INTELLIGENCE UPDATE

Agentic AI shows promise — but also carries risk


John O'Brien

19 Feb 2025

The release of DeepSeek's V3 and R1 large language models (LLMs) in January exemplified how disruptive this new AI era of LLM technologies is becoming (see *The DeepSeek paradox: more efficiency, more infrastructure?*).

Agentic AI, a form of AI that can operate autonomously and relies on LLM technology, has been gaining significant momentum, over the past 12 months in enterprise software/software as a service (SaaS). Companies such as Microsoft, ServiceNow and Salesforce — familiar names to data center operators — now offer customers both the tools to build their own AI agents and other ready-to-use agents delivered as a service. Data center vendor Siemens launched an industrial copilot in January 2025, in partnership with Microsoft, which now provides engineers with real-time querying, metrics and reports such as cycle time performance and machine error alerts.

Agentic AI offers the potential for data center operators to radically improve their facility efficiency, performance and availability. It could also reduce risks resulting from human error and other unknown factors, which may be only identifiable through advanced AI learning.

Although many operators may be hesitant to trust an LLM with data center management, there are compelling reasons to evaluate its potential. In this report, Uptime will discuss the opportunities, as well as the potential risks.

What is agentic AI?

Uptime defines agentic AI for the data center as "the autonomous control of an IT and/or OT environment by a generative AI system, which is under the control of an AI software agent(s)."

The agent would autonomously control facility assets and environments by interacting with a LLM and domain-specific data for its breadth and depth of knowledge. Real-time feedback from the environment would reinforce best practices and help adapt the agent's responses as needed. The agent can also respond to user queries and perform user-directed tasks via natural language copilots (as described above).

Uptime's research shows how AI-enabled control software can already dramatically improve the efficiency and performance of data center environments such as cooling and power (see *Using optimization software for cooling and capacity gains* and *Data center management and control software: an overview*). Using advances in LLMs and AI agents could augment control systems more quickly than through traditional machine learning, while broadening and deepening the range of activities. There are three core elements to Uptime's definition:

- **Agentic AI refers to the framework, architecture and technologies used by the AI agent(s)**. These may be proprietary to the enterprise (using a custom-built LLM) or owned and operated by a third-party software/SaaS provider.
  Both approaches have strengths and weaknesses. In-house agentic AI will require significant capital investments in infrastructure, computer processing and inferencing, but may be more secure in terms of data protection. SaaS is less complex, but lacking ownership of the intellectual property and processing data off-premises may make it less appealing to some operators.
- **An agentic AI system is adaptive and self-correcting**. It can learn from its environment by engaging with live system data and via feedback from human-user interactions. The AI agent learns and responds to these changes and adjusts its action autonomously.
- **Human operators can monitor and interact with the system via an AI copilot**. They can ask questions and receive information and reports (e.g., relating to system performance, health, and incident trends and resolutions). There should also be fail-safes and other escalation capabilities in the event of system error or failure.

Most definitions of agentic AI refer to the agent's autonomous abilities to learn and adapt, but the ability of an AI agent to perform control is often overlooked. In the data center, for example, agentic AI control could improve power or cooling system availability by continuously optimizing the environment. Human error, which is often a factor in outages, could be reduced if, for example, AI control replaced human tasks where skilled staff are unavailable. It could also provide back-up for operators during times of work stress.

How does agentic AI work?

AI agents access information from pre-trained LLMs, which can be custom-built or based on general-purpose models, such as ChatGPT or Llama. This information is typically stored in a vector database that specializes in storing and retrieving "unstructured data" used by LLMs, such as text, images, video and sounds.

Retrieval augmented generation (RAG) techniques allow agents to pull external "structured data" into the LLM to improve the context and specificity of responses (see *How generative AI learns and creates using GPUs*).

LLMs are trained on hundreds of billions of public and private data points, providing them with a broad knowledgebase. However, for data center applications, they need to be domain specific. This is achieved by integrating additional data sources, such as data from building management systems, supervisory control and data acquisition (SCADA), data center infrastructure management (DCIM) software, and real-time environmental and sensor data.

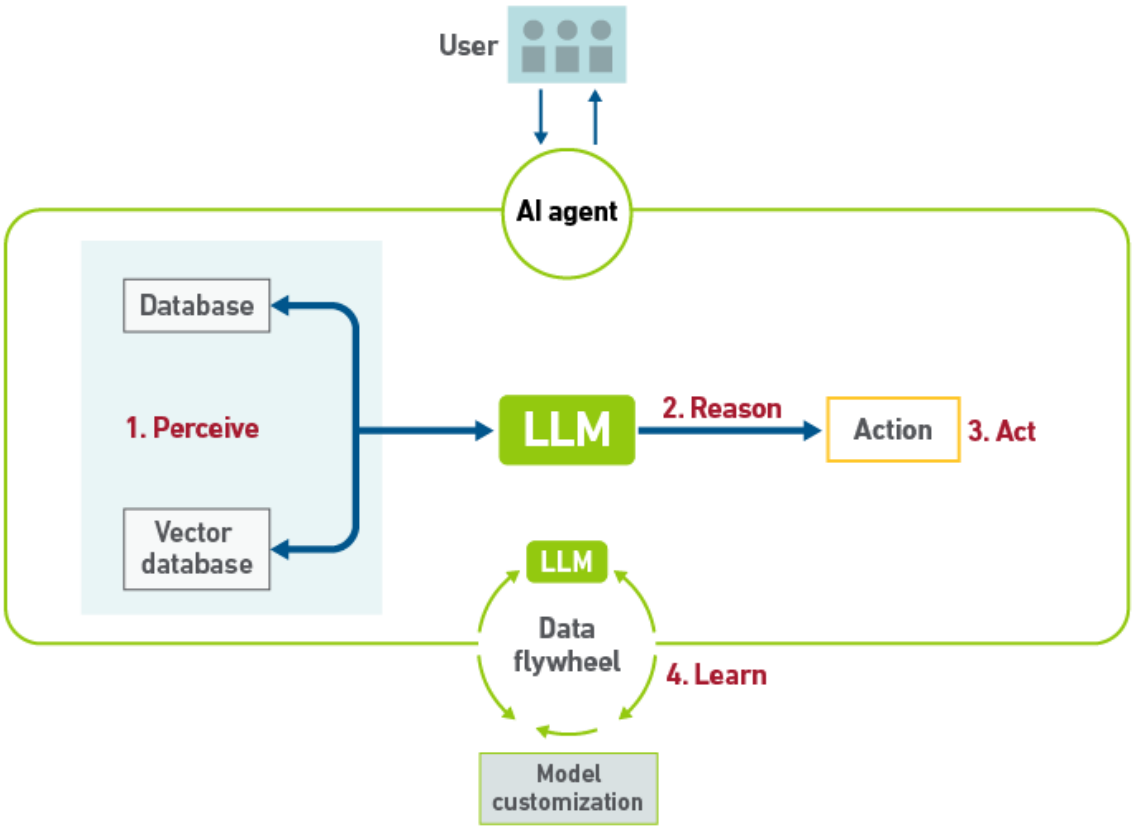
An AI agent could theoretically integrate data from all data center domains where data is available, including server and IT, power and cooling, and mechanical and electrical equipment. An LLM would also interpret questions and statements from human operators (such as data center managers). It would access these data sources (through RAG) to resolve the user's request. If the user requests an action, the LLM will understand the human statement and use API calls to execute external technologies (such as servers). An agent would take the learning and logic from this reasoning process to perform a control action.

There are four steps to the problem-solving process:

- **Perceive.** AI agents ingest data from various sources, such as sensors, databases (traditional and vector databases) and user applications.
- **Reason.** This involves using the LLM to understand tasks and generate and coordinate responses. RAG is used to access domain-specific, often proprietary data and deliver more accurate, task-specific outputs.
- **Act.** This involves executing tasks based on the LLM's reasoning and logic. Guardrails and fail-safes should prevent AI errors so that more complex steps or risks are escalated to humans.
- **Learn.** This involves continuous improvement of the process via a data flywheel, where data generated from the agent's interactions is fed back into the system to improve the LLM on an ongoing basis.

Figure 1 shows Nvidia's architectural approach to agentic AI problem solving. The AI agent responds to (and orchestrates) the actions performed by the agentic AI system, including user interactions and actions. In line with Uptime’s definition, this will also include control responses. Since agentic AI is a nascent technology, operators should be aware that frameworks are continually being refined.

Figure 1 Nvidia’s approach to agentic AI



UPTIME INSTITUTE 2025 (Adapted from Pounds, *What is agentic AI?* Nvidia blog, Oct 2022)

Opportunities and risks

There is significant potential for agentic AI in complex data center systems. The LLM and agent could integrate more data and perform more reasoning, fine-tuning and inferencing. Over time this could build a vast, continually expanding LLM for data center management to support more applications for autonomous monitoring, management and control.

A key area for investigating early opportunities could be in data center management software tools, such as DCIM, cooling and power optimization software, and hybrid IT management software.

However, at this point, many of these opportunities and benefits remain theoretical and several challenges remain, such as:

- **Agentic AI requires high-quality data.** The LLM needs this high-quality data to perform refining, reasoning and inferencing. Even with well-trained data, generative AI models can still hallucinate or provide inaccurate outputs.
- **Agentic AI systems need to be reliable**. Uptime's research shows that trust in AI for operational decision-making has been declining for the past three years since the introduction of generative AI (see *Uptime Annual Global Data Center Survey 2024*). These systems will need to be tested and proven before more operators seriously consider adoption.
- **LLM and AI software security is a growing concern for many organizations.** Microsoft's Data Security Index Report 2024 reveals that 40% of organizations experienced data security incidents from AI applications in 2024. Meanwhile, two-

thirds (65%) admit that their employees use unauthorized AI tools.

- **Technical complexity.** Integrating an LLM into a structured API can be highly complex. Most customized applications will require change management or rearchitecting for compatibility.
- **Uncertain costs and business value.** The cost of developing, training and inferencing via agents is difficult to predict. As the number of agents and prompts increases, these services' costs could accumulate significantly.
- **Sustainability concerns.** AI models can increase energy consumption and carbon emissions due to the computational power needed. Minimizing prompts to achieve accurate results while reducing environmental impact will be an ongoing challenge.
- **Data confidentiality and proprietary IP risks.** Organizations in competitive or highly regulated sectors (e.g., finance and health care) may face concerns over data confidentiality and intellectual property leakage, mainly when using open source or cloud-based LLMs.
- **Ensuring guardrails are in place.** Strict governance and fail-safes need to be in place so that mistakes caused by AI do not threaten the facility's availability or uptime. This is likely to require extensive testing in a secure non-live environment.

Agentic AI is not a simple “plug and play” technology. It requires ongoing investment in learning, refining and reasoning in the software, as well as in the supporting infrastructure. The outcome itself is also likely to evolve — requiring a continuous improvement loop and “data flywheel” that improves the data, learning, refining and reasoning.

The Uptime Intelligence View

Given the rapid advances in AI, it is likely that agentic AI will be adopted by many industries over the next decade, including data centers. Ongoing advances in LLMs and domain-specific models may eventually see agentic AI applied to some of the most complex operator problems — today, however, many risks and challenges remain. Building a solid foundation will be critical for most operators: this is likely to mean radically improving the quality and availability of data, and more advanced security and governance practices for AI.

Other related reports published by Uptime Institute include:
The impact of AI on data center operations (Part I)
How generative AI learns and creates using GPUs
Using optimization software for cooling and capacity gains

Uptime Institute experts consulted for this report:
Owen Rogers, Senior Research Director of Cloud Computing, Uptime Institute
Max Smolaks, Research Analyst, Uptime Institute

ABOUT THE AUTHOR

John O'Brien

John is Uptime Institute’s Senior Research Analyst for Cloud and Software Automation. As a technology industry analyst for over two decades, John has been analyzing the impact of cloud migration, modernization and optimization for the past decade. John covers hybrid and multi-cloud infrastructure, sustainability, and emerging AIOps, DataOps and FinOps practices.

jobrien@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. Its Tier Standard is the IT industry's most trusted and adopted global standard for the proper design, construction, and operation of data centers – the backbone of the digital economy. For over 25 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions.

With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.