uptime INTELLIGENCE

The impact of AI on data center operations (Part I)

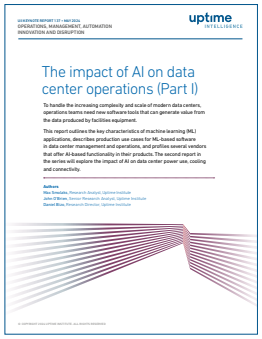
To handle the increasing complexity and scale of modern data centers, operations teams need new software tools that can generate value from the data produced by facilities equipment.

This report outlines the key characteristics of machine learning (ML) applications, describes production use cases for ML-based software in data center management and operations, and profiles several vendors that offer AI-based functionality in their products. The second report in the series will explore the impact of AI on data center power use, cooling and connectivity.

Authors
Max Smolaks, Research Analyst, Uptime Institute
John O'Brien, Senior Research Analyst, Uptime Institute
Daniel Bizo, Research Director, Uptime Institute

Key points

- A new generation of startups is applying ML-based systems in data center management, focusing on optimization of data center cooling equipment settings and IT power consumption.
- Data center operators need to set realistic expectations about the results of any ML deployments, and this requires understanding the core technical principles that define the capabilities and limitations of ML-based systems.
- The propensity to confidently give false information likely disqualifies generative AI from operational decision-making. However, this type of AI, with human supervision, could enhance other aspects of data center management.
- The data center management maturity model, proposed by Uptime Institute in 2019, remains valid; however, it has been expanded beyond data center infrastructure management (DCIM) to describe a wider variety of data center management software.



Contents

Figures

Uptime Institute Intelligence is an independent unit of Uptime Institute dedicated to identifying, analyzing and explaining the trends, technologies, operational practices and changing business models of the mission-critical infrastructure industry. For more about Uptime Institute Intelligence, visit uptimeinstitute.com/ui-intelligence or contact research@uptimeinstitute.com.

Introduction

AI research is going through a new peak in interest and investment. Intense activity in generative AI and neural network architecture development is pushing more organizations to experiment with novel approaches to automation. As part of this trend, machine learning (ML) models are being embedded into mainstream business and desktop applications, driving an unexpected demand for additional data center capacity.

AI, in many forms, is finding its way into the software tools that play an important supporting role in data center design, development and operations. Managers should track the technology and understand where and how it can be safely used in their facilities.

This report — the first of a two-part series — explores the impact of innovation in AI on data center owners and operators. It outlines the terminology required to discuss AI implementations and investigates the role of ML as the foundation for a new generation of data center management tools. It also evaluates the suitability of generative AI applications for mission-critical settings.

ML presents a suitable approach for dealing with the growing size and complexity of modern facilities, and the adoption of ML-based tools for data center management is likely to accelerate as organizations get accustomed to using such tools in other business areas.

The second report in this series will explore AI as a workload and analyze the likely impact of this technology on digital infrastructure footprints in the next few years.

What is the role of AI in digital infrastructure management?

While technological progress continues, a broad consensus on AI — what it is, what it is not, and what it can and cannot do — is still out of reach. This lack of agreement has only become more widespread since generative AI burst onto the scene, creating diverging, even polarizing opinions.

This introduction explores the fundamentals of AI as understood by Uptime Intelligence. It examines the relationship between AI and ML; the differences between ML and traditional software; and outlines some of the terms and definitions used to discuss the impact of AI on data center infrastructure.

An illusion of intelligence

The term “artificial intelligence” was coined in 1956 as the name for a new research discipline that attempted to simulate human intelligence in machines. Today, AI is used as an umbrella term that describes any computer software that is capable of perception and learning. These properties make AI systems suitable for tasks that cannot be expressed effectively (whether mathematically or logically) by using traditional programming, which relies on humans to develop a set of closed-loop logic.

Uptime Intelligence sees AI as a technology-led service that can simulate human tasks requiring cognition or intelligence, such as learning, predicting and problem-solving.

Such tasks include a vast array of compute problems, from optimizing airflow in a data hall and finding abnormal patterns in data to recognizing spoken words and identifying objects or people in images.

AI-based applications can appear to respond to their environment — they can adjust heating and lighting in the home; recalibrate control settings in the data center; and recommend the next steps to a call center agent when confronted with an angry customer. But this is the result of rules, logic and pattern spotting rather than cognition. Historically, many of these functions have been programmed into appliances and applications using traditional software development techniques.

ML is a field of study within AI that is concerned with the development of statistical algorithms that are "trained" on large datasets. This process enables ML models to analyze new data and produce the desired outcomes without being explicitly programmed to do so.

Today, ML is widely deployed and relatively well understood. The technology powers most product recommendation engines, email spam filters, automated translation and transcription tools, and image recognition applications. ML has been used in data center management software for at least 15 years but is currently experiencing a resurgence in popularity as more organizations become confident enough to experiment with AI.

Not all AI activities involve ML — some legacy approaches to AI rely on complex, hand-coded logic to express real-world objects, events or relationships. Examples of what is sometimes called "good old-fashioned artificial intelligence" include early chatbots, "expert systems" used in daily business activities in the 1980s and techniques used to emulate human players in video games. These approaches to AI are markedly different from ML and do not appear to have applications in data centers.

Making sense of algorithms and models

All ML applications are composed of algorithms, which are mathematical procedures implemented in software code. The algorithms analyze many examples of matching inputs and outputs to generate the pattern-spotting logic that can then be used to analyze new inputs and produce the most suitable outputs.

These algorithms are applied to training datasets to create an ML model. The model is the program that is saved after running multiple algorithms on the training data. It represents the rules, numbers and any other algorithm-specific data structures required to make classifications or predictions — the two key problem areas that dominate ML.

The life cycle of an ML model can be split into two distinct parts. The first is the training phase. This is when an ML model is created or trained by running algorithms on a dataset and configuring them to find the desired patterns. Training is a compute-intensive operation, often requiring specialized IT hardware and human assistance through either labeling the data or supervising (and correcting) the process.

The second phase is inference, when the model is deployed into production. Inference is the process of running new datapoints through a trained ML model to calculate an output such as a single numerical score, a string of text or an image.

For example, an ML model can be trained on a dataset containing thousands of examples of cooling equipment sensor readings from a single facility. The model analyses trends and extracts patterns to understand how each cooling unit affects the temperature across the data center. During inference, it can use what it has learned to suggest (and sometimes implement) equipment settings that improve overall cooling performance.

Inference requires considerably less computational resource than training — many ML models can run on any digital device capable of general-purpose processing. The size of the production model is typically much smaller than the size of the training dataset. Many of the models used in data centers today are small, easy to run and relatively simple.

Who needs machine learning?

While conventional software is programmed to perform a task, ML systems are programmed to learn to perform a task.

Traditional, logic-based software systems are deterministic and designed to operate in stable environments. A software system consists of explicit instructions that tell a computer what to do — it accepts specific inputs to produce predetermined outputs.

In contrast, ML programming is non-deterministic and designed for dynamic environments where inputs and outputs are unpredictable. This unpredictability can be caused by human interaction, a facility environment with unknown equipment at the time of model training, or the result of forces of nature, fluid dynamics and energy flows. Results are therefore based on statistical probabilities rather than certainties. When presented with new data, these systems can make the equivalent of an educated guess — something traditional software cannot do.

These properties suggest that ML systems can be integrated into software tools for data center management. Data centers are dynamic environments that consist of many different systems forming complex interdependencies, often so complex that they elude human understanding. Data centers are also equipped with sensors that can serve as sources of training data and tools that collect, categorize and store this data, such as building management systems and data center infrastructure management software.

The goal of the current generation of ML-based systems for the data center is not to take over the running of the facility but to enhance human operators through gradual improvements to existing functions and processes.

The three primary near-term applications of ML in data centers are:

- **Equipment setting optimization.** Enabling power and cooling systems to adapt to changing IT workload requirements in near real-time.
- **Predictive analytics.** Enabling predictive maintenance of data center equipment to minimize the risk of downtime and assisting in more accurate capacity planning.
- **Anomaly detection.** Supporting operations teams to quickly identify the root causes of failures or underperforming equipment, giving them a better chance at preventing an outage. AI-based upgrades to facility monitoring and physical security systems will also be able to identify “out-of-place” personnel or non-compliant activities.

Following conversations with Uptime Institute members, it is clear that for the next few years, most of the ML systems deployed in data centers will be used to augment employees in their tasks via automated insights or recommendations; this is sometimes called the human-in-the-loop approach.

Once data center owners and operators are confident of the results, they will likely progress to deploying fully autonomous or closed-loop AI systems that replace human involvement in a particular task altogether.

However, one key limitation of ML might slow down the pace of adoption: since ML systems are reliant on mathematical algorithms, they are typically more difficult to understand and interpret than traditional software code. This is often described as the "black box problem." Trust in decisions made by AI systems is vital to the widespread adoption of this technology. If ML is to become mainstream, software vendors will have to get better at tackling this problem and explaining how AI systems reach their recommendations and decisions.

AI startups innovate in cooling and IT operations

The optimization of data center cooling and IT power consumption are the two main areas that a new generation of AI software startups is addressing, Uptime Intelligence has found. Cooling and IT equipment are attractive candidates for AI-based optimization because of the complexities involved in making them run efficiently. Together, they account for most of the energy use in a data center — so even a relatively small gain from optimization can deliver significant savings over the life cycle of the infrastructure.

The challengers

This chapter explores the capabilities of four startups: Phaidra, TycheTools, Coolgradient and QiO Technologies (see **Table 1**). These companies, all established in the past five years, are focused on improving cooling and/or IT energy performance using ML and, in some cases, deep neural networks (DNNs) to support their recommendations, predictions and optimizations. None of these startups use generative AI at this point, although one has plans for this technology on its roadmap.

This is not an exhaustive list, and there are other startups, private and public companies that are applying AI in data center management.

Table 1

Four AI startups that have emerged in the past five years

Vendor (headquarters, year launched)	Focus / proposition	Business case / business value claims / ROI claims	Installations / customers	Partners / funding
Phaidra (Seattle, US, 2019)	• Reinforcement learning AI makes incremental changes to equipment configuration. • Software as a service only. • Cooling systems optimization: CRAH, fan coil units, chiller plants and condenser systems. • Real-time insights and dashboards.	• Closed-loop autonomous control. Claims: • Improved thermal stability in chiller plants. • Reduced energy costs in chiller plants (15% to 30%). • Lower equipment runtimes (up to 50%).	Unnamed pharma company with a large data center footprint.	Developed by former Google DeepMind engineers. Chief executive previously led Google Cloud PUE team. Raised $25m series A (2022) led by Starshot Capital.
Coolgradient (Amsterdam, Netherlands, 2018)	• Neural network machine learning. • On-premises only. • Optimization of CRAH, CRAC, direct expansion units, UPS systems and more. • Real-time insights and dashboards. • Recommendations on energy efficiency, downtime root cause analysis. • Advisory and monitoring services.	• Open-loop control only. • Humans implement AI recommendations (e.g., adjustments to overworked chillers, valve settings, fan speeds or floor pressure). • ROI claimed in under a year. • 3,360 MWh annual savings in a 20 MW data center.	Deployed in 18 Digital Realty data centers to date. New deals being signed in Asia-Pacific.	Platform can integrate with other tools and workflows. Academic partnerships with Delft University of Technology and University of Twente.
QiO Technologies (London, UK, 2022)	• Data center product called Foresight Optima DC+. • On-premises only. • Telemetry data taken from IT, storage and networking equipment, DCIM and BMS. • Automated control to improve energy usage at CPU level. • Tracks greenhouse gas Scope 1, 2 and 3 emissions. • Taxonomy reporting dashboard for the EU's CSRD.	• Open or closed-loop control. • Up to 52% IT load energy reduction claimed. • Reporting on ISO 30134 for compliance with the EU's CSRD. • 19% reduction in energy consumption by servers. • 14% energy savings and two-month payback at a large colocation facility.	BT Group, WWP.	Chip power management developed in partnership with Intel. Additional partnerships with BT, Exertis, Tech Mahindra, Mitsubishi Electric. Raised $10m series B (2023) from Wave Equity Partners.
TycheTools (Madrid, Spain, 2018)	• AI models and sensor-based telemetry provide monitoring, recommendations, and control of cooling equipment. • Software as a service only. • Proprietary sensors collect data on temperature, humidity and pressure.	• Open or closed-loop control. • Reduction in energy consumption (20% or more). • Compliance with ISO 14001 for environmental management systems, and ESG objectives.	Unnamed customers.	Winner of the 2022 data center energy efficiency startup program PERSEO, operated by Iberdrola. Academic partnership with Polytechnic University of Madrid.

BMS, building management software; CRAC, computer room air conditioning; CRAH, computer room air handler; CSRD, Corporate Sustainability Reporting Directive; DCIM, data center infrastructure management

What is different this time?

Uptime Intelligence's research shows that 64% of data center managers would now trust AI to make operational decisions (see *Uptime Institute Global Data Center Survey 2023*). Some colocation providers and enterprise data center operators have already started to deploy AI technologies in their facilities. There are several key factors driving this shift:

- Established vendors of data center infrastructure management (DCIM) and building management software (BMS) have prioritized adding incremental features requested by customers over developing completely new, AI-enhanced functionality or advanced analytics. This has left the door open to innovative startups.
- Data center operators and original equipment manufacturers are willing to establish partnerships with AI startups. This supports operational data sharing — overcoming a major obstacle in training AI models for data center applications.
- Tough new regulatory mandates for improved energy efficiency and sustainability, such as the EU's Corporate Sustainability Reporting Directive and the Digital Operational Resilience Act, will require the reporting of additional data center metrics. Promising new software tools may help with the reporting of energy use and carbon emissions.
- Human error contributes in some way to most data center outages, for example, when employees fail to follow correct standard operating procedures. More sophisticated automated controls using AI have the potential to reduce these incidents and prevent mistakes.
- Broader awareness of AI advances, such as generative AI, is pushing more organizations to experiment with novel approaches, even if the underlying technology is not generative or even new.

High-quality, relevant data sources are critical to ensure that AI systems produce reliable insights. Today, operators are in a much better position than ever before to acquire and process rich sets of operational data to find valuable patterns.

ML techniques can be applied to these datasets to look for interdependencies and anomalies. This differs from traditional business intelligence (BI) and analytics tools, which primarily look for linear trends and comparisons. BI and analytics tools are sometimes found lacking when analyzing complex systems like data centers, where an issue occurring in one part of the facility may have its cause in another location.

Startups thrive on partnerships

It takes months, or even years, for a startup to train and test ML models before they can be turned into commercial products or services. During this time, the relationship between the startup and established industry organizations is critical when building recognition and validation of the technology. All four AI startups tracked by Uptime Intelligence have forged strategic relationships within the data center industry to boost the development and commercial viability of their products:

- **QiO and Intel:** Chipmaker Intel entered a partnership with industrial AI startup QiO in 2021 to optimize the power consumption of its server chips. QiO's technology applies proprietary ML algorithms to manage processor power consumption through cycles of low activity (called C-states) and during code execution (P-states). The software monitors processor utilization patterns to predict the required minimum performance level while still meeting quality of service requirements (such as response times), then chooses the appropriate idle or performance state for the processor to adopt. These states largely govern voltage and clock speed selections, which helps reduce energy use. The tool claims an average 24% power consumption improvement per server without affecting application performance.
- **Phaidra and Google:** Phaidra was established by a team of Google DeepMind engineers who had previously worked on AlphaGo and Google's renewable energy forecasting technology. A precursor to the startup's technology has been deployed in Google's data centers since 2016. According to the startup, the cooling optimization algorithm reduced cooling energy use by 40%, which, in turn, improved the power usage effectiveness by 15%.
- **Coolgradient and Digital Realty:** AI-driven data center optimization startup Coolgradient entered a partnership with Digital Realty to support the AI training process and help shape its product for the market. The Coolgradient platform is now being used across 18 Digital Realty data centers.
- **TycheTools and Iberdrola:** Founded by José M Moya, associate professor of supercomputing at the Polytechnic University of Madrid, TycheTools uses proprietary sensors and a redundant wireless network to gather continuous environmental data across the data center. In June 2022, TycheTools won a data center energy efficiency startup challenge run by renewable energy specialist Iberdrola in partnership with Schneider Electric and Microsoft.

Predictive maintenance considerations

Ensuring optimal system performance is a key use case for ML models. Data center equipment has a long tail of gradual decay before it breaks down. This means a slow decline in efficiency and performance, which can often go undetected outside of scheduled maintenance or until the equipment fails.

Predictive maintenance can manage these risks by identifying and correcting equipment issues ahead of time. For example, it can identify when systems might be overworked, at risk of failure or when they can use less power, which can help extend the equipment's life span.

However, these specific predictive maintenance applications come with their own issues that need consideration (see **Table 2**).

Table 2

Example predictive maintenance tools and their related issues

Use case	Consideration
Identification of under-performing switches and servers, as well as recommendation of which ones to replace (QiO's tool). This can help data center operators avoid needlessly replacing healthy equipment and support more sustainable ways of prolonging IT life cycles.	Vendor equipment agreements typically involve the periodic replacement of parts, regardless of their condition. Data center operators will need to renegotiate these contracts to gain this level of flexibility.
Detecting differences in the operation of chiller modulating valves and harmonizing imbalances in CRAC fan speeds (as offered by Coolgradient). Both issues can lead to overworking and, ultimately, the failure of cooling equipment. Optimizing equipment performance reduces the risk of downtime and improves energy efficiency.	Maintenance schedules for data center equipment need to be followed to ensure warranties remain valid.

CRAC, computer room air conditioning

UPTIME INSTITUTE 2024

Condition-based maintenance is becoming a reality

Another approach to maintenance that increasingly leverages ML is condition-based maintenance (CBM). This strategy aims to maximize the life span of equipment through continuous monitoring and predictive analytics techniques. However, it is only possible for equipment that is connected to the network.

Whereas predictive maintenance uses aggregated sensor data and trends to predict future equipment degradation and failure, CBM uses real-time sensor data, such as temperature, pressure and vibration, to perform maintenance at the exact moment it is required and before a critical failure occurs.

This enables data center operators to reduce the number of hours spent on maintaining equipment and the number of times it must be taken out of operation, which lowers the risk of failure and human error.

Unlike other types of ML-based functionality that can be supported by startups and private companies, only the largest and most well-established equipment vendors can offer CBM because it relies on a global base of support staff who can respond to maintenance requests at short notice.

CBM is increasingly acknowledged as a viable data center maintenance model by regulatory and standards bodies. It is used in other sectors such as manufacturing, utilities, oil and gas, and maritime industries, but its use remains rare in data centers. Currently there are several large equipment vendors are developing data center-specific CBM services.

Achieving greater autonomous control

Many of the use cases described in **Table 2** require a human operator to monitor the software system, interpret the results and, if necessary, implement any changes. This so-called human-in-the-loop system is the most common approach to using AI in mission-critical settings.

Whereas a closed-loop system uses feedback within the environment to reduce errors and improve stability without any external involvement. A closed-loop ML model has the potential to fully automate the control of equipment without human intervention.

Reinforcement learning models learn through a system of penalties and rewards that are based on the feedback received from the environment.

Many organizations remain concerned about closed-loop autonomy. Nonetheless, exploring greater autonomy has the potential for significant benefits in the data center — last year alone, the Uptime Institute Resiliency Survey 2023 found that four in 10 data center owner operators suffered an IT service outage caused by human error.

Several of the AI startups investigated by Uptime Intelligence aim to offer autonomous control of facilities equipment, pending acceptance from their customers. Phaidra claims to offer Level 4 to 5 (the highest) autonomous control today, as defined by the Uptime Institute's data center maturity model (see **Table 3**). This means the software can operate data center cooling infrastructure with a level of human intervention, but it will yield the greatest benefits when it runs autonomously.

To enable this level of autonomy, Phaidra uses reinforcement learning (RL), a type of ML that has proven highly effective in the robotics and video game industries. Key members of Phaidra's team have previously helped develop AlphaGo, an AI model trained in the rules of board game Go that went on to beat the world champion in 2017.

RL models learn through a system of penalties and rewards that are based on the feedback received from the environment. Data center documentation, such as standard operating procedures (SOPs) and method of procedures (MOPs), can provide the rules for training the AI software. These rules would define the best operating practices in the data center and the optimal settings for systems and equipment.

Since environmental factors constantly change (such as performance, temperature and water pressure), equipment can deviate from its optimal settings. It then requires intervention to adjust these settings so it can operate correctly again. Based on feedback, the AI application will test and retest system settings until it achieves the right fit for the operating conditions. In a closed-loop system, the RL model can become self-learning, supporting continuous improvement without human oversight of its control decisions.

The attraction of such AI tools is only going to grow; they will be especially useful in running remote, "lights out" data centers that do not have any permanent staff on site.

Evolution is inevitable

Uptime Institute has long argued that data center management tools need to evolve toward greater autonomy. The data center maturity model (see **Table 3**) was first proposed in 2019 and applied specifically to DCIM. Four years later, there is finally the beginning of a shift toward Level 4 and Level 5 autonomy, albeit with a caveat — DCIM software alone will likely never evolve these capabilities. Instead, it will need to be combined with a new generation of data-centric tools.

Table 3

The Data Center Management Maturity Model

Level 1	Level 2	Level 3	Level 4	Level 5
No integration of infrastructure data. Basic monitoring is provided by the equipment vendor software and the BMS.	Software installed to monitor environmental and equipment power use. Able to adjust basic controls (e.g., cooling) in line with demand.	Software can track physical data center equipment characteristics, location and operational status. Energy and environmental data are used to reduce risks and waste.	Machine learning models are used for prediction, service management and multiple views, optimizing the data center in near real time. AI is applied to DCIM-based data lakes for advanced analytics.	AI-driven integrated management software adjusts data center behavior and makes the best use of resources according to goals, rules and service requirements throughout its life cycle.

UPTIME INSTITUTE 2024

Not every organization will, or should, take advantage of Level 4 and Level 5 functionality. These tools will provide an advantage to the operators of modern facilities that have exhausted the list of traditional efficiency measures, such as those achieving PUE values of less than 1.3.

For the rest, being an early adopter will not justify the expense. There are cheaper and easier ways to improve facility efficiency that do not require extensive data standardization efforts or additional skills in data science.

At present, AI and analytics innovation in data center management appears to be driven by startups rather than established software vendors. Few BMS and DCIM developers have integrated ML into their core products, and while some companies have features in development, these will take time to reach the market — if they ever leave the lab.

Uptime Intelligence is tracking multiple early-stage or private companies that use facilities data to create ML models and already have products or services on the market. It is likely more will emerge in the coming months and years.

These businesses are creating a new category of software and services that will require new types of interactions with all the moving parts inside the data center, as well as new commercial strategies and new methods of measuring the return on investment. Not all of them will be successful.

The speed of mainstream adoption will depend on how easy these tools will be to implement. Eventually, the industry will arrive at a specific set of processes and policies that focus on benefiting from equipment data.

New tools bring new challenges

The adoption of ML-powered tools for infrastructure management will require owners and operators to recognize the importance of data quality. They will not be able to trust the output of ML models if they cannot trust their data — and that means additional work on standardizing and cleaning their operational data stores.

In some cases, data center operators will have to hire analysts and data scientists to work alongside the facilities and IT teams.

Data harvesting at scale will invariably require more networking inside the data center — some of it wireless — and this presents a potentially wider attack surface for cybercriminals. As such, cybersecurity will be an important consideration for any operational AI deployment and a key risk that will need to be continuously managed.

What role might generative AI play in the data center?

Advances in AI are expected to change the way work is done across numerous organizations and job roles. This is especially true for generative AI tools, which are capable of synthesizing new content based on patterns learned from existing data.

This section explores the rise of large language models (LLMs) and generative AI. It examines whether data center managers should be as dismissive of this technology as many appear to be and considers whether generative AI will find a role in data center operations.

Trust in AI is affected by the hype

Data center owners and operators are starting to develop an understanding of the potential benefits of AI in data center management. So long as the underlying models are robust, transparent and trusted, AI is proving beneficial and is increasingly being used in areas such as predictive maintenance, anomaly detection, physical security, and filtering and prioritizing alerts.

At the same time, a deluge of marketing messages and media coverage is creating confusion around the exact capabilities of AI-based products and services. Conversations with data center managers show that there is widespread caution in the industry.

Machine learning, deep learning and generative AI

Deep learning — an advanced approach to ML inspired by the workings of the human brain — makes use of deep neural networks (DNNs) to identify patterns and trends in seemingly unrelated or uncorrelated data.

Generative AI is not a specific technology but a type of application that relies on the latest advances in DNN research. Much of the recent progress in generative AI is down to the transformer architecture — a method of building DNNs unveiled by Google in 2017 as part of its search engine technology and later used to create tools such as ChatGPT, which generates text, and DALL-E for images.

Deep learning — an advanced approach to ML inspired by the workings of the human brain — makes use of deep neural networks (DNNs) to identify patterns and trends in seemingly unrelated or uncorrelated data.

Transformers use attention mechanisms to learn the relationships between datapoints — such as words and phrases — largely without human oversight. The architecture manages to simultaneously improve output accuracy while reducing the duration of training required to create generative AI models. This technique kick-started a revolution in applied AI that became one of the key trends of 2023.

LLMs which are based on transformers, like ChatGPT, are trained using hundreds of gigabytes of text and can generate essays, scholarly or journalistic articles and even poetry. However, they face an issue that differentiates them from other types of AI and prevents them from being embraced in a mission-critical setting: they cannot guarantee the accuracy of their output.

The good, the bad and the impossible

With the growing awareness of the benefits of AI comes greater knowledge of its limitations. One of the key limitations of LLMs is their tendency to "hallucinate" and provide false information in a convincing manner. These models are not looking up facts; they are pattern-spotting engines that guess the next best option in a sequence.

This has led to much-publicized news stories about early adopters of tools like ChatGPT landing themselves in trouble because they relied on the output of generative AI models that contained factual errors. Such stories have likely contributed to the erosion of trust in AI as a tool for data center management — even if these types of issues are exclusive to generative AI.

It is a subject of debate among researchers and academics whether hallucinations can be eliminated entirely from the output of generative AI, but their prevalence can certainly be reduced.

One way to achieve this is called grounding and involves the automated cross-checking of LLM output against web search results or reliable data sources. Another way to minimize the chances of hallucinations is called process supervision. Here, the models are trained to reward themselves for each correct step of their reasoning rather than the right conclusion.

Finally, there is the creation of domain-specific LLMs. These can be either built from scratch using data sourced from specific organizations or industry verticals, or they can be created through the fine-tuning of generic or "foundational" models to perform well-defined, industry-specific tasks. Domain-specific LLMs are much better at understanding jargon and are less likely to hallucinate when used in a professional setting because they have not been designed to cater to a wide variety of use cases.

The propensity to provide false information with confidence likely disqualifies generative AI tools from ever taking part in operational decision-making in the data center — this is better handled by other types of AI or traditional data analytics. However, there are other aspects of data center management that could be enhanced by generative AI, albeit with human supervision.

Is there a place for generative AI in data center management?

First, generative AI can be extremely powerful as an outlining tool for creating first-pass documents, models, designs and even calculations. For this reason, it will likely find a place in those parts of the industry that are concerned with the creation and planning of data centers and operations. However, the limitations of generative AI mean that its accuracy can never be assumed or guaranteed, and that human expertise and oversight will still be required.

Generative AI also has the potential to be valuable in certain management and operations activities within the data center as a productivity and administrative support tool.

The Uptime Institute Data Center Resiliency Survey 2023 reveals that 39% of data center operators have experienced a serious outage because of human error, of which 50% were the result of a failure to follow the correct procedures. To mitigate these issues, generative AI could be used to support the learning and development of staff with different levels of experience and knowledge.

Generative AI could also be used to create and update the method of procedures (MOPs), standard operating procedures (SOPs) and emergency operating procedures (EOPs), which can often get overlooked due to time and management pressures. Other examples of potential applications of generative AI include the creation of:

- Technical user guides and operating manuals that are pertinent to the specific infrastructure within the facility.
- Step-by-step maintenance procedures.
- Standard information guides for new employee and/or customer on-boarding.
- Recruitment materials.
- Risk awareness information and updates.
- Q&A resources that can be updated as required.

In conversations with Uptime Institute members and other data center operators, some said they used generative AI for purposes like these — for example, when summarizing notes made at industry events and team meetings. These members agreed that LLMs will eventually be capable of creating documents like MOPs, SOPs and EOPs.

It is crucial to note that in these scenarios, AI-based tools would be used to draft documents that would be checked by experienced data center professionals before being used in operations. The question is whether the efficiencies gained in the process of using generative AI offset the risks that necessitate human validation.

The market is moving quickly

The development of new ML hardware, supporting software platforms and model architectures is moving at a rapid pace. Adoption of ML-based tools is on the agenda for countless businesses around the world, and data centers are no exception.

This technology enables owners and operators to improve their efficiency and sustainability profile, and commercial viability without having to make any physical changes to the facilities themselves. While the impact of generative AI on data center management is uncertain, there are use cases where this technology could prove useful, if applied with caution.

AI in data center management is increasingly seen as a differentiator. Many owners and operators have embarked on their first ML-powered optimization projects in 2023, and more organizations will join them in 2024.

Appendix

12 useful terms for understanding AI

1: AI summers and AI winters

AI research goes through periods of intense interest and excitement when new technological advances are made, often called "AI summers." These might be followed by "AI winters" — periods of stagnation when developers reach the technological limits of the day. There have been at least two AI winters to date: in the 1970s and then in the late 1980s. In both instances, funding for research dried out once it became apparent that AI would be unable to deliver on expectations without more powerful and less expensive compute. .

2: Data labeling

In most cases, training data needs at least some context to be useful. Data labeling is the process of identifying raw data (images, text files, videos, etc.), organizing it and adding meaningful labels so that a ML model can learn from it.

3: Data quality

For an ML system to function correctly, it needs high-quality data — this applies to both the data used for model training and the data that is fed to the model in production, during inference. Low-quality data can lead to ML outputs that are prone to errors and inconsistencies.

Examples of low-quality data include out-of-date, inaccurate or incomplete data; unlabeled or wrongly labeled data, which makes it difficult to categorize information correctly; and duplicate, outlier and biased data, which can lead to undue weighting of certain properties, resulting in distorted or prejudiced results.

4: Supervised learning

Supervised ML occurs when an algorithm is trained using labeled data and is rewarded or optimized to generate specific outputs. Supervised learning is widely deployed in image recognition.

5: Unsupervised learning

Unsupervised algorithms use unlabeled data (i.e., raw data) for training. The algorithm is provided with an input dataset but is not rewarded or optimized for specific outputs. Instead, it is trained to group objects by common characteristics. Product recommendation engines often rely on ML models that have been trained using unsupervised learning.

6: Semi-supervised learning

Semi-supervised learning uses a mix of labeled and unlabeled data to train an algorithm. In this process, the algorithm is first trained on a small amount of labeled data before it is trained on a much larger amount of unlabeled data.

7: Reinforcement learning

Reinforcement learning is an ML technique in which positive and negative values are assigned to desired and undesired actions. The goal is to encourage models to avoid the negative training examples and seek out the positive, learning how to maximize rewards through trial and error. Autonomous driving is one of the most popular examples of reinforcement learning.

Appendix *(cont)*

8: Deep neural networks

The latest wave of innovation in AI started in the 1990s with the popularization of deep neural networks (DNNs), which are ML models that mimic the structure of neurons in the brain. DNNs consist of several interconnected processing layers (hence the term "deep") that extract high-level features from the data provided. Each processing layer passes on a more abstract representation of the data to the next layer, with the final layer translating the output back into a more human-like insight. Unlike traditional ML models, which require data to be labeled, deep learning models can ingest large amounts of unlabeled data.

9: Generative AI

Generative AI is not a specific technology but a type of application that relies on the latest advances in DNN research. Generative AI models are capable of synthesizing new content based on patterns learned from existing data. For example, large language models (LLMs) are trained using hundreds of gigabytes of text and can generate essays, scholarly or journalistic articles and even poetry.

10: Federated learning

Federated learning is an approach to ML that addresses the issues of data privacy and security by enabling multiple organizations to collaborate on training a single ML model without sharing any data.

11: Foundational models

Foundational models are ML models that are designed to make it faster to build other models. This is important because it drives down the cost and resources needed for training.

12: Domain-specific LLMs

Domain-specific LLMs can be either built from scratch using data sourced from specific organizations or industry verticals, or by fine-tuning foundational models to perform specific tasks. Because they are designed to cater to a limited variety of use cases, domain-specific LLMs are better at understanding jargon and are less likely to hallucinate when used in a professional setting.

About the authors

Max Smolaks

Max Smolaks is a Research Analyst at Uptime Intelligence. His expertise spans digital infrastructure management software, power and cooling equipment, and regulations and standards. He has 10 years' experience as a technology journalist, reporting on innovation in IT and data center infrastructure.

msmolaks@uptimeinstitute.com

John O'Brien

John O'Brien is Uptime Intelligence'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, automation, and emerging AIOps, DataOps and FinOps practices.

jobrien@uptimeinstitute.com

Daniel Bizo

Daniel Bizo is Uptime Intelligence's Research Director. He has been covering the business and technology of enterprise IT and infrastructure in various roles, including more than a decade as an industry analyst and advisor.

dbizo@uptimeinstitute.com

All general queries

Uptime Institute
405 Lexington Avenue
9th Floor
New York, NY 10174, USA
+1 212 505 3030

info@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.

Uptime Institute is headquartered in New York, NY, with offices in Seattle, London, Sao Paulo, Dubai, Singapore, and Taipei.

For more information, please visit **www.uptimeinstitute.com**