FIDE PARTNERS

The DeepSeek Moment

https://fidepartners.com/
Tel: +34 910 244 113
info@fidepartners.com

February 2025

Implications for Global Data Center Infrastructure

Authors: Jim Andrew, Felipe Sarmiento, Camilo Panqueva.

Inside

Infrastructure requirements for frontier AI models remain substantial despite efficiency gains

DeepSeek's breakthrough in AI model efficiency challenges traditional infrastructure assumptions and claims to deliver state-of-the-art performance at a fraction of the cost.

This analysis examines the technical innovations driving this achievement, assesses the true infrastructure requirements, and explores the implications for future global Data Center demand.

While the initial market reaction has focused on cost reduction, the long-term implications suggest accelerated AI adoption and sustained infrastructure demand.

The DeepSeek moment: implications for global Data Center infrastructure

Introduction

The emergence of DeepSeek in late 2024 marked a potential inflection point in artificial intelligence infrastructure development.

In December 2024, High-Flyer, a Chinese hedge fund known for AI trading algorithms, launched DeepSeek V3, achieving performance comparable to OpenAI's GPT-4o at a claimed fraction of the traditional training costs.

The subsequent January 2025 release of DeepSeek R1 further challenged industry assumptions by matching or exceeding the capabilities of leading models like OpenAI's O1 and Anthropic's Claude 3.5.

DeepSeek's technical approach, combining advanced Mixture of Experts (MoE) architecture, novel reinforcement learning techniques, and significant engineering optimizations, alongside its open weights policy under MIT license, represented a marked departure from the closed model approaches of major competitors.

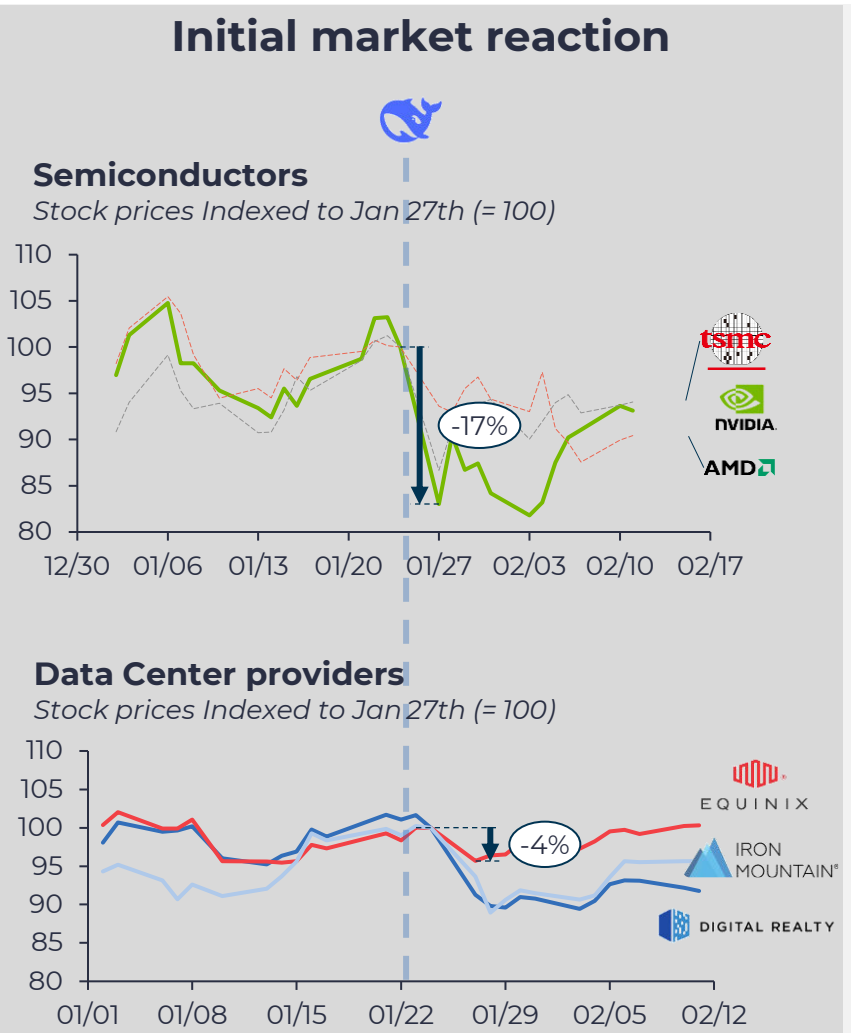
This development has triggered a widespread reassessment of AI infrastructure costs and deployment strategies.

The reported USD5.6 million training cost for DeepSeek R1, compared to the previous frontier model costs of hundreds of millions, initially suggested a fundamental shift in computational requirements for state-of-the-art AI models.

However, this figure obscures the true infrastructure demands and total investment required for state-of-the-art AI development

As the industry grapples with these claims, data center operators and infrastructure providers face critical questions about future capacity requirements, investment strategies, and the evolution of AI-ready facilities.

This analysis examines the technical reality behind DeepSeek's efficiency claims, assesses their market implications, and explores the lasting impact on global data center infrastructure.



Market over reaction to DeepSeek's efficiency claims

In the first few days after the R1 announcement, NVIDIA experienced a 17% decline, representing a USD600 billion in market value, reflecting investor concerns about potential disruption of the AI chip market.

DC REITs and AI infrastructure providers experienced increased volatility as markets processed the implications for DC demand and deployment strategies.

OpenAI implemented two rapid price cuts and accelerated its O1 reasoning model deployment to the general public.

In particular, the training cost claims, and the Open Weights announcement impacted the market by suggesting that SOTA models could be built using less compute and a potential shift to locally deployed models.

Novel architecture and training approach achieves o1-level performance

DeepSeek's technical breakthrough begins with its V3 model's fundamental architectural innovations. At its core, the model employs a Sparse Mixture of Experts (MoE) architecture with 671 billion parameters, of which only 37 billion are active at any time.

This selective activation approach, combined with Multi-head Latent Attention (MLA) for memory efficiency and a novel shared/routed expert system, creates a foundation that matches GPT-4o performance while enabling more efficient training and inference.

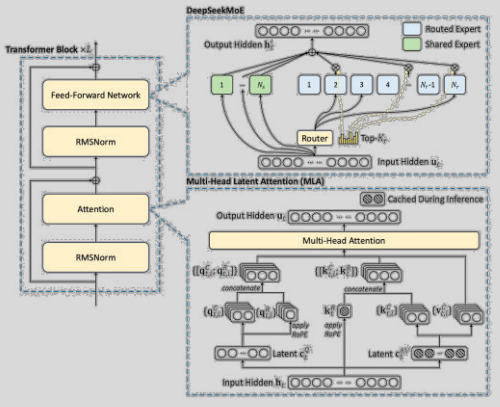
While these innovations are largely evolutions of previously known techniques, their implementation represents a significant optimization of existing approaches. The architecture's efficiency gains stem from its ability to activate only the most relevant neural pathways for specific tasks, effectively reducing computational requirements without sacrificing model capabilities.

Building on V3's efficient architecture, DeepSeek R1 introduced a revolutionary training approach that inverts traditional fine-tuning pipelines. Unlike conventional models that begin with supervised fine-tuning followed by reinforcement learning, DeepSeek prioritizes RL before supervised training.

This methodology incorporates three key innovations: an RL-first training pipeline, cold start data utilization, and Group Relative Policy Optimization (GRPO).

The most significant advancement comes from the R1-Zero variant, which removes humans from the feedback loop entirely, allowing the model to develop its own chain-of-thought reasoning patterns autonomously. This capability has proven particularly effective in mathematical reasoning tasks.

DeepSeek V3 Architecture



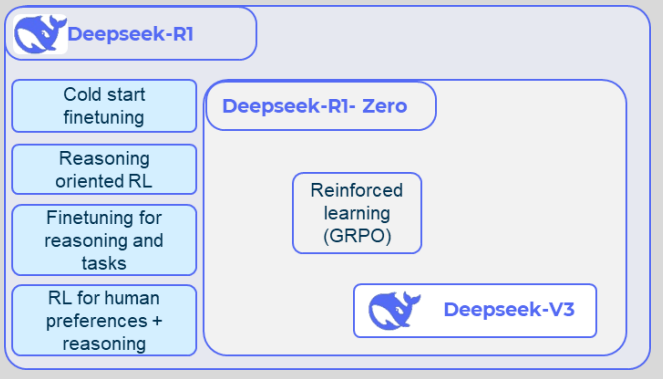
While models like LLaMA v3 use dense architectures (all parameters active) and GPT-4o employs limited sparsity (approximately 1:4 expert activation ratio), V3's selective expert activation achieves a remarkable 1:18 ratio, using only 37bn of 671bn parameters actively.

This extreme sparsity through MoE, combined with optimized multi-head latent attention, enables GPT-4o-level performance with significantly reduced computational demands for training and inference

R1 inverts the traditional AI training paradigm used by OpenAI, Anthropic, and others, where supervised fine-tuning (SFT) precedes reinforcement learning (RL).

By prioritizing RL before SFT and introducing autonomous reasoning development, R1 eliminates the human feedback bottleneck that typically constrains training efficiency.

DeepSeek R1 training pipeline



Technical Foundations

Performance analysis

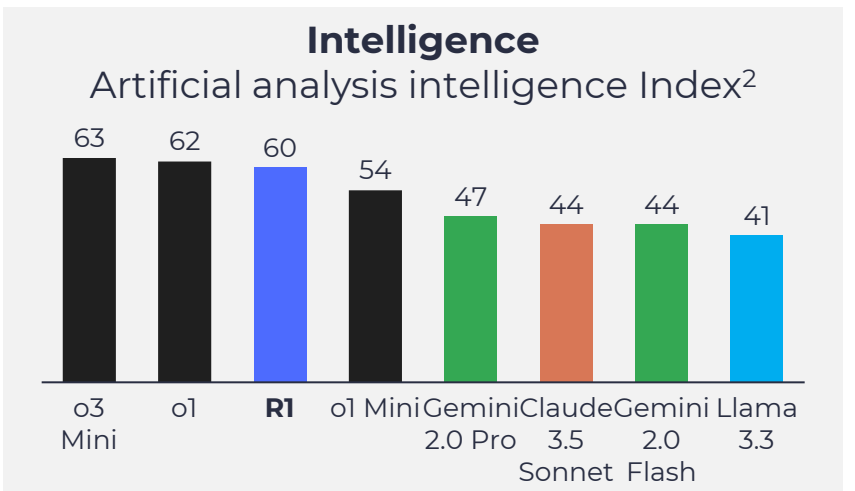
DeepSeek R1 demonstrates performance parity with leading models across major benchmarks, particularly excelling in mathematical reasoning and code generation. The model's efficiency-focused architecture enables these results while maintaining substantially lower infrastructure requirements.

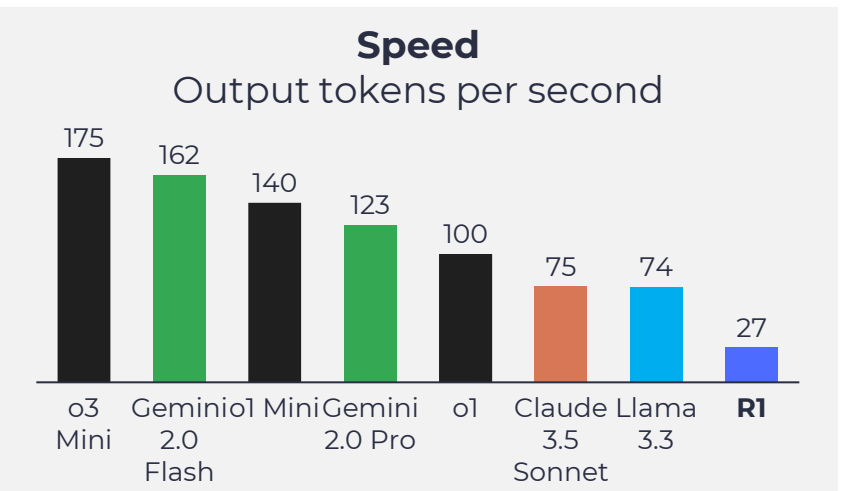
While performance varies by domain, with R1 showing strength in code generation and Chinese language tasks, it occasionally lags in some English language benchmarks.

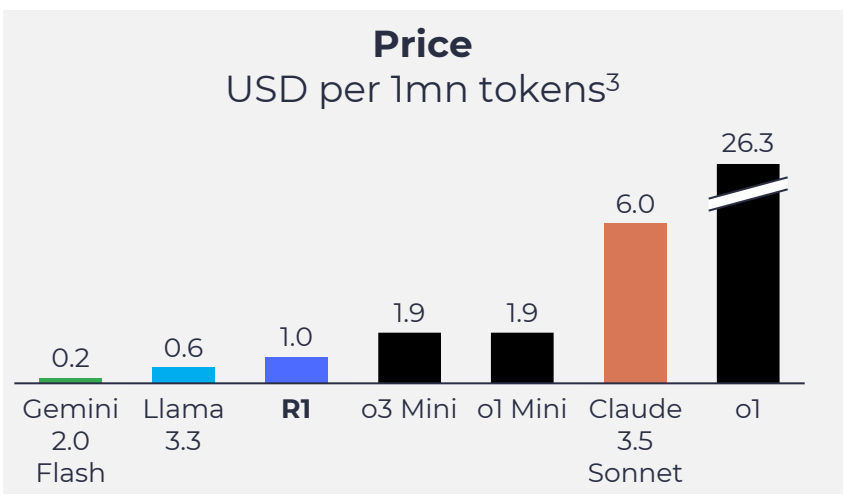
These results validate DeepSeek's architectural and training innovations. They demonstrate competitive performance across diverse tasks while maintaining significantly lower reported training costs.

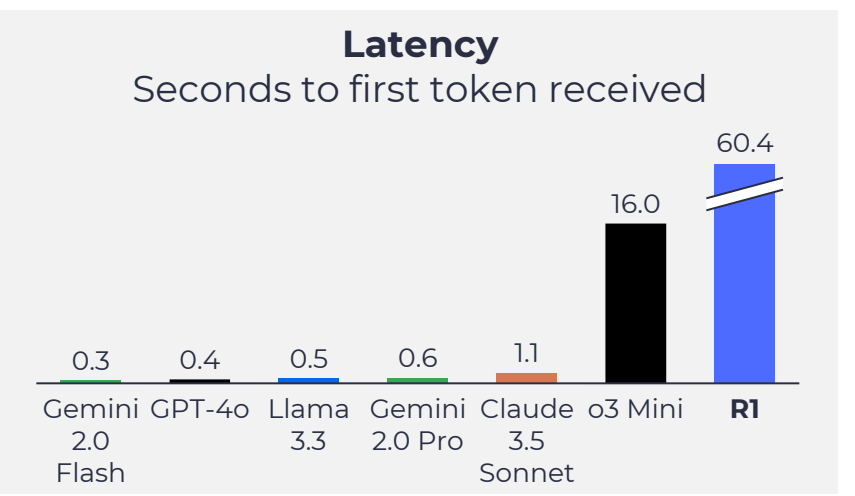
However, the significantly higher latency and lower throughput suggest that inference performance faces substantial compute constraints, the performance metrics indicate that achieving theoretical efficiency gains in practice requires substantial infrastructure investment, particularly for maintaining consistent service quality across a growing user base.

DeepSeek R1 achieves parity with frontier models across some use cases









Note 1: All metrics reflect performance as of Q1 2025 and may change over time due to model updates or API revisions. Underlying metrics presented here are sourced from artificialanalysis.ai (Q1 2025).

Note 2: The "Artificial Analysis Intelligence Index" is a composite metric covering multiple dimensions of intelligence, including (but not limited to) MMLU-Pro, GPQA Diamond, Humanity's Last Exam, SciCode, AIME, and MATH-500.

Note 3: Displayed as USD per 1 million tokens, blending input and output costs at a ratio of 3:1.

DeepSeek's USD5.6mn training cost claim masks ~USD2.0bn true infrastructure investment

While DeepSeek's reported USD5.6mn training cost has captured market attention, this figure represents only the direct GPU compute costs for the pre-training run.

The development of core technological innovations like Multi-Head Latent Attention demanded significant resources beyond pure training costs that require months of iterative testing, architectural experimentation, and continuous GPU utilization for ablation studies.

The full cost structure reveals a substantially larger investment: specialized sources estimate total server capex approaching USD1.6 billion, with associated operational costs of approximately USD900 million.

This infrastructure investment includes a distributed network of roughly 20,000 high-end GPUs, split between H800s and H100s (acquired before export restrictions), with additional orders of China-specific H20 GPUs in process.

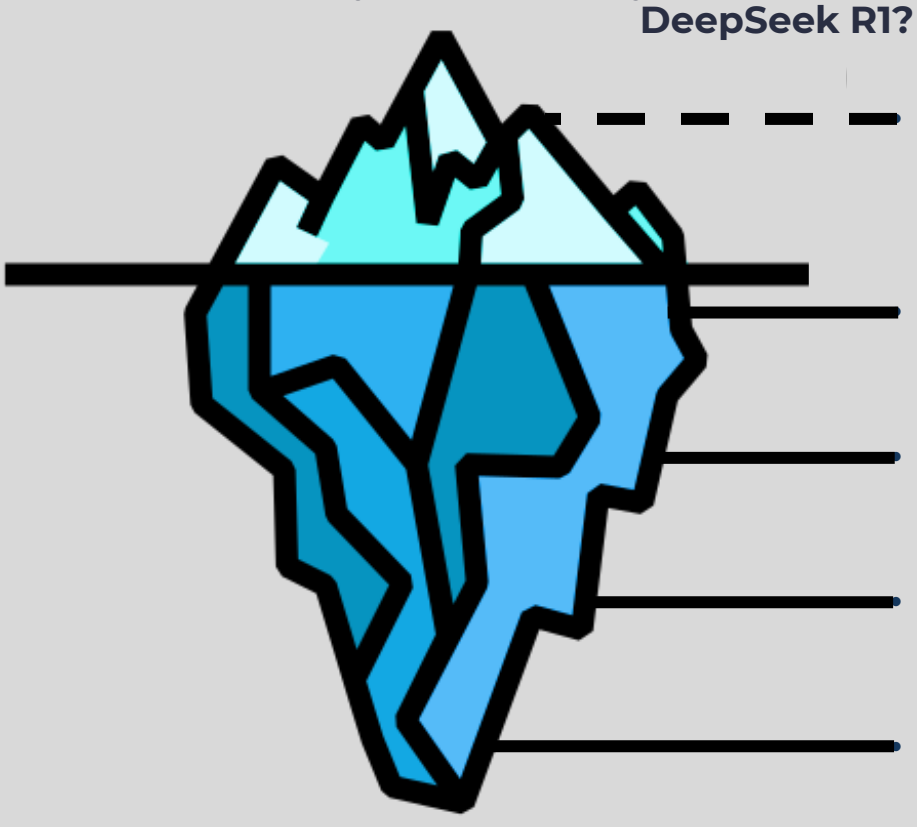
DeepSeek's infrastructure footprint significantly exceeds initial market estimates, comprising approximately ~60,000 AI capable GPUs across multiple generations (~10,000 each of A100s, H800s, and H100s, plus ~30,000 H20s).

The company also co-locates capacity in distributed data centers, sharing resources between AI development and High-Flyer's trading operations.

These requirements demonstrate that even with architectural efficiency gains, frontier AI development demands massive infrastructure investment.

The company's emphasis on training efficiency metrics may be strategically motivated by current GPU export controls and heightened scrutiny of Chinese AI infrastructure development, as DeepSeek has limited incentive to disclose its full computational capabilities.

Beyond efficiency claims what is the true cost of DeepSeek R1?



Training cost
(USD5.6mn - allegedly)

Server capex - last 4 years
(USD1.6bn)

Server operation costs - 4 years
(~USD900mn)

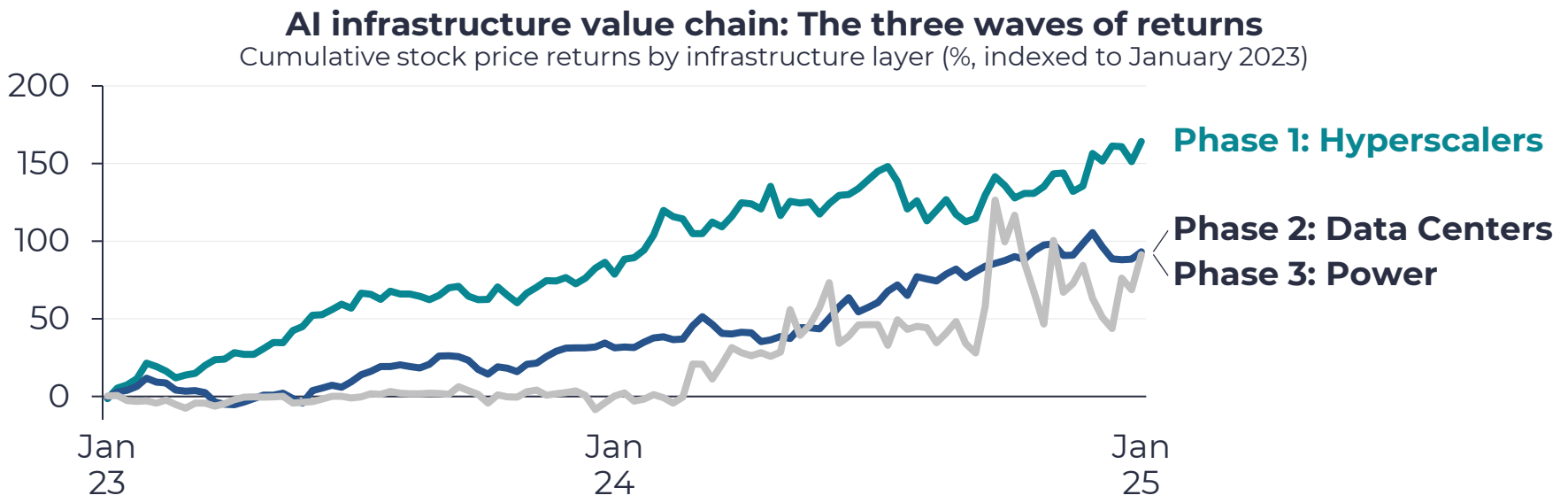
Research cost
(?)

HR costs
(?) - DeepSeek is hiring personnel offering salaries above USD1mn, more than other Chinese tech and AI labs. IT currently counts with ~150 FTEs, and has plans to grow keep growing

Source: Fide Partners research and analysis, semianalysis

Impact on global Data Center infrastructure

Will DeepSeek slow the data center deployment thrive?



Note 4: Based on market data as of February 14, 2025, indexed to the weighted average stock price of January 2023. Hyperscalers include selected CSPs with proprietary LLMs. Data Centers include selected data center REITs worldwide. Power includes selected energy players in the US.

Since the introduction of ChatGPT, LLMs have increased the value of AI value chain players. While hyperscalers and LLM developers have captured the largest capital return, LLMs have dramatically impacted data center operators and power providers.

The infrastructure requirements for inference and training have led to an increased deployment of AI-ready data centers. Despite announced AI data center investments by hyperscalers, DeepSeek's introduction has raised questions about a possible reduction in infrastructure deployment due to achieved efficiencies, which we see unlikely.

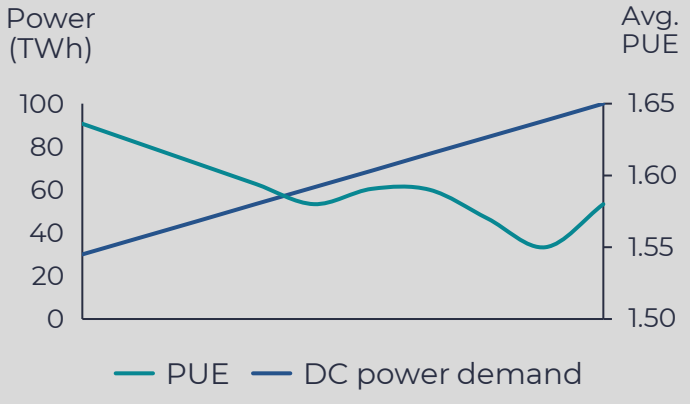
Although DeepSeek has achieved notable efficiencies in training and inference processes, potentially reducing computational needs per model, these gains are likely to accelerate AI adoption following the Jevon's paradox - improved efficiency leads to increased total consumption.

Early market reactions focused on potential infrastructure cost reductions, but as discussed in the previous sections, evidence indicates efficiency gains enable new AI applications rather than reducing overall compute needs.

Even considering a widespread adoption of DeepSeek innovations, frontier models still require massive data center deployments. Training demands extensive compute clusters with appropriate power and cooling infrastructure, while inference needs geographically distributed facilities for latency optimization.

We expect HS and LLM developers to complete their planned DC deployments and announce further campuses in the future, continuously increasing their capital intensity ratio (CIR) despite further achieving efficiency gains mirroring historical trends of previous technological advances like semiconductors.

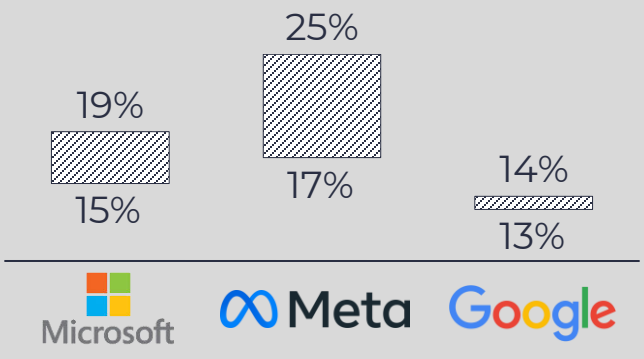
Data Center power efficiency and demand in the US (2014-2023)



Source: Uptime Institute, LBL ETA

HS's Capital Intensity Ratio[5]

capex/total revenues, % - bear case to base case



Note 5: Based on JP Morgan's estimate for CIR by 2027, using Bloomberg's data as of January 2025. The "bear" case reflects 2015-2019 averages for each company

FIDE PARTNERS

Conclusions: expected infrastructure evolution in the wake of DeepSeek

DeepSeek's emergence represents a significant evolution in AI model development, though its implications for infrastructure requirements are more nuanced than initial market reactions suggested.

While architectural innovations and training methodology improvements demonstrate potential for increased efficiency, the reality of frontier AI development still demands massive infrastructure investment. The reported USD5.6 million training cost masks a total infrastructure investment exceeding USD2.5 billion, highlighting the sustained importance of robust AI-ready facilities.

Export controls have accelerated efficiency innovation but also created a bifurcated infrastructure landscape between unrestricted and restricted markets. Despite efficiency gains, historical efficiency gains suggests that these improvements will drive increased AI adoption and expanded infrastructure requirements rather than reduced demand.

The capital intensity ratios of major hyperscalers remain high, indicating continued substantial infrastructure investment despite architectural efficiency improvements.

Contact:

Jim Andrew
Partner
Mail: jim.andrew@fidepartners.com
LinkedIn: Jim Andrew

Felipe Sarmiento
Manager
Mail: felipe.sarmiento@fidepartners.com
LinkedIn: Felipe Sarmiento

Camilo Panqueva
Consultant
Mail: camilo.panqueva@fidepartners.com
LinkedIn: Camilo Panqueva

Direction:

Boston:
50 Milk Street,
Planta 15, C.P. 02109.
Boston, MA

London:
Aldwych House
London, WC2B 4HN
United Kingdom

Madrid:
C/Don Ramón de la Cruz, 6, 1°
28001 - Madrid
Spain

Bogotá:
Carrera 11A #98-50
Ofc. 704, Edificio Punto99
110221, Bogota
Colombia

Mexico