JPMorganChase

2025 Emerging Technology Trends

Global Technology Strategy, Innovation and Partnerships

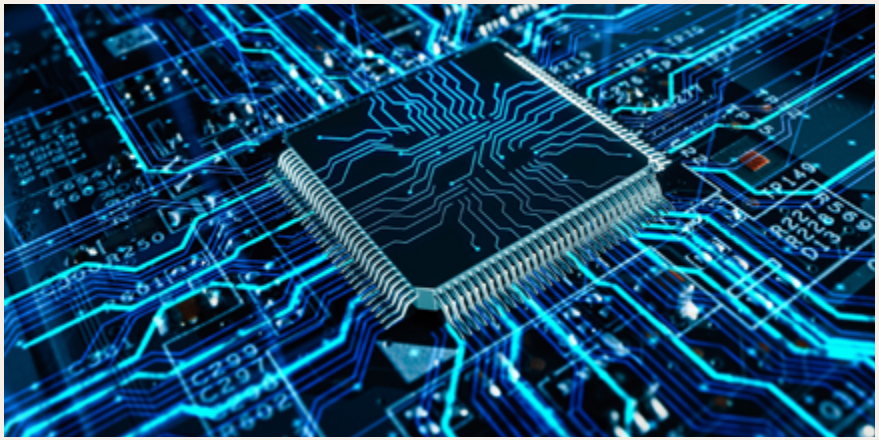
Executive Summary

Through collaborative efforts with technology leaders across the firm and continuous connectivity with the external technology ecosystem, the Global Technology Strategy, Innovation, and Partnerships team helps JPMorganChase remain connected to innovative and emerging technology trends. Each year, the team identifies a collection of the most meaningful emerging technology trends to shape and influence our Global Technology strategy and future roadmap. This document outlines the top trends for 2025, with an overview of each trend and insights from the external market.

In 2024, the excitement and innovation around Generative AI (GenAI) continued, offering transformative potential across industries. We saw AI capabilities start to become embedded in horizontal applications for content creation, workflow automation, and data analysis, and in vertical applications for users across software engineering, sales, marketing, finance, fraud, and risk. GenAI also gained traction for tasks like data management or in streamlining business processes with agents and orchestration. All of this is powered by foundational models while requiring specialist computing infrastructure to train these models. In addition, several trends emerged across the model providers including proprietary models versus open models, domain-specific models (e.g., coding, image and video creation), and the use of smaller models to provide broader deployment options with optimized performance, including for mobile and edge devices. In the security space, capabilities emerged to ensure the secure use of AI technology and to protect against new threat actors using GenAI and agentic AI.

Our 2025 report maintains a clear focus on the innovation around GenAI and AI agents across technology domains. Looking toward the future of technology modernization, we cover the emergence of key trends like agentic software development, AI clouds and data center design. Considering the breadth of innovation that is impacting employee and customer experiences, we examine how GenAI is automating workflows, introducing new interaction modalities and transforming areas such as content creation and marketing. Moving into data and AI infrastructure, we highlight trends such as the evolution of Retrieval Augmented Generation (RAG) for integrating proprietary data into model responses and multi-agent systems. Sustainability remains a key focus of innovation across the tech stack including solutions to reduce the power demands of AI models, which leads us to examine strategies for the next generation of data center design and use of smaller models. Finally, as adoption of AI accelerates and threat vectors evolve, we examine key technology trends for protecting agentic systems, confidential AI and agentic security capabilities.

Table of Contents



01 Innovative Products & Experiences

02 Delivering Data & AI/ML at Scale

03 Technology Modernization

04 Protect the Firm

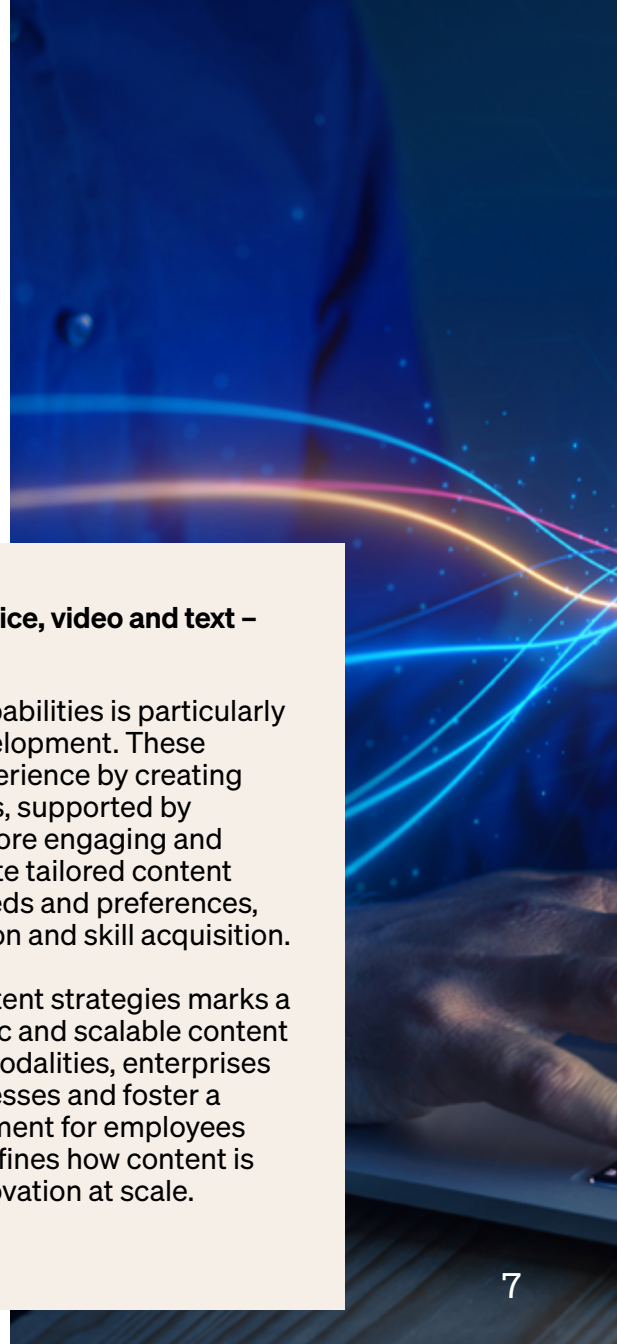
01 Innovative Products & Experiences

The convergence of advanced innovation is reshaping the way businesses engage with both customers and employees. Our clients and customers demand frictionless interactions, while our approximately 320K employees seek experiences that mirror the intuitive and efficient technologies they enjoy in their personal lives.

GenAI is at the forefront of transforming user experiences, automating workflows, and introducing new interaction modalities. Central to this shift is the Agent Operating System (OS), which will autonomously execute tasks, redefining efficiency and seamlessness.

This innovation extends to hardware, where on-device AI will provide more responsive applications and allow users to harness GenAI's power both online and offline. These advancements are setting a new standard for dynamic and frictionless experiences.

AI Generated Videos and Avatars



Multimodal models are opening new ways to autogenerate and scale content across modalities – voice, video and text – to create and scale rich content experiences across the enterprise

Multimodal models signify a major advancement in AI technology, offering capabilities far beyond those of traditional models, which are limited to processing a single type of data. Multimodal models can analyze and interpret information from diverse sources, allowing users to interact with them in ways that reflect their everyday experiences—using images, text, voice, and other media. This innovation enables employees to create and consume dynamic, engaging content on a large scale, transforming the enterprise content supply chain.

This is especially advantageous for long-form content, which can be repurposed into various formats to reach a broader audience, significantly saving employees time. For example, a comprehensive report can be condensed into a series of engaging videos or podcasts, making it more accessible and easier for employees to digest.

The impact of GenAI's multimodal capabilities is particularly evident in areas like learning and development. These models can enhance the learning experience by creating personalized and interactive materials, supported by AI avatar coaches, making learning more engaging and effective. This technology can generate tailored content that addresses individual learning needs and preferences, thereby improving knowledge retention and skill acquisition.

Integrating GenAI into enterprise content strategies marks a significant shift towards more dynamic and scalable content ecosystems. By leveraging multiple modalities, enterprises can streamline content creation processes and foster a more engaging and inclusive environment for employees to learn and grow. This approach redefines how content is produced and consumed, driving innovation at scale.

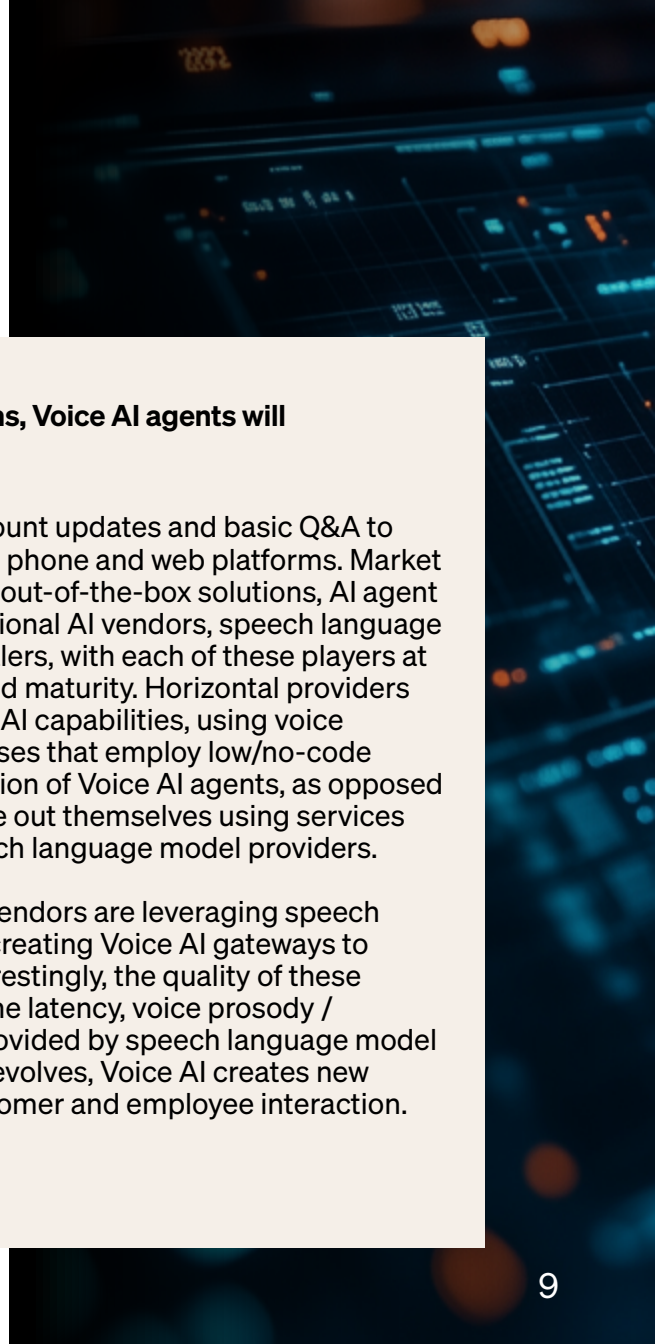
Market and Industry Perspectives

Considering the growing demand for video content, the market for multimodal models is positioned for substantial growth. Large tech players continue to invest in this space, aiming to develop and refine multimodal AI capabilities. This investment is driving rapid innovation, leading to the development of more sophisticated models that can handle complex tasks across various modalities.

In this increasingly saturated market, personalized and realistic full-sized avatars are becoming key differentiators among competitors. We are observing leading technology companies continue to enhance their models, by extending video output duration time and enhance resolution to enable high-quality video creation at scale.

Many enterprise sectors are increasingly adopting multimodal AI technologies to enhance their content creation and consumption processes. Companies are leveraging these models to improve customer engagement, streamline operations, and offer more personalized experiences.

Voice AI Agents



As the industry shifts from traditional methods of speech processing to LLM-driven conversations, Voice AI agents will uplift existing experiences like IVR to drive enhanced customer and employee servicing

Interactive Voice Response (IVR), a common form of contact center automation for years, primarily relies on predetermined scripts and complex menu options that require a keypad or simple voice input. GenAI has begun to revolutionize this space and transform customer connection.

Companies have started leveraging Voice AI to replace or augment traditional IVR capabilities. The industry is shifting away from workflows leveraging previously separate but connected services like Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), and Text-to-Speech (TTS) and moving towards speech-to-speech models that can generate speech at the appropriate pace, emotion, rhythm, and context, in response to acustomer's question.

Voice AI enhances voice capabilities and automates workflows through APIs, accessing customer data to streamline tasks. As the market gains momentum, Voice AI—or Voice AI agents—open new opportunities for customer and employee experience strategies, significantly improving call performance and operational efficiency at contact centers and help desks. Voice AI applications span from account updates and basic Q&A to call routing, deployable across phone and web platforms. Market capabilities include horizontal out-of-the-box solutions, AI agent servicing platforms, conversational AI vendors, speech language model providers and hyperscalers, with each of these players at different levels of capability and maturity. Horizontal providers leverage out-of-the-box Voice AI capabilities, using voice vendors, tailored for specific uses that employ low/no-code interfaces to simplify the creation of Voice AI agents, as opposed to organizations building these out themselves using services from the hyperscalers or speech language model providers.

Traditional conversational AI vendors are leveraging speech language model providers or creating Voice AI gateways to integrate various models. Interestingly, the quality of these experiences often hinges on the latency, voice prosody / intonation, and naturalness provided by speech language model providers. As this technology evolves, Voice AI creates new opportunities for modern customer and employee interaction.

Market and Industry Perspectives

The AI voice generator market is projected to grow from $3B in 2024 to $20.4B by 2030, driven by increased demand in the retail, healthcare, and automotive sectors. Advances in neural networks and deep learning are enhancing the authenticity of artificial voices. [1]

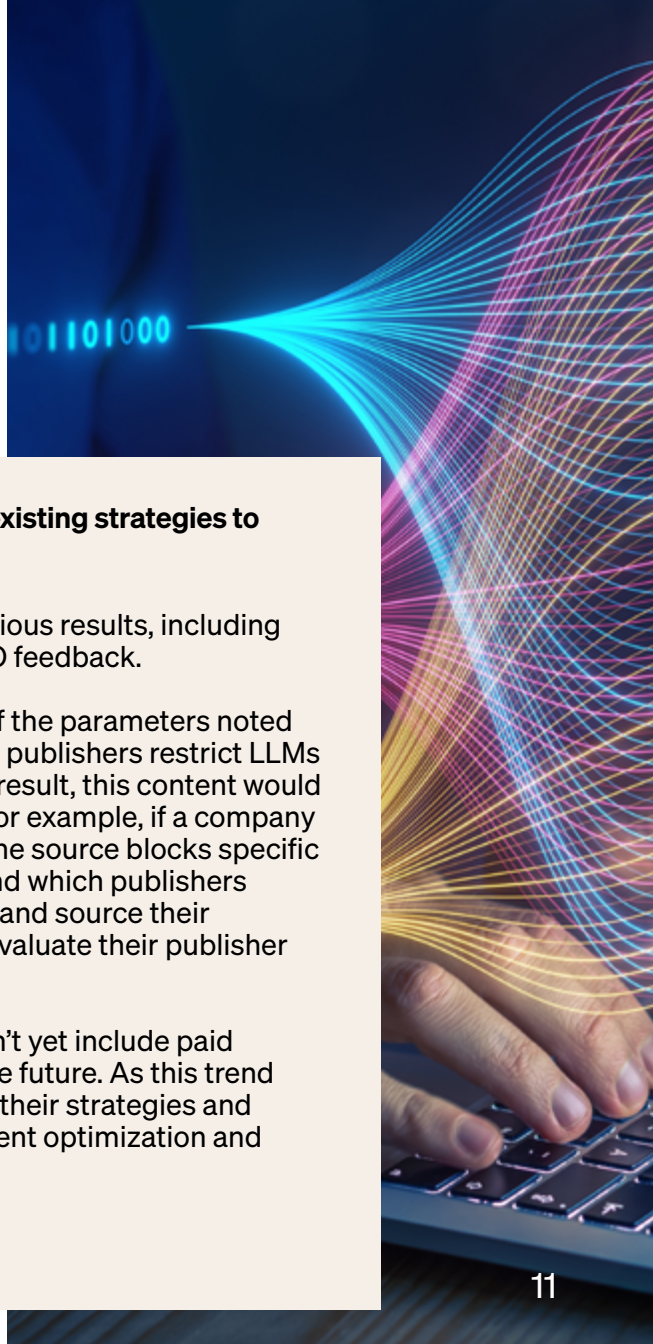
Horizontal vendors provide out-of-the-box, no-code platforms and workflow builders used to deploy humanlike voice AI agents using conversational pathway builders. This space is seeing significant growth, with many vendors launching in the past couple of years.

Conversational AI players have also started building out their offerings to include Voice AI. Although conversational AI has traditionally been bound to intent-based conversations, the use of multimodal models has created a new wave of vendors that are able to service customers across channels – chat, e-mail and voice – in natural language and execute actions on their behalf.

Adjacent services have started to open up, primarily around Voice AI agent observability, how to ensure optimal agent performance in areas like speed and latency, and to better understand the customer experience as seen with recent launches in this space.

Emotional AI is also appearing as a trend within voice with new models designed to enable human-like voice and conversations and understand users' tones of voice.

Generative Engine Optimization



With the use of GenAI search engines gaining prominence, brands will need to re-evaluate their existing strategies to optimize their content to appear in LLM-driven results

The future of search is expanding beyond traditional engines. While Search Engine Optimization (SEO) enhances website rankings in traditional search engines, Generative Engine Optimization (GEO) focuses on optimizing content for visibility in AI-powered search engines.

As AI-driven search grows, with platforms like those emerging in the industry, GEO is becoming crucial for product discovery. Companies are now analyzing how often Large Language Models (LLMs) recommend brands based on factors like brand perception, advertising effectiveness, price, and product features. Marketers must understand how LLMs rank consumer preferences and adjust their SEO strategies accordingly. For instance, a model prioritizes introductory offers or rewards in a search for the best student credit card, financial institutions should tailor their marketing content to include these criteria; however, changes will not be immediately adopted but only once a model is refreshed. It will be during these periods between these refreshes where marketers will need to review their content strategies to optimize previous results, including using SEO Agents for real-time SEO feedback.

Advertising effectiveness, as one of the parameters noted above, will play a critical role. Some publishers restrict LLMs from crawling their pages and as a result, this content would be less likely to appear in results. For example, if a company advertises with a news outlet, but the source blocks specific models, they will need to understand which publishers allow their selected model to crawl and source their content, leading a company to re-evaluate their publisher partnership strategies.

While AI-driven search engines don't yet include paid advertising, this could change in the future. As this trend develops, brands need to reassess their strategies and consider how GEO will impact content optimization and publisher partnerships.

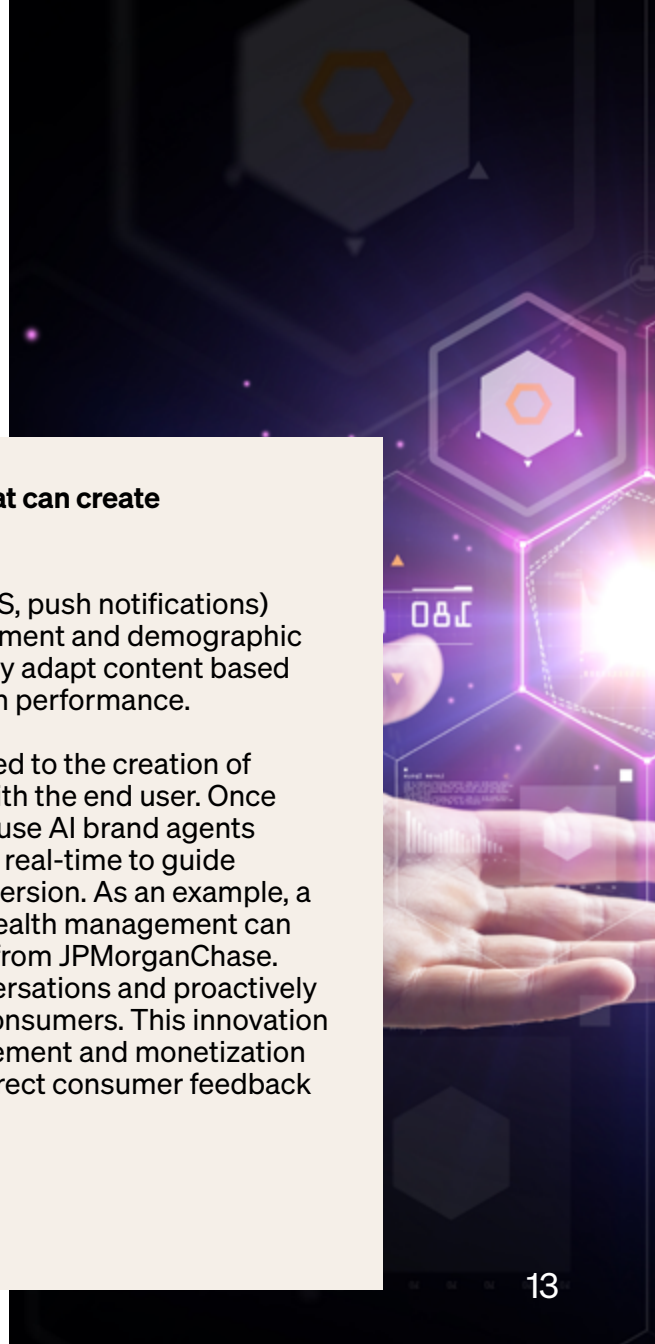
Market and Industry Perspectives

Gartner predicts that by 2026, traditional search engine volume will drop by 25% as market share shifts to AI chatbots and AI-driven search. Additionally, Gartner forecasts that by 2028, organic search traffic will decrease by 50% or more.[2]

Specific providers offer brand analytics that focus specifically on what AI models say about a particular brand or product, the likelihood that a LLM would recommend the brand to consumers, and how companies compare to their competitors - they also identify strengths and weaknesses. This information allows organizations to optimize SEO for LLMs as they transition towards a GEO strategy.

Some providers use Conversational SEO Assistants or 'SEO Agents' to interact with an AI assistant to receive real-time feedback and insights on keyword performance.

AdTech Automation



GenAI is transforming marketing by automating end-to-end campaign workflows with agents that can create personalized, multimodal content and engage consumers in conversational campaigns

When GenAI entered the market a couple of years ago, providers were initially focused on leveraging LLMs as copilot solutions to support independent tasks such as copy generation across blog or social posts.

Since then, there has been a notable shift from using LLMs purely for text-based content creation to automating and simplifying end-to-end marketing campaigns (including ads) by using AI agents to complete a task on a user's behalf or leveraging 'marketing copilots' to work simultaneously with a user to provide real-time feedback. These end-to-end platforms can produce multimodal, dynamic content across asset types, personalize communications that target unique audience segments, localize marketing messages and drive derivative content (meta-descriptions, social posts) all within a single platform.

As an example, a marketer can input a high-level marketing objective e.g., "Drive credit card sign-ups in New York" where the AI agent can automatically plan and generate all branded assets (e.g., emails, SMS, push notifications) personalized to each audience segment and demographic within a single view and dynamically adapt content based on user engagement and campaign performance.

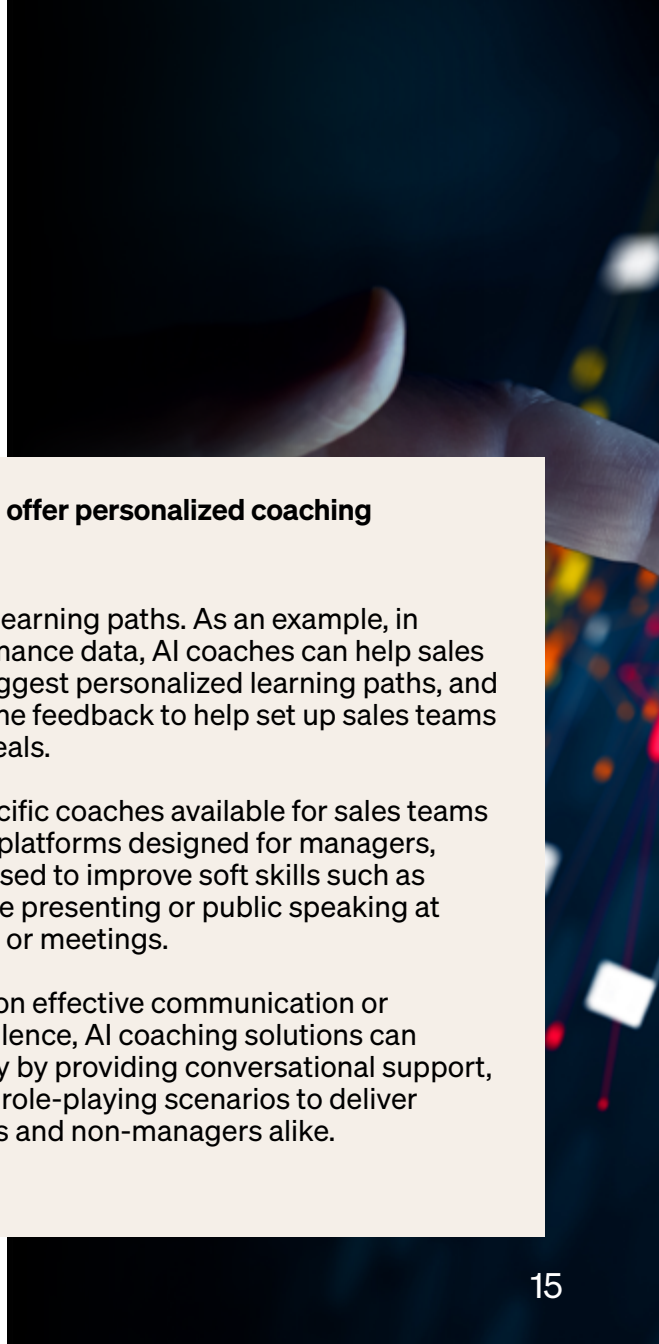
The agentic experience is not limited to the creation of the ad but includes engagement with the end user. Once a campaign is live, advertisers can use AI brand agents that adapt messaging and offers in real-time to guide consumers smoothly towards conversion. As an example, a reader who finishes an article on wealth management can engage with an AI-powered agent from JPMorganChase. These brand agents guide 1:1 conversations and proactively suggest personalized content to consumers. This innovation offers publishers increased engagement and monetization opportunities, while brands gain direct consumer feedback and valuable first-party data.

Market and Industry Perspectives

The AdTech market is expected to grow by 60% over the next five years, with total global spending reaching $43.5B by 2029.[3]

Although this is a nascent space, conversational campaigns go beyond simple chatting to include guided interactions and suggested reading, with the AI agent learning the end user's preferences along the way. This allows publishers and brands to connect directly with consumers in real-time and continuously optimize their personalized experience. While there are several vendors focused on automating aspects of a marketer's workflows such as auto-generating design briefs, social media posts or blogs, these players are limited to the creation of content vs. automating the end-to-end workflow.

AI Driven Coaching



As the demand for real-time feedback increases within enterprises, AI coaches have emerged to offer personalized coaching to improve employees' skills throughout their career journeys

AI coaches are digital and automated solutions that provide personalized guidance, support and training to employees or users within an organization, either as on-demand agents available on a needs basis or as embedded components within core enterprise systems for real-time assistance and insights.

Employee performance is expected to be optimized through automated, real-time feedback provided by AI coaches. This feedback can be embedded directly within end-user applications, such as video, voice, and chat-enabled collaboration platforms, or delivered via separate platforms, depending on the use case. Both customer-facing roles, such as contact center agents and sales teams, and employee-facing scenarios, such as communications and performance management, will benefit from this technology.

These coaches are tailored to individual employee needs by collecting and learning from data (e.g., baseline skills, work experience, career goals) and offering recommendations to drive skill acquisition via tailored learning paths. As an example, in looking at individual performance data, AI coaches can help sales teams identify skill gaps, suggest personalized learning paths, and provide customized, real-time feedback to help set up sales teams for success to close more deals.

Although there are role-specific coaches available for sales teams or contact center agents or platforms designed for managers, horizontal coaches can be used to improve soft skills such as communications in areas like presenting or public speaking at virtual and in-person events or meetings.

Whether offering guidance on effective communication or enhancing managerial excellence, AI coaching solutions can boost employee productivity by providing conversational support, personalized guidance, and role-playing scenarios to deliver tailored advice for managers and non-managers alike.

Market and Industry Perspectives

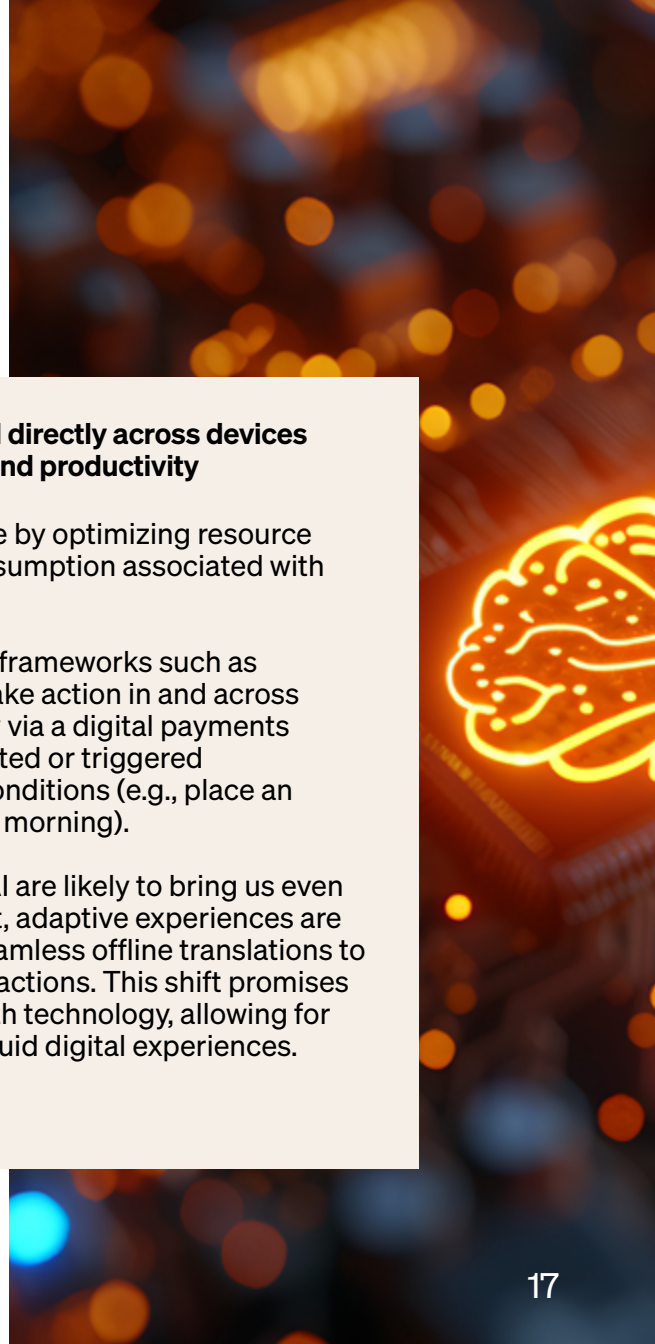
The global online coaching market is estimated to grow from just over $3B in 2022 to $11.7B by 2032.[4]

Existing HR platforms, learning solutions and new market entrants focused on skills coaching and existing job specific solutions (e.g., call center or sales technology) are all coming to market with AI coaching capabilities within the flow of work.

AI coaching has been largely driven by chat-based interactions, but new modalities are being introduced such as voice, video and avatars that can help provide real-time coaching feedback or the ability to roleplay performance management scenarios with AI-generated avatars.

Realistic practice scenarios for learners are gaining traction in the market, with these scenarios utilizing LLM-generated "employees" to simulate real-life situations, such as compliance policy training, agent-customer interactions, or managers navigating difficult conversations with employees.

On-device AI



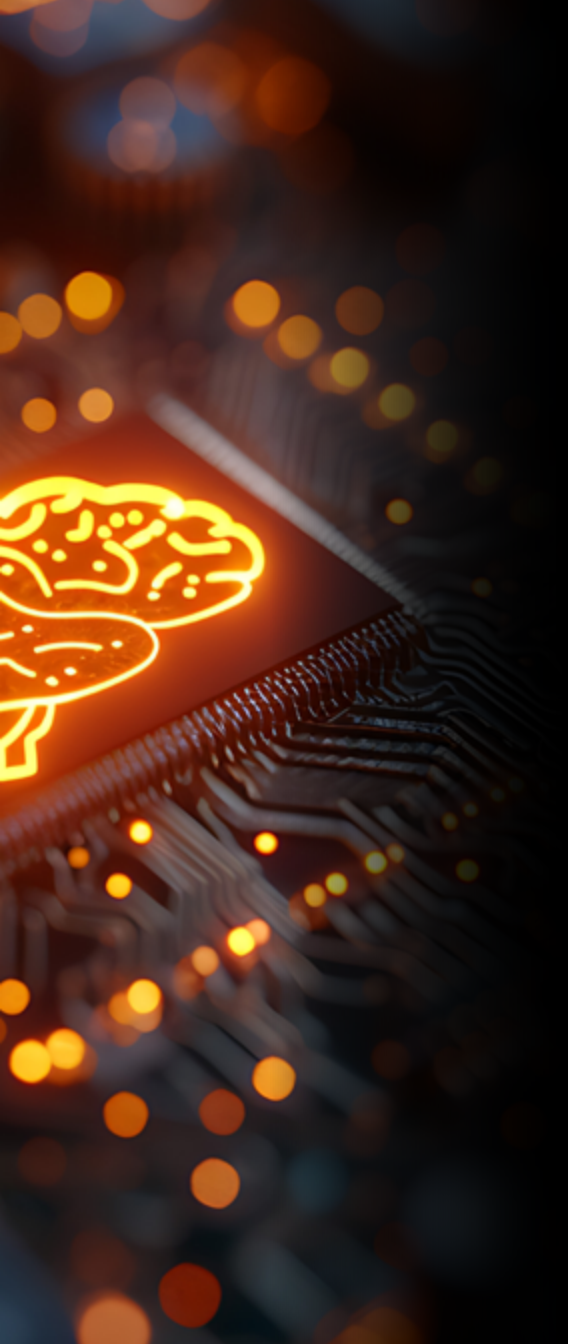
On-device AI represents a transformative step in computing where AI capabilities are embedded directly across devices like smartphones and laptops, without internet connectivity, to drive greater levels of efficiency and productivity

On-device AI represents a transformative step in computing, where AI capabilities are embedded directly within personal devices such as smartphones, laptops, and IoT devices. By processing data locally, on-device AI reduces reliance on the cloud, offering faster response times, enhanced privacy, and greater energy efficiency. Innovations in hardware—such as specialized AI chips and computing frameworks—and advancements in software now allow devices to handle complex tasks like natural language processing, image recognition, and predictive analytics autonomously, without constant internet connectivity.

The benefits of on-device AI extend to user experience and security. With data processed on the device itself, there is less dependency on remote servers, leading to quicker, more responsive applications. This setup provides stronger privacy safeguards, as sensitive data does not need to be transmitted to external servers, which is critical in sectors where data security and user trust are paramount. Furthermore, local processing can extend battery life by optimizing resource use and reducing the energy consumption associated with frequent data transfers.

Similarly, leveraging on device AI frameworks such as App Intents will enable users to take action in and across applications (e.g., sending money via a digital payments network) that can be voice-activated or triggered automatically based on certain conditions (e.g., place an order for coffee at this time every morning).

Future innovations in on-device AI are likely to bring us even closer to a world where intelligent, adaptive experiences are standard across devices, from seamless offline translations to real-time augmented reality interactions. This shift promises to enhance how users interact with technology, allowing for more personalized, private, and fluid digital experiences.



Market and Industry Perspectives

The on-device AI market size is projected to reach $114.4B by 2031, growing at a CAGR of 49.35% during 2024-2031.[5]

Market players are all making inroads in this space to advance on device AI to provide faster response times without relying on cloud servers, provide offline functionality while adapting to user behavior for tailored experiences that are both private and secure.

As on-device AI grows, AI Operating Systems (OS) will further transform how we interact across devices, creating more personalized user experiences.

Platforms now include enhanced AI processing capabilities with their Neural Processing Unit (NPU), enabling advanced tasks like real-time translation, image enhancement, and AR/VR applications directly on devices.

Agentic Financial Transactions

AI agents are poised to innovate digital commerce by autonomously conducting financial transactions, enhancing efficiency and user experiences – necessitating robust regulatory oversight

This trend could eventually impact the landscape of digital commerce by enabling AI agents to autonomously manage and execute financial transactions as part of an end-to-end process. This potential shift may reduce the need for human intervention, streamline processes, and enhance customer/ consumer experiences. By integrating advanced automation with secure payment technologies, AI agents might handle tasks ranging from scheduling and customer support to complex financial operations. As a natural evolution of agent capabilities, this development could open new avenues for businesses to improve efficiency, create innovative revenue streams, and offer seamless user experiences. As AI agents become more capable and reliable, they may increasingly be entrusted with activities of real-world economic significance, evolving into autonomous commercial actors.

These new forms of commerce offer significant potential, but as with every wave of innovation in payment systems, they come with risks. The evolution of AI agent payments will require robust controls, guardrails, and regulations to ensure safe and responsible operations.

This development signals a broader push by tech companies to create AI agents that can handle everything from product research to price comparisons and even purchases. This could reshape how consumers interact with eCommerce platforms and raise questions about the future role of human sales representatives and customer service agents.

AI-agent payments could transform the way consumers and businesses conduct transactions by automating tasks ranging from routine payments to complex negotiations. For example, consumer agents can optimize utility rates, manage recurring payments, enhance savings and investments, assist with travel bookings, inform product restocking, and handle gift purchases and returns. Business AI agents can negotiate supplier pricing, handle invoicing and payments, manage financial portfolios, and execute trades autonomously.

As the trend evolves, agent-to-agent fully autonomous AI workflows are likely to become more prevalent, further expanding the capabilities and reach of agentic financial transactions.

As the trend evolves, agent-to-agent fully autonomous AI workflows are likely to become more prevalent, further expanding the capabilities and reach of agentic financial transactions.

Market and Industry Perspectives

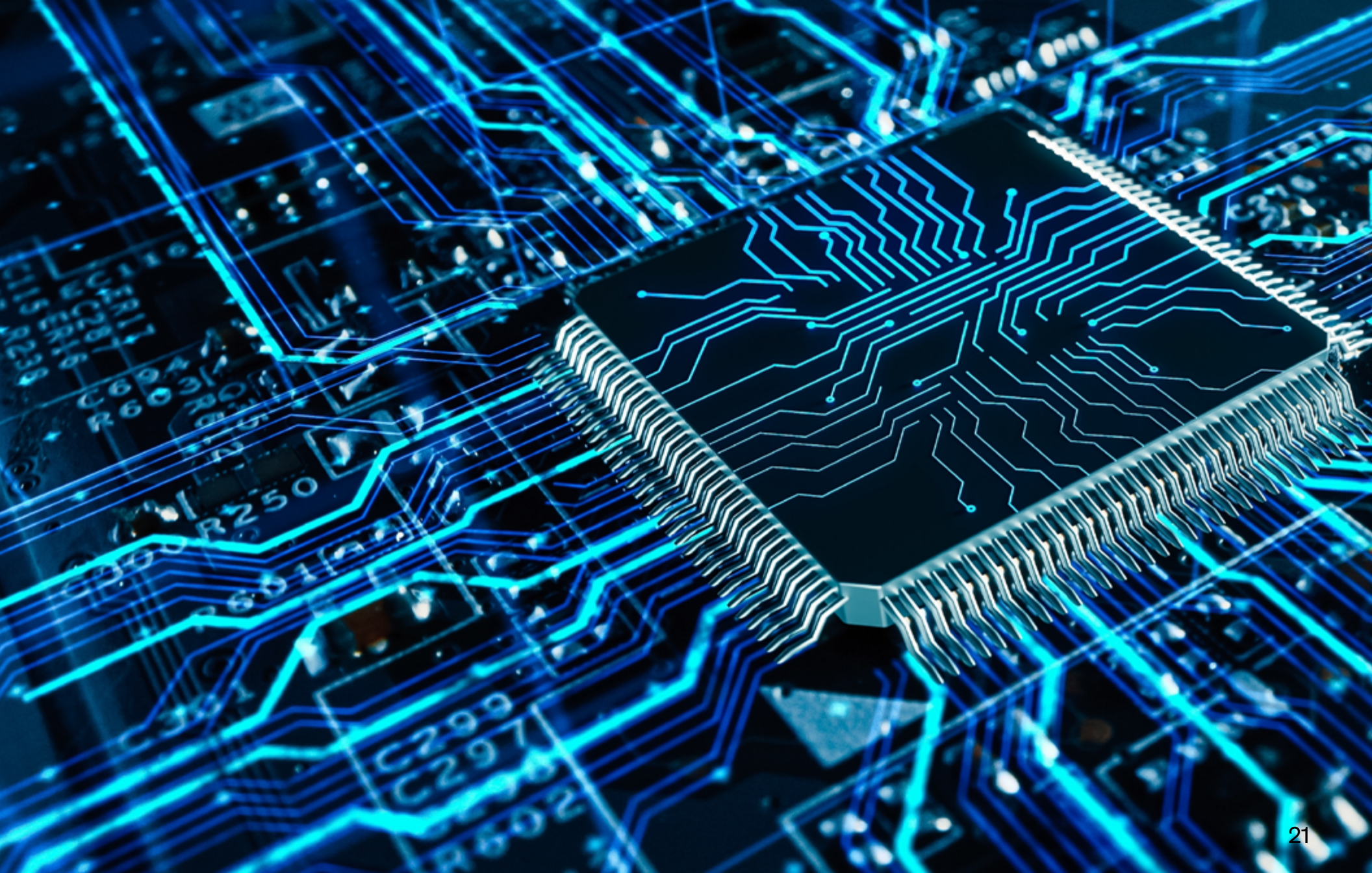
AI agents have launched to facilitate transactions, pay third parties using virtual credit cards, and to serve as shopping assistants (e.g., handling product questions, orders, and autonomous purchases).

The market is evolving, where AI agents will have the ability to use a computer similar to a human to complete tasks on a user's behalf (e.g., writing code, booking travel, browsing the web independently).

Efforts are underway to build the financial infrastructure to enable autonomous agentic AI commerce, with AI agents making and receiving payments instantly and globally without human intervention. By providing trusted identities, global financial rails, and governance through custom business rules, this infrastructure has the potential to unlock new revenue streams for enterprises.

Payment APIs are being developed to facilitate secure transactions for AI voice agents, enabling them to be multi-modal, where they can accept payments over the phone securely and facilitate transactions end-to-end (e.g., seamlessly booking hotel rooms and paying via credit card).

Open banking platforms have introduced agentic capabilities for financial institutions, which offer consumer agents that can be embedded into AI and customer-facing product suites to help individuals optimize their finances.



02 Delivering Data & AI/ML at Scale

In today's rapidly evolving technological landscape, data and AI are at the forefront of innovation, driving transformative changes across industries. As foundational models evolve, AI trends focus on optimizing inference-time compute to unlock new capabilities, leveraging synthetic data for model advancements, and enhancing data retrieval for more context-aware applications. Additionally, the shift from single-agent to multi-agent systems promises to revolutionize business operations, enabling organizations to fully harness AI-driven insights and automation.

Rise of Inference-Time Compute & Reasoning Models

The evolution of the Scaling Laws has led to the rise of inference-time compute and reasoning models, resulting in significant performance gains for GenAI models

The rapid advancement in the capabilities of foundation models, which underpin GenAI, has been largely driven by leading model providers' following the Scaling Laws (i.e., equivalent of "Moore's Law" to AI). These laws propose that model performance is influenced by three factors: the size of the model, measured by the number of parameters, which reflects the model's capacity to represent complex relationships in data; the amount of training data, which enables the model to learn from a variety of examples and contexts; and the compute power used for training, typically measured in terms of GPUs. In essence, increasing model size, training data, and compute power enhances model performance.

To date, model providers have adhered to the Scaling Laws during the pre-training phase of model development. This initial phase involves training the model on vast amounts of data to help it learn language patterns, grammar, and context. The objective of pre-training is to equip the model with foundational knowledge, enabling it to accurately predict the next token in a sequence.

While pre-training has yielded significant performance gains, the industry now faces challenges in scaling further, primarily due to limited compute resources for training larger models. Although this issue is expected to persist in the short term, supply and demand are likely to balance with the introduction of new GPUs, hyperscaler ASIC investments, and semiconductor innovations; however, the industry now faces the looming challenge of running out of training data.

Faced with these challenges, some in the industry have speculated about the decline of the traditional Scaling Laws. However, while we anticipate that model providers will continue investing in scaling the pre-training phase, the Scaling Laws are evolving, with a new focus centered on inference-time compute. In essence, this means that instead of encoding all knowledge during pre-training, models are being optimized to reason through problems step by step during inference.

Through innovations in models, providers like OpenAI have discovered that increased inference-time compute results in better performance. This has led to the development of "reasoning models," which are designed to "think" before responding. Traditionally, models generate tokens linearly without the ability to revise. In contrast, reasoning models use a "Chain-of-Thought" approach to break complex queries into smaller steps, enabling them to correct mistakes, revisit steps, and refine their strategy. This approach has resulted in significant gains in model performance, demonstrating that extended "thinking" with increased compute during inference time enhances outcomes, marking the next phase of the Scaling Laws.

Market and Industry Perspectives

In September, OpenAI unveiled a preview of o1, their reasoning model using the inference-time compute approach, followed by the full release in December.

Other major model providers have been exploring similar approaches, releasing models with advanced agentic capabilities. For instance, Google DeepMind published a paper in August 2024 titled “Scaling LLM Test-Time Compute Optimally Can Be More Effective than Scaling Model Parameters.” [6]

This approach is gaining industry mindshare. At the annual Microsoft Ignite event, CEO Satya Nadella remarked, “If anything, we are seeing the emergence of a new scaling law with test time or inference time compute.” [7] With the initial waves already evident, we anticipate that leading model providers will invest heavily in testing inference-time compute, aiming to develop more performant models with advanced reasoning capabilities.

Impact of Synthetic Data in Post-Training

The evolution of Scaling Laws is shifting focus from pre-training to post-training, leveraging synthetic data to overcome data constraints, thereby enhancing model capabilities and driving significant performance gains in GenAI development

As highlighted in the previous trend, the traditional Scaling Laws are evolving, with a shift away from a sole focus on pre-training due to challenges like limited compute resources and data availability. While massive investments from hyperscalers, companies are addressing compute constraints, the scarcity of training data has prompted model providers to innovate, particularly by leveraging synthetic data in post-training to enhance model capabilities.

In pre-training, the objective is to "predict the next token" using vast internet data, equipping models with foundational language understanding. Post-training, however, involves smaller, curated datasets designed to refine models for specific tasks, enhancing their practical utility. Techniques like Reinforcement Learning with Human Feedback (RLHF) and supervised fine-tuning (SFT) have been foundational in this phase. However, constructing diverse, high-quality datasets for SFT remains a significant challenge. This, combined with the limitations of scaling human input in RLHF, has shifted the focus towards using model-generated synthetic data to drive further advancements.

While model-generated synthetic data has been used in previous post-training efforts, applying it alongside reasoning models that create "chains-of-thought" to derive answers offers significant potential for performance gains. By scaling the number of reasoning paths a model explores, and using models to evaluate the optimal path, providers can generate high-quality synthetic data for post-training. Training on these optimal paths enhances the model's ability to "think" and reason through functional and verifiable tasks in areas like math, coding, engineering, and physics, while also improving generalization across diverse domains.

By training models to "think" and reason on functional and verifiable tasks, their overall capabilities are significantly enhanced. This shift in focus from pre-training to post-training and inference-time compute marks a new phase in AI development, where the ability to adapt and optimize through synthetic data and reasoning models becomes a key driver of progress. As the industry continues to explore these new dimensions of scaling, we can expect further breakthroughs in model performance and application.

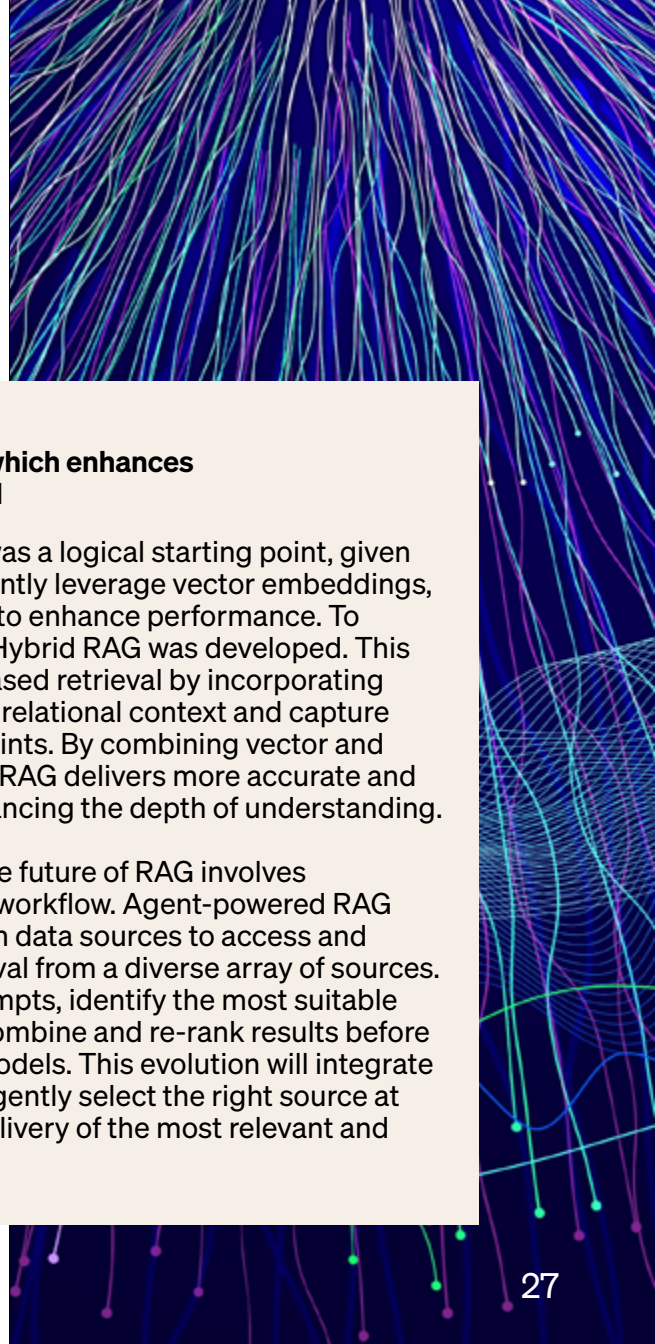
Market and Industry Perspectives

In line with the industry's focus on synthetic data, Gartner forecasts that 60% of the data used for AI and analytics projects this year will be synthetically generated. [7]

At his latest keynote at NeurIPS, Ilya Sutskever, OpenAI's co-founder & former Chief Scientist, stated "Pre-training as we know will unquestionably end. We've achieved peak data and there'll be no more. We have to deal with the data that we have. There's only one internet." [8]

Throughout the past year, virtually every leading model provider has openly acknowledged using model-generated synthetic data in the training process for models like OpenAI's GPT-4o , Anthropic's Claude Sonnet 3.5[9], Meta's Llama 3.1 405B[10], and Microsoft's Phi-4. [11] Additionally, Nvidia has released Nemotron-4 340B, a family of models specifically designed to generate high-quality synthetic data. [12]

Evolution of Retrieval Augmented Generation



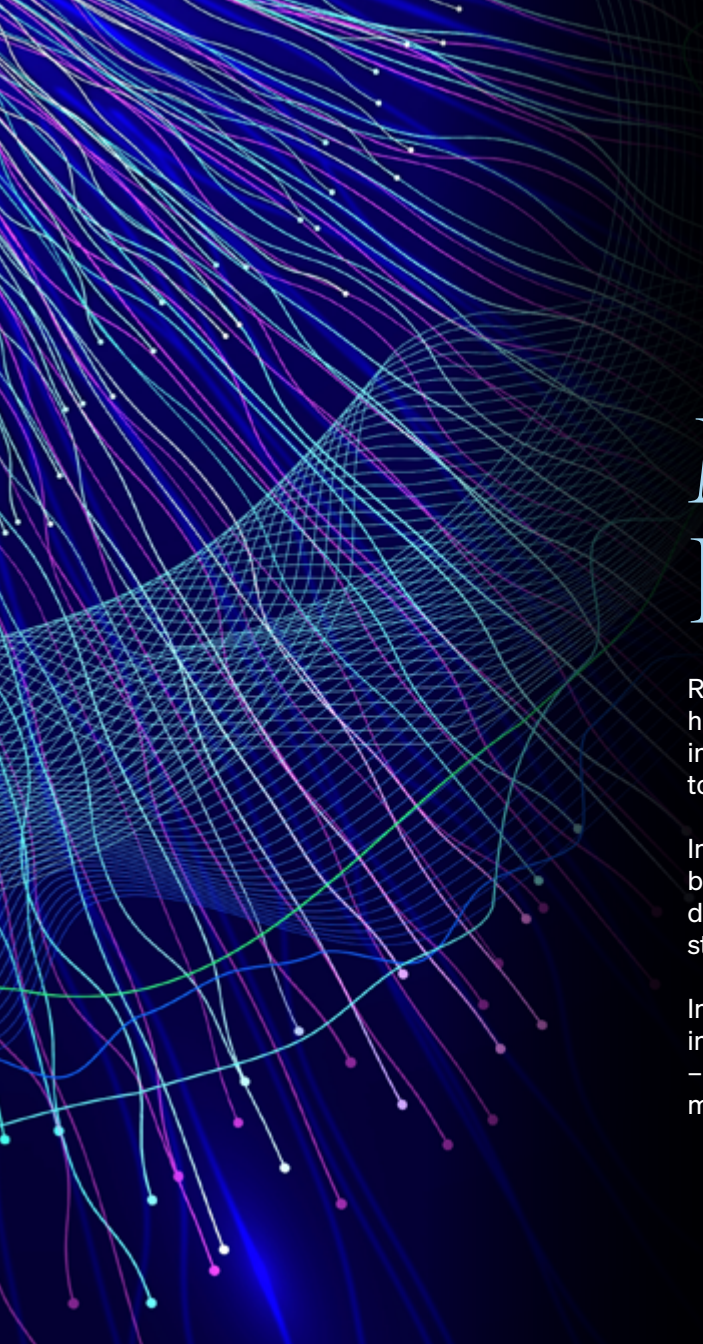
Retrieval Augmented Generation (RAG) is evolving from vector-based methods to Hybrid RAG, which enhances accuracy, and is evolving towards agent-powered systems for smarter, context-rich data retrieval

Retrieval Augmented Generation (RAG) has emerged as the primary technique for integrating proprietary data into model responses. This is crucial because foundation models, such as GPT-4 and Claude, are trained on public data and therefore lack specific business context. RAG enables the enrichment of model outputs by incorporating relevant, proprietary information.

The initial RAG approach focused on vector-based retrieval, where data is stored in vector databases for fast and efficient access. In this method, data is transformed into high-dimensional vectors using an embedding model, which captures the semantic relationships within the data. These vectors are then stored in a vector database. When a query is made, the same embedding model typically transforms the query into vector embeddings, allowing the system to efficiently search the vector database for relevant information. While effective, this approach is primarily suited for unstructured data, such as PDFs and PowerPoint presentations.

While vector-based retrieval was a logical starting point, given that foundation models inherently leverage vector embeddings, there remained opportunities to enhance performance. To address these opportunities, Hybrid RAG was developed. This innovation builds on vector-based retrieval by incorporating knowledge graphs, which add relational context and capture relationships between data points. By combining vector and graph-based retrieval, Hybrid RAG delivers more accurate and precise answers, thereby enhancing the depth of understanding.

As data sources proliferate, the future of RAG involves embedding agents within the workflow. Agent-powered RAG will be able to determine which data sources to access and orchestrate information retrieval from a diverse array of sources. These agents will analyze prompts, identify the most suitable data stores, and seamlessly combine and re-rank results before passing them to generative models. This evolution will integrate multiple data types and intelligently select the right source at the right time, ensuring the delivery of the most relevant and context-aware outputs.



Market and Industry Perspectives

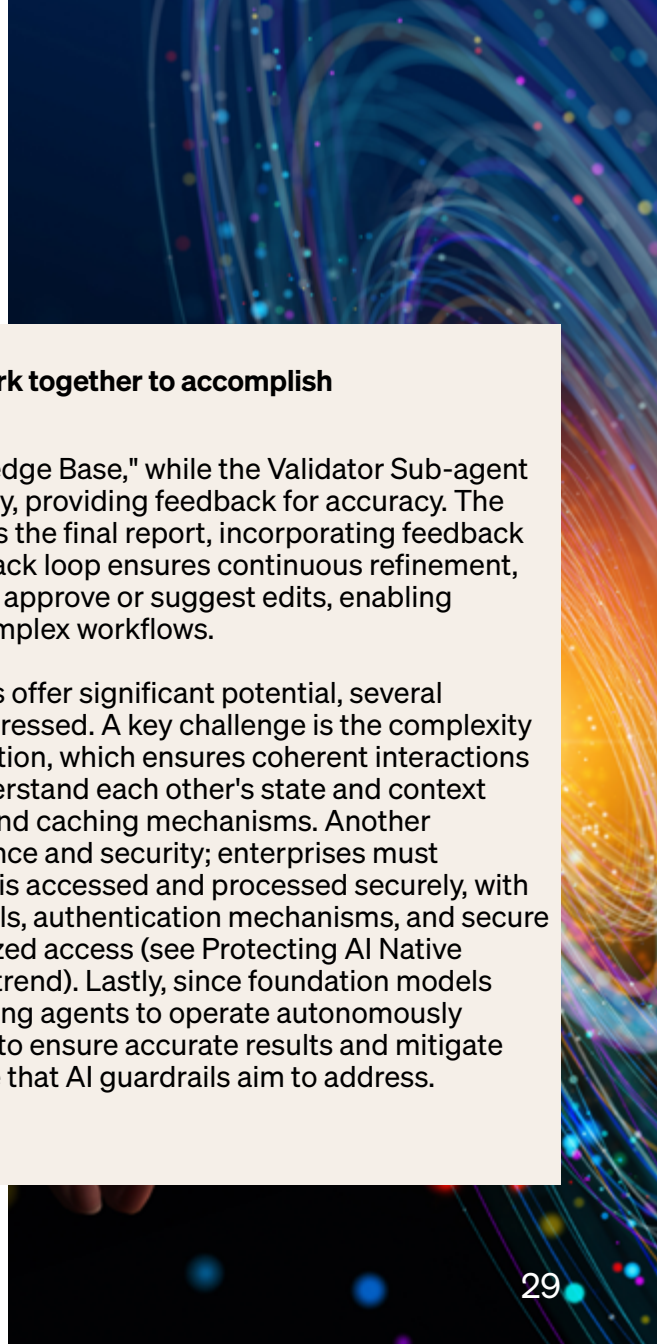
Recognizing RAG as essential for accurate and high-performing GenAI applications, hyperscalers are investing in comprehensive tooling to encourage users to build on their platforms and drive compute revenue.

Initially, these platforms focused solely on vector RAG, but they are now evolving, as demonstrated by new developments in both graph RAG and structured retrieval.

In addition to hyperscalers, smaller players are heavily investing in GenAI application-building capabilities – particularly around RAG, given their integral role in managing and accessing data. This is evident through in-house innovations from these companies through product enhancements and innovations, which offer unstructured data transformation and specialized search engine capabilities.

Lastly, companies are focused on developing intelligent retrieval models that excel at retrieving the most relevant information and accurately ranking, summarizing, and combining it for generative models. We anticipate further innovations in retrieval capabilities from these players.

Transition from Single Agent to Multi-Agent Systems



AI agents are advancing from single-agent to multi-agent systems, where specialized agents work together to accomplish complex tasks, providing greater automation and efficiency

Last year's trends discussed the emergence of AI agents, which have the ability to understand complex queries, effectively reason and plan, decompose tasks into smaller subgoals, reflect on and learn from mistakes, and ultimately take action. AI agents are underpinned by foundation models, which have access to components such as memory – both short-term (e.g., context window) and long-term (e.g., knowledge store) – as well as tool use, enabling them to call external tools like web APIs, Excel, or PowerPoint.

While 2024 introduced the concept of single-agents, 2025 will see the rise of multi-agent systems, consisting of multiple interacting agents that collaborate and coordinate to achieve a collective goal. These agents run on agentic frameworks, where an 'orchestrator agent' manages the process by breaking down tasks into subgoals and delegating them to specialized sub-agents that work in parallel.

For example, imagine a multi-agent system designed for comprehensive research assessments. A human initiates a task via a user interface, prompting an orchestrator agent to break it down into a detailed plan and delegate steps to specialized sub-agents. The Researcher Sub-agent gathers information using tools like "Web Search" and "Knowledge Base," while the Validator Sub-agent evaluates the draft's quality, providing feedback for accuracy. The Writer Sub-agent compiles the final report, incorporating feedback for improvement. A feedback loop ensures continuous refinement, and a human reviewer can approve or suggest edits, enabling efficient automation of complex workflows.

While multi-agent systems offer significant potential, several challenges need to be addressed. A key challenge is the complexity of inter-agent communication, which ensures coherent interactions by allowing agents to understand each other's state and context through shared memory and caching mechanisms. Another challenge is data governance and security; enterprises must ensure that sensitive data is accessed and processed securely, with appropriate access controls, authentication mechanisms, and secure APIs to prevent unauthorized access (see Protecting AI Native and Agentic Applications trend). Lastly, since foundation models are not deterministic, scaling agents to operate autonomously requires careful oversight to ensure accurate results and mitigate hallucinations, a challenge that AI guardrails aim to address.

Market and Industry Perspectives

Virtually every technology company, from hyperscalers and data platforms to model providers and startups, is aiming to capture value in the multi-agent ecosystem, especially as projections indicate that by 2028, one-third (33%) of interactions with GenAI will use autonomous agents to complete tasks.[13]

Model providers are heavily investing in innovations around reasoning and planning. Emerging startups are competing across multiple layers of the stack – offering open-source frameworks for building flexible agentic applications. Emergence is focused on enhancing orchestration agents with novel reasoning and planning techniques. Some companies provide low-code agent builders and customizable registries.

Alongside innovations enabling users to build multi-agent applications, nearly every existing SaaS vendor is integrating these capabilities into their core products and services.

Lastly, there is an entire ecosystem of vendors aimed at solving challenges around governance, security, and guardrails.

Automation Platforms

Automation is shifting from a siloed to unified approach, the platforms of tomorrow will serve as centralized hubs which will call on a variety of task optimized automation approaches to enhance end-to-end orchestration and business process management

Automation has evolved from traditional deterministic approaches to more recent GenAI based non-deterministic approaches, which include LLMs, Large Action Models (LAMs), and AI agents.

Examples of traditional deterministic approaches include Robotic Process Automation (RPA) which solved for legacy back office integration by recording tasks, orchestrating tasks, and repeating with bots to deliver to source systems (e.g., ERP, CRM), Business Process Automation (BPA) which enabled pre-defined and complex workflows, tasks, and decisions, Low Code App Platforms (LCAPs) which enabled business users to develop apps with ease (e.g., drag and drop, scalability), and integration Platform as a Service (iPaaS).

Traditional approaches are often focused on automating siloed tasks, and are typically implemented on business needs to generate cost savings. However, today we are at an inflection point. The rapid pace of innovation from GenAI non-deterministic players has caused the market to coalesce, where GenAI players seek to orchestrate similar processes as traditional players, but with GenAI approaches.

In response, traditional players and new market entrants have begun to shape the future of automation platforms, which offer a combination of both deterministic and non-deterministic capabilities.

There are several benefits to combining both deterministic and non-deterministic capabilities, such as full end-to-end automation (non-supervised / non-deterministic) to call on a variety of automation approaches (e.g., decision intelligence, bots, LLMs, Agents, LAMs, process mining, workflow design) based on specific business needs. In addition, orchestration is at the core, with unified governance to standardize on and connect into source systems.

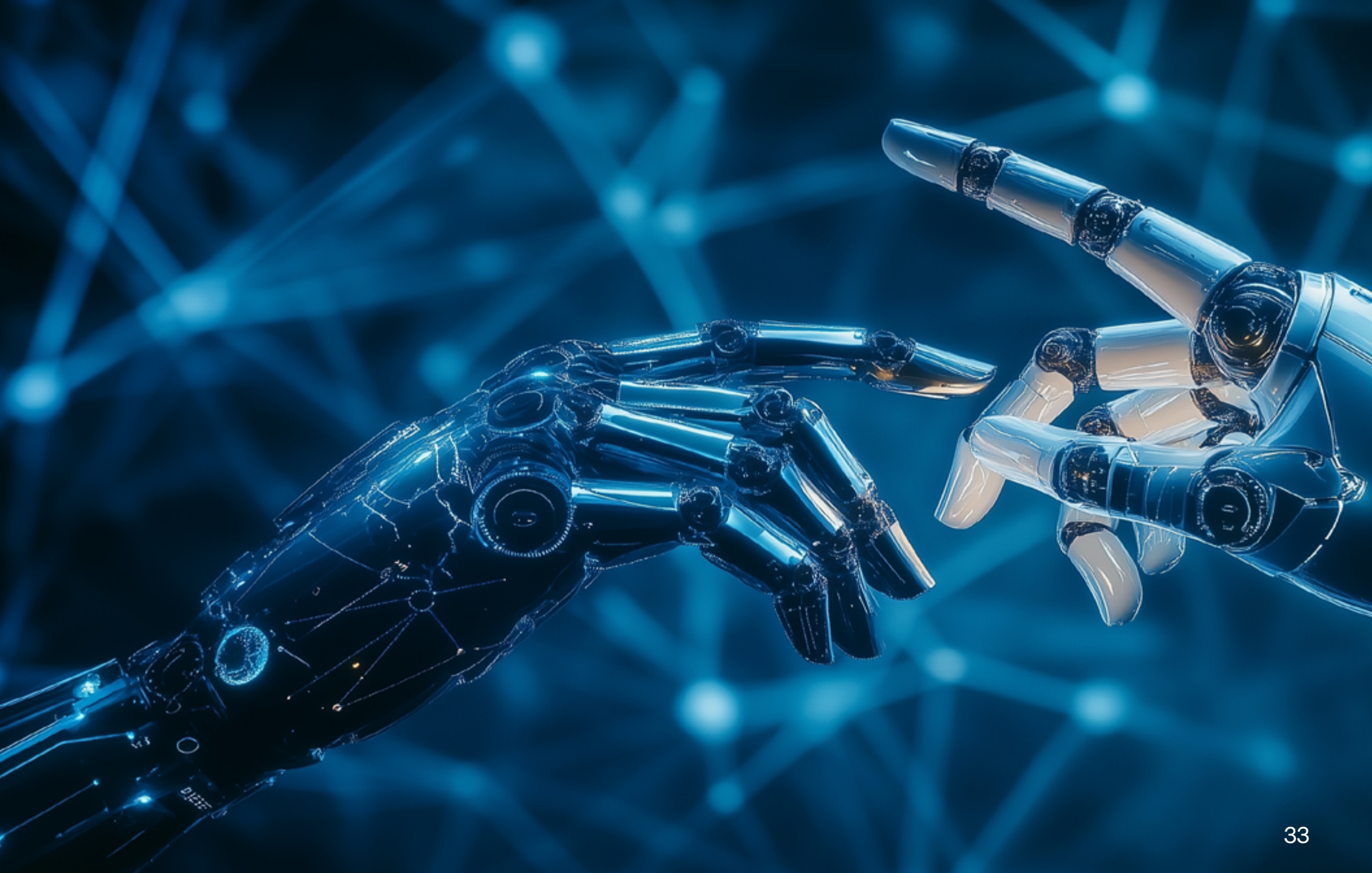
Market and Industry Perspectives

Traditional players and new market entrants are starting to converge capabilities to orchestrate and govern both deterministic and probabilistic automation techniques to support end-to-end business processes.

Emerging players have begun launching AI agents and enterprise automation platforms which operate by observing, learning, automating and adapting to new complex scenarios. These platforms typically leverage LAMs to interpret, predict, and execute actions.

Similarly, traditional players have begun embedding agent, orchestration, and autopilot capabilities into complex processes to deliver more granular levels of automation than co-pilots.

Gartner has termed the phrase “Business Orchestration and Automation Technologies” (BOAT) to describe this trend of convergence of automation platforms.[14]



03 Technology Modernization

The industry for cloud computing and modern engineering has evolved rapidly to keep up with the growing demand for GPUs to support AI workloads, and more efficient and faster ways for our developers to build. Innovations in the data center have emerged to support the hardware requirements of these workloads, including new power sources, advanced cooling techniques, and more efficient storage and networking solutions. In parallel, new software techniques have been developed to manage these underlying resources and optimize usage for improved workload performance.

AI continues to drive new paradigms in software development, transitioning from in-line code suggestions and chat interfaces to more autonomous development through software developer agents. This new method aims to transform how developers work and ultimately uplift the developer experience.

Next Gen Data Center Design

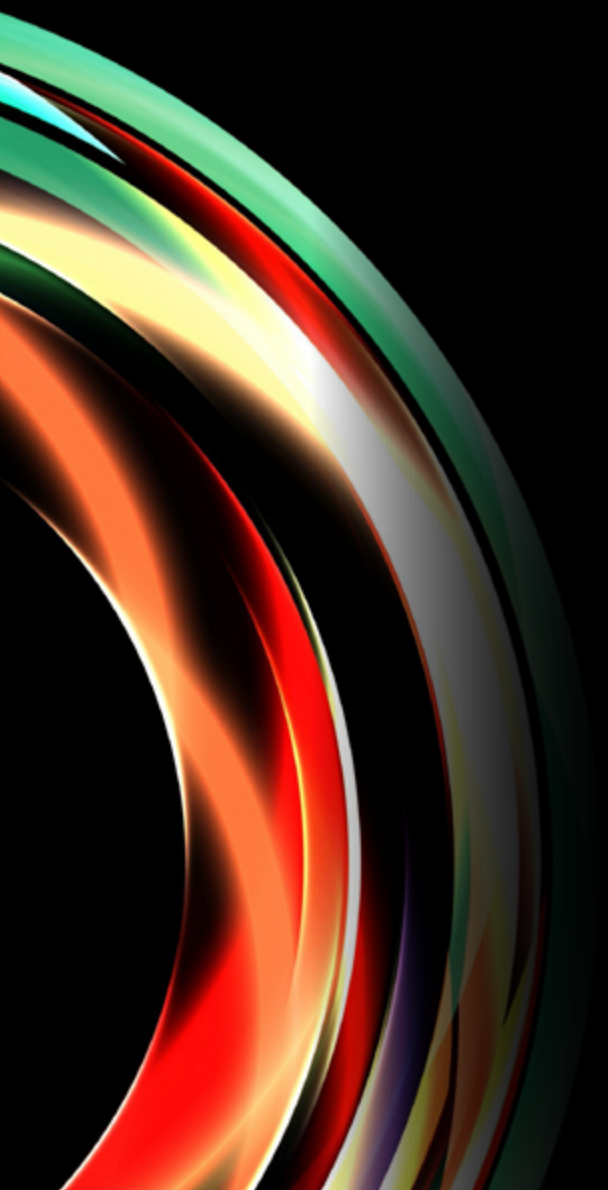
The evolution of data centers is being influenced by AI and workloads that require high-performance compute systems that support increased cabling, power consumption, heating, and cooling

The demand for elastic compute and evolving workloads continues to have a major impact on the overall data center market, as both capital investments and supply constraints have encouraged innovation. Every aspect of the data center is being reconsidered to accommodate new high-performance systems that support optimal workloads (e.g., AI), including server racks, innovative cooling techniques, geographical locations, and power sources.

For instance, server racks are being modified to accommodate heavier systems with complex cabling requirements while minimizing the risk of overheating. New approaches to liquid cooling are being applied to manage systems that use massive amounts of power and generate significant heat. These include methodologies like direct-to-chip liquid cooling (circulating liquid coolant directly to CPUs or GPUs), immersion cooling (where servers are submerged in a non-conductive liquid coolant), and in-row cooling (where liquid coolants are placed between server racks to absorb heat).

As new data centers continue to be built across the world, we are seeing large cloud providers strategically constructing data centers next to novel power sources, such as nuclear, to ensure access to energy capacity. Additionally, many cloud providers have made energy and sustainability commitments and view these new power sources as a potential way to reduce their carbon footprint. As a result of these new builds, the demand for materials will affect supply chain trends for years to come.

These new approaches are intended to rapidly satisfy existing demand for AI, but they will inevitably lead to new advances that will impact data center designs for years to come.



Market and Industry Perspectives

Digital giants and cloud providers such as AWS, Microsoft, and Google are developing new data center designs and techniques. While some of these innovations will remain proprietary, many are being contributed as open-source standards that will drive the industry forward.

Colocation providers are bringing new facilities online to attract both web-scale and enterprise customers seeking AI compute capabilities.

Original equipment manufacturers (OEMs) with industry chip manufacturing giants, are promoting common approaches while striving to differentiate themselves. Meanwhile, specialist providers are known for liquid cooling solutions and are experiencing significant demand.

AI Platforms-as-a-Service / AI Clouds

Simplifying the way infrastructure is accessed and scaled will optimize the way enterprises adopt AI technologies

Just as we saw with web-scale infrastructure, AI infrastructure is creating demand for consumption models and abstraction layers to simplify the management and deployment of workloads for developers and data scientists. While the large public cloud providers continue to invest in services to simplify access to Graphics Processing Units (GPUs) with Machine Learning Operations (MLOps) capabilities, several new players are focused solely on this problem statement. There are two main categories: AI Platforms as a Service (AI PaaS) and AI Clouds.

AI PaaS technologies provide a consistent platform on AI infrastructure (e.g., GPU clusters) to help provision, orchestrate, and deploy workloads while providing MLOps tooling to run training and inference jobs. These technologies typically offer several optimization techniques to improve model performance and reduce infrastructure costs.

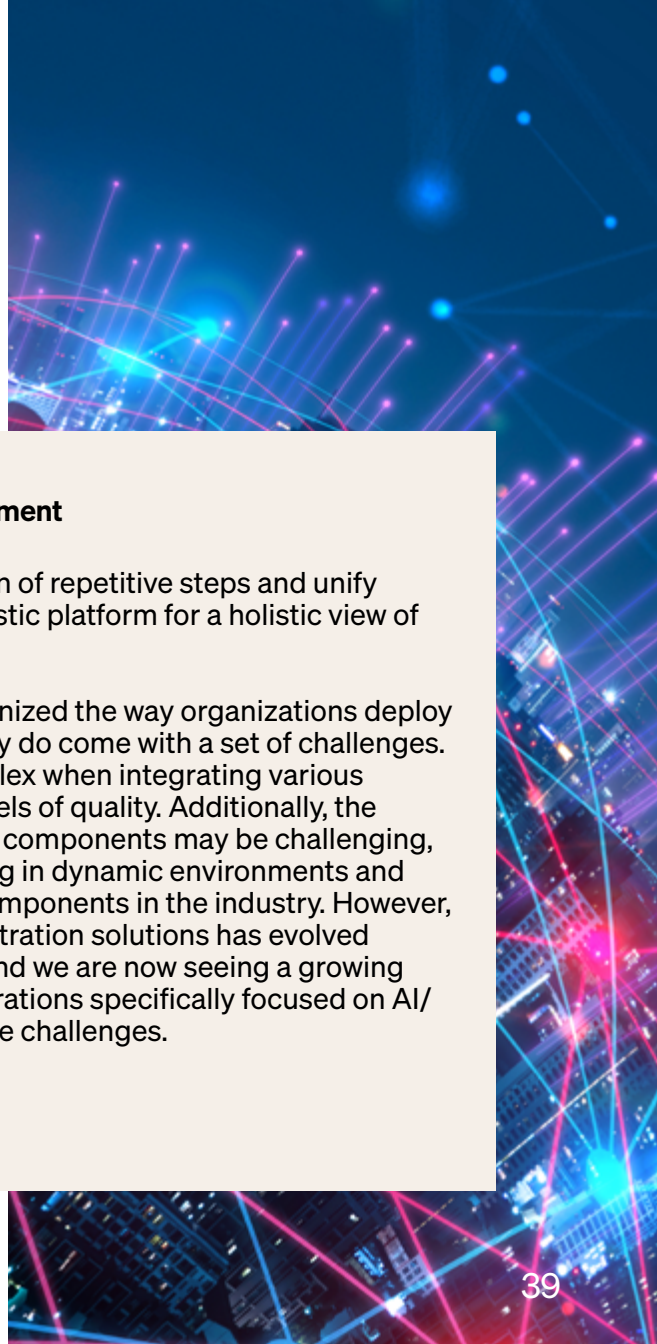
AI Clouds similarly offer abstraction layers, but their core focus is typically on reselling the underlying GPU environments, so their investment in software may be less than that of platform providers that aim to run on any AI infrastructure.

As the demand for efficient management and deployment of AI workloads grows, both established cloud providers and emerging companies are innovating to meet these needs. AI Platforms as a Service (AI PaaS) and AI Clouds represent two distinct approaches to addressing these challenges, each with its own strengths and focus areas. While AI PaaS emphasizes comprehensive platform solutions with robust MLOps capabilities and optimization techniques, AI Clouds concentrate on providing scalable GPU environments with varying levels of software investment. As the landscape continues to evolve, these solutions will play a crucial role in empowering enterprises to harness the full potential of AI technologies, ultimately driving innovation and efficiency across industries.

Market and Industry Perspectives

AI Clouds provide software (e.g., Slurm on Kubernetes) to help teams deploy and manage their workloads using common frameworks. Startups are taking different approaches to the same problem—helping data scientists optimize their use of expensive AI infrastructure through simpler resource management or improved usage optimization. Given the vast amounts of capex being spent on AI infrastructure and the known failure rates of these highly complex systems, the market will continue to focus on new approaches to manage and optimize these important resources.

AI Workload Orchestration



Orchestration tools will help streamline and automate AI/ML workflows and application management

The rise of GenAI has led to a surge in infrastructure investment. While much attention is given to hardware innovations, such as GPUs and AI silicon, the cloud-native ecosystem is equally focused on extending the benefits of workload scheduling and orchestration solutions into AI infrastructure and the overall machine learning lifecycle. These orchestration tools efficiently connect various AI tools and systems to help streamline the end-to-end AI lifecycle and optimize compute resources.

The benefits of orchestration tools can be seen today with Kubernetes, which has become the de facto standard for deploying and managing containerized workloads. Kubernetes allows instances to dynamically scale up and down based on demand and can automatically terminate pods with errors. The same principles can be applied to AI/ML workloads, which require massive amounts of computing power. Given their dynamic nature, AI Workload Orchestration tools can help streamline AI/ML tasks (such as fine-tuning and inference) through automation of repetitive steps and unify workflows into a vendor agnostic platform for a holistic view of your workloads.

While these tools have modernized the way organizations deploy and manage applications, they do come with a set of challenges. Data integration can be complex when integrating various data sources, formats and levels of quality. Additionally, the interoperability of different AI components may be challenging, with changes rapidly occurring in dynamic environments and varying standards of these components in the industry. However, the ecosystem around orchestration solutions has evolved dramatically over the years, and we are now seeing a growing community of tools and integrations specifically focused on AI/ML workloads to combat these challenges.

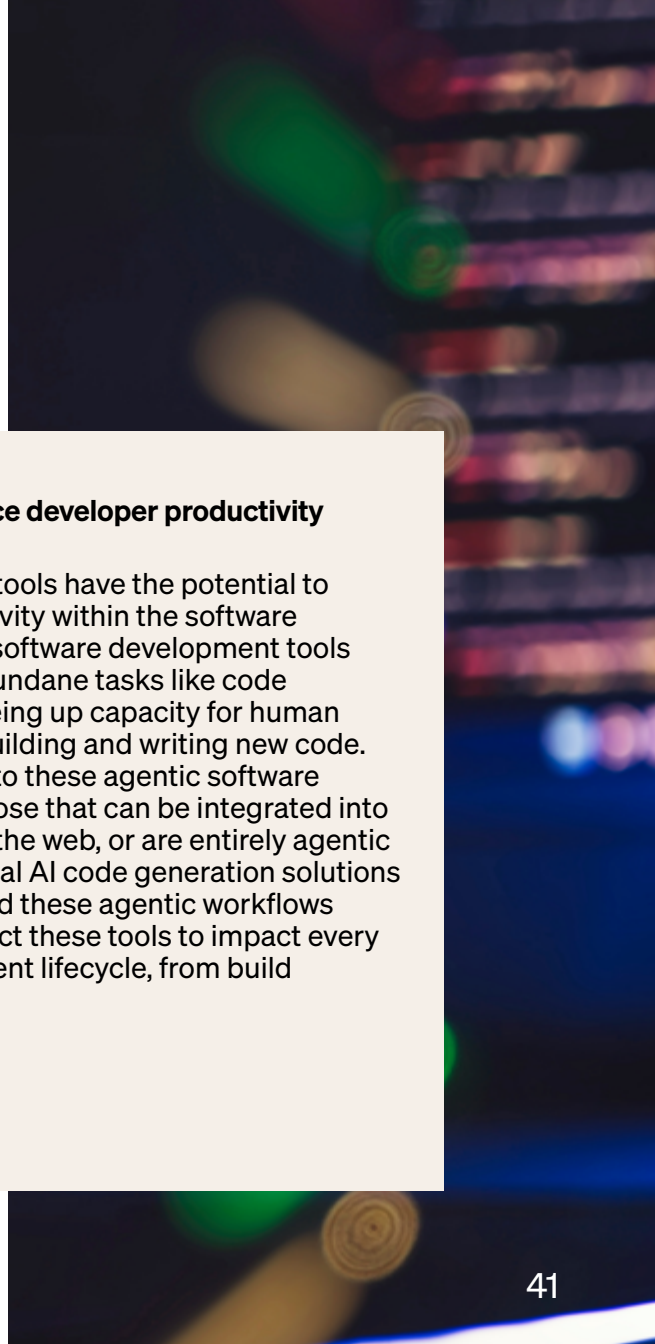
Market and Industry Perspectives

"Research teams can now take advantage of the frameworks we've built on top of Kubernetes, which make it easy to launch experiments, scale them by 10x or 50x, and take little effort to manage." - Christopher Berner, Head of Infrastructure at OpenAI.[15]

By adopting Kubernetes, OpenAI enjoys enhanced portability, enabling easy movement of research experiments between clusters. The consistent API provided by Kubernetes simplifies this process. Furthermore, OpenAI can leverage their own data centers in conjunction with Azure, resulting in cost savings and increased availability.

Large Kubernetes (K8s) deployments are effectively managing GPU resources from cloud providers like CoreWeave. Additionally, AI-specific operators, such as Azure's Kubernetes AI Toolchain Operator, are being developed. The Kubeflow portfolio of tools now includes support for techniques like hyperparameter tuning.

Agentic Software Development



Autonomous software tools for tasks like code generation, testing, and review will further enhance developer productivity

GenAI has significantly impacted how users and enterprises approach software development, with growing adoption of tools for code generation, test generation, documentation and code reviews. These tools have matured over the past year, but new tools have emerged to deliver a more autonomous approach known as Agentic Software Development.

In this form of development, users interact with a chat interface – similar to those that exist today in integrated development environments (IDEs) – and can ask an agent to build or run something in plain English with specific requirements. These "agents" can interpret the prompt, create a step-by-step plan on how to execute the task, and run it autonomously with access to build environment tools like terminals, compilers and programming languages; human developers remain in the loop but do not need to be involved at every step. Additionally, the agent can check in from time to time to confirm that its approach aligns with the developer's expectations.

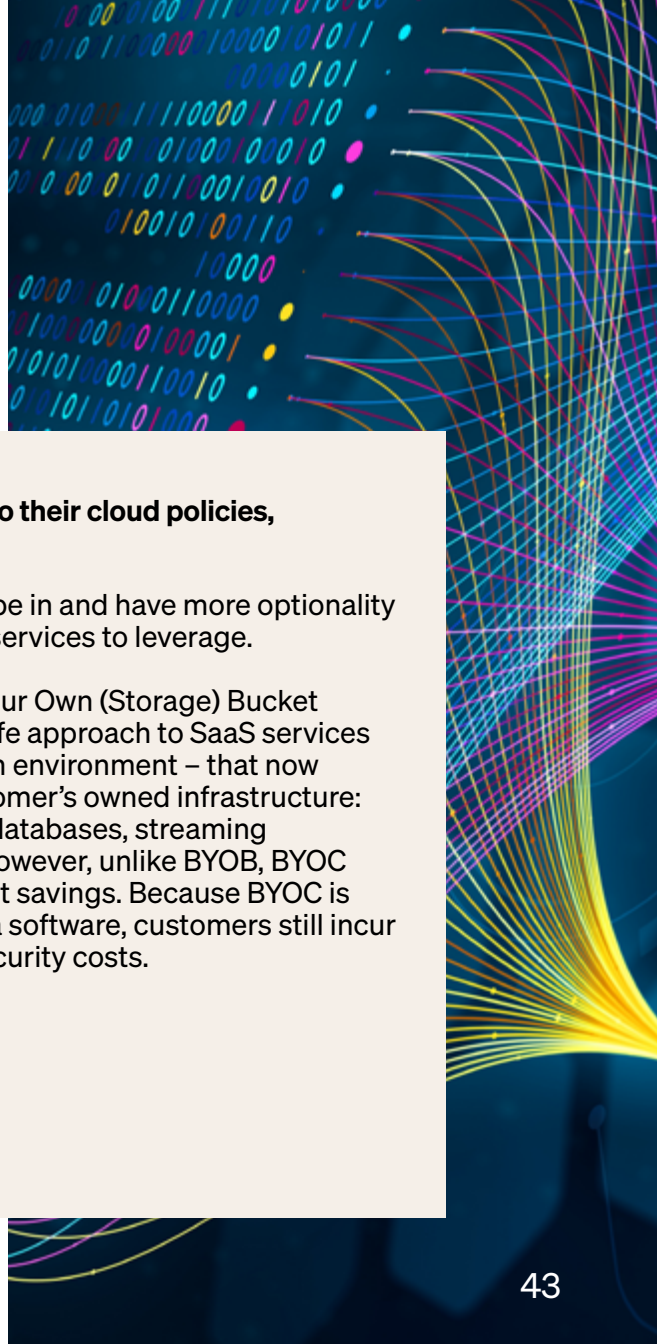
While still in early stages, these tools have the potential to increase efficiency and productivity within the software development lifecycle. Agentic software development tools can focus on more repetitive, mundane tasks like code renovations and migrations, freeing up capacity for human developers to concentrate on building and writing new code. There are different approaches to these agentic software development tools, including those that can be integrated into a developer's IDE, accessed via the web, or are entirely agentic IDEs themselves. Many traditional AI code generation solutions in the ecosystem have integrated these agentic workflows into their offerings, and we expect these tools to impact every stage of the software development lifecycle, from build to deployment.

Market and Industry Perspectives

The global AI code tools market size is expected to reach USD $27B by 2032, according to a study by Polaris Market Research.[16]

Large funding announcements from Agentic Software Development solutions are indicative of growing market adoption and interest. Emerging startups are each building their own large language models to power their agentic solutions, have received millions of dollars in investments. Despite their large funding amounts, neither solution is generally available today, suggesting strong investor confidence in this space. Industry leaders have announced agentic capabilities alongside their existing AI code generation tools. While the maturity of these tools varies, this suggests that the original AI code generation tools in the ecosystem continue to innovate and are expanding into other parts of the software development lifecycle to optimize processes. While the maturity of these tools varies, this suggests that the original AI code generation tools in the ecosystem continue to innovate and are expanding into other parts of the software development lifecycle to optimize processes.

Bring Your Own Cloud



New hosting constructs will give enterprises optionality when deploying workloads that adhere to their cloud policies, resiliency patterns and controls

Software hosting models have shifted over the years and range across self-hosted, Software as a Service (SaaS) self-hosted, and fully managed SaaS. The decision around which hosting model to leverage depends on the function of the service, and while more companies are adopting fully managed SaaS today, concerns around security, cost, resiliency and data sovereignty continue. Conversely, a self-hosted deployment ensures full control over the data but also requires customers to manage and maintain their own infrastructure. A new hosting model called Bring Your Own Cloud (BYOC) sits between these two models and looks to address issues while also providing the benefits of both.

In BYOC, a vendor's software is deployed in a customer's cloud environment (often called a Virtual Private Cloud or VPC), which is where the data also stays, while the vendor remotely manages the deployment. This allows for more data privacy and sovereignty. Customers can pick which cloud provider and the region they would prefer to be in and have more optionality and flexibility in terms of cloud services to leverage.

BYOC is an evolution of Bring Your Own (Storage) Bucket (BYOB) – a cost effective and safe approach to SaaS services using object storage in your own environment – that now expands to all aspects of a customer's owned infrastructure: identity providers, networking, databases, streaming infrastructure, AI models, etc. However, unlike BYOB, BYOC does not necessarily ensure cost savings. Because BYOC is still based on a subscription to a software, customers still incur infrastructure and additional security costs.

Market and Industry Perspectives

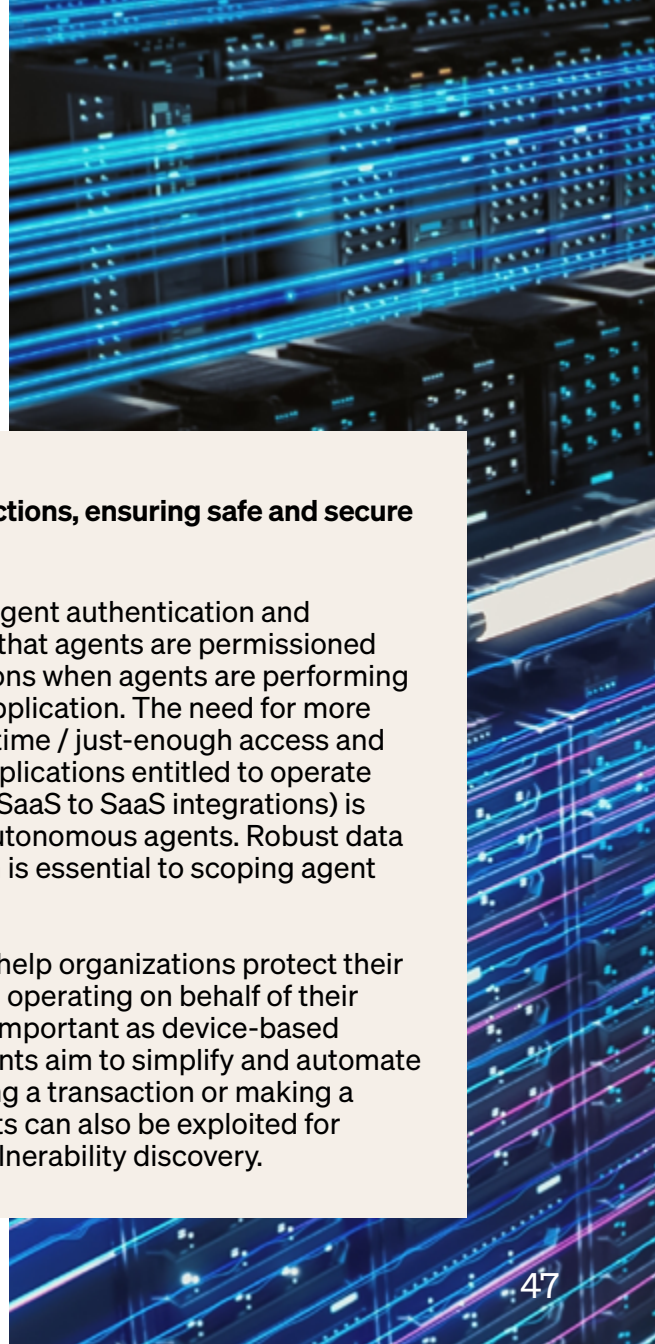
Several vendors offer BYOC support as a core product offering.

With data privacy concerns related to information being shared with foundational model providers, many vendors are offering Bring Your Own Model (BYOM) solutions where either proprietary or open models can be deployed within the customer's cloud account.

04 Protect the Firm

In today's fast-evolving digital landscape, cybersecurity must adapt swiftly to keep pace with technological advancements. As AI continues to revolutionize various sectors, trends in security are focused on safeguarding AI-native applications, AI-generated content, and sensitive data during AI computations. Additionally, the innovative use of AI by cybersecurity operations teams enables them to stay ahead of emerging threats.

Securing Agentic Applications



Emerging security solutions focus on governing AI agents to prevent unauthorized access and actions, ensuring safe and secure interactions by enforcing policies, monitoring behaviors, and managing permissions

The use of AI agents, which can operate dynamically and autonomously, is accelerating and driving new techniques to secure and govern AI-driven interactions and workflows. If safety and security guardrails are not considered, AI agents can expose sensitive data or be granted overly permissive and unmonitored access. They may also generate and execute malicious code and take other disruptive or potentially destructive actions if misdirected by threat actors.

Emerging security solutions are focused on improving the security posture and governance of agents across an organization. These solutions can discover and profile agents, enforce configuration policies and audit / monitor agent behaviors. Specific controls will also be required for certain classes of agents (e.g., integrated vulnerability scanning for software development agents).

Another emerging focus is on agent authentication and authorization of tools and data that agents are permissioned to access. This includes situations when agents are performing actions on behalf of a user or application. The need for more stringent controls (e.g., just-in-time / just-enough access and dynamic revocation) around applications entitled to operate on the enterprise's behalf (e.g., SaaS to SaaS integrations) is magnified by the potential of autonomous agents. Robust data management and classification is essential to scoping agent permissions appropriately.

Solutions are also emerging to help organizations protect their external interfaces from agents operating on behalf of their customers. This is particularly important as device-based agents and browser-based agents aim to simplify and automate tasks for users such as disputing a transaction or making a payment. However, these agents can also be exploited for malicious purposes, such as vulnerability discovery.

Market and Industry Perspectives

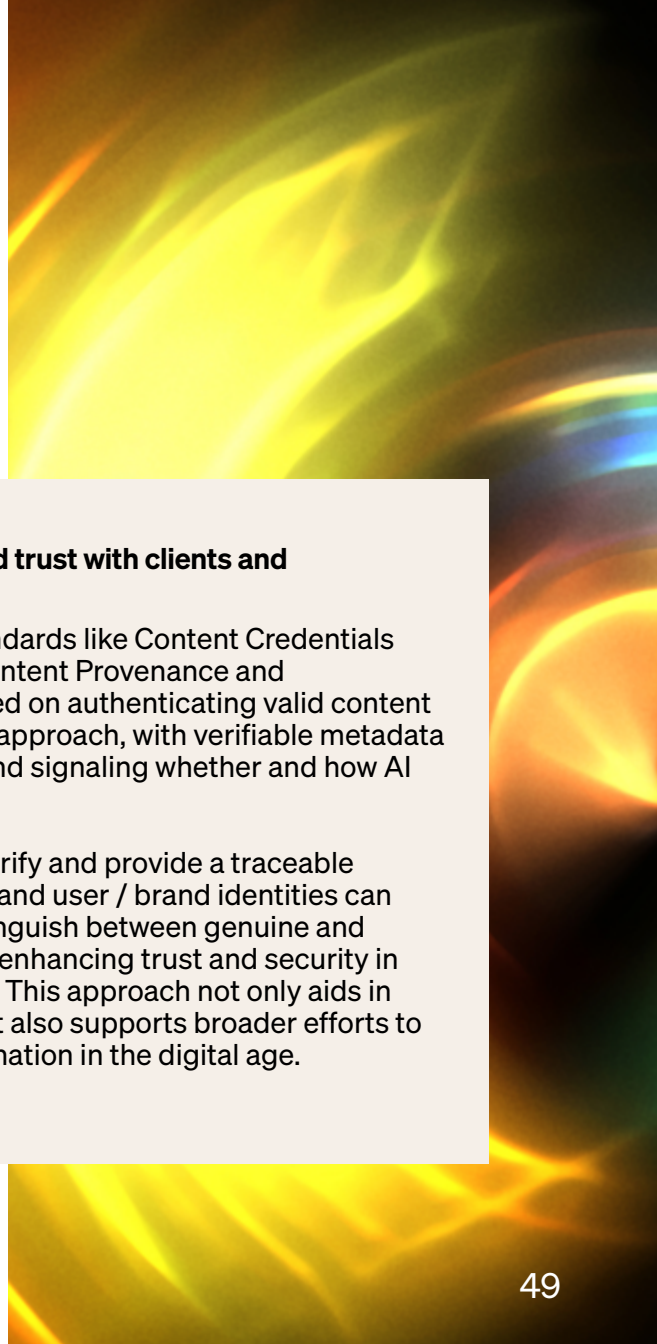
Startups are beginning to solve for these challenges by introducing new capabilities that expand upon cloud, SaaS and data security posture management controls and apply to the unique risks of AI and agentic platforms. Companies are also targeting governance and guardrail solutions for copilot and agent studio applications, which seek to make it easy for anyone to create an agent.

AI and agent platforms from the major cloud and platform providers are also introducing safety and security guardrail capabilities unique to their services but often lack the ability to be leveraged across multi-platform / cloud deployments.

AI gateway and proxy-based solutions continue to evolve from AI runtime protection companies to enforce preventive policies on AI prompts and agent flows. Other cloud security companies and open-source efforts are extending application proxy capabilities to profile and enforce policies on data flows between AI native applications and agentic services.

Identity and access management companies are seeking ways to expand their identity governance, dynamic and just-in-time provisioning / de-provisioning, and agent-based authorization policies to support the unique risks of agent-based models, particularly accounting for the dynamic nature of an agent where its role can evolve over time.

Detecting Deepfakes and Verified Credentials



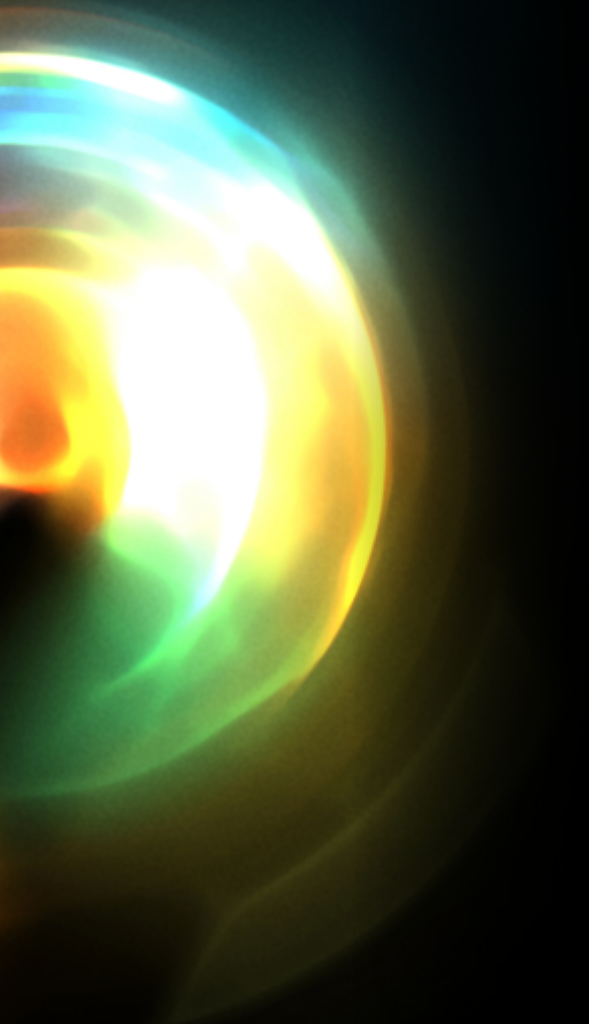
Tools emerging to help enterprises detect malicious AI generated content while enhancing brand trust with clients and customers through enhanced verification

Advancements in AI continue to enable easier creation of highly realistic "deepfake" audio, image, and video content. In response, an array of new capabilities are being introduced to help organizations protect against deepfakes targeting customer and employee communication channels. Deepfakes can be used to commit fraud and social engineering and used in digital and social media channels to manipulate public opinion, impersonate employees, and damage reputations.

Deepfake detection tools use advanced content analysis techniques to identify synthetically generated or manipulated digital media across multiple modes of digital content (e.g., text, audio, image, video). Emerging tools also aim to distinguish between legitimate AI-generated content (e.g., for marketing or accessibility purposes) and content with malicious intent.

Other tools and emerging standards like Content Credentials created by the Coalition for Content Provenance and Authenticity (C2PA) are focused on authenticating valid content through a "nutrition label" like approach, with verifiable metadata (e.g., time and date created) and signaling whether and how AI may have been used.

Techniques to continuously verify and provide a traceable history around digital content and user / brand identities can help users and platforms distinguish between genuine and manipulated content, thereby enhancing trust and security in digital media and interactions. This approach not only aids in the detection of deepfakes but also supports broader efforts to maintain the integrity of information in the digital age.



Market and Industry Perspectives

GenAI market leaders have identified opportunities to partner with deepfake detection vendors to help advance the development of detection models and deter bad actors from using their tools.

Companies that provide digital communication and authentication tools have developed new approaches to identify potential deepfake content as well as introduced new visual components to establish higher trust and verification (e.g., Brand Indicators for Message Identification checkmark in email), especially as these companies also roll out increasingly visual and audio realistic AI avatars. The C2PA consortium includes major players across the content lifecycle – from camera manufacturers to GenAI companies to social media platforms and digital giants. Many of these players have announced tools to generate content credentials in their platforms as well as plans to adopt the content credentials "icon of transparency," a mark that will provide creators, marketers and consumers around the world with the signal of trustworthy digital content, with a goal to make it as universally recognized as the copyright symbol.

Agentic Cybersecurity Operations

Enhances security operations by automating routine tasks, identifying threats, and providing insights that allow analysts to focus on strategic decisions and improving response efficiency in an evolving threat landscape

Like other verticals, AI-assisted copilots and agentic models have emerged as a significant trend in cybersecurity, driven by the increasing complexity and volume of cyber events, operational triage and the proliferation of tools that technology teams and security analysts face.

These AI-powered tools act as intelligent assistants, helping security operations analysts by automating routine tasks, proactively identifying potential threats and providing contextual insights around security incidents. AI security copilots can quickly analyze vast amounts of internal data and external threat intelligence to triage alerts by collecting relevant data and performing appropriate response and subsequent actions, thereby enhancing the efficiency and effectiveness of security operations. This allows human analysts to focus on more strategic decision-making and complex threat investigations.

Additionally, AI security copilots / agents can continuously learn from new data and evolving threats, improving their accuracy and adaptability over time as they ingest more context. As organizations strive to protect their digital assets in an increasingly transforming cyber environment, AI security copilots offer a solution to meet the pace of innovation in threats and streamline security workflows. While early adoption has largely come in the form of simpler use cases like summarization, tools are evolving into semi-autonomous agents that can act continuously on behalf of an analyst. These agents can conduct investigative steps and enable organizations to respond more efficiently and effectively to potential security issues.

Market and Industry Perspectives

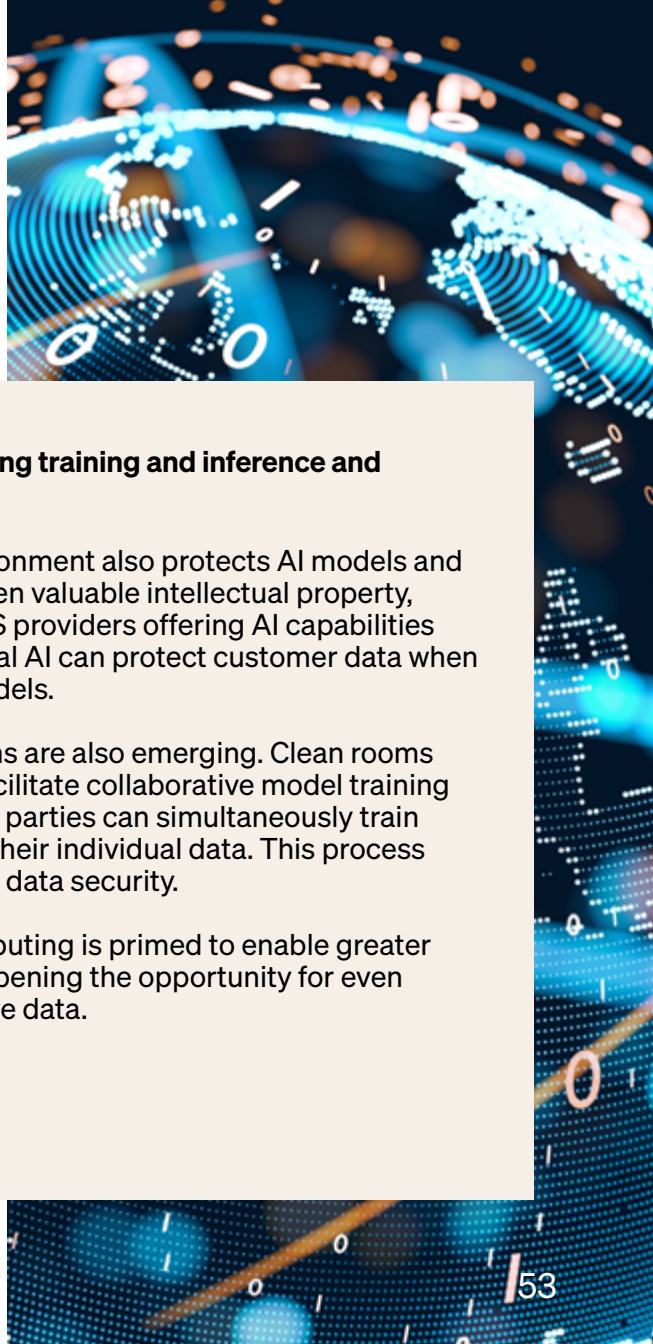
Incumbent security providers have increased their focus in this area – evolving their AI for security capabilities, including exposing their as a service for partners and customers to build on. Startups have also emerged in 2024 to challenge market incumbents with the ability to leverage an organization's breadth of security tools vs. single provider centric.

Startups are building on the foundational capabilities of Security Orchestration and Automated Response (SOAR) tools to drive more of an agentic approach to automate security operations, such as threat hunting and vulnerability management, while leveraging the existing tool and data integrations available in existing SOAR playbooks.

AI platform players have also focused on cybersecurity by developing vertical specific capabilities like an insider threat model that can baseline an organization's accepted behavior for each user and identify abnormal activity to flag risks.

Startups are creating AI agents or "digital employees" that can be curated for specific focuses to autonomously interoperate with human employees and each other, including for other security-related use cases like access management.

Confidential AI



Confidential AI leverages secure enclave technology to protect sensitive data and AI models during training and inference and enabling collaborative AI use cases while maintaining privacy

The key to unlocking the power of GenAI is in the integration of curated enterprise data. However, organizations face security challenges when implementing sensitive data to train models due to the risk of cloud / model providers extracting intellectual property or later downstream in the inferencing process. Confidential AI techniques are emerging with enhanced capability to protect sensitive data end-to-end when training or inferencing models while also maintaining the integrity and privacy of data during processing.

Confidential computing uses specialized hardware to protect data and code from being accessed during processing in memory. This is achieved through the use of hardware-based Trusted Execution Environments (TEEs), which create secure enclaves that isolate data and AI models from the rest of the system, preventing unauthorized access.

By safeguarding data in use, confidential computing allows AI systems to process sensitive information without exposing it.

This secure processing environment also protects AI models and model weights, which are often valuable intellectual property, from theft or tampering. SaaS providers offering AI capabilities are exploring how Confidential AI can protect customer data when being processed by their models.

The concept of AI clean rooms are also emerging. Clean rooms leverage confidential AI to facilitate collaborative model training or fine-tuning where multiple parties can simultaneously train AI models without exposing their individual data. This process preserves confidentiality and data security.

As a result, confidential computing is primed to enable greater trust in AI applications and opening the opportunity for even more innovation with sensitive data.

Market and Industry Perspectives

Cloud service providers continue to build out their own native confidential computing capabilities within their cloud services.

Startups have emerged to provide cross-cloud functionality and innovative solutions for secure GenAI processing, such as data clean rooms that create confidential clean rooms across any cloud provider, application, AI model, tool or code without any development uplift. Others are developing Confidential AI certifications to provide cryptographic proof that a model comes from a trustworthy and unbiased training procedure – an effort supported by OpenAI's Cybersecurity Grant Program.

The confidential computing consortium is a community forum that was created to spread awareness and accelerate the adoption of confidential computing practices through open collaboration. Members of the consortium range from major technology companies and cloud providers to startups.

References

[1] "Volume of Data/Information Created, Captured, Copied, and Consumed Worldwide from 2010 to 2023, with Forecasts from 2024 to 2028." Statista (2024). https://www.statista.com/statistics/871513/worldwide-data-created/

[2] "Predicts 2024: How GenAI Will Reshape Tech Marketing." Gartner (2024). https://www.gartner.com/en/newsroom/press-releases/2024-02-19-gartner-predicts-search-engine-volume-will-drop-25-percent-by-2026-due-to-ai-chatbots-and-other-virtual-agents

[3] "AdTech Spend to Reach $43.5B Globally by 2029." Juniper Research (2024). https://www.juniperresearch.com/press/adtech-spend-to-reach-435-billion-globally-by-2029/

[4] "Online Coaching Market Size, Share, Competitive Landscape and Trend Analysis Report, 2023-2032." Allied Market Research (2024). https://www.alliedmarketresearch.com/online-coaching-market-A06528

[5] "On Device AI Market Size and Forecast." Verified Market Research (2024). https://www.verifiedmarketresearch.com/product/on-device-ai-market/

[6] "Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters." Cornell (2024).

[7] "CoPilot is the UI for AI." (2024). Microsoft Ignite. https://www.thehindubusinessline.com/info-tech/copilot-is-the-ui-for-ai-satya-nadella/article68888483.ece

[8] "Sequence to Sequence Learning with Neural Networks." (2024). https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf

[9] "Scaling Laws – O1 Pro Architecture, Reasoning Training Infrastructure, Orion and Claude 3.5 Opus "Failures." (2024).

[10] "Introducing Llama 3.1: Our most capable models to date." (2024) https://ai.meta.com/blog/meta-llama-3-1/

[11] "Introducing Phi-4: Microsoft's Newest Small Language Model Specializing in Complex Reasoning." (2024).

[12] "Gartner Identifies Top Trends Shaping the Future of Data Science and Machine Learning." (2023).

[13] "Intelligent Agents in AI Really Can work Alone. Here's How." Gartner (2024).

[14] "Quick Answer: Beyond RPA, BPA, and Low Code – The Future is BOAT." Gartner Research (2024).

[15] Cloud Native Computing Foundation. "OpenAI – Launching and Scaling up Experiments, Made Simple." (2018).

[16] "AI Code Tools Market Size Worth $27.17B by 2032." Polaris Market Research (2024).

JPMorganChase

Global Technology Strategy, Innovation and Partnerships