IDC MarketScape

IDC MarketScape: Worldwide GenAI Life-Cycle Foundation Model Software 2025 Vendor Assessment

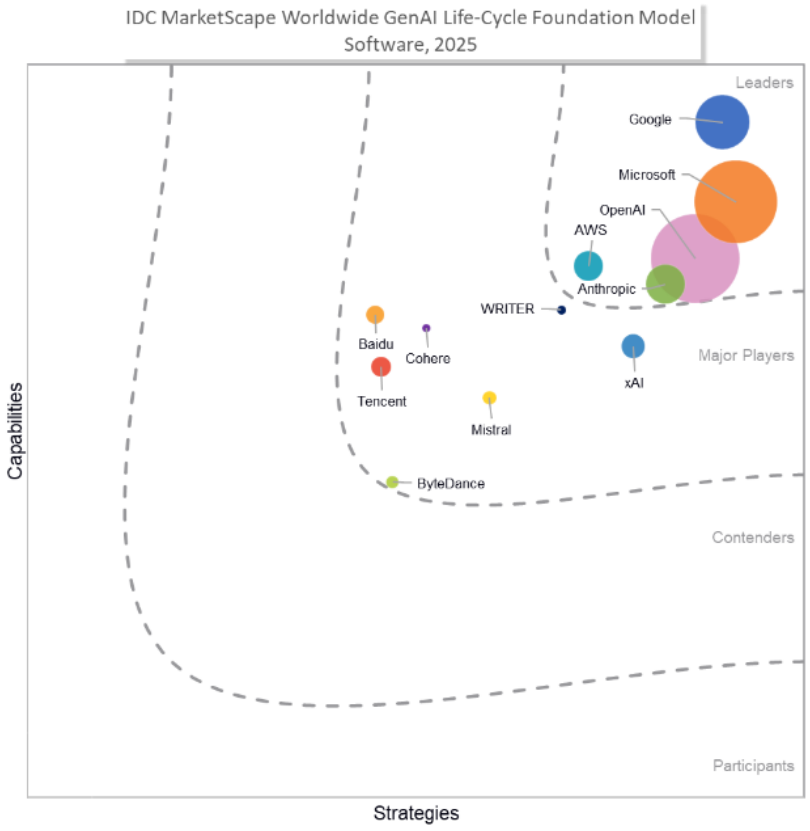
Tim Law Nancy Gohring

THIS EXCERPT FEATURES GOOGLE AS A LEADER

IDC MARKETSCAPE FIGURE

FIGURE 1

IDC MarketScape Worldwide GenAI Life-Cycle Foundation Model Software Vendor Assessment



Source: IDC, 2025

October 2025, IDC #US53007225e

See the Appendix for detailed methodology, market definition, and scoring criteria.

ABOUT THIS EXCERPT

The content for this excerpt was taken directly from IDC MarketScape: Worldwide GenAI Life-Cycle Foundation Model Software 2025 Vendor Assessment (Doc # US53007225).

IDC OPINION

This study focuses on "model families" from the major model providers that incorporate foundation models (including multimodal models), large language models (LLMs), and large reasoning models (LRMs). It includes LLMs, multimodal models, and reasoning models from each provider that have a common architecture. It does not include models built specifically for coding, image outputs, image classification, video, or audio generation.

The model landscape has experienced rapid and accelerating innovation based on increasing the scale of training data and parameters. The largest foundation models exceed multiple trillions of parameters and are pretrained on web-scale, public, proprietary, and synthetic data sets that reach into the trillions of tokens, requiring massive amounts of GPUs. Training models at this scale unlocked a virtuous cycle of mutually reinforcing innovations across the foundation model landscape that has been fueled by a flurry of academic research and massive investments in commercial research and development (R&D) and spurred on by innovations in the open source community. This focus on innovations was driven in part by operational concerns about model efficiency, speed, and costs. But it enabled architectural innovations and breakthroughs in pretraining and post-training techniques that have made models more efficient. In turn, this unlocked innovations in memory, reasoning, thinking, and other "cognitive" capabilities, which allowed foundation models to think like humans on many levels, solving structured problems, generalizing from limited data, transferring insights across domains, identifying causal relationships, and breaking down complex problems into intermediate steps.

Innovations have led to more capable, safer, more efficient, and more cost-effective models. Enterprises should understand and inquire with individual model providers about the architectural innovations they have adopted. Some of the critical areas of innovation are the following:

- **Multimodal:** Development of multimodal attention mechanisms, multimodal processing, and multimodal context handling

- **Reasoning:** Advanced reasoning with improved logical inference and stepwise logic
- **Metacognition:** Models reasoning about their own reasoning and adapting behavior
- **Reflection:** Reflection mechanisms allowing models to assess their own logic
- **Attention:** Advanced attention mechanisms like sparse, grouped, and multi-query attention
- **Agentic:** Agentic capabilities like planning, tool calling, function calling, and decisioning
- **Self-improvement:** Models using reward mechanisms such as reinforcement, reward estimation, and optimization

The innovations in LLMs across many areas have been mutually reinforcing. For example, innovations in attention mechanisms enabled multimodal processing, and innovations in multimodal processing led to improvements in instruction following. Innovations in both core model capabilities and adjacent capabilities fed this virtuous loop of model improvements.

Foundation models have followed a virtuous cycle of improvement rivalling any other in the history of enterprise AI. But given the nondeterministic nature of these models, much work is still needed to enable these models to meet the requirements of mission-critical enterprise applications and workflows safely and reliably.

The ability to quickly innovate and deliver new capabilities across these core components of model architectures and adjacent capabilities and across the generative AI (GenAI) stack has a profound impact on the competitive dynamic among foundation model providers. The race to dominate the foundation model market hinges on the ability of model providers to smartly invest in R&D and deliver significant innovations across all these areas to meet the requirements of production-grade enterprise deployments.

In parallel, foundation models have quickly evolved over the past two years from predominantly large language models at the center of generative AI applications to large "reasoning" models at the center of agentic AI systems, becoming autonomous AI agents that can use tools, search the web, plan, and execute workflows. Agentic systems powered by foundation models can reason, plan, and act, beyond simply generating content. The new reasoning capabilities enable enterprises to adopt an agentic approach to workflow automation to increase productivity, speed time to market, create better customer experiences, and reduce costs.

IDC MARKETSCAPE VENDOR INCLUSION CRITERIA

Vendors must be the primary creator and maintainer for one or more pretrained foundation models under active development that were released before March 31, 2025:

- Foundation models generate English language responses, although they may also be capable of generating other natural languages.
- Foundation models are text based or multimodal but may not be designed primarily to work with only images, video, audio, or speech.
- Foundation models were available for commercial use before or on March 31, 2025.

ADVICE FOR TECHNOLOGY BUYERS

Enterprises have a profound influence on the direction of model development and need to continue to advocate for innovations, richer enterprise capabilities, and safe, aligned, and responsible AI. Furthermore:

- **Institutionalize model selection and testing as a core competency.** There are hundreds of models available serving enterprise use cases. To scale the use of foundation models, your enterprise must have a formal, dedicated, and disciplined way of assessing models across a spectrum that includes safety, accuracy, performance, cost, and many other variables to ensure it aligns with the intended use case. Ensure you can evaluate models thoroughly and select the right fit for the use case and budget.
- **Collaborate with model providers that are committed to enterprise requirements.** Large model providers focus on multiple markets across enterprises and consumers. Enterprises need to assess what percentage of a vendor's R&D focus and spend is aimed at enterprise capabilities. As importantly, enterprises need to assess whether model providers have an enterprise software "mindset" and are organized around the task of delivering value to enterprises in every aspect of model development, delivery, and operations.
- **Focus on the right aspects of vendor innovation.** The ability to rapidly innovate is an important consideration in selecting a model provider. However, enterprises need to ensure they filter the hype that has driven valuations over the past several years and demand a detailed view of model provider road maps. Enterprises should select model providers that are aligned with the virtuous cycle of innovation described previously. It is critical to get transparency into the model providers' road map, commitment to delivering enterprise capabilities, and funding streams to continue investing in development. In the past two years, several prominent model providers either deprecated their foundation models

or suspended development. Enterprises need to ensure that their model providers are aligned with business requirements and have the wherewithal to deliver key model enhancements over the long term.

- **Note that extensive and aligned partner ecosystems are critical.** Model providers with extensive, committed, and aligned ecosystem partners are most likely to thrive. While foundation models may be the heart of generative and agentic systems, model value and performance are tightly tied to adjacent software capabilities, infrastructure, and hardware. Enterprises should seek model providers with extensive internal and external ecosystems.
- **Maintain control over model operations.** Enterprises must maintain control over key aspects of model operations to control costs and achieve desired business outcomes. They should leverage models that provide built-in controls over model performance. Such controls include manually or automatically adjusting how much time or budget the model will use in a reasoning task, which enables enterprises to control both latency and cost. Other controls involve trade-offs for latency and accuracy, such as toggles for adjusting reasoning levels. Be wary of completely automated controls. While some controls can be safely automated over time, in the near term, enterprises need to ensure controls are available for managing both cost and risk. Such controls include control over model selection and routing (e.g., routing simple inquiries to smaller, more efficient models), rate limiting, quotas, and other aspects that impact costs. Enterprises should collaborate with model providers that offer an optimal mix of direct controls and optional automation.
- **Focus on innovators in model safety and security.** Model providers quickly learned the importance of model alignment and safety in mission-critical enterprise applications. Enterprise adoption opened the aperture on existing risks and exposed new, unknown risks to enterprises. Enterprises should select model providers that can comprehensively address safety risks by innovating core capabilities, working with adjacent software and frameworks, and leveraging partner ecosystems. While additional guardrails are important, model safety begins at the core of the model.

VENDOR SUMMARY PROFILE

This section briefly explains IDC's key observations resulting in a vendor's position in the IDC MarketScape. While every vendor is evaluated against each of the criteria outlined in the Appendix, the description here provides a summary of each vendor's strengths and challenges.

Google

Google is positioned in the Leaders category in the 2025 IDC MarketScape for worldwide GenAI life-cycle foundation model software vendor assessment.

The Google Gemini model family is a set of multimodal models that feature advanced reasoning capabilities and cover numerous use cases. Built on a sparse mixture-of-experts architecture, the models feature a 1 million token context window for processing of multimodal inputs, including text, vision, and audio.

Strengths

- Gemini models are widely available, including via the Gemini API in the Google Cloud Vertex AI platform, which features many related services for designing generative AI and agentic AI applications. Google supports supervised fine-tuning via Vertex AI for customization.
- The native multimodal capabilities differentiate Gemini from many other model families. Gemini features deep integration into the Google ecosystem, such as Google Workspace, Android, and other Google Cloud offerings.
- Google's vast research and development with DeepMind enables it to quickly innovate, and its access to rich data sources provides a competitive advantage.

Challenges

- Gemini is still catching up to the large user base of ChatGPT, which gained rapid adoption and an early lead with individual users in enterprises and with consumers.
- While featuring advanced reasoning capabilities, Google should provide developers with additional tools to balance reasoning depth against compute and inference costs.
- Previous Gemini models fell short relative to other competitors in coding tasks in the market. Google's recent Gemini 2.5, the introduction of Gemini CLI, and the joining of Windsurf staff show progress in closing the gap. Google should continue to invest in interactive coding capabilities.

Consider Google When

Use cases requiring extended thinking, complex reasoning, and native multimodal capabilities can be a good fit for Gemini models. Gemini should be considered for enterprises developing applications within the Google ecosystem and that are already using Vertex AI as a development platform for building agentic AI or generative AI applications or leveraging Google data infrastructure. Gemini is a good fit for many use cases requiring advanced multimodal capabilities.

APPENDIX

Reading an IDC MarketScape Graph

For the purposes of this analysis, IDC divided potential key measures for success into two primary categories: capabilities and strategies.

Positioning on the y-axis reflects the vendor's current capabilities and menu of services and how well aligned the vendor is to customer needs. The capabilities category focuses on the capabilities of the company and product today, here and now. Under this category, IDC analysts will look at how well a vendor is building/delivering capabilities that enable it to execute its chosen strategy in the market.

Positioning on the x-axis, or strategies axis, indicates how well the vendor's future strategy aligns with what customers will require in three to five years. The strategies category focuses on high-level decisions and underlying assumptions about offerings, customer segments, and business and go-to-market plans for the next three to five years.

The size of the individual vendor markers in the IDC MarketScape represents the market share of each individual vendor within the specific market segment being assessed.

IDC MarketScape Methodology

IDC MarketScape criteria selection, weightings, and vendor scores represent well-researched IDC judgment about the market and specific vendors. IDC analysts tailor the range of standard characteristics by which vendors are measured through structured discussions, surveys, and interviews with market leaders, participants, and end users. Market weightings are based on user interviews, buyer surveys, and the input of IDC experts in each market. IDC analysts base individual vendor scores, and ultimately vendor positions on the IDC MarketScape, on detailed surveys and interviews with the vendors, publicly available information, and end-user experiences in an effort to provide an accurate and consistent assessment of each vendor's characteristics, behavior, and capability.

Market Definition

A general-purpose foundation model is any model pretrained on broad data (usually featuring transformer architecture, attention mechanisms, and self-supervised learning) that can be adapted using fine-tuning or RAG for either general or domain-specific downstream tasks. A model family is a group of models offered by a single vendor that typically share a common origin or structure but may be optimized for

different use cases and objectives. This IDC MarketScape addresses the market for proprietary models and does not include open source models such as those from Meta, NVIDIA, and IBM.

This study examines the capabilities of proprietary foundation models. Foundation models are large AI models that are pretrained on massive data sets, including the entirety of accumulated human knowledge available from online sources as well as specialized data sets, both real and synthetic, from domains such as science, math, literature, medicine, and law.

These models are at the heart of the generative AI and agentic AI revolution, which promises to completely rewrite the business and social landscape and usher in an era of "super intelligence." Foundation models are general-purpose models built to adapt to many tasks. They can be fine-tuned to specific tasks or domains and can be distilled into smaller, more efficient models.

The power of these models is their ability to generate novel outputs from existing sources. While early models processed and generated only text, foundation models evolved in the past several years to include the ability to process multimodal inputs (text, images, audio, and video) and output multiple formats. The rapidly evolving multimodal capabilities enabled enterprises to build cross-functional applications, for example, automating both the generation of advertising copy and the creative design of an ad campaign or website.

Because they are general-purpose models, foundation models serve a wide range of enterprise use cases. Some of the top categories of enterprise use cases are:

- **Content generation:** Enterprises use foundation models to generate content, such as text for emails and marketing materials, and to summarize documents. Foundation models have evolved to be able to process and generate structured content such as tables and graphs from text prompts and, as mentioned previously, generate images, video, and audio. Leading models can now process multimodal inputs and generate multimodal outputs.
- **Knowledge management and retrieval:** Foundation models are used to retrieve relevant information from knowledge sources based on user prompts. These sources can include public sources like web content or proprietary knowledge sources. Foundation model families often include embedding models, which prepare data for semantic search and retrieval-augmented generation.
- **Customer service:** A common use case for foundation models is enabling customer support with efficient search and retrieval of policies, documentation, and other company information to answer customer inquiries. When deployed as agents, they can autonomously resolve customer inquiries and enable customer self-service.

- **Software development and IT automation:** Foundation models are trained on coding languages and can generate code based on user prompts. They can be deployed as coding assistants to generate new code, debug code, and explain code. These models have evolved to the point of being able to construct entire applications from user prompts.
- **Business process automation:** Foundation models are being used to automate entire workflows from simple one-step processes to complex multiagent workflows. AI agents underpinned by reasoning foundation models can, for example, process insurance claims notifications, extracting details from call transcripts, submitted documents, and images; assess claims severity; validate coverages; and initiate settlement processes without human intervention.
- **Research and analysis:** Foundation models with advanced capabilities for deep research are automating processes across many domains, including legal, medical, and other areas of research. The mathematical reasoning capabilities of foundation models are increasingly being leveraged for analytical tasks, including time series and forecasting. For specialized domains, agentic capabilities of domain models can be leveraged to call tools for specialized analysis, such as tax calculators and similar tools.

Innovations in Core Model Capabilities

Foundation models have benefited from the self-reinforcing innovations in both core architecture and adjacent capabilities. Continued and rapid core architectural innovations include:

- **Attention mechanisms:** While attention mechanisms introduced the transformer architecture, they unleashed rapid innovations in the attention layer and spurred research into novel model architectures. Transformer-style attention mechanisms have evolved rapidly, reducing memory usage, increasing processing speed, and enabling longer context. Innovations like cross-attention helped enable multimodal capabilities, while alternative nonattention architectures like state-space models (SSMs) are leading to new hybrid architectures. Additional innovations borrowed from architectures like mixture of experts (MoE) have delivered unified multimodal attention, and more innovations are expected in both attention and nonattention approaches, along with hybrid approaches.
- **Instruction following:** This allows the model to interpret human intent and generate outputs aligned with the prompt. Before 2022, models required significant prompt engineering to produce aligned outputs. With significant enhancements, including instruction tuning, reinforcement learning with human feedback (RLHF), and multimodal processing, there have been large improvements in processing multiturn and conditional task instructions.

- **Multimodal processing:** Foundation models have moved beyond text-based tasks to incorporate multimodal capabilities that can process text, video, image, and audio inputs and likewise generate content in multiple modalities. This has unlocked additional use cases and enabled automation of existing processes across functions. Even more improvements are anticipated with the introduction of unified multimodal models that take any input and process outputs to any modality.
- **Context handling:** While there have been innovations in context handling for text, innovations in this component of LLMs were accelerated by the need to efficiently and cost effectively scale context lengths to process multimodal inputs. High-resolution images, audio, and video inputs can yield many thousands or millions of tokens. Innovations like sparse attention mechanisms, innovations in compression (downsampling features), and external memory and retrieval were critical.
- **Expanded context windows:** This is a related but separate issue from context handling. Context windows can be thought of as short-term or working memory. There has been exponential growth in the size of context windows from 128,000 tokens to up to 1 million–2 million tokens today, allowing users to include more instructions and upload larger and more documents and have longer conversations with the LLM. Context caching has been a recent development that allows expansion of the context window, and future improvements in tokenization will define how efficiently the context window is utilized.
- **Reasoning components:** Early LLMs were not as capable at logical inference and step-by-step thinking as current state-of-the-art models. Along with massive scaling of training data sets and developments like chain of thought (CoT), tree of thought (ToT), and reinforcement learning (RL) fine-tuning, these have helped models achieve stepwise logic. Along with greater specialization, these enhancements are now embedded in the model architecture.
- **Reflection mechanisms:** These are designed to enable the model to critique its performance on a task or step in a workflow, then iterate to improve the output. It improves upon the first-pass solution (reasoning) by either validating the output or revising it and improving it over time with self-critiquing loops.
- **Self-improvement:** The most recent releases of the major foundation models have demonstrated incredible capabilities to act autonomously and increasingly incorporate capabilities toward self-improvement with rewards and reinforcement and the ability to "evolve" in a sense.
- **Long-term memory:** Traditional LLMs were stateless, which meant they could not recall past conversations or any information across sessions. Beyond persistent storage with vector databases and RAG, researchers continue to

innovate on episodic memory, memory tokens, selective memory and forgetting, and other innovative approaches to more scalable memory architectures.

- **Agentic capabilities:** Planning, function calling, tool calling, agentic reasoning, and autonomous actions have significantly expanded the use of foundation models, which drive agentic applications and workflows. These allow connections to back-end business services or business logic (e.g., function calling for a database query) or connections to external tools or other agents (e.g., executing Python code to analyze data or generate a graph).
- **Model safety:** Reinforcement learning with human feedback remains a critical component of model safety and alignment, but innovations like reward signals for safety, rule-based feedback in reward functions, constitutional AI, safe completions (e.g., model admits it doesn't know rather than hallucinate or fabricate a response), self-reflection for verification, and other internal model mechanisms can mitigate unsafe responses. Many of these safety mechanisms happen at the model pretraining and tuning phase. New threats to the mental health of individual users have recently come into the spotlight, such as sycophancy, where models can negatively impact individuals by flattery or by reinforcing harmful or biased views. Model safety is a constantly expanding landscape that will require model providers to continue to invest in the safety of enterprises and individual employees.

Innovations in Adjacent Capabilities

Core model capabilities also benefit from innovations in adjacent software capabilities like orchestration, retrieval-augmented generation, inferencing (both at the software and hardware layer), data preparation and management, chunking and semantic indexing, vector databases, context assembly, multimodal preprocessing, and hardware and other infrastructure innovations. They benefit symbiotically from integrations with other enterprise solutions like enterprise search, productivity tools, and others. More innovations across all these areas will be required to deploy generative AI and agents on a massive scale. In detail:

- **Integrations:** Model providers have steadily been integrating with enterprise productivity tools, enterprise search, web search, and other applications to enhance their model capabilities and expand the use cases they can address. Model providers with large software application portfolios have been able to rapidly expand model capabilities and response richness via these integrations.
- **Retrieval:** Many model providers have directly integrated various retrieval capabilities into their models, including various patterns of retrieval-augmented generation, such as GraphRAG or other information search and retrieval mechanisms, to help ground models in enterprise data. Model providers in

general continue to benefit from innovations in adjacent retrieval and search capabilities.

- **Cost and FinOps:** On the cost front, foundation model providers are investing in innovations in cost controls (e.g., offering smaller, faster variants in their model family for simple queries and tasks), embedding capabilities within the model (cost-efficient architectures, cost-aware inference), and enabling their models to exploit infrastructure optimizations. Investments in more granular cost management provide cost signals that identify opportunities for more model innovations.
- **Guardrails and safety APIs:** While model safety is a core architectural component and the front line against unsafe model behaviors, adjacent capabilities like guardrails and safety APIs are critical lines of defense as well. Guardrails include filters to prevent unsafe prompts and filter unauthorized outputs. Safety APIs are a software layer between generative models and users that monitor and filter inputs and outputs for problems like malicious content.
- **Evaluation platforms for GenAI and agents:** Some model providers are closely partnering with third-party evaluation platform providers that measure and evaluate model and generative AI application performance across the entire life cycle in several areas, from retrieval relevancy to accuracy to function calling to tool usage. This allows for a more comprehensive review of model performance for both pre-production testing and in-production deployment for continuous monitoring. Again, these platforms provide important signals and insights that feed the cycle of model improvements.

LEARN MORE

Related Research

- *Asia/Pacific AI Enterprise Applications and Customer Experience Strategies* (IDC #IDC_P36949, September 2025)
- *Bumpy Launch for GPT-5, But OpenAI Ushers in an Era of Faster, Smarter, and Safer Models* (IDC #lcUS53744625, August 2025)
- *WRITER Introduces a New Autonomous "Super Agent": A Digital Coworker to Unlock Productivity* (IDC #US53741425, August 2025)
- *Insights from the AWS Summit NYC 2025: Agentic Portfolios Meet Rising Demand* (IDC #US53731825, August 2025)
- *AI View* (IDC #IDC_P44382, July 2025)
- *IDC MarketScape Criteria: Worldwide Document Scanner 2025 Vendor Assessment* (IDC #US53670825, July 2025)

- *IDC Market Glance: AI Life-Cycle Tools and Technologies, 3Q25* (IDC #US53695825, July 2025)
- *Global GenAI Technology Trends Survey, May 2025* (IDC #US53177225, July 2025)
- *IDC Market Glance: AI-Enabled Marketing Platforms for Large Enterprises, 2Q25* (IDC #US53555425, June 2025)
- *Microsoft Positions Itself as the Foundational Leader for Agentic AI* (IDC #US53563125, June 2025)
- *Agentic AI Impact on Enterprises: From the Tech Stack to the Future of Work and Services* (IDC #US53272524, April 2025)
- *Amazon Introduces SageMaker Unified Studio Marking the Race Toward Streamlined AI and GenAI Development* (IDC #US53184725, February 2025)

Synopsis

This IDC study evaluates 12 proprietary foundation model families. It emphasizes innovations in core capabilities like attention mechanisms, multimodal processing, and reasoning, alongside adjacent capabilities such as retrieval-augmented generation (RAG) and integrations. The study provides vendor profiles, strengths, challenges, and recommendations for enterprises to select models aligned with their needs, focusing on safety, efficiency, and enterprise-grade capabilities for generative AI and agentic AI applications.

"Foundation models revolutionized AI, evolving from simple text generation tasks to autonomous agents, driving unprecedented innovation in reasoning, multimodal processing, and enterprise-grade capabilities," said Tim Law, research director, AI and Automation, IDC. "A virtuous cycle of innovations in core model architecture, adjacent capabilities, GenAI hardware, and infrastructure has enabled widespread enterprise adoption and has rapidly altered the enterprise AI landscape."

ABOUT IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com