AI & GenAI Model Provider Market Landscape

SPOTLIGHT REPORT

May 2025

AI & GenAI Model Provider Market Landscape Overview

Since Microsoft's strategic investment in high-profile startup OpenAI and the subsequent evolution of ChatGPT, generative AI (GenAI) has taken the technology market by storm and is now a focus of all vendors covered by TBR. TBR's *AI & GenAI Model Provider Market Landscape* focuses on some of the more influential AI startups, including OpenAI, that are making GenAI a reality for many enterprises as well as their cloud delivery partners, which play a critical role in this new market.

This research also includes analysis of alliance relationships, specifically how AI vendors (both large language model [LLM] providers and GenAI facilitators) are working with the major hyperscalers and SaaS vendors and where AI startups are investing; key trends, such as the emergence of multimodal models; and what we can expect from these vendors in the coming quarters.

Publish date of latest edition: April 18, 2025

Click here to view a full list of the report's research topics and vendor coverage.

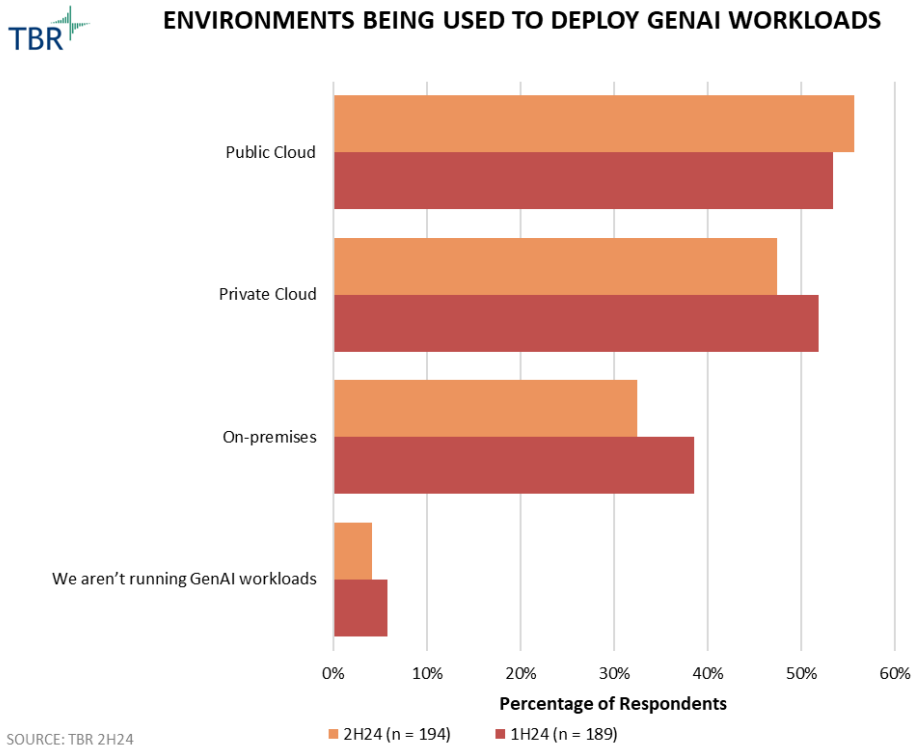
"Innovation in AI models is progressing rapidly, but the improvements require greater compute intensity. Cost-optimization will become an important consideration in the technology's next chapter, addressed by deeper looks into smaller models and Mixture of Experts architectures."

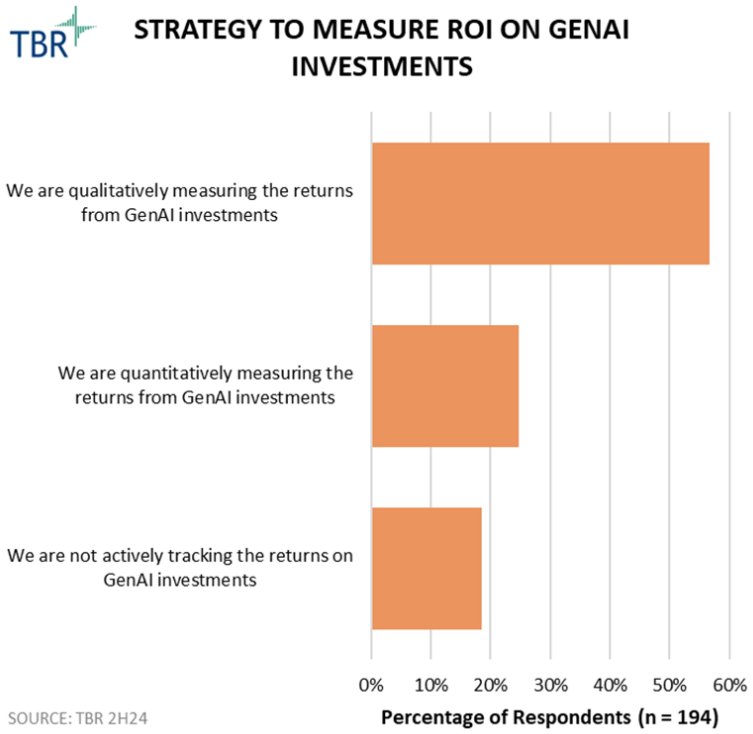
— Analyst Alex Demeule

Executive Summary Excerpt

Interest in AI capabilities has not waned as enterprises view the technology as critical to long-term competitive positioning

The buzz around GenAI persists as enterprise interest is leading to adoption. Yet it is still early days, and many enterprises remain in exploration mode. Some use cases, such as data management, customer service, administrative tasks and software development, have already moved from the proof-of-concept stage to production. Still, the exploration phase of AI adoption will be a slow burn as enterprises seek opportunities beyond these low-hanging fruit. As seen in the graph to the right, most enterprises are evaluating AI qualitatively, forgoing quantitative measures to keep up with peers based upon the assumption that the technology will bring transformational improvement to business operations.



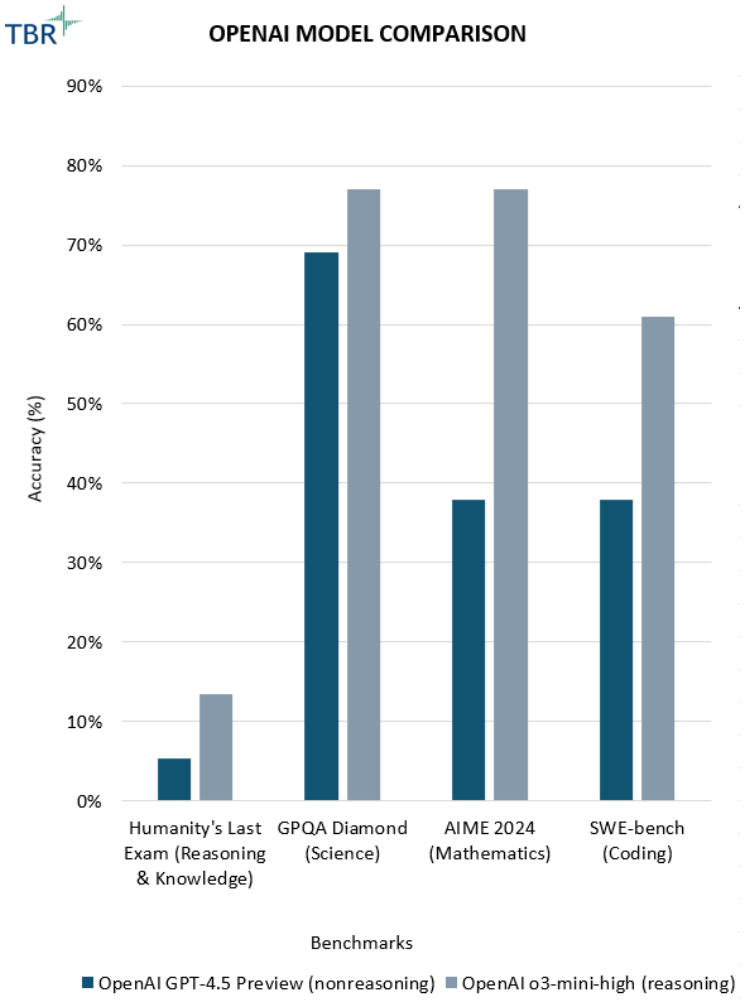


Reasoning models excel at performing complex, deterministic tasks, and have become the most popular models at the back end of agentic AI

The capability improvement brought by the iterative inferencing process has made reasoning models the focal point of frontier model research. In fact, most of the models sitting atop established third-party benchmarks are reasoning models, except for OpenAI's GPT-4.5, which the company stated would be its last nonreasoning LLM. Put simply, the difference in output quality is too pronounced to ignore, especially regarding complex, deterministic tasks. As seen in the graph, reasoning models outperform their nonreasoning predecessors across the board, with the greatest distinction appearing in coding and math benchmarks. The strength in complex, deterministic tasks makes reasoning models particularly adept at powering agentic AI capabilities, offering a wider range of addressable use cases and greater accuracy. In addition, reasoning frameworks can be leveraged at any parameter count, with available reasoning models ranging from fewer than 10 billion parameters to more than 100 billion.

As SaaS vendors continue to build proprietary, domain-specific SLMs [small language models] to power their agentic capabilities, incorporating reasoning frameworks will be an important part of their development strategies. Although the capabilities of reasoning models are impressive, the models bring new challenges and are not necessarily the best choice for every application.

Simple content generation and summarization, for instance, do not necessarily require iterative inferencing. Moreover, the greater compute intensity caused by repeated processing at the transformer layer will compound existing challenges to scaling AI adoption. Not only will these models be more expensive to run for the customer, but they will also exacerbate the persistent supply shortages facing cloud infrastructure providers. Microsoft has noted infrastructure constraints as a headwind to AI revenue growth in the past several quarters, and the emerging need for test-time compute adds to these infrastructure demands. As discussed in TBR's special report, Sheer Scale of GTC 2025 Reaffirms NVIDIA's Position at the Epicenter of the AI Revolution, NVIDIA's CEO Jensen Huang stated that reasoning AI consumes 100 times more compute than nonreasoning AI. Of course, this was a highly self-serving statement, as NVIDIA is the leading provider of GPUs powering this compute, but we are dealing with magnitudes of difference. For the use of reasoning models to continue scaling, this high compute intensity will need to be addressed.



SOURCE: THIRD-PARTY BENCHMARKS

Go-to-market Dynamics Excerpt

Model providers that have successfully productized their models are seeing the greatest revenue gains, while growth remains more limited for the rest of the pack

Model providers

Most model providers covered in this report are heavily reliant on third-party capital to sustain their operations, with only a select few establishing substantial revenue streams from the enterprise. The vendors that are generating meaningful revenue are doing so primarily through the consumer market. Given the broad popularity of ChatGPT, OpenAI offers the best example.

Today, the vendors generating the most revenue through their AI models have done so by productizing their models. Microsoft CEO [Satya] Nadella recently made headlines by calling OpenAI a product company, not a model company. Of course, OpenAI sells access to its models through Model as a Service engagements via Microsoft and Oracle, but the company is generating more revenue through consumer and commercial use of online tools like ChatGPT.

While Nadella's comment is apt, TBR suspects this is a byproduct of where the technology sits today. ChatGPT offers out-of-the-box capabilities accessible instantly by anyone, making it easy to monetize quickly. This is hardly the only opportunity, however. Enterprises are folding AI into their existing digital transformation journeys, but broad adoption will take time to manifest as the technology matures. At some point, licensing fees associated with custom AI development within the enterprise will become a strong revenue driver for model developers, setting these companies up well to drive diversified growth.

Infrastructure Providers

After NVIDIA and the chip providers, the hyperscalers have excelled at monetizing the AI opportunity. These vendors — AWS [Amazon Web Services], Microsoft (Azure), Google (Google Cloud Platform) and Oracle — have already seen meaningful contributions to cloud growth via training workloads, and they are set to see inference workloads scale over the coming year. Microsoft leads the pack, generating $13 billion of annualized revenue through its AI business.

As the revenue opportunity becomes more balanced across training and inference workloads, the revenue-sharing agreements held between hyperscalers and model developers resemble those governing ISVs selling through the cloud marketplace. The cloud infrastructure provider captures consumption-based revenue for the compute used to run the model, as well as the Model as a Service offering, while the model developer captures a separate consumption-based fee.

SaaS

The cloud application vendors are all-in on agentic AI and are monetizing the technology using a consumption model similar to that employed by the hyperscalers. Salesforce, for instance, charges around $2 per conversation within Agentforce. Overall, cloud application vendors have been slower to monetize their AI investments compared to the hyperscalers.

Salesforce's AI-related revenue is estimated at roughly $200 million to $300 million in annualized revenue, compared to Microsoft's $13 billion. The difference can be attributed to the fact that application providers — unlike hyperscalers — are not able to capture revenue opportunities in AI training, barring them from early monetization.

Still, AI adoption is reaching a point where operational inferencing has begun, and TBR expects cloud application vendors to see their AI-related revenues grow rapidly as inferencing scales.

Ecosystem Developments Excerpt

SaaS vendors will need to get on board with the new Model Context Protocol to ensure customers can use their model of choice

SaaS Vendor Strategy Assessment

From a strategic positioning perspective, TBR does not expect the rising popularity of the Model Context Protocol to have an outsized impact, primarily because we anticipate all application vendors will adopt the framework to ensure customers can leverage the model of their choice. Furthermore, cloud application vendors are positioned to benefit from the standardization of API calls between models and their workloads. Through a standardized API calling framework, these vendors will be better positioned to drive cost optimization and improve workload management for embedded AI tools.

Recent Developments

The Model Context Protocol is becoming the standard: The idea of the Model Context Protocol (MCP) has been steadily gaining popularity following its release by Anthropic in November 2024. At its core, MCP aims to address the emerging challenge of building dedicated API connectors between LLMs and applications by introducing an abstraction layer that standardizes API integrations. This abstraction layer — commonly referred to as the MCP server — would establish a default method for LLM function calling, which software providers would need to incorporate into their applications to access LLMs.

This standardization offers several benefits for model vendors, such as eliminating the need to build individual connectors for each service and promoting a modular approach to AI service integration, potentially unlocking long-term advantages in areas such as workload management and cost optimization.

For SaaS vendors, there is little reason to resist the shift toward MCP, and its growing popularity may make adoption inevitable. Application vendors like Microsoft and ServiceNow have already begun implementing the protocol by establishing MCP servers for the Copilot suite and Now Assist, respectively, and TBR expects other vendors to follow.

It is important to recognize, however, that this approach better suits vendors taking a model-agnostic stance — meaning they aim to empower enterprises to use any LLM to automate agentic capabilities. A possible exception lies with vendors that are less model-agnostic. For instance, Salesforce's emphasis on proprietary models reduces the need for MCP and favors the company's focus on native connectors between Customer 360 workflows and xGen models.

Ultimately, TBR expects Salesforce to adopt MCP, but there is an important distinction in how different SaaS vendors may approach standardization. Today, the BYOM [bring your own model] philosophy remains a priority for Salesforce, but if the company were to eventually push customers to use its proprietary models exclusively with Customer 360, its commitment to MCP could be deprioritized in favor of tighter customer lock-in.

Vendor Profiles Excerpt

Google enhances AI capabilities with the launch of Gemini 2.5 Pro, revolutionizing search functionality, healthcare solutions and multimodal content generation

Recent Developments

- Google recently introduced Gemini 2.5 Pro, an advanced AI model designed to improve complex reasoning, long-context understanding, and multimodal capabilities across text, code and media. Built to handle tasks in coding, math and science, Gemini 2.5 Pro currently features a 1 million token context window, with expansion on the horizon. It powers real-time content generation across formats such as text, images, audio and video. Gemini 2.5 was initially available through the Gemini Advanced subscription, but Google has since made Gemini 2.5 Pro free to all users as part of its ongoing effort to expand AI accessibility and compete with leading models from OpenAI and Anthropic.
- To enhance search experiences, Google has introduced AI Mode in Search, leveraging Gemini 2.0 to provide more personalized and conversational results. This expansion builds on AI Overviews, which now serve over 1 billion users monthly across more than 100 countries. By integrating advanced AI capabilities, Google aims to improve information retrieval and user interactions.
- Google is advancing healthcare with AI-powered tools designed to improve accessibility, diagnosis and patient outcomes. These innovations include AI models that analyze medical images, predict health risks and assist in clinical decision making. Through partnerships with global healthcare institutions, Google is working to integrate these solutions into real-world settings, ensuring they enhance medical research and patient care.

Recent Flagship Model Specifications

Gemini 2.5 Pro
Parameter Count: Unknown
Active Parameters: Unknown
Context Window: 1 million tokens

Gemma 3
Parameter Count: 1 billion, 4 billion, 12 billion, 27 billion
Active Parameters: Unknown
Context Window: 32,000 tokens for 1 billion, 128,000 tokens for 4+ billion

TBR Assessment

Google remains differentiated in the AI landscape through the deep integration of its proprietary models across a broad product ecosystem, including Search, YouTube, Android and Workspace. Although many competitors focus on niche capabilities or open-source development, Google positions Gemini as a comprehensive, multimodal foundation model designed for widescale consumer and enterprise adoption. Google's infrastructure, proprietary TPUs (Tensor Processing Units), and access to vast and diverse data sources provide a significant advantage in training and deploying next-generation models. Gemini 2.5 Pro is a testament to this strength, offering the best performance and largest context window available on the market. Although TBR expects the top spot to continue exchanging hands, we believe Google's models will remain among the frontier leaders for years to come.

OpenAI advances AI development with GPT-4.5, cutting-edge agent tools and a premium ChatGPT Pro subscription to expand capabilities and improve user experiences

Recent Developments

- OpenAI recently introduced GPT-4.5, a research preview model available to Pro users and providers. Building on previous versions, GPT-4.5 enhances unsupervised learning, improving pattern recognition and creative insight generation. Early reviews highlight the model's ability to assist in complex tasks, such as operational planning for biological threats. OpenAI plans to release GPT-4.5 to the public later in 2025 following development updates.
- OpenAI has introduced a new suite of tools designed to simplify the development of AI agents. The suite includes the Responses API, which merges the ease of the Chat Completions API with the tool-use features of the Assistants API. It also offers built-in functionalities like web search, file search and computer use to enhance agent capabilities. The suite's Agents SDK helps manage single and multi-agent workflows, while observability tools allow for detailed tracking and inspection of agent execution.
- In December OpenAI introduced ChatGPT Pro, a subscription plan offering enhanced access to advanced AI models and tools. For $200 a month, subscribers receive unlimited use of OpenAI's most sophisticated model, o1, along with o1-mini, GPT-4o and Advanced Voice. The plan also features o1 pro mode, which utilizes additional computing power to deliver more accurate and comprehensive responses to complex queries.

Recent Flagship Model Specifications

GPT-4o
Parameter Count: over 200 billion
Active Parameters: Unknown
Context Window: 128,000 tokens

GPT-o3-mini
Parameter Count: ~200 billion
Active Parameters: Unknown
Context Window: 128,000 tokens

GPT-o3-mini-high
Parameter Count: ~200 billion
Active Parameters: Unknown
Context Window: 128,000 tokens

GPT-4.5 (preview)
Parameter Count: ~10 trillion
Active Parameters: ~600 billion
Context Window: 128,000 tokens

TBR Assessment

OpenAI is the most valuable model developer in the market today, largely due to the company's success in productizing its models via ChatGPT. The mindshare generated by ChatGPT is benefiting the company's ability to reach custom enterprise workloads, though OpenAI must be mindful of the widening gap in price to performance relative to peers. From a sheer performance perspective, TBR believes the company's emphasis on securing compute infrastructure via the Stargate Project, as well as its ongoing partner initiatives to gain access to high-quality training data, will ensure its models remain near the top of established third-party benchmarks over the long term.

Table of Contents and Vendor Coverage

Publish date of latest edition: April 18, 2025

Vendor Coverage

Foundation Model Vendors
- AI21 Labs
- Anthropic
- Cohere
- Hugging Face
- Meta
- Mistral AI
- OpenAI
- Stability AI

SaaS Providers
- Salesforce
- SAP
- ServiceNow

Hyperscalers
- Amazon Web Services (AWS)
- Google Cloud
- Microsoft
- Oracle

Technology Business Research, Inc. is a leading independent market, competitive and ecosystem intelligence firm specializing in the business and financial analyses of hardware, software, professional services, and telecom vendors and operators. Serving a global clientele, TBR provides timely and actionable market research and business intelligence in formats that are tailored to clients' needs. Our analysts are available to address client-specific issues further or information needs on an inquiry or proprietary consulting basis.