Hype Cycle. While these areas hold great promise, their measured pace suggests that they face complex technical challenges and may require significant breakthroughs before they can be widely adopted in successful AI solutions.

Last but not least, computer vision has graduated from the Artificial Intelligence Hype Cycle this year; it has become an established technology with widespread integration in real-world applications.

Figure 1: Hype Cycle for Artificial Intelligence, 2025



Gartner

The Priority Matrix

This year's Priority Matrix reflects an AI landscape dominated by innovations of high or transformational benefit. Compared with 2024, there is a sharper emphasis on operational scalability and real-time intelligence, with a gradual pivot from GenAI as a central focus toward the foundational enablers that support sustainable AI delivery.

Within the next two years, edge AI, composite AI and responsible AI are anticipated to achieve mainstream adoption. Edge AI will enable more efficient data processing by bringing computation closer to data sources and further increase AI's potential applicability. Composite AI will enhance the robustness of AI systems by integrating multiple AI techniques. Responsible AI will ensure that these advancements are implemented ethically and transparently, paving the way for trust and accountability in AI systems.

In the two- to five-year time frame, several key AI technologies are expected to reach the Plateau of Productivity, driving significant advancements across industries. These include foundational technologies like AI engineering and ModelOps, which streamline the integration and management of AI systems, and emerging techniques such as generative AI and multimodal AI, which enhance content creation and data interpretation. Additionally, AI governance platforms and AI TRiSM are set to play crucial roles in ensuring ethical and secure AI deployment. Together, these developments will enable more robust, innovative and responsible AI applications, transforming how businesses and organizations operate.

Over the next five to 10 years, AI-ready data and AI simulation are expected to significantly enhance how AI is deployed across industries. AI-ready data will ensure that datasets are optimized for AI applications, enhancing accuracy and efficiency. Meanwhile, AI simulation will improve predictive capabilities and scenario testing, and FinOps for AI will streamline the financial management of AI investments, collectively driving more effective AI implementations.

Longer-term innovations, such as artificial general intelligence, hold disruptive potential but face scientific, ethical and geopolitical hurdles. Sustained research and coordinated ecosystem growth will be critical to their evolution.

AI leaders must balance ambitious exploration with operational discipline — advancing transformational capabilities while ensuring near-term investments are grounded in scalable, fit-for-purpose AI solutions.

Table 1: Priority Matrix for Artificial Intelligence, 2025

(Enlarged table in Appendix)

Benefit ↓	*Years to Mainstream Adoption*			
	Less Than 2 Years ↓	*2 to 5 Years* ↓	*5 to 10 Years* ↓	*More Than 10 Years* ↓
Transformational	Composite AI Responsible AI	AI Engineering Decision Intelligence Embodied AI First-Principles AI Foundation Models Generative AI ModelOps Multimodal AI	AI-Native Software Engineering AI-Ready Data World Models	Artificial General Intelligence
High	Edge AI	AI Agents AI Governance Platforms AI TRiSM Causal AI Cloud AI Services Knowledge Graphs Model Distillation Neurosymbolic AI Sovereign AI Synthetic Data	AI Simulation FinOps for AI	
Moderate				
Low				Quantum AI

Source: Gartner (June 2025)

Off the Hype Cycle

On the Rise

AI-Native Software Engineering

Analysis By: Manjunath Bhat, Mark Driver

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

AI-native software engineering is an emerging set of practices and principles that are optimized for using AI-based tools to develop and deliver software applications. This entails AI autonomously or semiautonomously performing a significant percentage of the tasks across the software development life cycle (SDLC). For example, developers use AI agents that proactively recommend and execute actions by sensing inputs from development environments with the goal of automating end-to-end workflows.

Why This Is Important

Today, developers use AI-based tools such as AI code assistants and AI testing tools that are limited to coding and testing activities. However, Gartner predicts a future where AI will be integral and native to most software engineering tasks, changing both the nature of developers' work and their way of working. AI-native software engineering practices are important because they enable developers to focus on meaningful tasks that require critical thinking, human ingenuity and empathy.

Business Impact

AI-native software engineering has the potential to deliver numerous benefits for developers, product teams and business stakeholders alike. AI-based tools will help developers go beyond handling drudgery. These tools will serve as ideation partners that boost human creativity and support product teams in generating new ideas. Product teams can use AI-enabled analysis to make data-driven product roadmap decisions that either validate or refute subjective human decisions. Then, teams can expedite creating prototypes to speed up feasibility studies and chart an informed path forward.

Drivers

- **Achieving step change in productivity and performance:** Software engineering has become AI-augmented with the emergence of AI-based developer tools, but the overall method and SDLC haven't yet transformed. AI augmentation has been applied to many tasks within the traditional SDLC to provide incremental improvements in lead time, cycle time and quality. This is driving software engineering leaders to explore ways to achieve a step change in productivity and performance.

- **Boosting human creativity:** Teams can leverage multimodal capabilities in AI-based development tools to boost human creativity. Instead of staring at a blank canvas, a product manager or UX designer can use AI design tools to convert text prompts, screenshots or paper sketches to images and visual prototypes. They can then iterate on the design and, once the design is finalized, UX designers can convert those designs to HTML and cascading style sheets (CSS).

- **Translating intent into action:** Early previews of AI-native application builders, such as Bolt from StackBlitz, Firebase Studio from Google, v0 from Vercel, and Lovable, demonstrate the ability to translate user intent into actions. Users can describe requirements from an end-user perspective rather than technical specifications. This signals a fundamental shift in how we express intent to computers — going from precise instructions to describing the desired state in natural language. Andrej Karpathy, the cofounder of OpenAI and founder of Eureka Labs, coined the term "vibe coding" to describe this new programming paradigm where you "forget the code even exists." Note that vibe coding as defined is currently limited to experimental prototypes and not suited to take applications to production.

- **Generating positive developer sentiment:** Developers, in general, are keen on experimenting with new tools. A recent Gartner survey of 5,112 software development team members from 51 organizations shows that 54% of software development team members consider having freedom to experiment and innovate as one of the most important aspects of the developer experience. Positive developer attitudes toward AI is driving the adoption of AI-native engineering practices within organizations.

Obstacles

- **Overly trusting AI outputs:** AI-native approaches create a new burden on developers and knowledge workers in general. Developers increasingly offload tasks to AI-based tools, which carry inherent risks of nondeterminism and hallucinations. Therefore, blindly trusting AI outputs without verification and explainability can potentially pose serious business risks, including reputational damage. GitClear's 2025 AI Copilot Code Quality Report shows four times more code cloning and a spike in the prevalence of duplicate code blocks between 2023 and 2024 compared to previous years.

- **Increased security risk:** AI tools expand the threat surface to include the chain of all events and interactions they initiate and participate in — including those that are invisible to human or system operators. This expanded threat surface increases the potential for unforeseen vulnerabilities and security breaches. Developers must account for agentic workflows as part of assessing and mitigating software supply chain risks.

- **Compounded risk of hallucinations in multiagentic workflows:** The risk of hallucinations compounds in multiagentic workflows, where AI-generated context is passed from one AI agent to another. Model overreach is also an issue with agentic systems — where the model does more than asked because in the training data it often goes further than just the bit you need. For example, you may ask for it to write a script for "search" and it creates one that does "search and replace."

User Recommendations

- Rethink developer workflows by taking advantage of emerging AI-based tools and technologies to improve productivity and enhance creative upstream work. Examples of upstream use cases include product discovery, product design, user sentiment analysis and feature prioritization.

- Unlock greater gains in productivity and minimize hallucinations by sharing context between AI-enabled developer tools (potentially AI agents) as they transition work from one tool to another.

- Prioritize low-risk and high-value use cases for agent-based tools by assessing their ability to automate repetitive work and keep humans in the loop for oversight, verification and explainability. To sustain high quality, human developers must review code, use test harnesses and establish security and compliance guardrails throughout the software development life cycle.

- Explore ways to benefit from autonomous improvement loops by segmenting tasks based on business criticality, risk threshold and task complexity. Software engineering leaders will need to look for opportunities where these autonomous loops deliver business value without increasing risks.

Sample Vendors

Anysphere; CodeStory; GitHub; Google; Lovable; Replit; StackBlitz; Vercel; Windsurf; Zed Industries

Gartner Recommended Reading

Innovation Insight for AI-Native Software Engineering

Why AI Will Not Replace Software Engineers, and What That Means for the C-Suite

How to Upskill Software Engineering Teams in the Age of AI

Magic Quadrant for AI Code Assistants

Emerging Tech: AI Developer Tools Must Span SDLC Phases to Deliver Value

Quantum AI

Analysis By: Chirag Dekate, Soyeb Barot

Benefit Rating: Low

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Quantum AI is an embryonic field of research emerging at the intersection of quantum technologies and AI. Quantum AI aims to exploit unique properties of quantum mechanics to develop new and more powerful AI algorithms that deliver better than classical performance, potentially resulting in new types of AI algorithms designed to run on quantum systems.

Why This Is Important

Quantum AI is an area of active research. Once commercialized, quantum AI could potentially help in:

- Enabling organizations to use quantum systems to address advanced AI analytics faster while using a fraction of the resources used in conventional AI supercomputing.
- Developing new AI algorithms that exploit quantum mechanics to deliver capabilities beyond ones that can be executed on classical systems.
- Unlocking disruptive applications that include drug discovery, energy industry and logistics.

Business Impact

While the business impact of the embryonic quantum AI field today is low, when validated techniques mature, quantum AI will enable competitive advantage across industries; for instance:

- **Life sciences:** Transform drug discovery by shortening timelines, lowering costs and improving outcomes.
- **Finance:** Optimize portfolios, minimize risk and improve fraud detection systems.
- **Material science:** Revolutionize energy transportation, manufacturing and create new revenue streams by discovering new materials.

Drivers

- Progress is steady in scaling quantum systems and improving error correction schemes.
- Hype around quantum technologies is driving more businesses and researchers to explore the intersection of quantum and AI.
- The accelerated pace of innovation in quantum systems (including a larger volume of higher quality qubits, and greater stability and reliability of quantum systems) is driving greater interest in applicability in areas, including quantum AI.
- Access to quantum computing as a service is lowering the barrier to entry, encouraging greater collaboration among researchers and enabling exploration of new algorithms and techniques.
- Governments and enterprises globally are increasing funding for quantum (and quantum AI) research, resulting in accelerated innovation.
- The halo effect of increased hype around GenAI is driving new focus on alternative research techniques, including quantum AI, that could potentially deliver new disruptive results.
- Universities and training programs are developing programs and curricula to develop a quantum-ready workforce.

Obstacles

- **Hardware limitations:** Current quantum systems, while getting stabler, are still error-prone and inherently noisy, limiting their utility and impact on practical quantum AI.
- **Algorithm limitations:** While several quantum AI algorithms have been proposed, very few have been vetted and proven, and they are nowhere close to being enterprise-ready.
- **Cost:** Despite their limited utility and widespread accessibility, rapidly evolving noisy intermediate-scale quantum (NISQ) systems are relatively expensive, which could inhibit research and development efforts needed to devise quantum AI algorithms.
- **Scalability of systems:** Scaling quantum systems to the level necessary for enterprise-ready quantum AI continues to be a major technical hurdle.
- **Compute paradigms:** Integrating traditional data and analytics pipelines with quantum is inherently challenging because quantum systems operate on a fundamentally different paradigm both from a data representation perspective and from a compute (non-von Neumann model) perspective.

User Recommendations

- Prioritize investments in AI and GenAI over any quantum AI investments. Quantum AI is too nascent to warrant focused investments and unlikely to yield material gains in the next two to three years.
- Partner with local universities by sponsoring academic research as a means of derisking your quantum AI investments and create a university-to-industry talent pipeline.
- Create a quantum AI opportunity radar that enables you to track progress of underlying technologies and quantum AI algorithms, enabling you to maximize value creation as the embryonic field of quantum technologies evolves.
- Diversify quantum use cases beyond a narrow AI context into other domains including materials simulations, search, optimization and other emerging algorithmic domains.

Sample Vendors

Amazon Web Services; Google; IBM; IonQ; Microsoft; Multiverse Computing; Pasqal; SandboxAQ

Gartner Recommended Reading

Leaders' Guide to Quantum Computing

Start Your Quantum Strategy Now in Response to Recent Quantum Processor Innovations

Executive Briefing on Emerging Technology: Quantum Computing

First-Principles AI

Analysis By: Erick Brethenoux, Svetlana Sicular

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

First-principles AI (FPAI; aka physics-informed AI) incorporates physical and analog principles, governing laws and domain knowledge into AI systems. In contrast, purely digital AI models do not necessarily obey the fundamental governing laws of physical systems and first principles — nor generalize well to scenarios on which they have not been trained. FPAI extends AI engineering to complex systems engineering and model-based systems, such as agent-based systems.

Why This Is Important

As AI expands in engineering and scientific use cases, it needs a stronger ability to model problems and better represent their context, but digital-only AI solutions cannot generalize well enough beyond training, which limits their adaptability. In contrast, FPAI instills a more reliable representation of the context and the physical reality, yielding more robust and adaptive systems. This leads to reduced training time, improved data efficiency, better generalization and greater physical consistency.

Business Impact

Physically consistent and scientifically sound AI models can significantly improve applicability, especially in engineering use cases. FPAI helps train models with fewer data points and accelerates the training process. As a result, models converge faster to optimal solutions. FPAI improves the generalizability of models to make reliable predictions for unseen scenarios, including applicability to nonstationary systems, and enhances transparency and interpretability, boosting trustworthiness.

Drivers

- **More flexible representation of system context and conditions:** FPAI approaches instill greater context flexibility, allowing developers to build more adaptive systems. Traditional business modeling approaches have been brittle. This is because the digital building blocks composing solutions cannot generalize well enough beyond their initial training data, therefore limiting those solutions' adaptability.
- **Additional physical knowledge representations:** As an example, FPAI approaches provide physics equations (e.g., partial differential equations) to guide or bound AI models. AI techniques, particularly in the machine learning (ML) family, have severe limitations — especially for causality and dependency analysis, admissible values, context flexibility and memory retention mechanisms. Asset-centric industries have already started leveraging FPAI in physical prototyping, predictive maintenance or composite materials analysis, for example.
- **Modeling challenges:** Complex systems like climate models, large-scale digital twins and complex health science problems are particularly difficult to model. Composite AI approaches provide more concrete answers and manageable solutions to these problems, but their engineering remains a significant challenge.
- **Simplified and enriched AI approaches:** First-principles knowledge defines problem and solution boundaries, reducing the scope of ML's traditional brute-force approach. For example, first-principles-based semantics can reveal deepfakes.
- **Need for more robust and adaptable business simulation systems:** With a better range of context modelization and more accurate knowledge representation techniques, FPAI simulations will be more reliable and account for a wider range of possible scenarios — all better-anchored in reality.
- **The advent of systems based on AI agents:** The capability of these systems to promote efficient combination of digital and analog models will, in turn, promote the use of FPAI.

Obstacles

- Development of systematic tests and standardized evaluations for these models across benchmark datasets and problems could slow the adoption of FPAI capabilities.
- Scaling of the training, testing and deployment of complex FPAI models on large datasets in an efficient manner will be a computational challenge.
- Collaboration across many diverse communities of physicists, mathematicians, computer scientists, statisticians, AI experts and domain scientists will pose a resource challenge.
- Brute-force approaches are prevalent in AI and easy to implement for data scientists, while first principles require additional fundamental knowledge of a subject that calls for a multidisciplinary team.
- Developers' difficulty to scope first-principle methods, without access to engineers or subject matter experts, can also prove a major obstacle for the introduction of these methods.

User Recommendations

- Set realistic development objectives by identifying errors that cannot be reduced and discrepancies that cannot be addressed, including data quality.
- Encourage reproducible and verifiable models, starting with small-scoped problems. Complex systems, in the scientific sense of the term, are generally good candidates for this approach.
- Enforce standards for testing accuracy and physical consistency for physics and first-principles-based models of the relevant domain, while characterizing sources of uncertainty.
- Promote model-consistent training for FPAI models and train models with data characteristics representative of the application, such as noise, sparsity and incompleteness.
- Quantify generalizability about how performance degrades with degree of extrapolation to unseen initial and boundary conditions and scenarios.
- Ensure relevant roles and education in a multidisciplinary AI team with domain expertise, so the team can develop effective and verifiable solutions.

Sample Vendors

Abzu; IntelliSense.io; MathWorks; NNAISENSE; NVIDIA; VERSES

Gartner Recommended Reading

Innovation Insight: AI Simulation

Go Beyond Machine Learning and Leverage Other AI Approaches

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

AI Governance Platforms

Analysis By: Lauren Kornutick, Sumit Agarwal, Avivah Litan

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

An AI governance platform provides a central view of AI applications and agents and each unique use case in the enterprise. It measures performance against agreed-to frameworks such as policy, regulations and industry standards. In addition to a central repository of use cases, it consolidates risk management activities, automates workflow approval for new uses and applies the appropriate level of technical enforcement required in a single-continuous tool for elements critical to AI.

Why This Is Important

AI is spreading rapidly. AI leaders are increasingly tasked with oversight of AI despite diverse subject matter expertise required to safely deploy. Oversight also ensures acceptable use and value based on predetermined criteria. AI governance platforms are emerging as a distinct market that combines traditional governance, risk and compliance product capabilities with automating and enforcing essential governance rules to verify that AI is safe, valuable and performing as intended.

Business Impact

We have observed that AI governance has become more centralized, although organizations may take different approaches. AI governance platforms are emerging to meet this need and address the governance, risk management and compliance requirements for deploying AI. AI governance platforms can significantly improve operational efficiency, engage cross-functional stakeholders effectively and enforce policies at runtime, all of which contribute to the overall success of AI in the enterprise at scale.

Drivers

- The increasingly pervasive nature of AI and a need for transparency and accountability highlight the need for strong AI governance.
- Global regulations that directly or adjacently target issues associated with AI, such as bias, security, safety and ethical concerns, and a lack of transparency and accountability, increase the complexity of compliance. A consolidated view is essential for implementing technical controls and oversight through approvals, validations, audits and observability to allow the enterprise to operate at scale.
- Organizations create AI governance principles and policies that lack monitoring and enforcement, which apply to use cases of AI in production.
- To effectively oversee AI, and deploy at scale, AI governance will require technology to implement AI governance controls to enact and enforce new enterprise AI requirements.
- Successful AI governance requires cross-functional subject matter expertise, including legal, compliance, risk, cybersecurity, IT and data analytics, to establish common goals, taxonomies and frameworks, which is difficult to achieve because of competing priorities. This reduces the effectiveness of governance and increases risk.
- Organizations already, and will continue to, do security assessments at the product level — that is insufficient. AI assessments must be conducted for each AI-enabled feature and use case.
- Stakeholders are increasingly demanding trust — and organizations leverage high-level responsible AI principles and AI acceptable use policies to drive confidence in an organization's products, services and the overall brand.
- Enterprise AI implementations are expanding, incorporating various AI models such as machine learning, computer vision, large language models and others, and driving new risks. Enterprises require a platform that provides a comprehensive view of the enterprise AI portfolio and record of decisions and accountability, its risks, risk mitigations and reliable performance.

Obstacles

- There is confusion over who should take ownership of AI governance (for example, legal, cybersecurity, data and analytics, IT.)
- AI governance teams will overlap with some, but not all, operational policies already monitored and enforced by data and analytics and IT governance, MLOps and cybersecurity teams. This necessitates clear perimeters within the organization while making space for this new function.
- The diverse nature of AI use cases and priorities requires organizations to decide if a single common platform of AI governance, or a combination of tools, is most appropriate for their needs.
- Vendors are latching onto “AI governance” as a go-to-market strategy due to low awareness of capabilities. This leaves buyers confused and with products that overpromise and underdeliver.
- Global regulations and standards vary wildly — and even though controls needed for compliance are similar, desired outcomes may be in opposition.
- Rapid progress makes some features obsolete and calls for new features that lag the progress.

User Recommendations

- Define acceptable use policies at a level granular enough to enforce for AI use cases. Acceptable use should be aligned to broader enterprise-responsible AI principles.
- Establish a review and approval workflow process for low-touch or high-touch reviews and attestation for new use cases.
- Establish an enterprise risk tolerance for deployment of AI within the enterprise that calculates both risk and value in a single risk score.
- Use the risk score to triage who will need to review and approve the AI and how the AI will be monitored, including technical controls.
- Evaluate the AIs performance against risk tolerance and deploy technical controls for AI. This should be done at some level for all use cases.
- Determine which governance processes need to translate to technology to effectively scale AI and to define your requirements for an AI governance platform.
- Check platform vendor references from your industry. Ensure that the references use the functionality you require too.

Sample Vendors

Collibra; Cranium; Credo AI; Holistic.AI; IBM; Monitaur; OneTrust

Gartner Recommended Reading

3 Key Steps to Build a Scalable AI Governance Framework

Market Guide for AI Trust, Risk and Security Management

Causal AI

Analysis By: Pieter den Hamer, Leinar Ramos, Ben Yan

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Causal AI identifies and utilizes cause-and-effect relationships to go beyond correlation-based predictive or generative models and toward AI systems that can prescribe actions more effectively and act more autonomously. It includes different techniques, such as causal graphs and simulation, that help uncover causal relationships to improve decision making.

Why This Is Important

AI's ultimate value comes from making better decisions and taking effective actions. However, the current correlation-based generative and predictive approaches have their limitations. Not only do they offer very limited transparency, generating or predicting an outcome is not the same as understanding what causes it and how to improve it. Causal AI is crucial when systems need to be more transparent and reliable in identifying and prescribing actions to achieve the right business outcomes.

Business Impact

Causal AI leads to:

- More reliable augmentation and autonomy in decision intelligence or AI-empowered decision making.
- More robustness and adaptabilty by leveraging causal relationships that remain valid in changing environments.
- The ability to extract causal knowledge, also known as causal discovery with various AI techniques, sometimes combined with simulations, reduces time and costs of real-world experiments (for example, A/B tests), although validation is still required.

Drivers

- In the context of decision intelligence, analytics demand is shifting from predictive to more prescriptive capabilities. A causal understanding of how to affect predicted outcomes and optimize decision making is increasingly important.
- AI — in particular agentic AI — systems increasingly need to act autonomously, particularly for time-sensitive and complex use cases where human involvement is not feasible. This will only be possible by AI understanding what impact actions will have and how to make effective interventions.
- Limited data availability for certain use cases requires more data-efficient techniques like causal AI, possibly combined with synthetic data. Causal AI leverages human domain knowledge of cause-and-effect relationships to bootstrap AI models in small-data situations.
- Growing complexity and dynamics of business require more robust AI techniques. Correlation-based AI models, trained with historical data, are brittle and lose accuracy when faced with gradual, let alone disruptive, changes. Causal structure changes much more slowly than statistical correlations, making causal AI more robust and adaptable in fast-changing environments.
- The need for greater AI trust and explainability is driving interest in models that are more intuitive to humans. Causal AI techniques, such as causal graphs, make it possible to be explicit about causes and explain models in terms that humans understand.
- Generative AI (GenAI) can accelerate causal AI implementation. GenAI is emerging as an aid to explore documents and other data sources for existing causal knowledge. This can then be used to generate candidate causal graphs, which, while still requiring human validation or completion, may reduce time-consuming manual work.
- The next step in AI requires causal AI. Current deep learning models, in particular large language models (LLMs) and “reasoning” models for GenAI and AI agents, have limitations in terms of reliability. A composite AI approach that complements for example LLMs with causal AI — in particular, causal knowledge graphs — offers a promising avenue to bring AI to a higher level.

Obstacles

- Causality is not trivial. Not every phenomenon is easy to model in terms of its causes and effects, with many factors potentially being relevant. Causality might be delayed, circular, unknown or hard to validate, despite the growing use of AI for causal discovery.
- The quality of a causal AI model depends on its causal assumptions and on the data used to build it. This data is susceptible to bias and imbalance and may be incomplete in terms of representing all causal factors, known or unknown.
- Causal AI requires technical and domain expertise to properly estimate causal effects. Building causal AI models is often more difficult than building correlation-based predictive models, requiring active collaboration between domain experts and AI experts.
- AI experts might be unaware of causality methods. If AI experts are overly reliant on data-driven models like machine learning (ML) or LLMs, organizations could get pushback when looking to implement causal AI.
- Limited experience with enterprise-scale applications. This represents a challenge when organizations run initial causal AI pilots and then want to scale them up to larger and possibly more complex causal models.

User Recommendations

- Apply causal AI to replace or complement existing AI approaches, including machine learning, generative AI and agentic AI, to achieve greater reliability and transparency.
- Use causal AI when more augmentation and automation is required. Examples include decision intelligence use cases in customer retention programs, marketing campaign allocation and financial portfolio optimization, as well as in smart robotics and autonomous systems.
- Select different causal AI techniques depending on the complexity of the specific use case. These include ML or LLMs for causal discovery, causal rule inferencing, causal graphs, Bayesian networks or simulation.
- Educate your AI teams on causal AI. Explain the difference between causal and correlation-based AI and cover the range of techniques available to incorporate causality.
- Closely involve domain experts in causal AI initiatives to help create, maintain or at least validate causal models.

Sample Vendors

Actable AI; Bayes Server; causaLens; Causality Link; Geminos; Howso; Parabole.ai; Scalnyx; Vizuro; Xplain Data

Gartner Recommended Reading

AI Design Patterns for Composite AI

When Not to Use Generative AI

Innovation Insight: AI Agents

Embodied AI

Analysis By: Pieter den Hamer

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Embodied AI is based on the view that intelligence and embodiment in a certain context are inextricably linked — one shapes the other. It is an approach where a physical or virtual AI agent's models are trained and co-engineered with its embodiment: the user interface, sensors, appearance, actuators or other capabilities required to perceive and interact with a specific, real or simulated environment. This enables more robust, resilient and adaptive execution of intelligent tasks.

Why This Is Important

Embodied AI aims to create AI agents that can act autonomously or augment humans in practical, dynamic contexts — much more so than current AI, including abstract large language and "reasoning" models with limited reliability and effectiveness in decision making and action taking. This is achieved through active perception and adaptive behavior, orchestrated by an AI agent's intelligence that is in symbiosis with the capabilities and constraints of the AI agent's host or body in a certain environment.

Business Impact

Embodied AI paves the way toward more robust, trustworthy, adaptive and actionable AI, widening its applicability and value creation. This is particularly the case where there is a need for more practical know-how, physical common sense, social and emotional intelligence, and a greater resilience to deal with the dynamics and unexpected events in real-world or virtual environments. Example use cases include autonomous vehicles and smart robots, but also virtual assistants or gaming characters.

Drivers

- Recent advances in GenAI and agentic AI are impressive, yet AI still has significant limitations, particularly with respect to its reliability in dealing with the dynamics and complexity of reality.
- Advances in realistic 3D/4D simulations, virtual/augmented/mixed reality and gaming. Combined with reinforcement learning for adaptive behavior training, this allows the co-evolving of baseline versions of both embodiment and intelligence of AI agents, before further deploying and improving them in a real environment, be it physical or virtual.
- Emerging approaches include world models, physics-informed or first-principles AI (representing, among others, the laws of physics or engineering heuristics), adaptive AI (learning during operations), emotion AI (understanding and expressing feelings in a social context), composite AI (e.g., using neurosymbolic AI for spatiotemporal reasoning) and causal AI (representing cause-and-effect relations).
- Innovation is ongoing in sensor technology, robotics engineering and, for example, new materials for more natural mechanics and haptic interfaces (relevant for embodied AI in physical contexts).
- Scientific insights about intelligence are evolving; intelligence is no longer seen as a centralized brain-only concept. Cognitive traits like perception, emotion, reasoning and behavior are often distributed and co-evolved in multiple parts of the body.
- Investments are being made in research to develop future artificial general intelligence, for which embodied AI is increasingly seen as a critical step, based on the view that intelligence is inseparable from its operational entity that interacts with the environment. This means it is not abstracted from but grounded in reality by design, holding the promise of providing intrinsic meaning or semantics to its knowledge representations and “native” common sense.

Obstacles

- The world is a very complex, unpredictable and even chaotic place. That is why the development of realistic simulations, effective robotics and — for example — truly autonomous cars has proven to be elusive.
- Real-world interaction requires real-time, highly responsive AI, even with limited energy and compute resources (e.g., on mobile or edge devices). However, more lightweight and energy-efficient AI are not easily achievable.
- Embodied AI holds the promise of more autonomous AI. Unfortunately, this may not only facilitate benevolent but also malevolent use. Effective regulation and risk management for responsible AI are, however, not a given.
- AI embodiments can be — depending on the use case — unnecessarily humanoid in their design, bringing in additional complexity and challenges.
- Embodied AI requires multidisciplinary collaboration between experts in areas as diverse as machine learning, GUI design and mechanical engineering.

User Recommendations

- Identify use cases that may benefit from applying embodied AI, both in more virtual domains, such as online customer interaction or knowledge worker augmentation, and in more physical domains, such as manufacturing, logistics, healthcare or facility management.
- Explore the value that embodied AI can add by reducing the limitations of current AI in terms of better interpretation of, for example, physical constraints in a warehouse or cultural norms in client interaction. This may result in increased safety or decreased bias in the use of AI, respectively.
- Extend the mindset of how AI agents should be developed or trained. Move from a modeling-only approach toward one that considers how intelligence can be a synergy between AI models and the design of the agent's embodiment. This could, for example, relate to the facial expression of virtual agents, or the coordination of movement in physical agents.

Sample Vendors

DEEP Robotics; Figure AI; Guerrilla Games; Intrinsic; NVIDIA; Qualcomm; Sereact; Toshiba; Unitree; Wayve

Gartner Recommended Reading

Innovation Insight: AI Agents

Innovation Insight: AI Simulation

Hype Cycle for Intralogistics Smart Robots and Drones, 2025

Top Strategic Technology Trends for 2025: Polyfunctional Robots

AI Simulation

Analysis By: Leinar Ramos, Jim Hare, Anthony Mullen

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

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 usually 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 value by broadening AI use to cases where data is scarce, private or confidential, using simulation to generate synthetic data (for example, synthetic data for autonomous driving training)

- Greater efficiency by leveraging AI to decrease the time and cost to create and use complex and realistic simulations
- Greater robustness by using simulation to generate diverse scenarios, increasing 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 techniques, like physics-based 3D simulation, are uniquely positioned to generate diverse AI training datasets. 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.
- **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.
- **Research in learned simulations (known as "world models") is driving interest in AI simulation.** Research is increasing on training world models that can learn to predict how the environment will evolve, based on its current state and agents' actions. These learned simulations could make AI simulation more feasible by not having to directly specify simulation parameters.
- **The emergence of embodied AI is increasing the need for AI simulation.** Simulation environments are often the primary way to train embodied AI (AI adapted to its physical or virtual context) via reinforcement learning. The increased interest in embodied AI systems, like robots, is driving AI simulation demand.
- **Increased technical debt in AI drives the need for the reusable environments that simulation provides.** 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, making 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 and 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. There will also be skepticism of the quality and accuracy of simulations, limiting potential adoption.
- **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 by creating realistic content for simulations, including images, videos and text. In addition, AI agents can be embedded in a simulation to make it more realistic.
- Create synergies between AI and simulation teams, projects and solutions to enable a new 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; The AnyLogic Company; Cosmo Tech; Epic Games; MathWorks; Microsoft; NVIDIA; Rockwell Automation; Unity

Gartner Recommended Reading

Innovation Insight: AI Simulation

World Models

Analysis By: Mike Fang, Nick Ingelbrecht, Sushovan Mukhopadhyay

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

World models are learned abstract representations of an environment. They enable AI systems to make predictions via simulating potential future states and helping to understand the consequences of the actions taken.

Why This Is Important

AI systems struggle to function effectively in physical environments due to challenges such as safety concerns, restricted data coverage, limited adaptivity to novel situations, and the absence of cause-and-effect reasoning capabilities. World models are fundamental for efficiently forming representations based on the environment, constructing plans, and simulating events and their outcomes. They offer insights into potential effects of actions in the environment, which are crucial for AI agents.

Business Impact

- By capturing the underlying principles and regularities of the environment, world models can enable the simulation and anticipation of future states and outcomes based on current conditions and actions. This allows AI systems to acquire knowledge, refine their models and apply learned insights to new situations for informed decision making, even in unfamiliar contexts.
- World models could provide AI applications with a controlled environment for experimentation, allowing researchers and developers to explore different strategies, algorithms and policies before deploying them in the real world.

Drivers

- World models have applicability across various sectors, from film production to autonomous vehicles and robotics. Their ability to enable simulation and anticipate complex interactions makes them invaluable tools for AI agents to achieve innovation and efficiency in diverse fields.
- World models empower AI to perform more sophisticated prediction and planning tasks, moving beyond mere pattern recognition in observed data. By simulating and understanding the dynamics of environments, AI can better handle uncertainty or missing information and therefore make informed decisions that account for future possibilities and contingencies.
- These models can be used to enhance the realism and credibility of generated video content by incorporating physical laws and constraints. This ensures that the produced visuals adhere to the principles of physics, resulting in more believable and immersive experiences.
- Trained on extensive multimodal datasets derived from robots functioning in real-world scenarios while combining first-principle AI capabilities, world models can guide robots in object manipulation and interaction with their environments.
- World models assist embodied AI in comprehending associations, counterfactuals, interactions and modeling the dynamics of the world. They go beyond summarizing observed content by efficiently simulating potential scenarios to predict outcomes, thereby enabling the selection of optimal actions.
- Research from control theory and cognitive science, such as Joint-Embedding Predictive Architecture (JEPA), has highlighted alternative approaches to construct world models.

Obstacles

- Simulating real-world environments and inferring causal relationships is one of the most challenging domains of AI, and therefore building world models. Counterfactual reasoning requires simulating hypothetical causes and predicting outcomes, but current models are limited.
- Simulating physical laws is challenging for world models, especially in capturing real-world physical rules. Existing synthetic video generation models like Sora simulate phenomena like object motion and light reflections, but struggle with complex physical effects like fluid dynamics and aerodynamics, lacking accuracy and consistency.
- Techniques supporting world models mainly interpolate data, not extrapolate. The real world has many uncertainties, making world models difficult to build.
- Unlike humans, world models need a very large amount of situational and contextual combinations, leading to high computational costs. Additionally, acquiring real-world data faces challenges like public availability, security and privacy issues.

User Recommendations

- Avoid relying solely on GenAI techniques for world modeling as a solution for every use case; instead, leverage a broad array of methods from both physical AI and cognitive science to create a more comprehensive and effective model.
- Utilize extensive multimodal datasets, including sensory inputs like images and sounds, to train or customize world models for better contextual understanding and decision making across diverse scenarios.
- Implement strategies to mitigate bias and ethical issues in world models, ensuring fair and unbiased decision making in AI systems.
- Manage expectations around these techniques, as they are still surrounded by hype. Begin by piloting them in more focused or non-mission-critical use cases through mini world models limited to constrained environments (like in game situations).

Sample Vendors

Covariant; Decart; Google; Meta; NVIDIA; OpenAI; VERSES; World Labs

Gartner Recommended Reading

Innovation Insight: Causal AI

Innovation Insight: AI Simulation

Emerging Tech: Generative AI's Path to Revolutionary Computer Vision Products

Decision Intelligence

Analysis By: David Pidsley, Pieter den Hamer, Erick Brethenoux

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Decision intelligence (DI) is a practical discipline that advances decision making by explicitly understanding and engineering how decisions are made, and how outcomes are evaluated, managed and improved via feedback. By digitizing and modeling decisions as assets, DI bridges the insight-to-action gap to continuously improve decision quality, actions and outcomes. DI is technology-agnostic and applies decision-centric frameworks like observe, orient, decide and act (OODA) and Gartner DI (GDI).

Why This Is Important

Agentic AI and generative AI (GenAI) hype, regulatory pressures on decision automation, and recent global crises have exposed the fragility of business processes and the predigital, implicit and suboptimal ways of decision making that remain incumbent. DI is positioned beyond the trigger, poised to address these challenges by making decisions more explicit, optimal, adaptable and auditable.

Business Impact

- Faster, higher quality decisions that are consistent, compliant and cost-effective while being complex, contextual and continuous, thus driving agility in facing opportunities and threats in domains like banking, healthcare and supply chain.
- Enduring, effective, efficient, explainable and ethical enterprisewide DI execution enhances timely stakeholder outcomes.
- Risk is mitigated through accurate, trustworthy, fair, privacy-protective and scalable decision-centric operationalization of AI to augment and automate decisions.
- Adaptability of decisions as assets strengthens decision governance and outcome predictability.

Drivers

- **Dynamic business complexity:** Unpredictable disruptions, chaotic environments and accelerating pace of digital competition demands near real-time decision models that can adapt. Decision services can be powered by the composition of multimodal data analysis, data science, optimization, expert knowledge and other AI techniques.
- **Decision silos:** DI curtails fragmented, localized and implicit decisions that undermine organizational efficacy and efficiency. It also addresses the demand for cross-functional alignment on decisions as assets, the need for harmonization on which action should be taken following a business decision, and outcome optimization that balances global efficiency and local adaptations.
- **Deluge of dashboards not driving action:** Despite proliferation of "data-driven" tools and interfaces, most of which fail to connect insights to actions, dashboard development delays create decision latency, ambiguous outcomes and inability to perceive a decision's impact harming organization efficiency.
- **Human-AI delegation and distrust:** AI adoption requires transparent, auditable decision models to address ethical concerns and ensure accountability. Automating human decisions has promoted disquiet and requires monitoring.
- **Regulatory scrutiny:** Data protection, AI and socio-environmental mandates compel explicit decision documentation for tighter compliance, risk awareness and mitigation. Explicit decision modeling and decision stewardship drive the analysis, management and control of the operational processes and observations needed to enforce decision governance policies and standards applied to decisions as assets.
- **Availability and innovation of enabling technologies:** Convergence of rule engines, simulation and optimization in DI platforms practically enables DI prototypes and pilots to become scalable DI implementations.
- **GenAI acceleration of DI:** Enriched context awareness via LLMs is accelerating composite AI model development for low-code/no-code business decision analysts and pilots of agentic decision automation.

Obstacles

- Business stakeholder apathy, limited urgency and low cultural readiness, ineffective change management, and lack of DI skills and AI literacy hinder adoption.
- Bridging the insight-to-action gap to improve outcomes requires a decision-centric vision beyond the data-driven dogma and the data-to-insight workflow. Technology centricity overlooks psychological and sociological factors in decision making.
- Weak collaboration, inadequate operating, delivery and organizational models (i.e., a DI center of excellence), and disconnected decision-making silos hamper DI effectiveness. Even advanced cross-silo DI practitioners struggle to impartially reconsider key decision flows.
- Unselective or overly enthusiastic adoption of decision automation introduces risks, including unintended consequences, loss of context and bias amplification. This undermines trust in DI and limits effective use of DI platforms.

User Recommendations

- **Define and model critical decisions** involving resource allocation, uncertainty or competing alternatives. Use these as pilots to build DI momentum and demonstrate value for enterprisewide adoption.
- **Inventory repetitive, high-impact decisions** and their key inputs. Adopt decision-centric modeling by articulating outcomes, decision logic, alternative courses of actions and required observations to drive continuous learning, improvement and transparency.
- **Maximize decision quality, resilience and traceability** through cross-functional DI fusion teams, fostering collaboration and alignment across departments. Delegate decision-making authority to those with the most relevant expertise and context.
- **Upskill staff in decision modeling, prescriptive analytics and optimization.** Investigate the roles of decision engineers, decision scientists and decision stewards. Experiment with agentic, GenAI and other composite AI, and DecisionOps to support organizationwide decision centricity and excellence.

Gartner Recommended Reading

Market Guide for Decision Intelligence Platforms

Innovation Insight: Decision Intelligence for Chinese Organizations

Quick Answer: Where to Reengineer Your Business Decisions

How to Manage the Risks of Decision Automation

When to Automate or Augment Decision Making

FinOps for AI

Analysis By: Jim Hare, Adam Ronthal, Andrei Razvan Sachelarescu

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

FinOps for AI is the application of financial operations best practices to help organizations increase visibility and manage the costs of AI services to ensure efficient usage and deliver maximum business impact. Using FinOps to track and measure AI spend and usage is crucial for optimizing costs, ensuring financial accountability and maximizing ROI.

Why This Is Important

Cost poses one of the greatest near-term threats to AI and GenAI success. AI workloads, especially in cloud environments, often use expensive GPU-based compute infrastructure and consume tokens in unforeseen ways, leading to unpredictable expenses if not monitored properly. Deploying and managing AI solutions generates other costs, including development, governance and change management. Using FinOps to track and measure AI spend and usage is crucial for optimizing costs, ensuring financial accountability and maximizing ROI.

Business Impact

FinOps helps businesses optimize AI spend by providing real-time cost visibility and control, enabling teams to allocate resources efficiently and prevent budget overruns while also preventing underprovisioning that can cause downtime or slowdowns. FinOps for AI also enhances collaboration between finance, engineering and operations teams, ensuring that AI investments align with business objectives while ensuring cost-efficiency. Using FinOps practices, organizations can maximize the ROI of AI initiatives and leverage cost-saving opportunities such as reserved instances, workload automation/optimization and usage-based pricing models.

Drivers

- AI adoption, especially AI applications and GenAI, is contributing to a spike in cloud costs for most enterprises. Hidden costs and unpredictable invoices make it difficult for organizations to deploy AI more broadly.
- Tracking AI costs and usage scaling is complex due to fluctuating computational demands, variable AI service pricing, hidden infrastructure costs and exponential scaling of model training and inference across users and applications.
- Organizations new to AI and/or the cloud are unlikely to be prepared for AI cost volatility and will need to adjust their legacy operating models and budget practices by adopting FinOps for AI. Many organizations face challenges in tracking and measuring AI costs against concrete business benefits.
- Engineering teams are often immature in their use of AI services and the many dynamic layers needed to achieve ongoing cost-effectiveness.
- The total cost of ownership (TCO) of AI use cases can differ from the cost of traditional software applications with fixed costs and purpose. Continuous training, switching to newer models, specialized infrastructure like GPUs and differences in processing costs for specific data types (text, image, video, audio) are part of ongoing AI costs.
- Many AI models and services are based on consumption pricing models and may be purchased in many versions or variants.
- Pricing may also fluctuate based on a variety of factors such as usage, model choice, accuracy and performance guarantees. The velocity of pricing volatility requires continuous and active assessments of price/performance and accuracy.

Obstacles

- Implementing FinOps for AI is challenging because of AI workloads' unpredictable and dynamic nature and the complexity and variety of cost factors that make cost estimation, budgeting and optimization more challenging compared with traditional cloud operations.
- Balancing performance and cost-efficiency is difficult because AI models often require specialized compute infrastructure resources, GPUs and large datasets that can lead to excessive cloud spending if they are not monitored and optimized effectively.
- Many organizations struggle with cross-functional collaboration among finance, operations and engineering teams, as aligning AI-specific cost insights with business objectives requires a cultural shift and enhanced visibility into AI-driven expenditures.

User Recommendations

- **Track the TCO of using AI:** Implement real-time tracking of total running costs, including cloud, infrastructure and labor costs. Use tagging and cost allocation strategies to assign expenses to specific AI projects or departments. Assign budgets to AI-related resource and service groups, and trigger cost alerts when consumption exceeds budget goals.
- **Optimize AI spend and workloads:** Track AI spend across packaged and custom SaaS, AI-leveraging commercial models (tokens via API calls), and compute from hosted models.
- **Understand the pros and cons of buying versus building models:** Closed models built by model providers may be considered more expensive at first glance, but they reduce delivery time, upfront development costs and the need for more expensive skills. Build models for truly strategic types of use cases.
- **Embrace an agile approach to model switching:** Regularly compare models in use with alternative options to see whether the same or better accuracy can be achieved at lower cost.
- **Invest in making data AI-ready:** Control data preparation and processing costs by investing in data cleansing and curation to produce smaller training and retrieval-augmented generation datasets of higher quality.
- **Implement proactive cost management controls and guardrails:** Integrate real-time anomaly detection and alerts with demand-throttling options to guard against unexpected cost spikes.

Sample Vendors

Airia; Exostellar; FinOps Foundation; Finout; Flexera; IBM

Gartner Recommended Reading

Beyond FinOps: Optimizing Your Public Cloud Costs

The State of FinOps for Data and Analytics, 2024

Emerging Tech: Data Management Solutions Need Augmented FinOps

Key Data and Analytics Platform Insights for FinOps Providers

Data and Analytics Essentials: FinOps and Cloud Operating Models

Neurosymbolic AI

Analysis By: Erick Brethenoux, Afraz Jaffri

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Neurosymbolic AI is a form of composite AI that combines probabilistic reasoning methods and symbolic systems to create more robust and trustworthy AI models. This fusion enables the combination of probabilistic models with logic-based techniques (such as rules and knowledge graphs) to enable AI systems to better represent, reason and generalize concepts. This approach provides a reasoning infrastructure for solving a wider range of business problems more effectively.

Why This Is Important

Neurosymbolic AI addresses limitations in current AI systems, such as incorrect outputs, lack of generalization to a variety of tasks and an inability to explain the steps that led to an output. The neurosymbolic approach leads to more powerful, versatile and interpretable AI solutions and allows AI systems to reason through more complex tasks. Generative AI systems are starting to leverage neurosymbolic methods to overcome their reasoning shortcomings.

Business Impact

Neurosymbolic AI will have an impact on the efficiency, adaptability and reliability of AI systems used across business processes. The integration of logic and multiple reasoning mechanisms brings down the need for ever larger AI models and their supporting infrastructure. These systems will rely less on the processing of huge amounts of data, making AI agile and resilient. Neurosymbolic approaches can augment and automate decision making with less risk of unintended consequences.

Drivers

- Neurosymbolic AI addresses the limitations of large reasoning models (LRMs), which are still plagued with a lack of symbolic abstraction when exclusively based on deep learning techniques.
- The need for explanation and interpretability of AI outputs is especially important in regulated industry use cases and in systems that use private data.
- Understanding the meanings behind words, not just their arrangement (semantics over syntax), is an increasing priority in systems that deal with real-world entities to ground meaning to words and terms in specific domains.
- The set of tools available to combine different types of AI models is increasing and becoming easier to use for developers and end users. The dominant approach is to chain together results from different models (composite AI) rather than using single models.
- The integration of multiple reasoning mechanisms necessary to provide agile AI systems eventually leads to adaptive AI systems, notably through blackboardlike mechanisms.
- Agentic AI advances also participate in advancing neurosymbolic methods, while agents using various composite AI techniques collaborate to solve problems.

Obstacles

- Most fundamental neurosymbolic AI methods and techniques are being developed in academia or industry research labs. Despite the increased availability of tools, implementations in business or enterprise settings are still limited.
- No agreed-upon techniques exist for implementing neurosymbolic AI, and disagreements continue between researchers and practitioners on the effectiveness of combining approaches, despite the emergence of real-world use cases.
- The commercial and investment trajectories for AI startups allocate almost all capital to deep-learning approaches, leaving only those willing to bet on the future to invest in neurosymbolic AI development.
- Currently, despite increasing exposure, popular media and academic conferences do not give as much exposure to the neurosymbolic AI movement as compared to other approaches (such as generative AI).

User Recommendations

- Adopt composite AI approaches when building AI systems by using a range of techniques that increase the robustness and reliability of AI models. Neurosymbolic AI approaches will fit into a composite AI architecture.
- Dedicate time to learning about neurosymbolic AI approaches, and to identifying use cases that can benefit from applying these approaches.
- Invest in data architecture that can leverage the building blocks for neurosymbolic AI techniques, such as knowledge graphs and agent-based techniques.
- Consider neurosymbolic AI architectures when the limitations of generative AI models prevent their implementation in the organization.
- Educate developers on the potential of neurosymbolic models by exploring the capabilities of neurosymbolic approaches while building learning AI agents.

Sample Vendors

Franz; Google DeepMind; IBM; Microsoft; RelationalAI; Wolfram|Alpha

Gartner Recommended Reading

Innovation Insight: AI Simulation

Go Beyond Machine Learning and Leverage Other AI Approaches

When Not to Use Generative AI

AI Zodiac: Mapping AI Use Cases to Techniques

Artificial General Intelligence

Analysis By: Pieter den Hamer, Philip Walsh

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Artificial general intelligence (AGI) is the (currently hypothetical) capability of a machine that can match or surpass the capabilities of humans across all cognitive tasks. In addition, AGI will be able to autonomously learn and adapt in pursuit of predetermined or novel goals in a wide range of both physical and virtual environments.

Why This Is Important

With AI's growing sophistication — including the recent advances in generative AI (GenAI) and agentic AI — a growing number of AI experts have shortened their predicted timelines for achieving AGI in the future or view AGI as no longer purely hypothetical. A clear, shared definition of AGI is necessary for evidence-based governance and realistic expectations. Achieving AGI would be a transformative tipping-point with profound consequences for productivity, employment, geopolitical power, legal, ethical and cultural norms — and society at large.

Business Impact

In the near term, anticipation of AGI fuels both overly optimistic expectations and existential fears, skewing investment, distorting trust and accelerating the emergence of new AI regulations. Over the longer horizon, the question of who builds and controls AGI — or other forms of increasingly powerful AI — looms large. Many experts see public stewardship as essential, a prospect that could upend private advantage and redraw entire markets.

Drivers

- Recent advances and growing interest in multimodal large language models (LLMs), so-called reasoning models and AI agents drive considerable hype about AGI. The massive scaling of deep learning and the availability of huge amounts of data and compute power largely have enabled these advances.
- AI's further evolution toward AGI, as defined here, is increasingly complemented by other partially new approaches, such as knowledge or causal graphs, world models, adaptive AI, embodied AI, composite and neurosymbolic AI, and likely other innovations yet unknown.
- A number of AI vendors are openly discussing and actively researching the field of AGI, creating the impression that AGI lies within reach. However, their definitions of AGI vary greatly and are often open to multiple interpretations. Moreover, other leading AI vendors and experts have dismissed AGI as hype, urging focus on the real impact of AI's growing capabilities.
- 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 nonhuman entities also applies to AI-powered machines. The humanlike responses of LLMs and the reasoning-like capabilities of recent AI models have fueled this tendency. Although many philosophers, neuropsychologists and other scientists consider this attribution highly uncertain or going too far, 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

- Little scientific consensus exists on the meaning of “human intelligence.” Any claims about AGI are hard to validate in the face of the enormous complexity of the human brain and mind and such a limited understanding of them.
- Unreliability, lack of transparency and limited abstraction and reasoning of pattern-based capabilities in current AI are not easy to overcome with deep learning’s intrinsically probabilistic approach. More data or more compute power for ever-bigger models is unlikely to resolve these issues, let alone to achieve AGI. To realize (and control) AGI will require further technological innovations. Therefore, AGI as defined here is unlikely to emerge in the near future.
- If AGI materializes, autonomous actors likely will emerge that, in time, will be attributed with full self-learning, agency, identity and perhaps even morality. This will open a bevy of considerations about AI’s legal rights and trigger profound ethical and even religious discussions. AGI also brings the risk of negative impacts on humans, from job losses to a new, AI-triggered arms race and more. A serious backlash may result, and regulations to ban or control AGI are likely to emerge in the near future.

User Recommendations

- Engage with stakeholders to address excessive optimism or unwarranted pessimism, and create or maintain realistic expectations around AGI. Ground AI strategy in concrete business problems rather than speculative AGI forecasts. Recalibrate the AI portfolio periodically as AI capabilities evolve, while leveraging the complementary strengths of human and artificial intelligence.
- Stay apprised of scientific and innovative breakthroughs that may indicate AGI’s possible emergence; however, be aware of the broad range of definitions and views regarding AGI, some strict and some less strict. Meanwhile, keep applying current AI to learn, reap its benefits and develop practices for its responsible use.
- Assess whether AI systems truly meet their specific use-case needs, rather than relying on generic measures of intelligence.
- Prepare for emerging AI regulations and promote internal AI governance to manage current and emerging AI risks. Because although AGI as defined here is not a reality now, current AI already poses significant risks regarding ethics, reliability and other areas.

Sample Vendors

Aigo; Amazon; Anthropic; Butterfly Effect; DeepSeek; Google; Microsoft; OpenAI

Gartner Recommended Reading

Artificial General Intelligence: Advanced Discussion Using 5 Future Scenarios

Video: Why Artificial General Intelligence (AGI) Matters — and Why It Doesn't

How to Think About Artificial General Intelligence

Composite AI

Analysis By: Erick Brethenoux, Pieter den Hamer

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Composite AI, also known as hybrid AI, refers to the combined application (or fusion) of different AI techniques to improve the efficiency of learning and broaden the level of knowledge representations. It broadens AI abstraction mechanisms and, ultimately, provides a platform to solve a wider range of business problems effectively.

Why This Is Important

Composite AI recognizes that no single AI technique is a panacea. It combines "connectionist" AI approaches, like machine learning (ML) and deep learning, with "symbolic" and other AI approaches, like rule- and logic-based reasoning, graph or optimization techniques. The goal is to enable AI solutions to generalize and learn, embodying more abstraction mechanisms. Composite AI is at the center of the generative AI (GenAI), decision intelligence (DI) platform and agentic AI markets.

Business Impact

Composite AI brings the power of AI to a broader group of organizations that do not have access to large amounts of historical or labeled data but possess significant human expertise. It helps to expand the scope and quality of AI applications addressing more types of reasoning challenges. Other benefits include better interpretability, embedded resilience and the support of augmented intelligence. The new wave of GenAI implementations heavily relies on composite AI.

Drivers

- The growing reliance on AI for decision making is driving organizations toward composite AI. The most appropriate actions can be determined by combining rule-based and optimization models — a combination often referred to as prescriptive analytics.
- Small datasets, or the limited availability of data, have pushed organizations to combine multiple AI techniques. Enterprises have started to complement scarce raw historical data with additional AI techniques, such as knowledge graphs and generative adversarial networks (GANs), to generate synthetic data.
- Combining AI techniques is much more effective than relying only on heuristics or a crudely "data-driven" approach. A rule- and logic-based technique can be combined with a deep learning model. Rules coming from human experts, or the application of physical/engineering model analysis, may specify that certain sensor readings indicate inefficient asset operations.
- Computer vision and natural language processing (NLP) solutions are used to identify and categorize people or objects in an image. This output can be used to enrich or generate a graph, representing the image entities and their relationships.
- The advent of DI platforms and AI agent-based systems will promote the efficient combination of multiple AI techniques, making composite AI (and in some instances neurosymbolic techniques) a de facto AI development approach.
- The development of multiagent systems is further empowering composite AI. A composite AI solution can be composed of multiple agents, each representing an actor in the ecosystem. Combining these agents into a "swarm" enables the creation of common situation awareness, more global planning optimization, responsive scheduling and process resilience.
- GenAI is accelerating the research and adoption of composite AI models through artifacts, process and collaboration generation, which are the foundation of DI platforms.

Obstacles

- **Lack of awareness and skills in leveraging multiple AI methods:** This could prevent organizations from considering the techniques particularly suited to solving specific problem types.
- **ModelOps deployment:** The ModelOps domain (that is, the operationalization of multiple AI models, such as optimization models, ML models, rule models and graph models) remains an art much more than a science. A robust ModelOps approach is required to efficiently govern composite AI environments and harmonize it with other disciplines, such as DevOps and DataOps.
- **Trust and risk barriers:** The AI engineering discipline is starting to take shape, but only mature organizations apply its benefits in operationalizing AI techniques. Security, ethical model behaviors, observability, model autonomy and change management practices must be addressed across the combined AI techniques.

User Recommendations

- Identify projects in which a crude "data-driven," ML-only approach is inefficient or ill-fitting. For example, in cases when enough of the right data is not available or when the pattern cannot be represented through current ML models.
- Capture domain knowledge and human expertise to provide context for data-driven insights by applying decision intelligence with business rules and knowledge graphs, in conjunction with ML and/or causal models.
- Combine the power of ML, image recognition or NLP with graph analytics to add higher-level, symbolic and relational intelligence.
- Extend the skills of ML experts, or recruit or upskill additional AI experts to cover graph analytics, optimization or other DI-relevant techniques for composite AI. For rules and heuristics, consider knowledge engineering skills, as well as emerging skills such as prompt engineering.
- Accelerate the development of DI solutions by encouraging experimentation with agentic and GenAI, which will in turn accelerate the necessity for deployment of composite AI solutions.

Sample Vendors

ACTICO; Aera Technology; Almawave; FICO; Filuta; Fujitsu; IBM; Indico Data; SAS

Gartner Recommended Reading

Go Beyond Machine Learning and Leverage Other AI Approaches

When Not to Use Generative AI

Top Strategic Technology Trends for 2022: AI Engineering

How to Choose Your Best-Fit Decision Management Suite Vendor

At the Peak

AI TRiSM

Analysis By: Avivah Litan, Jeremy D'Hoinne, Bart Willemsen, Lauren Kornutick, Max Goss, Andrew Bales, Sumit Agarwal

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

AI trust, risk and security management (TRiSM) comprises four layers of technical capabilities that support enterprise policies for all AI use cases and help assure AI governance, trustworthiness, fairness, safety, reliability, security, privacy and data protection. The top two layers — AI governance, and AI runtime inspection and enforcement — are new to AI and are, in part, consolidating into a distinct market segment. The bottom two layers represent traditional technology focused on AI.

Why This Is Important

AI brings new trust, risk and security management challenges that conventional controls do not address. Enterprises are most concerned with data compromise, third-party risks, and inaccurate or unwanted outputs, and need to ensure enterprise AI behavior and actions align with enterprise intent. Enterprises must further retain independence from any single AI model or hosting provider to ensure scalability, flexibility, cost control and trust, as AI markets rapidly mature and change.

Business Impact

Organizations that do not consistently manage AI risks are exponentially inclined to experience adverse outcomes such as project failures, AI misperformance and compromised data confidentiality. Inaccurate, unethical or unintended AI outcomes, process errors, uncontrolled biases, and interference from benign or 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 or incorrect business decisions.

Drivers

- The increasing use of AI and GenAI is limited by a lack of trust in AI as a safe and ethical option for supporting critical business processes.
- Enterprises face multiple AI risks and are most concerned with data compromise, third-party risks, and inaccurate or unwanted outputs.
- Malicious hacks against enterprise AI are still uncommon, while incidents of unconstrained harmful chatbots are well-documented, and internal oversharing data compromises are prevalent.
- User demand for GenAI TRiSM solutions is steadily increasing, and providers of all sizes are competing for this new enterprise business.
- Some organizations are mostly concerned with security and risk mitigation, while others also focus on supporting ethical or safe practices and regulatory compliance.
- AI trust, risk and security issues surface organizational silo issues, pushing teams to realign to solve problems that cross departmental boundaries and to implement technical measures that address them.
- The introduction of LLM-based AI agents with differing degrees of agency allows AI to take actions with or without human intervention.
- The rapid proliferation of AI agents will create more need for governance than human-in-the-loop oversight can fulfill alone.
- Regulations for AI risk management are driving businesses to institute measures for managing AI risk. Such regulations define new compliance requirements that organizations will have to meet on top of existing ones, like those pertaining to privacy protection.

Obstacles

- Adopting AI TRiSM technology is often an afterthought. Organizations generally don't consider it until AI applications are in production, when it becomes challenging to retrofit.
- Many enterprises are resource-constrained and don't have the skills or capacity to implement TRiSM technology.
- Enterprises often rely on their incumbent vendors to provide AI functionality for TRiSM capabilities though they often lack it, and must rely on vendor licensing agreements to ensure their confidential data remains private in the host environment.

- Off-the-shelf software that embeds AI is often closed and does not support APIs to third-party AI TRiSM products that can enforce enterprise policies.
- Most AI threats and risks 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.
- While AI TRiSM can integrate life cycle controls, this poses significant implementation challenges especially with embedded AI that often lacks API support.

User Recommendations

- Set up an organizational unit to manage AI TRiSM that is integrated with the organization's overarching governance initiative. Include members with a vested interest in AI projects.
- Define acceptable use policies at a level granular enough to enforce.
- Discover and inventory all AI used in the organization, leveraging the capabilities of AI TRiSM vendors who support this.
- Revisit and implement data classification, protection and access management across all enterprise information — both structured and unstructured — that can potentially be utilized by AI. Collaborate across different teams involved in information governance.
- Work with legal and compliance to contractually define accountability and responsibility for unacceptable AI use or behavior in third-party-embedded AI applications.
- Obtain vendor attestation to meet legal requirements.
- Evaluate and implement layered AI TRiSM technology to continuously support and enforce policies across all AI entities in use. This includes enterprise-owned AI TRiSM services, along with controls offered by frontier model providers, but do not solely rely on the latter.

Gartner Recommended Reading

Market Guide for AI Trust, Risk and Security Management

Top Strategic Technology Trends for 2024: AI Trust, Risk and Security Management

The Cybersecurity Leader's Playbook: Navigating the EU AI Act

Multimodal AI

Analysis By: Nick Ingelbrecht, Sushovan Mukhopadhyay, Yogesh Bhatt

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

Multimodal AI models are trained with multiple types of data (also known as modalities) simultaneously, such as images, video, audio and text. This enables them to create a shared data representation to improve performance in different tasks. At runtime, they can handle more than one modality, either in their inputs, their outputs or both.

Why This Is Important

Multimodal AI adds significant new technology capabilities such as greater accuracy to existing software. It spurs new specialized applications, enables new use cases such as visual question answering of image frames, manufacturing optimization, and fraud detection in banking and finance, and creates new value outcomes. The physical world and the data it generates are inherently multimodal. By integrating and analyzing diverse data sources, a more comprehensive evaluation of complex environments and tasks can be achieved compared with unimodal models, helping users make sense of the world and opening up new avenues for AI applications.

Business Impact

Gartner forecasts that:

- Over the next five years, multimodal AI will become increasingly integral to capability advancement in every application and software product across all industries.
- By 2027, 40% of generative AI (GenAI) solutions will be multimodal (text, images, audio and video), up from 1% in 2023.

Drivers

Multimodal AI adoption will generate cross sector transformational opportunities. Key drivers include:

- A paradigm shift from traditional, linear processes to dynamic, AI-driven systems where humans and machines collaborate seamlessly. And further evolution of agentic AI will involve increasing integration with multimodal AI techniques to handle the complexity and richness of real-world data and tasks.
- Recent AI breakthroughs, particularly in the realm of large language models (LLMs) and vision language models (VLMs), are highly relevant to multimodal AI. These advancements have catalyzed a renaissance in natural language processing and computer vision.
- Intelligent applications, by their nature, are context-rich and designed to adapt to constantly changing scenarios. This makes multimodal AI a crucial component for their development and evolution.
- World models are a significant driver for multimodal AI because they inherently require the ability to process and understand information from various modalities to accurately represent and simulate the complexity of the real world.
- Broader availability of AI/GenAI multimodal models, both proprietary and open-source, lowers the barriers to entry and adoption via AI marketplaces.
- There is a demand for multimodal domain-specialized models in areas such as healthcare, where multimodality extends or enriches use cases.

Obstacles

Multimodal AI is powerful in understanding and processing from various modalities, but faces several primary obstacles to adoption:

- Integrating diverse data types — such as text, images, audio and video — is challenging due to differences in format and time stamps, risking inaccurate interpretations. Multimodal AI models are complex, combining various modality-specific subnetworks, which can obscure transparency and explainability.
- Architectural complexity, increased data volume and the need for data fusion lead to inference latency, hindering reliable operation where immediate decision making is crucial.
- Dataset bias originates from leveraging training datasets like text, images, videos and speech, which may inadvertently reflect societal or cultural biases. This can result in making unfair or inaccurate predictions/decisions.

- Handling sensitive data across modalities increases breach risks and privacy violations. This complicates compliance with regulations like General Data Protection Regulation (GDPR) or Health Insurance Portability and Accountability Act (HIPAA), as multimodal AI exposes new attack surfaces and heightens privacy risks with diverse data types.

User Recommendations

Organizations looking to implement multimodal AI should:

- Identify AI use cases where multimodal AI can enhance business value beyond unimodal AI foundation models.
- Run pilots with off-the-shelf multimodal models to demonstrate not only technical feasibility but the business value.
- Build a strong model evaluation by assessing the quality of relationships between modalities such as comparing generated captions from images to ground-truth labels/descriptions.
- Prioritize building or accessing robust data infrastructure supporting the collection, storage and processing of diverse data types (text, images, audio and video).
- Build or acquire expertise to handle the technical complexities of processing and integrating multimodal data with legacy and existing workflows.
- Create or extend AI governance strategies and policies to address challenges with multimodal datasets and ensure compliance.
- Incorporate multimodality into technology roadmaps and create migration paths for multimodal AI in systems procurement or product development plans.

Sample Vendors

Aimesoft; Google; Hugging Face; Jina AI; Meta; Midjourney; NVIDIA; OpenAI; Stability AI; TwelveLabs

Gartner Recommended Reading

Emerging Tech: Generative AI's Path to Revolutionary Computer Vision Products

Emerging Tech: Multimodal Generative AI Interfaces Transform User Experiences

Innovation Insight: Multimodal AI Explained

Sovereign AI

Analysis By: Lydia Clougherty Jones, Clementine Valayer

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Sovereign AI is the effort by nation-states to invest in and progress their own development and use of AI to advance their unique sovereign objectives. Given the wide variances across sovereign AI innovation to risk ratios, sovereign AI impacts international relationships, global trade, local public/private partnerships and economic markets in unexpected ways.

Why This Is Important

Sovereign AI reflects the rapid acceleration of nation-states' investment to advance adoption of AI techniques for their own use, including to improve alignment of their internal government functions with operational goals. While sovereign AI could enhance an individual state's military defense, AI use by other nations could undermine those national security efforts. Sovereign AI aims to maximize AI value while decreasing AI risk, including for those sovereign states that collaborate to achieve common goals such as decreasing the impact of AI-generated deepfakes in political environments.

Business Impact

Sovereign AI impacts nearly all aspects of government and the enterprises with which it interacts. It improves the effectiveness of operations by automating tasks like those within government contact centers. Because sovereign AI modernizes the business of government, it can improve employee experience and accelerate citizen engagement. When sovereign states control their own AI systems, they reduce their dependence on other sovereign states and on the private tech market.

Drivers

- An increasing number of countries are actively planning and building their own AI infrastructure and capabilities to increase their competitiveness and safeguard their future, including by developing sustainable sovereign AI.
- Known and unknown risks of harms to sovereign objectives and citizen welfare from irresponsible uses of AI drive sovereign states to desire greater control over the development of AI systems, and more so over generative AI (GenAI) use cases.
- Need for sovereign entities to self-regulate, including how their data is used to train large language models (LLMs), is growing. For example, nation-states are increasingly using AI tools to make important government decisions, but clients report that these decisions are often outsourced to private companies without public input or oversight. This lack of transparency and accountability drives sovereign states to develop the AI tools themselves to address concerns about unwanted biases and conflicts of interest in these critical decision-making processes.
- Nation-states desire to increase their independence from other nations and the tech market, especially to remedy underrepresentation of cultural and linguistic inputs.
- Sovereign AI plays an important role in digital sovereignty as it focuses on the sovereign control of AI data and systems, including control over computing capacity, data storage, access to human resources and proprietary knowledge for AI application development. Digital sovereignty also can significantly impact sovereign AI development, with the availability of locally stored data, for example, to train AI models.
- Sovereign AI differs from sovereign data strategies. Its core focus is being the developer and user of AI technologies, not the regulator of them. Sovereign data strategies reflect state efforts to regulate data about and AI use by its citizens, private industry and its economy.
- Combating threats to political stability from the proliferation of deepfakes can drive nation-states to adopt sovereign AI.
- Upskilling government workers today for a more AI-ready government workforce tomorrow will assist sovereign AI advancement.
- Demand to advance local and national defense strategies is increasing.

- Progressing and maintaining leadership in emerging technologies space will accelerate scaling AI.

Obstacles

- Preparation of an AI-ready IT infrastructure.
- Development of an AI-skilled government workforce.
- Modernization of government culture to embrace advanced analytics and automation.
- Limited capability of already-taxed IT infrastructure and fragmented business networks that would need to serve a sovereign AI system.
- Lack of the right AI-ready data for training LLMs, resulting in AI output with varying levels of utility.
- Lack of technically skilled humans to loop into the AI development and use life cycle, resulting in an increase of unintended negative outcomes from AI use and GenAI artifacts.
- Lack of AI-readiness across people, talent, data management techniques, and enabling tools and technologies, impeding nation-states from accelerating AI adoption across government functions.
- Differences in political and cultural values that will create inconsistent AI-value versus AI-harms analysis, leading to unpredictable impacts on international trade and global markets.
- Legal, regulatory or fiscal policies that can impede sovereign AI initiatives.
- Development of AI by nation-states across the world that leads to a fragmentation and possible contradiction in the requirements for AI solutions, many of which cannot be met by either the public or private sector.

User Recommendations

Sovereign states seeking a self-governed and controlled approach to the development of AI systems aligned with their strategic objectives should:

- Start small and prioritize AI uses aligned with maximum advancement of their stakeholder and government business goals.

- Build an AI strategic roadmap that progresses from internal use cases to citizen-facing ones.
- Ensure the AI strategy identifies key value opportunities and risks.
- Develop early in the planning process a metric of success and include various metrics within the strategy to track successes and course-correct unwanted impacts.
- Cultivate public-private collaborations that foster and accelerate data and analytics, AI upskilling, technology innovation and adoption.
- Monitor and learn from sovereign AI already underway, including from New Zealand, the European Commission, India, the United States, Canada and the United Kingdom.
- Collaborate with friendly sovereign states to accelerate the learning curve, sharing failure analysis and positive narratives of unexpected success.
- Use Gartner's The Pillars of a Successful Artificial Intelligence Strategy to guide the nation toward self-governance of AI development, creating tangible value and achieving competitive national leader status.

Gartner Recommended Reading

Top Trends in AI Public Policy and Regulations for 2024

The Future of AI: Reshaping Society

Quick Answer: Why Is Empathy Critical for Postdigital Government?

Government Insight: U.S. Federal AI Executive Order Opportunities and Risks

Seize the Opportunities: How Can Will U.S. Policy on AI Impact Business Strategy?

AI Agents

Analysis By: Tom Coshow, Haritha Khandabattu

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

AI agents are autonomous or semiautonomous software entities that use AI techniques to perceive, make decisions, take actions and achieve goals in their digital or physical environments.

Why This Is Important

AI agents have the ability to make decisions and take action in their target environment to achieve organizational goals. Using AI practices and techniques such as LLMs, organizations are creating and deploying AI agents to achieve complex tasks.

Business Impact

AI agents have the potential to:

- Revolutionize a broad range of industries and environments with their ability to automate tasks from consumer, industrial, data analytics, content creation and logistics.
- Make informed decisions and interact intelligently with their surroundings.

Drivers

- **Generative AI breakthroughs:** Reasoning models and LAMs advance the ability to plan a complex series of actions.
- **Multimodal understanding:** The ability to use diverse modalities like vision, audio and language enables more general and flexible AI agents. This allows automatic adaptation to changes in the workflow, user interface or API. With this, one can create advanced workflows without explicit programming, significantly reducing the development time and effort for automation.
- **Increased decision-making complexity:** AI is increasingly used in real-world engineering problems containing complex systems, where large networks of interacting parts exhibit emergent behavior that cannot be easily predicted. AI agents can learn, plan and execute in complex environments.
- **Composite AI, including neurosymbolic models:** Advances in models that improve planning and problem solving are enabling more complex AI agents. AI agents can utilize a wide variety of AI practices to forecast, make decisions and plan.

Obstacles

- **Vulnerabilities:** Due to the complexity of the AI agent system, all components face various potential vulnerabilities such as access security, data security and governance.
- **Lack of trust:** Users are unsure whether they can trust the technology to accurately predict and execute tasks independently. Without a human in the loop, agents may take multiple consequential actions in rapid succession and bring about significant impacts before a human notices.
- **Interpretability and oversight:** Action policies may be opaque and have poor explainability, requiring mechanisms for human interpretability, oversight and control.
- **Pace of change:** The technologies deployed, from models to tooling, and the framework options available are changing rapidly, making it difficult for organizations to define their roadmap.

User Recommendations

- Incorporate AI agents into strategic planning by investing in understanding their capabilities and potential applications in various environments, considering their increasing autonomy and wide-ranging usability.
- Investigate the possibilities of utilizing multiagent systems, collectives of AI agents, that can operate both collaboratively and independently, enhancing adaptability and flexibility in response to different tasks and scenarios.
- Promote the development and integration of the use of a variety of AI practices, enabling learning, negotiation and decision-making capabilities.

Sample Vendors

Amazon Web Services; Autogen; CrewAI; Google; LangChain; Maisa; OneReach.ai; Salesforce; UnifyApps; Zhipu AI

Gartner Recommended Reading

Innovation Insight: AI Agents

Build AI Agent Services to Revolutionize Client Operations

Innovation Insight for the AI Agent Platform Landscape

AI Agents: Are You Ready to Set Your AI Free?

Emerging Patterns for Building LLM-Based AI Agents

AI-Ready Data

Analysis By: Roxane Edjlali, Mark Beyer, Svetlana Sicular, Ehtisham Zaidi

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

AI-ready data is determined through the data's ability to prove its fitness for use for specific AI use cases. Proof of readiness comes from the assessment of its representativeness evaluated by its alignment to the use case, support for continuous data qualification, and ensuring data and AI governance. As a result, AI-ready data can only be determined contextually to the AI use case and the AI technique used, which forces new approaches to data management.

Why This Is Important

With the rise of pretrained off-the-shelf models and hype from generative AI (GenAI), data management leaders are at the forefront of creating data strategies for AI. Chief data and analytics officers and data management leaders must quickly respond to rising AI-ready data demands by delivering AI-ready data to support AI use cases. Organizations not investing in AI-ready data practices can increasingly fail to deliver to business objectives and face data and AI governance issues that can lead to erroneous results and financial risk.

Business Impact

Organizations that invest in AI at scale need to evolve their data management practices and capabilities not only to preserve the evergreen classical ideas of data management but also to extend them to AI. It will be critical to provision AI-ready data iteratively to cater to existing and upcoming business demands, ensure trust, avoid risk and compliance issues, preserve intellectual property, and reduce bias and hallucinations.

Drivers

- Data is becoming the main source of differentiation and value from these pretrained models. Models, especially for GenAI, increasingly come from vendors rather than delivered in-house.
- According to the 2024 Gartner Evolution of Data Management Survey, 57% of organizations estimate their data is not AI-ready, and among the remaining 43% that do, the readiness assessment demonstrates gaps.
- According to the 2024 Gartner AI Mandates for the Enterprise Survey, participants report that data availability or quality is the No. 1 barrier to successful AI implementation.
- Rapid progress of AI poses new challenges in organizing and managing AI data. A cycle of augmented data management techniques better suited for meeting AI data requirements is expected. Data ecosystems on the foundation of data fabric architecture indicate the beginning of this new cycle.
- Augmented data management capabilities and tools greatly benefit from AI. AI techniques offer new data-centric approaches, such as automated feature engineering or assisted data engineering and code generation using retrieval-augmented generation.
- GenAI is removing the distinction between structured and unstructured data, thereby requiring data management to adapt to new uses.

Obstacles

- The AI community remains mostly unaware of data management capabilities, practices and tools that can greatly benefit AI development and deployment. The lack of information can lead to challenges when scaling prototypes in production. Traditional data management also ignores AI-specific considerations such as data bias, labeling and drift; this is changing but slowly.
- Responsible AI requires new governance approaches for both the data and AI model. These AI-specific data practices are not yet part of traditional data governance in most enterprises.
- Assuming AI models have addressed all potential data management issues once deployed is a fallacy. Deployment considerations such as ongoing drift monitoring require ongoing data management activities and practices.
- AI developers are focused on the use case context as opposed to independent validation and reuse, affecting effective production use and reusability across use cases.

User Recommendations

- Formalize AI-ready data as a dedicated practice as part of your overall data management strategy. Implement active metadata management, data quality, observability, integration and fabric as foundational components of this strategy.
- Train data engineers in support of AI and train AI specialists in data management.
- Support AI model development in a data-centric way due to the dependency of AI models on representative data. Diversify data, models and people to ensure AI value and avoid involuntary bias.
- Utilize data management expertise, AI engineering, DataOps and MLOps approaches to support the AI life cycle. Include data management requirements when deploying AI models.
- Develop data monitoring and data governance metrics to ensure that your AI models produce the correct output continuously.
- Define and measure minimum data standards for AI readiness of data early for each use case and continuously prove data fitness when taking AI to scale. These include checking lineage, quality and governance assessment, versioning and automated testing.
- Investigate data management tools rich in augmented data management capabilities that can integrate well with AI tools that have created disruptive data-centric AI capabilities, like multimodal data fabric.

Gartner Recommended Reading

A Journey Guide to Deliver AI Success Through AI-Ready Data

Quick Answer: What Makes Data AI-Ready?

Innovation Insight: How Generative AI Is Transforming Data Management Solutions

Become AI-Ready by Focusing on Foundational Data Tools and Technologies

How to Evaluate AI Data Readiness

AI Engineering

Analysis By: Soyeb Barot, Cuneyd Kaya, Leinar Ramos, Anthony Mullen

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

AI engineering is the foundational discipline for enterprise delivery of AI and generative AI (GenAI) solutions at scale. The discipline unifies DataOps, ModelOps and DevOps pipelines to create a coherent development, deployment (hybrid, multicloud, edge), and operationalization framework for AI-based systems.

Why This Is Important

The demand for AI solutions has dramatically increased, driven by the unrelenting adoption of GenAI techniques. Few organizations have built the data, AI model management and DevOps foundations required to move pilot projects to production, much less operate portfolios of AI solutions at scale. In addition, an appropriate operating model is required that enables cross-collaboration across teams with diverse skills to build AI systems. Hence, to meet the demands for scaling AI solutions, enterprises must establish consistent pipelines supporting the development, deployment, reuse, governance and maintenance of AI models (statistical, machine learning, generative, deep learning, graph, linguistic and rule-based).

Business Impact

AI engineering enables organizations to establish and grow a high-value portfolio of AI solutions consistently and securely. Almost every AI solution is built using an ensemble of models, using multiple AI techniques, thereby resulting in composite AI systems. There are significant engineering, process and culture challenges to address as part of building and deploying composite AI systems at scale. With defined AI engineering approaches — DataOps, ModelOps and DevOps — it is possible to deploy models into production in a structured, repeatable factory model. AI engineering also enables establishing process flow to ensure the AI engineering team works closely with business stakeholders throughout the development process of AI-based solutions.

Drivers

- The elimination of traditional siloed approaches to data management and AI engineering doubles the data engineering effort and reduces impedance mismatches across data ingestion, processing, model engineering and deployment, which inevitably drift further once the AI models are in production.
- DataOps, ModelOps (including LLMOps) and DevOps provide best practices for moving artifacts through the AI development life cycle. Standardization across data and model pipelines accelerates the delivery of AI solutions irrespective of the approach — retrieval-augmented generation (RAG) or fine-tuning techniques alongside implementations with models built using diverse AI techniques.
- AI engineering enables discoverable, composable and reusable data and AI artifacts (such as data catalogs, knowledge graphs, code repositories, reference architectures, feature stores and model stores) across the enterprise technical architecture. These are essential for scaling AI enterprisewide.
- AI engineering practices, processes and tools must be adapted to address GenAI. GenAI-specific adaptations include support for prompt engineering, vector DBs/knowledge graphs, architecting and deploying multiagent systems, and interactive deployment models.
- AI engineering tools can be subdivided into model-centric and data-centric tools. Terms such as DataOps, LLMOps, or broader terms such as ModelOps or MLOps, are used frequently. However, we believe they are all a subset of AI engineering that are DevOps best practices implemented to operationalize specific portions of the AI development life cycle.
- Finally, AI engineering makes it possible to orchestrate solutions across hybrid, multicloud, edge AI or Internet of Things (IoT).

Obstacles

- The rapid expansion of the Ops family has led to an influx of newer, yet marginally nuanced, Ops terms that are overwhelming AI leaders. Because of this ambiguity, AI leaders often fail to prioritize and pursue the right Ops disciplines, leading to an inferior AI operational state.
- AI engineering requires simultaneous development of pipelines across domains alongside maturity across the platform infrastructure.
- It requires integrating full-featured solutions with specific tools, enabling operationalization capabilities, to address enterprise architecture gaps with minimal functional overlap. These include gaps around extraction, transformation and loading stores, feature stores, model stores, model monitoring, pipeline observability, and governance.
- AI engineering requires cloud maturity and possible rearchitecting, or the ability to integrate data and AI model pipelines across deployment contexts. Potential complexity and management of analytical and AI workloads alongside costs may deter organizations that are in the initial phases of AI initiatives.
- Enterprises often seek “unicorn” experts to productize AI platforms and solutions. Spot-fix vendor solutions will bloat costs and potentially complicate already intricate integration and model management tasks.

User Recommendations

- Maximize business value from ongoing AI initiatives by establishing an AI engineering practice that streamlines data, models and implementation pipelines.
- Simplify data and analytics pipelines by identifying the capabilities required to operationalize end-to-end AI development platforms and build AI-specific toolchains.
- Use point solutions sparingly and only to plug feature/capability gaps in fully featured DataOps, MLOps, ModelOps and tools. Open-source technologies such as Airflow and Kafka, play an important role knitting together a comprehensive solution.
- Leverage cloud service provider environments as foundational to build AI engineering. At the same time, rationalize your data, analytics and AI portfolios as you migrate to the cloud.
- Adopt a platform approach with GenAI by investing in centralized AI engineering tools for automation, governance and use-case enablement across a broad set of AI models and cloud service providers.
- Adopt software delivery best practices by establishing a common understanding of the various aspects of your solution, avoiding silos and engaging with business stakeholders.
- Upskill data engineering and platform engineering teams to adopt tools and processes that drive continuous integration/continuous development for AI artifacts.

Sample Vendors

Amazon Web Services; Anyscale; CoreWeave (Weights & Biases); Dataiku; DataRobot; Domino Data Lab; Google; Microsoft; NVIDIA (OctoAI); Unstructured

Gartner Recommended Reading

Top Strategic Technology Trends for 2022: AI Engineering

Demystify the Ops Landscape to Scale AI Initiatives: A Gartner Trend Insight Report

Case Study: AI Model Operations at Scale (Fidelity)

AI Engineering Best Practices for Data Scientists

Responsible AI

Analysis By: Svetlana Sicular, Philip Walsh

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

Responsible artificial intelligence (RAI) is an umbrella term for aspects of making appropriate business and ethical choices when adopting AI. These include business and societal value, risk, trust, transparency, fairness, bias mitigation, explainability, sustainability, accountability, safety, privacy and regulatory compliance. RAI encompasses organizational responsibilities and practices that ensure positive, accountable and ethical AI development and operation.

Why This Is Important

Wide adoption of AI resulted in separation of RAI mostly into individual focus areas, AI governance, and trust, risk and security management (TRiSM). Furthermore, jurisdictions and industry regulations concretize practices that were less defined under the RAI umbrella. While the term responsible AI is still in use, enterprises will continue emphasizing specific areas and focus on their nuanced AI goals, such as risk, privacy, compliance, ethics, AI applications evaluation, and ensuring AI-ready data.

Business Impact

Responsible attitudes toward AI are required from every role in the organization. RAI assumes accountability for AI development and use at the individual, organizational and societal levels. If AI governance and TRiSM are practiced by designated groups, RAI extends its reach to all stakeholders involved in the AI process. Concrete RAI practices, such as preserving privacy and debiasing AI, protect organizations by ensuring AI technologies are beneficial, safe, ethical and trustworthy.

Drivers

RAI helps AI participants develop, implement, utilize and address the various drivers they face. With widening AI adoption, the RAI drivers are becoming more important and are better understood by vendors, buyers, society and legislators:

- The adoption of generative AI (GenAI) raises new concerns, such as hallucinations, leaked sensitive data, copyright issues and reputational risks, that bring new actors in RAI (for example, in security, legal and procurement).

- Leading vendors are offering indemnification of their GenAI offerings, making customers more confident as part of their RAI approaches: although a good step, these are still incomplete.
- The organizational driver of RAI assumes the need to strike a balance between the business value and associated risks within regulatory, business and ethical boundaries. This includes considerations such as reskilling employees to adapt to AI technologies and safeguarding intellectual property.
- The societal driver includes resolving AI safety for societal well-being versus limiting human freedoms. Existing and pending legal guidelines and regulations, such as the EU's Artificial Intelligence Act, make RAI a necessity.
- The customer/citizen driver is based on fairness and ethics and requires reconciling privacy with convenience. Customers/citizens may be willing to share their data in exchange for certain benefits.
- AI affects all ways of life and touches all societal strata; hence, the RAI challenges are multifaceted and cannot be easily generalized, therefore, organizations address concrete items under the RAI umbrella. New problems will continue to arise with rapidly evolving technologies and their uses.

Obstacles

- RAI may look good on paper, but poorly defined accountability for RAI renders it ineffective in reality.
- Organizations lack awareness of AI's unintended consequences. Many turn to RAI practices only after they experience AI's negative effects, whereas prevention is simpler.
- Most AI regulations are still in draft. AI products' adoption of regulations for privacy and intellectual property makes it challenging for organizations to ensure compliance and avoid all possible liability risks.
- Rapidly evolving AI technologies, including tools for explainability, reasoning, bias and hallucinations detection, privacy protection and some regulatory compliance, lull organizations into a false sense of responsibility, while mere technology is not enough. A disciplined AI risk, ethics and governance approach is necessary, in addition to technology.
- Creating RAI principles and operationalizing them without regularly measuring the progress makes it hard to sustain RAI practices.

User Recommendations

- Identify and prioritize RAI focus areas for your AI strategy. Publicize consistent approaches across all RAI focus areas. Typical areas of RAI in the enterprise are fairness, bias mitigation, ethics, risk management, security, privacy, reliability, sustainability and regulatory compliance.
- Designate a champion for each use case who will be accountable for the responsible development and use of AI.
- Define the AI life cycle framework. Address RAI in all phases of this cycle. Address hard trade-off questions.
- Provide training for applicable RAI focus areas to personnel. Include AI literacy and critical thinking as part of the training.
- Operationalize RAI principles. Ensure diversity of participants and enable them to easily voice AI concerns.
- Participate in industry or societal AI groups. Learn best practices and contribute your own because everybody will benefit from this exchange. Ensure policies account for the needs of any internal or external stakeholders.

Gartner Recommended Reading

Employee Activism Drives Adoption and Norms for Responsible AI: Trends in Action

Create an AI Literacy Roadmap to Drive Responsible and Productive AI

Critical Insights: Impact of U.S. Federal Policy Changes on Responsible AI

ModelOps

Analysis By: Cuneyd Kaya, Soyeb Barot

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Model operationalization (ModelOps) is primarily focused on the end-to-end governance and life cycle management of advanced analytics, AI and decision models, such as models based on machine learning (ML), generative AI (GenAI) and knowledge graphs.

Why This Is Important

ModelOps helps companies in standardizing, scaling and augmenting their analytics, AI and GenAI initiatives. It helps organizations to move their models from the lab environments into production. MLOps primarily focuses on monitoring, operationalization and governance of ML models, while ModelOps assists with the same for all advanced analytics, decision and AI models, including GenAI and retrieval-augmented generation (RAG) systems.

Business Impact

ModelOps as a practice:

- Provides the capability for the management and operationalization of diverse AI (including generative AI), analytics and decision intelligence systems.
- Enables the complex subsystems required for such systems with observability, versioning, monitoring, data and model orchestration, guardrails, experimentation and explainability.
- Ensures collaboration among a wider business, development and deployment community and the ability to associate model outcomes with business KPIs.
- Is an important tech capability part of AI engineering; ModelOps' best practices make it possible to deploy models into production in a structured, repeatable factory model.

Drivers

- Modern AI systems are being built with a symbiotic combination of generative and classic AI models, agents and intelligent software capabilities. As the number of advanced analytics, AI and decision models at organizations increases, organizations will have to manage different types of prepackaged or custom-made models in production. In doing so, it becomes crucial for organizations to ensure these models are working as intended, maintaining accuracy and reliability across various applications.
- Organizations want to be more agile and responsive to changes within their advanced analytics and AI pipelines not only with models, but also with data, application and infrastructure.
- ModelOps provides a framework to separate responsibilities across various teams for how models (including GenAI, foundational models, analytics, ML, physical, simulation and symbolic) are built, tested, deployed and monitored across different environments (for example, development, test and production). This enables better productivity and collaboration and it lowers failure rates.
- ModelOps provides tools to address model degradation via drift and bias. In other scenarios, enabling model runtime enforcements, explainability and integrity is paramount.
- Organizations don't want to deploy an unlimited number of open-source offerings to manage ModelOps, but there are few comprehensive solutions that provide end-to-end capabilities in every domain of model operationalization. Moreover, not every capability is required immediately. Often, versioning, monitoring and model orchestration precede the full implementation of feature stores, pipelines and observability.
- GenAI will require an increased focus on testing and the introduction of capabilities to version, manage and automate prompts, routers and retrieval-augmented generation systems. Fine-tuning will also require enhanced ModelOps capabilities to manage transformations for complex domain and function training datasets.

Obstacles

- Organizations using different types of models often don't build the right Ops, model usage policy and management capabilities until they already have a chaotic landscape of unmanaged advanced analytic and AI systems.
- Not all analytical techniques currently benefit from mature operationalization methods. Because the spotlight has been on ML techniques, MLOps benefits from a more evolved AI practice, but some models, like agentic modeling and optimization techniques, require more attention in ModelOps practices and platforms.
- Organizations are struggling to get GenAI into production, due to their data not being consumable by AI models, along with data privacy, security and regulatory concerns.
- Organizations may adopt ModelOps platform capabilities that they don't immediately need. At the same time, organizations that are siloed and fail to adopt a comprehensive ModelOps strategy create redundancy in effort with respect to operationalization.

User Recommendations

- Buy ModelOps capabilities integrated into your primary AI platforms. Enrich these capabilities with best-of-breed open-source or proprietary ModelOps offerings where unique problems, like feature stores or observability, require enhanced solutions.
- Utilize ModelOps best practices across composite AI, data, models and applications to ensure transition, reduce friction and increase value generation.
- Recruit/upskill additional engineers who can master ModelOps on AI systems that utilize unstructured data, search and retrieval, graph and optimization.
- Encourage collaboration between data science and development and deployment teams; empower teams to make decisions to automate, scale and bring stability to the analytics and AI pipeline.
- Collaborate with software engineering teams to scale ModelOps. Offloading operationalization responsibilities to production support teams enables increased ModelOps specialization and sophistication across the ecosystem of complex AI-enabled applications.

Sample Vendors

Amazon Web Services; CoreWeave (Weights & Biases); Databricks; Dataiku; DataRobot; Domino Data Labs; Google Cloud; IBM; Microsoft Azure; ModelOp

Gartner Recommended Reading

Demystify the Ops Landscape to Scale AI Initiatives: A Gartner Trend Insight Report

Launch an Effective Machine Learning Monitoring System

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

Reference Architecture Brief: MLOps Architecture

Critical Capabilities for Data Science and Machine Learning Platforms, AI Engineering

Sliding into the Trough

Foundation Models

Analysis By: Arun Chandrasekaran

Benefit Rating: Transformational

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

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 are increasingly becoming multimodal. 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 natural language processing and have also been applied to computer vision, audio and video processing, and software engineering use cases.

Business Impact

With their potential to enhance applications across a broad range of enterprise use cases, foundation models are having 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 customer experience, and enabling rapid, cost-effective creation of new products and services.

Drivers

- **Quicker time to value:** Foundation models can effectively deliver value through prebuilt APIs, prompt engineering, retrieval-augmented generation or further fine-tuning. While fine-tuning may enable more customization, the other three options are less complex, quicker and cheaper.
- **Superior performance across multiple domains:** The difference between foundation 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 and are increasingly becoming better at reasoning tasks.
- **Fast-paced innovation:** 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 and driving quicker enterprise adoption.
- **Productivity gains:** Foundation models are having an impact across broad swaths of enterprise business functions as their ability to automate tasks gets wider. Business functions such as marketing, customer service and IT (especially software engineering) are areas where clients are seeking initial gains.

Obstacles

- **Flawed 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 and users, particularly for prompt engineering and fine-tuning.
- **Data integration:** Adapting foundation models to specific enterprise use cases often requires building RAG pipelines, fine-tuning or significant prompt engineering automation. While the know-how of how to implement this is growing, highly effective techniques to ground these models remain both technically complex and expensive.

User Recommendations

- **Plan to introduce foundation models into existing speech, text or coding domains.** If you have any older language processing systems, moving to a transformer-based model could significantly improve performance. Knowledge search, summarization and content generation are popular emerging use cases across industries.

- **Start with models that have superior ecosystem support** and adequate enterprise guardrails around security and privacy and are more widely deployed.
- **Be objective** about the adequate balance between accuracy, costs, security and privacy, and time to value when selecting foundation models to determine the appropriate model needed. Beware of building models from scratch, given the complexity and steep costs.
- **Educate** developers and data and analytics teams on prompt engineering and other advanced techniques needed to steer these models.
- **Designate an incubation team** to monitor industry developments, communicate the "art of the possible," experiment with business units and share valuable lessons learned companywide.

Sample Vendors

Alibaba Group; Anthropic; Cohere; DeepSeek; Google; IBM; Meta; Microsoft; Mistral AI; OpenAI

Gartner Recommended Reading

Innovation Guide for Generative AI Models

Answering IT Leaders' Top 10 Questions on Open Generative AI Models

Explore Small Language Models for Specific AI Scenarios

Synthetic Data

Analysis By: Arun Chandrasekaran, Alys Woodward, Anthony Mullen

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

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 (ML) development, data sharing, and data monetization.

Why This Is Important

Obtaining and labeling real-world data for AI development is a time-consuming and expensive task. Specifically for use cases such as training models for autonomous vehicles, collecting real data for 100% coverage of edge cases is either practically impossible or incredibly hard. These challenges can be remedied with synthetic data. It can be generated quickly, cost-effectively, and without personally identifiable information (PII) or protected health information (PHI), making it a valuable technology for privacy preservation. The rise of frontier AI models has highlighted synthetic data as a cost-effective means to build scalable models.

Business Impact

Adoption is increasing across various industries, and Gartner predicts a further rise as synthetic data:

- Avoids using PII when training AI models via synthetic variations of original data or synthetic replacement of parts of data.
- Reduces cost and saves time in ML development.
- Improves AI performance as more “fit for purpose” training data leads to better outcomes.
- Enables organizations to pursue new use cases for which minimal real data is available.
- Addresses fairness issues, such as bias and toxicity, more efficiently.
- Enables software testing on realistic yet private data, without legislation risks or in the absence of such data.

Drivers

- In regulated industries such as healthcare and finance, buyer interest is growing, as synthetic tabular data can be used to preserve privacy in AI training data.
- To meet the increasing demand for synthetic data for natural language automation training, especially for chatbots and speech applications, vendors are bringing new offerings to market, often to train domain generative AI (GenAI) models. 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.
- Transformer and diffusion architectures, the architectural foundations for GenAI, are enabling synthetic data generation at quality and precision levels not seen before. AI simulation techniques are improving synthetic data quality by better recreating real-world representations.
- There is scope for 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 and multimodal AI. 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.
- As data providers for training frontier AI models raise their data access costs, synthetic data is gaining traction as an economic alternative.

Obstacles

- Synthetic data can have implicit bias problems, miss natural anomalies, be complicated to develop, or not contribute any new information to existing, real-world data.
- While synthetic data reduces privacy risks, some industries (e.g., healthcare, finance) still face regulatory uncertainty about whether it can fully replace or augment real data for compliance purposes.
- Synthetic data generation methodologies lack standardization.
- Validating the accuracy of synthetic data is difficult. It can be challenging to know for sure whether a synthetic dataset accurately captures the underlying real-world environment.
- Buyers are still confused about when and how to use the technology due to the lack of skills.
- Enterprises have legacy data pipelines, and integrating synthetic data into existing data lakes, analytics systems and AI workflows often requires additional effort, tooling and infrastructure.
- 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.
- Establish clear policies on synthetic data life cycle management, storage and access control.
- Work with legal and compliance teams to ensure synthetic data aligns with industry regulations, such as General Data Protection Regulation, Health Insurance Portability and Accountability Act and the California Consumer Privacy Act.
- Educate internal stakeholders through training programs on the benefits and limitations of synthetic data. Institute guardrails to mitigate challenges such as user skepticism and inadequate data validation.
- Measure and communicate the business value and the success and failure stories of synthetic data initiatives.

Sample Vendors

Anyverse; Betterdata; Bifrost; Gretel; MOSTLY AI; Parallel Domain; Rendered.ai; SAS; Tonic.ai; YData

Gartner Recommended Reading

Innovation Guide for Generative AI Models

Executive Briefing on Emerging Technology: Synthetic Data

Market Guide for Data Masking and Synthetic Data

Edge AI

Analysis By: Eric Goodness

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

Edge AI is the use of AI techniques embedded in non-IT products (consumer/commercial, industrial), Internet of Things (IoT) endpoints, gateways and edge servers. Capabilities span consumer, commercial and industrial uses, such as mobile devices, autonomous vehicles, enhanced medical diagnostics and streaming analytics. While predominantly focused on AI inference, more sophisticated systems include local training capabilities to optimize models at the edge.

Why This Is Important

Many edge computing use cases are latency-sensitive and data-intensive, and require a level of autonomy and data sovereignty, for local decision intelligence. Such needs have resulted in AI deployment in a wide range of edge computing solutions. Edge AI allows industries, in hazardous and/or highly regulated environments, to apply various AI and machine learning (ML) techniques in operational environments. These applications include distributed, resource-constrained assets, an ability to benefit from improved decision support providing a close-feedback loop, with automated machine tasks, that enables asset reliability.

Business Impact

- Real-time data analysis and decision intelligence.
- Improved operational efficiency, such as visual inspection systems for quality management, output and process efficiency.
- Enhanced customer experience (CX) from AI feedback embedded within products.
- Reduced connectivity costs with fewer data journeys between the edge and cloud.
- Persistent functionality, independent of connectivity.
- Reduced storage demand as only prioritized data is passed to core systems.
- Preserved data privacy at the endpoint.

Drivers

Overall, edge AI has benefited from improvements in the capabilities of AI, including:

- The maturation of MLOps and ModelOps tools and processes that support ease of use across a broader set of features spanning the wider MLOps functions. Initially, many companies came to market with a narrowcast focus on model compression.
- The improved performance of combined ML techniques and an associated increase in data availability (such as time-series data from industrial assets).

Business demand for new and improved outcomes, solely achievable from the use of AI at the edge, include:

- Reducing full-time equivalents with vision-based solutions used for surveillance or inspections.
- Improving manufacturing production quality by automating various processes.
- Optimizing operational processes across industries.
- New approaches to CX, such as personalization on mobile devices or changes in retail from edge-based smart check-out points of sale.
- Privacy-preserving edges.

Additional drivers include:

- An increasing number of users are upgrading legacy systems and infrastructure in “brownfield” environments. By using MLOps platforms, AI software can be hosted within an edge computer or a gateway (aggregation point) or embedded within a product with the requisite compute resources.
- More manufacturers are embedding AI in the endpoint as an element of product servitization. In this architecture, IoT endpoints, such as in automobiles, home appliances and commercial building infrastructure, are capable of running AI models to interpret data captured by the endpoint and drive some of the endpoints’ functions.
- Rising demand for R&D in training has decentralized AI models at the edge for adaptive AI. These emerging solutions are driven by explicit needs such as privacy preservation or the requirement for machines and processes to run in disconnected (from the cloud) scenarios.

Obstacles

- Edge AI is constrained by the limitations of the equipment deployed, such as form factor, power budget, data volume, decision latency and security.
- Systems deploying AI techniques can be nondeterministic. This will impact edge AI applicability in certain use cases, especially where safety and security requirements are primary.
- The autonomy of edge AI-enabled solutions, built on some ML and deep learning techniques, often presents questions of trust, especially where the inferences are not readily explainable. As adaptive AI solutions increase, these issues will increase if identical models deployed to equivalent endpoints begin to evolve diverging behaviors.
- The lack of quality and sufficient data for training is a universal challenge.
- Deep learning in neural networks is a compute-intensive task, often requiring the use of high-performance chips with corresponding high-power budgets. This limits deployment locations where small-form factors and low-power requirements are paramount.

User Recommendations

- Determine whether the use of edge AI provides suitable cost-benefit improvements or whether traditional centralized data analytics and AI methodologies are adequate and scalable.
- Evaluate when to consider AI at the edge versus a centralized solution. Good candidates for edge AI are applications that have high communications costs, are sensitive to latency, require real-time responses or ingest high volumes of data at the edge.
- Assess the different technologies available to support edge AI and the viability of the vendors offering them. Many potential vendors are startups that may have interesting products but limited support capabilities.
- Use edge gateways and servers as the aggregation and filtering points to perform most of the edge AI and analytics functions. Make an exception for compute-intensive endpoints, where AI-based analytics can be performed on the devices themselves.

Sample Vendors

Edge Impulse; IFS (Falkonry); Johnson Controls; Pratexo; Synadia Communications

Gartner Recommended Reading

Emerging Tech Impact Radar: Edge Artificial Intelligence

Innovation Insight for Edge AI

Emerging Tech: Differentiate With an Edge AI Benchmarking Strategy

Market Guide for Edge Computing

Emerging Tech: Empower Outcome-Centric IoT With AI

Generative AI

Analysis By: Svetlana Sicular

Benefit Rating: Transformational

Market Penetration: More than 50% of target audience

Maturity: Adolescent

Definition:

Generative AI (GenAI) technologies can generate new derived versions of content, strategies, designs and methods by learning from large repositories of original source content. Generative AI has profound business impacts, including on content discovery, creation, authenticity and regulations; automation of human work; and customer and employee experiences.

Why This Is Important

GenAI is becoming real in enterprises. AI leaders from the 2024 Gartner AI Mandates for the Enterprise Survey reported an average spend of $1.9 million in fiscal year 2024 on GenAI initiatives, which reflects a belief in further GenAI potential. Governments are committing large funds to GenAI; vendors continue fast innovation, advancing model performance, multimodality, reasoning and agentic capabilities. Research of training data, explainability, fine-tuning, distillation and other aspects of GenAI exploitation is fast-paced and is reflected in commercial and open-source solutions.

Business Impact

GenAI has a strong momentum for expansion and deeper integration into business workflows across various business functions and industries. Fully integrated tools, accompanied by AI governance practices, robust education and IT support, enable enterprises to tackle critical business processes. Multimodal GenAI opens new opportunities in life sciences, transportation and education. The current focus for GenAI application is on productivity, automation and evolving job roles.

Drivers

- GenAI is proving its worth in life sciences, manufacturing, finance, law and entertainment. It is becoming more specialized and optimized for domains such as coding assistance, scientific discovery, research, diagnostics, legal analysis and financial modeling. Additionally, 78% of enterprises surveyed by Gartner have integrated or are planning to integrate the use of GenAI into some areas (see Technology Spending Drivers, Business Outcomes and Challenges for CIOs Across Industries for more information).
- Businesses aim to automate tasks, generate content and enhance customer experience by integrating GenAI into their processes. Prompt engineering is the main approach for custom GenAI use cases.
- Governments, spurred by the GenAI promise, are increasing investments in national AI strategies.
- Agentic AI is a top driver of a GenAI value proposition this year due to automation benefits and combining GenAI with other techniques.
- Fierce GenAI model competition continues. GenAI providers are introducing model quality and performance improvements, as well as more sophisticated reasoning and handling of image and video inputs. Galloping leaderboards list hundreds of large language models (LLMs), including a variety of smaller models that demonstrate precision and cost-effectiveness in specific domains and tasks, such as time series. Distillation, truncation and other methods to derive smaller models from large ones result in reduced latency and lower costs. Open-source LLMs democratize access to GenAI and stimulate ecosystem innovation.
- Technology vendors and service providers compete on GenAI applications and model offerings, and their enterprise readiness, pricing, infrastructure, safety and indemnification. Vendors and open-source communities offer better tooling for training, fine-tuning, evaluation and life cycle.
- Infrastructure innovations and investments are on the rise. Hyperscalers and some enterprises are building supercomputing systems that combine innovations in computational accelerators, high-speed networks and performance-optimized storage. Meanwhile, innovations like DeepSeek stimulate ideas efficiently with less advanced chips and lower costs.

Obstacles

- Estimating GenAI's value is challenging, with less than 30% of the AI leaders from the 2024 Gartner AI Mandates for the Enterprise Survey reporting that their CEOs praise AI investment returns. Organizations face productivity leakage, where GenAI adoption doesn't directly yield value.
- Technical challenges include security, model evaluation, data availability and quality, and managing compute for inferencing.
- Low maturity organizations have difficulty in identifying suitable use cases and face unrealistic expectations for GenAI initiatives.
- Advanced organizations struggle to find skilled professionals. New users necessitate GenAI literacy.
- Governance challenges include hallucinations, bias, fairness and establishing a governance operating model. Government regulations may impede GenAI initiatives.
- GenAI licensing and pricing are inconsistent among providers. Pricing remains confusing and constantly evolving, often catching customers by surprise.

User Recommendations

- Focus on problems that GenAI can solve effectively. Develop methods to identify impactful GenAI use cases that align with business objectives and offer tangible benefits.
- Design solutions to be loosely coupled with GenAI models to enable flexible model selection and combinations. Investigate GenAI vendor roadmaps to avoid spending your own resources on the capabilities that vendors will deliver in the future.
- Develop an AI-ready data strategy around your GenAI portfolio. Plan to incorporate your proprietary data into GenAI via retrieval-augmented generation or similar methods. Ensure data is relevant, clean, and accessible for GenAI models.
- Invest in AI literacy and talent upskilling for working with GenAI tools and technologies.
- Establish GenAI governance operating model, policies, controls and technical oversight. Consider both your and your vendors’ responsible AI practices.
- Plan for the cost of running GenAI initiatives, including infrastructure, compute resources and ongoing maintenance.

Sample Vendors

Alibaba Cloud; Amazon Web Services; Anthropic; DeepSeek; Google; Hugging Face; IBM; Meta; Microsoft; OpenAI

Gartner Recommended Reading

Generative AI: The Basics

Solution Path for Implementing Generative AI Systems

AI Technology Sandwich: A Conceptual Framework for Executing AI

10 Best Practices for Scaling Generative AI Across the Enterprise

Climbing the Slope

Cloud AI Services

Analysis By: Jim Scheibmeir, Bern Elliot

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Cloud AI services provide AI model-building tools, APIs for prebuilt services and associated middleware that enable the building/training, deployment and consumption of machine learning (ML) and generative AI models running on prebuilt infrastructure as cloud services. These services include pretrained vision, language and other generative AI services, and automated ML and fine-tuning to create new models and customize prebuilt models.

Why This Is Important

The use of cloud AI services continues to increase due to the interest and hype around generative AI technologies. Vendors have established large language model (LLM) APIs and solutions with fully integrated MLOps pipelines. The addition of low-code tools aids ease of use. Applications regularly use AI cloud services in language, vision and tabular data to automate business processes. Developers are increasingly using both prebuilt and customized ML models in applications.

Business Impact

Cloud AI services impact business-running applications, allowing developers to enhance application functions with intelligent automation. Generative AI is the newest category to these solutions. Data-driven decisioning mandates the inclusion of ML models to add application functionality. Some AI technologies are maturing, but generative AI includes less mature capabilities. Cloud AI services enhance applications with models that score, forecast and generate content, enabling data-driven business operations.

Drivers

- **Demand for conversational interactions:** The emergence of generative AI and LLMs facilitates conversationally enabled applications where users can use LLMs with data sources to gain insights.

- **Opportunities to capitalize on data investments:** The wealth of data from both internal and third-party sources is highly useful for building predictive ML models that enable data-driven decision intelligence in applications.
- **Need to meet business key performance indicators (KPIs):** There is a need for businesses to automate processes to improve accuracy, improve responsiveness and reduce costs by deploying both AI and ML models.
- **Reduced barriers to entry:** The ability to use pretrained generative AI models and fine-tune them has reduced the need for large quantities of data to train models. Access for developers and citizen data scientists to AI and ML services, due to the availability of API-callable LLMs, will further expand the use of AI by development teams.
- **Automated ML as an enabler for custom development:** The use of automated ML to customize packaged services to address specific business needs is much more accessible and doesn't require data scientists.
- **Wide range of cloud AI services:** A range of specialized providers in the market offer cloud AI services, including orchestration layers to streamline deployment of solutions.
- **Emerging AI model marketplaces:** New marketplaces should help developers adopt predictive and foundation models.

Obstacles

- Developers and citizen data scientists lack understanding about how to adapt cloud AI services to specific use cases.
- Grounding generative AI models is a challenge, requiring well-crafted retrieval-augmented generation (RAG) solutions that often include vector embeddings and other capabilities to implement. Many cloud AI developer services (CAIDS) providers are offering these capabilities as part of their generative AI offering.
- Pricing models for usage-based cloud AI services present a risk for businesses as the costs associated with their use can accrue rapidly. There is a need for comprehensive cost modeling tools to address this issue.
- There is an increased need for packaged solutions that utilize multiple services for developers and citizen data scientists.
- There is a lack of skills such as prompt engineering and fine-tuning for developers to effectively implement these services in a responsible manner.

User Recommendations

- Improve the chances of success of your AI strategy by experimenting with AI techniques, including the use of generative AI models such as LLMs and multimodal models, and other cloud services. Ensure that generative AI models are loosely coupled as the technology continues to evolve rapidly.
- Use cloud AI services to build less complex models, giving the benefit of more productive AI while freeing up your data science assets for higher-priority projects.
- Empower non-data-scientists with features such as automated algorithm selection, dataset preparation and feature engineering for project elements. Leverage existing expertise on operating cloud services to assist technical professional teams.
- Utilize pretrained generative AI models to allow for rapid prototyping and deployment of LLM-enabled solutions.
- Develop cost modeling tools that allow the enterprise to effectively predict both usage and management costs as AI models are broadly deployed in applications across the business.

Sample Vendors

Alibaba Cloud; AWS; Baidu; Google; H2O.ai; IBM; Microsoft; NVIDIA; Oracle; Tencent Cloud

Gartner Recommended Reading

Critical Capabilities for Cloud AI Developer Services

Magic Quadrant for Cloud AI Developer Services

Knowledge Graphs

Analysis By: Afraz Jaffri

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Knowledge graphs are machine-readable representations of the physical and digital environments. They include entities (people, companies and digital assets) and their relationships, which adhere to a graph data model — a network of nodes (vertices) and links (edges/arcs).

Why This Is Important

Knowledge graphs capture information about the world in a visually intuitive format yet are still able to represent complex relationships. Knowledge graphs act as the backbone of a number of products, including search, smart assistants and recommendation engines. Knowledge graphs support collaboration and sharing, exploration and discovery, and the extraction of insights through analysis. Generative AI models can be combined with knowledge graphs to provide context for more accurate outputs in a technique becoming known as GraphRAG.

Business Impact

Knowledge graphs can drive business impact in a variety of different settings, including:

- Digital workplace, such as collaboration, sharing and search
- Automation, such as ingestion of data from content to robotic process automation
- Machine learning (ML), such as augmenting training data
- Investigative analysis, such as law enforcement, cybersecurity and risk management
- Digital commerce, such as product information management and recommendations
- Data management, such as metadata management, data cataloging and data fabric

Drivers

- The need to complement AI and ML methods that detect only patterns in data (such as the current generation of foundation models) with the explicit knowledge, rules and semantics provided by knowledge graphs.
- The desire to make better use of unstructured data in documents, correspondence, images and videos, using standardized metadata that can be related and managed and provide the foundation for AI-ready data.
- The increased usage of knowledge graphs with large language models to provide enhanced contextual understanding when answering questions on large quantities of enterprise data.
- The increasing awareness of the use of knowledge graphs in consumer products and services, such as smart devices and voice assistants, chatbots, search engines, recommendation engines and route planning.
- The emerging landscape of Web3 applications and the need for data access across trust networks, leading to the creation of decentralized knowledge graphs to build immutable and queryable data structures.
- The need to manage the increasing number of data silos where data is often duplicated, and where meaning, usage and consumption patterns are not well-defined.
- The use of graph algorithms and ML to identify influencers, customer segments, fraudulent activity and critical bottlenecks in complex networks.

Obstacles

- Awareness of knowledge graph use cases is increasing, but business value and relevance are difficult to capture in the early implementation stages.
- Moving knowledge graph models from prototype to production requires engineering and system integration expertise. Methods to maintain knowledge graphs as they scale — to ensure reliable performance, handle duplication and preserve data quality — remain immature.
- Organizations want to enable the ingestion, validation and sharing of ontologies and data relating to entities such as geography, people and events. However, making internal data interoperable with external knowledge graphs is a challenge.
- In-house expertise, especially among subject matter experts, is lacking, and identifying third-party providers is difficult. Often, expertise resides with vendors of graph technologies. Skills in scalability and optimization are also hard to acquire.

User Recommendations

- **Create a working group of knowledge graph practitioners and sponsors** by assessing the skills of data and analytics (D&A) leaders, practitioners and business domain experts. Factors like use case requirements, data characteristics, scalability expectations, query flexibilities and domain knowledge of knowledge graphs should be addressed.
- **Run a pilot to identify use cases that need custom-made knowledge graphs.** The pilot should deliver not only tangible value for the business but also learning and development for D&A staff.
- **Create a minimum viable subset that can capture the information of a business domain to decrease time to value.** Assess the data, both structured and unstructured, needed to feed a knowledge graph, and follow agile development principles.
- **Utilize vendor and service provider expertise** to validate use cases, educate stakeholders and provide an initial knowledge graph implementation.
- **Include knowledge graphs within the scope of D&A governance and management.** To avoid perpetuating data silos, investigate and establish ways for multiple knowledge graphs to interoperate and extend toward a data fabric.

Sample Vendors

Altair; Diffbot; eccenca; Fluree; Graphwise (Ontotext); Neo4j; Stardog; TopQuadrant

Gartner Recommended Reading

How to Build Knowledge Graphs That Enable AI-Driven Enterprise Applications

3 Ways to Enhance AI With Graph Analytics and Machine Learning

Model Distillation

Analysis By: Birgi Tamersoy, Yogesh Bhatt

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Model distillation is a technique where a smaller, simpler model (the "student") is trained to replicate the behavior of a larger, more complex model (the "teacher"). This is achieved by having the student model learn from the predictions of the teacher model. The goal is to retain the performance of the teacher model while benefiting from the efficiency and reduced resource requirements of the student model.

Why This Is Important

Organizations often face a trade-off between accuracy and efficiency when deploying AI models. Larger models, or model ensembles, tend to have higher accuracy but need more computational resources and may have latency issues. Smaller models are more efficient and require fewer computational resources, but tend to be less accurate when trained from scratch. Model distillation helps by transferring the information content of large models into smaller ones, maintaining performance while improving efficiency.

Business Impact

Model distillation can provide:

- **Cost reduction and energy efficiency:** Smaller models require fewer computational resources during deployment, resulting in reduced capital investments, operational costs and energy consumption.
- **Faster inference times:** Efficient models result in quicker inference times, enabling real-time applications and improving user experience.

- **Compliance and privacy:** Smaller models can be deployed on-premises or on a device, which can help in meeting data privacy regulations.

Drivers

- **Developments in foundation models:** As foundation models grow larger and more capable, they offer the ability to generate high-quality synthetic data, but at the same time, pose cost challenges for deployment. These developments create both an enabler and a need for model distillation.
- **Growth in edge computing and mobile applications:** The rise of edge computing and AI-powered mobile applications require AI models that can operate efficiently on devices with limited computational power. Model distillation helps create smaller models that are suitable for deployment on such devices.
- **Increased focus on sustainability:** As organizations strive to reduce their carbon footprint, there is a growing emphasis on energy-efficient AI solutions. Model distillation contributes to sustainability by reducing the computational resources and energy required for model deployment and operation.
- **Advancements in transfer learning:** Developments in transfer learning have enhanced the ability to transfer knowledge from large models to smaller ones, improving the effectiveness of model distillation techniques. This allows distilled models to achieve higher accuracy while remaining efficient.
- **Regulatory pressure and data privacy:** Increasing regulations around data privacy and protection are driving the need for AI models that can be deployed on-premises or on a device, minimizing data transfer and exposure. Model distillation supports compliance by enabling the deployment of efficient models in secure environments.

Obstacles

- **Loss of model accuracy:** Student models may lose some accuracy compared to teacher models, especially when the student model has significantly fewer parameters compared to the teacher model.
- **Complexity of the distillation process:** The process may require higher levels of technical expertise, which can be a barrier for some organizations.
- **Limited generalization across domains:** Distilled models might not generalize well across various domains, limiting their applicability in diverse tasks.
- **Dependence on high-quality teacher models:** The success of distillation depends on the quality of the teacher model; any biases, inefficiencies or potential intellectual property infringement risks can be passed to the student model.
- **Licensing restrictions on teacher models:** Some providers impose restrictions on the use of teacher model outputs, prohibiting their use in the training of other models, which can limit the feasibility of distillation efforts.

User Recommendations

- **Optimize deployment costs:** Use model distillation to reduce computational resources and lower operational expenses, while maintaining model performance.
- **Prioritize high-quality teacher models:** Ensure that your teacher models are well-aligned for your target task, well-optimized and free from biases. This will improve the quality of distilled models.
- **Invest in expertise:** Invest in skilled personnel or training programs for implementing effective model distillation. This will ensure that your AI solutions are accurate and efficient.
- **Align with regulatory compliance:** Use model distillation to create efficient models that can be deployed on-premises or on a device. This approach reduces privacy and safety concerns and eases regulatory compliance.

Sample Vendors

Amazon Web Services; Google; IBM; Microsoft; OpenAI; Predibase; Snorkel AI

Gartner Recommended Reading

Technical Professionals Need to Track 5 Important LLM Developments

Appendixes

See the previous Hype Cycle: Hype Cycle for Artificial Intelligence, 2024

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 it as it enters this phase.
Years to Mainstream Adoption	The time required for the innovation to reach the Plateau of

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 (June 2025)

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 constraints replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source: Gartner (June 2025)

Document Revision History

Hype Cycle for Artificial Intelligence, 2024 - 17 June 2024

Hype Cycle for Artificial Intelligence, 2023 - 19 July 2023

Hype Cycle for Artificial Intelligence, 2022 - 8 July 2022

Hype Cycle for Artificial Intelligence, 2021 - 29 July 2021

Hype Cycle for Artificial Intelligence, 2020 - 27 July 2020

Hype Cycle for Artificial Intelligence, 2019 - 25 July 2019

Hype Cycle for Artificial Intelligence, 2018 - 24 July 2018

Hype Cycle for Artificial Intelligence, 2017 - 24 July 2017

Hype Cycle for Smart Machines, 2016 - 21 July 2016

Hype Cycle for Smart Machines, 2015 - 24 July 2015

Hype Cycle for Smart Machines, 2014 - 18 July 2014

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 2024 Hype Cycle Builder

10 Best Practices for Scaling Generative AI Across the Enterprise

When Not to Use Generative AI

Explore Small Language Models for Specific AI Scenarios

How to Interpret and Act on 2025 AI Index Report Insights

Frontier AI Models Like DeepSeek Will Transform I&O Strategies

Table 1: Priority Matrix for Artificial Intelligence, 2025

Benefit ↓	*Years to Mainstream Adoption*			
	Less Than 2 Years ↓	*2 to 5 Years* ↓	*5 to 10 Years* ↓	*More Than 10 Years* ↓
Transformational	Composite AI Responsible AI	AI Engineering Decision Intelligence Embodied AI First-Principles AI Foundation Models Generative AI ModelOps Multimodal AI	AI-Native Software Engineering AI-Ready Data World Models	Artificial General Intelligence
High	Edge AI	AI Agents AI Governance Platforms AI TRiSM Causal AI Cloud AI Services Knowledge Graphs Model Distillation Neurosymbolic AI Sovereign AI Synthetic Data	AI Simulation FinOps for AI	
Moderate				

Benefit	*Years to Mainstream Adoption*			
↓	*Less Than 2 Years* ↓	*2 to 5 Years* ↓	*5 to 10 Years* ↓	*More Than 10 Years* ↓
Low				Quantum AI

Source: Gartner (June 2025)

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 it as it enters this phase.
Years to Mainstream Adoption	The time required for the innovation to reach the Plateau of Productivity.

Source: Gartner (June 2025)

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 (June 2025)

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 constraints replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source: Gartner (June 2025)