Financial Issues are Driving the AI Agenda

Financial Issues are Driving the AI Agenda

Executive Summary

The artificial intelligence sector stands at a curious inflection point. Whilst the technology continues to advance and capture imaginations, the economic foundations upon which the industry has been built are beginning to show cracks. What emerges is not simply a story of technical evolution, but rather a fundamental reordering driven by the hard realities of cost, competition, and capital allocation.

The Limits of Scale

The pathway to superior AI performance seemed straightforward: larger models, more data, greater computing power. This scaling paradigm delivered remarkable results, transforming foundation models from academic curiosities into commercial juggernauts.

Recent evidence suggests this approach is encountering diminishing returns. Researchers studying advanced reasoning systems in 2025 found that adding more computational steps no longer delivered proportionate improvements. The performance curve has flattened whilst the cost curve continues its relentless ascent.

The industry's response has been to pivot towards inference scaling – essentially making models "think" longer at the point of use rather than simply pre-training them to be larger. Whilst this approach can yield performance gains, it carries significant cost implications. Each query consumes more compute, translating directly into higher operational expenses.

The economics quickly become challenging, especially for organisations running billions of queries daily. Moreover, inference scaling cannot substitute for fundamental knowledge gaps in the underlying models—it merely provides additional processing time to work with what the model already knows.

Commoditisation

As the capability differences narrow between leading foundation models, the sector is experiencing what can only be described as commoditisation. When flagship models offer broadly similar performance across most tasks, competition inevitably shifts to price. The average cost of inference has been falling at approximately 86% annually, driven by both fierce competition and economies of scale. Vendors who once priced models on a cost-plus basis to recoup training investments now find themselves in a race to the bottom.

This commoditisation creates a peculiar dynamic. OpenAI is reportedly on track to lose $5 billion in 2025, whilst Anthropic projects losses exceeding $2.7 billion. These are not small companies struggling to gain traction – they are leaders in the field, burning through capital at extraordinary rates.

Strategic Entrenchment

Faced with commoditisation pressures, leading players are pursuing partnership strategies designed to entrench their positions within specific sectors and user segments. OpenAI's multi-year agreement with AMD to deploy 6 gigawatts of GPU capacity, government partnerships with AI laboratories, and embedding within broader service offerings all represent attempts to build defensible moats beyond mere model performance. The logic is clear: if the technology itself becomes commoditised, value must be captured through integration, distribution, and sector-specific applications.

This strategic shift has accelerated interest in smaller, specialised models addressing economically viable use cases. Rather than pursuing ever-larger general-purpose models, organisations are increasingly focused on domain-specific applications where AI can demonstrably justify its costs. Healthcare diagnostics, financial risk assessment, and supply chain optimisation represent areas where the value proposition is clearer and more immediate.

Data as the New Differentiator

Proprietary data has emerged as perhaps the most critical competitive asset, as foundation models converge in capability. Experts increasingly argue that companies controlling exclusive, high-quality datasets will define AI's future trajectory, not those merely developing algorithms. An AI model trained on a decade of proprietary customer service logs or unique scientific research delivers insights that generic models trained on public data simply cannot replicate.

This shift in value capture has profound implications. Data providers may increasingly dictate terms to model developers rather than the reverse. The economics reverse: instead of organisations paying model providers for access to cutting-edge algorithms, they may soon find model providers competing for access to valuable proprietary datasets. For many enterprises, this represents a more favourable positioning than being perpetual consumers of commoditised inference services.

The Infrastructure Dilemma

The technology giants are embarking upon unprecedented infrastructure spending. Microsoft, Amazon, Google, and Meta collectively plan to invest approximately $320 billion in AI infrastructure during 2025, with Goldman Sachs estimating total global AI-related infrastructure spending could reach $3-4 trillion by 2030. These figures dwarf any previous technology buildout in history.

A troubling question looms: will these investments generate commensurate returns? The disparity between capital expenditures and revenue growth has become increasingly apparent, with sales-to-capex ratios among major technology firms deteriorating sharply.

Several of the large players have begun to issue bonds in recent weeks to fund infrastructure investments; they've turned away from investing their own cash and are now spending other people's money. The current economics of foundation model development may be fundamentally unsustainable without radical shifts in either pricing structures or cost bases. As one analyst noted bluntly, "I find it hard to see how there can be a good return on investment given the current maths".

The Depreciation Dilemma

Uncertainty about depreciation timescales for the high end chips that power AI is compounding these concerns. Traditional enterprise servers remain relevant for three to five years, but AI-specific hardware faces a far more aggressive obsolescence cycle. Nvidia's rapid pace of innovation – with major architecture releases annually or more frequently – threatens to devalue existing hardware investments faster than anticipated. Nvidia's CEO remarked facetiously that when Blackwell GPUs became available, "you couldn't give Hoppers away" – a comment that, whilst exaggerated, captures the essence of the problem.

If premium chips depreciate more rapidly than the five-to-six-year schedules currently used by hyperscalers, the accounting implications could be severe. Conversely, defenders of current depreciation practices note that older GPUs retain value for inference workloads and remain economically viable for throughput-oriented tasks. The truth likely lies somewhere between these extremes, but the uncertainty itself represents another material risk for those in the sector.

Circular Flows and Concentrated Risk

Hundreds of billions in infrastructure financing now flow through circular arrangements where customers are suppliers, suppliers are investors, and dependencies intertwine. Nvidia invests in AI startups who then commit to purchasing Nvidia chips; hyperscalers provide cloud infrastructure to AI companies whilst simultaneously competing with them. Whilst such arrangements aren't inherently problematic – vendor financing exists across many industries – the scale and concentration in AI are unprecedented.

The concern is not merely circular revenue recognition, which most sophisticated investors can identify, but rather systemic risk. If growth expectations falter or a major player stumbles, the interconnectedness could amplify losses across the ecosystem. The entire structure depends upon continued strong demand for AI services and sustained belief in future revenue potential.

The ROI Problem

That revenue potential remains frustratingly elusive for many organisations. Only 51% of companies can confidently evaluate AI return on investment, and most organisations have yet to see meaningful returns from AI investments. This disconnect between spending and measurable value creation represents perhaps the most fundamental challenge facing the sector.

Many executives deploy and even accelerate AI investments not primarily because current projects are delivering strong returns, but rather because they fear falling behind competitors.

Talent and Energy Constraints

Two additional cost pressures compound these challenges.

First, AI talent demand exceeds supply by more than x3 globally, with over 1.6 million open positions and only 518,000 qualified candidates. This mismatch drives extraordinary salary inflation, with AI roles commanding 67% higher compensation than traditional software positions and 38% year-over-year growth. For organisations beyond the technology giants, such talent costs can quickly render AI initiatives economically unviable.

Second, energy consumption presents both immediate cost burdens and broader societal implications. Data centres consumed 4.4% of U.S. electricity in 2023, a figure that could triple by 2028. AI-specific operations are projected to consume 165-326 terawatt-hours annually by 2028 – enough to power 22% of American households. Beyond the direct costs, this consumption raises questions about grid capacity, carbon emissions, and the sustainability of continued AI expansion at current trajectories.

The DeepSeek Disruption

Into this environment of mounting costs and uncertain returns comes DeepSeek, whose claimed achievement of training competitive models for $6 million rather than $80-100 million represents a potential paradigm shift. Through architectural innovations including mixture-of-experts approaches and optimised distillation techniques, DeepSeek has demonstrated that step-function improvements in cost-effectiveness remain possible. If validated and replicated, such breakthroughs could reshape competitive dynamics, rendering obsolete massive capital investments predicated on different cost assumptions.

The emergence of dramatically more efficient approaches highlights a critical uncertainty: today's infrastructure buildouts may be optimised for techniques that prove transitory. Capital deployed today could find itself stranded by tomorrow's innovations.

Regulatory Fog

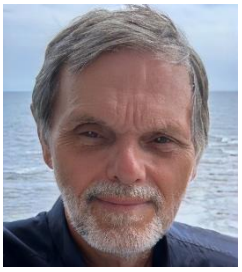
Finally, regulatory uncertainty and complication pervades the sector. The EU AI Act imposes significant costly and complex compliance requirements with severe penalties. In the United States, there are campaigners at state level arguing for greater safeguards, and lobbying by Silicon Valley has pushed Trump’s administration towards outlawing these at federal level. Organisations operating internationally must navigate fragmented and shifting requirements with different compliance obligations, liability risks, and enforcement mechanisms.

For many firms, regulatory uncertainty ranks as a top policy concern, creating hesitation about deployment strategies and difficulty forecasting compliance costs. The challenge is not merely keeping pace with regulations—it is planning investments when the rules themselves remain in flux.

An Uncertain Equilibrium

What becomes clear from examining AI's economic landscape is that technical capability and commercial viability are separating. The technology continues to advance, but the business models supporting its development face mounting strain. Enormous capital commitments proceed alongside troubling questions about depreciation, returns, and systemic risk. Costs for talent and energy escalate whilst most organisations struggle to demonstrate clear value from their AI investments.

The sector may yet resolve these tensions – through new monetisation models, dramatic efficiency improvements, or applications that finally deliver transformative returns at scale. But the era when scaling alone guaranteed progress has ended. What follows will be determined not by algorithms, but by economics.



Peter G. Osborn
Peter.Osborn@PlannedData.com
+44 (0)7802-666758
https://www.PlannedData.com