GENERATIVE AI WEEK

November 11 – 12 2025
Sheraton Austin Georgetown, Texas

THE GLOBAL STATE OF GENERATIVE AI IN ENTERPRISE

INDUSTRY REPORT 2026

Visit Site | Register

Contents

2 Table of Contents and Foreword

3 List of Tables

3 List of Figures

4 Enterprise Market & Technology Landscape

10 GenAI in Core Industries

- 10 GenAI in Financial Services
- 11 GenAI in Creative Industries
- 14 GenAI in Retail
- 17 GenAI in Manufacturing
- 20 GenAI in Healthcare
- 24 GenAI in Education
- 25 GenAI in Transportation

28 AI Industry Trends

- 28 AI Infrastructure & Architecture
- 29 Agentic AI
- 35 AI Governance – Risk, Compliance, Responsible AI

37 GenAI in Enterprise: Case Studies

39 GenAI Technology

43 GenAI and Investments

45 GenAI Infrastructure Development

47 Value Creation through GenAI

48 Vendor Landscape

50 Appendix

57 Bibliography

Foreword

When we released our first state of the industry report in 2023, enterprises were going through a wave of experimentation, trying to identify Gen AI transformative use cases across workflows. 2024's report showed an explosion in pilot studies and proof of concepts, with enterprises seeking to define governance policies, infrastructure requirements and value creation.

This year's report shows how fast the landscape is evolving, as enterprises moving from pilot into full-scale production, effectively deploying and scaling Generative AI initiatives to deliver tangible business value.

In this report, we seek to highlight the key forces shaping Generative AI adoption:

- ✓ **Generative AI in Core Industries:** How sector-specific use cases are evolving and what's working
- ✓ **AI Industry Trends:** Where the technology is heading and what's driving the next wave of innovation
- ✓ **AI in the Enterprise:** What best-in-class operationalisation looks like - from architecture to governance
- ✓ **Gen AI Investments:** Where capital is flowing and how it's reshaping the competitive landscape
- ✓ **Gen AI Infrastructure:** How leaders are building scalable, flexible, and cost-effective platforms for AI deployment

As we convene at Generative AI Week, this report is designed to ground our conversations in real data, real strategy, and real outcomes. It's not just a snapshot of where we are today - it's a guide to what's next for enterprise leaders seeking to implement Generative AI across E2E workflows.

Sam Lehmann
Event Director
Generative AI Week

Visit Site Register

Tables

6 **TABLE 1**
Business functions where enterprises are using GenAI by industry (%)

13 **TABLE 2**
GenAI use cases across the creative industries value chain

16 **TABLE 3**
Impact of GenAI on the retail value chain

19 **TABLE 4**
GenAI applications across the manufacturing value chain

20 **TABLE 5**
Categorization of GenAI models in manufacturing

20 **TABLE 5**
Agentic AI vs GenAI vs Traditional AI

30 **TABLE 6**
Agentic AI vs GenAI vs Traditional AI

36 **TABLE 7**
Notable RAI policymaking milestones

40 **TABLE 8**
Significant model and dataset releases

41 **TABLE 9**
Leading GenAI models and specifications

42 **TABLE 10**
Illustrative capabilities of GenAI platforms from select frontier labs

43 **TABLE 11**
Top private equity deals in Gen AI – Q1' 2025

44 **TABLE 12**
Top private equity deals in Gen AI – Q1' 2025

48 **TABLE 13**
Significant AI model and dataset releases, 2024 onwards

49 **TABLE 14**
Leading vendors: GenAI

Figures

7 **FIGURE 1**
GenAI impact on business revenues

8 **FIGURE 2**
GenAI implementation status

8 **FIGURE 3**
Global enterprise GenAI market by segments in US$ billions, 2025-2030

9 **FIGURE 4**
Global enterprise GenAI market by region in %, 2025-2030

9 **FIGURE 5**
Enterprise GenAI: Market share of LLMs in 2024 in %

10 **FIGURE 6**
Gen AI opportunity by function in US$ billion: Banking

15 **FIGURE 7**
Air concept shoe by GenAI

24 **FIGURE 8**
Potential with GenAI in education

27 **FIGURE 9**
GenAI adoption and impact in transportation

29 **FIGURE 10**
GenAI infrastructure funding in 2024

31 **FIGURE 11**
Global Agentic AI market size in US$ billions, 2025-2030

33 **FIGURE 12**
Evolution to multimodal GenAI agents

34 **FIGURE 13**
GenAI vs Agentic AI approach to task completion

34 **FIGURE 14**
Comparative scoring of leading Agentic AI solutions

35 **FIGURE 15**
Investment in responsible AI by company revenue, 2024

42 **FIGURE 16**
Leading GenAI AI chatbots market share and user growth in the U.S., April 2025

43 **FIGURE 17**
GenAI spending vs economic potential of the industry

44 **FIGURE 18**
VC investments in GenAI, 2014-2024, US$ Millions

Visit Site Register

Enterprise Market & Technology Landscape

According to a 2025 U.S.-focused study by McKinsey, as many as 71% of the organizations use GenAI in at least one business function, up from 65% in early 2024. Therefore, it is no surprise that global GenAI spending in the enterprise is estimated to grow from US$4.0 billion in 2025 to US$19.2 billion in 2030 at a CAGR of 36.8%.

While 2024 marked the year that GenAI became a strategic imperative for the enterprise, as companies scaled and learned from their pilots, 2025 has begun to witness efforts to deliver a tangible return on investment (ROI) by deploying GenAI at scale. However, senior decision makers are not expected to demand tangible value and financial results immediately and are operating with a medium to long-term timeline.

After all, despite GenAI's meteoric rise over the last two years, it is still very much in its nascent stages of development and usage, as is evident from the fact that 60% of enterprise GenAI investments today come from innovation budgets. However, with 40% of the spending sourced from more permanent budgets, 58% of which is redirected from existing allocations, businesses are demonstrating a growing commitment to AI transformation. Another reason GenAI will take long to deliver tangible value is that companies need to deploy their limited resources across various competing transformational priorities and a complex and ever-changing regulatory landscape.

Another point to consider is that not all enterprise GenAI investments will be fruitful. In fact, according to estimates by Gartner, at least 30% of GenAI projects will be abandoned after proof of concept by the end of 2025 due to poor data quality, inadequate risk controls, escalating costs and power requirements, or unclear business value. In fact, according to Carly Davenport, VP at Goldman Sachs, the U.S. will have to spend over US$7 billion annually in capital investment to facilitate GenAI-related new power generation alone. Additionally, they will also need to build the supporting infrastructure, such as the transmission wires that transport electricity over long distances and distribution cables that carry electricity to homes, making the overall investment much higher.

While 2024 marked the year that GenAI became a strategic imperative for the enterprise, as companies scaled and learned from their pilots, 2025 has begun to witness efforts to deliver a tangible return on investment (ROI) by deploying GenAI at scale. However, senior decision makers are not expected to demand tangible value and financial results immediately, and are operating with a medium to long-term timeline.

Visit Site Register

Even though investments in foundation models still dominate enterprise GenAI spending, the application layer is now growing faster. The top three areas of GenAI application spending are mentioned below.

1 Code copilots

The intersection of AI and coding has become one of the hottest areas in the technology world regarding VC investments. AI coding tools can automate various routine development tasks such as code generation, testing, and debugging, which has proven to be particularly useful given the huge global demand for software and the shortage of skilled developers. GitHub Copilot's rapid rise to a US$300 million revenue run rate validates this trajectory, while emerging tools like Codeium and Cursor are also growing fast. Beyond general coding assistants, enterprises are also investing in task-specific copilots like Harness' AI DevOps Engineer and QA Assistant for pipeline generation and test automation, along with AI agents like All Hands that can perform more end-to-end software development.

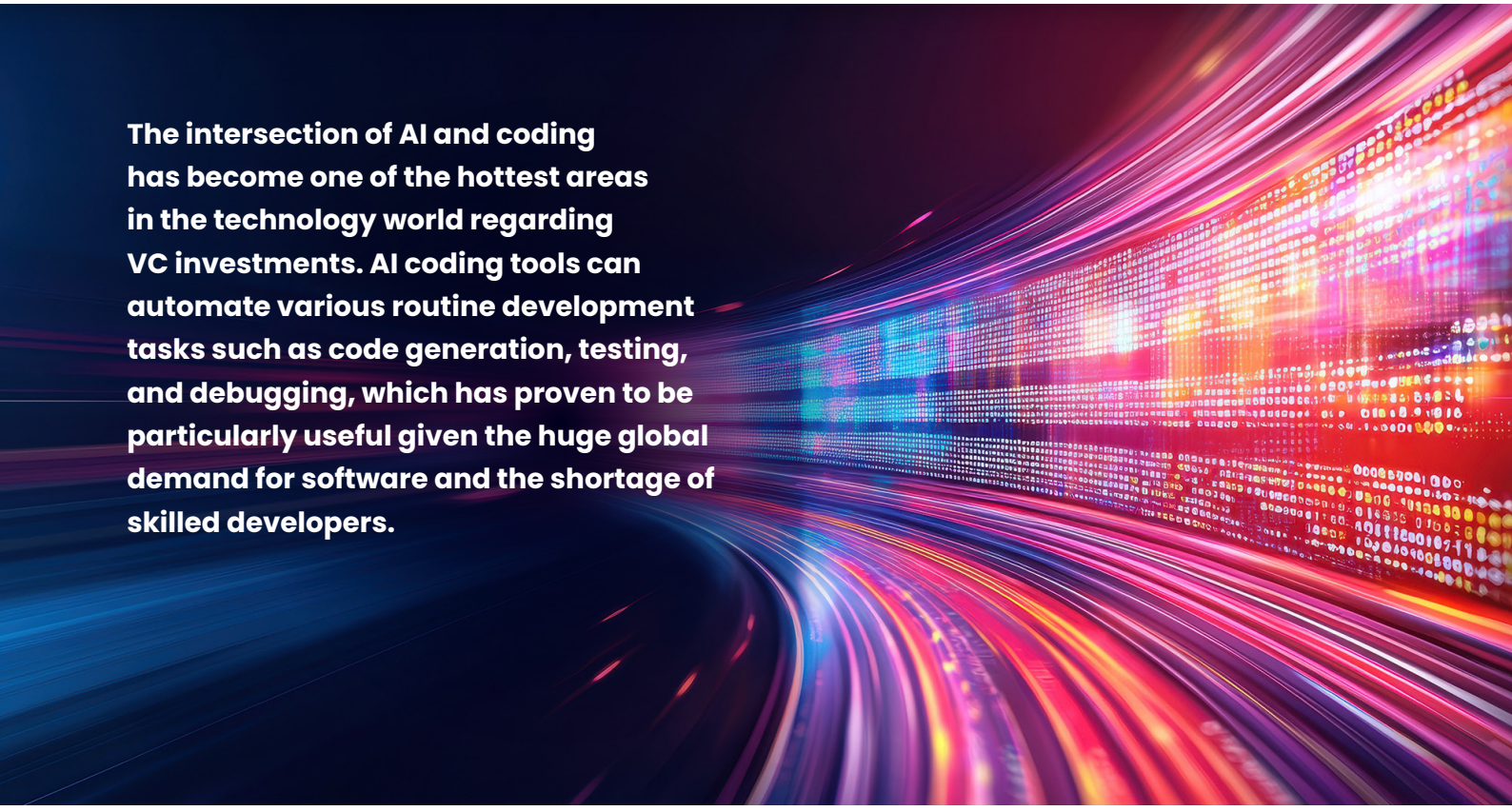
2 Support chatbots

According to the Menlo Ventures study, support chatbots attracted 31% of enterprise adoption in 2024. A good example is global bank ING, which has managed to resolve around 45% of its 85,000 weekly customer interactions in the Netherlands alone through chatbots. Aisera, Decagon, and Sierra are examples of agents that interact directly with end customers, while Observe AI supports contact center agents with real-time guidance during calls.

3 Enterprise search & retrieval and data extraction & transformation

enterprises are investing significantly in these solutions to unlock and harness the knowledge often hidden within data silos across organizations. Good examples are solutions such as Glean and Sana that connect to emails, messengers, and document stores to enable unified semantic search across disparate systems and deliver AI-powered knowledge management.

The intersection of AI and coding has become one of the hottest areas in the technology world regarding VC investments. AI coding tools can automate various routine development tasks such as code generation, testing, and debugging, which has proven to be particularly useful given the huge global demand for software and the shortage of skilled developers.



Visit Site Register

Table 1: Business functions where enterprises are using GenAI by industry (%)

INDUSTRY / BUSINESS FUNCTIONS	Technology	Professional Services	Advanced Industries	Media and Telecom	Consumer Goods and Retail	Financial Services	Healthcare, Pharma, Medical	Energy and Materials	Overall
Marketing and sales	55	49	48	45	46	40	29	33	42
Product and/or service development	39	41	39	26	21	25	22	17	28
IT	31	16	26	22	20	24	30	26	23
Service operations	30	23	24	37	13	26	14	13	22
Knowledge management	26	34	17	26	12	16	24	13	21
Software engineering	36	9	17	30	8	20	13	8	18
Human resources	16	17	13	22	8	11	7	16	13
Risk, legal, and compliance	12	9	6	6	11	21	5	9	11
Strategy and corporate finance	14	14	21	10	7	7	6	5	11
Supply chain/inventory management	10	4	15	3	14	4	2	6	7
Manufacturing	5	3	13	3	8	0	5	7	5
Using gen AI in at least 1 function	88	80	79	79	68	65	63	59	71

Note: Global survey conducted between July 16-31, 2024, with 1,491 participants at all levels of the organization

Source: McKinsey

Market size

The global enterprise GenAI market is estimated to grow from US$4.0 billion in 2025 to US$19.2 billion in 2030 at a CAGR of 36.8%. One of the main reasons for the technology's growing popularity across the enterprise is the public availability of advanced and breakthrough GenAI tools such as ChatGPT, Google's Gemini, and Microsoft Copilot, which have made professionals comfortable with the potential use cases for more industry-centric use.

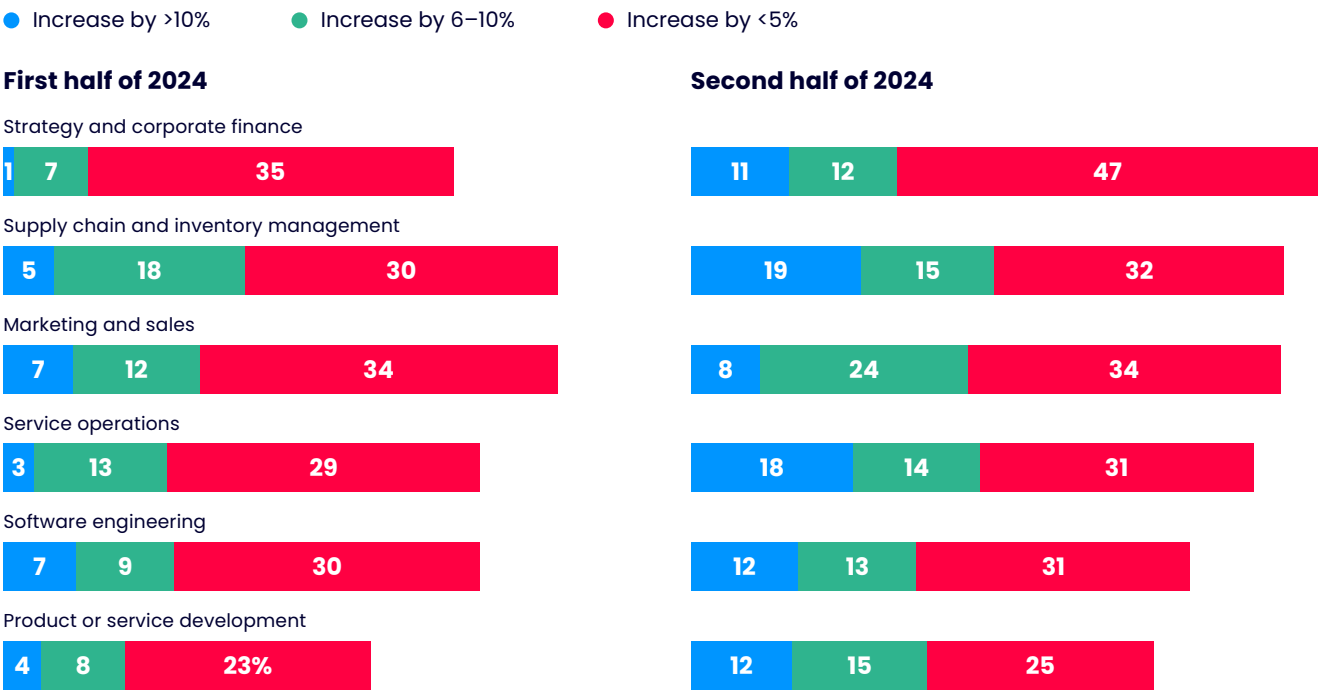
Even though there is consistent adoption across industries, some of them, such as information technology (IT), cybersecurity, operations, marketing, and customer service, are more mature than others. Moreover, enterprises that reported higher ROI for their most scaled initiatives are broadly further along in their GenAI journeys.

Visit Site Register

The software segment is expected to account for the largest 67% share in 2025, with services accounting for the rest. The emergence and expected meteoric rise of AI agents is the primary driver of the software segment over the short to medium term, as the technology gains interest as a breakthrough innovation with the potential to unlock the full potential of GenAI. However, it should be noted that agentic AI cannot be considered a silver bullet, and all the broader challenges currently facing GenAI still apply.

The global enterprise GenAI market is estimated to grow from US$4.0 billion in 2025 to US$19.2 billion in 2030 at a CAGR of 36.8%. One of the main reasons for the technology's growth is the public availability of advanced and breakthrough GenAI tools such as ChatGPT, Google's Gemini, and Microsoft Copilot, which have made professionals comfortable with the potential use cases for more industry-centric use.

Figure 1: GenAI impact on business revenues



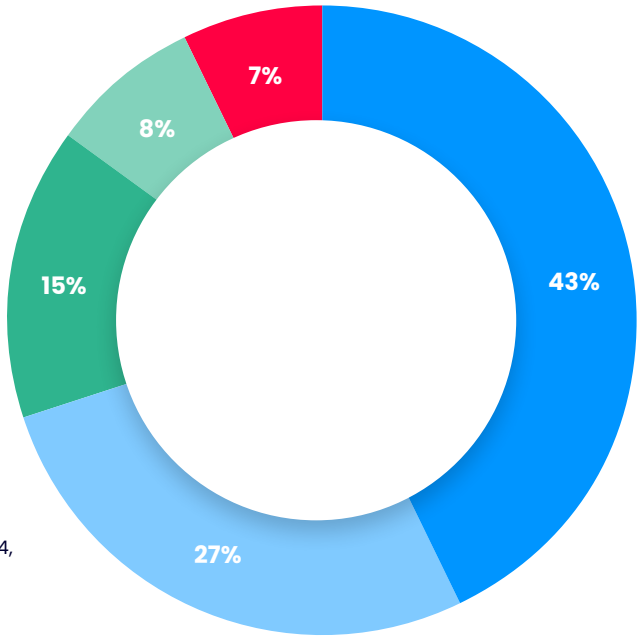
Note: Global survey conducted between Feb 22- Mar 6 (H1 2024) and Jul 16-31 (H2 2024). A question was asked of those who said their organizations regularly use GenAI in a given function.

Source: McKinsey & Company

Visit Site Register

Figure 2: GenAI implementation status

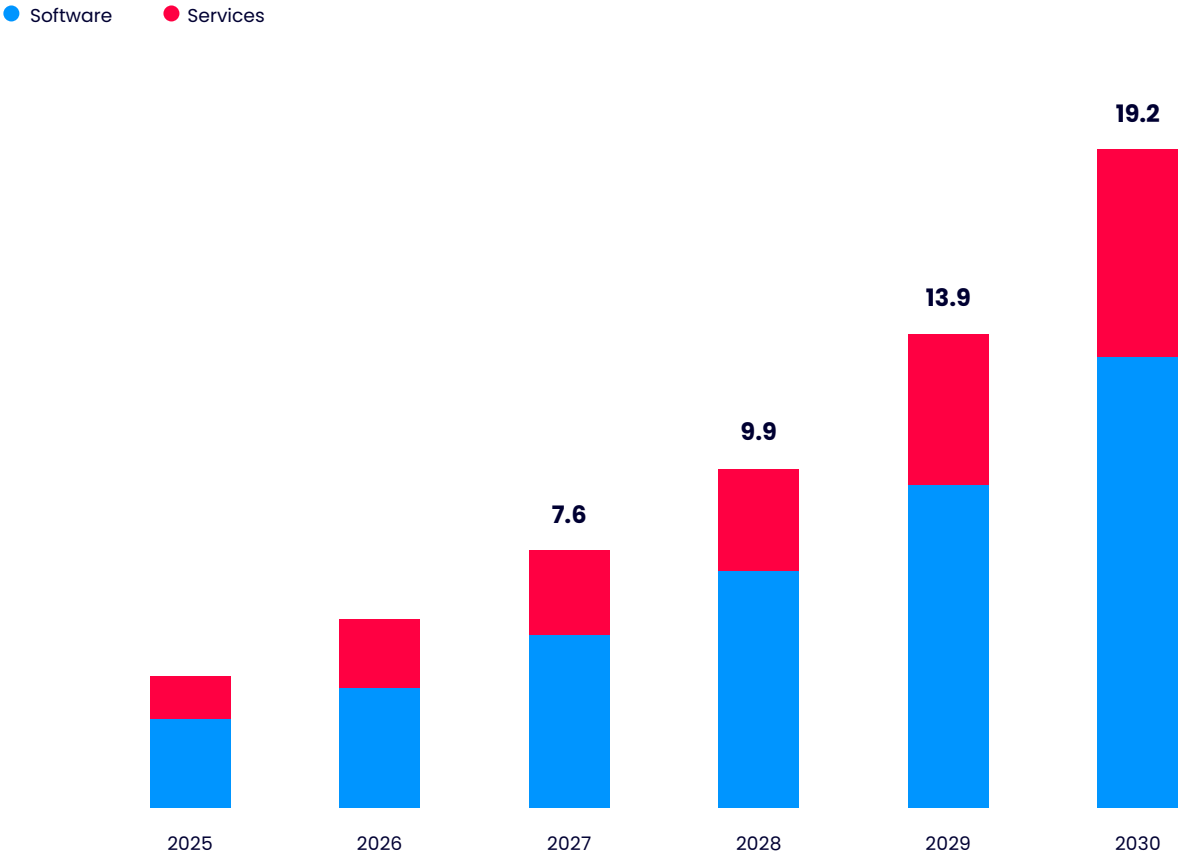
- Live/pilot
- Moving towards full implementation
- Fully implemented,
- Evaluation stage
- Not live (Testing phase)



Note: The ISG GenAI Use Case Study, conducted in August 2024, surveyed 2,000 companies globally

Source: ISG 2024 MarketLens AI Study

Figure 3: Global enterprise GenAI market by segments in US$ billions, 2025-2030

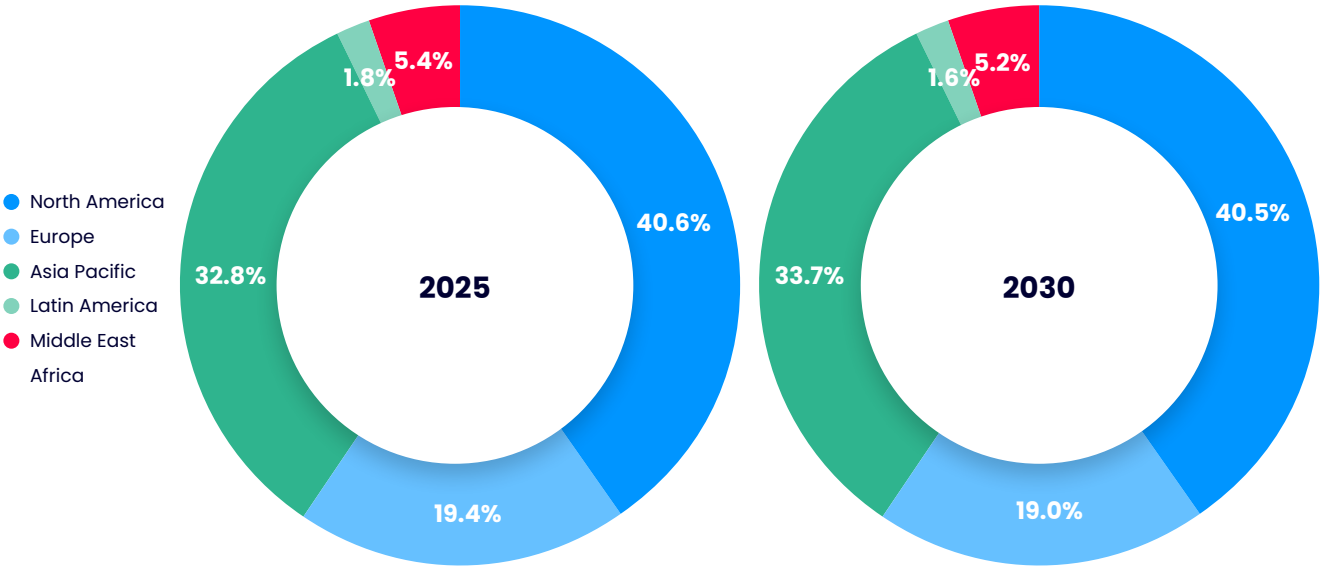


Source: AgileIntel

Visit Site Register

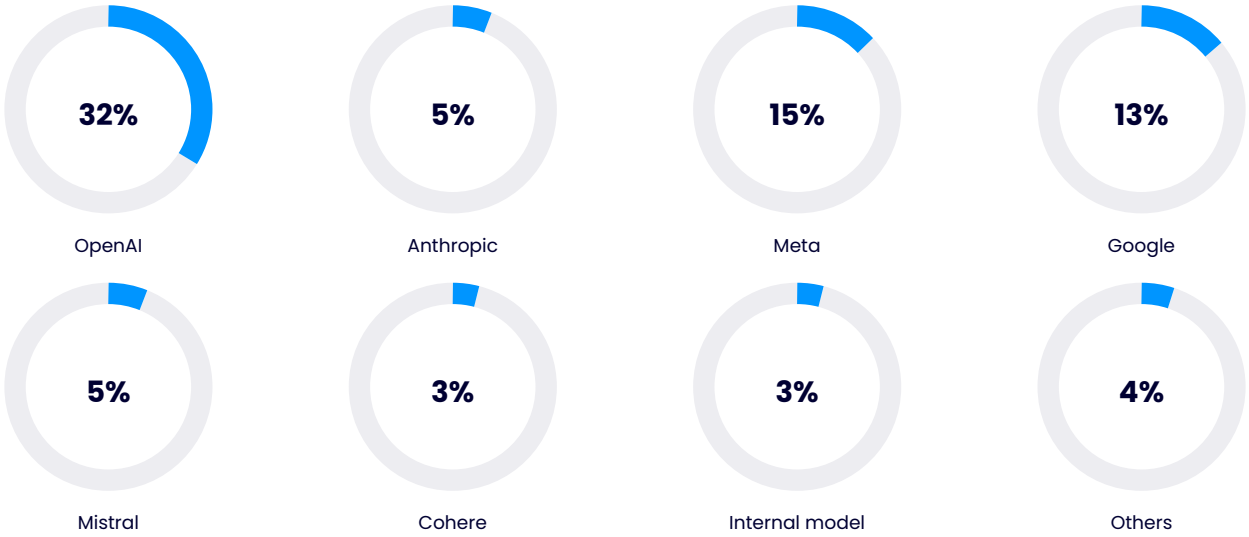
North America is expected to dominate the global market, accounting for approximately 41% of the market share by 2025. Meanwhile, pthe Asia Pacific is projected to be the fastest-growing region from 2025 to 2030, with significant contributions from China, Japan, South Korea, and India, driven by substantial government initiatives. OpenAI dominates the market with a share of 32%, followed by Anthropic (25%), Meta (15%), Google (13%), and Mistral AI (5%).

Figure 4: Global enterprise GenAI market by region in %, 2025-2030



Source: AgileIntel

Figure 5: Enterprise GenAI: Market share of LLMs in 2024 in %



Note: Meta's Llama 3 and Mistral are open-source LLMs
AgileIntel

Visit Site Register

GenAI in Core Industries

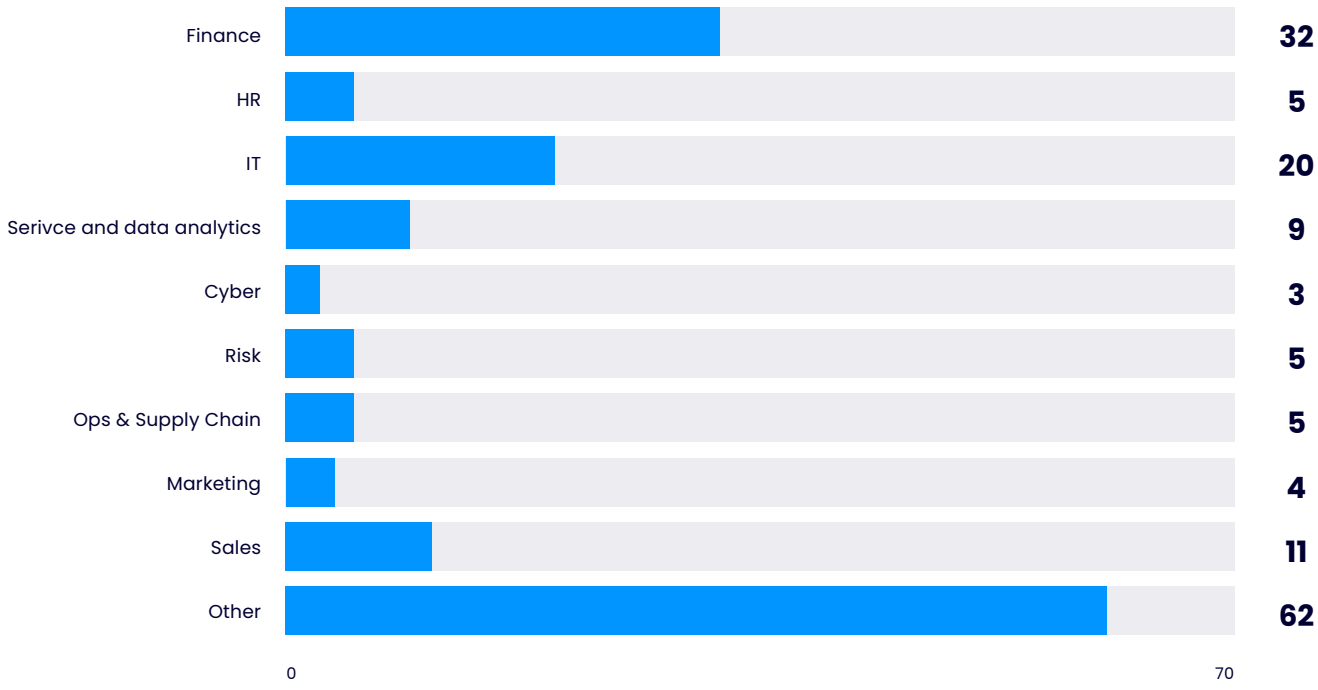
GenAI in Financial Services

The global financial services industry continues to operate in volatile macroeconomic conditions characterized by sudden interest rate hikes and heightened trade tensions. While European and Indian banks are reaping the rewards of rising interest rates, North American banks face a mixed bag of results due to more polarized outcomes. On the other hand, just when Japanese banks had begun to show signs of recovery, U.S. tariff fears resulted in the country's banking index plunging over 20% in the week ending April 4, 2025.

Amidst such uncertainty, only the banks that adapt will thrive while the others risk being left further behind. One key adaptation strategy employed by the global financial services industry is the integration of GenAI, which has become a core enabler of banking transformation. The technology has the potential to not only enable operational transformation and reinvent business models but also save costs, generate higher revenues, and address risk and compliance requirements.

Moreover, as the industry becomes more digitized, GenAI offers opportunities to automate complex processes, deliver customized customer experiences, and strengthen security measures, thereby allowing them to compete with nimbler digital-first competitors. This is especially important in today's volatile macroeconomic environment, which has placed significant pressure on global financial organizations to deliver adequate returns to stakeholders. According to McKinsey estimates GenAI could add between US$200 billion to US$340 billion to the global banking sector annually.

Figure 6: Gen AI opportunity by function in US$ billion: Banking



Source: KPMG, February 2025

Visit Site Register

Even though several financial services companies have already successfully implemented GenAI and started realizing efficiencies, only a few have reported achieving revenue growth from their GenAI investments. Therefore, they now face pressure from shareholders to show immediate ROI on their investments, with a 2025 study by KPMG pegging this at around 70%.

Achieving ROI

Even though several financial services companies have already successfully implemented GenAI in their operations and started realizing efficiencies, only a few have reported achieving revenue growth from their GenAI investments. Therefore, they now face significant pressure from shareholders to show immediate ROI on their investments. However, despite these pressures and the uncertainties created by the rapid evolution of AI technologies, global financial institutions are poised to increase their GenAI budgets over the short to medium term. In fact, according to a 2025 study by BCG, one in three banks plan to spend over US$25 million on ramping up their GenAI capabilities in 2025. However, there is a significant shift in how GenAI is being deployed across the banking industry as banks and other organizations shift from broad experimentation to a strategic enterprise approach that prioritizes targeted applications, especially at the interface between institutions and customers. GenAI-powered tools now support autonomous chat agents that transcend predefined scripts, real-time loan approvals, and automated processing of submitted documentation.

Interestingly, enterprises view the potential value of GenAI in the financial services industry not only as a downstream application but as a tool that complements other machine learning (ML) models and applications. Therefore, they are integrating GenAI not as stand-alone silo models but as a part of a network of models and technologies including robotic process automation (RPA) and autonomous agentic AI solutions. Here, the insights and outputs from one are used to inform the function and direction of another.

This approach has already started to deliver results in the form of 24/7 virtual advisors, providing customized financial guidance, automating routine transactions, and proactively managing customer needs based on real-time data and predictive insights. Additionally, back-office processes, such as fraud detection, compliance monitoring, and risk assessment, are getting streamlined by analyzing vast amounts of data with enhanced speed and precision.

GenAI in Creative Industries

The creative industry has historically relied heavily on human intuition, emotion, and originality, protecting it from disruption by AI and related technologies. However, GenAI has opened up many opportunities, with the sector now ripe for an imminent disruptive impact.

Among GenAI's most promising applications in the creative industries is the use of conversational interfaces to create novel content or translate existing ones. For example, the technology can be used to generate videos or podcasts from articles and blog posts, or to generate variations of a script or storyboard, enabling creators to explore options faster.

Visit Site Register

This is mainly due to the technology's ability to not only automate repetitive tasks such as resizing images, removing backgrounds, and generating design variations, but also provide a new palette for creative individuals to experiment with. This includes generating personalized content, pictures, and videos that are virtually indistinguishable from those made by humans, enhancing operational efficiency, and enabling companies to quickly adapt to evolving trends.

In fact, according to a June 2024 article by BCG, GenAI can now create high-quality content at near-zero marginal cost that allows companies to deliver on the promise of personalization at scale. Another study by the World Economic Forum (WEF) showed that GenAI tools can save creative professionals up to 11 hours per week on tasks such as brainstorming, prototyping, and refining content. These benefits empower more people, including those without deep technical or artistic skills, to join the creators' board.

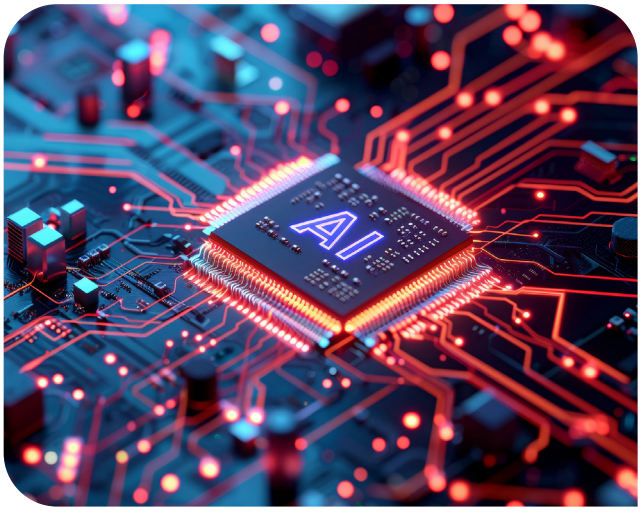
Among GenAI's most promising applications in the creative industries is the use of conversational interfaces to create novel content or translate existing ones. For example, the technology can be used to generate videos or podcasts from articles and blog posts, or to generate variations of a script or storyboard, enabling creators to explore options faster. Text-to-video GenAI models such as OpenAI's Sora have spurred a tectonic shift in the advertising industry, with brands and agencies innovating at a rapid pace to leverage AI-generated video content in their advertising.

LLMs, Generative Adversarial Networks (GANs), Deep Reinforcement Learning (DRL), and Multi-Modal GenAI are the four main GenAI technologies that underpin much of this disruption. LLMs can generate human-quality content, such as poems or scripts, much faster than people can, and also translate languages. GANs go a step further by pitting two neural networks against each other, with one creating new content and the other evaluating its authenticity. They can also produce advanced imagery, ranging from photorealistic landscapes to abstract compositions. DRL employs a reward-based, trial-and-error system through which AI agents can create content that aligns with specific aesthetic preferences or user behavior patterns. Multi-modal AI works by learning patterns and the association between text descriptions and corresponding images, videos, or audio recordings.

The impact of GenAI in creative industries is already visible. A good example is Adobe integrating related capabilities throughout its Creative Cloud suite, with tools like Generative Fill and Text to Image, which are changing how designers work. According to Scott Belsky, Adobe's former Chief Product Officer, the company is aiming to have a language user interface for all of its applications over the short to medium term. Another is graphic design software company Canva's Magic Studio, which has democratized design by making sophisticated AI tools accessible to non-designers.

In 2025, the use cases that are expected to gain traction include:

1. Applications such as U.S.-based Runway AI's text-to-video tool and Cinelytic's analytics and predictive film intelligence platform are designed to plug into production workflows, enabling studios and filmmakers to streamline production tasks and make informed business decisions.



Visit Site Register

2 Tools such as Pencil AI, which is built on OpenAI's GPT LLMs, can create high-quality, low-cost ads quickly, with predictive analytics to test performance. ChatGPT also provides analytical capabilities, allowing industry players to create audience archetypes to test new TV programs.

3 From a post-production perspective, AI applications that provide dubbing and subtitling solutions are expected to witness increased usage. Platforms like Speechify, ElevenLabs, and Panjaya.ai simplify and expedite the process of dubbing audio and creating closed captioning. This enables distribution companies to generate incremental revenues in areas where localization costs have historically been expensive.

4 GenAI-based music generation tools such as MuseNet, Magenta Studio, and Musicfy that can assist in composing music by learning complex musical patterns, predicting the next word or music note in a sequence, and mixing specified instruments. They can also change one type of sound into another, such as from whistling to the violin or from the flute to the saxophone. This capability is beneficial for artists who may not be proficient in playing all the instruments they wish to incorporate, saving both time and costs. This space has advanced rapidly due to unsupervised learning on large datasets and the use of transformers.

5 Image generation tools such as Stable Diffusion, Midjourney, DALL·E, and Ideogram, based on diffusion models (DMs), are fast gaining traction. These open-source tools are developed with the Multimodal Diffusion Transformer (MM-DiT) architecture, which is beneficial for both text and image.

Table 2: GenAI use cases across the creative industries value chain

Commercial Adoption	Pre-production	Production	Post-production	Commercial Strategy	Business Operations
Low	• Concept development for marketing campaigns • Market analysis • Market testing	• Media content for publishing (text & image) • Audio content generation	• AI dubbing for content localization • Content moderation	• Personalized content discovery • Dynamic and personalized advertising	• Customer service chatbots • Content moderation
Medium	• AI-integrated VFX workflows (storyboarding, motion capture) • Movie predictive analysis • Script analysis	• News article generation • Music composition • AI-based virtual reality experience • AI rendering	• Voice cloning • Creating realistic sound effects for film, TV, or games • Video editing process automation	• Conversation summarization tool	• Cybersecurity and protection • Streaming optimization
High	• Game prototyping • Script writing	• AI news broadcaster • Autocompleting code to assist in-game programming • AI game NPCs	• Coloring and grading • Visual effects (VFX) workflow	• Conversation summarization tool	• Budget management

Source: Alix Partners

Visit Site Register

GenAI in Retail

According to an April 2024 study by McKinsey involving many Fortune 500 retail executives, as many as 82% of the respondents said that even though they were still in the piloting and testing phase, the technology had big potential, mainly in augmenting their internal value chains. In 2025, most of the pilots and proofs of concept are expected to assume a larger scale and start delivering ROI, especially in terms of faster, real-time actionable insights in minutes or days, compared with weeks or months earlier. The technology, especially conversational AI, is democratizing data analysis, allowing non-technical users to derive meaningful insights without the need for specialized skills. This not only speeds up internal decision-making but also enables more flexible and innovative use of information across the retail industry.

GenAI is also expected to impact other areas of the retail value chain with automation of routine tasks such as employee scheduling, predictive maintenance, customer inquiries, and onboarding new employees, witnessing maximum disruption.

According to McKinsey estimates, GenAI is poised to unlock between US$400 billion to US$600 billion in economic value for retailers and resolve billions of dollars in inefficiencies. It is also expected to reduce forecasting errors by up to 50%, helping retailers keep up with consumer trends. Therefore, it is no surprise that 45% of global retail marketing leaders plan to invest in GenAI over the next 12-24 months, according to a recent study by Deloitte. Another study by research and advisory company IHL Group found that GenAI is poised to increase retail sales by 51% and gross margins by 20% between 2023-2029, while reducing selling and administrative (S&A) costs by 29%

The main challenge facing the industry in terms of GenAI deployment is that most of the companies are heavily reliant on existing, general-purpose tools. A late 2024 report by PYMNTS Intelligence involving over 500 C-suite employees in the U.S. retail industry found that 61% of them are using just existing baseline models, limiting their ability to achieve more transformative ROI. Comparatively, sectors such as information and manufacturing were ahead in developing proprietary solutions, with 70% and 69% doing so, respectively.

Key use case opportunities

Retail Media: presents a high-margin opportunity for retailers who are increasingly selling their data to brands that can then leverage it to reach consumers closer to the point of purchase. Advances in GenAI are expected to augment retail media by automating ad campaign creation and optimization and helping brands enhance their return on ad spend (RoAS). It is also likely to improve both self-serve and programmatic ad-buying infrastructure due to its ability to process millions of data points within seconds, helping media buyers select the optimal ad format, including the time and location the ad will air. According to a January 2025 study by Coresight Research, the U.S. retail media market is expected to reach US$67.8 million by the end of 2025, ultimately increasing to US$106.4 billion in 2028, at a CAGR of 16%.

Visit Site Register

Product development: GenAI models offer brands various ways in which they can improve their creative processes in terms of new product development. While multimodal models, such as Midjourney, have offered image-generation features for some time, new GenAI applications allow creative professionals to deploy these models without learning how to design prompts and interact with them. Additionally, applications such as Digital Wave Technology's Maestro allow brands to generate creative new product ideas that are more consistent with the brand story and avoid hallucination and toxicity. Below is an image from NIKE, revealing the artistic possibilities of GenAI for new product ideation. Further, GenAI models can facilitate new product development by mining social media posts for major or emerging customer trends or analyzing product reviews, which can then be input into image-generation applications for new product ideas. 2025 is expected to witness the availability of applications that can manage and control multiple GenAI models, which will democratize the use of image-generation technology, making it accessible to a wider base of non-technical users.

Figure 7: Air concept shoe by GenAI

Note: Air concept for tennis player Zheng Qinwen
Source: NIKE

Voice commerce: 2025 is expected to witness the expansion of GenAI-powered voice-based shopping or V-Commerce, allowing users to complete purchases, receive customized recommendations, and manage orders using voice commands. A good example is Apple Intelligence, which has integrated advanced natural language capabilities in Siri to offer highly customized shopping recommendations and even predict future purchases. Another disruptive example is SoundHound AI, which is integrated in vehicles, allowing drivers and passengers to order takeout for pickup directly from the car's infotainment system through voice commands.

Examples of retailers using GenAI

Amazon: has developed an AI virtual assistant called Rufus that is trained on the company's product catalog and customer reviews, among other resources. The application leverages Amazon Web Services (AWS) chips Trainium and Inferentia, and a custom-built LLM that allows it to answer product-related questions and compare products, in a personalized setting.

CarMax: a U.S.-based car retailer, was one of the first in the industry to start using GenAI and has since evolved the technology's usage to create detailed car comparisons with specifications, features, benefits, and customer reviews. Its internal tool, called Rhode, simplifies access to company knowledge for associates, while Skye augments customer experience during vehicle transactions.

The North Face: has deployed IBM's Watson-powered GenAI model to offer a conversational shopping assistant on its online shopping platform. The AI assistant asks customers questions about their preferences, planned activities, and intended usage for outdoor gear, and then delivers product recommendations based on the responses.

Visit Site Register

eBay: The company's GenAI-powered shopping assistant, ShopBot, helps customers navigate through over a billion listings using text, voice, or even by sharing a photo to indicate what they're searching for. The bot can also initiate further conversations to enhance its understanding of the customer's requirements, thereby allowing it to offer tailored suggestions.

Shopify: has launched a GenAI tool called Magic that uses automatic text generation to create automated content such as product descriptions, email subject lines, and headers for an online store. It also allows merchants to modify photo backgrounds to suit their branding, without needing expertise in complex software like Photoshop.

Table 3: Impact of GenAI on the retail value chain

Retail value chain	Before generative AI	After generative AI
Procurement	• Manual handling of supplier negotiations (including end-to-end contract creation), often resulting in overlooked details • Tedious supplier assessments based on limited data, leading to suboptimal choices	• GenAI chatbots handle initial rounds of supplier negotiations • GenAI-powered briefs and summaries of supplier terms assist procurement associates in closing deals.
Distribution	• Individuals handling communication with third-party logistics providers • Delayed response to distribution disruptions due to the complexity of supply chain operations	• Initial communication and email messages to third-party logistics handled by Gen AI chatbots • Returns management process, along with a response to distribution disruption, supported by Ge AI
In-store operations	• Information searches, such as price, in-store location, and stock level handled manually by associates, leading to delayed customer service	• People use GenAI-powered assistants for instant voice access to information
E-commerce	• Hundreds of hours spent on the generation of e-commerce content • Manual rule-based website personalization, consuming employees' resources	• Automated generation of e-commerce content (eg, product profiles, descriptions) within a few minutes • E-commerce customer experience personalized spontaneously by automated front-end development techniques
Marketing	• One-size-fits-all marketing approach due to limited customer insights derived from structured data • Creation of marketing materials through a lengthy, iterative process	• Unlimited insights extracted from different unstructured sources (eg, product reviews) • Fully personalized marketing materials generated with increased efficiency for every customer
Back office	• Time-consuming administrative processes, such as HR and payroll, are prone to errors and inefficiencies	• The next-generation "white collar" lean—transferring administrative processes of support functions to GenAI-powered chatbots and interfaces, such as development copilots, HR/financial copilots.

Source: McKinsey & Company

According to McKinsey estimates, GenAI is poised to unlock between US$400 billion to US$600 billion in economic value for retailers and resolve billions of dollars in inefficiencies. It is also expected to reduce forecasting errors by up to 50%, helping retailers keep up with consumer trends.

Visit Site Register

GenAI in Manufacturing

Over the past couple of years, GenAI has transitioned from a futuristic concept to a tangible transformative force, shaping the manufacturing landscape in previously unimaginable ways. The technology now allows manufacturers to automate and enhance factory activities by supporting functions such as programming and machine maintenance (including predictive maintenance), autonomous factory management, intelligent quality control, smart supplier contract management, and product R&D. A good example is German manufacturer Bosch which is using GenAI to create a comprehensive dataset of synthetic product defect images to train its AI system for optimal quality control.

According to a 2025 study by Deloitte titled Future of Manufacturing involving 600 manufacturers globally, as many as 87% reported that they had initiated a GenAI pilot already, while 24% indicated that they had adopted GenAI use cases in at least one of their facilities. Further, 50% of the respondents said that GenAI solutions ranked among the top-priority solutions for their organizations over the next 24 months, higher than other highly sought-after technologies such as digital twins, the omniverse, and the metaverse.

Another 2025 study by technology company NTT DATA involving over 500 manufacturing leaders and decision makers in 34 countries, a staggering 95% of them said that GenAI was already directly improving efficiency and bottom-line performance. Interestingly, 94% expect the integration of IoT data into GenAI models to significantly improve the accuracy and relevance of AI-generated outputs. Manufacturers are also using GenAI to personalize operations by training LLMs on smaller datasets from their internal industrial IoT (IIoT) devices, instead of the conventional large datasets. This enables seamless information exchange between legacy machines and equipment not using open-source AI tools and GenAI systems. Additionally, these smaller language models can be fine-tuned to operate closer to the edge (end-user), where latency and security are important to IIoT solutions.

GenAI-powered robots are used in manufacturing through the use of natural language prompts that are inherent in the technology. This allows machine operators who are not necessarily trained in robotics or software code to communicate with the machines using natural language.

Key benefits of using GenAI in the manufacturing industry:

Faster product rollouts: GenAI tools allow manufacturers to bring products to market faster by automating and optimizing different stages of product development, including innovation, design, prototyping, and testing. Once a GenAI model has been trained on a product's bill of materials, raw material usage, process parameters, internal research data, and other data (such as product patents or previous product trials), it can identify the ingredients that may be best suited for a new product, predict the product's benefits, and recommend formula recipes. A good example is AstraZeneca, which is using GenAI to automate and quicken the drug development process. The technology has already helped the company reduce development lead times by 50% and the use of active pharmaceutical ingredients in experiments by 75%. Another leading pharmaceutical company is using GenAI to analyze production line bottlenecks and optimize its tablet packaging process. It has resulted in boosting production efficiency by 20% while minimizing material waste.

Visit Site Register

Digital twins: manufacturers are using GenAI algorithms to create accurate digital representations of their products, production lines, or entire factories. Real-time data is taken from sensors and other sources to improve design, test new processes, and create new products without disrupting the production process. A good example is Indian specialty chemical manufacturer Jubilant Ingrevia, which has reduced process variability by 63% by deploying digital twins to model, forecast, and manage operations in real time.

New product development: GenAI tools analyze vast information on prevailing market trends, consumer preferences, and past performance of products, to give manufacturers a clearer picture of new and advanced product designs, and even discover new business models. In terms of novel designs, GenAI enables manufacturers to visualize concepts in high fidelity much earlier in the design process and get precise feedback from customers, thereby allowing them to create a previously unimagined product. McKinsey estimates that GenAI could unlock US$60 billion annually in productivity in product research and design alone. Additionally, through synthetic data augmentation, GenAI can enable accurate simulations, aligning product development with stringent requirements and customer preferences, thereby saving time and resources.

Predictive maintenance: Previously, manufacturers prevented breakdowns by performing scheduled maintenance according to fixed cycles or periods. With the advent of AI and ML, they began using data from various sensors to identify patterns, predict breakdowns, and then proactively conduct maintenance. GenAI has further improved this process by automatically creating text or images that provide detailed instructions, including lists of required spare parts. This system enables maintenance personnel to spend more time on the actual tasks instead of preparing instructions, enhancing productivity, and reducing costs. Owing to its comprehensive nature, it also allows inexperienced technicians to repair or maintain equipment more effectively.

Customization at scale: GenAI allows for the efficient customization of products at scale, catering to the unique preferences of individual customers without compromising efficiency. By using this technology, manufacturers can readily adjust designs and processes to meet customer demands in real time. AI-driven insights allow for the integration of unique product features on a large scale, without a significant increase in costs. As the technology evolves, the potential for personalized products will expand, optimizing design, performance, and functionality based on specific customer preferences. Industries already in the advanced stages of integrating GenAI in their manufacturing processes include consumer electronics, automotive, and fashion.

Visit Site Register

Table 4: GenAI applications across the manufacturing value chain

Planning – product development	
●	Create product concepts and engineering drawings to reduce R&D and prototyping times.
●	Discover new materials by testing to define their fit and function as alternative raw materials.
●	Predict product-market fit with qualitative consumer/market data.
Planning – production planning and procurement	
●	Develop production plans based on available materials, equipment, and resources.
●	Discover new supplier profiles across sources.
●	Pre-screen, summarize, and extract clauses of interest across contracts and assess risks.
●	Automatically action ERP exception messages to achieve optimal inventory levels
Production – performance, maintenance, and health and safety	
●	Create employee training videos and maintenance troubleshooting role-plays.
●	Write standard operating procedures and policies, and translate documents into other languages.
●	Identify hazardous working conditions and notify key stakeholders about required measures.
●	Automate root cause analysis to identify causes of nonconformances without manual data analysis
●	Predict exact machine failure modes and automatically develop intervention plans.
●	Adjust production orders in real time based on LoT, RFID, and order-tracking data.
●	Receive performance updates, priorities, and advice from AI chatbots.
Supply chain – warehousing and logistics	
●	Automate route design, using routing algorithms to reduce cost and lead time
●	Provide updates on shipments and delivery times via chatbot interface.
●	Generate and verify the required documents for transportation.
●	Provide an interactive virtual assistant for drivers to augment typical services provided (eg, route navigation)
●	Improve yard management processes based on sensor and camera data.
●	Optimize warehouse design to streamline order-picking routes.
●	Automate materials reordering to minimize stockouts and inventory levels.

Source: McKinsey & Company

● Content Generation ● Insight Generation ● Interaction

GenAI has transitioned from a futuristic concept to a tangible transformative force, shaping the manufacturing landscape in previously unimaginable ways. According to a 2025 study by Deloitte titled Future of Manufacturing involving 600 manufacturers globally, as many as 87% reported that they had initiated a GenAI pilot already, while 24% indicated that they had adopted GenAI use cases in at least one of their facilities.

Visit Site Register

Table 5: Categorization of GenAI models in manufacturing

Generative AI Model	Application in Manufacturing
Generative Adversarial Networks (GANs)	Creation of digital twins, virtual replicas of physical assets or processes based on real-time sensor data for product design and optimizing manufacturing processes. **Pros:** High-quality realistic images and data augmentation, processing sequential data in parallel **Cons:** Difficult to train, limited and repetitive outputs, difficult to find the right balance between the generator and discriminator.
Variational Autoencoders (VAEs)	Prediction of equipment failures through machine learning algorithms trained on machine data. **Pros:** Generating data similar to training data, overcoming limitations of traditional image processing methods **Cons:** Less flexible than GAN, unable to tackle sequential data, difficult to control the quality
Transformer-Based Models	Simulation of production scenarios, prediction of demand, defect detection, and material fracture mechanics. **Pros:** Processing sequential data in parallel, handling multiple data types, Powerful for diverse multimodal tasks **Cons:** Requiring large amounts of high-quality training data, slow and computationally intensive process

Source: ScienceDirect

GenAI in Healthcare

GenAI is rapidly transforming the healthcare industry. As many as 85% of respondents in McKinsey's Q4 2024 survey of U.S.-based payers, health systems, and healthcare services and technology (HST) groups are already implementing the technology across the enterprise. Another study by Deloitte conducted towards the end of 2024 demonstrated similar results, with as many as 75% of the companies in the healthcare space already experimenting with GenAI.

The widespread acceptance of the technology is driving the rapid evolution of the industry in the face of many years of plateaued growth in the areas of telemedicine and digital therapeutics.

Visit Site Register

As GenAI matures, it is resulting in the creation of novel solutions, especially to address gaps in areas pertaining to chronic conditions such as heart failure, diabetes, and mental health. Every aspect of healthcare, ranging from personalized care to automated workflows, is expected to be disrupted at various levels by GenAI in 2025. The industry is expected to witness a greater adoption of multimodal GenAI models that can analyze and generate text, images, genomics data, and even real-time patient vitals simultaneously, compared to the single modality models that were dominant in 2024.

According to a 2025 study published in the Journal of Medical Internet Research (JMIR), patients receiving care powered by GenAI attended 42% more therapy sessions and achieved a 25% higher recovery rate compared to other treatments. These findings showcase GenAI's ability to improve clinical outcomes and the overall standard of care.

If 2023 was about GenAI experimentation and 2024 was about point solutions, 2025 is expected to be about value delivery through end-to-end transformation. Instead of isolated GenAI tools fulfilling specific tasks like physician note-taking or scheduling, the industry is expected to witness the proliferation of integrated systems that automate entire workflows ranging from patient intakes to treatment plans. These intelligent agents will coordinate across departments, learning from each interaction to improve efficiency and outcomes. For example, in the pharma industry, key processes that will be transformed with GenAI include clinical trials, regulatory submissions, medical legal regulatory review, and omnichannel engagement.

Overall, with the global healthcare industry grappling with challenges such as labor shortages, clinician burnout, declining profitability, and worsening health outcomes, GenAI offers a transformative enterprise approach to address these problems. The technology is primed to address the healthcare industry's greatest pain points by democratizing knowledge, increasing interoperability, expediting drug discovery, and enabling hyper-personalization of the care experience. Among the various areas that could witness significant disruption over the medium to long term are patient and member experience, daily administrative tasks, and clinician and clinical productivity.

Despite the enthusiasm around the technology's large-scale integration in the healthcare industry, an early 2025 BCG report predicts that over 33% of ongoing GenAI programs will fail to deliver value in 2025. These failures are ultimately likely to pave the way for more sustainable and impactful transformations, driving a sharper focus on integrating GenAI into existing health care workflows. GenAI applications in the short to long term:

If 2023 was about GenAI experimentation and 2024 was about point solutions, 2025 is expected to be about value delivery through end-to-end transformation. Instead of isolated GenAI tools fulfilling specific tasks like physician note-taking or scheduling, the industry is expected to witness the proliferation of integrated systems that automate entire workflows.

Visit Site Register

Short term: the immediate applications of the technology are focused on the use of natural language processing (NLP) in healthcare settings, enabling functions such as ambient scribing to lessen the burden on manual clinical documentation. Other use cases are automated consumer messaging, clinical message autoreply, and document auto-generation.

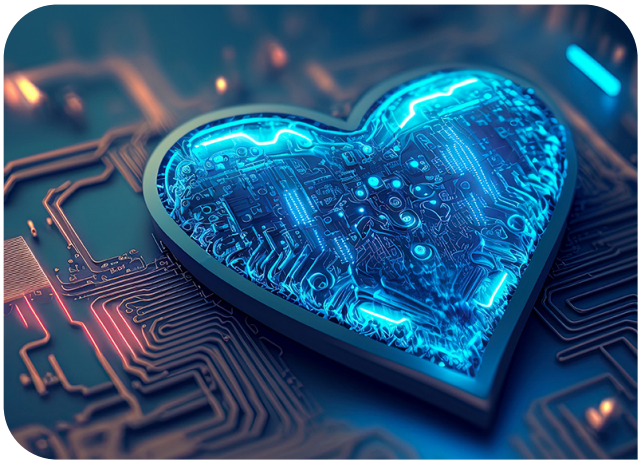
Medium-term: over the medium term, the technology is expected to facilitate the integration of data science within various hospital functions to extract relevant insights from data sources such as medical records, research studies, and patient-generated data, resulting in more personalized and effective treatment plans.

Long term: over the long term, GenAI is expected to replace the doctor in diagnosis and prognosis. In fact, in some cases, AI and ML have already reached a 98.4% accuracy in certain cancer diagnoses, paving the way for quick disruption in the future.

Key use cases of GenAI in the healthcare industry

Drug and treatment discovery: In 2024, AI-powered drug discovery made many gains. However, in 2025, GenAI is expected to bring about rapid disruption by facilitating the design of novel drug compounds in real time. Pharmaceutical and biotech companies are increasingly using customized language models to augment their understanding of disease biology and accelerate processes to identify promising compounds. Both commercial and open GenAI models can already analyze vast biomedical data sets to suggest novel molecular structures, predict drug interactions, and design custom compounds tailored to a specific target or disease. Many of these compounds have been hard to discover through traditional methods. This mitigates formidable costs and time constraints. When used together with causal modeling approaches, the models allow companies to identify clues previously undiscovered or underrepresented in clinical data, unveiling previously ignored therapeutic opportunities. According to a recent BCG report, in 2025, this trend will further shorten discovery cycles and reveal more promising candidates to test in clinical settings.

Drug development: In addition to discovery, GenAI can enhance the drug development process across all areas, such as preclinical testing, clinical study design, and regulatory submissions. For preclinical testing, GenAI models can estimate the toxicity of a drug compound by analyzing chemical structures and potential risks associated with candidate therapies. They can also forecast pharmacokinetic properties and ADME (absorption, distribution, metabolism, and excretion) properties of drug candidates, which can predict the effect of a drug on its target and related safety levels. In terms of clinical study design, GenAI increases the chances of success by identifying the most relevant patient populations, endpoints, and dosing regimens. And finally, the technology can expedite the regulatory submissions process by automating compliance checks and proactively performing checks against guidelines. Additionally, GenAI tools are poised to transform manufacturing operations by processing engineers to optimize workflows to manufacture therapeutic products, including monoclonal antibodies and cell therapies. A good example is Exscientia, a drug design and development company that uses Google Cloud GenAI capabilities to enable faster drug discovery through a Design-Make-Test-Learn (DMTL) cycle.



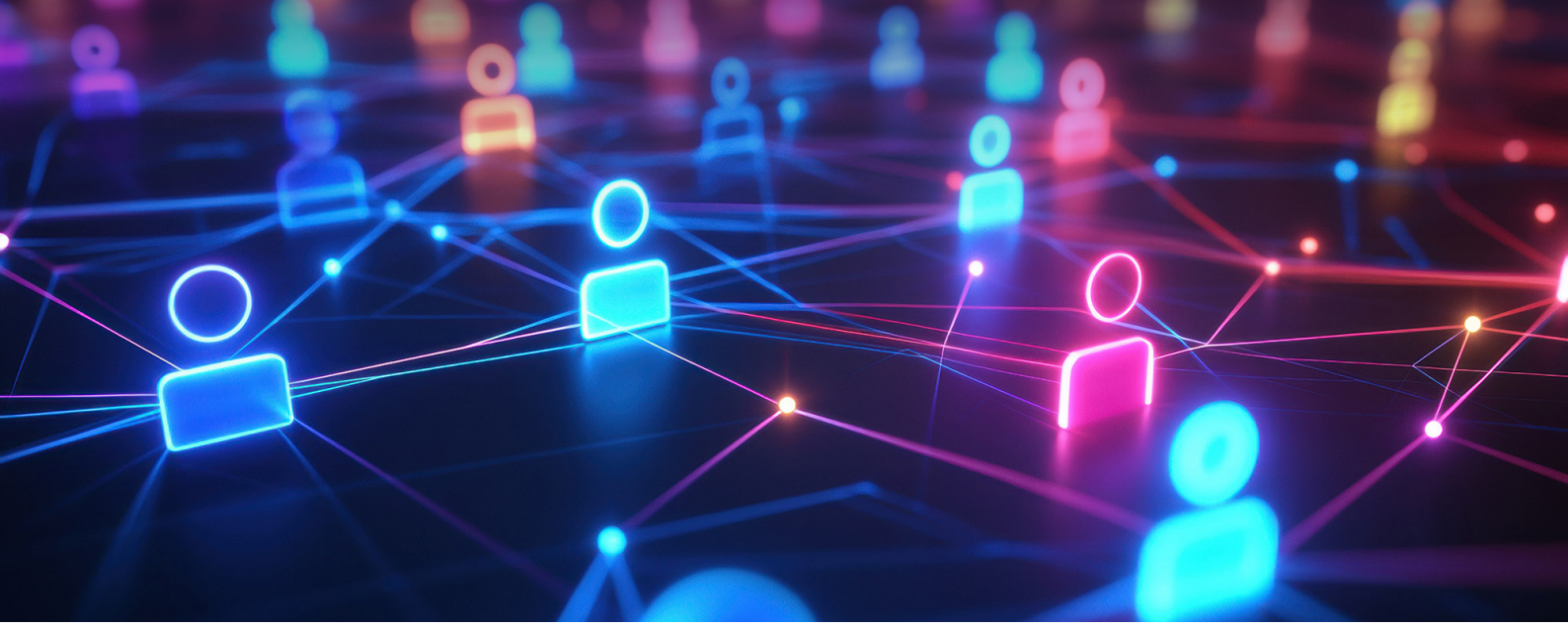
Visit Site Register

Quality control: GenAI is now playing a bigger role in quality control for pharmaceuticals and medical device products by standardizing the manufacturing processes and improving the detection and mitigation of related deviations. This approach to quality control will allow manufacturers to adjust processes, reduce waste, improve yield, and increase product quality. For example, an issue-resolution GenAI model trained with historical data can enable organizations to identify the effects of minor changes on product outcomes and thereby reimagine processes without extensive and often manual trial-and-error tests.

Chatbots: According to an August 2024 study by health policy research company KFF, just over 16% of the adult respondents said that they use AI chatbots at least once a month to find health information or advice, rising to 25% for adults under 30 years old. As GenAI-powered chatbots evolve and improve, these consumer behavior patterns will most likely force established online health information gateways to offer their bespoke AI tools or risk losing web traffic. This will enable health providers to start realizing significant operational efficiencies and competitive advantage by using these trained chatbots to attract patients and routing them to the most appropriate sources of care, while reducing the burden on humans who staff the 24/7 triaging capabilities that they offer.

Personalized care: Recent advancements in agentic AI are driving personalized treatment plans by analyzing large datasets containing patient-specific data such as genetic profiles, medical records, and live health data. This helps healthcare professionals to recommend targeted therapies such as chemotherapy, radiation, or surgery, depending on each patient's unique profile. According to a study published in the ScienceDirect journal in March 2025, GenAI-powered personalized treatments improved cancer patient survival rates by 20% compared to standard care and extended progression-free periods by 15%.

According to an August 2024 study by health policy research company KFF, just over 16% of the adult respondents said that they use AI chatbots at least once a month to find health information or advice, rising to 25% for adults under 30 years old.



Visit Site Register

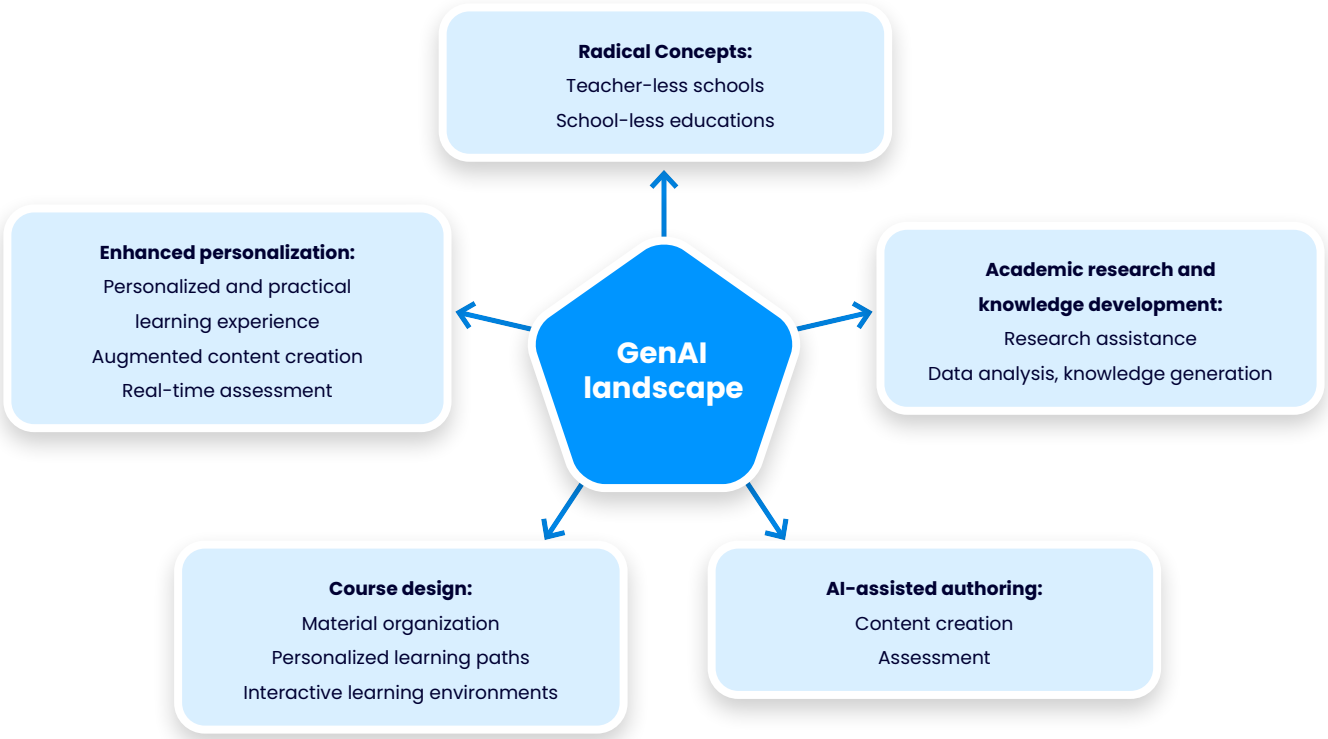
GenAI in Education

GenAI is transforming the education industry by disrupting traditional teaching methods, improving student support systems, and reorganizing the overall ecosystem. A late 2024 report by American education technology company Cengage Group found that as many as 49% of higher education instructors in the U.S. are already using GenAI, up from 44% in 2024 and just 24% in 2023.

The technology's core capabilities, which include creating and disseminating information, make it ideal for disrupting the education space. Over the last year or so, LLMs have showcased their ability to answer questions on a range of subjects, write cogently, and even create images. Moreover, ChatGPT and similar models have proven their expertise in cracking tough examinations in fields such as law, medicine, history, and even operations management.

Education technology companies and students have already started using GenAI tools such as ChatGPT, TutorAI, and the Poe app that stimulate creativity by assisting in brainstorming sessions and generating fresh ideas. Additionally, GenAI models have started assisting teachers in creating homework and assignments, explaining complex concepts to students simply, designing courses, and creating gamified learning experiences and personalized learning plans for each student. A good example is Speechify, which offers text-to-speech or speech-to-text generation capabilities that are particularly useful for students with learning disabilities such as dyslexia or ADHD. Another is Kahoot!, which uses GenAI to design games that align with curriculum goals, making learning both fun and effective.

Figure 8: Potential with GenAI in education



Source: Attri.ai

Visit Site Register

Key use cases:

Personalized adaptive learning experience: GenAI-powered intelligent learning platforms analyze various types of student data, such as historical performance, skills, and teacher feedback, to offer personalized and adaptive learning experiences. By analyzing large datasets, educators can identify knowledge gaps and provide recommendations and guidance. GenAI tool can create exercises, quizzes, and practice questions customized to each student's learning needs. Additionally, through the use of GenAI tools, teachers can offer real-time assistance, progress monitoring, and adjust teaching strategies to optimize learning.

Curriculum creation and design: educators are using GenAI to create course and teaching materials such as syllabi, quizzes, exercises, and concept summaries. This not only saves time through the automated generation of content, but also improves resource variety. GenAI also enables the rapid creation of e-learning capsules, micro-videos, and interactive multimedia elements, personalized to the needs of different courses. Moreover, platforms providing courses for language learning can use GenAI to correct grammar and create related exercises and questions.

Virtual experiments: GenAI, together with virtual reality, is being used to make simulations and virtual environments to enable students to conduct experiments, observe outcomes, and test predictions in real time.

Automated assessment and grading: GenAI tools such as ChatGPT and the Intelligent Essay Assessor can reliably review and grade written coursework and provide feedback, thereby ensuring speed, consistency, and objectivity. Various studies have demonstrated that these tools can reduce grading time and deliver accurate and consistent results.

GenAI in Transportation

GenAI is expected to be one of the primary growth drivers of the global transportation and logistics industry, which is expected to increase at a significant CAGR of 44% between 2023 and 2032, to a value of almost US$19 billion. As the industry grapples with shifting trade flows, margin pressures, rising need for sustainable practices, and increasing demands from shippers and regulators, GenAI offers significant transformative potential.

According to a February 2024 global study by IDC, over 50% of transportation companies were already using GenAI with knowledge management, marketing (better shipper/lead conversion, increased dynamic pricing/quoting), and product/service creation, accounting for over 70% of use cases. Another study conducted by Deloitte in July 2024 of over 200 executives found that almost all of them (99%) expect the technology to transform their industry, but over two-thirds (71%) expect this transformation to take more than three years. The transportation use cases witnessing the highest adoption and impact are asset management, route optimization, and warehouse operations. Interestingly, over half of the companies surveyed were found to be running GenAI initiatives within each of these use cases, with around 80% of adopters reporting extremely high or high economic value in each use case.

Visit Site Register

One long-standing challenge for trucking and freight forwarding companies has been the planning of efficient transportation routes. GenAI models present an opportunity to solve this problem by analyzing data related to tariffs, trade agreements, traffic patterns, public transportation, and other variables to generate optimal routes and minimize costs.

Major transportation companies have already started making investments in use cases related to contract consulting, transportation execution, strategy, and customer experience. With the technology still very much in its nascent stages, it promises to disrupt every link in the transportation and logistics value chain over the medium to long term.

Key use cases:

Route optimization: One long-standing challenge for trucking and freight forwarding companies has been the planning of efficient transportation routes. GenAI models present an opportunity to solve this problem by analyzing data related to tariffs, trade agreements, traffic patterns, public transportation, and other variables to generate optimal routes and minimize costs. One of the main benefits of GenAI in the industry lies in the dynamic optimization of transportation networks in real-time through the analysis of traffic data, pedestrian crossings, and emergency vehicle locations. International shipping companies such as DHL are integrating GenAI models into their processes and analyzing data pertaining to shipment volumes, vessel capacity, and port capacities to determine cost-effective and environmentally friendly delivery methods.

Dynamic inventory management: With efficient warehousing of goods key to a successful transportation enterprise, dynamic inventory management assumes critical importance. This is especially true if the volume of goods being handled is large. Therefore, inventory control managers are increasingly using GenAI to analyze data gathered from lead times, demand, stock levels, and other sources, to improve product visibility and prevent stockouts and overstock surpluses. Moreover, GenAI-powered systems can dynamically organize warehouse layouts according to product popularity and order forecasts of certain items, thereby reducing trip time and boosting efficiency.

Autonomous vehicles: GenAI can create various realistic virtual driving scenarios to train autonomous cars and advanced driver-assistance systems (ADAS) for unpredictable circumstances. Additionally, the technology can improve autonomous vehicles' decision-making abilities by creating simulations of different weather patterns and road conditions.

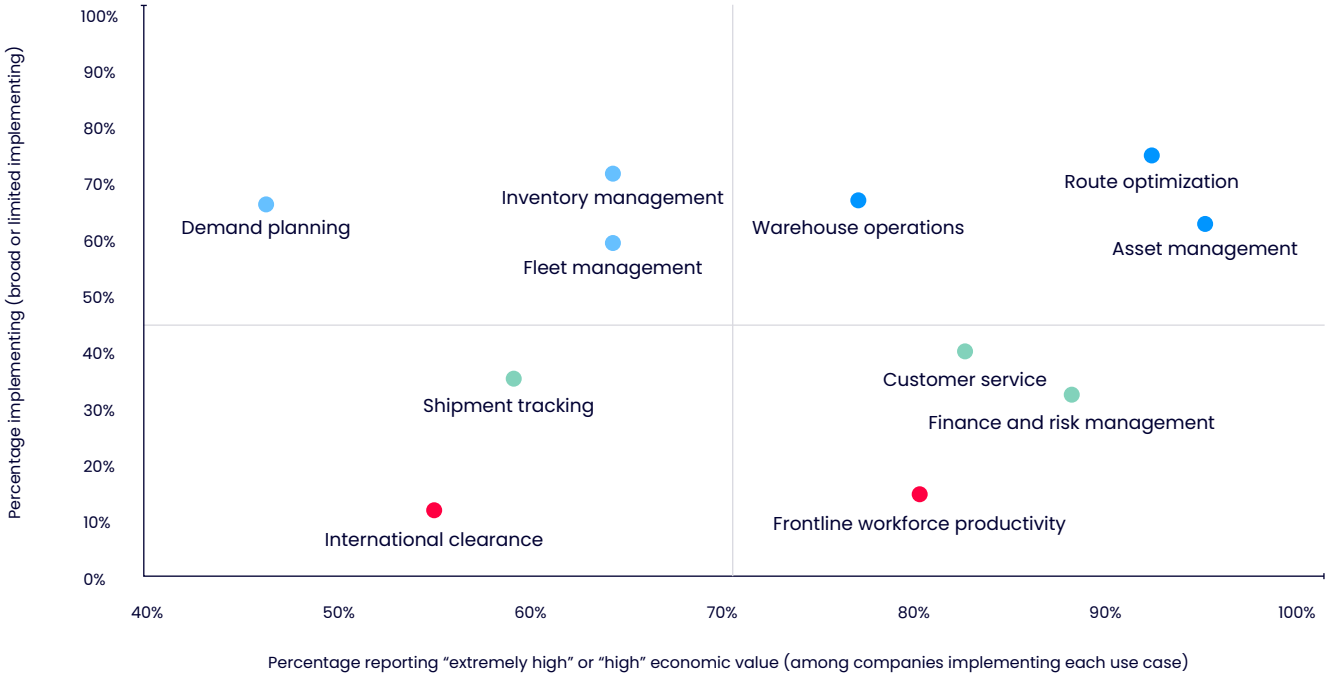
Predictive maintenance and demand forecasting: GenAI can also predict infrastructure and vehicle maintenance requirements before they arise, making it possible for transportation companies to take preventative action and avoid malfunctions and shutdowns. Supply chain managers are increasingly using the technology to analyze historical data related to elements such as seasonality, promotions, customer sentiment, and economic situations. This enables them to create efficient ordering patterns, precisely forecast future trends, and identify hazards.

Visit Site Register

Figure 9: GenAI adoption and impact in transportation

- High adoption, high impact
- High adoption, low impact
- Low adoption, high impact
- Low adoption, low impact
- Adoption too low to measure impact

incomplete - missing info from the source content



Note: GenAI in Transportation Survey carried out among 200 executives worldwide, July 2024.
Source: Deloitte

Visit Site Register

AI Industry Trends

AI Infrastructure & Architecture

As AI and related technologies continue to evolve, enterprises are making significant investments to develop robust, scalable, and efficient AI infrastructure. According to a 2025 study by S&P Global Market Intelligence, GenAI-related investments exceeded US$56 billion in 2024, almost double from US$29 billion in 2023. An area of interest for investors is the infrastructure layer, which includes semiconductors, graphics processing unit (GPU) cloud, photonic fabrics, high-density compute solutions, edge computing, software tools, and sustainable GenAI infrastructure. Investment in GenAI infrastructure nearly quadrupled in 2024 to almost US$26 billion, up from US$6.86 billion in 2023. The top five GenAI infrastructure trends include:

Disaggregated and composable infrastructure: with conventional monolithic architectures becoming expensive and inflexible, enterprises are moving towards disaggregated, software-defined infrastructure, in which compute, storage, and networking resources are dynamically allocated based on workload needs. This includes composable GPU workspaces, particularly in multi-tenant environments, that are fast replacing traditional data centers due to their ability to decouple compute, storage, and networking resources, enabling organizations to reallocate GPU power according to current workloads. For stakeholders, the strategic advantages of investing in composable GPU workspaces include cost efficiency, operational agility, enhanced ROI, and future-proofing IT.

Photonic networking for AI Acceleration: the growing size and complexity of GenAI models require ultra-fast, low-latency networking. Cluster sizes are having to quickly scale from just a few AI processors in a server to tens of processors in a single rack and thousands of processors across multiple racks, all while relying on high-bandwidth, low-latency network connectivity to handle huge data transfers. Photonic fabrics are setting new standards for AI clusters, significantly reducing data transfer times and eliminating network congestion. These platforms allow AI compute to be networked seamlessly, from within processor packages to servers across multiple racks.

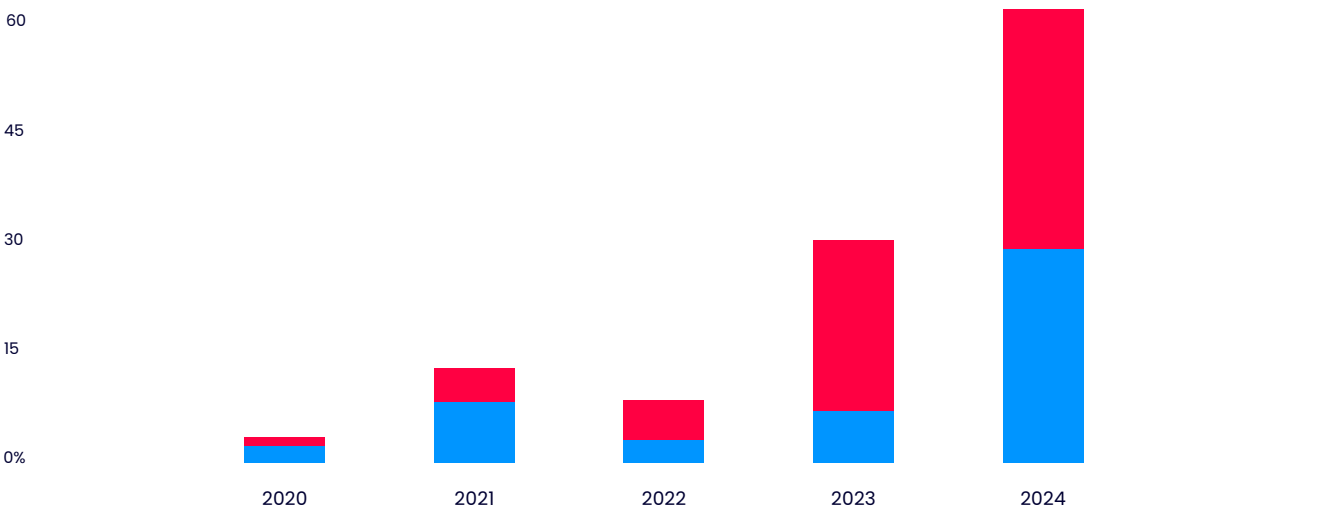
High-density compute solutions: according to recent estimates by Deloitte, continuous improvements in AI and data center processing efficiency could yield an energy consumption level of approximately 1,000 TWh by 2030. These levels of AI workloads demand large-scale hardware infrastructure, making high-density compute solutions critical to achieve maximum output while optimizing power, cooling, and physical space. These solutions are ideal for enterprise GenAI, high-performance computing (HPC), and data center operations.

Edge computing: the shift towards real-time AI processing is driving the need for edge computing solutions. GenAI models often require significant computational resources and memory with large model parameters and deep neural networks (DNNs). Edge computing addresses the limitations of traditional cloud-centric architectures by distributing computational resources closer to the data source, reducing latency and bandwidth consumption.

Visit Site Register

Sustainable infrastructure: GenAI needs massive computational power, rendering it an energy-intensive technology. The production of graphics processing units (GPUs) requires rare earth metals, the mining of which contributes to greenhouse gas (GHG) emissions. Recent estimates suggest that Gen AI could be responsible for creating between 1.2 to 5.0 million metric tons of e-waste by 2030, which is around 1,000 times more e-waste than was produced in 2023. Technology companies are undertaking various initiatives to make GenAI more sustainable. These include energy-efficient chips, smaller models, right-sizing AI/Gen AI workloads, and investments in low-carbon energy sources. A good example is Nvidia's new Blackwell chip that has 30 times improved performance for LLM workloads and 25 times lower energy consumption than the preceding iteration. Another example is Google's TensorFlow and Hugging Face, which have incorporated quantization techniques to reduce the size of models, thereby reducing power and resource requirements.

Figure 10: GenAI infrastructure funding in 2024



Source: S&P Global, as of Jan 10, 2025

Agentic AI

While traditional LLMs are trained on enormous collections of text, images, audio, video, and numbers, and respond to specific human prompts, AI agents (Agentic AI), which build on advanced GenAI models, can act independently, and reason and learn without constant human intervention. Agentic AI technology is gaining traction simply because computers are becoming better at recognizing images and understanding language, mainly due to the evolution of transformer-based technology. Just like humans, these agents work collaboratively using advanced reasoning and planning skills to solve complex, multi-step problems, with LLMs acting as their "brains" for decision-making. What makes them even more attractive is their ability to not only draw from databases and networks but also learn from user behavior and improve over time. Releases such as OpenAI's GPT model family, Anthropic's Claude, and Microsoft's Copilot are driving the current buzz around Agentic AI.

Visit Site Register

Table 6: Agentic AI vs GenAI vs Traditional AI

Feature	Agentic AI	Generative AI	Traditional AI
Primary Function	Goal-oriented action & decision-making	Content generation (text, code, images, etc.)	Focused on automating repetitive tasks
Autonomy	High – Operates with minimal human oversight	Variable – May require user prompts or guidance	Low – Relies on specific algorithms and set rules
Learning	Reinforced Learning – Improves through experience	Data-driven learning – Learns from existing data	Relies on predefined rules and human intervention

Source: AISERA

2025 and beyond

According to Maryam Ashoori, Director of Product Management, IBM Watsonx.ai, 2025 is expected to be the year when companies begin exploring and deploying agentic AI solutions. An early 2025 U.S.-focused study by IBM and business intelligence company Morning Consult involving 1,000 developers building AI applications for enterprise found that as many as 99% were exploring or developing AI agents. Another study by Deloitte conducted in late 2024 predicted that 25% of the companies that use GenAI will launch agentic AI pilots or proofs of concept in 2025, growing to 50% in 2027. Moreover, some of these applications, in some industries, and for some use cases, could see actual adoption into existing workflows in 2025, especially by the back half of the year. Yet another global study conducted by Capgemini found that 50% of the respondents will implement AI agents in 2025, with the number expected to rise to 82% by 2028.

However, Vyoma Gajjar, an AI technical solutions architect, cautions against unbridled optimism, saying that the technology's proliferation requires more than just better algorithms. It needs significant advancements in contextual reasoning and testing for edge cases, and a lack of capabilities in these areas is one of the main hurdles to widespread adoption. Moreover, while the technology is garnering significant attention and investment globally, current Agentic models are prone to making mistakes and getting stuck in loops. In multi-agent systems, "hallucinations" can often spread from one agent to another, which results in a loop of incorrect actions and results. A good example is the AI agent Devin, which was launched by Cognition Software in March 2024 to perform programming jobs unassisted, based on natural language prompts from human programmers. In a recent benchmarking test, Devin was able to resolve nearly 14% of GitHub issues from real-world code repositories, which, even though twice as good as LLM-based chatbots, was still far from being fully autonomous.

Visit Site Register

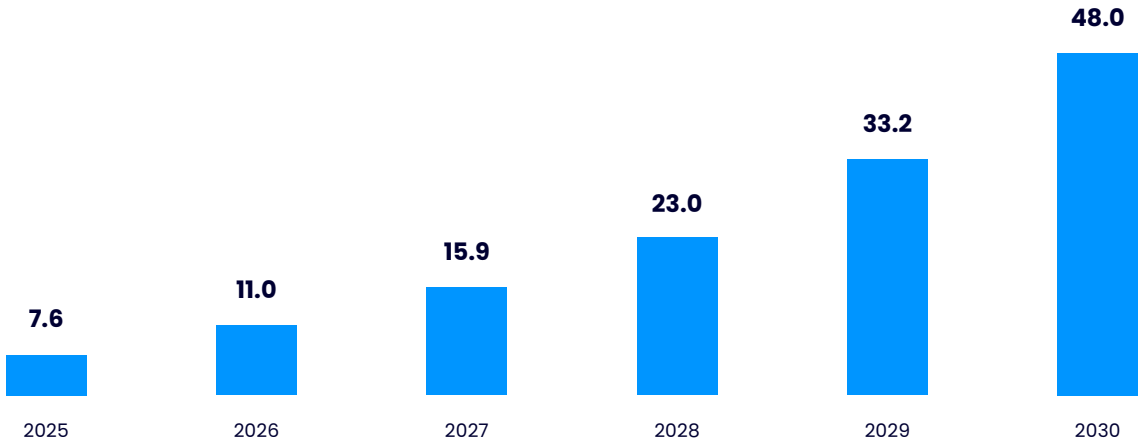
Despite current limitations, the vision of Agentic AI is compelling. The technology is developing at a rapid pace, with some of the latest Agentic AI models employing chain-of-thought functions that, while slower and more deliberative as compared to the more conventional large-scale models, can conduct higher-order reasoning on complex problems. Moreover, multimodal data analysis has the potential to make agentic AI more flexible by increasing the kinds of data that can be analyzed and created. Multimodal AI also shows that agentic AI can be even more powerful when combined with other kinds of AI technologies, such as computer vision (image recognition), transcription, and translation.

The global Agentic AI market is estimated to grow from US$7.6 billion in 2025 to US$48 billion in 2030 at a CAGR of 44.5%.

Use cases

Customer service: American startups such as Sierra, Ema, and Decagon are developing agentic AI chatbots that can act independently according to their understanding of customer intent and emotions. They operate with multiple specialized agents, each responsible for different aspects of the conversation, such as intent recognition, knowledge retrieval, and emotional understanding. For example, an AI agent could anticipate a delayed delivery, notify the customer proactively, and offer a discount to improve satisfaction. It could also transform customer interaction with conversational support that is empathetic and personalized. Agentic AI chatbots can be of various types: reactive, memory-augmented, tool-using, semi-autonomous, multi-agent networks, and self-improving.

Figure 11: Global Agentic AI market size in US$ billions, 2025-2030



Source: AgileIntel

Despite limitations, the vision of Agentic AI is compelling. The technology is developing at a rapid pace, with some of the latest Agentic AI models employing chain-of-thought functions that, while slower and more deliberative as compared to the more conventional large-scale models, can conduct higher-order reasoning on complex problems.

Visit Site Register

Procurement: While current procurement tools focus on data analysis and guided automation, Agentic AI systems such as Zip are already able to function autonomously, guiding employees through complex purchasing decisions by reviewing company policies and requirements.

Sales support: Agentic CRMs such as Rox not only store customer data but also help companies get a better understanding of their customers by predicting their needs and proactively engaging with them. U.S.-based 11x has developed two Agentic AI systems, Alice and Mike. While the former functions as a digital sales development representative that autonomously identifies key decision makers and schedules meetings, Mike automates inbound and outbound calls in 28 languages in a personalized, low-latency phone call.

Scientific and materials discovery: even though machine learning and non-agentic AI have been used in areas such as drug discovery and new material creation for a long time, Agentic AI is poised to disrupt the field. Agents can not only analyze the properties of specific materials but also propose new materials or combinations based on the characteristics the user is seeking. Moreover, it can also identify optimal suppliers based on priorities such as cost or timing and even order necessary materials. One promising use is ADME (Absorption, Distribution, Metabolism, Excretion) profiling, which predicts drug behavior in the body. A major hurdle is drug candidate failure in later stages due to poor ADME properties or toxicity popping up. Agentic AI can predict these properties early by analyzing molecular structures and historical data, filtering out unfavorable candidates and prioritizing promising ones.

Entertainment: Fully autonomous AI agents are already being used in the gaming industry owing to their ability to provide human-like behavior and gameplay for non-player characters (NPCs). For example, researchers created a small virtual town populated with AI by building a sandbox setting similar to The Sims with 25 agents called “Stanford AI Village”. In this village, users can observe and interact with agents as they share news, build relationships, and arrange group activities.

Application and Cybersecurity: According to a report by Skybox Security Research Lab, over 30,000 new vulnerabilities were identified in the year leading up to June 2024. As cyber threats grow in number and sophistication, Agentic AI is assuming a critical role in bolstering security postures. This is mainly because the technology outperforms conventional security systems, such as firewalls and antivirus software, to provide a new level of automated defense. It not only analyzes factors like application code, network traffic, user behavior, and system logs to detect anomalies, but also prioritizes these vulnerabilities by risk level and automatically applies patches or recommends fixes.

Fully autonomous AI agents are already being used in the gaming industry owing to their ability to provide human-like behavior and gameplay for non-player characters (NPCs). For example, researchers created a small virtual town populated with AI by building a sandbox setting similar to The Sims with 25 agents called “Stanford AI Village”.

Visit Site Register

Figure 12: Evolution to multimodal GenAI agents

The evolution can be broken down into three key phases:

2000s

Integration of Machine Learning (ML)

Learning from data: The integration of ML allowed agents to learn from large datasets, improving their ability to make decisions and perform tasks. This was a significant step forward from rule-based systems, as agents could now adapt to new information and improve over time.
Natural Language Processing (NLP) enabled user interactions: Advances in NLP enabled agents to understand and generate human language more effectively, making interactions more natural and intuitive.

2010s

Introduction of multimodality

Combining text, images, and audio: Multimodal agents emerged, capable of processing and integrating information from various sources. For instance, an agent could analyze a text description, recognize objects in an image, and understand spoken commands. This multimodality made agents more versatile and capable of handling complex tasks.
Enhanced user interactions: Multimodal agents could interact with users in more dynamic ways, such as providing visual aids in response to text queries or understanding context from a combination of spoken and visual inputs.

2000s present

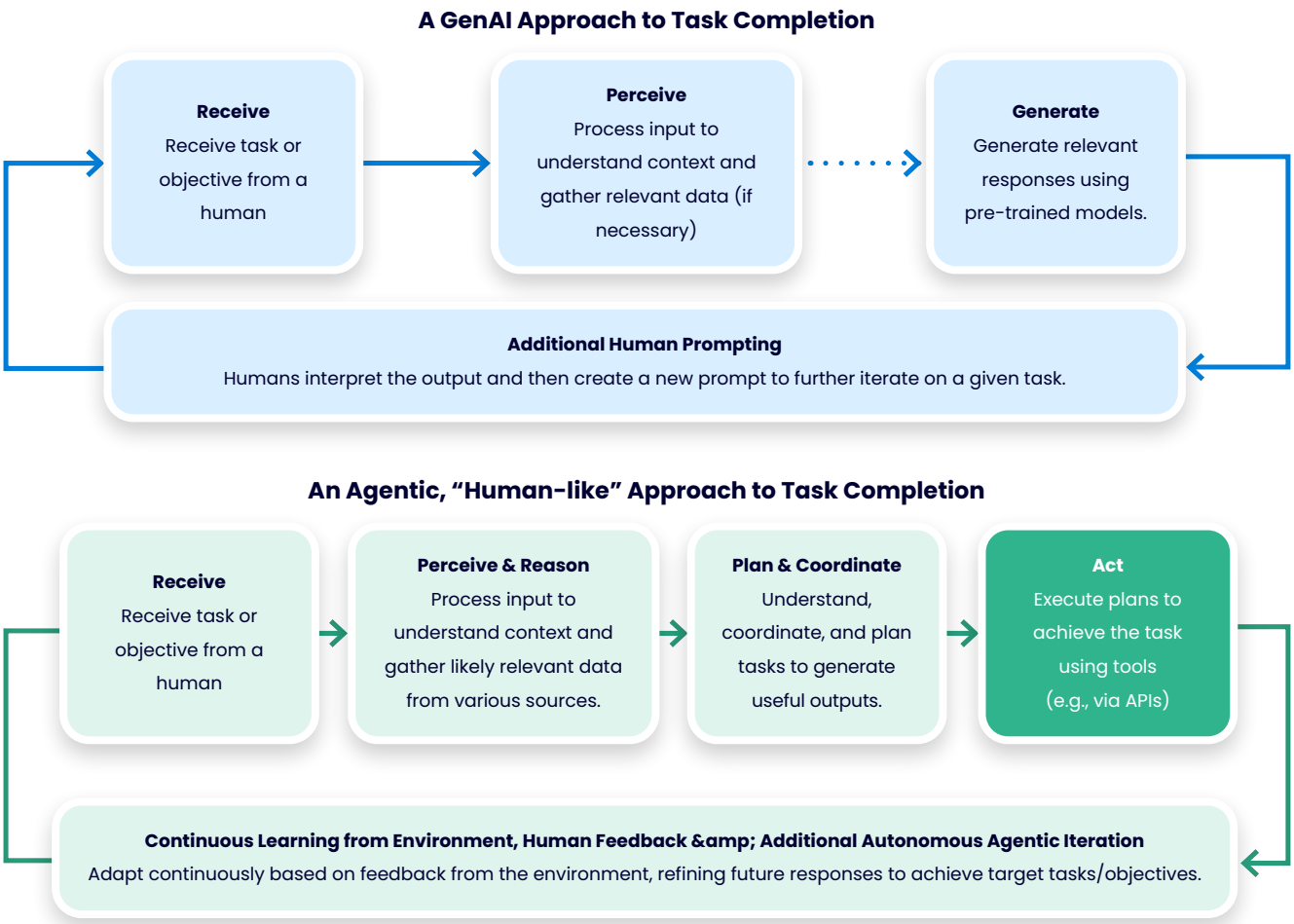
Advanced autonomy and real-time interactions

Advanced autonomy: Agents can operate independently, rationalize and set their own goals, develop path(s) to attain these goals, and make independent decisions without constant human intervention, leveraging data from multiple sources or synthetic datasets. In a multi-agentic orchestration system, the first set of agents focus on mimicking human behavior (e.g. ChatGPT-4o), that is, thinking fast to come up with solution approach, while the second set of agents focus on slow reasoning (e.g. ChatGPT-1o) to come up with a vetted solution5 . Combining thinking fast and slow reasoning, agents can process information and make optimal decisions in real-time – crucial for applications like autonomous vehicles, real-time customer service, and various mission-critical business processes. This autonomy makes agentic AI particularly powerful in dynamic and complex real-world environments.
User interactions within an ethical and responsible AI-controlled environment: With increased capabilities, there has also been a focus on ensuring that agentic systems operate ethically and responsibly, considering factors such as bias, transparency, and accountability.

Source: AgileIntel

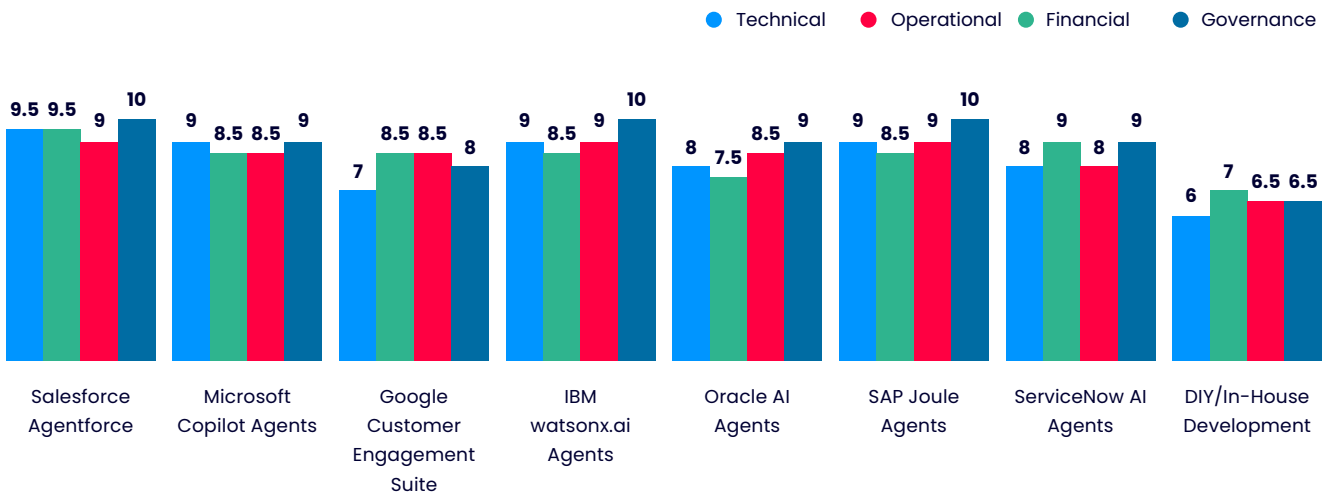
Visit Site Register

Figure 13: GenAI vs Agentic AI approach to task completion



Source: Cambridge Centre For Alternative Finance

Figure 14: Comparative scoring of leading Agentic AI solutions



Source: The Futurum Group

Visit Site Register

AI Governance – Risk, Compliance, Responsible AI

Responsible AI

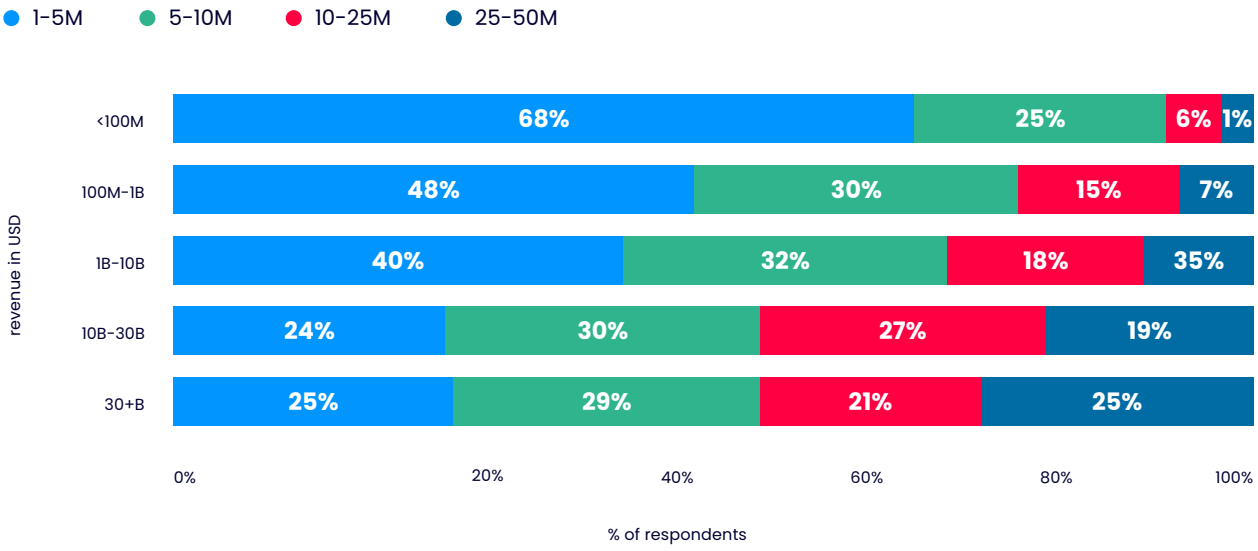
According to a 2024 McKinsey report, global GenAI use doubled in the year leading up to the study, with ChatGPT boasting 200 million weekly active users as of August 2024, double the number from 2023. Another study by Thomson Reuters conducted in 2025 showed that 95% of the respondents believe GenAI to be central to their organization's workflow within the next five years. Pertinently, the pace of GenAI adoption is quicker than that of personal computers and the internet.

However, it is the allure of GenAI's potential that has led organizations to dive headfirst into adoption without mitigating risks adequately. This has proven to be particularly challenging given the nascency of the technology as a whole. Moreover, GenAI's greater sophistication as compared to traditional AI poses a huge challenge from a technical standpoint. After all, AI models have evolved from just a few parameters with ML, to tens of thousands with deep learning, and now to millions, billions, and at times trillions with the LLMs.

Therefore, companies and organizations are increasingly designing GenAI applications responsibly, addressing potential risks and transparently sharing lessons learned to help establish best practices. According to a 2025 McKinsey report, companies that have been able to capture significant value from the technology's use have consistently paid more attention to address known risks and identify and prevent new ones.

According to a 2024 McKinsey report, global GenAI use doubled in the year leading up to the study, with ChatGPT boasting 200 million weekly active users as of August 2024, double the number from 2023. Pertinently, the pace of GenAI adoption is quicker than that of personal computers and the internet.

Figure 15: Investment in responsible AI by company revenue, 2024



Note: The survey was conducted among business leaders from over 30 countries, N=759

Source: Stanford AI Index Report 2025

Visit Site Register

Responsible AI (RAI) is a comprehensive and holistic framework that guides companies and other organizations to implement AI in a way that enables them to benefit from AI systems while mitigating risk and remaining consistent with corporate values. For GenAI to be integrated across industries at scale, companies must implement the principles of RAI across the full application life cycle by governing their data, protecting company intellectual property (IP), preserving user privacy, and complying with laws and regulations. One way of doing it is by automating and scaling parts of AI governance, security, and risk management programs to detect and monitor configured guardrails and controls more efficiently. Another way is to adopt a risk-tiered approach that applies different monitoring standards to AI systems based on risk and impact on customers, partners, and employees.

Table 7: Notable RAI policymaking milestones

Date	Stakeholders	Scope	Description
May 2024	OECD	Global	The OECD updated its AI principles and refined its framework to reflect the latest advancements in AI governance. These principles emphasized building AI systems that take into account inclusive growth, transparency, and explainability, as well as respect for the rule of law, human rights, and democratic values.
May 2024	Council of Europe	Europe	The Council of Europe adopted a legally binding AI treaty (The Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law). This treaty was drafted to ensure that the activities within the life cycle of AI systems align with human rights, democracy, and the rule of law.
Jun 2024	European Union	Europe	The EU passed the AI Act (EU AI Act), the first comprehensive regulatory framework for AI in a major global economy. The act categorizes AI by risk, regulating them accordingly and ensuring that providers—or developers—of high-risk systems bear most of the obligations.
Jul 2024	African Union	Africa	The African Union launched its Continental AI Strategy (AU AI Strategy), outlining a unified vision for AI development, ethics, and governance across the continent. The strategy emphasizes the ethical, responsible, and equitable development of AI within Africa.
Sep 2024	United Nations	Global	The United Nations updated its Governing AI for Humanity report (U.N. AI Advisory Body), outlining efforts to establish global AI governance mechanisms. The report recommends developing a blueprint to address AI-related risks and calls on national and international standards organizations, technology companies, civil society, and policymakers to collaborate on AI standards.
Oct 2024	G7	Global	The G7 Digital Competition Communiqué (G7 AI Cooperation) reaffirmed commitments to fair and open AI markets, stressing the need for coordinated regulatory approaches. Previous discussions focused on competition and the regulatory challenges posed by AI's rapid growth.
Oct 2024	ASEAN and the US	Asia and the US	Following the 12th ASEAN-United States Summit, ASEAN-U.S. leaders issued a statement on promoting safe, secure, and trustworthy AI. They committed to cooperating on the development of international AI governance frameworks and standards to advance these goals.
Nov 2024	International Network of AI Safety Institutes	Global	The first International Network of AI Safety Institutes was established, bringing together nine countries and the EU to formalize global AI safety cooperation. The network unites technical organizations committed to advancing AI safety, helping governments and societies understand the risks of advanced AI systems, and proposing solutions.
Feb 2025	Arab League	Arab Nations	The Arab Dialogue Circle on "Artificial Intelligence in the Arab World: Innovative Applications and Ethical Challenges" was launched at the Arab League headquarters, focusing on AI innovations while placing a strong emphasis on ethical considerations.

Source: Stanford AI Index Report 2025

Visit Site Register

GenAI in Enterprise: Case Studies

WestRock's GenAI integration has yielded higher productivity and lower costs.

Paul McClung, VP of internal audit at WestRock, a global sustainable, fiber-based packaging solutions company, first heard about GenAI in 2022 but dismissed its use to augment the company's audit function. However, the company's IT department developed a secure GenAI platform in late 2023 for all internal departments to experiment with.

One of the first applications of the technology was on the front end of the audit process to draft objectives. When this proved to be successful, Paul decided to automate the entire audit process by ingesting data and running it through a seamless model with a click of a button. However, the team found that linking several tasks instead of executing them individually would prove to be more effective. Another effective strategy was to integrate a high level of standardization within internal processes, starting with standard prompts to write audit objectives and execution methods. This enabled WestRock to automate the process of creating sample risk and control matrices, draft audit programs, and even suggest technology tools and scripts for the company to consider.

Some of the early value captured through the use of GenAI has unsurprisingly been higher productivity and lower costs. However, Paul cautions that these benefits are still very much in their nascent stages, and to realize optimum value, the company will have to fully reengineer its processes, timelines, milestones, and resource deployment models. It will also have to move away from its previous strategy of getting its programmers to develop scripts based on requirements, to a more iterative process that involves developing scripts in real time and adjusting as needed. This requires a team approach where multiple people challenge the results of the GenAI models, but in a condensed time frame based on the technology's speed.

According to Paul, WestRock's future with GenAI technology involves integration with Agentic AI to add a learning mechanism to the platform that builds on historical lessons to improve and expand its scope of operations in the future. Another immediate goal is to leverage the company's learnings with data analytics to improve the implementation of GenAI. This includes leveraging the platform with continuous monitoring and full population assessments rather than just sampling.

WestRock's future with GenAI technology involves integration with Agentic AI to add a learning mechanism to the platform that builds on historical lessons to improve and expand its scope of operations in the future. Another immediate goal is to leverage the company's learnings with data analytics to improve the implementation of GenAI.

Visit Site Register

Over the medium term, the company is expected to develop a dynamic system in which risk assessment questions are generated based on changes in industry and external environment data gathered and analyzed in real time. This will likely result in follow-up reporting, action tracking, and trending being automated through interactive chatbots.

McDonald's much-touted conversational AI solution was withdrawn from the U.S. in June 2024

McDonald's acquisition of voice-based conversational AI technology company Apprente in 2019 marked the beginning of the company's exploration with GenAI. Apprente specialized in developing sophisticated speech recognition and natural language processing (NLP) systems designed to handle complex, multi-lingual, and context-sensitive interactions. These solutions were expected to automate McDonald's drive-thru systems and streamline the order-taking process. In October 2021, the company forged a strategic partnership with IBM to leverage its AI and cloud computing expertise to expand the deployment of AI-powered drive-through systems across more locations.

However, despite the integration of sophisticated technology, the GenAI-enabled system frequently misunderstood customer orders with background noise, varied accents, and complex orders, leading to significant misinterpretations. In fact, many videos of the AI's failures were recorded and widely shared on social media, causing much negative publicity for McDonald's. In June 2024, the fast-food chain withdrew the automated systems from over 100 locations around the U.S. The key reasons for failure are mentioned below:

Real-world testing: One of the main reasons for this failure was the lack of real-world testing to ensure the systems can handle the variability of actual customer interactions. This includes simulating different accents, background noises, and complex order scenarios. Moreover, the system wasn't trained on exhaustive and diverse datasets that were updated regularly to keep it adaptable to new linguistic patterns and customer behaviors.

User-centric design and feedback loops: The company failed to incorporate user feedback into the development cycle to continually refine and improve the system. This is especially important for a company like McDonald's, in which understanding user needs and expectations is crucial for designing AI systems. AI systems should be continuously updated and refined based on real-world performance data and user feedback. Establishing feedback loops allows for ongoing improvement and adaptation to changing conditions and user behaviors. This iterative process helps maintain the system's relevance and effectiveness over time.

One of the main reasons for this failure was the lack of real-world testing to ensure the systems can handle the variability of actual customer interactions. This includes simulating different accents, background noises, and complex order scenarios.

Visit Site Register

GenAI Technology

The latest LLMs such as GPT-4 (1.8T parameters), Claude 3 (2T parameters), and Meta's LLaMA 3 (405B parameters), are now being trained on billions, or even trillions of parameters, resulting in significant advancements in natural language understanding, code generation, and reasoning. In fact, some of these models are now operating at or near human-level accuracy on functions such as reading, image recognition, speech recognition, and language understanding.

Some of the top current LLMs include:

Claude: created by Anthropic, Claude focuses on constitutional AI and has three primary branches - Opus, Haiku, and Sonnet. Its latest iteration is the Claude 3.5 Sonnet that can decipher nuance, humor, and complex instructions better than previous versions. The LLM also has broad programming capabilities that make it ideal for application development. In October 2024, Claude added a computer-use AI tool that allows it to use a computer like a human does.

DeepSeek-R1 is an open-source reasoning LLM that uses reinforcement learning to deliver mathematical problem-solving and logical inference capabilities. DeepSeek-R1 can perform critical problem-solving through self-verification, chain-of-thought reasoning, and reflection.

Ernie: released by Chinese technology company Baidu in August 2023, Ernie is said to have 10 trillion parameters and has garnered 45 million users globally.

Gemini: a product of the Google family of LLMs, Gemini models are multimodal and available as a web chatbot, the Google Vertex AI service, and via API. They are available in three variants Ultra, Pro, and Nano. Ultra is the largest and most capable, Pro is the mid-tier model, and Nano is the smallest model, designed for efficiency with on-device tasks. The latest version of Gemini, the Gemini 1.5 Pro, was released in May 2024.

Llama: developed by Meta, Llama was first released in 2023 and then subsequently in July 2024 as both a 405 billion and 70 billion parameter model. The most recent version is Llama 3.2, which was released in September 2024, initially with smaller parameter counts of 11 billion and 90 billion. Llama uses a transformer architecture and was trained on many public data sources, including webpages from CommonCrawl, GitHub, Wikipedia, and Project Gutenberg.

ChatGPT continues to be the market leader, but its growth has slowed as Google and Microsoft introduce enhancements to their AI assistants. Among startups, general-purpose AI chatbots are experiencing gradual but consistent user acquisition, while business-focused Claude AI is currently leading in terms of growth.

Visit Site Register

Table 8: Significant model and dataset releases

Feature	Copilot (Microsoft)	ChatGPT (OpenAI)	Gemini (Google)	Llama (Meta)
Developer	Microsoft	OpenAI	Google DeepMind	Meta AI
Latest Model	Microsoft 365 Copilot (2025)	GPT-4.5 (2025)	Gemini 2.5 (2025)	Llama 4 (2025)
Primary Focus	Integration of AI in Microsoft apps	General AI, conversation, coding	Multimodal AI, Google ecosystem	-
Training Data	Built on OpenAI's GPT-4, Proprietary	Broad, multimodal, diverse	Web-scale, multimodal	-
Key Features	Deep integration with Microsoft 365 and GitHub	Web browsing, DALL-E, document analysis, and voice interactions	Multimodal (text, images, audio, video), Google services integration	Open-source LLM optimized for research and on-device deployment
Code Generation	Excellent in Python, JavaScript, C++, Java	Excellent (Python, JS, SQL)	Strong (Python, JS, SQL)	Average
Multimodal Support	Supports text and image generation	Strong (images, text)	Very strong (text, images, audio, video)	Text-only; no native multimodal support for images, audio, or video.
Memory Feature	Yes	Yes (for Plus users)	Yes	Yes
API Availability	No	Yes	Yes	Yes
Free Version	Yes (Microsoft Edge)	Yes (GPT-3.5)	Yes (Gemini 1.0)	Yes (Llama 2 and Llama 3.2)
Strengths	Code-specific assistance	Versatile, reliable general AI, Strong conversational abilities, Integrated with plugins	Best multimodal AI, Google ecosystem integration, Strong reasoning	Privacy & mobile deployment
Weaknesses	Over-reliance risks	No real-time browsing in the free version, can generate hallucinations, and Advanced features are behind a paywall.	Requires Google integration, some accuracy issues	Limited complexity handling
Best For	Software development	General-purpose AI, chatbots, writing, and research	Multimodal tasks, search, and productivity	Offline, low-resource environments
Cost Structure	Included with Microsoft 365 subscriptions	Subscription-based (Plus/ Team)	Free with a Google account	Subscription-based

Source: Swiss German University, Web Search

Visit Site Register

Table 9: Leading GenAI models and specifications

Model	Creator	Context Window	Artificial Analysis Intelligence Index	End-to-End Response Time
Command-R	Cohere	128k	15	6.97
Jamba 1.6 Mini	AI21 labs	256k	18	2.89
DBRX	Databricks	33k	20	NA
Codestral (May '24)	Mistral AI	33k	20	5.01
LFM 40B	Liquid	32k	22	3.23
Qwen3 0.6B	Alibaba	32k	23	NA
Yi-Large	Alibaba	32k	28	7.78
Nova Micro	Aws	130k	28	1.79
Tulu3 405B	Ai2	128k	40	NA
Phi-4	MS Azure	16k	40	12.98
Phi-4	MiniMax	4m	40	16.51
Sonar Pro	Perplexity	200k	43	7.98
Reka Flash 3	Reka	128k	47	45.06
Claude 3.7 Sonnet	Anthropic	200k	48	7.44
GPT-4o	OpenAI	128k	50	3.83
Llama 4 Maverick	Meta	1m	51	4.21
Grok 3	X.AI	1m	51	10.36
DeepSeek V3	Deepseek	128k	53	22.47
Gemini 2.5 Pro	Google	1m	68	39.73
OpenChat 3.5	Openchat	8k	NA	10.87
Arctic	Snowflake	4k	NA	NA
Solar Mini	Upstage	4k	NA	38.52

Note: Context windows - Maximum number of combined input & output tokens., Artificial analysis Intelligence Index - a comprehensive benchmark used to evaluate and compare the intelligence of language models

Source: Artificial Analysis

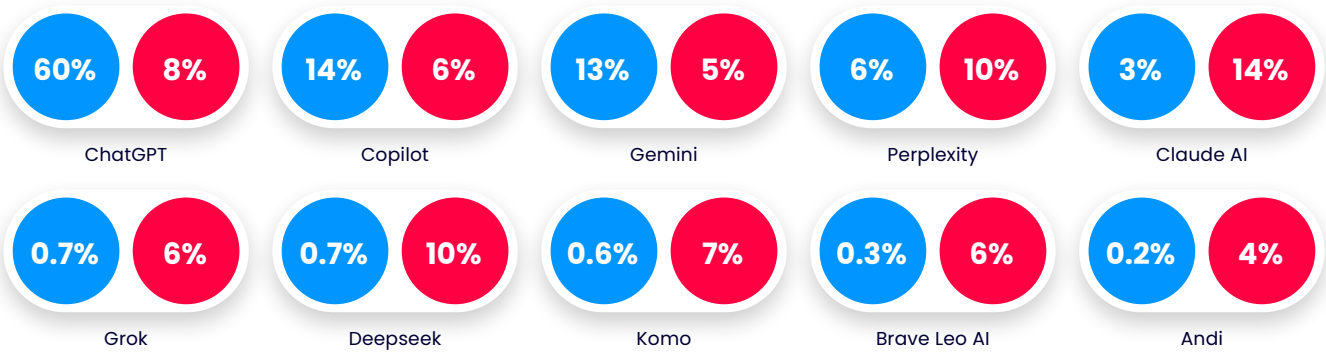
Visit Site Register

Table 10: Illustrative capabilities of GenAI platforms from select frontier labs

	2022-23	Jan 2025
Anthropic	**Claude** • Not multimodal (text only) • Limited contextual understanding (difficulty with complex conversations) • No tool usage	**Claude 3.5** • Multimodal (text, audio, and images) • Enhanced contextual understanding and coherence during long interactions • Experimental computer usage capability for some users
Google Gemini	**Google Bard** • Not multimodal (text only) • Fair reasoning • Limited contextual understanding (difficulty with complex conversations) • Limited real-time data integration • Low personalization (limited adaptability)	**Gemini 2.0 Flash** • Multimodal (text, audio, and images) • Advanced reasoning (capable of multistep problem-solving and nuanced analysis) • Enhanced contextual understanding (maintains coherence in long dialogues) • Real-time data integration (from Google Search) • Advanced personalization (user context)
Meta	**Llama 1** • Not multimodal (text only) • Fair reasoning • Limited contextual understanding (difficulty with complex conversations) • No API access	**Llama 3.3** • Text-based (earlier versions were multimodal, LLaMa 3.2) • Advanced reasoning (capable of multistep problem-solving and nuanced analysis) • Enhanced contextual understanding (maintains coherence in long dialogues) • API access (tools for model and agent development)
Meta	**Microsoft Phi-1** • Not multimodal (text only) • Fair reasoning (i.e., limited to coding tasks) • Focused training (smaller, coding-focused data set)	**Phi-4** • Multimodal (text, audio, and images) • Advanced reasoning (capable of multistep problem-solving and nuanced analysis) • Comprehensive training (diverse data)
OpenAI	**GPT-3.5** • Not multimodal (text only) • Fair reasoning ability (e.g., scored high on SAT, but bottom 10% on bar examination) • Limited contextual understanding (difficulty with coherence in complex conversations) • Standard API access (for text generation)	**OpenAI 0** • Multimodal (text, audio, and images) • Advanced reasoning (e.g., top 10% on bar examination) • Enhanced contextual understanding and coherence during long interactions • Advanced API access (supports multimodal inputs)

Source: McKinsey

Figure 16: Leading GenAI AI chatbots market share and user growth in the U.S., April 2025

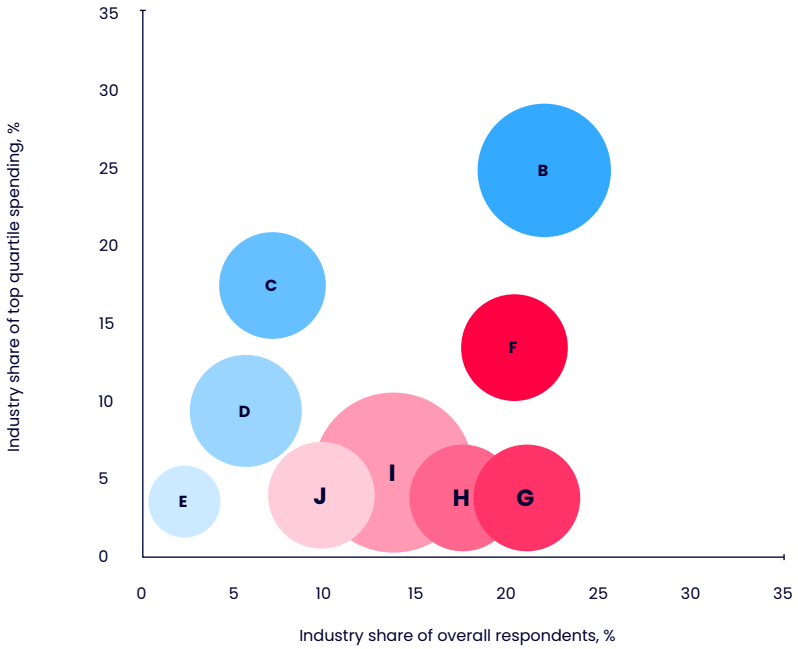


Source: FirstPageSage

Visit Site Register

GenAI and Investments

Figure 17: GenAI spending vs economic potential of the industry



Circle size = economic potential, US$ billion

Higher spend Share in top quartile is higher than share in overall survey

- **A** Healthcare
- **B** Technology
- **C** Media and telecom
- **D** Advanced industries?
- **E** Agriculture

Lower spend Share in top quartile is lower than share in overall survey

- **F** Financial services
- **G** Energy and materials
- **H** Consumer goods and retail
- **I** Hardware engineering and construction
- **J** Travel, transportation, and logistics

Note: McKinsey US CxO survey, Oct-Nov 2024, n=118

Source: McKinsey

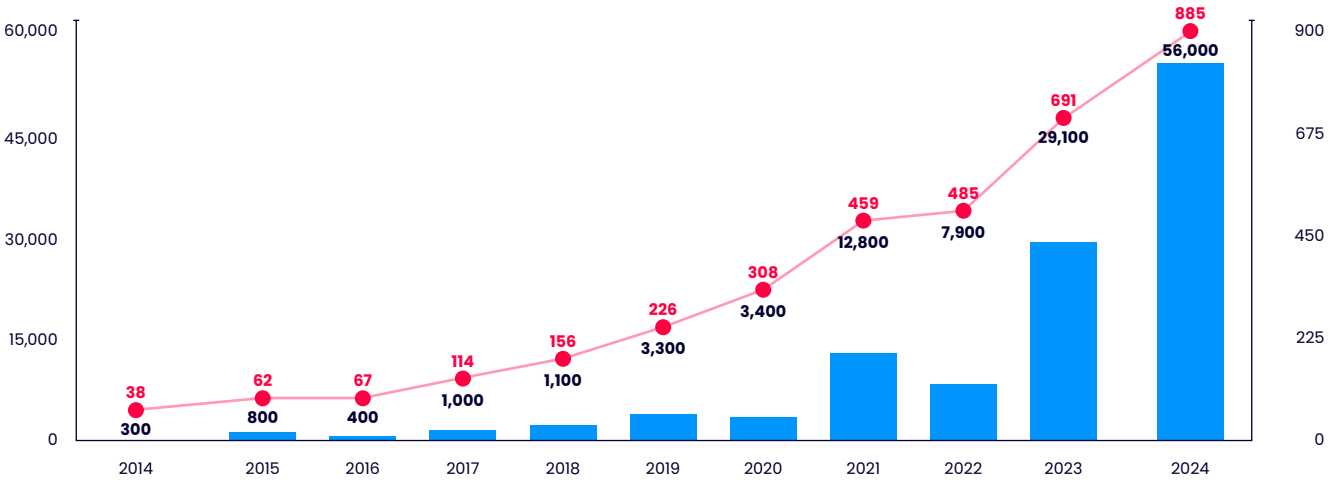
Table 11: 10 most active investors in GenAI

Company	VC Investments	VC-Backed Exits	Median deal size (US$ Mn)	Select Portfolio Companies
Sequoia	84	7	16	OpenAI, xAI, Glean, Safe Superintelligence
Gaingels	76	6	10	Adbridge, Cerebras Systems, People.ai, Figure
Pioneer Fund	74	2	3.6	Moonvalley, Agentic Labs, Model ML
Andreessen Horowitz	73	4	30	Mistral, Cursor, OpenAI
Khosla Ventures	57	3	15	OpenAI, Curai Health, Replika
Soma Capital	55	5	4.2	Artisan, Imbue, Moonvalley
Alumni Ventures	54	3	11.7	Cohere, Lambda, Groq
General Catalyst	41	7	26.3	Cohere, Lambda, Groq
Lightspeed	40	2	32.4	Anthropic, Granola, xAI
Lightspeed	40	-	0.2	Zealth-ai, Omma, Banqora

Source: PitchBook, June 06, 2025

Visit Site Register

Figure 18: VC investments in GenAI, 2014-2024, US$ Millions



Source: TechCrunch, as of January 3, 2025

Table 12: Top private equity deals in Gen AI – Q1' 2025

Company	Round Amount	Round Date	Round Valuation	Select Investors	Country
OpenAI	US$40.0B	Venture Capital 2025-03-31	US$300.0B	SoftBank, Altimeter Capital, Coatue, Microsoft, Thrive Capital	US
Anthropic	US$3.5B	Series E 2025-03-03	US$61.5B	Lightspeed Venture Partners, Bessemer Venture Partners, General Catalyst, Menlo Ventures, Salesforce Ventures	US
Safe Superintelligence	US$2.0B	Series B 2025-03-09	US$30.0B	Greenoaks, Andreessen Horowitz, Sequoia Capital	US
Groq	US$1.5B	Undisclosed 2025-02-10	N/A	Kingdom of Saudi Arabia	US
Anthropic	US$1.0B	Corporate Minority 2025-01-22	N/A	Google	US
Isomorphic Laboratories	US$600.0M	Series A 2025-03-31	N/A	Thrive Capital, Google Ventures, Alphabet	UK
Saronic	US$600.0M	Series C 2025-02-18	US$4.0B	Elad Gil, Andreessen Horowitz, General Catalyst, 8VC, Caffeinated Capital	US
Lambda	US$480.0M	Series D 2025-02-19	US$2.5B	Andra Capital, SGW, 1517 Fund, Crescent Cove Advisors, Super Micro Computer	US
Apptronik	US$403.0M	Series A 2025-02-12	N/A	B Capital, Capital Factory, Korea Investment Partners, ARK Invest, Atinum Investment	US
CoreWeave	US$350.0M	Corporate Minority 2025-03-10	N/A	OpenAI	US

Source: CB Insights, May 01, 2025

Visit Site Register

GenAI Infrastructure Development

The recent debut of DeepSeek's R1 model, which can deliver advanced performance at much lower costs compared to frontier models, has introduced a significant shift in the GenAI landscape. Suddenly, hyperscale data centers, which require large investments, are no longer the limiting factor in GenAI progress. In fact, industry experts are arguing that the R1 has ushered in an era where leading models are trained and deployed with significantly lower resource requirements, potentially putting an end to the trillion-dollar arms race in GenAI infrastructure.

However, experts and industry players view DeepSeek's efficiency gains as a catalyst for even more aggressive GenAI deployment, with computing power and related infrastructure fast emerging as one of this decade's most critical resources. This is because millions of servers continuously run to process the foundation models and ML applications that underpin AI and related technologies. This is why Sam Altman, CEO of OpenAI, has reportedly discussed creating a US$7 trillion fund for GenAI investment before 2030.

According to a 2025 study by McKinsey, by 2030, capital investments to support AI-related data center capacity demand are expected to range from about US$3 trillion to US$8 trillion. Around 15% of this investment is expected to be directed towards builders for land, materials, and site development; 25% to energizers for power generation and transmission, cooling, and electrical equipment; and the largest share of 60% to technology developers and designers, which produce chips and computing hardware for data centers.

This investment is fueled by the mass-adoption of GenAI, the integration of AI-powered applications across various industries, and the enterprise race to build competitive infrastructure. Moreover, governments are now increasingly investing heavily in AI infrastructure to improve their security and economic posture, and technological independence.

The recent debut of DeepSeek's R1 model, which is able to deliver advanced performance at much lower costs compared to frontier models, has introduced a significant shift in the GenAI landscape. Suddenly, hyperscale data centers, which require large investments are no longer the limiting factor in GenAI progress.

Visit Site Register

Investment challenges:

Technological uncertainties: disruptions or advancements in model architectures, including efficiency gains in compute utilization, could result in a reduction in expected hardware and energy demand.

Supply chain constraints: labor shortages, supply chain bottlenecks, and regulatory hurdles could delay grid connections, chip availability, and data center expansion, slowing overall AI adoption and innovation.

Geopolitical tensions: the recent tariffs imposed by the Trump administration and technology export controls have ushered in an era of uncertainty in computing power demand, potentially impacting infrastructure investments and AI growth.

Governance and ROI: AI governance issues, including bias, security, and regulation, could add additional layers of complexity, thereby slowing development. Additionally, with AI inference expected to become the dominant workload by 2030, it poses a major unpredictable cost component. Therefore, companies could face problems demonstrating clear ROI from related AI investments.

Market supply and demand: Global semiconductor manufacturing is controlled by only a few firms, stifling competition. Therefore, the ability of the market to build capacity remains insufficient to meet current demand, while at the same time, shifts in AI model training methods and workloads make it difficult to predict future demand for specific chips.

Competitive advantage: To gain a competitive advantage in an increasingly crowded market, companies are creating custom models, fine-tuning existing models, or using retrieval augmented generation (RAG) embedding to give GenAI systems access to up-to-date and accurate corporate information. These endeavors require significant investments in infrastructure for training and deploying these systems.

Grid weaknesses: Powering data centers could face disruptions due to existing grid weaknesses and heat management challenges arising from rising processor densities.

Visit Site Register

Value Creation through GenAI

According to a September 2024 BCG article, GenAI transformation can yield a 1-to-2 percentage point increase in revenue and an 8% to 12% cost reduction compared with the baseline. A more recent 2025 BCG study found that GenAI can address 30% to 50% of industry-agnostic IT costs, thereby sparking potential savings of up to 10% in the technology function. A 2024 McKinsey study offers a quantitative view, with estimates pegging GenAI to result in cost savings opportunities of US$1.4 trillion to US$2.6 trillion across functions such as customer service, R&D, manufacturing, supply chain, and procurement.

Yet another study by Google Cloud involving over 2,500 C-suite leaders of U.S. companies with more than US$10 million in revenue found that a total of 86% of those who implemented GenAI saw their revenue increase by more than 6%. Additionally, 77% witnessed an improvement in their leads and customer acquisition, 45% saw employee productivity at least double, 56% reported improved cybersecurity, and 71% said that they were able to resolve issues faster.

Moreover, the emergence of increasingly capable small LLMs has lowered the inference cost for a system performing at the level of GPT-3.5 over 280-fold between November 2022 and October 2024. At the hardware level, costs have reduced by 30% annually, while energy efficiency has improved by 40% each year. Open-weight models are closing the gap with closed models, reducing the performance difference from 8% to just 1.7% on some benchmarks in a single year. Together, these trends are rapidly lowering the barriers to advanced AI.

GenAI tools can deliver enterprise cost savings by automating repetitive tasks, accurately forecasting next-generation spending, creating detailed simulations of various operational scenarios, monitoring real-time spending, categorizing strategic spending, and enhancing supplier management. For example, GenAI-powered chatbots can process large volumes of customer queries, reducing operational costs. GenAI models can analyze historical spending data and provide accurate predictions for future expenditures, enabling companies to manage budgets more effectively, avoid unnecessary costs, and allocate resources effectively.

GenAI models can also create detailed simulations that generate realistic scenarios, helping businesses test the impact of various decisions in a virtual environment before implementing them in a real-life setting. Finally, the technology can monitor real-time spending and flag deviations that prevent cost overruns.

The emergence of increasingly capable small LLMs has lowered the inference cost for a system performing at the level of GPT-3.5 over 280-fold between November 2022 and October 2024. At the hardware level, costs have reduced by 30% annually, while energy efficiency has improved by 40% each year.

Visit Site Register

Vendor Landscape

Table 13: Significant AI model and dataset releases, 2024 onwards

Date	Name	Category	Creator (s)
Sep 11, 2024	NotebookLM Podcast Tool	Text-to-podcast	Google Labs
Sep 12, 2024	o1-preview	Language, math, biology	OpenAI
Sep 17, 2024	NVLM (D, H, X)	Vision, Language	Nvidia
Sep 19, 2024	Qwen2.5	LLM	Alibaba
Oct 16, 2024	Ministral	LLM	Mistral
Oct 22, 2024	Anthropic Computer Use	Agentic Capability	Anthropic
Oct 28, 2024	Apple Intelligence	iPhone feature	Apple
Dec 3, 2024	Nova Pro	Multimodal	Amazon
Dec 11, 2024	Gemini 2	LLM	Google DeepMind
Dec 12, 2024	Sora	Text-to video	OpenAI
Dec 13, 2024	Global MMLU	Dataset	Cohere
Dec 20, 2024	o3 (beta)	Multimodal	OpenAI
Dec 27, 2024	DeepSeek-V3	LLM	DeepSeek
Feb 3, 2025	Deep research	Multimodal	OpenAI
Feb 5, 2025	Gemini 2.0 Flash	LLM	Google DeepMind
Feb 6, 2025	Le Chat	LLM	Mistral
Feb 18, 2025	Grok-3	Chatbot	xAI
Feb 24, 2025	Claude 3.7 Sonnet	LLM	Anthropic
Feb 27, 2025	GPT 4.5	LLM	OpenAI
Mar 4, 2025	Aya Vision	Multimodal	Cohere
Mar 25, 2025	Gemini 2.5 Pro	LLM	Google DeepMind
Apr 16, 2025	o3, o4-mini/high	Multimodal	OpenAI
Apr 17, 2025	Gemini 2.5 Flash	LLM	Google DeepMind
Apr 22, 2025	Gemini Veo 2	LLM, Text-to-Video	Google DeepMind
Apr 30, 2025	Nova Pro	Foundation model	Amazon
May 1, 2025	Mellum	LLM	JetBrain
May 30, 2025	Veo 3	LLM, Text-to-Video	Google DeepMind

Source: Stanford AI Index Report 2025, Company Websites

Visit Site Register

Table 14: Leading vendors: GenAI

Vendors	Country	Expertise
OpenAI	US	Develops advanced language models, including GPT, for generative AI tasks.
Microsoft	US	Provides AI tools and cloud services with a focus on enterprise AI solutions.
AWS	US	Offers cloud-based AI services, including machine learning and NLP models.
Google	US	Develops AI and machine learning technologies, including language models like BERT.
Anthropic	US	Focuses on developing safe and steerable AI models, with an emphasis on alignment.
AI21 labs	Israel	Builds advanced language models and generative AI solutions for enterprises
Cohere	Canada	Develops private, scalable AI solutions with a focus on natural language processing
Alibaba Cloud	China	Provides AI services and cloud computing infrastructure, including NLP models
Baidu	China	Develops AI solutions with a focus on natural language processing and autonomous systems
Aleph Alpha	Germany	Specializes in advanced AI research and development of language models
Meta	Llama	An open-source AI model that can be customized and deployed based on user requirements.
Hugging Face	US	Focuses on providing open-source AI models and tools for natural language processing.
Synthesia	UK	Offer a powerful set of tools for fast, professional video creation.
Guidde	US	Helps teams create and share video-based documentation quickly and easily.
DeepSeek	China	Focused on developing large language models (LLMs).
Perplexity AI	US	Combines traditional web search with large language models to deliver conversational answers, complete with source citations.

Source: Company Websites

Visit Site Register

Appendix

Profiles of leading GenAI vendors

OpenAI

Founding year: 2015
Headquarters: US
No of employees: 5,328 (Apr 2025)
CEO: Sam Altman
Revenue: US$10 Billion (2025)

OpenAI, founded in December 2015 and based in San Francisco, California, is a leading private research organization focused on developing artificial intelligence products. The company is led by CEO Sam Altman, with Bret Taylor as Chairman and Greg Brockman serving as President.

Since its inception, OpenAI has made significant strides with groundbreaking products. These include the GPT series (advanced language models like GPT-3 and GPT-4), DALL·E (a tool that creates images from text prompts), OpenAI Codex (which powers code-writing tools), and ChatGPT (a conversational AI that engages users in human-like dialogue). With these innovations, OpenAI is transforming industries such as tech, entertainment, and education, while continuing to drive AI development toward a future where it benefits all sectors of society.

OpenAI has successfully raised a total of US$21.9 billion across 10 funding rounds, with its most recent round being a Debt Financing on October 3, 2024. The company is backed by 39 investors, with notable recent contributions from Citi and JP Morgan Chase.

Microsoft

Founding year: 1975
Headquarters: US
No of employees: 228,000 (Jun 2024)
CEO: Satya Nadella
Revenue: US$245.1 Billion (Jun 2024)
Revenue Intelligent Cloud: US$105.4 Billion (Jun 2024)

Founded on April 4, 1975, by Bill Gates and Paul Allen, Microsoft is a global leader in technology, headquartered in Redmond, Washington. As a publicly traded company on the Nasdaq under the ticker MSFT, Microsoft is a key player in the information technology industry. The company's diverse range of products and services includes software development, consumer electronics, cloud computing, social networking, and video games. Notable brands and services include Windows, Microsoft 365, Azure, Xbox, LinkedIn, and GitHub.

Under the leadership of CEO Satya Nadella, Microsoft has seen significant growth, with a revenue of US$245.1 billion and a net income of US$88.1 billion in 2024. The company operates worldwide, with subsidiaries like LinkedIn, GitHub, and Skype Technologies, and a workforce of over 228,000 employees.

As of early 2025, Microsoft has acquired a total of 256 companies. Notable acquisitions include the US$68.7 billion purchase of Activision Blizzard in 2022 to strengthen its gaming division and the US$190 million acquisition of Fungible in 2023 to expand its cloud and AI capabilities. Other examples include the acquisition of Xandr and Ally.io, further boosting its portfolio in advertising technology and workforce solutions.

AWS

Founding year: 2002, (Cloud Computing – 2006)
Headquarters: US
No of employees: 1,556,000 (Dec 2024)
CEO: Matt Garman
Revenue: US$638.0 Billion (Dec 2024)
Revenue AWS: US$107.6 Billion (Dec 2024)

Amazon Web Services (AWS), founded in 2002 and a key subsidiary of Amazon, has grown to become a dominant player in the cloud computing and web services industry. In 2023, AWS generated

Visit Site Register

US$90.8 billion in revenue and US$24.6 billion in operating income. The division is known for providing comprehensive cloud solutions, including computing power, storage, and machine learning services.

AWS has several subsidiaries that help expand its services and capabilities. For instance, Annapurna Labs develops custom chips for cloud computing, improving AWS's hardware. AWS Elemental provides video processing tools for media companies to stream high-quality content. NICE Software offers data analytics and decision-making solutions. Wickr, a secure messaging platform, strengthens AWS's focus on security and privacy.

As of January 2025, the company has made a total of 145 investments, with 97 of them as lead investors. The company primarily invests in AI, cloud computing, and tech startups. Additionally, it has acquired 9 companies, with its most recent acquisition being Wickr on June 25, 2021.

Google

Founding year: 1998, (Google Cloud – 2008)
Headquarters: US
No of employees: 183,323 (Dec 2024) Google Cloud; 54,000 (Jan 2024)
CEO: Sundar Pichai, (Google Cloud – Thomas Kurian)
Revenue: US$350.0 Billion (Dec 2024)
Revenue Google Cloud: US$43.2 Billion (Dec 2024)

Founded on September 4, 1998, by Larry Page and Sergey Brin, Google is a subsidiary of Alphabet Inc., headquartered in Mountain View, California. The company offers a broad range of products and services across multiple categories. Its web-based tools include search engines like Google Search, Google Maps, and Google Drive, as well as productivity tools like Gmail and Google Docs. The company also provides advertising services through platforms such as Google Ads and AdSense, while offering communication tools like Google Meet and Google Voice.

In hardware, Google produces devices such as Pixel smartphones, Google Nest smart home products, and Fitbit wearables. Additionally, Google is heavily invested in cloud computing and AI, with products like Google Cloud and TensorFlow.

As of early 2025, Google has made 266 acquisitions, with its most recent being Cameyo on June 5, 2024. The company frequently acquires organizations in fields like AI, cloud computing, and security. Notable acquisitions include Mandiant for US$5.4 billion, Raxium for US$1 billion, and Alter for US$100 million, all in 2022. Additionally, the company has made 306 investments, with its most recent one on December 24, 2024, in Hazeltree, a treasury management solution provider.

Anthropic

Founding year: 2021
Headquarters: US
No of employees: 1,097 (2025)
CEO: Dario Amodei
Revenue: US$1.4 Billion (2025)

Anthropic is a U.S.-based AI startup founded in 2021 by former OpenAI employees, Dario and Daniela Amodei. The company is focused on advancing the safety and reliability of artificial intelligence, particularly through large language models (LLMs). Anthropic's flagship product is Claude, a family of AI models designed to prioritize safe, transparent, and human-aligned outputs.

Over the past few years, the company has secured significant investments, including US$4 billion from Amazon and US$2 billion from Google, underscoring its increasing influence in the AI sector. As of January 2025, it has raised a total of US$13.7 billion across 12 funding rounds. Other prominent investors include Ventioneers, Manhattan Venture Partners (MVP), Stackpoint Ventures, Alliance Global Partners, and TeleSoft Partners.

AI121 Labs

Founding year: 2017
Headquarters: Israel
No of employees: 268 (May 2025)
CEO: Ori Goshen
Revenue: US$35 Million (May 2025)

Visit Site Register

Founded by Amnon Shashua, Yoav Shoham, and Ori Goshen in 2017, AI21 Labs creates advanced AI systems and models to help businesses use generative AI in real-world applications. Over the years, the company has launched various products, including Wordtune Spices, a generative AI tool designed to enhance writing, and AI21 Studio, a developer platform to build various applications and services.

As of January 2025, AI21 Labs has raised a total of U$326.5 million over 8 funding rounds. The largest investments came from the Series C rounds, with US$155 million raised in August 2023 and $53 million in November 2023. The company has attracted funding from notable investors such as Comcast Ventures, Intel Capital, Samsung NEXT, and Pitango VC, to name a few.

Cohere

Founding year: 2019
Headquarters: Canada
No of employees: 796 (May 2025)
CEO: Aidan Gomez
Revenue: US$35 Million (May 2025)

Cohere is an AI company that prioritizes data security, creating scalable and private AI solutions designed to solve practical business problems. It develops AI solutions across various industries, including financial services, manufacturing, energy and utilities, and healthcare. In the financial sector, Cohere enhances efficiency by automating tasks, improving risk management, and offering real-time insights. In healthcare, it advances patient care by connecting data sources, accelerating research, and streamlining workflows for better patient outcomes.

In January 2025, Royal Bank of Canada partnered with the company to develop generative AI products for the financial industry, specifically targeting risk management and security.

The company has secured US$1.1 billion in funding across 7 rounds, with its most recent round being a Grant on December 6, 2024. The company is supported by 34 investors, including the Government of Canada and NVIDIA. Additionally, Cohere has made two investments, the latest being a US$1.5 million investment in Questflow on July 8, 2024.

Alibaba Cloud

Founding year: 1999, (Alibaba Cloud – 2009)
Headquarters: China
No of employees: 124,320 (Mar 2025)
Alibaba Cloud: 4,656 (2025)
CEO: Eddie Wu
Revenue: US$137.3 Billion (Mar 2025)
Revenue Cloud Intelligence Group: US$16.3 Billion (Mar 2025)

Launched in 2009, Alibaba Cloud is a leading global cloud computing provider and a subsidiary of Alibaba Group. The company offers a diverse range of products and services designed to meet various business needs. Among the offerings include Elastic Compute Service (ECS) for high-performance virtual servers, Object Storage Service (OSS), and Elastic GPU Service for scalable computing power. Other notable products include Web Application Firewall (WAF) for security, Cloud Enterprise Network (CEN) for seamless connectivity, and DingTalk Enterprise for team collaboration.

Among the more recently launched solutions include Secure Access Service Edge (SASE) for network security, Intelligent Media Services (IMS), and Alibaba Cloud Model Studio for AI model development. Additionally, Alibaba Cloud provides specialized services such as ApsaraDB for SelectDB and Short Message Service (SMS) for communication.

As of January 2025, Alibaba Cloud has made 9 investments, with its most recent being in a Chinese cloud operating systems provider, Sealos, on December 12, 2024. Additionally, the company has acquired 3 companies, focusing mainly on cybersecurity and tech-related acquisitions. Over time, it has raised a total of US$1.2 billion in funding across two rounds, primarily from Alibaba Group.

Visit Site Register

Baidu

Founding year: 2000
Headquarters: China
No of employees: 35,900 (Dec 2024)
CEO: Robin Li
Revenue: US$18.2 Billion (Dec 2024)
Revenue Cloud Services: US$3.0 Billion (Dec 2024)

Founded on January 18, 2000, Baidu is a leading Chinese multinational offering a wide range of products and services across various sectors, focusing on internet services, AI, and cloud computing. Its mobile ecosystem includes the Baidu App, Haokan (short video platform), and Quanmin (flash video app), while knowledge-based platforms like Baidu Encyclopedia and Baidu Knows provide expert-driven content and user-generated Q&A.

The company also delivers AI-driven services like Smart Mini Programs, Baijiahao, and Managed Pages for businesses. In addition, Baidu's intelligent driving division, led by the Apollo platform, is a market leader in autonomous driving technology in China. Other offerings include Baidu Health for healthcare services and DuerOS, a smart assistant platform.

As of January 2025, Baidu has raised a total of US$26.2 million over three funding rounds. The company is supported by nine investors, with Venture TDF and ePlanet Capital being the most recent. Additionally, the company has made 128 investments, 3 diversity investments, and 32 exits, with notable acquisitions including healthcare data provider, GBI, in February 2023.

Aleph Alpha

Founding year: 2019
Headquarters: Germany
No of employees: 298 (2025)
CEO: Jonas Andrulis
Revenue: US$14.7 Million (2025)

Aleph Alpha GmbH is a pioneering German AI startup founded by Jonas Andrulis and Samuel Weinbach. The company has recently launched its next-generation Control-Models, designed to provide more human-like interaction and solve complex tasks using large language models. These models are equipped with enhanced natural language processing, making them perfect for applications like chatbots and digital assistants. Additionally, they feature Explainable AI technology, enabling traceability and verification of AI-generated content. This breakthrough ensures transparency, reduces hallucinations, and supports compliance with upcoming EU regulations. Through these innovations, Aleph Alpha combines high performance, trust, and efficiency, setting a new benchmark in generative AI.

As of January 2025, the company has raised a total of US$533.6 million across 6 funding rounds. The largest round was a Series B in November 2023, securing US$500 million, with lead investors Bosch Ventures, Innovation Park Artificial Intelligence, and Schwarz Group. Previous rounds include a Series A in 2021 with US$25.4 million and a seed round in the same year with US$5.83 million. The latest funding was a Secondary Market round in November 2024. The company has 15 investors, with Schwarz Group and Burda Principal Investments being the most recent.

Meta

Founding year: 2004
Headquarters: US
No of employees: 74,067 (Dec 2024)
CEO: Mark Zuckerberg
Revenue: US$164.5 Billion (Dec 2024)

Meta has positioned itself as a major global investor in artificial intelligence, allocating around US$40 billion annually towards AI and virtual reality research. This significant investment underscores their commitment to pushing the boundaries of digital interaction.

A key product of this investment is the Meta AI chatbot, launched in late 2023 and integrated into WhatsApp, Instagram, and Facebook Messenger.

Visit Site Register

This chatbot offers contextual understanding, multilingual communication, image generation, and real-time information processing to provide conversational assistance and creative support to users.

Furthermore, Meta is actively developing generative AI tools like AI Image Editing, AI Studio for custom AI characters, and experimental Text-to-Video generation. The company has also partnered with Google Cloud to offer its Llama models.

Hugging Face

Founding year: 2016
Headquarters: US
No of employees: 534 (2025)
CEO: Clément Delangue
Revenue: US$46.8 Million (2025)

Hugging Face, headquartered in New York City, offers open-source tools for machine learning, with a primary focus on natural language processing (NLP). Renowned for its popular transformers library, the platform enables users to build, train, and share machine learning models, datasets, and projects.

As of January 2025, the company has raised US$395.2 million over 7 funding rounds, with the latest Series D round on January 16, 2024. The company is supported by 38 investors, including Bossa Invest and PremjiInvest. In addition, the company has made 4 acquisitions, with the most recent being XetHub on August 8, 2024. The company frequently acquires organizations in areas related to machine learning, natural language processing, and AI tools. Notable acquisitions include Argilla in June 2024, Gradio in December 2021, and Sam in September 2017.

Synthesia

Founding year: 2017
Headquarters: UK
No of employees: 511 (Jun 2025)
CEO: Victor Riparbelli
Revenue: US$35 Million (Jun 2025)

Founded by Lourdes Agapito, Matthias Niessner, Steffen Tjerrild, and Victor Riparbelli, Synthesia is a UK-based leader in AI-driven video creation technology. The company's platforms offer a powerful set of tools for fast, professional video creation. For instance, its AI video editor and screen recorder make it easy to produce and edit content directly in the browser, while brand kits and a centralized media library help maintain consistency. Moreover, users can choose from over 230 AI avatars—including personal and selfie avatars—and generate voiceovers in 140+ languages with options like voice cloning. Localization is simple with one-click translations, AI dubbing, and closed captions. Built-in features like review workflows, live collaboration, and version control support efficient team production and feedback.

As of 2025, Synthesia has raised US$$336.6 million across seven funding rounds, with major investments from top-tier firms including Accel, Kleiner Perkins, NEA, FirstMark, Seedcamp, and Adobe Ventures. Its most recent Series D round in January 2025 brought in US$180 million, led by New Enterprise Associates.

Guidde

Founding year: 2020
Headquarters: US
No of employees: 52
CEO: Yoav Einav
Revenue: US$5.5 Million (2025)

Guidde is a California-based startup, founded in 2020, that helps teams create and share video-based documentation quickly and easily. With AI-powered tools like Guidde Create and Guidde Broadcast, users can capture workflows with one click, generate step-by-step guides, and share professional tutorials in minutes. The platform supports over 100 languages, includes smart editing tools, and ensures content security with automatic blurring. The platform is widely used by customer success, product, and pre-sales teams to improve onboarding, reduce support tickets, and boost productivity.

Visit Site Register

The company has raised a total of US$26.6 million across four funding rounds to support its growth and innovation. The company's funding journey began with seed investments in 2021, including backing from Entrée Capital. In 2023, it secured US$11.6 million in a Series A round led by Norwest Venture Partners, followed by another Series A round in early 2025, raising US$15 million from Qualcomm Ventures and other investors.

DeepSeek

Founding year: 2023
Headquarters: China
No of employees: 200 (Jan 2025)
CEO: Liang Wenfeng
Revenue: US$200 Million (2024)

DeepSeek is a Chinese AI company based in Hangzhou, Zhejiang, focused on developing large language models (LLMs). It was founded in July 2023 by Liang Wenfeng. The company gained widespread prominence in January 2025 when it released its own AI chatbot and the DeepSeek-R1 model.

Its lineup includes advanced LLMs like DeepSeek R1, V2, and V3, along with specialized tools such as DeepSeek Coder for programming, DeepSeek Math for solving mathematical problems, and DeepSeek VL for vision and language tasks. These models are accessible through the DeepSeek App, DeepSeek Chat, and the DeepSeek Platform, with users able to integrate the technology via API. The company also provides service performance data to help users monitor reliability and usage.

The company has experienced rapid growth, reaching 100 million users within just 14 days of launch, making it one of the fastest-growing platforms in history.

Perplexity AI

Founding year: 2022
Headquarters: US
No of employees: 1,292 (2025)
CEO: Aravind Srinivas
Revenue: US$100 Million (2024)

Perplexity AI is an American AI-powered search engine company, founded in 2022 by Aravind Srinivas, Denis Yarats, Johnny Ho, and Andy Konwinski. Headquartered in San Francisco, the company combines traditional web search with large language models to deliver conversational answers, complete with source citations. Users can ask follow-up questions, making the experience feel more like a dialogue than a standard search.

The platform launched on December 7, 2022, and is available via web, Google Chrome extension, and mobile apps for iOS and Android. It uses Microsoft Bing for search results and runs on Microsoft Azure. The free version is powered by OpenAI's GPT-3.5, while the Pro subscription offers access to more advanced models, including GPT-4.

The company has raised a total of US$665 million across five funding rounds, with major investors including NVIDIA, IVP, SoftBank Vision Fund, NEA, and Bessemer Venture Partners. In its most recent round in December 2024, the company secured US$500 million and reached a valuation of US$9 billion, making it one of the fastest-growing startups in AI search.

In May 2025, the company announced a partnership with PayPal to enable in-chat payments, letting U.S. users make purchases like travel and event tickets directly through its AI chat, marking a move toward AI-driven e-commerce.

Visit Site Register

GENERATIVE AI WEEK

November 11 – 12 2025 | Sheraton Austin Georgetown, Texas

North America's premier event for scaling Generative AI in the enterprise

Generative AI is transforming industries, but how do you move from experimentation to enterprise-wide impact? Generative AI Week 2025 is North America's premier summit for business and technology leaders seeking to operationalize, implement, and scale Generative AI solutions with measurable business value. This exclusive two-day event brings together top AI and data executives from leading enterprises, to share insights on how they are successfully embedding AI into their operations. Whether you are exploring AI adoption or optimizing existing AI strategies, #GenAIWeek 2025 provides the roadmap, expertise, and real-world case studies you need to drive transformation at scale.

DOWNLOAD THE AGENDA

Visit Site Register

Bibliography

Style followed:

Author's Last Name, First Name. "Page Title." Website Name. Month Day, Year. URL.

1. Allen, Leanne, Höck, Benedikt and Clamp, Adrian. "A blueprint for creating value through AI-driven transformation." KPMG. **https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2025/02/intelligent-banking-report.pdf.**

2. Artificialanalysis AI. "LLM Leaderboard - Comparison of GPT-4o, Llama 3, Mistral, Gemini, and over 30 models." **https://artificialanalysis.ai/leaderboards/models.**

3. Bailyn, Evan. "Top Generative AI Chatbots by Market Share – May 2025." Firstpagesage. May 9, 2025. **https://firstpagesage.com/reports/top-generative-ai-chatbots/.**

4. Bajpai, Rahul, Tiwari, Arpan and Sarer, Baris. "The future of Edge AI." Deloitte. **https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology-media-telecommunications/deloitte-the-future-of-edge-ai.pdf.**

5. Bantourakis, Minos and Venturini, Francesco. "The impact of GenAI on the creative industries, and the ethics and governance we must put in place." Weforum. January 21, 2025. **https://www.weforum.org/stories/2025/01/the-impact-of-genai-on-the-creative-industries/.**

6. BCG. "How Digital and AI Will Reshape Health Care in 2025." January 2025. **https://web-assets.bcg.com/8c/f8/ae51ffb44ca59cb8abd751940441/bcg-how-digital-and-ai-solutions-will-reshape-health-care-in-2025.pdf.**

7. Belcic, Ivan and Stryker, Cole. "AI agents in 2025: Expectations vs. reality." IBM. March 04, 2025. **https://www.ibm.com/think/insights/ai-agents-2025-expectations-vs-reality.**

8. Bobier, Jean-François, Chatterjee, Abhik, and Ebeling, Ruth. "The CIO's Role in AI Value Creation." BCG. February 19, 2025. **https://www.bcg.com/publications/2025/cios-role-in-ai-transformation-and-productivity.**

9. Business Wire. "First Real-World Multisite Study Shows GenAI-Powered Mental Health Treatment Outperforms Standard of Care." March 10, 2025. **https://www.businesswire.com/news/home/20250310848349/en/First-Real-World-Multisite-Study-Shows-GenAI-Powered-Mental-Health-Treatment-Outperforms-Standard-of-Care.**

10. Capgemini. "Capgemini accelerates enterprise adoption of agentic AI for industries with NVIDIA." Mar 19, 2025. **https://www.capgemini.com/news/press-releases/capgemini-accelerates-enterprise-adoption-of-agentic-ai-for-industries-with-nvidia/.**

11. CB Insights. "State of AI." **https://www.cbinsights.com/reports/CB-Insights_Artificial-Intelligence-Report-2024.pdf.**

12. Cengage Group. "GenAI in Higher Education - Positive Sentiment Builds with Rapid Transformation." April 07, 2025. **https://www.cengagegroup.com/news/perspectives/2025/genai-in-higher-education---positive-sentiment-builds-with-rapid-transformation/.**

13. Cetin, Enver. "Agentic AI and the Future of Personalized Healthcare." Ciklum. May 19, 2025. **https://www.ciklum.com/resources/blog/future-of-personalized-healthcare.**

14. Chandrasekaran, Arun. "3 Bold and Actionable Predictions for the Future of GenAI." Gartner. April 12, 2024. **https://www.gartner.com/en/articles/3-bold-and-actionable-predictions-for-the-future-of-genai.**

15. Cognizant. "Generative AI in the transportation and logistics industry." **https://www.cognizant.com/en_us/industries/documents/generative-ai-in-transport-logistics-industry.pdf.**

16. Coshow, Tom, and Gao, Arnold. "Top Strategic Technology Trends for 2025: Agentic AI." Gartner. October 21, 2024. **https://www.gartner.com/doc/reprints.**

17. Dataspan. "GenAI in Manufacturing: 7 Real-World Use Cases." December 27, 2024. **https://www.dataspan.ai/blog/7-use-cases-of-genai-in-manufacturing.**

18. Dilmegani, Cem. "Generative AI in Retail: Use Cases, Examples & Benefits in 2025." AI Multiple Research. May 05, 2025. **https://research.aimultiple.com/generative-ai-in-retail/.**

19. Dilmegani, Cem. "Top 10 Use Cases of Generative AI in Education in 2025." AI Multiple Research. May 06, 2025. **https://research.aimultiple.com/generative-ai-in-education/.**

20. Dilmegani, Cem. "Agentic AI: 8 Use Cases & Real-life Examples in 2025." AI Multiple Research. April 28, 2025. **https://research.aimultiple.com/agentic-ai/.**

21. Done for you. "AI Model Comparison: Which AI Reigns Supreme in 2025?." **https://doneforyou.com/ai-model-comparison-which-ai-reigns-supreme-in-2025/.**

Visit Site Register

22. Drut. "The Future of AI Infrastructure: Trends to Watch in 2025." February 19, 2025. https://drut.io/drut-blog/f/the-future-of-ai-infrastructure-trends-to-watch-in-2025.

23. Dudley, Brian, and DelMastro, Thomas. "The Next Frontier: The Rise of Agentic AI." Adams Street Partners. March 12, 2025. https://www.adamsstreetpartners.com/insights/the-next-frontier-the-rise-of-agentic-ai/.

24. Duk, Vitalii. "Generative AI: Practical Ways for Enterprises to Cut Costs and Boost Sales." Get Dynamiq. February 06, 2025. https://www.getdynamiq.ai/post/generative-ai-practical-ways-for-enterprises-to-cut-costs-and-boost-sales.

25. Endemano, Mark and Brien, Catherine. "AI in Creative Industries: Enhancing, rather than replacing, human creativity in TV and film." Alix Partners. January 10, 2025. https://www.alixpartners.com/insights/102jsme/ai-in-creative-industries-enhancing-rather-than-replacing-human-creativity-in/.

26. ESA Automation. "Generative AI Powers Smart Manufacturing." March 27, 2025. https://www.esa-automation.com/en/generative-ai-powers-smart-manufacturing/.

27. Fernandez, Joaquin. "The leading generative AI companies." IOT Analytics. March 04, 2025. https://iot-analytics.com/leading-generative-ai-companies/.

28. Garcia, Cyril, Charpiot, Vincent and Andrillon, Florent. "Developing sustainable Gen AI." Capgemini. https://www.capgemini.com/dk-en/wp-content/uploads/sites/7/2025/02/Final-Web-Version-Report-Sustainable-Gen-AI-2-1.pdf.

29. Gartner. "Gartner Experts Answer the Top Generative AI Questions for Your Enterprise." https://www.gartner.com/en/topics/generative-ai.

30. Gaus, Tim. "Beyond automation: How Generative AI is redefining manufacturing." Deloitte. April 22, 2025. https://www2.deloitte.com/us/en/blog/business-operations-room-blog/2025/generative-ai-in-manufacturing.html.

31. Goldman Sachs. "Gen Ai: Too Much Spend, Too Little Benefit?." June 25, 2024. https://www.goldmansachs.com/images/migrated/insights/pages/gs-research/gen-ai--too-much-spend,-too-little-benefit-/TOM_AI%202.0_ForRedaction.pdf.

32. Gough, Jonathan D. "Top 10 Agentic AI Examples and Use Cases." Converge Technology Solutions. May 06, 2025. https://convergetp.com/2025/05/06/top-10-agentic-ai-examples-and-use-cases/.

33. Hitchcock, Larry, Garza, Mauricio, and Crowley, Eileen. "Gen AI transforming transportation: Lessons from the frontier of an emerging technology." Deloitte. November 21, 2024. https://www2.deloitte.com/us/en/insights/focus/transportation/ai-in-transportation.html.

34. Hörmann, Fabiane. "Generative artificial intelligence takes Siemens' predictive maintenance solution to the next level." Siemens. February 05, 2024. https://press.siemens.com/global/en/pressrelease/generative-artificial-intelligence-takes-siemens-predictive-maintenance-solution-next.

35. Intellias. "Generative AI in Retail: Use Cases, Examples, and Implementation." March 31, 2025. https://intellias.com/generative-ai-in-retail/.

36. ISG. "Enterprise Spending on GenAI Expected to Rise 50% in 2025, as Focus Shifts From Efficiency to Expertise." September 23, 2024. https://ir.isg-one.com/news-market-information/press-releases/news-details/2024/Enterprise-Spending-on-GenAI-Expected-to-Rise-50-in-2025-as-Focus-Shifts-From-Efficiency-to-Expertise/default.aspx.

37. Kaur, Jagreet. "Building Chatbots with Agentic AI." Xenonstack. April 03, 2025. https://www.xenonstack.com/blog/chatbot-agentic-ai.

38. Kerner, Sean Michael. "25 of the best large language models in 2025." Techtarget. January 31, 2025. https://www.techtarget.com/whatis/feature/12-of-the-best-large-language-models.

39. Korolov, Maria. "As AI scales, infrastructure challenges emerge." CIO. October 23, 2024. https://www.cio.com/article/3577669/as-ai-scales-infrastructure-challenges-emerge.html.

40. Kudumala, Aditya, Israel, Adam and Lella, Sai. "Realizing Transformative Value from AI & Generative AI in Life Sciences." Deloitte. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/us-realizing-transformative-value-from-AI-GenAI-in-life-sciences-032124.pdf.

41. Laddagi, Navin. "Top 5 Reasons Why Enterprises Need To Own Their GenAI Platform in 2025." Quantiphi. February 04, 2025. https://quantiphi.com/top-5-reasons-why-enterprises-need-their-own-

42. Lawton, George. "8 top generative AI tool categories for 2025." TechTarget. January 07, 2025. https://www.techtarget.com/searchenterpriseai/tip/Top-generative-AI-tool-categories.

43. Lee, Wai Yee. "Empowering Future Manufacturing: AI and Operational Technologies for 2025 and Beyond." IDC. February 10, 2025. https://blogs.idc.com/2025/02/10/empowering-future-manufacturing-ai-and-operational-technologies-for-2025-and-beyond/.

Visit Site Register

44. LoDolce, Matt, and Moran, Meghan. "Gartner Forecasts Worldwide GenAI Spending to Reach $644 Billion in 2025." Gartner. March 31, 2025. **https://www.gartner.com/en/newsroom/press-releases/2025-03-31-gartner-forecasts-worldwide-genai-spending-to-reach-644-billion-in-2025.**

45. Lucente, Ida. "Generative AI in Healthcare: Use Cases, Benefits, and Challenges." John Snow Labs. May 22, 2025. **https://www.johnsnowlabs.com/generative-ai-healthcare/.**

46. Maniar, Shweta. "How GenAI will transform life sciences in 2025." Pharma Phorum. January 07, 2025. **https://pharmaphorum.com/digital/how-genai-will-transform-life-sciences-2025.**

47. Markham, Isobel. "How WestRock Harnessed GenAI to Enhance Internal Audit." Deloitte. March 23, 2024. **https://deloitte.wsj.com/riskandcompliance/how-westrock-harnessed-genai-to-enhance-internal-audit-f0926363.**

48. Martin,Carlos Pardo, Lamb, Jessica and Dahab, Amine. "Generative AI in healthcare: Current trends and future outlook." Mckinsey. March 26, 2025. **https://www.mckinsey.com/industries/healthcare/our-insights/generative-ai-in-healthcare-current-trends-and-future-outlook.**

49. Miglio, Andrea Del, Giovine, Carlo, and Hauser, Stephanie. "Banking on innovation: How ING uses generative AI to put people first." Mckinsey. **https://www.mckinsey.com/industries/financial-services/how-we-help-clients/banking-on-innovation-how-ing-uses-generative-ai-to-put-people-first.**

50. Morrison, Paul. "Retail 2025: 6 Trends Re-defining the Future of Shopping." WNS. **https://www.wns.com/perspectives/articles/retail-2025-6-trends-re-defining-the-future-of-shopping.**

51. Noffsinger, Jesse, Patel, Mark and Sachdeva, Pankaj. "The cost of compute: A $7 trillion race to scale data centers." Mckinsey. April 28, 2025. **https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers.**

52. NTT Data. "A 'Complete Revolution' in manufacturing: NTT DATA research reveals GenAI's transformative potential and impact on core functions." May 01, 2025. **https://www.nttdata.com/global/en/news/press-release/2025/may/050100.**

53. Petruk, Maksym. "How to Compare AI Models from OpenAI, Google, and More." We Soft You. July 01, 2024. **https://wesoftyou.com/ai/how-to-compare-ai-models/.**

54. Pratt, Mary K. "10 real-world agentic AI examples and use cases." TechTarget. Mar 07, 2025. **https://www.techtarget.com/searchenterpriseai/feature/Real-world-agentic-AI-examples-and-use-cases.**

55. Pratt, Mary K. "The future of generative AI: 10 trends to follow in 2025." TechTarget. February 04, 2025. **https://www.techtarget.com/searchenterpriseai/feature/The-future-of-generative-AI-Trends-to-follow.**

56. Ramamurthy, Shanker, and Sironi, Paolo. "2025 Global Outlook for Banking and Financial Markets." IBM. **https://www.ibm.com/downloads/documents/us-en/115dcc7faf363f21.**

57. Riemer, Stiene, Coppola, Matteo, and Rogg, Jürgen. "For Banks, the AI Reckoning Is Here." BCG. May, 2025. **https://web-assets.bcg.com/3e/6f/9dfa63434eb7a00e1cf1cdcb3754/for-banks-the-ai-reckoning-is-here-may-2025.pdf.**

58. Robbins, Jacob. "Meet the 10 most active investors in generative AI." Pitchbook. June 12, 2024. **https://pitchbook.com/news/articles/top-generative-ai-vc-investors-list.**

59. Sai, Moguloju. "ChatGPT vs Gemini AI Pro vs Llama vs Copilot vs DeepSeek R1." Medium. February 07, 2025. **https://medium.com/@saimoguloju2/chatgpt-vs-gemini-ai-pro-vs-llama-vs-copilot-vs-deepseek-r1-9ce268b3492d.**

60. SGU. "A Comparison of Leading AI Models: DeepSeek AI, ChatGPT, Gemini, and Perplexity AI." February 07, 2025. **https://sgu.ac.id/a-comparison-of-leading-ai-models-deepseek-ai-chatgpt-gemini-and-perplexity-ai/.**

61. Sharma, Suraj. "9 Ways Generative AI in Transportation is Enhancing the Sector." Nextgen Invent. **https://nextgeninvent.com/blogs/generative-ai-in-transportation-enhancing-the-sector/.**

62. Shubham. "Agentic AI: An Introduction to Autonomous Intelligent Systems." Learn Open CV. February 11, 2025. **https://learnopencv.com/agentic-ai/.**

63. Singla, Alex, Sukharevsky, Alexander, and Yee, Lareina. "The state of AI: How organizations are rewiring to capture value." Mckinsey. March 12, 2025. **https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.**

64. Stratt, Luri "GenAI funding hits record in 2024 boosted by infrastructure interest." SP Global. January 22, 2025. **https://www.spglobal.com/market-intelligence/en/news-insights/articles/2025/1/genai-funding-hits-record-in-2024-boosted-by-infrastructure-interest-87132257.**

65. Talkai Info. "A Comparative Analysis of the Best Language Models: ChatGPT, Gemini, Claude, and Llama." https://talkai.info/blog/comparative_analysis_of_chatgpt_gemini_claude_llama/.

66. Tully, Tim, Redfern, Joff, and Xiao, Derek. "2024: The State of Generative AI in the Enterprise." Menlovc. November 20, 2024. https://menlovc.com/2024-the-state-of-generative-ai-in-the-enterprise/.

67. Tyrone. "Designing Composable GPU Workspaces in Multi-Tenant Environments: A Blueprint for Agile AI Infrastructure." March 31, 2025. https://blog.tyronesystems.com/designing-composable-gpu-workspaces-in-multi-tenant-environments-a-blueprint-for-agile-ai-infrastructure/.

68. Vals AI. "GPQA Benchmark." March 26, 2025. https://www.vals.ai/benchmarks/gpqa-03-26-2025.

69. Virtasant. "AI in Creative Industries: End of Creativity as We Know It?." April 22, 2025. https://www.virtasant.com/ai-today/ai-in-creative-industries-end-of-creativity-as-we-know-it.

70. Warren, Zach, Abbott, Mike, and Leach, Lucy. "2025 Generative AI in Professional Services Report." Thomson Reuters. https://www.thomsonreuters.com/content/dam/ewp-m/documents/thomsonreuters/en/pdf/reports/2025-generative-ai-in-professional-services-report-tr5433489-rgb.pdf.

71. Zimmerman, Vicktery. "Deloitte Global's 2025 Predictions Report: Generative AI: Paving the Way for a Transformative Future in Technology, Media, and Telecommunications." Deloitte. November 19, 2024. https://www.deloitte.com/global/en/about/press-room/deloitte-globals-2025-predictions-report.html.

Visit Site Register