uptime
INTELLIGENCE

INTELLIGENCE UPDATE

# Enterprises will deploy inference in-house — if they can

Max Smolaks 23 Mar 2026

The development of large language models (LLMs) is a complex process, requiring specialized infrastructure and skills, as well as the ability to differentiate the result — since there is little value in replicating the work done by others.

It is becoming clear that few organizations will choose to train their own LLMs, opting instead to rely on the growing number of commercial and open-weight models. This is a sign of market maturity: when organizations need enterprise software, they don't build it from scratch; they purchase it from established vendors or deploy open-source alternatives.

After choosing a model, the next step is to determine where it will be hosted to perform inference — using a copy of the model to generate outputs in response to user inputs. Here too, the choices are becoming clearer, with distinct benefits and drawbacks to delivering AI on-premises, in colocation, cloud, or as a service.

A recent Uptime Intelligence report explored the economics of inference across data center venues (see *Where to deploy AI inference: a guide to the economics*). Based on its findings, we can make three further observations:

### 1: Running generative AI in your own data center can be the lowest-cost option — but few organizations will be able to achieve these savings.

The report demonstrates that consistently high hardware utilization is the key to cost-efficient AI compute. Hyperscale cloud providers can deliver lower-cost services primarily due to the economies of scale and their ability to maximize the use of their virtualized IT infrastructure. Enterprises can only match them by leveraging existing infrastructure and in-house skills (essentially sunk costs), while maintaining high utilization of the hardware.

In practice, however, few organizations track server utilization, and even fewer manage it. The Uptime Institute IT and Power Efficiency Survey 2024 found that 53% of respondents had no utilization objective for their overall server fleet. Among those that did, only 29% reported average utilization above 65% — the threshold at which AI infrastructure becomes more cost-effective on-premises.

In addition, hardware utilization for LLM-based services is notoriously difficult to predict; it is shaped by the hidden system prompts provided by developers, the complexity of the end-user inputs, and the number of tokens generated in response. This inherent unpredictability of the workload makes the public cloud more attractive, as it allows customers to only pay for the capacity they use.

**2: The choice of the model will sometimes dictate the choice of infrastructure — and vice versa.**

While latency, data locality, governance, and operational control are important considerations, enterprises also need to account for limited model portability.

Organizations planning to use flagship models from providers such as OpenAI or Anthropic are restricted to consuming them as a service or via a cloud platform (which is more expensive). These models cannot be deployed on-premises or in a colocation environment. Furthermore, opting to use efficient inference hardware developed by a cloud vendor (e.g., Google TPUs, AWS Trainium or Microsoft Maia) locks an organization into purchasing the vendor's cloud services and using models that have been adapted to run on that hardware. Therefore, decisions about hardware and data center venue should not be taken separately from decisions about the choice of LLMs.

Freely distributed open-weight models paired with GPU-based servers from vendors such as Dell, HPE, Lenovo and Supermicro offer the greatest flexibility in deployment. This path is likely to emerge as the preferred option for enterprises requiring full control over their data during inference.

**3: Smaller models are a little less capable and a lot cheaper to run.**

The gap between the facility requirements of typical corporate IT — averaging around 7 kW per rack in 2025 — and the demands of dense AI compute necessary to deploy the so-called "frontier" models is widening. If an AI cluster cannot be accommodated within existing data halls and necessitates additional space, any cost savings over public cloud evaporate.

Not all models require dense, liquid-cooled infrastructure. Smaller, less complex LLMs are capable of delivering functionality such as transcription, translation and summarization, and can be deployed on-premises within existing facilities with minimal changes to cooling and power distribution (see *Why bigger is not better: gen AI models are shrinking*).

For organizations starting small, leveraging in-house facilities or existing colocation space can be the more attractive option; the cost per token remains low even without achieving high hardware utilization. This is likely why, despite the commercial attraction of public cloud, Uptime Intelligence surveys show on-premises data centers as the most popular venue for AI workloads.

## ABOUT THE AUTHOR

### Max Smolaks

24 Mar 2026

Max is a Research Analyst at Uptime Institute Intelligence. Mr Smolaks' 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**

## About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. With over 4,000 awards issued in over 122 countries around the globe, and over 1,100 currently active projects in 80+ countries, Uptime has helped tens of thousands of companies optimize critical IT assets while managing costs, resources, and efficiency. For over 30 years, the company has established industry-leading benchmarks for data center performance, resilience, sustainability, and efficiency, which provide customers assurance that their digital infrastructure can perform across a wide array of operating conditions at a level consistent with their individual business needs. Uptime's Tier Standard is the IT industry's most trusted and adopted global standard for the design, construction, and operation of data centers.

Offerings include the organization's Tier Standard and Certifications, Management & Operations reviews and assessments including SCIRA-FSI financial sector risk assessment, the Sustainability Assessment, and a broad range of additional risk management, performance, availability, and related offerings. Uptime Education training programs have been successfully completed by over 100,000 data center professionals, such as the much-valued ATD (Accredited Tier Designer) and AOS (Accredited Operations Specialist). The Uptime Education curriculum has been expanded by the acquisition of CNet Training Ltd. In 2023.

Uptime Institute is headquartered in New York, NY, with offices in London, Sao Paulo, Dubai, Riyadh, and Singapore, and full-time Uptime professionals based in over thirty-four countries around the world.

For more information, visit www.uptimeinstitute.com