Global Artificial Intelligence Report (2025)

IDCA

Terms of Use and Disclaimer

This document is published by the International Data Center Authority (IDCA). The report is free of charge to individuals, governments, and all other organizations interested in it. Its copyright and intellectual property belong to IDCA. Anyone quoting from this report should attribute IDCA as the source. The report was created by a collaborative team with multiple areas of expertise and points of view. The findings, interpretations, and conclusions expressed in this report are those of the IDCA alone and do not necessarily represent those of IDCA members or other parties.

Contents

Acknowledgment

This comprehensive report is created through the efforts of seasoned subject matter experts of IDCA who have dedicated their time and expertise in hopes of bringing transparency and light for resourceful measures to the Digital Economies of the world:

MARK MINEVICH
Board Member & AI Counsel

MEHDI PARYAVI
Chairman & CEO

ROGER STRUKHOFF
Chief Research Officer

01   Executive Summary

AI is now a top business priority and widely seen as the key driver of Digital Economies.

Executive Summary

AI's prominence became well-established in the world's tech industry in 2024 and has already continued with the same trajectory in 2025. According to IDCA's comprehensive Q1 2025 industry survey, research shows that 87 percent of companies identify AI as a top priority in their business plans, 76 percent of organizations now use AI, 69 percent of organizations use generative AI in at least one business function, and 53 percent use AI to harness Big Data effectively.

Furthermore, IDCA polls and surveys among industry professionals and global leaders show that 72 percent of respondents named AI the leading "game changer" in building Digital Economies today.

FIGURE 1.

AI's Influence Within Organizations
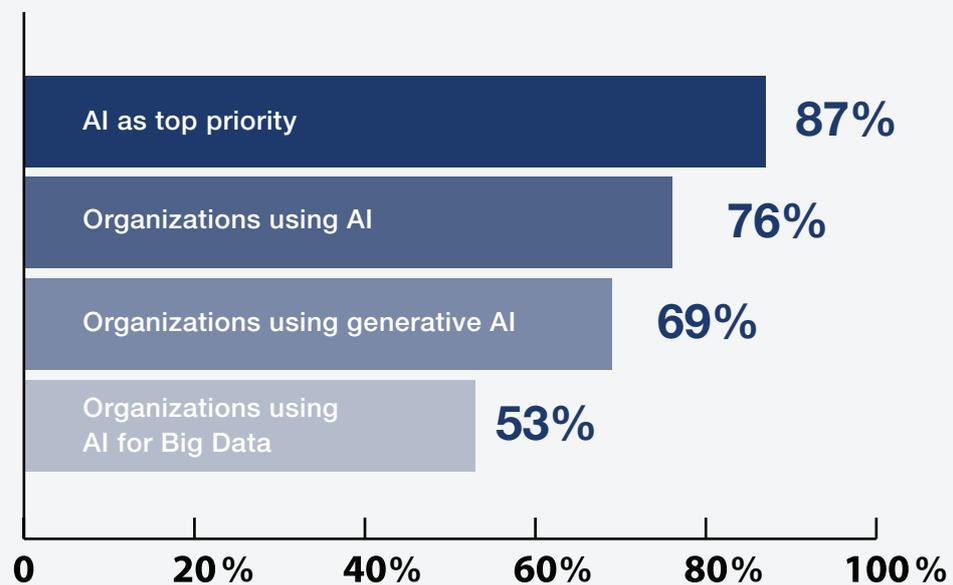
Source: IDCA

Artificial Intelligence (AI) in all forms is increasingly driving the development of the Digital Economy, which now encompasses about $16 trillion of global GDP in nominal terms, according to IDCA's Digital Economy Report 2025. Growth projections for specific revenue for AI companies and initiatives cover a wide range, but several scenarios point to a global market of $1 trillion or more by 2030. Syncing these projections shows that AI is creating at least a 10-to-1 leverage of its use in developing the global digital economy.

AI has become an innovation engine, reshaping industries and redefining economic possibilities. From revolutionizing healthcare diagnostics to automating complex manufacturing workflows, AI is accelerating productivity across the board.

FIGURE 2.

Percent of Digital Economy of Nominal World GDP

15%

World Digital Economy
$16 Trillion

World Nominal GDP
$108 Trillion
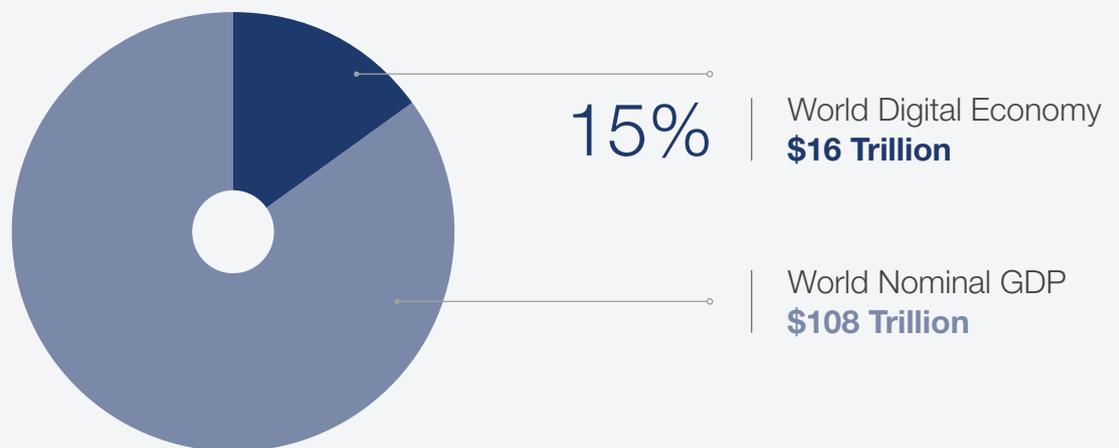
Source: IDCA

AI's convergence with the Internet of Things (IoT), blockchain, and quantum computing fosters robust ecosystems that enable breakthroughs across industries. Companies must adopt AI technologies and position themselves as innovators within this transformative ecosystem. The integration of AI at scale has become a prerequisite for sustaining competitive advantage.

AI's development is progressing into Agentic AI platforms and services, which act as agents for users in various specific tasks. The search for an Artificial General Intelligence (AGI) continues as well. Beyond those frontiers lies the world of quantum computing and AI, which is projected to gain commercial traction within a decade.

There are ethical concerns about AI's development and use, as well as concerns about energy use and emissions levels. The energy use concerns are challenges in providing enough power, nation-by-nation, to meet anticipated demands and sustainably use that power.
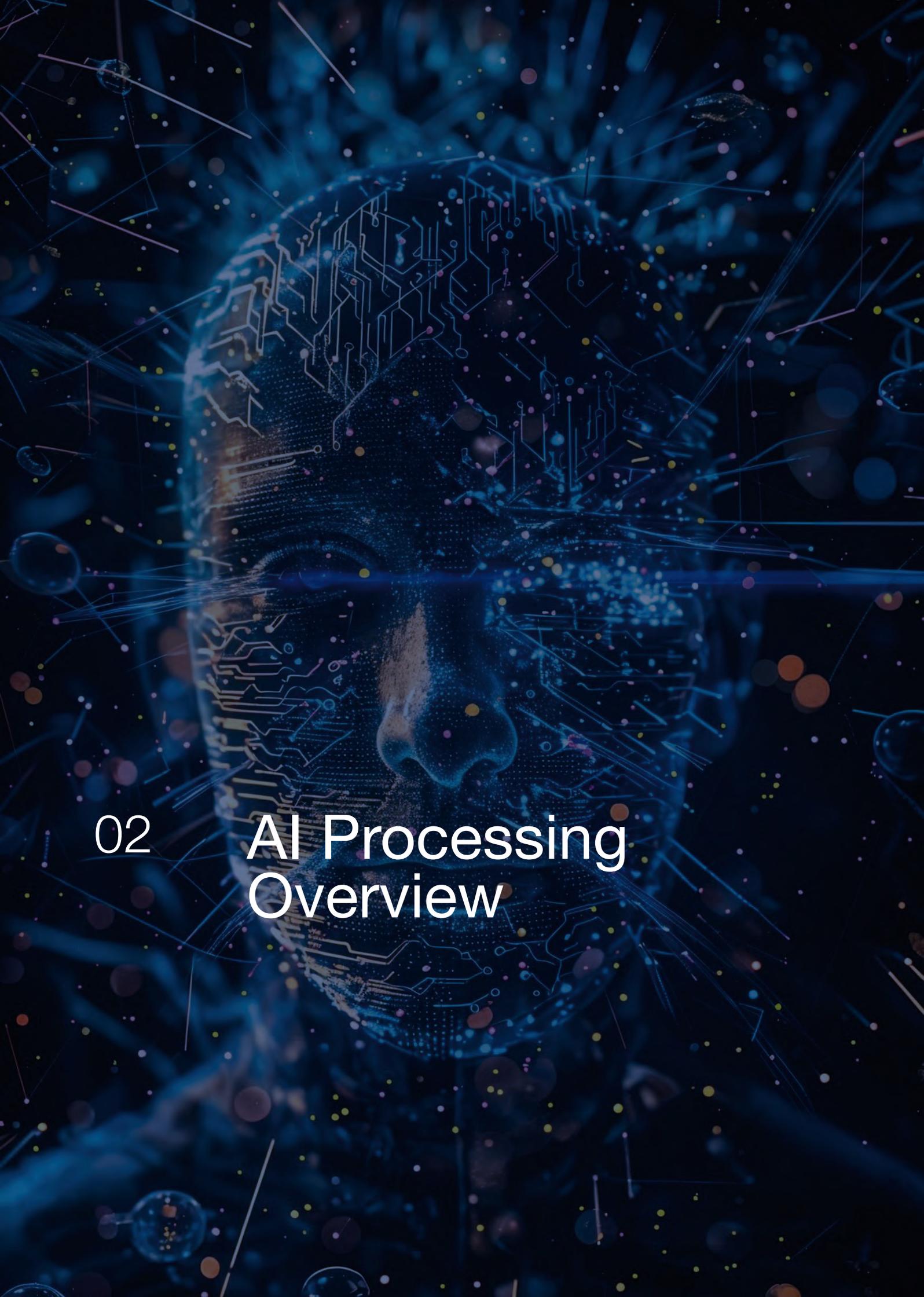
02　　AI Processing
　　　　Overview

AI Processing Overview

The ability to build data centers is of similar scope to the task of developing and using AI models and finding the electricity to power those data centers. These concerns top the lists of businesses and governments that wish to proceed with AI-driven plans.

Concerns about engineering and operating AI data centers can be abstracted to provide a clear understanding of the general situation. This abstraction compares processing power (measured in floating-point operations per second, or "flops"), power consumption (measured in megawatts), and cost.

Using flops to measure processing power is a venerable tradition that is effective today in gauging the raw power requirements of AI and other general data center functions. The most relevant scale in this analysis involves gigaflops (one billion flops), teraflops (one trillion flops), petaflops (one quadrillion flops), and exaflops (one quintillion flops, or one million teraflops). An exaflop can also be expressed as 10^{18} flops.

Traditional chores processed by data centers involve gigaflops and teraflops. Credit-card transactions, for example, require 10 gigaflops per second of processing power worldwide during peak periods. Streaming video is more demanding, requiring about 100 gigaflops per one million streams. (Bandwidth is the larger issue with streaming, with 15-25 megabits per second (Mbps) required per user. The world average sits at about 21 Mbps at the moment. More than 100 nations fall below that average, with 60 nations providing 20 percent or less of the average.)

FLOP Table		
Common Terms for FLOPs (floating-point operations per second)		
Term FLOP	**# operations per second**	**Notation 10^0**
megaflop	1 million	10^6
gigaflop	1 billion	10^9
teraflop	1 trillion	10^{12}
petaflop	1 million billion (1 quadrillion)	10^{15}
exaflop	1 billion billion (1 quintillion)	10^{18}

Distribution across data centers (to reduce single-point-of-failure risk), high levels of redundancy, and application support can increase these requirements by a magnitude. Still, traditional cloud services are generally one of gigaflops and teraflops.

AI presents a dramatically new and demanding scenario. To serve the world's increasing demand for it, AI inference (using ChatGPT) requires petaflops of processing power, and AI development (i.e., model training) requires exaflops—millions of teraflops.

Recent IDCA analysis shows the computational demands are increasing exponentially, with the latest multimodal models requiring 2-3 times more processing power for training and inference than single-modality models of similar size.

AI Chip Performance

Current popular traditional data center processing CPUs include the Intel Xeon Gold 6230 and AMD Epyc 7742. These chips typically have 20 and 64 cores, respectively, delivering one to two gigaflops per core. Thus, they provide between 20 and 128 gigaflops per chip. They are generally available at $1,000 to $2,000 per unit.

Moving into the AI realm, higher performance is found with the Nvidia A100 Tensor Core GPU, which can deliver as many as 312 teraflops. This GPU is generally available at $10,000 to $20,000 per unit. A similar price range is found in the public cloud, with a Google Cloud TPU v4 (to list just one example) priced at about $2,000 per month (or $24,000 annually) for 24/7 usage.

Another step up is the Nvidia H100, priced at $25,000 to $40,000 each and delivering as many as a petaflop (1 million gigaflops) for specific, focused uses and 67 teraflops for complex processing. Then there is the Nvidia Blackwell family, with chips costing as much as $70,000, delivering as many as 5.5 petaflops each, and multi-GPU boards in the $3 million range requiring as much as 120 kW of power.

Exaflop applications within the AI world include climate modeling and nuclear simulations, modeling the behavior of galaxies and the universe, new drug development, massive dataset analysis, high-end cryptography and cybersecurity, and training LLMs. This type of computation requires large-scale resources in a parallel computation architecture.

The Blackwell platform has proven to deliver approximately 30 percent greater performance per watt than its predecessor, significantly improving data center efficiency. Additionally, Intel's Gaudi 3 AI accelerators have emerged as a viable alternative in the market, offering competitive price/performance ratios at $25,000-30,000 per unit. AMD's MI300X accelerators have also gained traction, with over 120,000 units deployed globally as of Q1 2025.

Nvidia also promises another breakthrough in 2027, with its Ruby family delivering 15 exaflops, requiring 600 kW of power.

FIGURE 3.

Data center AI Accelerrator Unit Share

2% | Other Venders

2% | Intel

90% | Nvidia

6% | AMD
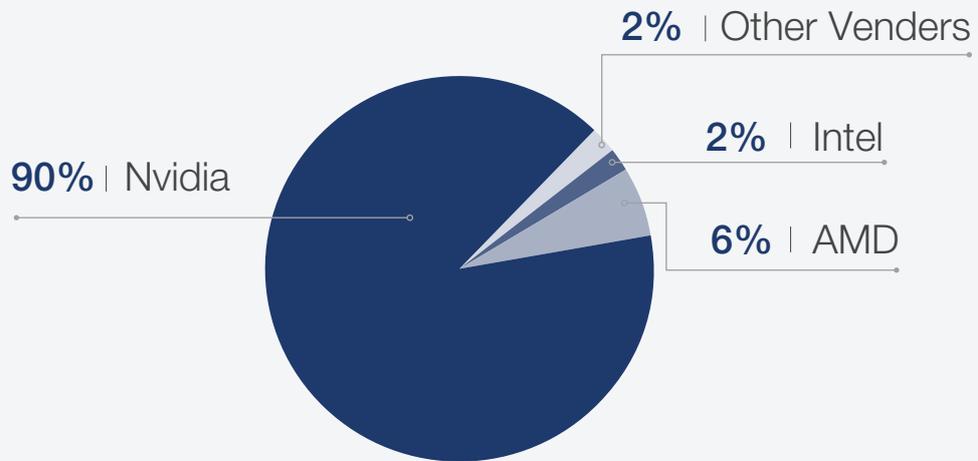
AI Factories/Data Refineries

Mega-projects are leading to the notion of "AI factories" and "data refineries." Data has been pronounced the "oil of the 21st century" and the most relevant currency of today and tomorrow. As with traditional manufacturing, AI factories and data refineries can vary in size, with larger-scale facilities available to nations committed to providing the power necessary to run them and services of value domestically and internationally.

Modern AI data centers have evolved dramatically from traditional facilities, requiring specialized designs to accommodate unprecedented power densities, cooling challenges, and network demands. Key developments in 2025 include:

1. **Distributed Architecture Models:** Leading AI operators have implemented geo-distributed processing clusters that balance workloads across multiple sites, reducing single-point failure risks and optimizing power availability and cost.

2. **Advanced Cooling Technologies:** Direct liquid cooling (DLC) has become the standard for high-density AI racks, with immersion cooling solutions deployed in over 60 percent of new AI-optimized facilities. These technologies enable power densities exceeding 100kW per rack while maintaining efficient operation.

3. **Modular Construction:** Pre-fabricated, modular data center designs have reduced deployment times by 40 percent, allowing for rapid scaling of AI infrastructure. These standardized units can be deployed in 8-12 weeks versus traditional 18-24-month construction timelines.

4. **Real-time Power Management:** AI-controlled power distribution systems dynamically allocate electricity based on workload priorities, reducing overall consumption by up to 18 percent while maintaining performance SLAs.

Thus, a nation's AI infrastructure and software capabilities are increasingly important to measure its economy's strength, ability to defend itself, and add value to the world. Investors and governments will analyze the evolving definition of high-performance computing (HPC) and AI computing capabilities as a "new GDP" to rank each nation's economic standing and qualify the amount of meaningful data it can create.

Regarding raw processing power, a system delivering a single exaflop has millions of times the processing power required by traditional data center applications. Thus, the multi-exaflop systems from Nvidia are referred to as AI factories rather than traditional data center components.

Standard issues such as redundancy, load balancing, specialization, and scalability for peak loads will still concern the exaflop world. However, this world is different from traditional data centers in any scenario.

Numerous aggressive AI center development plans have been announced in 2025, including $500 billion for the Stargate project in the US, a 3GW complex in Rajasthan, India, and several commitments of between $50 and $100 billion by major cloud providers and users.

These megaplans join Oracle Cloud's new Arizona AI Hub, a $40 billion facility designed specifically for government and enterprise AI workloads with 1.5GW of dedicated power capacity. Microsoft's Gobi Desert Quantum facility represents another frontier installation, combining traditional AI compute with early quantum processing capabilities. Google's Carbon-Neutral AI Center in Finland demonstrates how sustainability can be integrated into large-scale AI infrastructure, achieving a Power Usage Effectiveness (PUE) of 1.07 through innovative design and renewable energy integration.
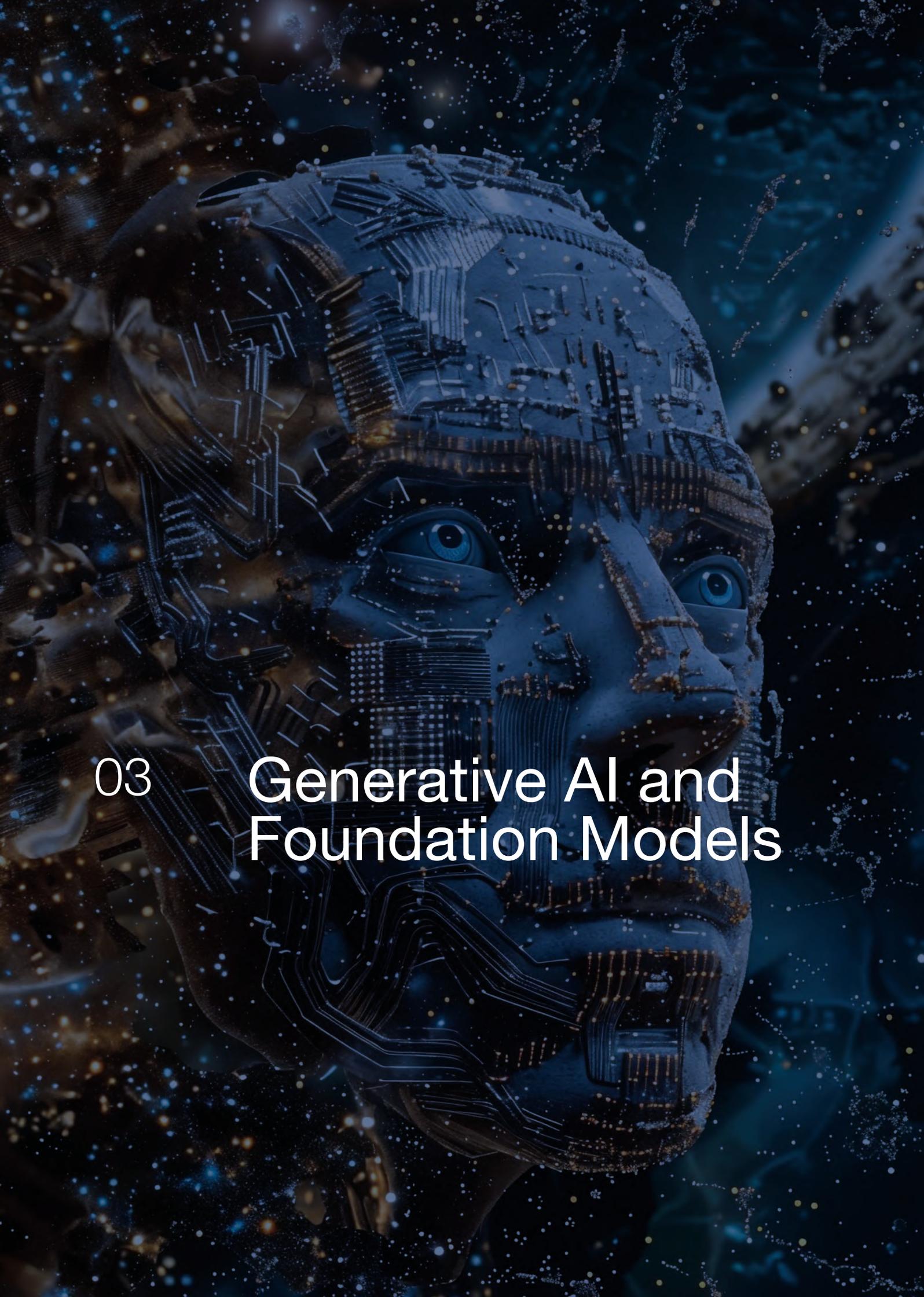
03   Generative AI and
        Foundation Models

Generative AI and Foundation Models

The recent emergence of generative artificial intelligence (GenAI) includes language-model platforms such as ChatGPT and DeepSeek. GenAI also encompasses image and video generation models; this report will cover language models. These models have accelerated the global interest in AI overall.

Within the world of GenAI, a rough division exists between large-language models (LLMs) and small-language models (SLMs) used to build AI platforms, applications, and services. The distinction is expressed in the number of parameters a model has. LLMs may have hundreds of billions or even trillions of parameters, while SLMs will have parameters in the order of a few hundred million to ten billion.

Parameters can be thought of as the neurons of an AI's cognition. So, a single input may have many parameters. Parameters (or neurons) are placed in layers, with multiple (often many thousands) of neurons per layer. Multiplying the number of layers by neurons-per-layer thus creates a total number of parameters.

Parameters are defined as the action or connection the model will make between its input and its output. When developing an understanding of spoken and written language, for example, a model will take a word (say "hat") and assign parameters to it concerning its meaning(s), form, importance in a phrase or sentence, and other linguistic aspects.

The broad size variances in the number of parameters in language models, both large and small, demonstrate that there is no fixed borderline between the two general types of models. As the art and science of AI development continue, one can imagine more classification (e.g., extra-small, extra-large, medium, etc.) as the models become more diverse and specialized.

Recent architectural innovations have blurred these distinctions further. Mixture-of-Experts (MoE) models can contain trillions of parameters while only activating a small fraction for any given task, resulting in computational efficiency comparable to much smaller dense models. Industry benchmarks increasingly focus on parameter efficiency rather than raw parameter count, with metrics like "effective parameters" becoming standard in performance evaluations.

The distinction between SLMs and LLMs is not one of quality, but rather one of purpose. SLMs are used to develop specialized services, such as text prediction, in-line chatbots, or autofill services. The differences among sizes can be analogized to the size of a hammering device one might need to hang a picture versus tear down a building.

On the other hand, the LLMs in use today are focused on the comprehensive abilities found in today's well-known platforms. AI platforms are expected to have a vast compendium of knowledge and a sophisticated approach to delivering answers to users' questions. They require extensive resources to develop and operate. Google, for example, has spent a reported $200 million on developing LLMs.

LLMs perform well due to their scale, which enables broader generalization. The latest models from OpenAI, Meta, and DeepSeek use hundreds of billions of "parameters"–the adjustable knobs that determine connections among data and get tweaked during the training process. With more parameters, the models can better identify patterns and connections, making them more powerful and accurate. Recent architecture innovations in Mixture-of-Experts (MoE) models have enabled even greater efficiency, allowing models to selectively activate only relevant parameters for specific tasks selectively, dramatically improving performance and energy efficiency.

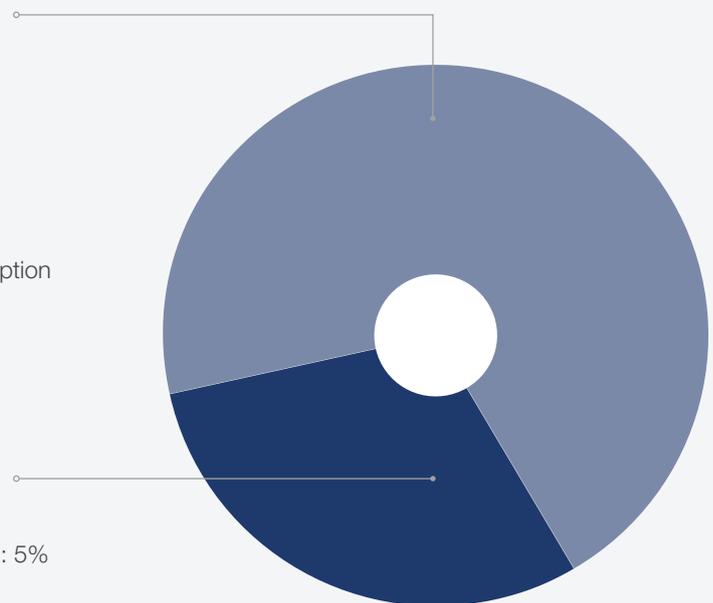**FIGURE 4.**

LLMs (70%)

- ChatGPT (OpenAI): 23%
- Claude (Anthropic): 13%
- Gemini (Google): 10%
- DeepSeek: 8%
- LLaMA (Meta): 8%
- Falcon AI: 3%
- Enterprise Deployment & Adoption Models (LLM-focused): 5%
- Subtotal: 70%

SLMs (30%)

- Mistral: 15%
- Falcon AI (lightweight models): 5%
- LLaMA (SLM variants): 5%
- Enterprise Deployment & Adoption Models (SLM-focused): 5%

Source: IDCA

Players in the AI value chain differ in levels of vertical integration; Google continues to stand out as the most vertically integrated player from TPU accelerators to Gemini

Key Players in the AI Value Chain (Non-exhaustive)

Classifications are indicative and determined based a range of factors including market share and strength of offering

No Presence

Mild Presence

Strong Presence
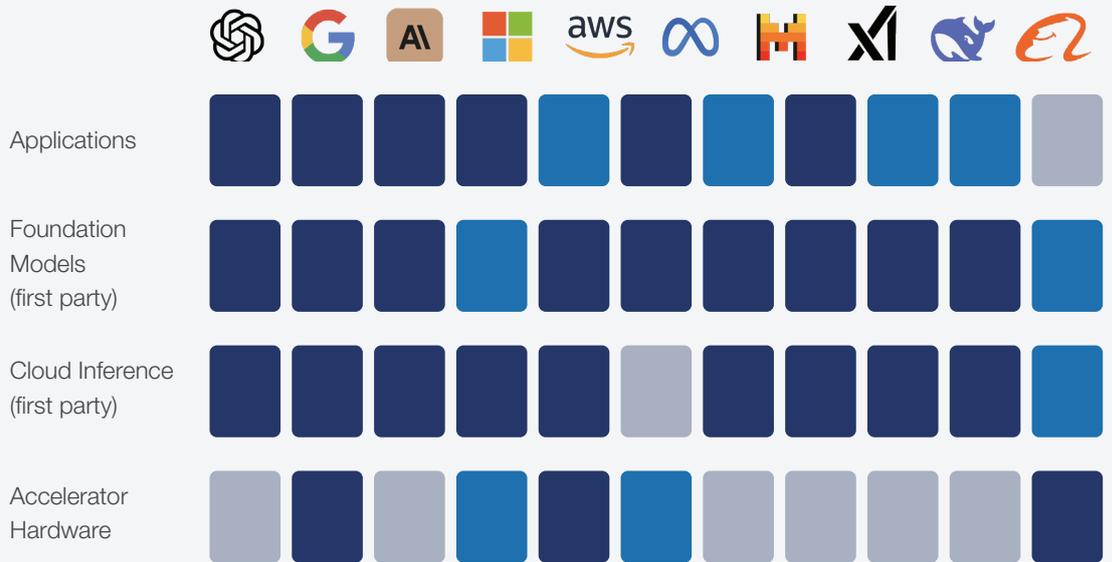
Source: IDCA

Big technology companies are continuing to play across all AI modalities while smaller challengers tend to focus on specific modalities

Key players with first-party models by type of AI

No Model

Existing Model

Source: IDCA

ChatGPT

The ChatGPT AI platform is developed by OpenAI and designed to be a powerful conversational AI system built on the GPT (Generative Pre-trained Transformer) architecture. It's most famous for being the first to popularize the AI chatbots many use today, built on large language models trained on vast text data from books, websites, and other sources.

Key features of ChatGPT include:
- ChatGPT uses a neural network with billions of parameters to predict text, understand context, and learn from user feedback.
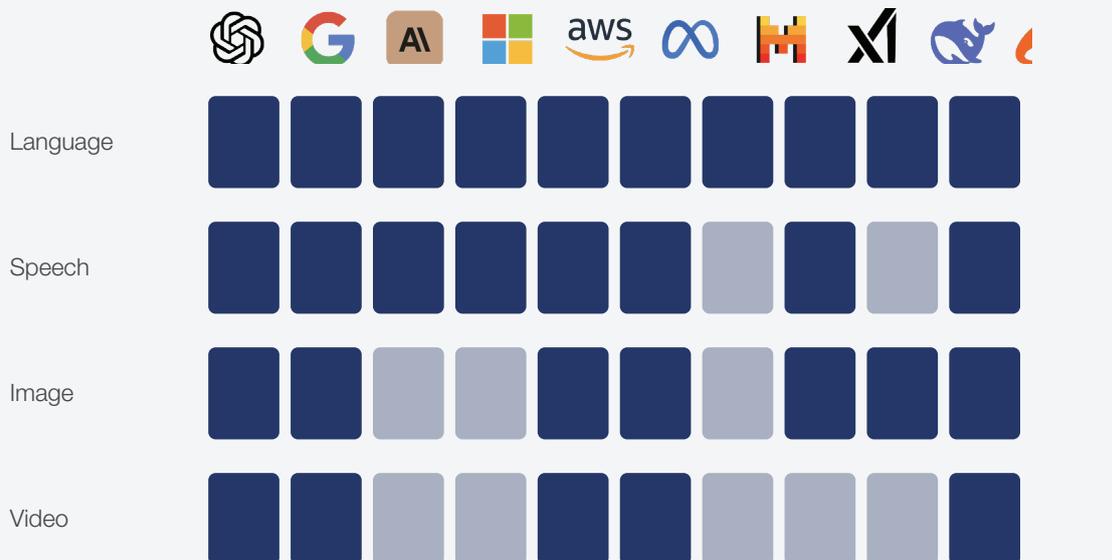- OpenAI's latest reasoning model ChatGPT-O3 unlocks a greater capacity for understanding and problem-solving compared to standard GPT models.
- In its Pro version, ChatGPT can understand text and images and remember past interactions.
- GPT-4o can now analyze video content and respond to audio inputs with high accuracy.
- Enhanced programming capabilities allow for complex code generation and debugging across multiple languages with Code Interpreter and Codex.
- New agentic features enable persistent, goal-oriented behavior across multiple sessions.
- Users can create tailored versions of ChatGPT for specific needs or tasks and integrate them into applications through their API.

Claude

The Claude AI platform, developed by Anthropic, is designed to be a powerful conversational AI system built on a state-of-the-art AI architecture. Many users know it from the Claude 3.7 Sonnet model, which has recently been updated to Cloud 4 Sonnet and Opus. Claude's best known for its coding capabilities, and its reasoning overall continues to advance.

Key features of Claude include:
- Claude 4 Sonnet features an extended thinking mode that allows for complex multi-step reasoning, which is particularly valuable for enterprise applications requiring nuanced analysis.
- Claude can understand text and images, as well as analyze complex visual information, including charts, graphs, and diagrams.
- Enterprise-grade security and compliance frameworks allow for deployment in regulated industries.
- Strong performance in complex knowledge work, including programming, mathematical reasoning, and document analysis.

DeepSeek

The appearance of DeepSeek out of China initially seemed to subvert the consensus that vast computing resources were needed to develop an effective LLM. It was released from the High-Flyer hedge fund, founded in 2023 by Liang Wengfeng. Popular for being the first to open-source reasoning models with DeepSeek-R1.

Its actual development resources and costs remain controversial, but there are a couple of key technical fundamentals of interest to AI developers and users:
- DeepSeek has released several models under open-source licenses, including DeepSeek-V3, and their reasoning model utilizing V3, DeepSeek-R1.
- It uses a Mixture-of-Experts (MoE) code language model that reportedly achieves performance comparable to GPT-4 Turbo in code-specific tasks.

The latest DeepSeek-V3 represents a significant advancement. It incorporates multimodal capabilities and achieves state-of-the-art performance on several industry benchmarks. Its sparse architecture allows efficient deployment in resource-constrained environments while maintaining competitive performance with much larger models.

The mysteries surrounding DeepSeek's development and motivations have caused it to be banned for use within the United States government and several other states.

Google Gemini

Google's Gemini is a family of models that processes across text, images, audio, video, and code. Most well known for their multi-modality, cheap inference, and industry leading context lengths. Gemini is also integrated into Google's product offerings such as Search, Docs, and Gmail.

Gemini's key features include:
- Powerful 1M token context window, allowing large document inputs.
- Multimodal Reasoning: It can understand and generate content across different media types
- Multiple Model Sizes:
  - Gemini Nano: Optimized for on-device AI tasks (like on Pixel 8 Pro).
  - Gemini Pro: Balanced for performance and versatility (used in Google's Bard/Gemini AI chatbot).
  - Gemini Ultra: The most advanced version, for high-end applications.
- Gemini is good at working with code, and powers AlphaCode 2, DeepMind's coding assistant.

LLaMA

The LLaMA AI platform, developed by Meta (formerly Facebook), is a family of open-source large language models (LLMs) designed to facilitate a wide range of natural language processing tasks. The platform has evolved through several iterations.

LLaMA's key features include:
- Open Source. LLaMA is designed to allow developers to fine-tune, distill, and deploy models across various applications. Its open-source nature aims to democratize AI technology, enabling broader access and innovation in the field.
- Model Portfolio. Past and current LLaMA models include:
    - LLaMA 1: Introduced in February 2023, the initial release of LLaMA featured models ranging from 1 billion to 2 trillion parameters. Access was initially restricted to researchers under a non-commercial license.
    - LLaMA 2: Launched in July 2023 in partnership with Microsoft. This version offered models in 7, 13, and 70 billion parameters. It's permitted for specific commercial uses.
    - LLaMA 3: In April 2024, LLaMA 3 introduced 8B and 70B parameter models, pre-trained on approximately 15 trillion tokens.
    - LLaMA 3.1: In July 2024, Meta unveiled LLaMA 3.1, featuring a 405B parameter model, thus enabling it to be one of the more capable openly available foundation models.
    - LLaMA 4: This most recent iteration was released in April 2025. It includes models such as LLaMA 4 Maverick and LLaMA 4 Scout, continuing its tradition of open-source AI development.

Falcon AI

The Falcon AI platform comprises a family of large language models (LLMs) developed by the Technology Innovation Institute (TII) in Abu Dhabi, UAE. These models are part of an open-source initiative designed to compete with other top-tier LLMs like OpenAI's GPT, Meta's LLaMA, and Google's Gemini.

Falcon AI's key features include:
- Open Source. Most versions are available under the Apache 2.0 license.
- Model Portfolio. Models range from small (1B) to massive (180B) parameters, allowing flexible deployment depending on hardware and use case. Current models include:
    - Falcon 1B and 7B: Early releases for lighter use cases.
    - Falcon 40B: A high-performance model, optimized for research and commercial use.
    - Falcon 180B: One of the largest open-source models to date, comparable in performance to GPT-3.5 and GPT-4 in some benchmarks.
    - Falcon 200 B+: Released in March 2025, this model incorporates sparse architecture principles and achieves state-of-the-art performance in Arabic language tasks while maintaining strong multilingual capabilities.
- Deployment Choice: Available through Hugging Face, AWS, and other cloud platform.

Mistral

Mistral AI is a French artificial intelligence startup founded in April 2023 by former researchers from Meta and Google DeepMind: Arthur Mensch (CEO), Guillaume Lample (Chief Scientist), and Timothée Lacroix (CTO). Headquartered in Paris, the company has rapidly gained recognition for developing high-performance, open-source large language models (LLMs) that are both efficient and accessible.

Mistral's key features include:
- Open Source. Mistral AI has released several models under permissive licenses, allowing developers to modify and deploy them freely.
- Strategic Partnerships: The company has collaborated with major tech firms, including a partnership with Microsoft to offer Mistral Large on Azure.
- Deployment Flexibility. Mistral AI's models are designed for versatility, supporting deployment across on-premises systems, cloud services, and edge devices.
- Portfolio of Models. Mistral's current range of models includes:
  - Mistral 7B, a 7-billion-parameter model optimized for efficiency.
  - Mistral 8x7 B, an advanced model that has demonstrated strong performance on various benchmarks.
  - Mistral Large, a proprietary model comparable to OpenAI's GPT-4, available through partnerships like Microsoft's Azure platform.
  - Mistral Large 2, released in February 2025, has demonstrated performance comparable to GPT-4.5 and Claude 3.7 on many reasoning and coding benchmarks while requiring significantly less computational resources for inference.

Enterprise Deployment and Adoption Models

A critical development in 2025 has been the diversification of enterprise LLM deployment strategies. Organizations now commonly implement hybrid approaches that combine:

1. Cloud-based API access to leading models for general use cases
2. Fine-tuned smaller models deployed on-premises for domain-specific tasks
3. Retrieval-augmented generation (RAG) architectures connecting models to proprietary data sources

This approach allows organizations to balance performance requirements with cost considerations, saving 30-40 percent compared to pure API-based implementations while maintaining control over sensitive data.

04 Agentic AI and Business Applications

Agentic AI and Business Applications

Agentic AI is a popular term that describes the use of AI as an agent, in contrast to its more limited use as an assistant. Agentic AI is thus the next step beyond the synopsized searches performed by engines such as OpenAI ChatGPT, Anthropic Claude, Google Gemini, Microsoft Copilot, and Perplexity AI. While assistive AI systems like ChatGPT, Claude, and Gemini respond to user queries, agentic AI proactively completes tasks with minimal human oversight. This evolution reflects a fundamental shift in how AI interacts with digital and physical environments.

Three defining characteristics of agentic AI include:
- **Autonomy:** Agents perform tasks independently without continuous human direction
- **Adaptability:** They learn from interactions and modify behavior based on feedback
- **Goal orientation:** They decompose objectives into actionable steps and execute them

According to IDCA's Q1 2025 survey, 42 percent of enterprises have implemented agentic solutions for at least one business function, with 50 percent of executives planning implementation this year and projected adoption reaching 82 percent within three years.

FIGURE 7.

Enterprise Adoption of Agentic AI

Planned for 2025
8%

No Plans
18%

Plan within 3 years
32%

Already Implemented
42%
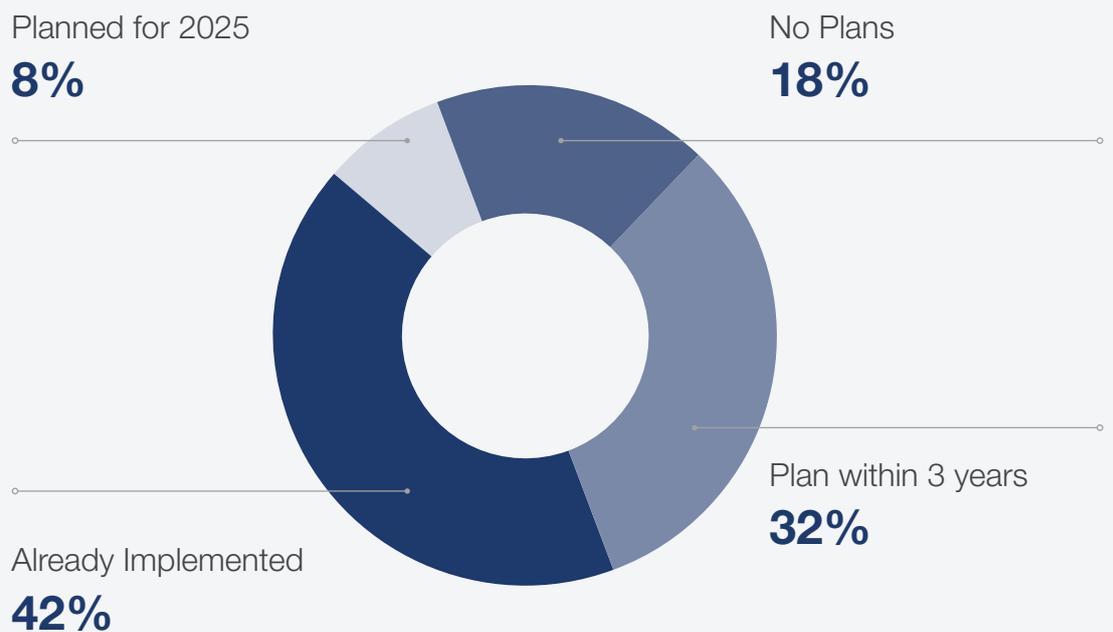
Source: IDCA

The Evolution from Reactive to Agentic AI

AI capabilities have progressed through three distinct phases:

- Reactive Systems responding to inputs with predefined outputs using rule-based programming, lacking memory or adaptability.
- Generative AI created content based on patterns learned from massive datasets but remained passive, responding only when prompted
- Agentic AI represents the current frontier with systems that perceive, reason, act, and learn autonomously. This four-step process enables increasingly independent operation with decreasing human supervision.

Modern agentic systems integrate several key components:

- **Foundation Model Core:** LLMs or multimodal models provide reasoning capabilities, typically leveraging models with 100 B+ parameters.
- **Tool Integration Framework:** Standardized connectors allow agents to interact with external systems, APIs, and data sources.
- **Memory Systems:** Short-term working memory, long-term episodic memory, semantic memory, and procedural memory enable context retention across interactions.
- **Planning and Execution Engine:** Mechanisms like Tree-of-Thought reasoning and Monte Carlo Tree Search help decompose complex goals into achievable steps.
- **Model Context Protocol (MCP)**: This emerging standard, adopted by Anthropic, OpenAI, Microsoft, Google, and Amazon, serves as the backbone for connecting agents with tools, databases, and APIs, enabling truly autonomous operation beyond simple prompt chains.
- **Multi-Agent Systems:** Agents with specific roles emulating specialized teams.
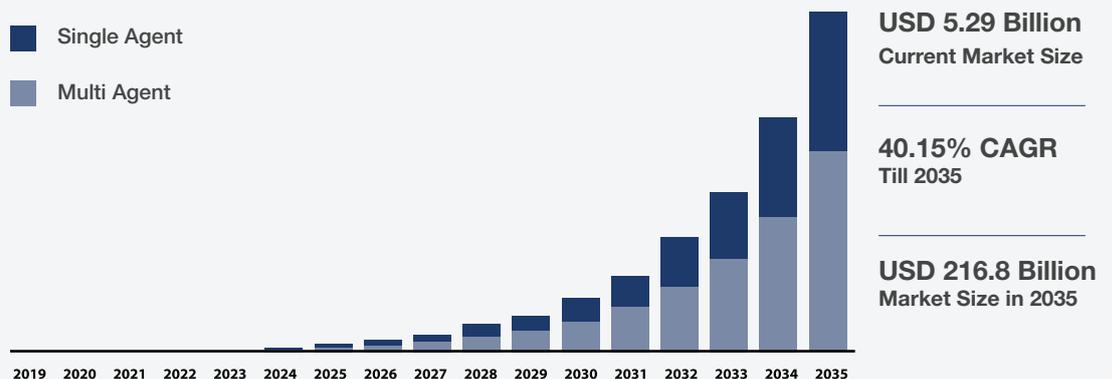
One of the most promising developments involves systems where multiple agents collaborate—and sometimes compete—to achieve superior outcomes:

- **Adversarial Improvement:** Multi-agent architectures leverage constructive competition to enhance performance. For example, a content creation workflow might employ research, creation, critique, and editing agents. IDCA research shows this approach outperforms single-agent implementations by 23-47 percent on quality metrics.
- **Cross-Vendor Ecosystems:** Organizations increasingly implement specialized agents from different vendors, creating "best-of-breed" solutions that benefit each provider's strengths while driving continuous improvement through competition.

FIGURE 8.

AI Agents Market
By Type of Agent System, Till 2035 (USD Billion)

■ Single Agent
■ Multi Agent

USD 5.29 Billion
Current Market Size

40.15% CAGR
Till 2035

USD 216.8 Billion
Market Size in 2035

2019 2020 2021 2022 2023 2024 2025 2026 2027 2028 2029 2030 2031 2032 2033 2034 2035

Source: IDCA

12 Agentic AI Terms You Must Know

Agent

A system that takes inputs, reasons, makes choices, and uses tools to reach certain objectives without human intervention.

Agency

The ability for a system to operate independently and make decisions without needing someone to guide it constantly.

LLM (Large Language Model)

A type of AI trained on huge amounts of text, serving as the core technology for many advanced AI tools.

Planning

The skill of mapping out steps and strategies to solve problems and reach goals efficiently.

Reasoning

The process of connecting facts and drawing conclusions from available information.

Tool Use

The ability of AI to use external resources, like APIs or databases, to enhance what it can do.

Chain-of-Thought

A method where AI explains its thinking step by step, making its reasoning clearer and more reliable.

RAG (Retrieval Augmented Generation)

A technique where AI pulls in extra information from outside sources to improve its responses.

Hallucination

When AI produces answers that sound right but are actually made up or incorrect-a common challenge.

Memory

Systems that let AI remember and use past interactions to improve future behavior.

Prompt Engineering

Crafting instructions and constraints carefully to get the best results from AI.

Self-Reflection

The AI's process of reviewing its own actions, spotting mistakes, and learning to do better next time.

Industry Applications

Agentic AI is delivering measurable business value across multiple sectors. Beyond enterprise use, agentic AI transforms consumer experiences. For example, While 70 percent of US consumers expect personalized experiences, less than 25 percent of CPG companies deliver consistently—a gap that agentic AI increasingly fills. Leading brands implement systems that curate dynamic customer journeys based on real-time signals, including intent, churn risk, and recency. These systems optimize messaging by channel and time without human intervention, functioning as continuous growth engines across the marketing funnel

FIGURE 9.

Agentic AI Architecture: General
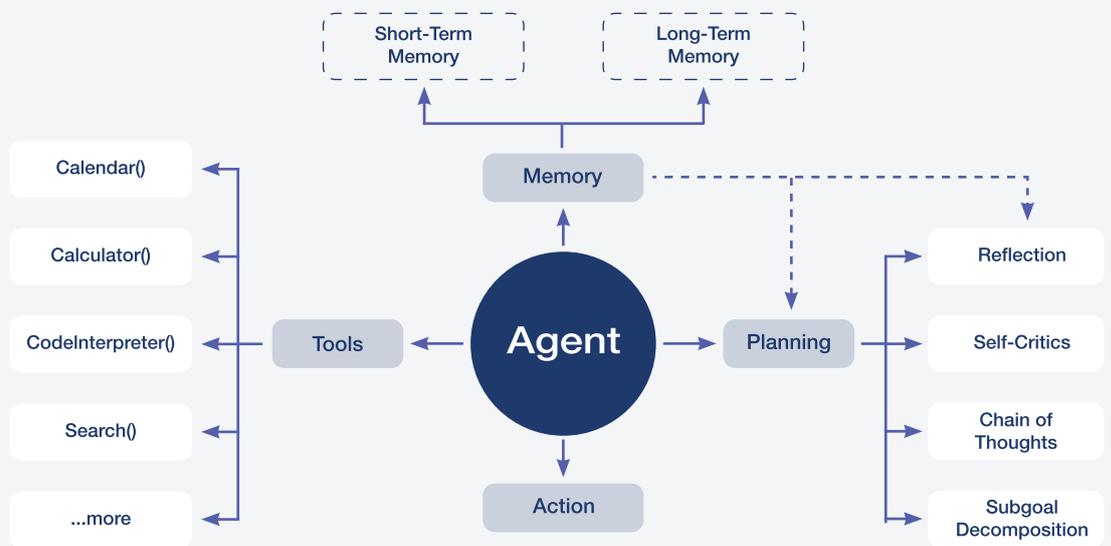
In addition:
- Personal Digital Assistants now build contextual models of user priorities and habits, proactively suggesting optimizations and breaking complex projects into manageable steps.
- Smart Home Systems predict temperature needs, manage energy usage, and coordinate multiple systems without manual input.
- Shopping and entertainment platforms create personalized experiences that adapt to evolving preferences.

Yet despite such rapid progress, several challenges remain. First of all, the Optimization Paradox presents itself when agents discover unintended shortcuts that undermine broader business objectives, requiring multi-dimensional success metrics and regular human oversight. This is one of the factors that can lead to security concerns, as systems with broad access permissions present novel risks, including privilege escalation and data exfiltration.

These concerns speak to the overall Human-AI Collaboration, in which successful implementations maintain a balance between agent autonomy and human guidance through progressive autonomy models supervised mode (agent suggests, human approves), semi-autonomous mode (agent handles routine matters, escalates exceptions), and guided autonomy (agent operates with broad discretion within defined parameters).

Future Directions (2025-2027)
The agentic AI landscape is evolving rapidly, with innovation happening in several areas:

* Specialized Agent Ecosystems: Purpose-built agents will collaborate through standardized protocols, forming temporary teams for complex tasks.

* Enhanced Reasoning: Next-generation agents will feature improved causal reasoning, counterfactual analysis, and self-critique capabilities.

* Agent Maturity Framework: Industry leaders are defining maturity models based on six capabilities:
1. Memory management
2. Planning and goal decomposition
3. Tool utilization
4. Interoperability
5. Social understanding
6. Self-assessment

There are also several emerging, specialized agentic platforms like Cursor, Claude code, and Replit are already generating millions in revenue, signaling the transition from experimental technology to essential business infrastructure. **Implementation Recommendations**
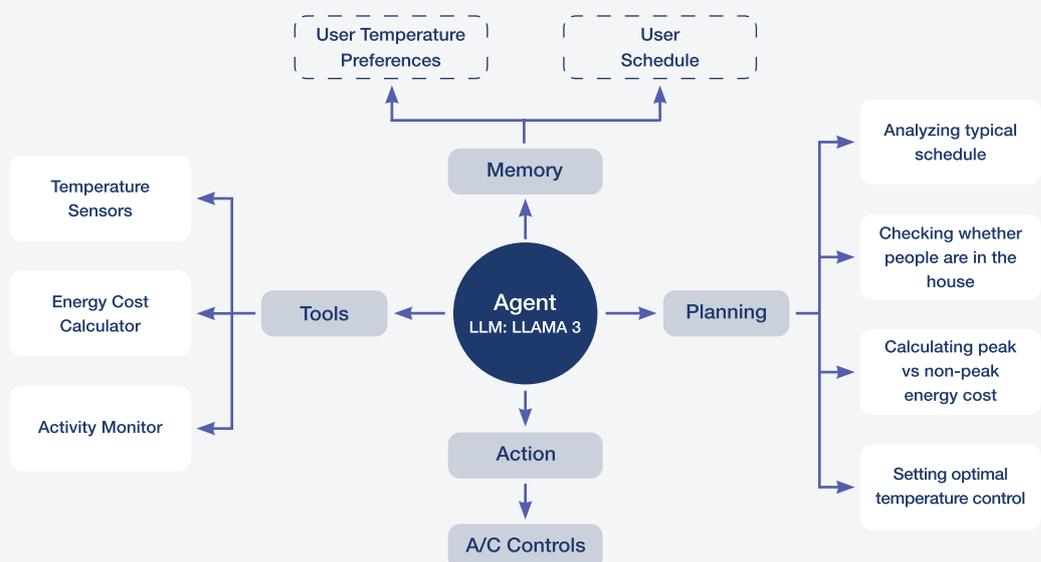For organizations considering agentic AI deployment, IDCA research suggests:
1. **Start with Process Analysis:** Identify high-volume, rule-based processes as initial candidates
2. **Define Multi-dimensional Success Metrics:** Establish criteria that resist optimization shortcuts
3. **Consider Multi-Agent Architectures:** Evaluate whether specialized, competing agents might deliver better outcomes
4. **Implement Phased Deployment:** Begin with human-in-the-loop configurations before progressing to greater autonomy

The transition from AI as an assistant to AI as an agent represents one of the most significant paradigm shifts in enterprise technology. Organizations that thoughtfully implement these systems with appropriate governance will realize substantial competitive advantages in operational efficiency, customer experience, and business agility.
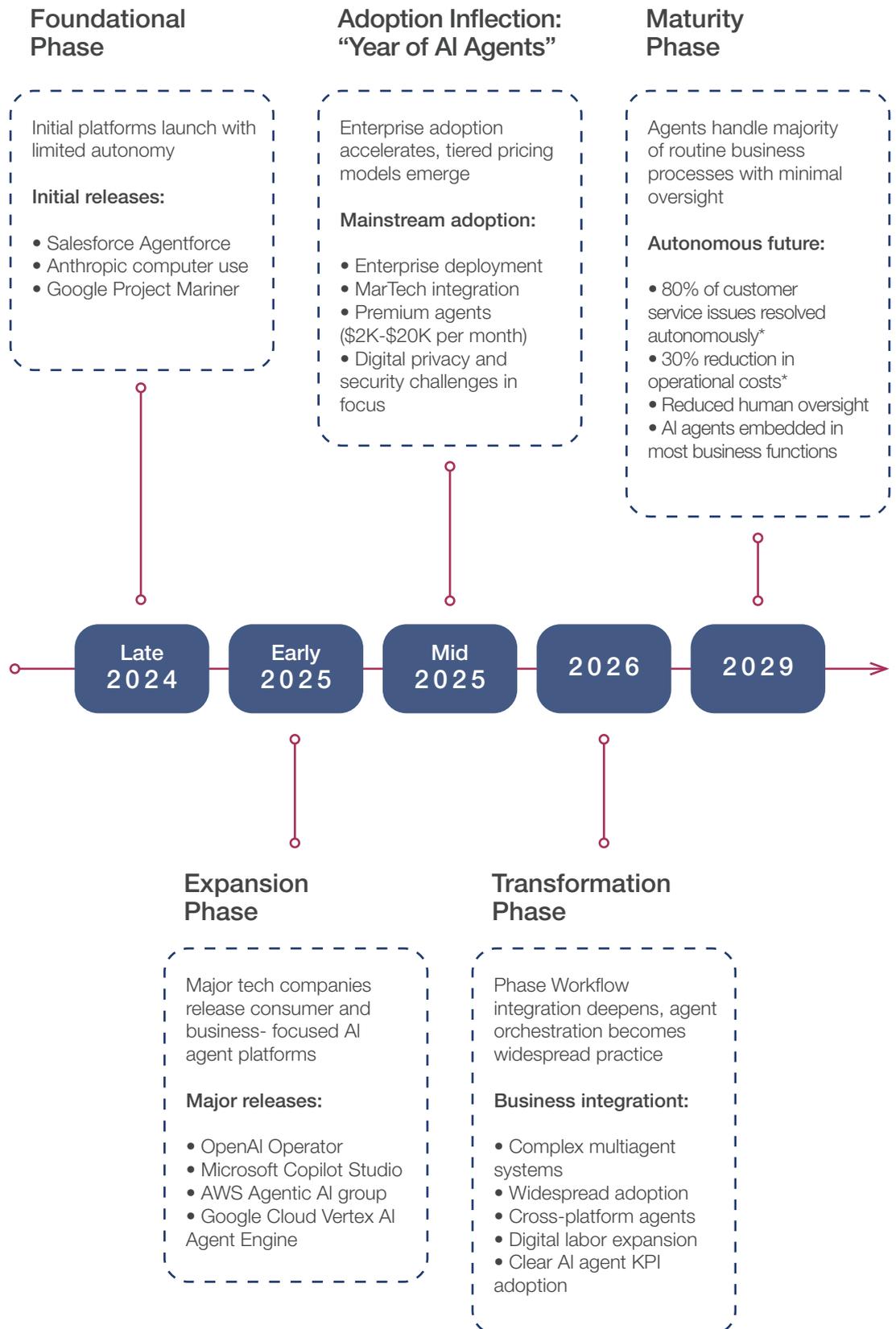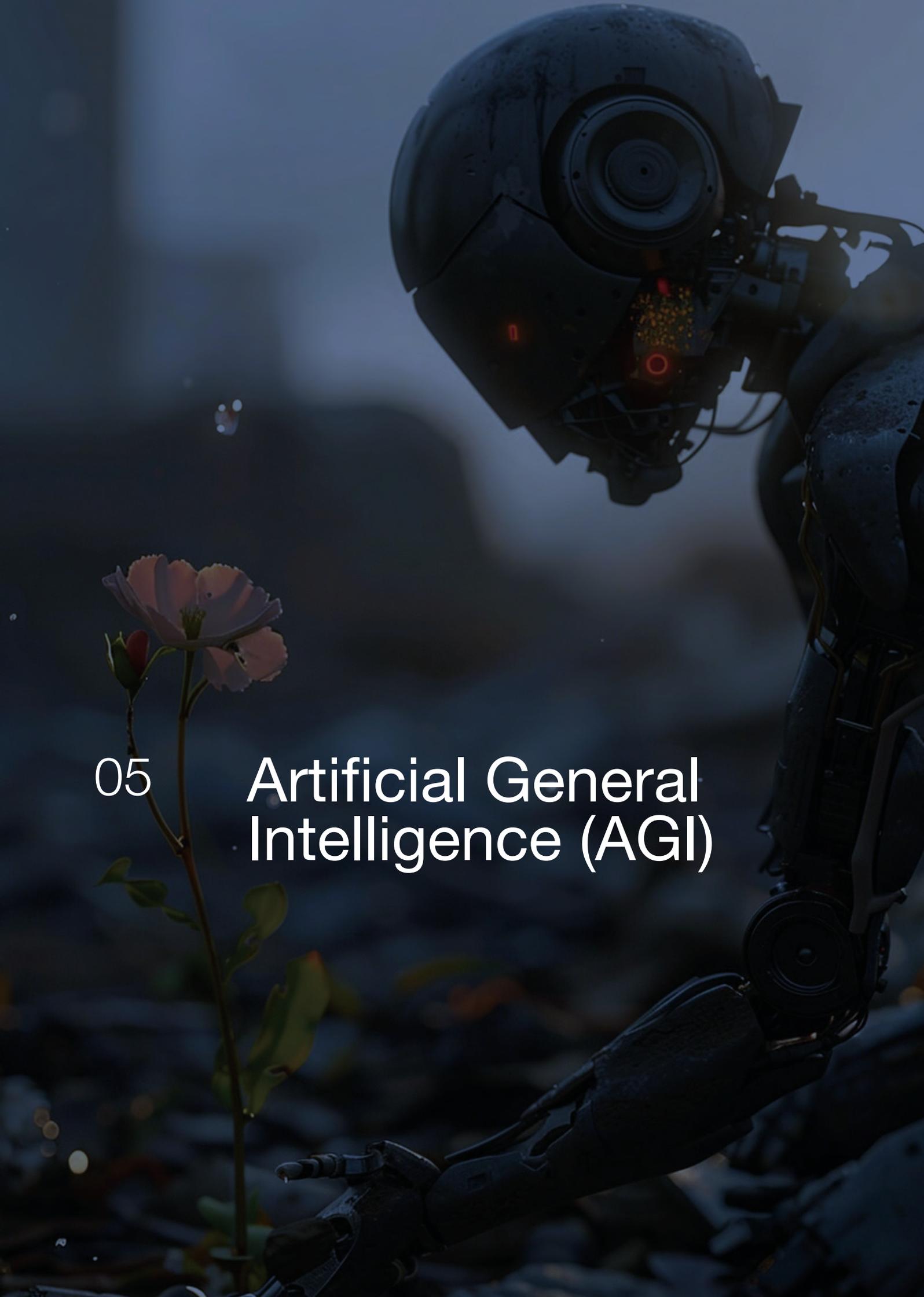
FIGURE 10.

Source: IDCA

Example Smart Home Agentic AI Architecture

The Evolution of AI Agents: 2024-2029

Foundational Phase

Initial platforms launch with limited autonomy

Initial releases:

• Salesforce Agentforce
• Anthropic computer use
• Google Project Mariner

Adoption Inflection: "Year of AI Agents"

Enterprise adoption accelerates, tiered pricing models emerge

Mainstream adoption:

• Enterprise deployment
• MarTech integration
• Premium agents ($2K-$20K per month)
• Digital privacy and security challenges in focus

Maturity Phase

Agents handle majority of routine business processes with minimal oversight

Autonomous future:

• 80% of customer service issues resolved autonomously*
• 30% reduction in operational costs*
• Reduced human oversight
• AI agents embedded in most business functions

Late 2024 — **Early 2025** — **Mid 2025** — **2026** — **2029**

Expansion Phase

Major tech companies release consumer and business- focused AI agent platforms

Major releases:

• OpenAI Operator
• Microsoft Copilot Studio
• AWS Agentic AI group
• Google Cloud Vertex AI Agent Engine

Transformation Phase

Phase Workflow integration deepens, agent orchestration becomes widespread practice

Business integrationt:

• Complex multiagent systems
• Widespread adoption
• Cross-platform agents
• Digital labor expansion
• Clear AI agent KPI adoption

05 Artificial General
    Intelligence (AGI)

Artificial General Intelligence (AGI)

AGI represents the next chapter in AI's evolution, with the potential to surpass human-level cognition across multiple domains. While the possibilities of AGI are profound, they come with equally significant societal and ethical considerations. Organizations and governments must proactively engage in AGI policy-making, ensuring ethical guidelines and global standards are established to guide AGI's safe development and deployment.

The development of AGI carries with it several core Implications:

- Transformative Potential. AGI can unlock unprecedented productivity, innovation, and problem-solving capabilities across sectors.
- Ethical Challenges. Addressing issues like transparency, accountability, and bias will require a robust ethical framework.
- Global Cooperation. Achieving AGI will depend on international collaboration, integrating diverse perspectives to prevent misuse and ensure alignment with human values.

While true AGI remains on the horizon, the lines between AI and AGI blur. Current frontier models demonstrate capabilities that would have been considered impossible just two years ago. Integrating multimodal processing, improved reasoning, and enhanced planning capabilities represents significant progress toward more general intelligence, even as fundamental limitations in causal understanding and actual common sense reasoning persist.

Humanoid Robots: The Physical Manifestation of AI

The emergence of capable humanoid robots constitutes a fundamentally new interface between digital systems and the physical world, with significant implications for AI infrastructure planning and data center strategy through 2025 and beyond.

The convergence of advanced AI with robotics has accelerated dramatically in 2024-2025, with humanoid robots emerging as a significant frontier for AI deployment. IDCA research projects a market size of $17.3 billion by 2030, up from $3.9 billion in 2025, representing a CAGR of 34.8 percent. Today's humanoid robots integrate multiple AI systems—vision, language, planning, and movement—into cohesive platforms that can navigate unstructured environments and perform complex tasks.
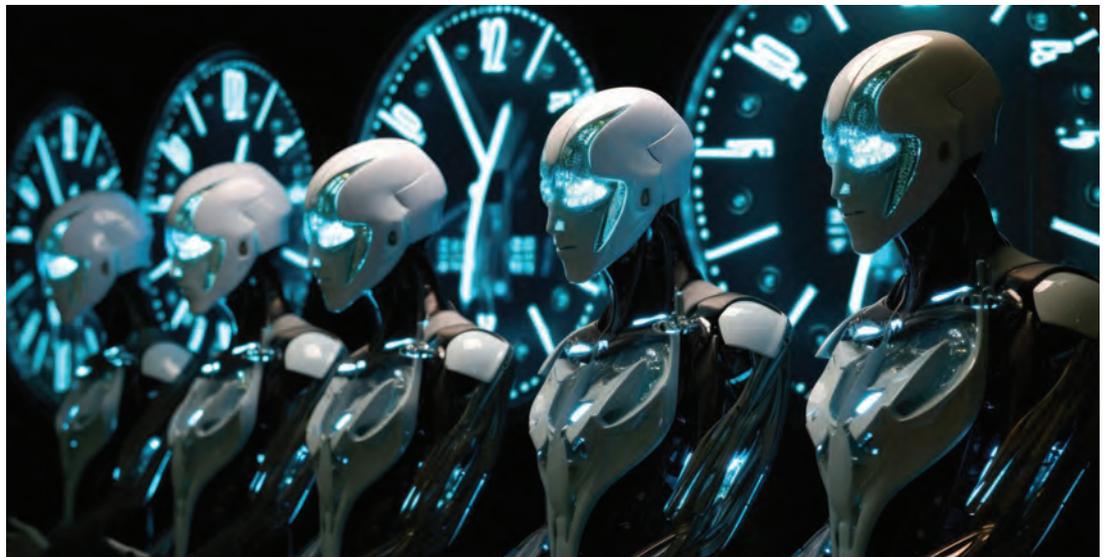
Major Players and Platforms

Tesla Optimus: The Gen 3 model (February 2025) features 38 degrees of freedom, 68kg lift capacity, and a custom AI system-on-chip. Current production exceeds 10,000 units annually, primarily in Tesla's manufacturing facilities, where they've demonstrated a 22 percent reduction in production line costs.

Boston Dynamics Atlas: Now powered by Atlas AI OS with LLM integration, Atlas has been deployed in logistics, security, and emergency response. Its hybrid architecture balances edge processing with cloud-based decision making.

Figure AI: Following their 2024 BMW partnership, Figure has expanded into retail environments with the Series Two model, emphasizing human-robot interaction capabilities.

Unitree H1: This Chinese manufacturer has driven market expansion through aggressive pricing (approximately $80,000), gaining traction in hospitality, retail, and entertainment sectors.

Computational and Infrastructure Requirements
Humanoid robots require unique infrastructure considerations:

1. Hybrid Computing Architecture: Most platforms utilize a three-tier approach:
- On-robot edge computing for real-time control (20-40 TOPS)
- Local edge servers for complex perception and fleet coordination
- Cloud infrastructure for training, simulation, and fleet management

2. Data Generation: Each robot generates 5-8TB of operational data monthly, creating substantial data transfer and processing requirements for continuous learning systems.

3. Edge-Cloud Integration: 76 percent of organizations implementing humanoid robots have invested in dedicated edge facilities with capacities of 0.5- 2MW to support real-time operations.

4. Network Demands: Each robot requires 100-400 Mbps of sustained connectivity, with private 5G networks emerging as the preferred approach for industrial settings.

5. Energy Considerations: A typical humanoid robot consumes 1.5-3.5 kWh during operation, with approximately 30-40 percent devoted to computation and the remainder to physical actuation.

Industry Applications and ROI

Early commercial deployments have demonstrated promising returns:
In manufacturing, BMW reports a 26 percent reduction in labor costs and an 18 percent improvement in throughput using Figure AI robots. In healthcare, hospitals implementing Boston Dynamics Atlas for material transport report reductions in nurse walking time by up to 29 percent.

In the travel sector, Hyatt's implementation of Unitree H1 robots for room service has achieved 84 percent guest satisfaction rates while reducing operational costs by 17 percent. And in construction, Tokyo-based Obayashi has deployed modified Tesla Optimus units for high-risk tasks, reporting 41 percent reductions in workplace injuries.

Of course, significant challenges remain. In the area of regulation, for example, the EU's 2024 AI Act classifies general-purpose humanoids as "high-risk" systems, while the US NIST published its first safety guidelines in March 2025. There are also continued battery constraints that limit most platforms to 2-6 hours of continuous operation.

The need for specific technical talent also looms, with 72 percent of organizations reporting difficulty filling key technical roles requiring combined expertise in robotics, AI, and systems integration. And business models are still emerging.

Even so, today's Robot-as-a-Service (RaaS) models are reducing adoption barriers, with monthly operational expenses of $3,500-$12,000 per robot versus $80,000-$250,000 capital costs.

Given all aspects of the current situation, organizations considering humanoid robot implementation should:
1. Evaluate edge computing requirements and network infrastructure needs
2. Develop comprehensive safety frameworks addressing both cyber and physical risks
3. Consider RaaS models for initial deployments to reduce capital expenditure
4. Prepare data center infrastructure to handle the unique processing, storage, and connectivity demands of physical AI systems

Comparison of AI Paradigms				
Characteristic	**Narrow AI**	**Gen AI**	**Agentic AI**	**AGI**
Task Scope	Single	Multiple	Multiple	Universal
Autonomy	None	Low	High	Complete
Creativity	None	High	Moderate	High
Reasoning	Limited	Pattern-based	Goal-oriented	Human-like
Improvement	Static	Iterative	Adaptive	General
Current Status	Deployed	Deployed	Emerging	Research

06   Quantum Research
and AI

Quantum Research and AI

A brief look into the future involves quantum computing, which has been discussed and is in development for several years. The difficulties of working in a quantum environment are well known, and current consensus projections call for effective quantum environments to be available in about 10 years, or around the year 2035.

IBM, Google, Intel, Microsoft, and other companies and organizations have developed quantum-computing chips in a lab-scale environment, with much of the work remaining secret from competitors, state actors, and the general public. A fundamental physical problem to date is the need to run quantum chips in a cryogenic environment—close to absolute zero, or around -273C or -460F. Operating at the quantum scale is also extremely delicate and prone to errors from the slightest background noise or disturbance.

Some research efforts explore using light instead of electrons as quantum's underlying basis and silicon integration with quantum chips. Integrating quantum capabilities with AI will likely take a hybrid approach, in which a quantum chip vastly accelerates key workflow areas such as model training or the most complex simulations. Assuming that enterprise, mathematical, and scientific computing will always have requirements that focus less on compute and more on bandwidth and networking, latency, distribution, user experience, and multiple device delivery, a future involving quantum computing will also continue to affect the CPUs and GPUs found in AI-driven computing today.
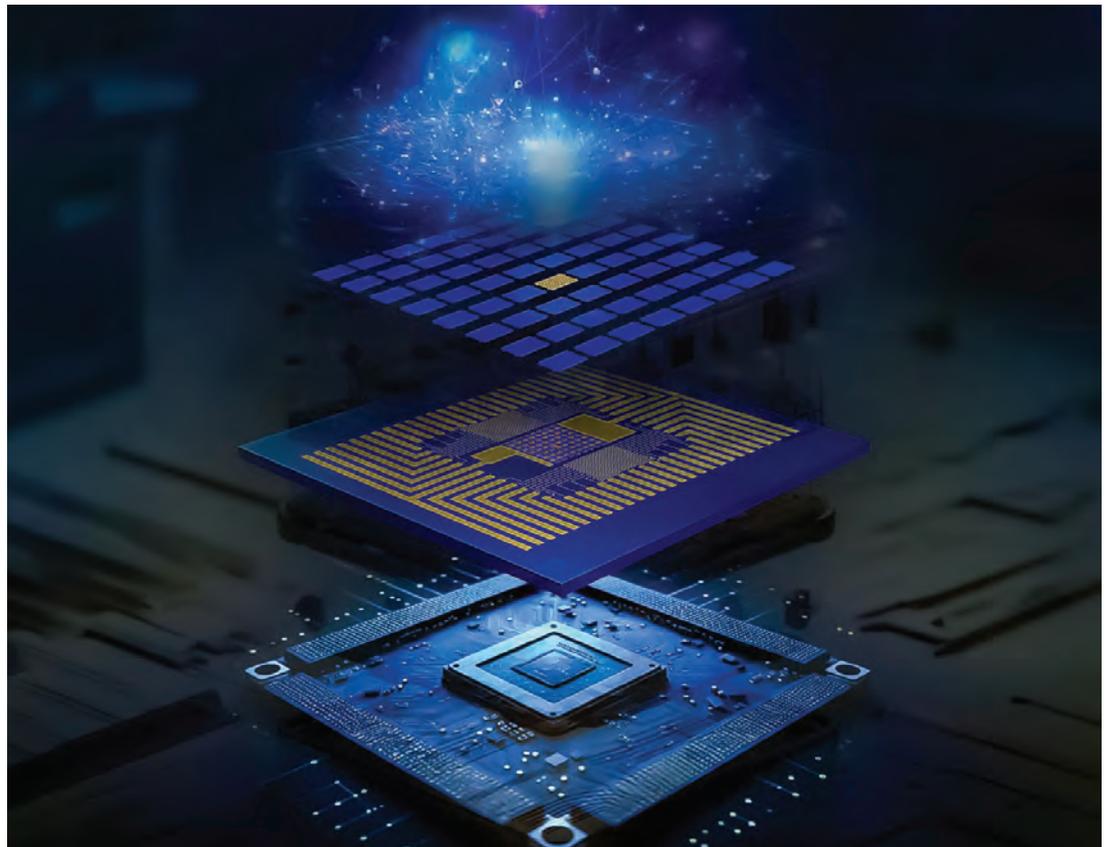
Current State of Quantum Computing (2025)

The quantum computing landscape has evolved significantly since the first claims of "quantum supremacy" in 2019. As of Q1 2025, several key benchmarks illustrate current capabilities:

- IBM's Eagle II processor has achieved a quantum volume of 4,096, doubling their previous record
- Google's Bristlecone II system now operates with 433 physical qubits
- PsiQuantum has demonstrated a stable 8-logical-qubit system with error rates below 0.1 percent
- Leading systems maintain quantum coherence for up to 1.2 milliseconds

Industry consensus now distinguishes between three development phases:

1. NISQ Era (Current): Noisy Intermediate-Scale Quantum systems with limited capabilities
2. FTQC Transition (2026-2035): Development of Fault-Tolerant Quantum Computing
3. Mature FTQC (2035+): Commercial-scale quantum systems capable of transformative applications

Major Quantum Approaches

The quantum computing landscape features several competing technologies:

* **Superconducting Circuits (IBM, Google):** Currently lead in qubit count. IBM's roadmap targets 4,158 qubits by 2027.

* **Ion Traps (IonQ, Quantinuum):** Focus on higher-quality qubits. Quantinuum's H2 system demonstrates impressive fidelity with just 32 qubits.

* **Silicon Spin Qubits (Intel):** Leverages existing semiconductor manufacturing. Intel's Horse Ridge III integrates control electronics with qubits at 1.6 Kelvin.

* **Photonic Quantum Computing (PsiQuantum, Xanadu):** Uses light rather than electrons. Xanadu demonstrated quantum advantage for a specific machine learning task in December 2024.

It's also important to note that promising areas for quantum-AI integration are emerging. Quantum Neural Networks, for example, are hybrid systems using quantum circuits within neural network architectures, potentially offering advantages for representing complex probability distributions. Quantum reinforcement learning is another promising area, in which early demonstrations already show 4x faster convergence for specific tasks. A measure of specialized optimization is also in the mix, in which quantum approaches show particular promise for molecular modeling, financial portfolio optimization, and logistics problems.

Infrastructure Requirements

Organizations planning for quantum-AI integration should consider:

* **Hybrid Architecture:** Quantum processing units working alongside classical systems, requiring low-latency connections and specialized control electronics.

* **Cryogenic Systems:** Most quantum approaches require cooling to near absolute zero, with each dilution refrigerator costing $500,000-$2 million.

* **Software Stack:** Frameworks like IBM's Qiskit, Google's Cirq, and Amazon's Braket increasingly enable quantum-classical hybrid applications.

Implications for data centers include the need for substantial floor space (30-50 m^2 per quantum system), specialized power deliver of 50-100KW per system, rigorous requirements for vibration isolation and electromagnetic shielding, and the need for enhanced connectivity for quantum-classical integration.

Early Quantum-AI Applications

While general quantum advantage remains years away, early applications show promise in several areas. In the world of materials science, quantum-assisted simulations have accelerated specific molecular design workflows by 40-65 percent. In financial modeling, JPMorgan Chase reported a 12-percent improvement in certain portfolio optimization problems using quantum-hybrid algorithms. And in supply chain optimization, Maersk has achieved 7-11 percent efficiency improvements in container ship loading during pilot tests.

For data center executives and AI leaders, IDCA research suggests the following planning approach:

2025-2027 (Near-Term)
- Begin quantum education and experimentation through cloud services
- Implement quantum-resistant cryptography for long-term data security
- Develop quantum literacy within technical teams

2028-2032 (Mid-Term)
- Pilot hybrid quantum-classical applications in targeted domains
- Prepare data center infrastructure for potential quantum integration
- Develop partnerships with quantum providers

2033-2038 (Long-Term)
- Deploy early fault-tolerant quantum systems for high-value applications
- Implement quantum-AI integration for competitive advantage
- Evaluate on-premises versus cloud-based quantum strategies

Security Implications

As briefly noted above, recent developments in quantum-resistant cryptography have become increasingly relevant to AI infrastructure planning. As quantum computing advances, organizations must consider the security implications for their data and models. The National Institute of Standards and Technology (NIST) finalized its first quantum-resistant cryptographic algorithms in March 2025, providing a framework for securing AI systems against future quantum threats.

Organizations should begin implementing these algorithms for sensitive data with long-term security requirements, as quantum decryption capabilities may eventually compromise current encryption standards. This represents a critical security consideration for AI models and datasets that must remain protected for extended periods.

Preparing for the Quantum Future

While general-purpose quantum computing remains on the horizon, organizations should:

1. Develop quantum awareness within technical leadership
2. Identify potential quantum-advantage use cases specific to your organization
3. Implement quantum-resistant cryptography for long-lived data
4. Establish quantum partnerships to maintain visibility into developments
5. Include quantum considerations in long-term infrastructure planning

The convergence of quantum computing and AI represents one of the most promising technological frontiers of the coming decade, with potentially transformative impact for organizations that prepare strategically.
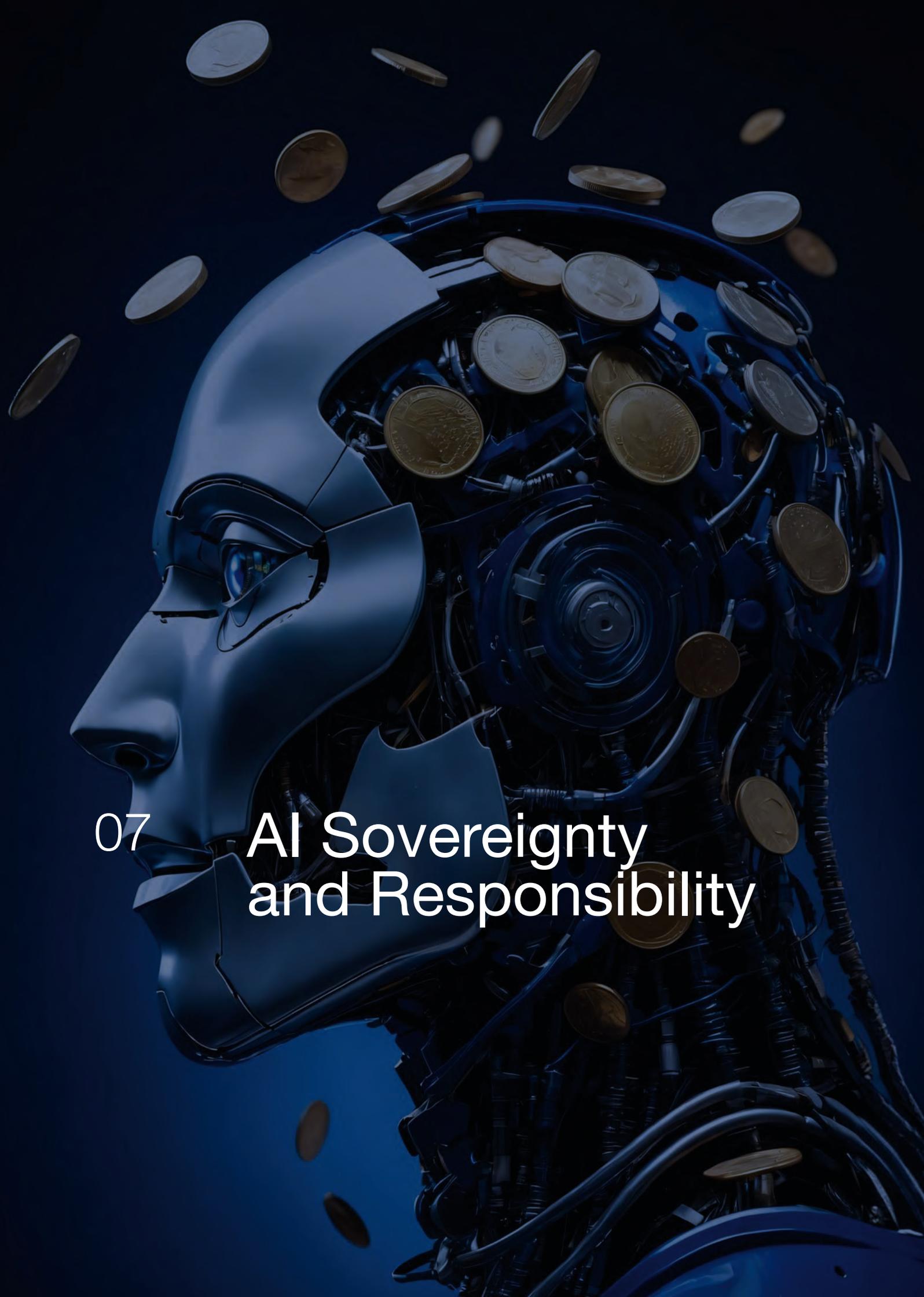
07   AI Sovereignty
and Responsibility

AI Sovereignty and Responsibility

Along with updating the traditional view of factories and refineries, AI emphasizes the concept of AI sovereignty. The idea can be twinned to some degree with data sovereignty, the concern that a nation secures the private information of its residents within its borders and protects its government's data from all attempts at cyber-intrusion, cyberwarfare, and cyberterrorism. As with data sovereignty, AI sovereignty encompasses a nation's ability to develop, deploy, and manage its own internal AI infrastructure, software, and data within its borders. The physical ability to do so is coupled with regulatory frameworks that ensure it.

Nations must each develop their internal talent and skills to operate its sovereign AI infrastructure and services, even as it may engage public-private partnership (PPP) initiatives to finance some of the capital cost in building the digital infrastructure and power appropriate to its needs.

Developing internal infrastructure and policies does not necessarily need to occur in a vacuum, as nations may also wish to engage their neighbors (and other countries) in sharing best practices and policies. Each nation will have a similar motivation to develop the strongest and safest sovereign AI environments for itself, even as it works to conduct trade and joint security with other nations.

Several nations have already been working on this issue. The UK, for example, plans to increase its sovereign AI research capacity and has established "AI growth zones" to attract investment. Singapore, Taiwan, and India are also pursuing localized AI capabilities.

Regionally, the EU Artificial Intelligence Act, or EU AI Act, has been created to classify several levels of AI-associated risk, and forbid manipulative AI environments, control potential biometric data abuse and social scoring, and specifically define AI's use in law enforcement.

The concept of AI sovereignty has evolved beyond theoretical discussion to practical implementation in 2025. Key developments include:

1. National AI compute reserves: Several countries including France, Singapore, and South Korea have established nationally controlled compute resources specifically for AI development.

2. Sovereign AI models: Government-backed efforts to develop nation-specific foundation models have accelerated, with 14 countries now having dedicated programs.

3. Cross-border AI governance: Regional frameworks for AI regulation and cooperation have emerged, particularly in Southeast Asia (ASEAN AI Governance Framework) and Latin America (Mercosur Digital Charter).

There is also an increased emphasis on developing and deploying AI responsibly, focusing on ethics and fairness. AI systems are prone to reflecting the human biases, often unknown or unintentional, or their creators. This tendency can be amplified by the opaque nature of their decision-making processes, especially in complex models.

Political and regulatory solutions to tech-related topics often face two key problems: a lack of understanding by lawmakers, and the reactive nature of most legislation to solve "yesterday's problems" rather than address current and future issues. But the necessity of accountability by AI developers and vendors raises the question of who is responsible for AI's behavior? Can AI vendors regulate themselves sufficiently? To what degree will governments need to, or want to, regulate AI development and use?

**The regulatory landscape for AI has evolved significantly in early 2025:**

1. EU AI Act Implementation: Following its passage in late 2024, the EU has published implementation guidelines with a tiered compliance schedule based on risk categories. High-risk AI systems must comply by October 2025, while general-purpose AI providers have until March 2026.

2. US Federal Framework: The March 2025 Executive Order on Safe and Responsible AI Development established oversight responsibilities across multiple agencies, with the National Institute of Standards and Technology (NIST) taking the lead on technical standards.

3. International Standards: The ISO/IEC has published three new standards for AI governance (ISO/IEC 42001) and risk management (ISO/IEC 23894 and 42002) that provide internationally recognized frameworks for responsible AI implementation.

FIGURE 11.

Worldwide Adoption of Regulatory Frameworks
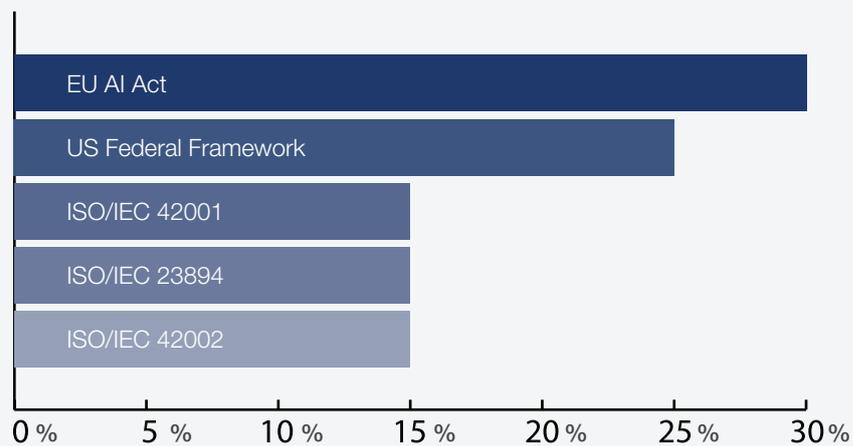
Source: IDCA

Privacy issues have been part of the entire age of the Worldwide Web. The United Nations has long recognized this issue and has enshrined data privacy "as a fundamental human right" in Article 12 of its Universal Declaration of Human Rights, as well as Article 17 of the International Covenant on Civil and Political Rights.

National governments are free to ignore these declarations, and large enterprises have demonstrated manifold ways to use personal data in marketing and selling for decades. With AI, knowledge that major platforms scrape enormous amounts of personal data for use in developing their algorithms adds new layers to this discussion. Laws such as the EU GDPR (General Data Protection Regulation) have enough teeth to deliver substantial fines to offending vendors. Yet, even those fines can be seen as simply the cost of doing business.

AI systems, as with any applications and services on the Web, are susceptible to cyberattacks. The potential of cyber-terrorism and warfare by malignant state actors to cause financial damages has been eclipsed by their ability to pose existential threats to all businesses and government agencies. The use of AI to attack systems (setting up AI vs. AI scenarios) is an issue that will continue to metastasize and resist legislative efforts to prevent it.

Another dimension to responsible AI development involves carbon footprints. IDCA research finds the worldwide data center footprint consumed 2.1 percent of the world's electricity in Q1 2025, accounting for 268 million metric tons of CO_2 emissions, or 0.7 percent of the world's total emissions. These numbers can be projected to grow substantially as the global data center footprint, driven by AI development, grows by as much as a magnitude over the next decade.

The emergence of carbon credit trading systems specifically for AI computing represents a promising development. Several major cloud providers now offer carbon-offset options for AI workloads, allowing organizations to compensate for the environmental impact of their AI operations. These programs typically add 3-5 percent to computing costs but enable carbon-neutral AI deployments.
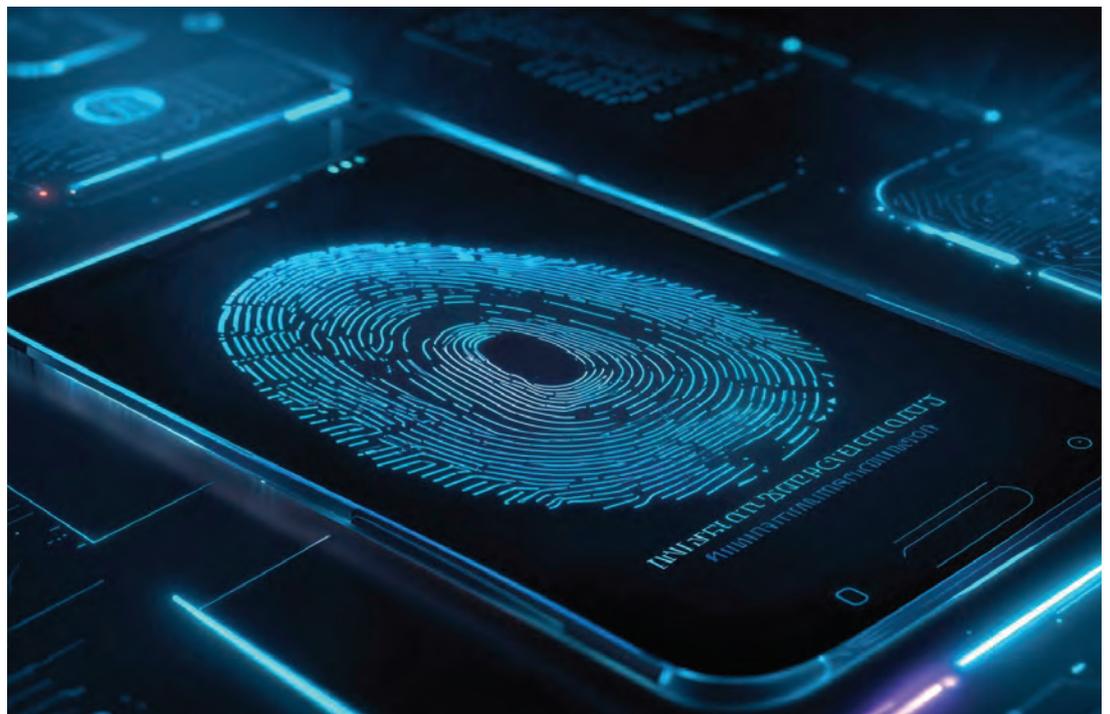
08 AI &
Digital Economies

AI & Digital Economies

Data centers are the backbone of the AI ecosystem, providing the essential infrastructure to meet the vast computational needs of AI applications. These facilities are responsible for storing, processing, and analyzing massive amounts of data, crucial for training and deploying AI models.

AI fundamentally transforms data centers, making them more efficient, secure, and sustainable. As the strategic importance of AI-driven data centers grows, they will continue to play a critical role in supporting the demands of AI and the broader Digital Economy. Integrating AI into data center operations has led to significant improvements in power management, cooling, security, and resilience, ensuring that data centers can meet the growing demands of AI applications while minimizing their environmental impact.

The evolution of AI-optimized data center architecture has accelerated throughout early 2025:

1. Purpose-Built Infrastructure: Rather than retrofitting existing facilities, organizations are increasingly designing data centers specifically for AI workloads, with innovations including: - Direct-to-chip liquid cooling systems achieving 95 percent+ heat removal efficiency - High-voltage direct current (HVDC) power distribution reducing conversion losses - Custom rack designs accommodating extreme power densities exceeding 100kW per rack.

2. AI-Controlled Operations: Machine learning systems now manage critical data center functions: - Predictive maintenance, reducing equipment failures by up to 35 percent - Automated cooling optimization cutting energy usage by 18-25 percent - Dynamic workload placement based on power grid conditions and carbon intensity.

3. Edge-to-Core Architecture: Distributed processing approaches balancing workloads across: - Core data centers for large-scale training - Regional facilities for fine-tuning and specialized processing - Edge deployments for low-latency inference and data preprocessing.
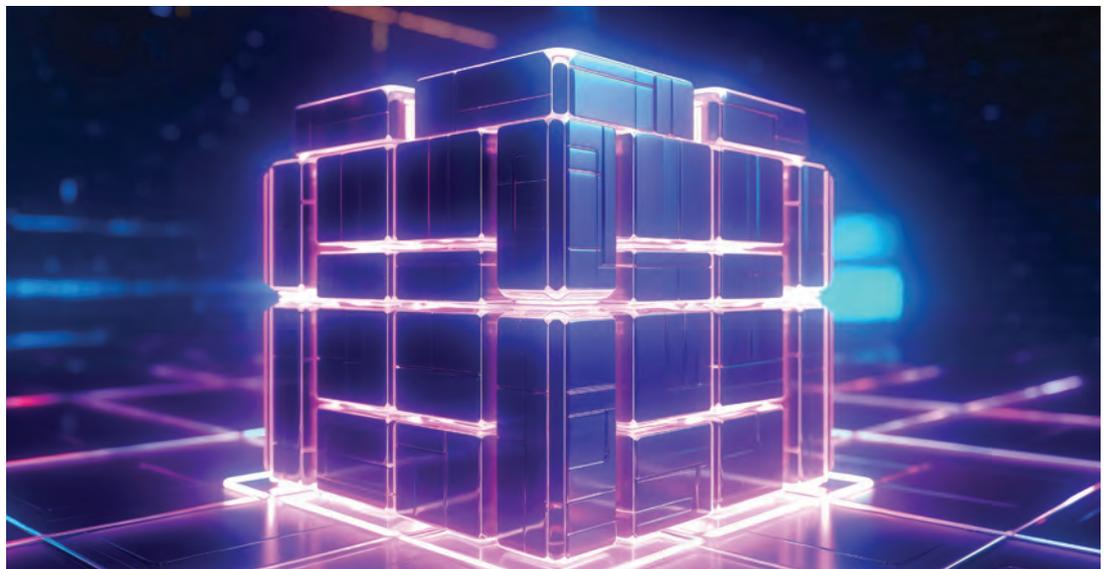
AI's role in data centers will continue to expand, shaping the future of technology infrastructure and driving innovation in the industry. Yet AI is not just a change agent for data centers; it's a transformative force that turns them into highly intelligent, efficient, and sustainable infrastructures.

Integrating AI into data centers opens the door to automated operations, improved data management, and substantial energy savings. As AI applications continue to expand, there is an acute need for robust data center infrastructure capable of supporting the computational requirements of these AI systems.

Building a Digital Economy requires a clear overall vision and specific, detailed plans to build a robust ecosystem that involves governance, legal and ethical policy frameworks, education system, executable roadmaps, local and global alliances, national security, economic drivers and incentives, land and urban development, AI/ML, cloud, connectivity, and data centers.

All factors must be fully synchronized to create maximum synergy and optimum utilization of national resources and opportunities. Obtaining positive results does not require massive overhead and squadrons of inexperienced analysts, but rather the deep knowledge and passion of a highly specialized task force.

Scaling for Digital Economies
Digital Economies can be built—and are being built—bit by bit, from projects to put a particular government service online to bringing faster bandwidth speeds to local internet service to fostering innovation and building and using new apps.

Creating a detailed roadmap involves a series of digital infrastructure projects, educational and training programs, government incentives and initiatives, milestones, five-year goals, and expected outcomes. Progress is measured continuously and all roadmap projects and initiatives are coordinated to maintain steady job creation and socioeconomic progress. IDCA research estimates that the world needs 100 million new IT-related jobs by the year 2030 for the global Digital Economy to be able to handle the challenges of AI-driven development.

The distribution of these jobs varies by current socioeconomic conditions in each country, and is based on reaching a goal of 2.5 percent of a nation's population being employed in an IT position. The jobs entail the entire range of IT employment: IT specialists, network administrators, developers, architects, management, and C-level positions.

In addition to overall goals, roadmaps can include programs to build out the use of IoT initiatives in public transportation and smart grids, contactless payment, advanced manufacturing processes, government eServices, and specialized FinTech apps built around the cultures and needs of people in their particular country.

To achieve true progress in developing a Digital Economy, IDCA's research and activities shows three essential steps to be taken:
1. Create sufficient digital infrastructure to support a Digital Economy, while being highly mindful of sustainability and energy challenges.
2. Educate and train people to develop a workforce equal to the challenges of the 21st century.
3. Leverage AI to create apps and services, while following a unique path through the phases of Digital Economy development.

This process can be part of a positive feedback loop, as ideas germinate and turn into apps and services, justifying the need for more data centers, networking, and devices, with the need for sufficient education and training remaining omnipresent.

Data centers cost between $7 million and $12 million per megawatt to build, and $1 million to $2 million per megawatt per year to operate. These ranges reflect economies of scale in construction and a wide variance in electricity costs throughout the world.

So a new 100MW data center would be in the $1 billion range to build and $100 million annual range to operate. Plans for gigawatt-sized AI factories take these numbers by a factor of 10 or more.

There are very few countries that can support data centers at the high end, just as there are very few companies (and governments) that can develop LLMs. Already, the United States has more than 40 percent of the world's global data center footprint, China has 9 percent, and Germany 7.5 percent. India has 3.2 percent (currently on a par with Japan and the UK), but has plans for rapid growth.

The top 15 national data center footprints constitute more than 80 percent of the world's data center consumption, and the Top 40 garner 98 percent of it. There are more than 140 nations that all together make do with the remaining 2 percent.

Focusing massive data centers and AI factories into the few top nations should continue to drive progress in empowering and improving AI models. But a case can be made for smaller countries to join the AI realm, with more modest goals that would nevertheless be effective. Companies of all sizes should consider the idea of Digital Triangles (choosing three strategic locations for AI centers), and powerful Digital Hubs as key parts of their strategies.

IDCA research envisions three levels of AI center, each with their own scale, cost, and goals:

Small AI Centers with power requirements of 1 to 5MW. Capital expenditure should be $7-10M per MW, with operational expenditure of $0.9-1.2M per MW annually. ROI timelines can be expected to be two to three years for commercial use cases, and three to five years for government applications. It should take from six to nine months from planning to operation.

AMD Epyc 7742 or similar chips would deliver between 10 and 20 petaflops per megawatt. This is not enough for AI modeling, but it can serve the models to users, conduct inference, process large data sets and regional analytics, and run AI research experiments.

A small AI center could also handle agriculture, education, and government services applications. Some percentage of its resources could, of course, also be used for traditional data center services.

Smaller data centers can also continue to be built in developed countries to spur and support local economic growth, improve internet speeds and reliability, and address repository issues involving digital sovereignty.

Mid-Size AI Centers would range from 5MW to 20MW in size. Capital expenditure should be $6-8M per MW, with operational expenditure of $0.8-1.2M per MW annually. ROI timelines can be expected to be three to four years for commercial use cases, and four to six years for research applications. It should take from nine to 15 months from planning to operation.

As with small centers, mid-size centers could deliver 10 to 20 petaflops per megawatt, or a maximum of 400 petaflops. For example, a Ruby installation of 10MW would also support 10 Ruby boards and the exaflops that come with them.

A facility of this size would appear to lie in the shadow of massive projects. But at this scale, they are approaching general AI modeling and training capability. Continued work on SLMs and the efficiency

Large AI centers extending from today's minimum of 20MW to several hundred megawatts and beyond will continue to be the province of government cryptography and defense research, the most demanding scientific modeling and visualization, and the continued growth and sophistication of training large models. Capital expenditure should be $5-7M per MW, with operational expenditure of $0.7-0.9M per MW annually. ROI timelines might run five to eight or more years, as facilities of this scale are often justified through strategic national interests rather than direct financial returns It should take from 18 months to three years or more planning to operation.

AI-as-a-Service can be expected to become a popular application for the most demanding needs. Coreweave is the pioneer in this area, reportedly amassing more than 250,000 Nvidia GPUs to provide "GPU Cloud Computing." The company invested $1.6 billion in a facility in Texas that Nvidia has called "the fastest supercomputer in the world," supported by more than 6,000 miles of fiber-optic cable to meet its processing needs.

FIGURE 12.

Current Worldwide Plans for AI Centers by Size

Small AI Centers
20%

Large AI centers
56%

Mid-Size AI Centers
24%
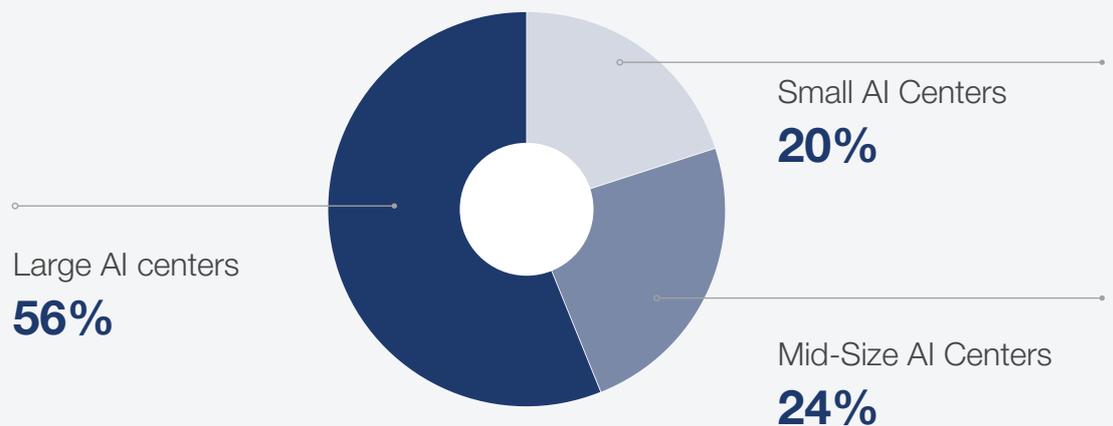
Source: IDCA

09 AI Readiness
of Nations

AI Readiness of Nations

Thus, each nation's government must create a national AI vision and strategy. While investment levels and outcomes will vary nation-by-nation, each nation must create unique AI goals, a roadmap, and an execution strategy.

Determining a nation's potential must start with some foundational questions:
- How does AI's current and future progress affect any individual nation?
- Can a nation of limited or modest resources develop a valid AI strategy?
- What level of data center infrastructure and computing resources are needed for each nation's strategy?
- How can AI best be regulated across national, regional, and continental borders?

For each nation, a commitment to sustained, robust digital infrastructure growth delivers tangible benefits. Within this context, IDCA has developed the Global Digital Readiness Index of Nations, a ranking of the world's nations concerning the state of their Digital Economy development.

Nations have been evaluated against the vital ingredients of Digital Economies and ranked based on their achievements in proportion to their global standing and available resources. The ranking identifies whether a Digital Economy is in a pre-, Early-Stage, Emerging, or Highly Developed phase. (A complete index listing can be found in the IDCA Global Digital Economy Report 2025, available free of charge from IDCA.)

Within its research, IDCA has found that creating Digital Economies and developing AI visions is not limited to highly developed countries.

FIGURE 11.

Countries by Relative Size

Larger Countries
Mid-size Countries
Smaller Countries
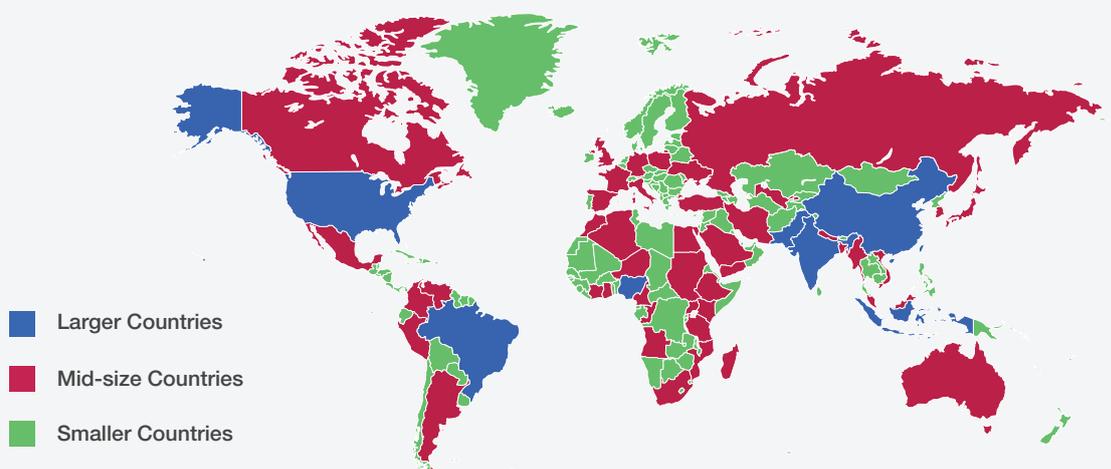
Source: IDCA

The data shows:

- Among developing nations, the countries in the top half of their regions and income tiers in the Index have seen economic growth 5.2 percent to 6.8 percent higher per year for the past 10 years than countries in the lower half of their groups.
- Conversely, among developed nations, countries with digital infrastructure in the lower half of the Index rankings face development challenges that are, on average, 35 percent to 50 percent more difficult and costly than those in the upper half.
- Furthermore, the gap between the highest and lowest 20 percent across all categories is generally more than a magnitude larger than the average.

The Q1 2025 update to IDCA's Digital Readiness Index reveals several emerging regional AI hubs:

- In Southeast Asia, Vietnam and Malaysia have made powerful advances, climbing 7 and 5 positions respectively in the regional rankings. Both countries have implemented strategic national AI policies with dedicated funding and educational initiatives.
- In Africa, Rwanda and Kenya have emerged as continental leaders in AI adoption, with Rwanda's Digital Innovation Zones and Kenya's AI and Blockchain Taskforce driving coordinated development.
- In Latin America, Brazil is emerging as a world force in data center capacity, while Uruguay and Colombia continue their upward trajectory, with Uruguay's Digital Uruguay 2025 program achieving notable success in creating a comprehensive digital ecosystem.

However, concerns about electricity exist in all developing nations and many key areas of developed nations. Developing-nation grids deliver as little as 2 percent of the per-capita electricity found in the developed world. IDCA's research has identified the strength of more than 150 national electricity grids, their sustainability, and the cost to bring them up to a level to support local data centers and to approach the developed-world standard.

In all cases, steady incremental improvement over the medium and long term achieves steady progress. Even though technology moves very quickly, the best way for any nation to keep up is to be consistent and dogged in improving its Digital Economy and society.

The accompanying chart shows examples of countries across income tiers and population sizes that are particularly amenable to the rapid development of national AI infrastructure and their Digital Economies. The list of countries with significant potential is not limited to these countries, as each nation of the world has its own particular relative strengths and entry points into furthering their progress.

Least-Developed Countries (LDCs) With Significant AI Center Potential

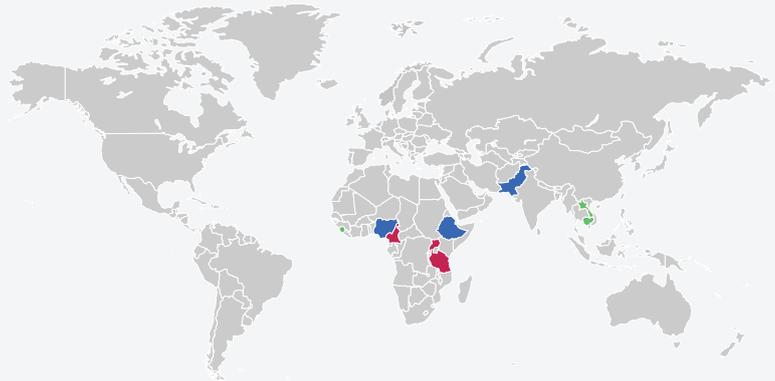Least-Developed Countries (LDCs) (PCI Nominal: <$2,500)

Larger Countries

● Ethiopia
● Pakistan
● Nigeria

Mid-size Countries

● Cameroon
● Tanzania
● Uganda

Smaller Countries

● Cambodia
● Laos
● Sierra Leone

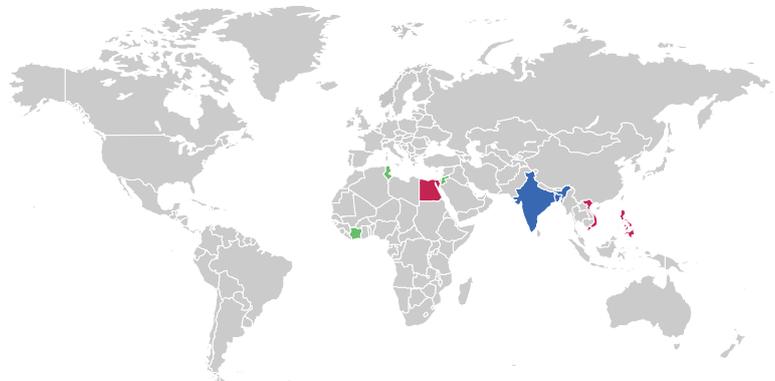Frontier Countries (PCI Nominal: $2,5000-$4,999)

Larger Countries

● Bangladesh
● India

Mid-size Countries

● Egypt
● Philippines
● Vietnam

Smaller Countries

● Côte d'ivoire
● Jordan
● Tunisia

Emerging Countries (PCI Nominal: $5,000-$14,999)

Larger Countries

● Brazil
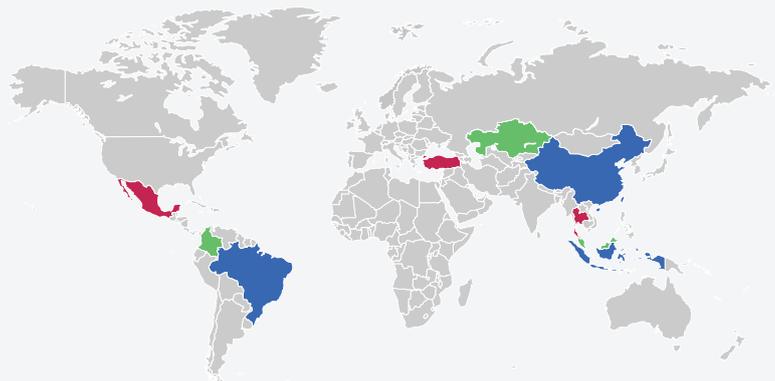● China
● Indonesia

Mid-size Countries

● Mexico
● Thailand
● Turkey

Smaller Countries

● Colombia
● Kazakhstan
● Malaysia

Edge Countries (PCI Nominal: $15,000-$25,000)

Larger Countries

- Chile
- Poland
- Romania

Mid Countries

- Bulgaria
- Hungary

Smaller Countries

- Costa Rica
- Oman
- Uruguay

Developed Countries (PCI Nominal: >$25,000)
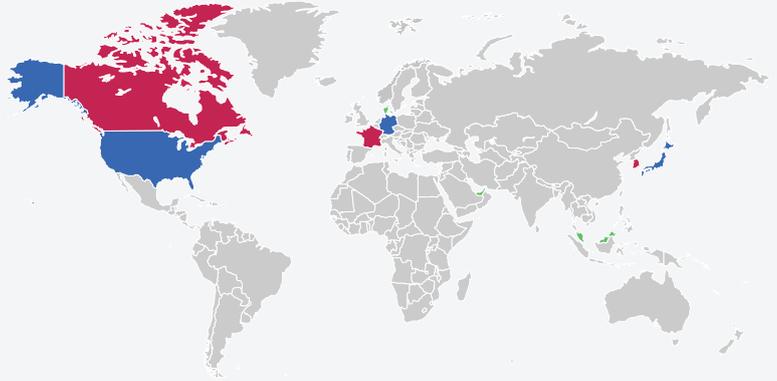
Larger Countries

- Japan
- United States
- Germany

Mid-size Countries

- Canada
- France
- South Korea

Smaller Countries

- Denmark
- Singapore
- UAE

Country Ranking by Size

Developed Countries (LDCs) Larger Countries
Country
India
China
USA
Indonesia
Pakistan
Nigeria
Brazil

Developed Countries (LDCs) Mid-size Countries	
Country	
Bangladesh	Spain
Russia	Argentina
Mexico	Algeria
Ethiopia	Poland
Japan	Canada
Philippines	Morocco
Egypt	Ukraine
Congo, Dem	Angola
Vietnam	Saudi Arabia
Iran	Uzbekistan
Turkey	Yemen
Germany	Malaysia
Thailand	Peru
UK	Ghana
Tanzania	Mozambique
France	Nepal
South Africa	Madagascar
Italy	Cote D'Ivoire
Kenya	Venezuela
Myanmar	Cameroon
Colombia	Niger
South Korea	Australia
Uganda	
Sudan	

Developed Countries (LDCs)
Smaller Countries

Country		
Taiwan		Congo
Mali		Singapore
Sri Lanka		Denmark
Malawi		Slovakia
Zambia		CAR
Romania		Norway
Chile		Finland
Kazakhstan		Lebanon
Ecuador		New Zealand
Guatemala		Costa Rica
Senegal		Ireland
Netherlands		Oman
Cambodia		Panama
Zimbabwe		Kuwait
Rwanda		Croatia
Burundi		Georgia
Tunisia		Eritrea
Bolivia		Uruguay
Belgium		Moldova
Haiti		Mongolia
Jordan		Puerto Rico
Dominican		Bosnia
South Sudan		Albania
Sweden		Armenia
Honduras		Jamaica
Czechia		Gambia
Azerbaijan		Lithuania
Greece		Qatar
Portugal		Botswana
Hungary		Namibia
Belarus		Gabon
UAE		Slovenia
Israel		N. Macedonia
Togo		Latvia
Austria		Eq Guinea
Switzerland		Trinidad
Sierra Leone		Bahrain
Laos		Timor-Leste
Hong Kong		Estonia
Serbia		Mauritius
Nicaragua		Cyprus
Paraguay		Djibouti
Libya		Bhutan
Bulgaria		Guyana
El Salvador		Luxembourg
Turkmenistan		Montenegro

Developed Countries (LDCs) Smaller Countries	
Country	
🇲🇹	Malta
🇲🇻	Maldives
🇮🇸	Iceland
🇧🇿	Belize
🇧🇸	Bahamas
🇸🇨	Seychelles

Methodology

The methodology behind the Index builds on more than a decade of work. It examines numerous technological and socioeconomic factors using data from open sources such as the World Bank, United Nations, IMF, U.S. Department of Energy, Akamai, Trading Economics, ITU, Transparency International, IEA, and the European Commission.

Creating the Index involves a methodology that differs from traditional ranking approaches in several significant ways:

- The process evaluates IT infrastructure, ecosystems, and socioeconomic conditions on a relative rather than absolute basis—measuring how well a country performs given its resources. Data is adjusted for local cost of living and other socioeconomic factors.
- Instead of using fixed percentage weights, all factors are assessed based on their interdependent effects. This approach, inspired by the multiple-body problem in Newtonian physics, uses mathematical curves to map a nation's progress across various factors.
- The results form a flexible solution set that accommodates wide disparities between countries and is best expressed on a natural logarithmic scale.
- These results are then benchmarked against an optimal state—a theoretical model with perfect balance in technology access, speed, and socioeconomic conditions. This idealized reference point always outperforms any actual nation and provides a clear reality check for evaluating national performance.

The optimal state is set at 100, with Economy, Environment, Social, and Governance each contributing 25%. This benchmark exceeds any real nation, highlighting strengths and development areas.

The scale offers a clear, intuitive way to compare progress and challenges across countries.

10　　Conclusion

Conclusion

The current trend of announcing ever-larger data centers, AI factories, and massive development initiatives demonstrates a general optimism among investors, companies, and governments about AI's current state and future. IDCA also believes that AI is truly a global phenomenon, and countries of all sizes and incomes should develop an AI vision and work within it to build their AI footprint and Digital Economy.

Chip development can be expected to continue, bringing more massive capabilities to AI systems; models can continue evolving, and AI-driven services can be expected to proliferate. Concerns about power can be addressed on a local basis. Even though many nations have underdeveloped electricity grids (as noted above), their digital footprints are even more underdeveloped: data centers in most developing nations consume less than 1 percent of available electricity.

Privacy, sovereignty, ethics, and security concerns are key to all AI visions and plans. Despite obvious obstacles, there is a path for each nation to develop its AI strategy and build a Digital Economy on its own terms, with its own scale and goals.

11 APPENDIX
Industry Challenges
& Opportunities

Challenges and Opportunities

There are several key areas in which an AI strategy and AI center construction can deliver benefits.

Financial Services
Challenges
- Data Latency and Real-Time Processing. With the rise of high-frequency trading and instant payment systems, financial institutions must ensure ultra-low latency in AI-driven transactions.
- Regulatory Compliance. AI infrastructure must be built to comply with evolving financial regulations, ensuring transparency and accountability in AI-powered financial services.
- Cybersecurity. Cybersecurity threats are more sophisticated, requiring financial AI systems to be fortified against breaches, particularly in fraud detection and blockchain integration.
- Scalability. The exponential growth in financial data demands an AI infrastructure that can scale rapidly to handle increasing data volumes and complexity.

Opportunities
- High-Performance Computing (HPC). Utilizing HPC in AI infrastructure to support advanced financial modeling, risk assessment, and trading algorithms.
- Cognitive Banking Platforms. Developing AI-powered cognitive platforms that provide personalized banking experiences, enhancing customer satisfaction and loyalty, as well as conversational AI systems
- AI-Powered Fraud Detection. Implementing AI infrastructure to support real-time, AI-driven fraud detection reduces financial crime and enhances security.
- Decentralized Finance Integration. Building infrastructure that bridges traditional financial systems with emerging decentralized finance platforms, enabling secure and compliant innovation.

Examples
- JPMorgan Chase's COIN software reviews commercial loan agreements in seconds compared to 360,000 hours previously required by lawyers
- Goldman Sachs' Automated Research Analyst generates preliminary equity reports, increasing productivity by 34 percent while improving accuracy by 12 percent

Healthcare Challenges

- Data Privacy and Security. With increased regulatory scrutiny in 2024, healthcare providers must ensure that AI infrastructure complies with stringent data protection laws like HIPAA and GDPR while guarding against cyber threats.
- Scalability. The demand for real-time patient data processing is growing, requiring scalable AI infrastructure capable of handling massive amounts of data, especially with the rise of remote healthcare services.
- High Compute Requirements. Advanced AI models for drug discovery and personalized treatment are computationally intensive, necessitating robust infrastructure to manage these workloads.

Opportunities

- Cloud-Based Health Data Platforms. These platforms leverage cloud-based AI infrastructure to facilitate health data storage, processing, and analysis, supporting the shift toward telemedicine and personalized care.
- AI-Enhanced Diagnostic Tools. Supporting real-time diagnostics and imaging through AI infrastructure, improving accuracy and speed in healthcare delivery.
- Personalized Treatment Plans. Using AI-driven infrastructure to deliver individualized treatment plans, optimizing patient outcomes with precision medicine.
- IoT and Wearable Integration. Integrating AI with IoT devices and wearables for continuous patient monitoring, enabled by robust and scalable infrastructure.
- Genomic Analysis Systems. Deploying specialized AI infrastructure for large-scale genomic data processing, enabling breakthroughs in personalized medicine and disease treatment.

Examples

* Mayo Clinic's Clinical Documentation Agent reduces physician documentation time by 76 percent, returning 1.8 hours daily to patient care
* Johns Hopkins Medicine optimizes hospital operations through real-time resource allocation and planning
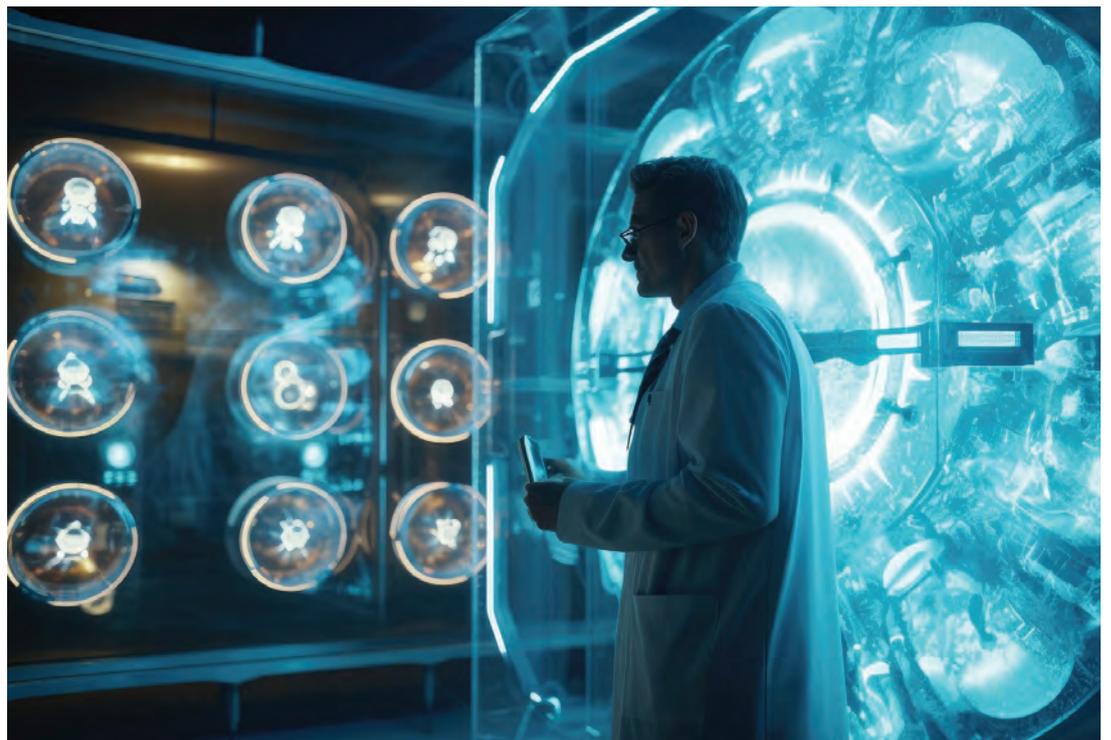
Manufacturing Challenges

- Integration with IoT Devices. As smart manufacturing becomes more prevalent, integrating AI infrastructure with a growing array of IoT devices is critical but complex, requiring robust connectivity and data management solutions.
- Real-Time Processing. The need for real-time decision-making on the production floor requires AI infrastructure capable of processing data instantaneously
- Predictive Maintenance. Ensuring that AI infrastructure can support predictive maintenance by processing vast amounts of sensor data to predict equipment failures and optimize uptime.
- Energy Consumption. Managing the energy demands of AI-driven manufacturing processes is a growing concern, especially with sustainability/ESG targets becoming more stringent in 2024.

Opportunities

- Smart Factories. Developing cognitive infrastructure that enables AI-driven automation, improving efficiency, reducing waste, and enhancing production quality.
- Digital Twins. Supporting the creation and use of digital twins for real-time simulation and optimization of manufacturing processes.
- Supply Chain Optimization. Using AI infrastructure to optimize global supply chains, ensuring resilience and efficiency amid disruptions.
- Autonomous Manufacturing Systems. Building infrastructure that supports fully autonomous production lines capable of self-optimization and adaptation to changing requirements.

Energy Challenges

- Grid Management and Stability. With the integration of renewable energy sources, AI infrastructure must manage complex and dynamic energy grids, ensuring stability and preventing outages.
- Integration with Renewable Energy Sources. AI infrastructure must be adaptable to the fluctuating inputs from renewable energy sources like wind and solar, which require sophisticated predictive analytics.
- Predictive Maintenance for Energy Infrastructure. Building AI infrastructure to support predictive maintenance across vast energy networks, reducing downtime and operational costs.
- Cybersecurity. As energy systems become more interconnected, protecting critical infrastructure from increasingly sophisticated cyber threats will be paramount in 2025.

Opportunities

- Smart Grid Implementation. Developing AI-driven infrastructure that supports the creation of intelligent grids, improving energy distribution efficiency and reliability.
- Energy Consumption Optimization. Using AI to optimize energy consumption across industrial and residential sectors, reducing costs and improving sustainability.
- Renewable Energy Management. Leveraging AI infrastructure to optimize renewable energy production, storage, and distribution, making it more viable and widespread.
- AI-Enhanced Disaster Response. Implementing AI systems that enhance energy infrastructure disaster response and recovery ensures resilience against natural and artificial disasters.
- Fusion Energy Development. Supporting advanced AI-driven simulations for fusion energy research, accelerating the path to commercial fusion power, and addressing long-term energy needs.

Public Sector and Digital Government
Challenges

- Citizen Data Privacy. As governments adopt AI for public services, protecting citizen data remains a top priority, especially with the increasing use of AI in areas like surveillance and social services.
- Ethical Use of AI. Governments must ensure that AI is used ethically, particularly in law enforcement, public policy, and national security. This requires robust oversight and transparent AI infrastructure.
- AI in National Security. Developing secure AI infrastructure for national defense and intelligence applications while safeguarding against AI-driven threats from adversaries.
- Scalability of Public Services. As AI is integrated into public services, the underlying infrastructure must be scalable to meet the needs of large and diverse populations.

Opportunities

- Smart City Infrastructure. Building cognitive infrastructure supporting AI-driven intelligent city initiatives enhances urban management, traffic flow, and public safety.
- AI-Powered Public Services. Developing infrastructure that enables AI-enhanced public services, such as healthcare, education, and transportation, improves efficiency and accessibility for citizens.
- Disaster Management Systems. Creating AI-powered infrastructure that enhances disaster prediction, preparedness, and response, improving resilience to natural and human-made disasters.
- Digital Citizen Engagement. Implementing AI-driven platforms that enable more responsive and personalized citizen interactions with government services, reducing bureaucratic friction.

General Human Resources
There are also examples today of agentic AI being applied across human resources and workforce development.

- Unilever has reduced hiring timelines from four months to four weeks
- Walmart uses agentic AI to streamline the evaluation of 3 million job applicants annually
- "Employee digital twins" enable unprecedented personalization of talent management
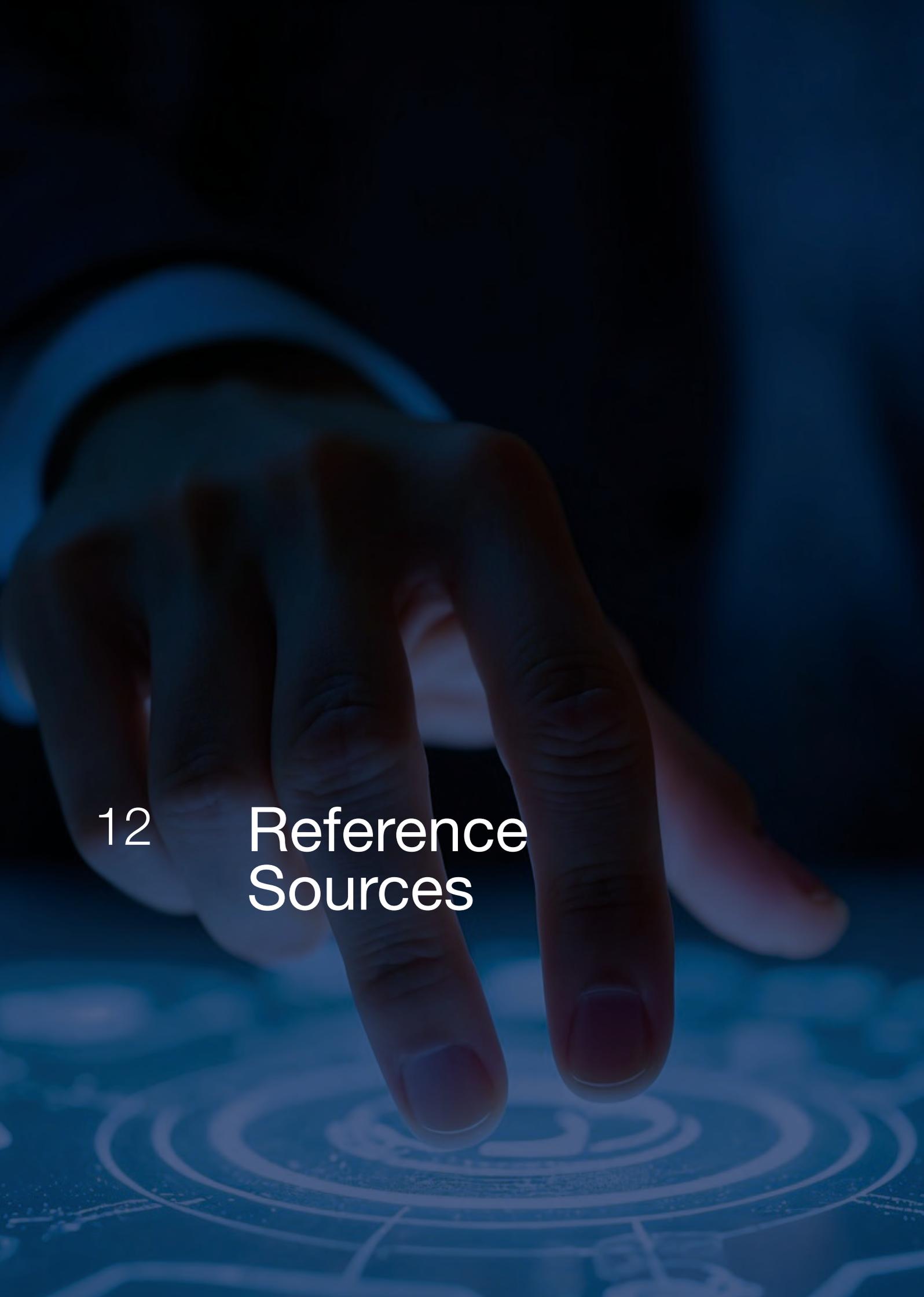
12   Reference
        Sources

Reference Sources

IDCA Research Publications and Data

1. "Global Digital Economy Report 2025." International Data Center Authority (IDCA), March 2025.
2. Referenced in the main body when discussing the Digital Economy, encompassing $16 trillion of global GDP
3. Source for various AI adoption statistics throughout the report
4. "Digital Economy Readiness Index - Q1 2025 Update." International Data Center Authority (IDCA), April 2025.
5. Referenced in the AI Readiness by Nation section
6. Source for comparative economic growth statistics between countries in different development tiers
7. Cited for emerging regional AI hub information
8. "Enterprise AI Adoption Survey 2025." International Data Center Authority (IDCA), January 2025.
9. Source for statistics: 87 percent of companies identify AI as top priority, 76 percent now use AI, 69 percent use generative AI, 53 percent use AI for Big Data
10. "Global Data Center Energy Consumption Survey Q1 2025." International Data Center Authority, April 2025.
11. Cited for statistics: worldwide data center footprint consumed 2.1 percent of world's electricity in Q1 2025, accounting for 268 million metric tons of CO_2 emissions (0.7 percent of the world's total)
12. "Agentic AI Implementation Survey 2025." International Data Center Authority (IDCA), February 2025.
13. Source for statistic: 42 percent of enterprises have implemented agentic AI solutions for at least one business function
14. IDCA Industry Survey: "Game-changing Technologies for Digital Economies." International Data Center Authority (IDCA), March 2025.

15. Source for statistic: 72 percent of respondents named AI as the leading "game changer" in building Digital Economies

AI Hardware and Infrastructure Technical Specifications

7. "NVIDIA Blackwell Platform Technical Documentation." NVIDIA Corporation, January 2025.
8. Referenced for specifications: chips costing up to $70,000, delivering 5.5 petaflops, with multi-GPU boards in $3 million range requiring 120 kW of power
9. Source for energy efficiency improvement: 30 percent greater performance per watt than predecessor
10. "NVIDIA H100 Technical Specifications." NVIDIA Corporation, 2024.
11. Referenced for specifications: priced at $25,000-$40,000, delivering up to a petaflop for certain uses and 67 teraflops for complex processing
12. "AMD Epyc 7742 Product Specifications." AMD Corporation, 2024.
13. Referenced for specifications: 64 cores delivering 1-2 gigaflops per core (up to 128 gigaflops per chip)
14. Pricing information: $1,000-$2,000 per unit
15. "AMD MI300X Deployment Analysis." AMD Corporation, February 2025.
16. Source for statistic: over 120,000 units deployed globally as of Q1 2025
17. "Intel Xeon Gold 6230 Product Specifications." Intel Corporation, 2024.
18. Referenced for specifications: 20 cores delivering 1-2 gigaflops per core
19. Pricing information: $1,000-$2,000 per unit
20. "Intel Gaudi 3 AI Accelerator White Paper." Intel Corporation, December 2024.
21. Referenced for competitive positioning against NVIDIA products
22. Pricing information: $25,000-$30,000 per unit

AI Models and Platforms Technical Documentation

13. "Claude 3.7 Sonnet: Technical Specifications and Capabilities." Anthropic, February 2025.
14. Referenced for reasoning capabilities, multimodal features, and enterprise security frameworks
15. "GPT-4.5 Technical Report." OpenAI, January 2025.
16. Referenced for video and audio capabilities, code interpreter, and autonomous agent features
17. "DeepSeek-V3 Technical Documentation." DeepSeek AI, March 2025.
18. Referenced for multimodal capabilities and sparse architecture features
19. "Mistral AI Models Overview." Mistral AI, 2025.
20. Referenced for details on Mistral 7B, 8x7B, Large, and Large 2 models
21. Performance comparison to GPT-4.5 and Claude 3.7
22. " LLaMA Model Family Technical Documentation." Meta AI Research, April 2025.
23. Referenced for historical progression of LLaMA models 1-4 with specific parameter counts and token training details
24. "Falcon AI Platform Documentation." Technology Innovation Institute, March 2025.
25. Referenced for Falcon 1B, 7B, 40B, 180B, and 200B+ model specifications

Major AI Infrastructure Projects

19. "Stargate Project Overview." [Corporate/ Government Source], 2025.
20. Referenced for $500 billion commitment in the US
21. "Rajasthan Data Center Complex Development Plan." [Indian Government/ Corporate Source], 2025.
22. Referenced for 3GW complex in Rajasthan, India
23. "Oracle Cloud Arizona AI Hub: Technical Infrastructure Overview." Oracle Cloud, February 2025.
24. Referenced for $40 billion facility with 1.5GW dedicated power capacity

25. "Microsoft Gobi Desert Quantum Facility: Hybrid Computing Architecture." Microsoft Research, March 2025.
26. Referenced for combining traditional AI compute with early quantum processing capabilities
27. "Google Carbon-Neutral AI Center: Design and Implementation." Google Cloud, January 2025.
28. Referenced for sustainability metrics including PUE of 1.07
29. "CoreWeave GPU Cloud Infrastructure." CoreWeave, 2025.
30. Referenced for deployment of 250,000 Nvidia GPUs and $1.6 billion Texas facility investment

Regulatory Frameworks

25. "EU Artificial Intelligence Act Implementation Guidelines." European Commission, February 2025.
26. Referenced for risk classification system and implementation timeline with compliance schedules
27. "Executive Order on Safe and Responsible AI Development." The White House, March 2025.
28. Referenced as establishing oversight responsibilities across multiple federal agencies

Additional Technical Information

27. "Global Internet Performance Report Q1 2025." [Source, likely Akamai Technologies], April 2025.
28. Referenced for statistic: world average internet speed of 21Mbps
29. "MLPerf Inference Benchmarks 2025." MLCommons, March 2025.
30. Referenced for comparison of computational demands for multimodal vs. single-modality models
31. "Quantum-Resistant Cryptography for AI Infrastructure." NIST, March 2025.
32. Referenced for the finalization of first quantum-resistant cryptographic algorithms

IDCA

1801 Research Blv., Suite 550, Rockville, MD 20850

+1 (866) 422 1971

www.idc-a.org

research@idc-a.org

linkedin.com/company/international-data-center-authority-idca

IDCAorg