ORIGINALLY PUBLISHED FEBRUARY 11, 2025

AVITY Investment Management Inc.

FEBRUARY 11, 2025

INSIGHTS

GenAI: The Path Forward
Part 2 – AI ROI ("*ROAI*")

EXECUTIVE SUMMARY

The emergence of generative artificial intelligence (GenAI) has catalyzed a surge of capital expenditures (capex) directed at building out AI infrastructure. The "Big Four" hyperscalers (Meta, Amazon, Microsoft and Alphabet) are expected to spend over $300 billion on capex in 2025, up from $225 billion in 2024. A significant portion of this will be directed to AI-related equipment, including semiconductors, data centers and the tools that power them. Some critical debates have emerged as investors attempt to forecast the industry landscape. In our view, the two most important factors that will shape the AI sector are (1) the ability of scaling laws to continue to advance large language models, detailed in Part 1 of this white paper series, and (2) the return on investment (ROI) from AI-related spending ("ROAI"), which will be the focus of this note. It remains very early in the adoption curve for AI, but there is encouraging evidence that AI solutions can significantly boost productivity. The recent news from China's DeepSeek strengthens our hypothesis that as models improve and costs decline, AI use cases will proliferate and deliver strong ROAI across the economy.

WHY FOCUS ON ROAI? HISTORICAL CONTEXT FOR TECHNOLOGICAL REVOLUTIONS

There is historical precedent for new technologies to be accompanied by periods of exuberance as investors extrapolate growth trends far into the future. The railroad mania of the 19th century and dot-com bubble of the late 20th serve as cautionary tales. In both instances, it was easy to envision the potential of the new technologies, and an arms race ensued to lay railroad track and fiber optic cables. While railroads and the internet would ultimately prove substantial industries, both sectors were oversupplied for a period of time. It took several years for widespread adoption to eventually fill the dormant railroad and fiber optic capacity. It's often been said that the original dotcom investors were right, just a decade or two early. Nonetheless, the timing mismatch between supply and demand can be painful for investors as financial market valuations adjust from a period of euphoria to one of disillusionment.

> "I think bubbles are interesting because a lot of the bubbles ended up being things that were very valuable over time and it's just more a question of timing… even the dot com bubble… there's all this fiber laid and it ended up being super valuable, but it just wasn't as valuable as quickly as people thought… It's hard to predict what's going to happen in the next few years. I think AI is going to be very fundamental. I think there's a meaningful chance that a lot of the companies are overbuilding now… but on the flip side, I actually think all the companies that are investing are making a rational decision because the downside of being behind is that you're out of position for the most important technology for the next 10 to 15 years." – *Mark Zuckerberg, Bloomberg Interview*

The familiar pattern that many technological revolutions have followed has been best described by academic and author Carlota Perez. In her influential book "Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages," Perez describes five recurring phases of each technological revolution: the *eruption phase*, when new products unveil their potential; the *frenzy phase*, when financial capital (and speculation) drives an intense expansion of infrastructure to support the new technologies; the *turning point*, when imbalances between potential production and lack of existing demand result in a bubble collapse; the *synergy phase*, or "golden age", a time when production can grow into

the installed infrastructure and drive productivity and prosperity; and lastly, the *maturity phase*, when new markets reach saturation.[1]

On more than one occasion over the last few rounds of quarterly earnings calls big tech CEOs have expressed the notion that underinvesting is a greater risk than overinvesting, giving some the impression that capex is being deployed due to "fear of missing out" (FOMO) with few clear monetization strategies.

> "I think the one way I think about it is when we go through a curve like this, the risk of underinvesting is dramatically greater than the risk of overinvesting for us here, even in scenarios where if it turns out that we are overinvesting… these are infrastructure which are widely useful for us, they have long useful lives… But I think not investing to be at the front here, I think, definitely has much more significant downside." – *Sundar Pichai, CEO Alphabet*

Part of the motivation behind these aggressive investments is the notion that first mover advantage in AI may lead to significant competitive advantages. Some believe the achievement of artificial general intelligence (AGI), a level of AI that is comparable to human-level intelligence, will be able to upend competitive dynamics across every industry. Determined to avoid being unseated from their thrones, incumbents (e.g., big tech) commenced an arms race to stay on the cutting edge of AI.

The rapid pace of investment has been scrutinized by those who believe the ROAI does not justify the spending. Indeed, quantifying the benefit of AI investments is challenging. Some 41% of organizations struggle to define and measure the exact impact of their investments (Deloitte), and 85% of business leaders don't believe they have the tools to appropriately measure ROI (KPMG). The lack of visible near-term ROAI and dearth of "killer apps" begs the question: will AI actually be as revolutionary as many believe? We believe the answer is yes, but timing AI's impact is more difficult to discern.

THE CHALLENGE OF VALUING INNOVATION

The cadence and magnitude of AI returns is critical to the investment case for AI-exposed businesses. In the case of AI (and the internet, and railroad), massive upfront costs are required to deploy infrastructure. Enthusiasm for new technology can lead to a period of overbuilding.[2] Meanwhile, revenue recognition is not always immediate. Before AI can start being broadly adopted, AI applications must find product-market fit. To-date, few obvious killer apps have emerged. However, use cases are beginning to surface, particularly productivity tools, and it is easy to visualize the enormous economic value that could be created in the next 5-10 years with the proliferation of "agentic AI" (AI agents that can execute tasks autonomously), domain-specific AI search engines (which could have specialization in different verticals such as medicine, law or finance) and low cost, ubiquitous content creation tools, to name a few. Even if these use cases aren't monetized for many years, the net present value of AI projects may still appear attractive in a discounted cash flow analysis given the size of the ultimate reward. Discounted cash flow analysis is the foundation for valuing any productive asset. Discounted cash flow dynamics explain why the long-term optimism around new technologies can support asset prices despite lack of near-term returns. However, since discounted cash flow models are highly sensitive to their various inputs (e.g., projected growth, interest rates and profit margins), prices for long duration assets can experience significant fluctuations as these variables change over time.

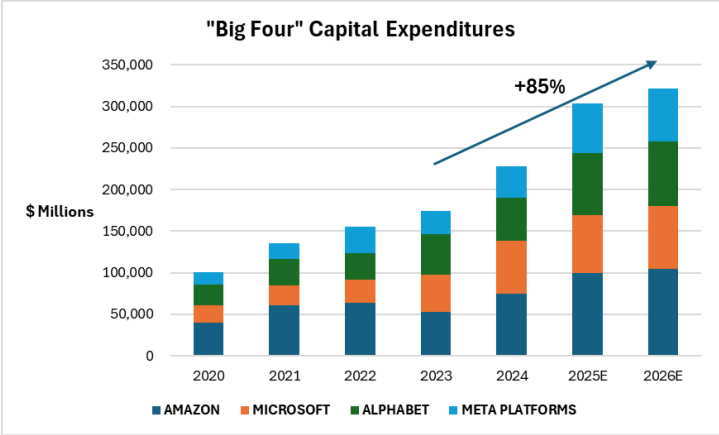
[1] It should be noted that bubbles, while treacherous for investors, are typically beneficial for innovation. In their new book "Bubbles and the End of Stagnation", Byrne Hobart and Tobias Huber argue that "financial bubbles provide the capital necessary to fund disruptive technologies at the frontier of innovation and accelerate breakthroughs in science, technology and engineering." This idea is complementary to the work of Carlota Perez.

[2] Capacity constraints for sought after technologies, such as GPUs, often incentivize "double ordering" as customers seek to acquire as much of a scarce asset as they can get. This can exacerbate supply chain volatility for new technologies by leading to false demand signals.

AI INFERENCE – WHERE THE RUBBER MEETS THE ROAD

Dozens of infrastructure companies are monetizing AI today. Infrastructure companies act as the "picks and shovels" vendors that sell the tools that enable the construction and use of LLMs. The most obvious example of AI picks and shovels players are semiconductor companies. With any new computing paradigm, semiconductors are the first to benefit, as chips must be deployed into the real world before applications can be built on top them. We are seeing evidence of this today from companies like NVIDIA, whose GPUs are used to train and run inference on LLMs, as well as Broadcom and Marvell, who are assisting other big tech companies in building custom silicon. The buildout of AI infrastructure is also benefitting the hardware ecosystem that surrounds semiconductors within a data center, including networking, cooling and electricity equipment. Cloud Service Providers (CSPs), the primary buyers of AI-semiconductors, can also be considered infrastructure providers. Amazon Web Service (AWS), Microsoft Azure and Google Compute Platform (GCP) are monetizing AI as customers consume the compute that these platforms provide through access to their data centers. (These CSPs are simultaneously leveraging AI to enhance their own operations and products.)

Importantly, the success of infrastructure companies in monetizing AI does not necessarily translate to use cases that are being adopted and creating value in the real economy. A large portion of AI spend has been earmarked for LLM training, which to at least some degree can be considered R&D (i.e., experimental). Inferencing, which refers to a trained model being fed a query and producing a response, is a better representation of end-users actually consuming AI solutions.



Source: Bloomberg

Inference trends appear to be growing rapidly, though this level of detail is not easily measured or reported in company filings. NVIDIA has stated that approximately 40% of its rapidly growing data center revenue is for AI inference, with some significant portion of that being attributed to recommender engines that fuel digital advertising and e-commerce. Weekly average users on OpenAI's ChatGPT has swelled to over 300 million, up from essentially zero prior to its launch just over two years ago. Google search usage is increasing as consumers become habituated to its Gemini powered AI overviews. Satya Nadella, CEO of Microsoft, recently stated that Microsoft's Copilot is experiencing accelerated adoption and suggested a significant portion of Azure AI consumption is inference.

> "This is going to be the fastest growth to $10 billion of any business in our history. It's all inference… we're not actually selling raw GPUs for other people to train on. In fact, that's sort of a business we turn away because we have so much demand on inference… we are not even participating in most of that because we are literally going to the real demand, which is in the enterprise space of our own products like GitHub Copilot or M365 Copilot. So I feel the quality of our revenue is also pretty superior in that context." – *Satya Nadella, CEO Microsoft, Q1 2025 Earnings Call*

Cloudflare, which runs a developer platform on its global network of data centers, has witnessed a material shift from AI training towards inference.

> "As we mentioned last quarter, the accelerating shift from AI training to AI inference has given us confidence to continue to increase our investment in our GPU rollout as we provision greater capacity to support demand in 2025." – *Matthew Prince, CEO Cloudflare, Q4 2024 Earnings Call*

Sundar Pichai, CEO of Alphabet, noted that the mix of spend dedicated to inferencing has been rising in recent years.

> "If you look at the trajectory over the past three years, the proportion of the spend towards inference compared to training has been increasing, which is good because obviously, inference is to support businesses with good ROIC."
> *– Sundar Pichai, CEO Alphabet, Q4 2024 Earnings Call*

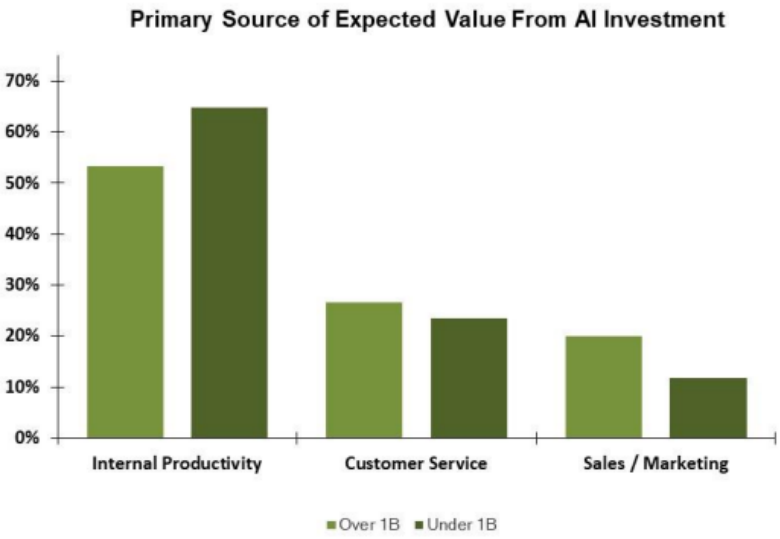
Andy Jassy, Amazon's CEO, presented a forward-looking view that included inference as a core building block of every future application.

> "It may be hard for some to fathom a world where virtually every app has generative AI infusing it with inference being a core building block just like compute, storage and database, and most companies having their own agents that accomplish various tasks and interact with one another. This is the world we're thinking about all the time…"
> *– Andy Jassy, CEO Amazon, Q4 2024 Earnings Call*

Recent inference trends indicate AI adoption is underway, suggesting end-users are finding value in these tools just as myriad productivity solutions are expected to hit the market in 2025.

PRODUCTIVITY – THE KILLER APP

Sequoia published a popular article in September 2023 (and a follow up published in June 2024) that theorized that there was a significant gap between the cost of the AI infrastructure buildout and the revenue generated from the technology. One challenge with quantifying ROAI in this fashion is that many of the initial use cases for AI have been harnessed by companies internally to wring productivity benefits. These productivity initiatives can take the form of accelerating software development, automating IT workflows, supplementing customer service teams with chatbots and improving one's cyber security posture. Many of these upgrades simultaneously enhance the employee experience and reduce employee turnover, a nontrivial cost savings. The ROAI from internal efficiency projects is harder for industry analysts to measure compared to revenue generating applications that are visible to the outside world. Nonetheless, there is a growing body of evidence that AI can deliver meaningful productivity benefits.



Source: Evercore ISI Research Survey

Industry analyst surveys indicate promising ROAI results from a broad swath of organizations. A recent Deloitte report showed 67% of organizations are increasing investment in AI given strong value delivered to-date, with improved productivity cited as the most common benefit achieved with AI. An IDC study surveyed over 3,000 organizations globally and found that companies are realizing an average 3.7x return on every $1 invested in AI. Furthermore, organizations classified as "AI leaders" are realizing an average return of 10.3x. Forbes found that 74% of AI early adopters are realizing ROAI and KPMG reported 78% of executives feeling confidence in their ROAI. A recent National Research Group study, which surveyed 2,500 executive level business leaders, found that 74% of enterprises using AI are generating ROAI on at least one use case. A Forrester Consulting study found that small and medium-sized businesses that employ Microsoft 365 Copilot can experience ROIs ranging from 132% to 353% over a three-year period.[3] Github, a software developer platform

[3] Priced at $30 per month, Copilot represents a meaningful revenue driver for Microsoft given its nearly 400 million users.

owned by Microsoft, estimates that Copilot allows coders to complete tasks 55% faster. Many of these studies reported higher employee satisfaction as a potent intangible benefit.

CASE STUDY ON INTERNAL EFFICIENCIES: AMAZON

Amazon is driving internal efficiencies across its entire business by leveraging AI. The company has been using transformer technology, the foundation of LLMs, for years to forecast which products customers will purchase. Amazon trains its machine learning systems on synthetic data to simulate edge cases, such as the peak holiday season. The company is also leveraging AI to optimize delivery routes, automate fulfillment operations with robotics ("physical AI") and augment customer service channels. AI powered chatbots on Amazon.com have improved customer satisfaction scores. More than 500,000 independent sellers globally are already using AI tools to generate product listings on Amazon.com. AI is also enhancing the company's internal software development. CEO Andy Jassy estimated that the company's "Q" assistant helped developers migrate 30,000 product applications to Java 17, saving 4,500 developer years and $260 million annually. Furthermore, Amazon developers shipped 79% of auto-generated code reviews without making any additional changes. Amazon has built or is in the process of building approximately 1,000 GenAI applications.

AGENTIC AI – IS 2025 THE YEAR OF THE AGENT?

It has been argued that for AI to begin delivering meaningful ROAI, chatbots must graduate to autonomous agents working in real-world domains. Agentic AI describes autonomous AI systems that can operate with minimal human oversight, learn and adapt from interactions, collaborate with other AI agents, and execute tasks. Whereas chatbots are mostly general purpose tools, AI agents can be built for specific workflows, making them more purposeful with a measurable ROAI. Gartner estimates that by 2028, 33% of enterprise software applications will include agentic AI, up from less than 1% today, and these agents will autonomously execute 15% of day-to-day decisions.

> "The next step is that these copilots will become pilots, autonomous agents, and the impact of that will be considerable. There's always a long tail of people using a technology for things you don't expect. LLMs are a bicycle for the mind that will unlock creativity for things we haven't even conceived of yet." – *Vivek Raghunathan, SVP of Engineering, Snowflake*

Indeed, enterprise software companies are hurrying to deliver agentic tools to their customers. Salesforce recently announced the newest version of its Agentforce, the agentic layer of Salesforce's platform which the company describes as a digital labor platform for enterprises. Microsoft has announced agentic features alongside Copilot. (In a recent Morgan Stanley Alphawise survey, 66% of Chief Information Officers cited Microsoft as the vendor they are most likely to leverage for their AI agent strategies.) ServiceNow, a software platform that allows large enterprises to automate workflows, has integrated agentic features into a premium priced SKU with an average pricing uplift of 30%. Customers are seeing the value, as ServiceNow's Pro Plus SKU has been the fastest adopted product in the company's history and recently posted growth of 150% quarter-over-quarter.

> *Karl Keirstead, UBS Software Analyst (2024 UBS Global Technology Conference):* When I talk to big enterprises about where they're pointing AI, like what the use cases are, I've been struck by the number that describe pointing AI at some internal function or internal business process that they're trying to automate with AI as opposed to pointing AI externally. So it feels to me like the bulk of activity is frankly in internal business workflow automation. And that's what ServiceNow does. So you're just in the right end market where a lot of the activity is occurring.
>
> *Gina Mastantuono, ServiceNow CFO:* I think your hypothesis is exactly right. Where folks are focusing AI right now initially is all about that business process automation, that business process, efficiency, cost savings… the conversations that we're having with all of our customers are 100% initially on that efficiency, productivity, and cost savings.

OpenAI has recently announced agentic features in its flagship ChatGPT tool. OpenAI's Operator uses its own browser to type, click and scroll webpages on behalf of its user. These are features that Anthropic, the Amazon-backed AI startup, has already launched. OpenAI's latest feature, dubbed Deep Research, uses a reasoning process to complete multi-step research tasks. Deep Research queries can take between 5 to 30 minutes to complete, "accomplishing in tens of minutes what would take a human many hours." For $200 per month, any worker can add a research assistant to their staff in the form of Deep Research.

Deep Research displayed a leading degree of accuracy on a recent evaluation that tested AI models across a range of expert-level questions.

Source: OpenAI

Model	Accuracy (%)
GPT-4o	3.3
Grok-2	3.8
Claude 3.5 Sonnet	4.3
Gemini Thinking	6.2
OpenAI o1	**9.1**
DeepSeek-R1*	9.4
OpenAI o3-mini (medium)*	10.5
OpenAI o3-mini (high)*	13.0
OpenAI deep research**	**26.6**

* Model is not multi-modal, evaluated on text-only subset.
**with browsing + python tools

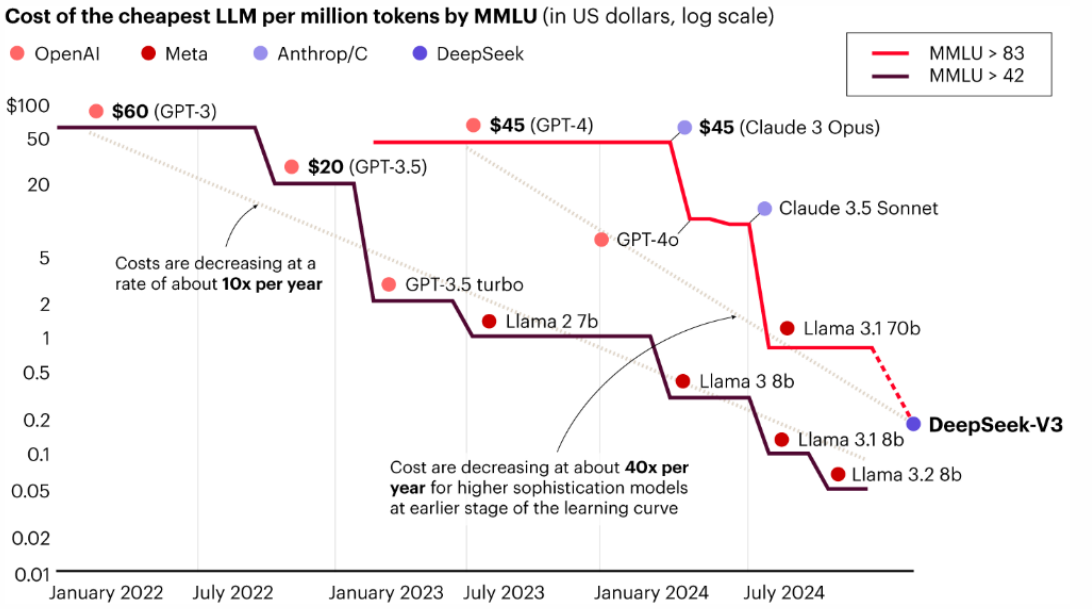
Productivity tools are likely to be the preliminary use cases for AI and AI agents appear well positioned to be the mechanism through which efficiencies are harnessed. While more "killer apps" will materialize over time, it is likely that productivity solutions are just that. The McKinsey Global Institute estimates that AI could add $2.6 to $4.4 trillion annually across 63 use cases analyzed, with cost efficiencies driving the majority of the benefit. These productivity tools are emerging at the perfect time as labor supply has been scarce in recent years and the working age population in the U.S. likely to continue contracting as baby boomers retire.

> "What an LLM does is itself the killer app… if you ask me what was the original killer application of spreadsheets, maybe it was finance. But that's maybe 10% of use cases throughout the enterprise. The killer app for the spreadsheet was the spreadsheet. It was a new way for people to be productive, and I guarantee you that there are industries now where you couldn't live without spreadsheets. Similarly, the killer app of a LLM is the LLM itself. It is a new way for people to be productive, and people are discovering new use cases of it every day." – *Vivek Raghunathan, SVP of Engineering, Snowflake*

COST CURVES AND COMMODITIZATION

The recent announcement of China's DeepSeek's open-source AI models sent shockwaves through markets as investors grappled with the implications of lower cost AI for the hardware ecosystem. (China emerging as a formidable AI competitor, despite significant export restrictions, was likely an additional source of market jitters.) DeepSeek used a process called distillation to train smaller "student" models by drawing on the responses from larger, pre-trained "teacher" models, like ChatGPT. Additionally, DeepSeek utilized a "Mixture of Experts (MOE)" architecture, in which the model activates only small fraction of relevant parameters for a given query, rather than activating every parameter. This approach results in a more efficient use of resources and significantly reduces inference costs. The DeepSeek news warrants certain caveats. First, it is very likely that the total cost of DeepSeek's R1 model exceeded the $6 million reported cost of its final training run. Additionally, the cost declines demonstrated by DeepSeek are not altogether surprising as they track with broader trends in

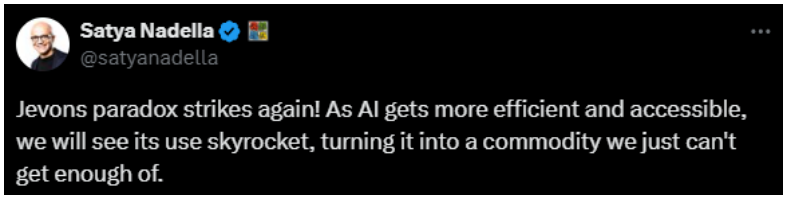
AI efficiency. Nonetheless, DeepSeek's innovation shouldn't be underappreciated and neither of these exceptions impact our forward looking thesis.



DeepSeek's cost performance follows the trend of falling AI inference costs across the industry.

Source: a16z, Bain analysis

We believe DeepSeek has two important implications for AI. First, capital and scale do not appear to be durable moats for AI model builders. This does not necessarily mean that leading AI labs won't continue to push frontier models forward, but it does suggest that models will trend towards commoditization over time. To the extent AI models become commoditized, technology companies that provide AI solutions (e.g., software and internet companies) are very well positioned as they will be able to sell inference at a lower cost while not being captive to a monopolistic AI model, which would otherwise have the leverage to extract high rents from the ecosystem. The commoditization of AI models and cheaper inference provides a clearer path to ROAI for many AI solutions.

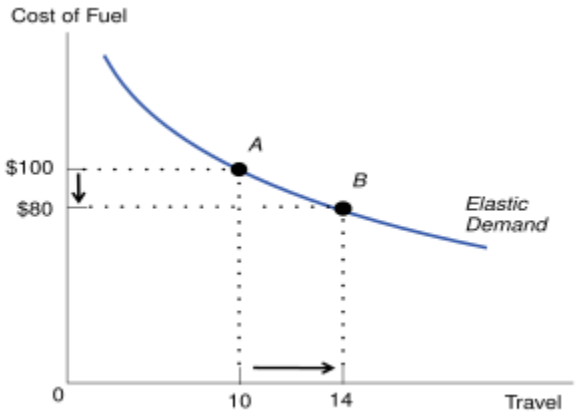


The model efficiency demonstrated by DeepSeek also has implications for AI adoption. Jevons Paradox, which has become cliché among the tech investor over the last three weeks, describes the tendency for total consumption to increase as the price of a resource drops. There is plenty of historical precedent to support Jevons Paradox. The semiconductor industry has grown from less than $1 billion to nearly $700 billion in the last 60 years even as the price of transistors has dropped by several thousand-fold. Public cloud infrastructure experienced similar commoditization debates just a decade ago, as the price of infrastructure-as-a-service (IaaS) plummeted, and yet Gartner estimates that total IaaS spending continues to grow in the mid-teens percent.

"We did the same thing in the cloud when we launched AWS in 2006, where we offered S3 object storage for $0.15 a gigabyte and compute for $0.10 an hour, which of course is much lower now many years later. People thought that people would spend a lot less money on infrastructure technology. What happens is companies will spend a lot less per unit of infrastructure and that is very, very useful for their businesses, but they then get excited about what else they could build that they always thought was cost-prohibitive before and they usually end up spending a lot more in total on technology once you make the per unit cost less." – *Andy Jassy, CEO Amazon, Q4 2024 Earnings Call*

A hypothetical example of elastic demand: A 20% increase in fuel efficiency results in a 40% increase in travel. Fuel consumption increases even as Jevons paradox occurs.

Source: Wikipedia, "Jevons Paradox"



The implications of cheaper and commoditized AI are a net positive for the ecosystem, as inference costs rapidly decline and unlock new use cases that can deliver attractive ROAI. Semiconductors stand to benefit from increasing volume of AI inference, while more value will begin to accrue to the application layer.

"As these models are being commoditized at a rapid rate and probably more rapid than anyone could have dreamed of, it's super, super exciting for a platform and app vendors like us, since the competitive differentiation will happen at our level in terms of the applications, the business processes, how you can help those companies run better and then ultimately get a business outcome." – *Bill McDermott, CEO ServiceNow, Q4 2024 Earnings Call*

CONCLUSION

For the AI field to advance from the experimentation stage to mass adoption, there is a growing need for end-users to demonstrate economic value from the technology. The inability for AI to find product-market fit and generate return on investment will have negative implications for the sector. However, tech-savvy firms have demonstrated strong evidence that AI can deliver on many of its anticipated productivity benefits. It is now incumbent upon application vendors to package these tools in a way to optimize end user ROAI and adoption, while enterprises modernize their data estates to be more compatible with the technology.

The adage that we tend to overestimate the impact of technology in the short-term and underestimate it in the long-term appears especially relevant for the emerging space of artificial intelligence. Every technological revolution runs the risk of overhype in the near-term but that does not mean long-term optimism is misguided. It is normal for the asset prices of new technologies to experience bouts of "price discovery" (volatility) as investors debate the outlook. Even as we interpret lower cost, commoditized models as a long-term positive for the industry, the market reaction to DeepSeek is a reminder that volatility should be expected when key assumptions change for long duration assets. Investors with a long-time horizon should be focused on ROAI and its implications, including the potential to reshape competitive dynamics across industries.