Institute for
New Economic Thinking

Working Papers
ineteconomics.org/workingpapers

The U.S. Is Betting the Economy on 'Scaling' AI: Where Is the Intelligence When One Needs It?

Servaas Storm[*]

Working Paper No. 244

December 1st, 2025

ABSTRACT

The AI industry is betting that 'scaling', *i.e.*, adding more and more data, GPUs, compute infrastructure and dollars, will lead to machine superintelligence or Artificial General Intelligence (AGI) — which in turn will lead to exponential growth of output, productivity and profits for the industry and the larger American economy. Focusing on AGI and generic LLMs, the point of this paper is plain: AI's 'scaling' strategy must fail and the AI data-center investment bubble will pop. The paper identifies four bottlenecks: (1) the planned $5 trillion investment in data center infrastructure (during 2026-2030) is not going to pay off; AI revenues will not increase enough and AI inference cost continue to rise faster than revenues; (2) AI firms will have to resort to *hyper-scale borrowing* from banks and investment-grade bond markets to fund their capex; this hyperscale borrowing will create a ticking time bomb on the balance sheets of AI firms, because the core capital expenditure on specialized GPUs and server risks becoming economically obsolete within two or three years; (3) it will be impossible to build the projected data center infrastructure

[*] [*] Department Economics of Technology and Innovation, Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands. S.T.H.Storm@tudelft.nl

The author is very grateful to Thomas Ferguson for critical comments and valuable suggestions on the paper. The author accepts the blame for whatever deficiencies remain.

fast enough, because upstream suppliers — producing everything from copper wire to turbines to transformers and switchgear — will run into labor shortages, long waiting times for power grid connections, material bottlenecks and regulatory blowback; and (4) the strategic bet of frontier AI firms that AGI can be achieved by building ever more data centers and using ever more chips is already going bad; AI products will continue to be untrustworthy for high-stake usage. As a result, the magical projections of exponential growth, which defy economic and financial logic and fatally ignore unforgiving real-world constraints will turn out to be wrong. The fact that the AI industry is the main source of growth in an otherwise sclerotic U.S. economy and is driven by a concentrated set of hyper-scalers engaging in 'circular' financial transactions based on aggressively optimistic long-term cash flow-generating potential should be a very serious cause for concern.

https://doi.org/10.36687/inetwp244

JEL codes: E24, F52, G01, O30, O33.

Keywords: Artificial intelligence; AGI; AI bubble; LLMs; circular financing; revenues; price-earnings ratio; leverage; scaling; inference cost; hallucinations; Chinese competition; AI-generated work-slop; misinformation; FOMO; price war.

> “If there must be madness, something may be said of having it on a heroic scale”
> — John Kenneth Galbraith (*The Great Crash 1929*)
>
> “The propensities to swindle and to be swindled run parallel to the propensity to speculate during a boom.”
> — Robert Z. Aliber, Charles P. Kindleberger & Robert N. McCauley (*Manias, Panics, and Crashes*)

This paper reviews four major factors that will (rather sooner than later) pop the irrational enthusiasm around the Artificial Intelligence (AI) infrastructure spending spree in the U.S.[1] The AI industry is betting the U.S. economy on its strategy of ‘scaling’, *i.e.*, adding more and more data, GPUs, compute infrastructure and dollars, in the belief that this will lead to machine superintelligence or Artificial General Intelligence (AGI) — and, of course, that the capital expenditures can be recouped, with a good rate of return, because the superintelligent Large Language Models (LLMs)[2] will raise productivity, speed up technological progress and innovation, generating exponential growth of output as well as profits for their owners and the larger American economy.

My main point is simple: the ‘scaling’ strategy to achieve AGI is already showing significant diminishing returns, because more data and more GPUs cannot fundamentally improve the performance of LLMs, after a point. The algorithms are not constructed on proper and robust world models, but instead are built to autocomplete, based on sophisticated pattern-matching. As a result, AGI will remain a dream and the magical projections of exponential growth, which defy economic and financial logic and fatally ignore unforgiving physical (analogue) constraints imposed by the electricity grid, the construction industry, and the shortage of (skilled) workers who are needed to help build the additional data centercenter infrastructure, will turn out to be wrong.

[1] Our focus in this paper is on the massive investments in U.S. data-center computing infrastructure for generic LLMs. It is a follow up to and update on my earlier analysis of the hype around the U.S. AI industry; see Storm (2025). Here, we zoom in on the unsustainable circularity of the AI industry growth model (based on generic LLMs) and on key macroeconomic and physical constraints that will block AI’s future growth.

[2] LLMs are, in essence, giant plagiarism machines, that are used by a few major well-placed players to accumulate significant wealth by illegally extracting value from other people’s creative work, personal data, or laborlabor (Bender and Hanna 2025).

However, the generic LLMs are different from specialized, domain-specific AI tools that are already being used to great effect in many scientific disciplines, such as protein science,[3] code generation, and pharmaceutical research. Targeted machine-learning systems can manage load on the electricity grid more effectively and help reduce carbon emissions in trucking, shipping, steelmaking and mining industries. Each use case should be scrutinized for its utility, the biases and risks it might introduce, and its propensity to destroy jobs that depend on human judgment. However, these domain-specific tools do not require the massive capital expenditures on GPUs and data-centercenter infrastructure that are needed for the LLMs; hence, investments in customized AI tools, trained on datasets specific to a particular domain, such as legal documents, financial reports, or technical literature, will likely pay-off, unlike the megalomanic expenditures on training, inference, and compute for the generic LLMs. This paper focuses on the latter.

The fact that the AI industry is not only the principal source of growth in an otherwise sclerotic U.S. economy (Storm 2025), but is also driven by a concentrated set of hyper-scalers engaging in 'circular' transactions based on aggressively optimistic long-term cash flow-generating potential should be cause for worry (see Arun 2025). We now laugh at the utter foolishness of some of the richest merchants in Amsterdam who at the height of Tulip Mania in February 1637 were buying exotic tulip bulbs at ridiculous prices of around 5,000 Dutch guilders per pound (Goldgar 2007) — approximately two million Euros today, equal to the cost of an upscale Amsterdam canal house.[4] But rest assured: we will be the butt of the jokes of future historians — for not recognizing the current AI mania for what it is: a bubble.[5]

[3] In research, machine learning has revolutionized protein structure prediction and design, facilitating medicine development (Saplakoglu (2024) offers a discussion of the uses and limits of AI tools in protein folding research).

[4] Most tulips were far cheaper. Economic historian Anne Goldgar (2018) found only 37 people who spent more than 300 guilders on bulbs, around the yearly wage of a master craftsman.

[5] The signs of a bubble are plain. In the first six months of 2025, AI start-ups that have no profits, no sales, no pitch and no product to speak of, have been securing billions of dollars of funding (Moore 2025). For example, pre-revenue, pre-product AI company 'Safe Superintelligence', founded by Ilya Sutskever, ex-chief scientist at OpenAI, raised $2 billion at a $32 billion valuation in April 2025 (Prabhu 2025). Similarly, 'Thinking Machines Lab', an AI research and product company launched by OpenAI's former chief technology officer Mira Murati, raised $2 billion at a valuation of $12 billion from investors such as Nvidia, AMD and Cisco in July 2025 (McPhee 2025). The company has not released a product, has no customers and has refused to tell investors what they're even trying to build. "It was the most absurd pitch meeting," one investor who met with Murati said. "She was like, 'So we're doing an AI company with the best AI people, but we can't answer any questions" (Levine 2025).

To be clear, cracks are showing and worries of an AI bubble are intensifying (e.g., Schmidt and Xu 2025; Thornhill 2025), and rapidly so. In recent weeks, several hedge funds trimmed their stakes in some of the largest seven tech firms (the so-called 'Magnificent Seven') (Sen 2025), especially after warnings from the CEOs of Goldman Sachs[6] and Morgan Stanley (Saini and Nishant 2025). More than 50 per cent of the fund managers surveyed by Bank of America (during November 2025), who between them manage around $500 billion in financial assets, said that AI stocks were already in a bubble; a net 20 per cent of these fund managers stated that corporations were spending too much on their data centercenter investments (Smith *et al*. 2025). In similar news, tech billionaire Peter Thiel's hedge fund sold off its entire stake in Nvidia worth around $100 million in September 2025 (Reuters 2025). And when Meta reported record revenue on October 29 (2025), its share price surprisingly plummeted by 11%; the reason: Meta CEO Mark Zuckerberg disclosed that he will "aggressively" increase capital spending on AI, drawing questions from analysts about how Meta plans to actually make money off the new technology.

The AI bubble will eventually pop because:

1. There is no world in which the enormous spending in data centercenter infrastructure (more than $5 trillion in the next five years) is going to pay off; the AI-revenue projections are pie-in-the-sky, as customers are unlikely to pay (enough) for the rather modest services offered by the LLMs and given the eventual oversupply of LLM services. At the same time, due to scaling, inference costs continue to rise — currently at a rate faster than that of revenues. Hence, most AI companies will realistically fail to turn a profit, as prices will fall (because of Chinese competition), while costs go up.
2. There is no way in which the AI industry can fund its capital expenditures out of revenues from paid subscribers or money from sovereign wealth funds. Hence, AI firms will have to resort to *hyperscale borrowing* from banks and investment-grade bond markets to fund their capex, laying the foundations for the next debt crisis. This hyperscale borrowing will create a ticking time bomb on the balance sheets of AI firms, because the core CapEx spending is on specialized GPUs and servers, which — because of unrelenting technological progress — risk becoming economically obsolete within two or three years.

[6] At the same time, the fear of missing out is real. As reported by Matt Wirz and Peter Ridegeair (2025), for the *Wall Street Journal* (November 16), days after Goldman Sachs CEO David Solomon voiced his concerns to analysts, Goldman formed a new team in its banking and markets group focused on AI infrastructure financing.

3. It will be impossible to build the projected data centercenter infrastructure in the next five years or so (which is the horizon of most AI investors). The lead time necessary to build a hyperscale data centercenter is currently around 2 years, but expect it to become much longer, say 7 or more years. Why? Upstream suppliers to the growth in data centercenters — the established industrial companies producing everything from copper wire to turbines to transformers and switchgear — have to expand production. These upstream suppliers will run into labor shortages, long waiting times for power grid connections, material bottlenecks and regulatory blowback – and all this will lengthen the lead times necessary to build a hyperscale data center.
4. The strategic bet of leading AI firms that Generative AI can be achieved by building ever more data centers and using ever more chips is already going bad. This scaling strategy is already exhibiting diminishing returns. It is the wrong strategy, since generic LLMs are not constructed on proper and robust world models, but instead are built to autocomplete, based on sophisticated pattern-matching.[7] LLMs will continue to make errors and hallucinate, especially when used outside their training data. Generic AI products are never going to actually work right and will continue to be untrustworthy.

The exuberance, and the fear of missing out (FOMO), in Silicon Valley and on Wall Street is, therefore, deeply irrational — and will prove to be socially costly to the larger U.S. economy, not just because of the inevitable correction, crash and recession (once the bubble pops), but more fundamentally, in terms of the scarce resources that have been and will be wasted on the hallucinogenic pipe dreams of a few entitled Ayn Randian billionaire tech brothers and sisters, who, quite in character, have already begun to hedge their bets by begging the taxpayer for subsidies and government loan guarantees (Cooper 2025).[8]

Once the AI market blows up, the blast radius will be wide, hitting not only Wall Street firms and the stock markets, but also pensions, mutual and exchange-traded funds, the energy and

[7] A June 2025 paper by researchers at Apple, titled "The Illusion of Thinking", demonstrates that the frontier LLMs, by their design, are incapable of reasoning, "over-think" easy problems, waste computational power on "wrong answers" (after finding the "correct" one), and completely collapse when the complexity of the issue under consideration increases and/or they have to generalize outside the space of their training data (Shojaee *et al*. 2025). To illustrate the point: the Apple researchers find that the largest billion-dollar generative AI system cannot solve the "Tower of Hanoi" puzzle, even when it is given the solution algorithm and all it has to do is follow the steps.

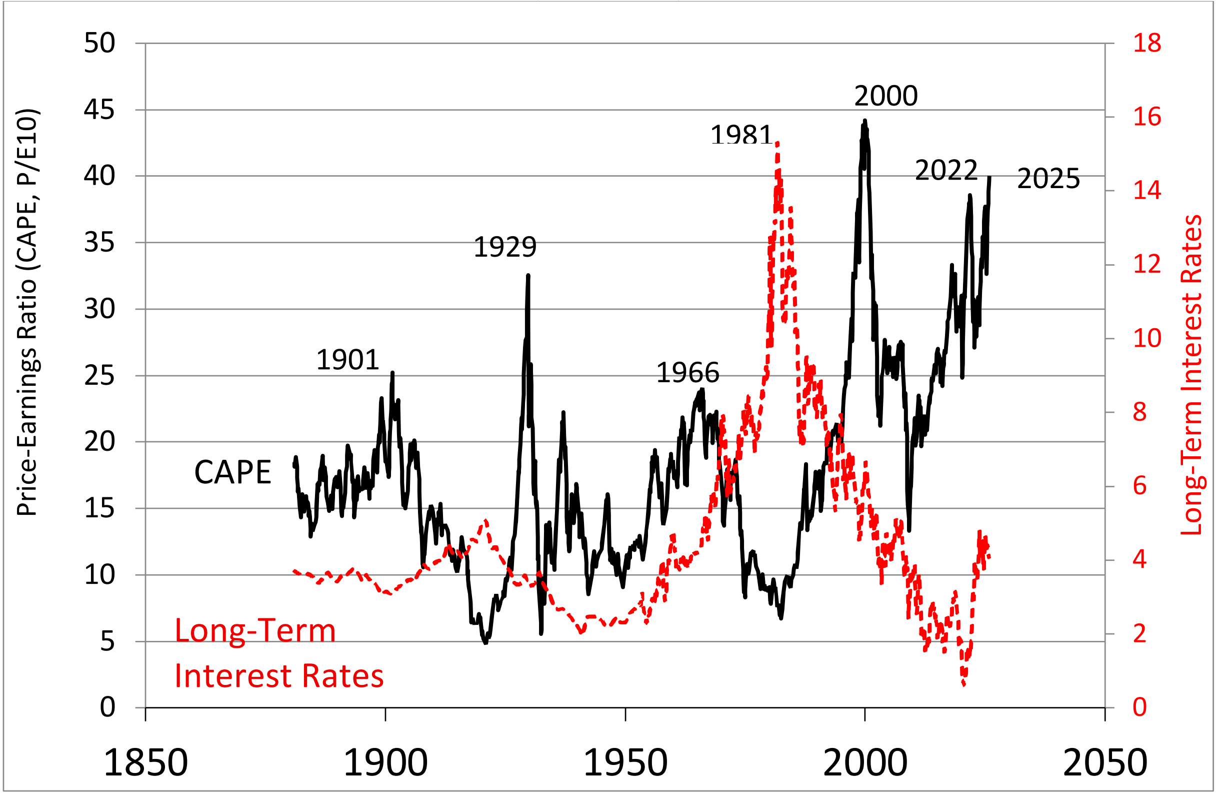
[8] At a recent *Wall Street Journal* tech conference, OpenAI Chief Financial Officer Sarah Friar suggested that a government loan guarantee might be necessary to fund the enormous investments needed to keep the company at the cutting edge (Wall Street Journal 2025).

construction industries, and individual investors and workers. All the massive resources wasted on these phantasmagorical pipe dreams could have been usefully spent on truly reviving manufacturing investment, necessary infrastructure (roads, bridges, the electricity grid), greening the energy system and improving social services. The opportunity costs of the current AI-driven model are enormous. All this is happening despite obvious signals that the U.S. economy finds itself already in bubble territory.

History Rhymes, Or Is This Time Different?

Figure 1 presents data on the S&P 500's Shiller P/E Ratio, which is commonly referred to as the cyclically adjusted P/E Ratio (CAPE Ratio), calculated as average inflation-adjusted earnings from the previous 10 years. Historically, a Shiller P/E Ratio above 30 has been a harbinger of speculative excess, followed by a bear market. In December 2023, the Shiller index rose to 30.45 and has remained above 30 ever since; in November 2025, the Shiller P/E ratio rose above 40. Since 1871, this is only the sixth instance in which the CAPE Ratio exceeded 30. The first time it happened was during August-September 1929 and we all know what came next: the Dow Jones Industrial Average lost 89% of its value (Williams 2024). The second time it happened occurred almost seven decades later: during the end-of-the-millennium dotcom bubble, when the Shiller P/E ratio recorded an all-time high of 44.19 in December 1999. Following the bursting of the dot-com bubble, the S&P 500 lost 49% of its peak value.

Figure 1
S&P500: Shiller Price-Earnings Ratio
(1870-2025)

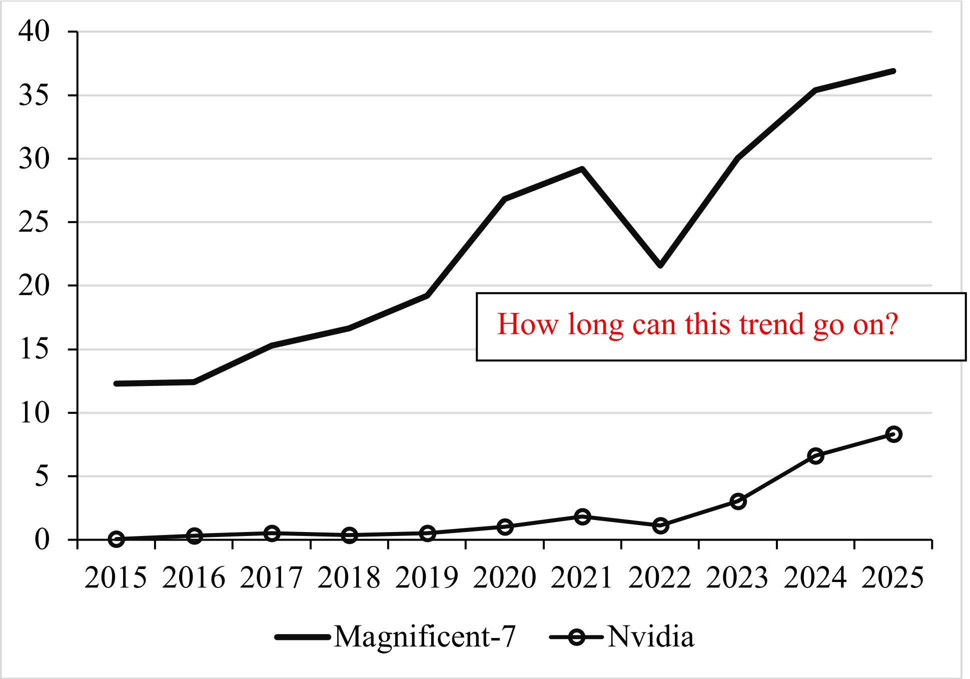


Source: Robert Shiller (2025), https://shillerdata.com/ (accessed on 15/11/2025)

The next three peaks above 30 in the Shiller P/E Ratio occurred very recently: during September 2017-November 2018; December 2019-February 2020; and August 2020-May 2022. Following these surges, the S&P 500 eventually dropped by anywhere between 20% and 33% (Williams 2024). We are currently living in the sixth such period of speculative excess.

The latest surge in the S&P500 P/E Ratio has been driven by the high P/E ratios of the so-called Magnificent-7 (Alphabet, Amazon, Apple, Meta, Microsoft, Nvidia and Tesla). The share prices and market capitalizations of these companies have increased much more strongly than the share prices of the other 493 corporations in the S&P500 Index, and as a result, the combined weight of the Magnificent-7 in the S&P500 Index has increased from 12.3% in 2015 to 36.9% on October 29, 2025 (**Figure 2**). The "Magnificent Seven" technology firms thus account for almost 37% of the S&P 500, a level twice as high as that of the top tech companies during the dotcom bubble (Carvão 2025).

Figure 2
Share in S&P500 Market Capitalization:
The Magnificent-7 & Nvidia (per cent)



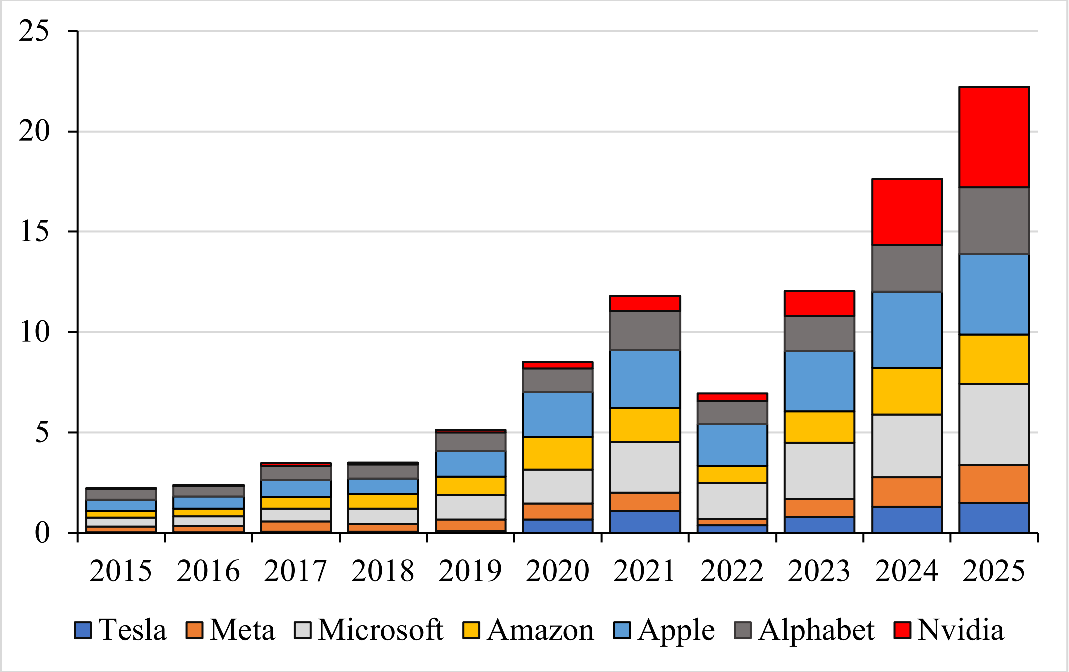
Source: L. Daly (2025), 'The Magnificent Seven's Market Cap Vs. the S&P 500', *The Motley Crew*, October 18; https://www.fool.com/research/magnificent-seven-sp-500/

Nvidia is the key firm in all of this, evolving from a rather small GPU-producing company largely known in gaming and graphics into the essential semiconductor engine behind the AI revolution. Nvidia had a weight of just 0.1% in the S&P500 Index in 2015, but now counts for 8.3% of this index — and, as the first company ever, its market capitalization has recently broken the $5 trillion mark (this happened on October 29, 2025; see **Figure 3**). NVIDIA does not own large commercial cloud data centers (like AWS, Google, or Microsoft), but it does provide complete data center architecture, high-performance, parallel-compute hardware (GPUs), software stacks, and reference designs to help cloud operators and enterprises deploy AI workloads. Nvidia's GPUs, more so than any other company's, have become the critical hardware that powers advanced AI models like ChatGPT — Nvidia's has a near monopoly in the GPU market, having a market share of around 94% (Mutjaba 2025).

While Nvidia remains at the heart of the U.S. AI industry, Big Tech peers Apple and Microsoft have also reached $4 trillion in market value in recent months (**Figure 3**). Since

ChatGPT's launch in November 2022, the Magnificent-7 have added an estimated $15.3 trillion in market value, reaching a combined market capitalization of $22.2 trillion on October 29, 2025. During November 2022-October 2025, AI-related stocks have delivered over 70% of the S&P 500's price returns and nearly 80% of its earnings growth.

Figure 3
Market Capitalization of the Magnificent-7
(2015-2025; Trillions of US dollars)

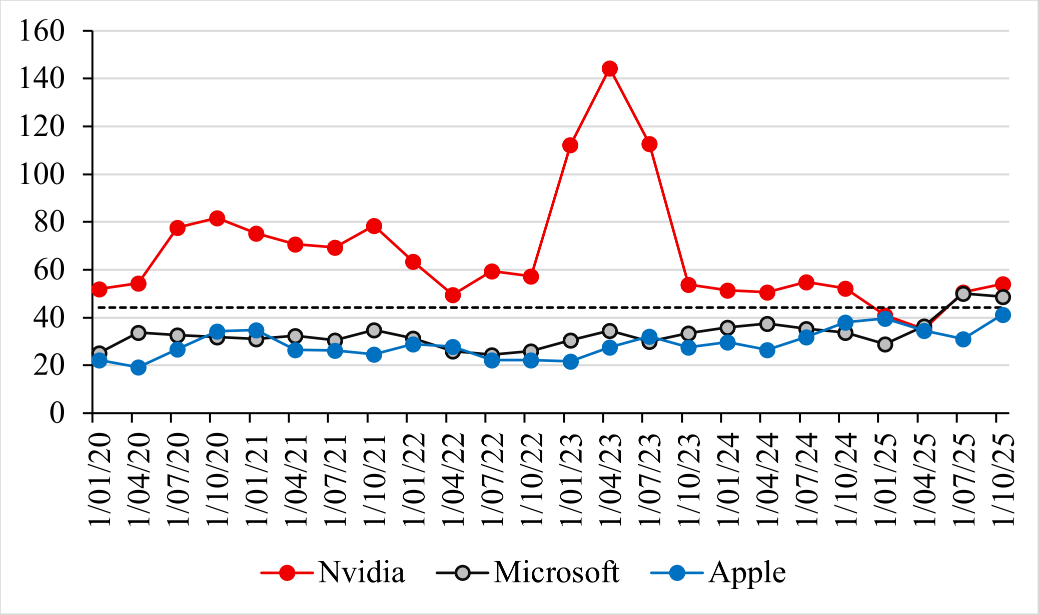


Source: L. Daly (2025), 'The Magnificent Seven's Market Cap Vs. the S&P 500', *The Motley Crew*, October 18; https://www.fool.com/research/magnificent-seven-sp-500/

Stock market valuations for AI companies are based on aggressively optimistic projections rather than current earnings. The stock market has priced AI as an exponential technology that will transform the economy as well as society in unprecedented ways, raising productivity, boosting innovation and generating profits.

The P/E ratios of major AI firms — Nvidia, Microsoft and Apple — have gone through the roof, surging to levels higher than 40 and, in Nvidia's case, higher than 50 (**Figure 4**). Financial investors are willing to buy shares at prices that are 40 to 50 times higher than earnings per share — which must mean that they are optimistically expecting soaring profits in the not-so-distant future, based on exponential increases in paid demand for AI-services.

Figure 4
Historical Price-Earnings Ratios: Apple, Microsoft & Nvidia
(2020q1-2025q3)

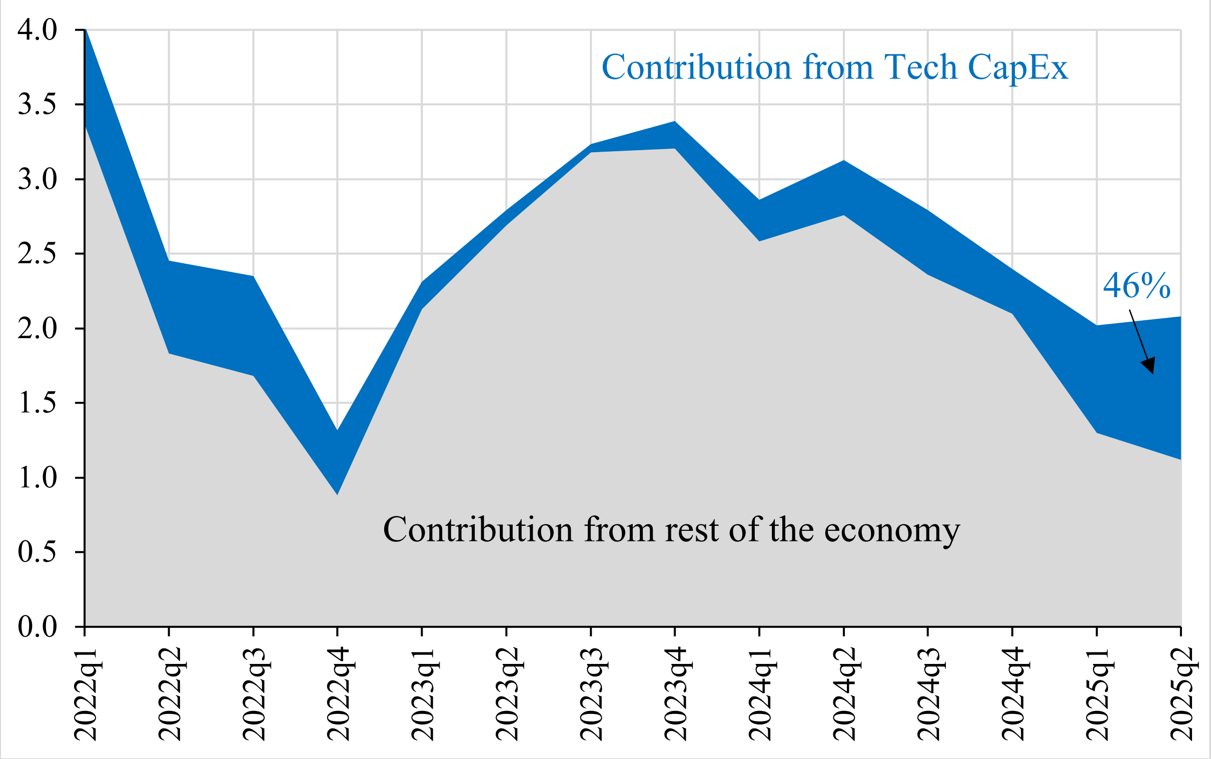


Source: https://www.macrotrends.net/stocks/charts/GOOGL/alphabet/market-cap (accessed on November 15, 2025). *Notes*: The dashed line represents the all-time high of the aggregate Shiller S&P500 P/E Ratio of 44.19 (reached in December 1999).

The capital expenditures by Big Tech are large enough to affect U.S. real GDP growth. As shown in **Figure 5**, the data-center related spending accounted for almost half of real GDP growth of the U.S. economy in the second quarter of 2025 (Sløk 2025a), up from just 14% during the years 2022-2024.[9] This is an outsized effect, given that investment in information processing equipment & software is only around 5% of GDP. It implies massive investment in just a small slice of the economy which, in turn, is diverting away funds from other — manufacturing — activities in the economy. The ebullient expectations concerning future rates of return in the AI industry have raised the hurdle rates of return for all other industries — and the result is that the AI bubble is inadvertently starving the rest of the American economy, big time (Kedrosky 2025a).

[9] Harvard economist Jason Furman calculated in a September 27, 2025 post on X, that annualized U.S. GDP growth in the first half of 2025 would have been just 0.1% without investments in information-processing equipment and software —categories largely driven by data center construction for AI infrastructure. He notes that investment in information processing equipment & software is 4% of GDP.

Figure 5
U.S. Annualized Real GDP Growth: Contribution from Tech Capital Expenditure
(2022q1-2025q2; percent)

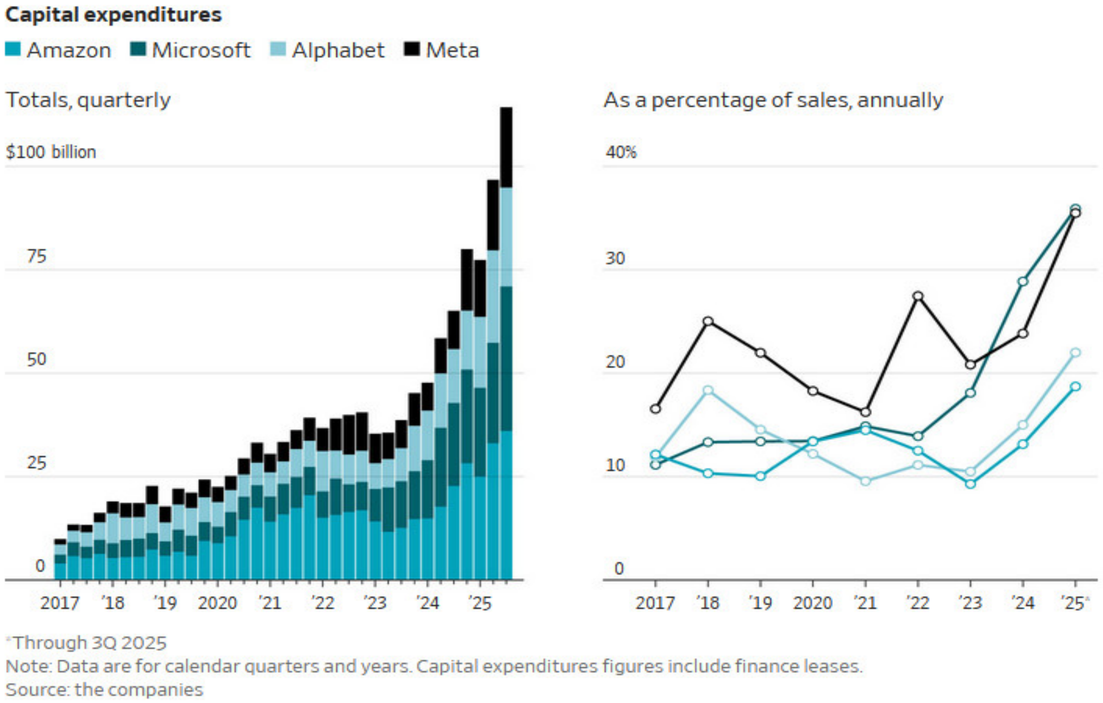


Source: Calculated based on BEA **Table 1.5.6.** Real Gross Domestic Product, Expanded Detail, Chained 2017 Dollars. *Note*: Capital expenditure by Big Tech is defined as real fixed investment in information processing equipment and software.

Fake It, Until You Make It!

The AI party is still in full swing. AI firms are racing to build out data center infrastructure for what they believe is virtually limitless demand for AI services. Among the headline announcements of the last couple of months: Microsoft, Alphabet, Meta and Amazon just tripled down on their AI infrastructure spending, reporting in October 2025 that their 2025 capital expenditures would collectively total roughly $380 billion (**Figure 6**). They expect that number to be substantially higher in 2026.

Figure 6
Capital Expenditures by Amazon, Microsoft, Alphabet and Meta
(2017q1 – 2025q2)


Capital expenditures

Totals, quarterly

As a percentage of sales, annually

*Through 3Q 2025
Note: Data are for calendar quarters and years. Capital expenditures figures include finance leases.
Source: the companies

Source: Christopher Mims (2025), 'When AI Hype Meets AI Reality: A Reckoning in 6 Charts.' *Wall Street Journal*, November 14.

In late September, Nvidia announced plans to invest up to $100 billion in OpenAI to fund a new generation of data centers (Nvidia 2025), while OpenAI pledged to lease millions of Nvidia chips for those 10 gigawatts (GW) facilities (OpenAI 2025a). Days later, OpenAI struck a similar multibillion-dollar 6GW deal with AMD (OpenAI 2025b) and a 10GW with Broadcom. OpenAI, Softbank and Oracle pledged to invest $500 billion in AI supercomputers. And in early November, OpenAI signed a seven-year, $38 billion deal to buy cloud services from Amazon.

JP Morgan Chase & Co, in a recent 58-page report '*AI Capex — Financing the Investment Cycle*', projects that 122GW of data center capacity will be built from 2026-2030 to satisfy the (arguably) astronomical demand for 'compute' (Wigglesworth 2025a). The additional 122GW of data center capacity is estimated to cost between $5-7 trillion. For 2026, the projected data center funding needs will be around $700 billion, which, according to the report, could probably be entirely financed by hyper-scaler *cash flows* and by High-Grade bond markets. “However, 2030 funding needs are in excess of $1.4 trillion, which will surpass current market capabilities, necessitating the search for alternative funding sources.”

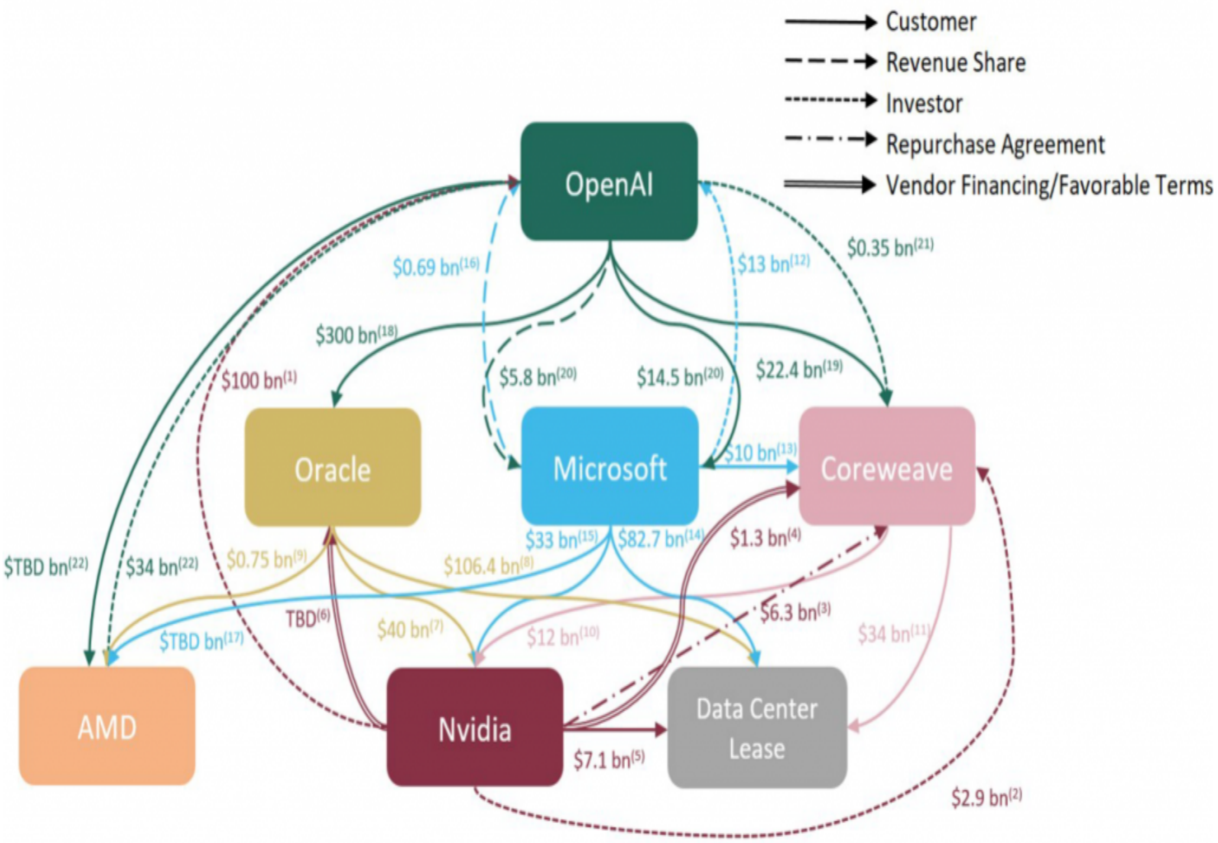
In particular, OpenAI has gone wild with compute/infrastructure deals this year (Seetharaman and Hu 2025): its CEO Sam Altman has said the startup is committed to spending $1.4 trillion in total to develop 30 GW of cloud computing and infrastructure capacity for AI training and inference — which would increase the current computing capacity of the U.S. (of 53.7 GW) by 55%. OpenAI, which is loss-making, does not have the funds to pay for these enormous capital expenditures. OpenAI will get $100 billion from Nvidia and another $40 billion from Japan’s Softbank. Altogether, this totals around $140 billion — in other words: merely 10% of OpenAI’s intended $1.4 trillion expenditure has been covered so far.

The flurry of mega-financing deals is mostly circular (Weil 2025), as is illustrated in **Figure 7**, and eerily resembling similar circular investment deals in the dot-com era. To give you a flavor: Nvidia invests in OpenAI and OpenAI is looking to buy millions of Nvidia’s specialized chips. OpenAI buys computing power from Oracle which buys Nvidia’s GPUs. Nvidia owns about 5% of CoreWeave and sells chips to CoreWeave. CoreWeave’s biggest customer is Microsoft, which is an investor in OpenAI, shares revenue with OpenAI, buys chips from Nvidia and has partnerships with AMD.

AMD, a rival to Nvidia, was so eager to land OpenAI as a customer that it issued warrants for OpenAI to buy 10% of AMD at a penny a share. OpenAI is a CoreWeave customer and also a shareholder. Nvidia has invested in xAI and will supply it with processors. And so on and so forth. The deals include revenue sharing across the stack and cross-ownership.

Nothing is transparent. It is not clear where the money needed for these deals is coming from. It is not clear what these opaque circular transactions imply for the valuations of the listed and non-public AI firms involved. It is not clear what all this means for the competition over hardware between chip producers (Nvidia versus AMD) and over AI services AI startups (OpenAI versus Anthropic versus xAI versus Microsoft). Not surprisingly, these astronomical circular financing deals are raising eyebrows. To many observers, they bring back traumatic memories of the circular financing arrangements of the late 1990s, when vendors and clients reinforced each other’s stock valuations without generating any real value.

Figure 7
The Circular Money Flows Among Major AI-Industry Companies:
"Robbing Peter to Pay Paul"



Source: Morgan Stanley (2025) Company Data, Morgan Stanley Research.

In response, AI industry leaders are doubling down on their message that the AI revolution is real and sustainable. Nvidia CEO Jensen Huang stated that there is no AI bubble and that exponentially growing AI demand is structural rather than speculative, comparing today's buildout to the "beginning of a new industrial revolution" (Kimball 2025). Asset-management firm Blackrock concurs (Barnette and Peterson 2025): "AI is not just a technological trend; it represents an infrastructure transformation with growing macroeconomic significance", adding that "unlike the speculative frenzy of the late 1990s and early 2000s, today's technology leaders are anchored by fundamental stability." Nouriel Roubini (2025), senior economic strategist at Hudson Bay Capital, chimes in, arguing that tech trumps tariffs and the U.S. economy will just do fine; note, however, that the same Roubini warned, in December

2024, that the era of stable growth and low inflation is over and will give way to a period of "secular stagflation" (Shaw 2024).

But not everyone is convinced: the cracks are starting to show (Zitron 2025a, 2025b; Schmidt and Xu 2025; Thornhill 2025). Slowing AI adoption rates (Storm 2025), surging capital and compute costs and elusive profits are fueling warnings that the boom may be headed for a hard reset. Hence, let us now consider the first factor that will contribute to the popping of the AI bubble.

First Problem: The Revenue Delusion

No, there is no world in which OpenAI's planned $1.4 trillion in infrastructure spending is going to pay off. Venture capitalist Tomasz Tunguz (2025) has calculated the implied revenue needed to support these spending levels at OpenAI's target gross profit margins, and finds that OpenAI would need to grow from an estimated $13 billion in revenue in 2025 to $983 billion in revenue in 2030 (*i.e*., an increase by 6500%). In line with this, AI-industry expert Ed Zitron (2025a) estimates that the AI industry will need $2 trillion in revenue by 2030 to meet their target profit margins, which he deems impossible. In its annual *Global Technology Report*, Bain & Co. (2025) concludes the same: by 2030, AI companies will need $2 trillion in combined annual revenue to fund the computing power needed to meet projected demand. Yet, Bain warns, *even under generous assumptions*, their revenue is likely to fall $800 billion short of that mark as efforts to monetize services like ChatGPT trail the spending requirements for data centers and related infrastructure. And finally, according to the recent model forecast by J.P. Morgan (2025), the AI industry would need to generate $650 billion in annual revenues in perpetuity just to deliver a modest 10% return on the massive investments expected through 2030 (Wigglesworth 2025b); the 10% rate of return is far below the AI industry's target rates of return — and JP Morgan's implicit bubble warning (Rizvi 2025) is that AI firms are grossly overvalued.[10]

[10] The JP Morgan report identifies two major threats to long-term profitability: monetization risk and technological obsolescence. Rapid innovation cycles could render current AI chips and systems outdated before they achieve profitability.

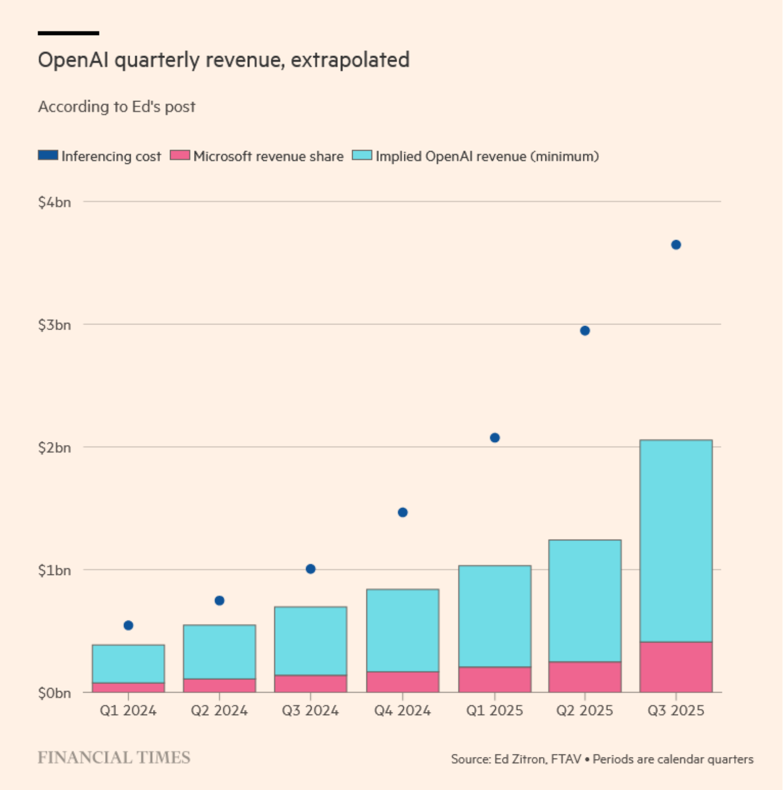
These revenue and profit targets are insane. To explain why, consider OpenAI's revenue target of $983 billion in 2030 in greater detail. For starters, let us look at the results of the very clever financial detective work done on OpenAI's quarterly revenue and inference costs by Ed Zitron (Elder 2025). OpenAI's financial reporting is notoriously opaque and convoluted. Zitron, therefore, uses the only available official data concerning OpenAI, published by Microsoft which has a major financial stake in OpenAI. Based on Microsoft's official numbers, Zitron, working together with financial experts of the *Financial Times Alphaville* (Elder 2025) concludes that OpenAI's quarterly revenues during 2024q1-2025q3 fall increasingly short of its inference cost (incurred at Microsoft's Azure web-hosting platform).

As **Figure 8** shows, OpenAI appears to have spent more than $12.4 billion at Microsoft Azure on inference compute alone in the last seven quarters. Its implied revenue for the period was a minimum of $6.8 billion. OpenAI's cumulative loss must thus have been around $5.6 billion. When asked about the accuracy of these disturbing numbers, OpenAI responded that the *FT* team should ask Microsoft (Elder 2025); Microsoft responded saying that "the numbers aren't quite right." Asked what exactly that meant, the spokeswoman said Microsoft would not comment and did not respond to subsequent requests (Elder 2025). One would have expected a strong response from Microsoft and OpenAI if the numbers were obviously incorrect.

Hence, it is reasonable to assume that the numbers in **Figure 8** are likely roughly right.

According to financial disclosures to shareholders, OpenAI's projected cumulative losses during 2025-2029 will amount to $35 billion. Zitron's (2025b) calculations suggest that OpenAI will remain loss-making probably for much longer, because, net of the 20% Microsoft revenue share, OpenAI's minimum estimated revenue would cover inference costs approximately never (Elder 2025). Let me repeat: approximately never!

Figure 8
Estimated Quarterly Revenue and Inferencing Cost: OpenAI
(2024q1 – 2025q3)



Source: Ed Zitron (2025b) & *Financial Times* (Elder 2025).

Zitron's (2025b) estimates fundamentally challenge the viability of OpenAI's business model. Going by his numbers, either running costs (of inference) have to be lowered much faster or customer charges have to rise dramatically (to raise revenues). However, lowering AI's running costs is going to be difficult, maybe even impossible, for the following two reasons.

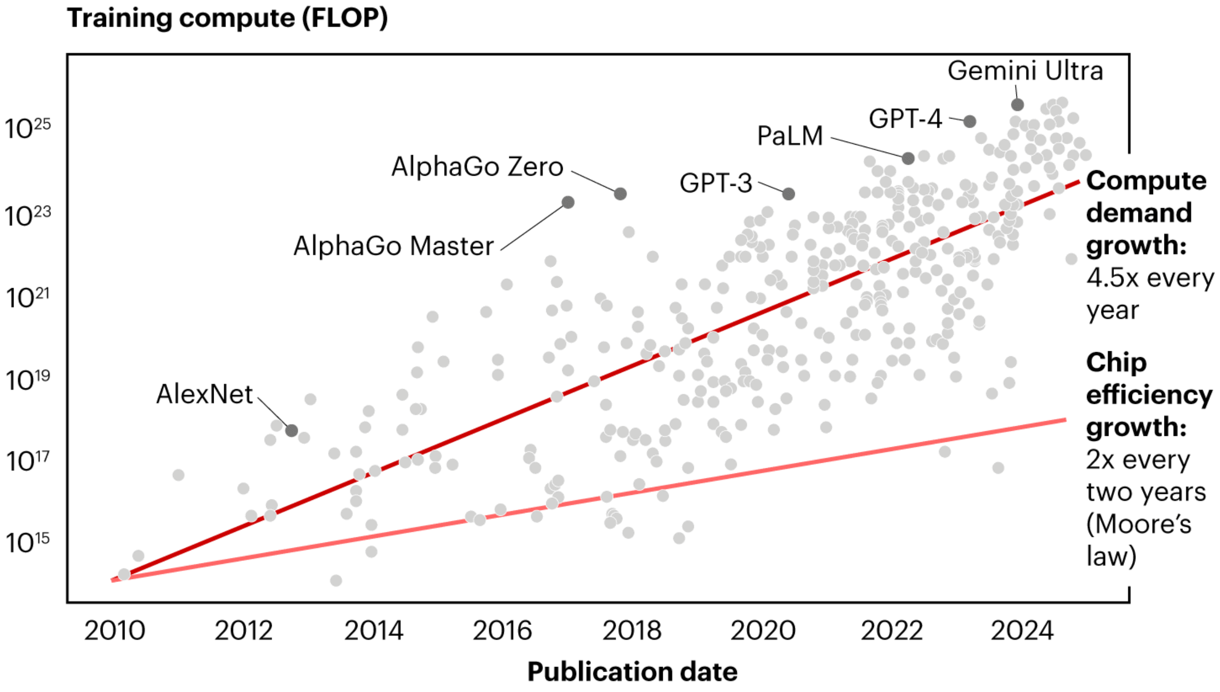
First, AI inference and training requirements R can be defined as:

$$(1) \qquad R = e \times n$$

where R = the number of floating point operations (FLOPs = the number of calculations a system performs, or the scale of AI compute); e = a measure of the 'performance' efficiency of the GPUs (which depends on transistor density), and n = the number of installed GPUs. According to Bain & Co. (2025), AI inference and training requirements (R) have grown at more than twice the rate of GPU 'performance efficiency' (*e*), with the result that data center operators have to rely on brute-force scaling, *i.e.*, increasing *n*.

The crux of Bain's argument is that compute demand is scaling faster than the efficiency of the tools that generate it. In fact, the growth rate for AI's compute demand (which is driven by the belief that the bigger the scale of AI computing power, the better will be its output) is more than twice the rate of Moore's law (**Figure 9**). As a result, the demand for GPUs by the AI industry is soaring — and because Nvidia has a near monopoly on GPUs, it is capable of forcing (cash-constrained) neocloud companies (CoreWeave, Lambda, Crusoe, Together.ai) to play by its rules, leasing GPUs, rather than buying them (Mallipatna 2025).[11]

Figure 9
Compute Demand Grows Twice As Fast As Chip Efficiency



Source: Bain & Co. (2025), *Global Technology Report*. *Notes*: Chip efficiency growth not shown to exact scale, with the rate of growth intended to be illustrative; FLOP = number of floating point operations, or the number of calculations a system performs.

[11] Hyperscalers (AWS, Microsoft Azure, Google Cloud, Oracle) typically purchase GPUs because they have the balance sheet and deep pockets to absorb multi-billion dollar capital expenditures and can depreciate assets over time. They also build the data centers, earning higher margins and maintaining control over pricing.

Similarly, cash-strapped AI start-up firms (OpenAI, Anthropic) typically rent compute by the hour from clouds or neo-clouds, which is a (pay-as-you-go) lease-on-a-lease model (that turns capital expenditures into operational expenses). Crucially, each (lease) layer adds a profit markup, which ultimately flows back as rent income to Nvidia, as the primary beneficiary, which remains the owner of the GPUs, locks customers into long-term dependency on its semiconductor technology, while maintaining its pricing power.

This core-dependency structure of the AI compute economy, with Nvidia comfortably in the core and neocloud companies and AI firms in the periphery, considerably raises the cost of AI inference and training. In fact, if we assume that the Nvidia H100 GPU costs ~$25.000 (approximate numbers), Nvidia can earn between $27,000–$70,000 per GPU over just three years by leasing (assuming high utilization). In the extreme case, the cost of compute increases by a factor of almost 3. For Nvidia, this turns hardware into a recurring financial flywheel, as Mallipatna (2025) writes. Lowering AI's running costs is simply not in the interest of Nvidia — the spider in the middle of the AI industry web.

However, the revenue outlook for the AI industry is probably even more problematic. This is related to the **second reason** why lowering AI's running costs is going to be difficult, maybe even impossible: AI's inference costs are rising (not declining), even though the cost per million tokens[12] of LLM inference has declined by ~98% — from $20 per million tokens in late 2022 to around $0.40 per million tokens in August 2025 (Barla 2025). Hardware efficiency improves roughly fourfold and software optimization threefold with each generation, and 'efficiency' improves by a factor of 12 as a result (Lishawa 2025). However, as **Figure 9** shows, the AI industry uses considerably larger training data sets now than one or two years ago, because the frontier AI models try to mimic 'reasoning', by breaking a user's input into component parts, then run inference on each one of those parts, and as a result 'consume' ever more tokens to generate one extra word (of output).

Frontier models allocate massive computational effort during inference itself in *test-time computing*. Test-time computing is the latest approach to scaling LLMs, which allows for

[12] Tokens, the smallest unit of textual data processed by AI models, are central to inference compute. Typically, one word corresponds to about 1.4 tokens. Here is how AI treats some common text: "Hello" → 1 token; "Fantastic" → 1 token; "Shouldn't" → 3 tokens (*"should", "n"* + *"'t"*); "Artificial intelligence" → 2 tokens (*"Artificial", "intelligence"*).

longer context windows (Torene 2025) and bigger suggestions from the models. Competition among AI firms to demonstrate agentic behavior amplifies the problem, as it requires far more tokens to sustain planning chains and memory states. As a result, a typical enterprise query in 2021 used fewer than 220 tokens; by 2025, models such as GPT-4 Pro and ChatGPT-5 process around 22,000 tokens in a single exchange based on test-time computing (Lishawa 2025). Therefore, frontier AI models now require over 100x more tokens than four years ago and at the current rate of expansion, tokens per query could rise to between 150,000 and 1,500,000 by 2030, depending on task complexity.

Each token interacts with every parameter in a model, requiring two floating-point operations (FLOPs) per token-parameter pair. This implies that 100x more token-parameter pairs require 200x more FLOPs and 200x more electricity usage. In effect, with a rapidly declining electricity cost per token and a much larger increase in token consumption per generated word, AI application inference costs have grown about 10 times over the last two years (Szyszka 2025). Any per token cost savings are therefore outweighed by the increased use of tokens in test-time computing — an approach which is deemed critical to the success of the scaling strategy to LLMs pursued by AI firms. Efficiency reduces cost per token but greatly expands token use and raises total cost and total energy use (Spark 2025), reproducing the rebound pattern familiar from earlier industrial energy systems.

In conclusion, inference cost are unlikely to go down, and this is true for LLM training cost as well. This means that the only feasible way for the AI firms to generate positive returns is to drastically raise customer charges. Here, things become pretty wild.

OpenAI claims to be making roughly $13 billion in annual revenue in 2025, but probably only around $3 billion comes from paying (every-day and enterprise) ChatGPT users. OpenAI boasts that ChatGPT has 800 million regular users per week in October 2025, but only 5% of them (~4 million users) are paying subscribers, paying (a minimum of) $20 per month. 95% of everyday users use the free product and remain unconvinced about the value of AI-powered devices and services. More money from usage comes from the 1.5 million enterprise customers that are using ChatGPT. However, by far the biggest payers for OpenAI's compute are Microsoft, Meta, Amazon and Nvidia, which are all key actors in the AI ecosystem. In other words, paid demand for OpenAI's compute coming from outside the AI industry remains limited. OpenAI generates revenue by selling compute to Nvidia, which in turn is generating revenue by selling or leasing GPUs to OpenAI. This illustrates the dangerous degree of circularity of the AI industry.

OpenAI will never be able to generate the $650 billion in revenues in 2030 needed to support the projected spending levels at a gross profit margin of 10%, based on the subscriptions of paying users outside Silicon Valley. There are around 1.5 billion active iPhone users and over 300 million paying Netflix subscribers — and as JP Morgan puts it (Morales 2025), it is unrealistic to assume that these iPhone users or Netflix subscribers would be willing to pay an extra $35 or $180 each month, respectively, for OpenAI's services. This will not happen, especially because user satisfaction with the performance of its AI tools is stagnating or even declining. Enterprise customers already becoming less enthusiastic about their AI tools: according to an August 2025 MIT report titled *The GenAI Divide: State of AI in Business 2025*, published by MIT's NANDA initiative, 95% of generative AI pilot projects in companies did nothing to raise revenue growth (Challapally *et al.* 2025).

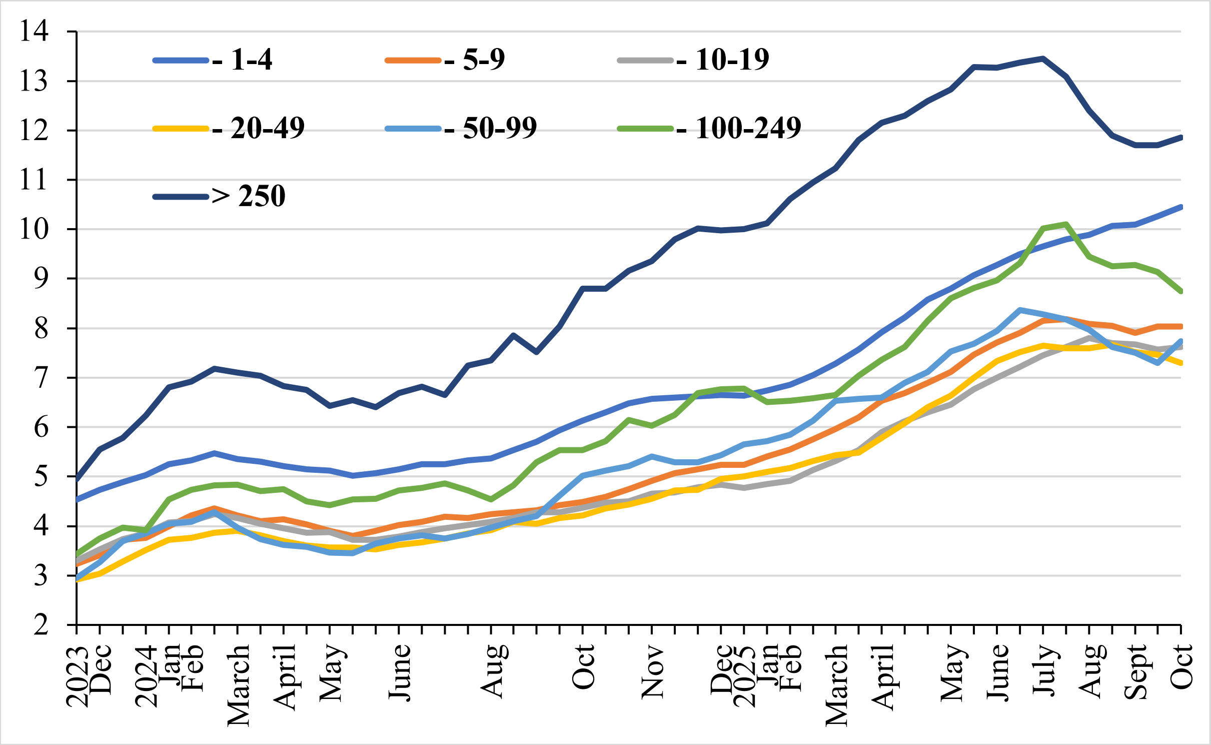
Concerns about AI-generated 'slop' hurting worker and firm productivity are mounting, and for good reason. Using two large-scale AI adoption surveys (late 2023 and 2024) covering 11 exposed occupations (25,000 workers in 7,000 workplaces) in Denmark, Humlum and Vestergaard (2025) show, in a recent NBER Working Paper, that the economic impacts of GenAI adoption are minimal: "AI chatbots have had no significant impact on earnings or recorded hours in any occupation, with confidence intervals ruling out effects larger than 1%. Modest productivity gains (average time savings of 3%), combined with weak wage pass-through, help explain these limited labor market effects." And recent U.S. Census Bureau data by firm size show that AI adoption has been declining among companies with more than 250 employees (**Figure 10**) (as noted by Sløk 2025b).

Why would individual users and enterprises pay huge subscription fees to the Silicon Valley AI industry, when they have the option to use much cheaper Chinese LLMs, such as DeepSeek and Alibaba Cloud, which achieve performance levels comparable to leading American AI models[13] and continue to slash prices? The glut in AI data centers is already

[13] While the U.S. maintains its lead in number of AI models, Chinese models have rapidly closed the quality gap: performance differences on major benchmarks such as MMLU and HumanEval shrank from double digits in 2023 to near parity in 2024. Meanwhile, China continues to lead in AI publications and patents. *Source*: *The 2025 AI Index Report*, Stanford University.

leading to price wars, while U.S. AI firms continue to channel more money into one single idea, *viz*. scaling ever larger language models, in the hopes that something magical will emerge.

Figure 10
AI Adoption Rates Are Starting to Decline for Larger Firms
(November 2023-October 2025)



Source: U.S. Census Bureau, *Business Trends and Outlook Survey* (BTOS) 2023-2025. *Notes*: The U.S. Census Bureau conducts a biweekly survey of 1.2 million firms. Businesses are asked whether they have used AI tools such as machine learning, natural language processing, virtual agents or voice recognition to help produce goods or services in the past two weeks. See Torsten Sløk (2025b), https://www.apolloacademy.com/ai-adoption-rate-trending-down-for-large-companies/

Chinese AI companies could flood the American market with cheap AI bots, with comparable performance, and pop the AI bubble on Wall Street (Marcus 2025c). This is a realistic scenario in a belligerently multipolar world. But even if it does not happen, a price war between AI companies, American and Chinese, leading to an industry shake-out, is inevitable, given the massive overcapacity that is currently under construction and in planning. Or, as

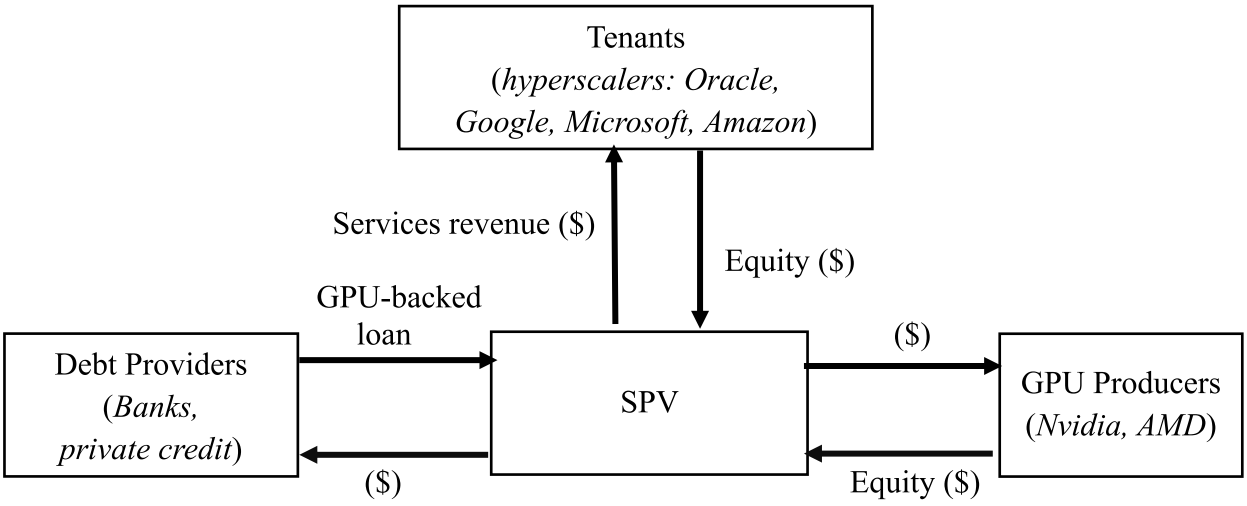
OpenAI’s Sam Altman puts it, . “Someone's gonna get burned there, I think […] Someone is going to lose a phenomenal amount of money – we don't know who …” (Heath 2025). This time, he is stating the truth.

Second Problem: The Ticking Time-Bomb of Hyperscale Borrowing

There is no world in which the AI industry can fund its capital expenditures of $5-7 trillion during 2026-2030 out of revenues from paid subscribers or money from sovereign wealth funds. OpenAI, Anthropic and other startups continue to lose money, and must fund most of the planned investment by selling off pieces of themselves to investors and by resorting to *hyperscale borrowing* from banks and investment-grade bond markets. Ignoring numerous red flags, particularly operating and financial leverage, virtually every Wall Street player is angling to get a slice of the action, from banks such as JPMorgan Chase and Morgan Stanley to asset managers such as BlackRock and Apollo Global Management. This Wall Street frenzy will in all likelihood lay the foundations for the next debt crisis (Fitch 2025).

The point is that the Big Tech firms are off-loading the risks associated with their gargantuan data center investments to special purpose vehicles or joint ventures (**Figure 11**). These off-balance-sheet vehicles raise equity (a small slice) and much more debt (the big slice); these special vehicles are therefore highly leveraged. Too-big-to-fail financial firms on Wall Street take a large minority ownership in these vehicles and also help issuing the bonds. The special purpose vehicles (SPVs) generally pay a significant premium — of up to 200 to 300 basis points, or 2 to 3 percentage points (see Kedrosky 2025b) — on its bonds over and above investment grade interest rates. This offers Wall Street players a bonanza. Investment-grade companies such as Meta agree to paying the interest rate premium (via the SPVs), because they do not want to show the massive data center debt as debt on their balance sheets. Meta completed nearly $300 billion in financing through SPV structure. Put differently, they are paying the premium for deliberately designed opacity and supposed balance sheet remoteness, in the full knowledge that a default on the data center investments will nonetheless impact their balance sheets. Wall Street players create customized financings for the AI firms (Platt *et al*. 2025), including structuring loans into tranches of asset-backed securities (ABS), collateralized by the data centers (which are to be build). These premium-paying bonds and data-center-backed securities (and associated risks) will be sold to gullible retail financial investors and other financial firms (insurers, corporate debt funds and almost every type of bond buyer) — and these actors will be left holding the bag, once things go south.[14]

Figure 11
Special Purpose Vehicles (SPV) for GPU Purchases or Leases



Source: Arun (2025).

The bubble is not confined to just the data centers. Power utilities say that they need two or three times more electricity within a few years to power the gigantic AI data centers. Power utilities also prefer often highly-leveraged SPVs as the financing vehicle for developing and constructing the large-scale energy and infrastructure projects. This is primarily because SPVs have separate legal personality, and do not have to be consolidated in their balance sheets. This 'corporate veil' obscures the involvement and risks of these utilities in the AI data-center boom.

Based on its — over-optimistic — projections of the future revenue streams of the AI industry, JP Morgan staff estimates that the AI industry will need approximately $1.5 trillion

[14] Insurance firms invest heavily in data-centers via private credit firms (also known as shadow banks). The established platform companies (often hyper-scalers)will not be the ones to take the eventual hit.

in investment-grade (IG) bond funding over the next five years (2026-2030). This astonishing bond funding requirement is supplemented by $150 billion in leveraged loans, plus high-yield debt, private credit, and new asset-backed securitization mechanisms for data center assets. For example, the Stargate project, backed by OpenAI, Oracle, Softbank and Emirati-owned investment firm MGX, aims to invest $500 billion in data center infrastructure exclusive for OpenAI in the U.S. by 2029. JPMorgan Chase and Mitsubishi UFJ agreed to lead a banking syndicate of more than 30 banks which will lend $2.3 billion to the Stargate project (Wirz 2025); the loan will be structured into asset-backed securities and sold to (retail) financial investors. Oracle needs to borrow billions more for its spending spree and its debt is rising above $100 billion. In response, Moody's Ratings and S&P Global Ratings are edging closer to reclassifying Oracle's bonds as junk debt (Tracy 2025).[15]

Similarly, xAI set up a joint venture with investment firm Valor Equity Partners which aims to raise between $15bn and $20bn in debt and equity to purchase chips on behalf of xAI needed for its Colossus 2 data center in Memphis, Tennessee. The joint venture will borrow from private-credit firms, including Apollo Global Management, which in turn will sell these loans as ABS to financial investors. Even cash-flush Meta is signing complicated, off-balance sheet, debt deals involving private-equity firms (Zitron 2024). In particular, Blue Owl Capital invested around $3 billion in an 80% stake in a joint venture with Meta to build Hyperion data centers (Wirz 2025).[16] The joint venture sold bonds worth $27 billion, with Pimco buying two-thirds of these bonds.

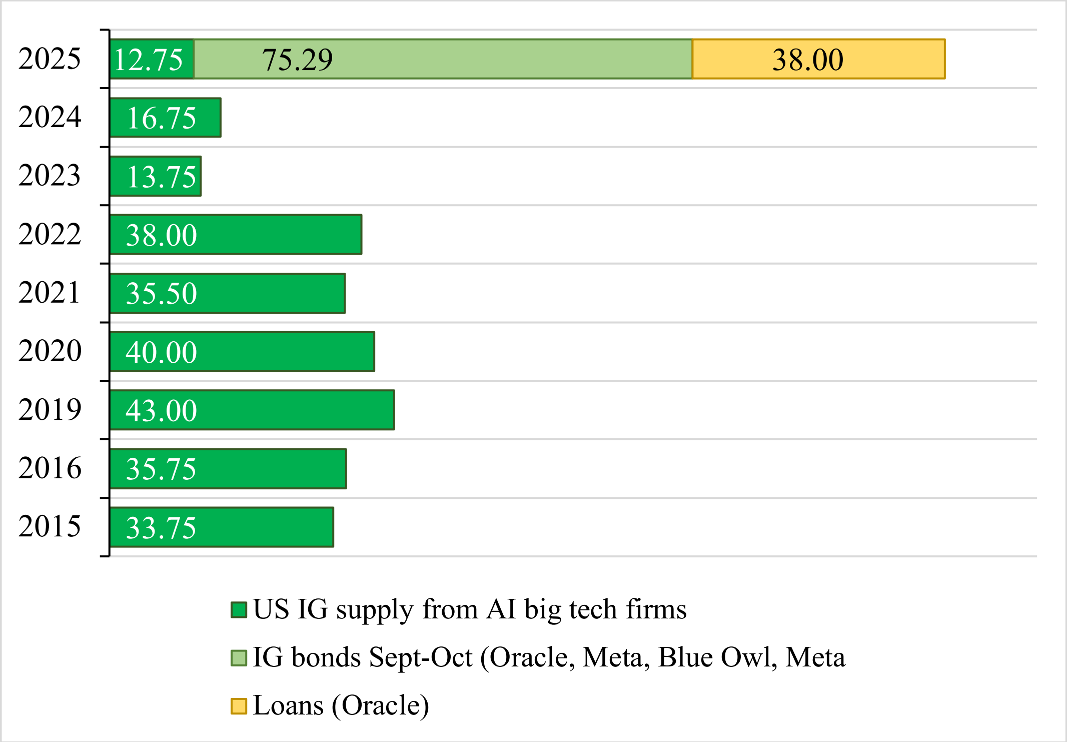
It is evident that investor appetite for data-center debt is still very strong (see **Figure 12**). Investment-grade borrowing by AI big tech firms during September-October 2025 amounted to $75 billion — compared to $32 billion on average per year during 2015-2024. IG bonds issued by the AI companies make up 14% of the American IG bond market in October 2025. Barclays estimated that cumulative AI-related investment could reach the equivalent of more

[15] Oracle's stock price has fallen by 40% during September 21-November 21, 2025.

[16] Blue Owl Capital was an upstart investment firm that lent money to midsize U.S. companies. These days, the firm is financing massive data centers costing tens of billions of dollars for Meta and Oracle. Blue Owl raised about $30 billion to build an AI data center for Meta in Louisiana, putting in $3 billion of its clients' money and borrowing the rest. The Wall Street Journal (November 16, 2025) notes that "the deal included a provision, considered extraordinary on Wall Street, giving Blue Owl's equity investment a debtlike guarantee in case the partnership falls apart" (Wirtz 2025).

than 10 per cent of US GDP by 2029, compared to circa 6% in the first six months of 2025 (Smith, Herbert and Rees 2025).

Figure 12
Borrowing for AI Data Center Construction
(2015-2025; billions of US dollars)



Source: Bank of America Global Research; chart created by Lucy Raitano (October 31 2025). *Notes*: IG = investment grade. The data for Blue Owl and Meta refer to a project-style holding company created by Blue Owl Capital to invest in a large-scale Hyperion data center joint venture with Meta.

One can reasonably worry whether these debts will ever be repaid. Cash-flow projections are very uncertain, particularly because the cost of providing AI inference continues to increase (**Figure 8**), while growing competition — also from China — will depress prices. Circular financing among the leading tech companies, AI service providers and big financial players creates an interlocking liability structure across the sector, creating concentration risks for lenders and shareholders. The lack of transparency in these debt-financed transactions and the interlocked liability structure between Silicon Valley and Wall Street are augmenting the system risks of the AI industry.

At the firm level, OpenAI's $1.4 trillion commitments represent complex, long-term contracts like Power Purchase Agreements (PPAs) for energy, which lock in electricity costs and capacity for decades to come, effectively pre-buying the energy needed to run its models. This financial engineering exposes OpenAI to immense physical-world risks, originating from grid stability, supply chain disruptions, and regulatory blowback (see *The Small Cap Strategist* 2025).

Even worse, the hyperscale data-center borrowing creates a ticking time bomb, waiting to explode, on the balance sheets of Big Tech and AI firms, because the capital expenditure is on specialized GPUs and servers, which — because of unrelenting technological progress — risk becoming economically obsolete within two or three years (Arun 2025; Smith 2025). Data servers, networking equipment and storage devices have a useful lifetime of 3-5 years and a corresponding annual depreciation rate of 20%-30%. These chips aren't general-purpose compute engines; they are purpose-built for training and running generative AI models, tuned to the specific architectures and software stacks of a few major vendors such as Nvidia, Google, and Amazon. These chips are part of purpose-built AI data centers — engineered for extreme power density, advanced cooling, and specialized networking. Together, they form a closed system optimized for scale but hard to repurpose.

Nvidia is building new generation GPUs each and every year and GPU prices are likely falling. The rate of economic decay of the AI compute infrastructure is high and the payback periods are correspondingly short. "You're investing in something that is a perishable good," economist David McWilliams told *Fortune* (Lichtenberg 2025), calling AI hardware "digital lettuce" that is "going to go off now." Michael Burry, who famously shorted the U.S. housing market before its collapse in 2008, similarly argued that AI companies' projected growth looks to be massively exaggerated (Tangermann 2025), accusing the AI firms in a November 10 tweet of assuming unrealistically low depreciation rates for AI hardware to boost the numbers. In fact, data centers depreciate twice as fast as the revenues are growing (Milmo 2025). The AI industry building generic LLMs will not be able to recoup the $1.5 trillion with an adequate rate of return.

Third Problem: Exponential Growth in an Analogue World

It will be impossible to build the projected data center infrastructure in the next five years or so (which is the horizon of most AI investors). The lead time necessary to build a hyperscale data center is currently around 2 years, but expect it to become much longer, say 7 or more years. Why? Upstream suppliers to the growth in data centers — the established analogue companies producing everything from copper wire to gas turbines to transformers and switchgear — have to expand production (Patel and Dean 2025). These suppliers of data-center components will have to build new factories and equipment, hire more workers and mobilize the funding to finance all this. Doing so, these upstream suppliers will run into labor shortages, long waiting times for power grid connections and securing construction permits, material (supply-chain) bottlenecks and regulatory dead-ends – and all this will lengthen the lead times necessary to build a hyperscale data center.

Consider just the example of gas turbines: as AI drives unprecedented data center growth, next-generation data center operators bypass traditional power grids, turning to on-site generation to meet their urgent energy demands. Around three out five data centers are exploring on-site power generation to boost energy efficiency or resilience (Robb 2025). Natural gas ends up being the fastest way to scale power generation: it is a mature technology and the nation has a vast pipeline network to take it to the sites; and it can be deployed on site. Meta's Hyperion project is already using three H-class natural gas turbines as part of its multi-gigawatt AI data center in in Louisiana. The Oracle/OpenAI Stargate project in Abilene Texas will be powered by a combination of gas turbines and fuel cells. xAI is ordering up to 60 gas turbines for its supercomputer facility in Memphis, Tennessee. Hence, the demand for natural gas turbines is rising very fast (Robb 2025; Johnson 2025).

The problem for the suppliers of these gas turbines is that if they build new factories (to produce these turbines), those factories have to be amortized over a period of two or three decades. At usual margins, those factories are only worth building if the AI demand lasts for 10-30 years. This is deeply uncertain. The Big Three gas turbine manufacturers — GE Vernova, Siemens Energy and Mitsubishi Electric, which together control over 75% of the global gas turbine market, with Mitsubishi Power alone commanding a 36% market share — have reasonable doubts that the data center boom will last this long (Johnson 2025). Hence, they are increasing their productive capacities only gradually. Mitsubishi Electric tells customers that turbines ordered today will not arrive until 2030, and also GE Vernova pushes deliveries past 2029.[17] According to the chief executive of GE Vernova, Scott Strazik, meeting the projected demands for U.S. electricity production is not something his industry can solve in the next five years, but rather in 10 or 15 years.

The Big Three gas turbine manufacturers are effectively rationing supply, creating temporary rents. The resulting cost escalation has been considerable (Johnson 2025). In 2022, combined cycle gas-fired construction costs averaged $722 per kilowatt, according to the U.S. *Energy Information Administration*. Today, gas turbine construction costs are $2,200 to $2,500 per kilowatt for combined cycle configurations, *i.e.*, three times higher than in 2022 (Johnson 2025).[18] This is another key factor driving up the costs of data center investments and lowering their returns.

McKinsey (2024) predicts that data center energy use in the U.S. will more than double — from TWh 224 in 2025 to TWh 606 in 2030. Data centers would claim roughly 10% of aggregate US electricity generation and the U.S. electricity grid will need to expand. In many cases the regional power grids are hopelessly inadequate to cope with power demand possibly surging over the next few years. Data center power demand is already driving up wholesale electricity costs, with prices today as much as 267 per cent higher than five years ago in areas near data centers (Bloomberg 2025). Electricity could well be the biggest obstacle to the technology's deployment and growth (Yoon 2025), according to Goldman Sachs (2025). Adding 100 GW or more of power in a timely manner is a remarkable challenge, particularly in light of the necessary grid upgrade requirements. It takes many years: the analogue buildout is out-of-sync with the exponential demands of Silicon Valley firms. Moreover, whether they like it or not, the AI Industry will need to come to terms with the fact that keeping retail electricity prices under control is a politically important and sensitive aspect of managing the data center boom.

[17] Scott Strazik, the chief executive of GE Vernova, one of the biggest U.S. manufacturers of equipment to generate electricity, such as transformers and natural-gas turbines, recently said on *The Wall Street Journal*'s Bold Names podcast that nearly all of the company's output is booked through 2028. which means that there is no spare manufacturing capacity to build that equipment any faster.

[18] This cost escalation makes alternatives for onsite electricity increasingly attractive by comparison. Investment in nuclear energy is increasing, with the money being spent on small modular reactor (SMR) development and improvements to existing plants. However, nuclear energy has extremely low operational expenditures, but very long lead times (of up to 20 years) and high capital expenditures.

FT Alphaville (2025) further notes that the nature of the AI-related power demand is particularly problematic (Wigglesworth 2025a). It cites a recent Nvidia (2025) report:

> "Unlike a traditional data center running thousands of uncorrelated tasks, an AI factory operates as a single, synchronous system. When training a large language model (LLM), thousands of GPUs execute cycles of intense computation, followed by periods of data exchange, in near-perfect unison. This creates a facility-wide power profile characterized by massive and rapid load swings. This volatility challenge has been documented in joint research by NVIDIA, Microsoft, and OpenAI on power stabilization for AI training data centers. The research shows how synchronized GPU workloads can cause grid-scale oscillations. The power draw of a rack can swing from an "idle" state of around 30% to 100% utilization and back again in milliseconds. This forces engineers to oversize components for handling the peak current, not the average, driving up costs and footprint. When aggregated across an entire data hall, these volatile swings — representing hundreds of megawatts ramping up and down in seconds — pose a significant threat to the stability of the utility grid, making grid interconnection a primary bottleneck for AI scaling."

The imperative of a stable utility grid thus constitutes a major obstacle for AI scaling.

But so do labor shortages: the surge in data center construction and the expansion of the electricity grid will create massive labor and skill bottlenecks. According to the *Willis Global Construction Rate Trend Report* for 2025Q1 (Rafetto 2025), there is shortage of skilled construction workers, with estimates indicating that an additional 500,000 workers (electricians, technicians, ironworkers, etc.) are urgently needed to meet the mounting construction demand.[19] The *National Association of Manufacturers* projected in 2025 that the U.S. could face a shortfall of 1.9 million manufacturing workers by 2033 (Rogers 2025).

This labor shortage is alarming given the need for skilled workers to build out the AI data center infrastructure. To illustrate the order of magnitude of the shortage: The 1.2 GW Stargate facility in Abilene has a construction workforce of over 5,000 people. Simple extrapolation suggests that the U.S. will need around 420,000 workers to build 100 GW of AI

[19] Recent natural disasters, including the devastating wildfires in California, have added to the demand for (re-)construction workers.

data center capacity during 2026-2030 (Patel and Dean 2025). Where will these workers come from?

Fourth Problem: AI Scaling is Hitting a Wall

The strategic bet of leading AI firms that Generative AI can be achieved by building ever more data centers and using ever more chips is already exhibiting diminishing returns. It is the wrong strategy, since LLMs are not constructed on proper and robust world models (Marcus 2025b), but instead are built to autocomplete, based on sophisticated pattern-matching (Zhao *et al*. 2025). The AI industry has focused on training their generic LLMs on lots and lots of examples, and then assuming that one will somehow be able to do induction to things one has not seen before using just the statistics of what one has already learned. The hope of AI scaling was that the abilities to generalize and abstract would come out of the statistics, but it has not worked, and will not work, as well as people had hoped and is now hitting a wall. Zhao *et al*. (2025), in their paper "*Is Chain-of-Thought Reasoning of LLMs a Mirage? A Data Distribution Lens*", conclude that "LLMs are not principled reasoners, but rather sophisticated simulators of reasoning-like text." Generic LLMs will therefore continue to make errors and hallucinate (Tauman Kalai *et al*. 2025), especially when used outside their training data. AI products are never going to actually work right and will continue to be untrustworthy.[20] Cognitive scientist and AI expert Gary Marcus (2025a) offers this as a summary:

> "As a fantastic, souped-up version of autocomplete, AI coding tools somewhat increase productivity. They can help a coder learn a new API, or maybe even a new programming language. But current AI tools don't replace an understanding of debugging, nor an understanding of system architecture, nor an understanding of what clients want. […] The idea that coders (and more generally, software architects) are on their way out is absurd. AI will be a tool to help people write code, just as spell-checkers are a tool to help authors write articles and novels, but AI will not soon replace people who understand how to conceive of, write, and debug code."

[20] See also Vafa *et al.* (2025).

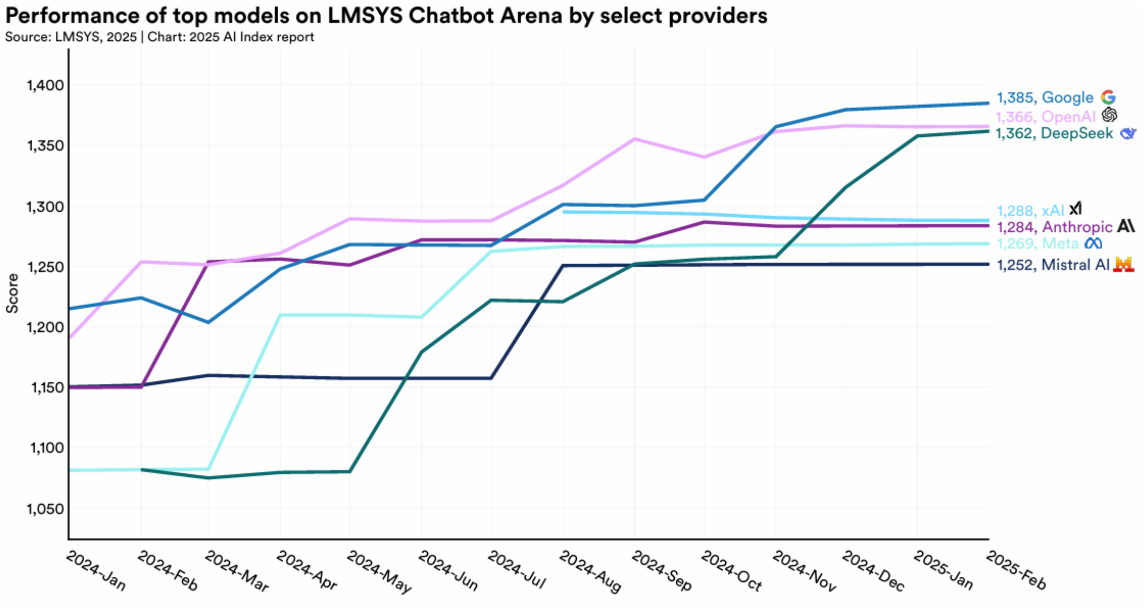
Or, think of LLMs, as Bender and Hanna (2025) suggest, as "synthetic text-extruding machines". "Like an industrial plastic process," they explain, text databases "are forced through complicated machinery to produce a product that looks like communicative language, but without any intent or thinking mind behind it". The same is true of other "generative" AI models that spit out images and music. They are all, the authors say, "synthetic media machines" – or, as I like to call them, giant plagiarism machines. "Both language models and text-to-image models will out-and-out plagiarize their inputs," the authors write.

As training and inference costs of AI scaling continue to outpace revenues in the AI industry — as argued above — the improvements in performance resulting from scaling are diminishing (Kedrosky 2025b). AI models have hit data saturation, where additional data yield diminishing returns. At this stage, the model has already learned most of the significant patterns, and the latest, significantly more expensive test-time computing strategy will only marginally improve performance — throwing more compute at the problem gets you, say, a 0.3% improvement in some performance indicators that costs $50 million and three months of GPU time. Simply throwing more compute at a problem during inference no longer guarantees better or more efficient results.

Rising training and inference costs, directly due to scaling, must eventually lead to higher user charges — and as users will have to pay more for AI services, they will more critically look around for "good value for their money'. Here, the problems for American AI companies are just starting. Chinese LLMs, such as (open-source) DeepSeek and Alibaba Cloud, achieve performance levels comparable to leading American AI models (**Figure 13**) but offer services at considerably lower prices (Stanford University 2025).

The user cost per 1 million tokens is $0.14 for DeepSeek R1 compared to $7.50 for ChatGPT o1. To illustrate the difference: A news website that uses AI for automated article summaries and processes 500 million tokens per month, would thus pay $70 for DeepSeek R1 versus $3,750 for ChatGPT o1. The cost difference is considerable and is increasingly difficult to justify in terms of superior performance of the OpenAI tool.

Figure 13


Performance of top models on LMSYS Chatbot Arena by select providers

Source: LMSYS, 2025 | Chart: 2025 AI Index report

Figure 2.1.39

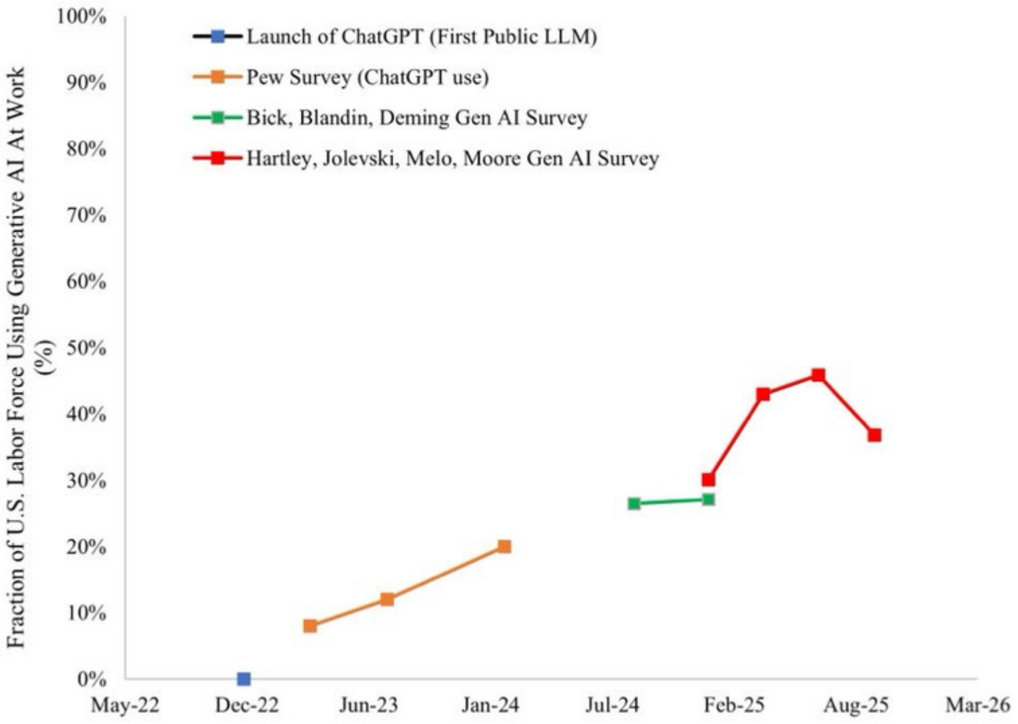
Source: Stanford University Human-Centered Artificial Intelligence, *Artificial Intelligence Index Report 2025*, Figure 2.1.39.

Disappointment concerning the usefulness of AI tools in businesses is growing and AI use has begun to decline (see **Figures 10** and **14**).[21] As mentioned earlier, a report from the *MIT Media Lab* found that 95% of organizations see no measurable return on their investment in these technologies (Challapally *et al.* 2025). According to research from *BetterUp* and *Stanford Social Media Lab*, around 40% of the 1,150 full-time U.S. desk workers who participated in their survey in September 2025 stated that they received AI-generated workslop in the last month; it is happening across industries but is especially prominent in professional services and technology.

[21] TechCrunch reported on October 17, 2025, that "ChatGPT's mobile app is seeing slowing download growth and daily use" based on data from Apptopia.

Figure 14
Fraction of the U.S. Labor Force Using Generative AI Tools At Work

Figure 2: Fraction of U.S. Labor Force Using Generative AI Use At Work



Notes: First survey (orange) from Pew Charitable Trusts (2023-2024). Second survey (green) from Bick, Blandin, and Deming (2024). Third survey (red) from this paper (Hartley, Jolevski, Melo, and Moore (2025))

Source: Hartley, Jolevski, Melo and Moore (2024).

Workslop refers to “AI-generated work content that masquerades as good work, but lacks the substance to meaningfully advance a given task.” Workslop has to be corrected and improved and the annual cost of this time waste for a 10,000-person company is estimated to equal $9 million.

A study (Becker *et al*. 2025) from the nonprofit *Model Evaluation and Threat Research* (METR) finds that in practice, programmers, using early 2025-AI-tools, are actually slower when using AI assistance tools, spending 19 percent more time when using GenAI than when actively coding by themselves. Programmers spent their time on reviewing AI outputs, prompting AI systems, and correcting AI-generated code. Using AI is also introducing new — unintended and significant — security risks into software development, which increases the need for tech workers to check software security. Untrained programmers writing and fixing code by describing what they want to an AI bot, may not only be introducing errors into their code, but also may be self-sabotaging by introducing severe cybersecurity risks to their code (see Marcus and Hamiel 2025).

The hardwired inclination to hallucinate (Metz and Weise 2025) limits the usefulness of AI in high-stakes activities such as healthcare, education and finance. According to recent research, chatbots spread false claims when prompted with questions about controversial news topics 35% of the time — almost double the 18% rate of a year ago (NewsGuard 2025). "The worst performer was Inflection AI, which provided false claims to news prompts 57% of the time. The rate for Perplexity was 47% and for Meta and ChatGPT it was 40%" (Woollacott 2025).[22] Potential liabilities resulting from the harm done by the decisions of autonomous unsupervised AI tools are simply too large in these high-stake activities — and this will restrict the adoption of and reliance on such AI tools.

For example, targeted AI tools are not and will not replace radiologists (Mousa 2025), because these tools can only diagnose abnormalities that are common in their (hospital-specific clinical) training data, but performance deteriorates significantly when the tools are used outside of their training domain (meaning, in other hospitals). AI tools tend to generate excessive 'false positives', which led to patients being called back more often. In contrast, having a double reading by two radiologists caught more cancers while slightly lowering callbacks. Health-cost insurers often exclude damages due to diagnoses generated autonomously by software, because a broken algorithm can harm many patients at once (Mousa 2025). Likewise, a survey by *National Nurses United* (NNU), the nation's largest union of registered nurses, found that AI technology often contradicts and undermines nurses' own clinical judgment and threatens patient safety (Bender and Hanna 2025). More generally, AI slop erodes scientific inquiry and scholarly discussion by design (Guest *et al*. 2025), leaving the door open to pseudoscience, exclusion, and surveillance. AI tools may therefore also poison scientific enquiry, when researchers refuse to pay attention.

[22] A recent study from *Columbia Journalism Review's Tow Center for Digital Journalism* tested eight AI-driven search tools by providing direct excerpts from real news articles and asking the models to identify each article's original headline, publisher, publication date, and URL. They find that the AI models incorrectly cited sources in more than 60% of these queries, raising serious concerns about their reliability in correctly attributing news content. "Perplexity provided incorrect information in 37 percent of the queries tested, whereas ChatGPTSearch incorrectly identified 67 percent (134 out of 200) of articles queried. Grok 3 demonstrated the highest error rate, at 94 percent" (source: Edwards 2025).

In an ironic twist, the supply of AI-slop will only increase in future, because due to the lack of 'authentic training data' (Al-Sibai 2024), LLMs will increase their input of 'synthetic' AI-generated artificial data — an incredible act of self-poisoning (Shumailov *et al*. 2024). The more AI-slop these models ingest, the greater the likelihood that their outputs will be junk: the "garbage-in, garbage-out" (GIGO) principle does hold. AI systems, which are trained on their own outputs, gradually lose accuracy, diversity, and reliability. This occurs because errors compound across successive model generations, leading to distorted data distributions and irreversible defects in performance. Veteran tech columnist Steven Vaughn-Nichols warns that "we're going to invest more and more in AI, right up to the point that model collapse hits hard and AI answers are so bad even a brain-dead CEO can't ignore it."

Part of the problem is that the integrity of AI bots, if at all existent, is compromised not just by the data poisoning (outlined above), but also because the direction of AI development is driven by corporate greed. As a result, AI tools are developed to lay-off workers, turn them into gig workers while putting them under corporate surveillance. Bender and Hanna (2025) recount, for example, how the National Eating Disorders Association in the US replaced their hotline operators with a chatbot days after the former voted to unionize. AI algorithms work very successfully to help landlords push the highest possible rents on tenants. Likewise, the health insurance industry uses AI automation and predictive technologies to systematically deny patients coverage for necessary medical care. AI also works for the military: defense company Anduril builds autonomous drones, virtual reality headsets, and other AI-powered technologies for the U.S. military. And private equity firms are hiring AI people to go through the companies they own and see how these should be restructured.[23] AI bots are used to maximize clicks and views — turning social media in giant misinformation machines.

Elon Musk could not care less: his AI chatbot Grok is actively generating AI-slop and fake information, for example, telling users that Musk is "in better shape" than LeBron James, "funnier" than Jerry Seinfeld, smarter than Albert Einstein and more handsome than François Civil. It has spread false information about the November 2015 Paris attack and told a user that Zyklon B — the deadly product used in the gas chambers of Nazi Germany's

[23] As pointed out by Thomas Ferguson, there is a growing overlap between firms and investors in high tech, defense and finance — and the political system (see French 2025).

extermination camps — was in fact designed for "disinfection" rather than for mass murder — language long associated with Holocaust denial. According to *Le Monde* (November 24, 2025), Grok is first and foremost a disinformation machine. Problem: other LLMs will 'eat' Grok's output and get infected as well (Woollacott 2025).

The degree of self-poisoning will grow exponentially, also because AI's test-time scaling strategy requires ever more data, authentic as well as machine-generated slop. It is not a coincidence that AI firms need more and more human workers to keep their AI models on the road. AI-firms including Google employ so-called 'AI-raters', tens of thousands of workers who use their judgment to moderate content and evaluate the responses generated by the AI bots and help chatbots sound more human (Bansal 2025ba, 2025b).[24] When checking AI responses, these AI-raters, often based in the Global South, also try their best to ensure that a chatbot does not spout inaccurate, unethical or harmful information (Bansal 2025a).[25] Many raters avoid using generative AI and advise family members and friends to not buy newer phones that have AI integrated in them, to resist automatic updates if possible that add AI integration, and to not tell AI anything personal (Bansal 2025b). "We joke that [chatbots] would be great if we could get them to stop lying," said one anonymous AI-rater who has worked with Gemini, ChatGPT and Grok (Bansal 2025b). This is, however, impossible.

The insatiable demand for AI compute originates mostly from the biggest players in the AI industry (OpenAI, Microsoft, Meta, Amazon and Nvidia), which (again) shows the dangerous circularity of the AI industry. These firms buy compute for training their models and inference. Demand for paid services coming from outside the AI-ecosystem remains surprisingly small. As Ed Zitron (2024) points out, The Information reported that customers of Microsoft's 365 suite are barely adopting its (paid-for) AI-powered "Copilot" products, with somewhere between 0.1% and 1% of its 440 million seats paying for the features ((at $30 to $50 per person). The Information quotes one firm testing the AI features as saying that "most

[24] AI firms are replacing human content moderators by AI-driven content moderation, but they are doing it faster than their bots are actually learning the job. AI content moderation is failing 70% of the time. False positives pile up, predators exploit loopholes, and harmful content slips through. The internet becomes less safe, not more. See D'Anastasio (2025).

[25] In December 2024, "over 140 Facebook moderators in Kenya filed suit against Meta and its subcontractor, alleging severe PTSD, depression, and anxiety from repeated exposure to graphic content, including murders, suicides, and child abuse. In April 2025, a similar lawsuit in Ghana highlighted psychological distress among moderators exposed to extreme content, emphasizing that unsafe moderation practices are a global concern" (Fogleman 2025).

people don't find it that valuable right now," and others saying that "many businesses haven't seen [breakthroughs] in productivity and other benefits" and they're "not sure when they will." The leader in productivity and business software cannot seem to find a product that people will pay for, in part because the performance is not worth the costs.

In an intriguing move, Microsoft, which owns a 27% stake in OpenAI, has recently partnered with Anthropic (not OpenAI) to integrate Claude AI models into its MS-365 applications and Copilot assistant for enterprise use (O'Brien 2025). Microsoft is moving away from its alliance with OpenAI, as OpenAI is increasingly trying to secure its own cloud capacity through big deals with Oracle, SoftBank and others. Deloitte and IBM have also chosen Anthropic (not OpenAI) to accelerate the development of enterprise-ready AI (in 2025). Is OpenAI going to lose the coming battle over (B2B) market shares?

Conclusion: Seeking Wealth Before Honor

The data center investment boom in the U.S. is a bubble, without doubt.

And all bubbles pop — at some point. The wealthy 17th-century merchants and skilled craftsmen in Amsterdam, living in a country rapidly expanding its wealth and trade networks, invested more and more money into tulip bulbs, believing that the trade in these exotic flowers would make them a fortune. "He who considers the profits that some make every year from their tulips will believe that there is no better Alchemy than this agriculture," Jean Franeau, a 17th-century poet, wrote.[26] But overnight, in February 1637, the mania stopped, probably because of fears of oversupply, as the obligations for that season's bulbs quickly became worthless. However, no one in the Low Countries went bankrupt because of this crash, and no one drowned themselves in the Amsterdam canals (Goldgar 2018). Those who lost money in the February 1637 crash did so only notionally: they might not get paid later. But all involved were rich enough to bear the loss. The Dutch economy was left completely unaffected. The U.S. will not be as lucky, when the AI bubble pops.

[26] Quoted in Goldgar (2007, p. 198).

Nonetheless, Tulip Mania did cause great, non-economic burdens (as argued by Goldgar 2007). Tulip bulb sellers primarily sold to people they knew: neighbors; colleagues; clients; doctors; shopkeepers; and booksellers. When buyers, who had promised to purchase the tulips, refused to pay, this destroyed the trust in established relationships and it damaged the values underlying ordered Dutch society:

> "“What is worse than cheating or being false?” Pieter de Clopper [a contemporary] wrote. “Let us know ourselves as liars all.” What was wrong with the tulip trade was not riches, not commerce, but the setting aside of an ordered society based on trust: seeking ‘inconstant wealth before honor.’ That was a message from the middling ranks. But it was not a message to the poor, but a warning to themselves.”
> (Goldgar 2007, p. 304)

The unavoidable AI-data-center crash in the U.S. will be painful to the economy (unlike the Tulip crash), even if some useful technology and infrastructure will survive and be productive in the longer run. There will be a reassessment of which (customized) AI tools are still useful, and these will provide the foundation for further development — also as open-source initiatives. However, given the unrestricted greed of the platform and other Big Tech corporations, this will also mean that AI tools that weaken the labor conditions — in activities including the visual arts, education, health care and the media — will survive. Similarly, generative AI is already entrenched in militaries and intelligence agencies and will, for sure, get used for surveillance and corporate control. All the big promises of the AI industry will fade, but the many harmful uses of the technology will stick around. The immediate economic harm done will look rather insignificant compared to the long-term damage of the AI mania. The continuous oversupply of AI slop, LLM fabricated hallucinations, clickbait fake news and propaganda, deliberate deepfake images and endless machine-made junk, all produced under capitalism’s banner of progress and greed, consuming loads of energy and spouting tons of carbon emissions will further undermine the trust in and the foundations of America’s economic and social order.

If you are looking for a happy ending, please go and see a Walt Disney movie.

References

Aliber, R.Z., C.P. Kindleberger and R.N. McCauley. 2023. *Manias, Panics, and Crashes. A History of Financial Crises*. London: Palgrave Macmillan. https://doi.org/10.1007/978-3-031-16008-0

Al-Sibai, N. 2024. 'AI Companies Running Out of Training Data After Burning Through Entire Internet.' *Futurism*, April 1.

Al-Sibai, N. 2025. AI Models Show Signs of Falling Apart as They Ingest More AI-Generated Data'.' *Futurism*, May 30.

Arun, A. 2025. 'Bubble or Nothing: Data Center Project Finance.' *Report*, November. New York: Center for Public Finance. https://publicenterprise.org/wp-content/uploads/Bubble-or-Nothing.pdf

Bain & Company. 2025. *How Can We Meet AI's Insatiable Demand for Compute Power? Technology Report*, September 23. Link: https://www.bain.com/insights/how-can-we-meet-ais-insatiable-demand-for-compute-power-technology-report-2025/

Barla, N. 2025. 'Token Burnout: Why AI Costs Are Climbing and How Product Leaders Can Prototype Smarter.' *Adaline Labs*, August 13. https://labs.adaline.ai/p/token-burnout-why-ai-costs-are-climbing

Bansal, V. 2025a. 'How thousands of 'overworked, underpaid' humans train Google's AI to seem smart.' *The Guardian*, September 11. https://www.theguardian.com/technology/2025/sep/11/google-gemini-ai-training-humans

Bansal, V. 2025b. 'Meet the AI Workers Who Tell Their Friends and Family to Stay Away from AI.' *The Guardian*, November 22. https://www.theguardian.com/technology/2025/nov/22/ai-workers-tell-family-stay-away

Barnette, C. and M. Peterson. 2025. 'Are We in a Bubble? The AI Boom in Context.' BlackRock Advisor Center, November 11. Link: https://www.blackrock.com/us/financial-professionals/insights/ai-tech-bubble

Becker, J. N. Rush, E. Barnes and D. Rein. 2025. 'Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity.' July 10. METR Model Evaluation & Threat Research. https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/

Bender, E.M. and A. Hanna. 2025. *The AI Con. How to Fight Big Tech's Hype and Create the Future We Want*. London: Penguin Books. ISBN: 9781529949902

Bloomberg. 2025. 'AI Data Centers Are Sending Power Bills Soaring.' September 30. https://www.bloomberg.com/graphics/2025-ai-data-centers-electricity-prices/

Carvão, P. 2025. 'Is The AI Bubble Bursting? Lessons From The Dot-Com Era.' *Forbes*, August 21.

Challapally, A. *et al.* 2025. The GenAI Divide. State of AI in Business 2025. July. Project NANDA. Cambridge, Mass.: MIT.

Clairouin, O. 2025. 'Why Grok is First and Foremost a Disinformation Machine.' *Le Monde*, November 24.

Cooper, R. 2025. 'OpenAI Is Manoeuvring for a Government Bailout.' *The American Prospect*, November 7.

Daly, L. 2025. 'The Magnificent Seven's Market Cap Vs. the S&P 500.' *The Motley Crew*, October 18.

D’Anastasio, C. 2025. ‘AI Is Replacing Online Moderators, But It's Bad at the Job.’ *Bloomberg*, August 22.

De Palmas, C. 2025. ‘All Eyes on Nvidia: Could This Week’s Earnings Define the AI Investment Era?’ *FXEmpire*, November 17. Link: https://www.fxempire.com/forecasts/article/all-eyes-on-nvidia-could-this-weeks-earnings-define-the-ai-investment-era-1561715

Edwards, ‘AI Search Engines Cite Incorrect News Sources at an Alarming 60% Rate, Study Says.’ *Ars Technica*, March 13.

Elder, B. 2025. ‘How High Are OpenAI’s Compute Costs? Possibly a Lot Higher Than We Thought.’ *The Financial Times*, November 12.

Fitch, A. 2025. ‘Debt Is Fueling the Next Wave of the AI Boom.’ *The Wall Street Journal*, September 29.

French, N. 2025. ‘Tech Capital Is Dominating American Politics. An Interview with Thomas Ferguson.’ Jacobin, September 17. https://jacobin.com/2025/09/tech-capital-american-politics-ferguson

Fogleman, J. 2025. ‘Humans Behind the Screens: Duty of Care in the Age of AI Content Moderation.’ September 28. https://john.fogleman.law/articles/humans-behind-the-screens-duty-of-care-in-the-age-of-ai-content-moderation/

Furman, J. 2025. Twitter, September 27. https://x.com/jasonfurman/status/1971995367202775284

Galbraith, J.K. 2009/1955 *The Great Crash 1929*. Harper Business. **ISBN**: 9780547248165

Goldgar, A. 2007. *Tulipmania – Money, Honor and Knowledge in the Dutch Golden Age*. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226301303.001.0001

Goldgar, A. 2018. ‘Tulip Mania: The Classic Story of a Dutch Financial Bubble is Mostly Wrong.’ *The Conversation*, February 12. https://doi.org/10.64628/AB.sanjfvkam

Goldman Sachs. 2025. ‘Bridging the Gap: How Smart Demand Management Can Forestall the AI Energy Crisis.’ August 11. https://www.goldmansachs.com/what-we-do/goldman-sachs-global-institute/articles/smart-demand-management-can-forestall-the-ai-energy-crisis

Goode, L. and W. Knight. 2025. ‘Meta, Google, and Microsoft Triple Down on AI Spending.’ *WIRED*, October 29. Link: https://www.wired.com/story/microsoft-google-meta-2025-earnings/

Guest, O. *et al*. 2025. ‘Against the Uncritical Adoption of 'AI' Technologies in Academia.’ https://zenodo.org/records/17065099

Hammond, G., H. Murphy and J. Fontanella-Khan. 2025. ‘Elon Musk’s xAI Nears $230bn Valuation in Fundraising Deal.’ *The Financial Times*, November 19.

Hartley, J., F. Jolevski, V. Melo and B. Moore. 2024. ‘The Labor Market Effects of Generative Artificial Intelligence.’ https://dx.doi.org/10.2139/ssrn.5136877

Heath, A. 2025. ‘I Talked to Sam Altman About the GPT-5 Launch Fiasco.’ *The Verge*, August 15.

Humlum, A. and E. Vestergaard. 2025. ‘Large Language Models, Small Labor Market Effects.’ NBER Working Paper 33777. Cambridge, Mass.: National Bureau of Economic Research. https://doi.org/10.3386/w33777

Jaźwińska, K. and A. Chandrasekar. 2025. 'AI Search Has a Citation Problem.' *Columbia Journalism Review*, March 6. https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php

Johnson, C. 2025. 'How AI Data Centers Are Driving a $42 Billion Onsite Generation Boom.' Energy Industry Insights from Avanza Energy, September 9. https://avanzaenergy.substack.com/p/how-ai-data-centers-are-driving-a

Kedrosky, P. 2025a. 'SPVs, Credit, and AI Datacenters.' June 30. https://paulkedrosky.com/weekend-reading-plus-spvs-meta-and-fiber-buildout-2-0/

Kedrosky, P. 2025b. 'Honey, AI Capex is Eating the Economy.' July 18. https://paulkedrosky.com/honey-ai-capex-ate-the-economy/

Kimball, S. 2025. 'Nvidia Shares Rise after CEO Huang Says AI Computing Demand is Up 'Substantially'.' CNBC, October 8. https://www.cnbc.com/2025/10/08/jensen-huang-nvidia-computing-demand.html

Levine, M. 2025. 'The Perfect AI Startup.' Bloomberg, September 29.

Lichtenberg, . 2025. 'One of the World's Most Popular Economists on Why AI is 'Undoubtedly Going to Crash': It's Built Off 'Digital Lettuce'—and the U.S. will be Just Fine Anyway.' *Fortune*, November 20.

Lin, B. 2025. 'Anthropic and IBM Partner in Bid for AI Business Customers.' *The Wall Street Journal*, October 7.

Lishawa, J. 2025. 'The Cost of Context: The Exponential Growth in Tokens.' *Illuminem Voices*, November 3. https://illuminem.com/illuminemvoices/the-cost-of-context-the-exponential-growth-in-tokens

Mallipatna, P. 2025. 'The AI Leverage Economy: When Compute Becomes Collateral.' *Pramodh's Substack*, October 20. https://pramodhmallipatna.substack.com/p/the-ai-leverage-economy-when-compute?triedRedirect=true

Marcus, G. 2025a. 'AI Coding Fantasy meets Pac-Man.' *Marcus on AI*, Substack, March 12. AI Coding Fantasy meets Pac-Man - by Gary Marcus

Marcus, G. 2025b. 'Generative AI's crippling and widespread failure to induce robust models of the world.' *Marcus on AI*, Substack, June 25. Generative AI's crippling and widespread failure to induce robust models of the world

Marcus, G, 2025c. 'Could China Devastate the US Without Firing a Shot? *Marcus on AI*, Substack, October 27. https://garymarcus.substack.com/p/could-china-devastate-the-us-without

Marcus, G. and N. Hamiel. 2025. 'LLMs + Coding Agents = Security Nightmare.' *Marcus on AI*, Substack, August 17. https://garymarcus.substack.com/p/llms-coding-agents-security-nightmare

McKinsey & Co. 2024. 'How Data Centers and the Energy Sector Can sate AI's Hunger for Power.' https://www.mckinsey.com/industries/private-capital/our-insights/how-data-centers-and-the-energy-sector-can-sate-ais-hunger-for-power

McPhee, A. 2025. 'Thinking Machines Lab — The $10 Billion Startup With No Product.' Medium, April 17.

Metz, C. and K. Weise. 2025. ‘A.I. is getting more powerful, but its hallucinations are getting worse.’ The New York Times, May 5. https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html

Milmo, D. 2025. ‘Boom or Bubble? Inside the $3tn AI Datacentre Spending Spree.’ *The Guardian*, November 2.

Mims, C. 2025. ‘When AI Hype Meets AI Reality: A Reckoning in 6 Charts.’ *The Wall Street Journal*, November 14.

Moore, E. 2025. ‘Who Needs Revenue When You’re a Multibillion-Dollar AI Start-Up?’ *The Financial Times*, March 13.

Morales, J. 2025. ‘J.P. Morgan Calls Out AI Spend, Says $650 Billion in Annual Revenue Required to Deliver Mere 10% Return on AI Buildout — Equivalent to $35 Payment From Every iPhone User, Or $180 From Every Netflix Subscriber 'In Perpetuity'.’ *Tom’s Hardware*, November 11.

Mousa, D. 2025. ‘AI Isn't Replacing Radiologists.’ Substack Understanding AI, October 1. https://www.understandingai.org/p/ai-isnt-replacing-radiologists

Mutjaba, H. 2025. ‘NVIDIA’s Discrete GPU Market Share Swells To 94%, AMD Drops To 6% In Q2 2025, 27% Increase In AIB Shipments.’ *WCCFTECH*, September 2.

NewsGuard. 2025. ‘AI False Information Rate Nearly Doubles in One Year.’ September 4. https://www.newsguardtech.com/ai-monitor/august-2025-ai-false-claim-monitor/

Nvidia. 2025. ‘OpenAI and NVIDIA Announce Strategic Partnership to Deploy 10 Gigawatts of NVIDIA Systems.’ Nvidia website, September.

Nvidia. 2025b. ‘Building the 800 VDC Ecosystem for Efficient, Scalable AI Factories.’ Nvidia website, October 13.

O’Brien, M. 2025. ‘Microsoft Partners with Anthropic and Nvidia in Cloud Infrastructure Deal.’ Associated Press, November 18. https://apnews.com/article/microsoft-ignite-anthropic-nvidia-a3e4d6ba75f475eb130d91c81e522f93

OpenAI. 2025a. ‘OpenAI and NVIDIA Announce Strategic Partnership to Deploy 10 Gigawatts of NVIDIA Systems.’ OpenAI website, September 2025.

OpenAI 2025b. ‘AMD and OpenAI Announce Strategic Partnership to Deploy 6 Gigawatts of AMD GPUs.’ OpenAI website, October 6.

Patel, D. and R. Dean. 2025. ‘Thoughts on the AI Buildout.’ Blog. https://www.dwarkesh.com/p/thoughts-on-the-ai-buildout

Platt, E., O. Barnes and H. Murphy. 2025. ‘Meta Seeks $29bn From Private Credit Giants to Fund AI Data Centres.’ *The Financial Times*, June 27.

Prabhu, A. 2025. ‘$2B Raise at $32B Valuation: 5 Facts OpenAI Co-founder’s Safe Superintelligence. Can It Outperform Anthropic and Google Deepmind?’ Techfundingnews.com, April 15.

Rafetto, C. 2025. ‘Willis Report: Construction Struggles with AI Data Centers and Labor Shortages.’ ConstructionOwners.com, March 24.

Reuters. 2025. ‘Peter Thiel's Fund Offloaded Nvidia Stake in Third quarter, Filing Shows.’ *Reuters*, November 17.

Rizvi, M. 2025. ‘J.P. Morgan Warns AI Industry Needs $650 Billion Annually for 10% Returns.’ *Tech Searchers*, November 12. Link: https://techsearchers.com/jp-morgan-ai-industry-warning-2030/

Robb, D. 2025. ‘Data Centers Bypassing the Grid to Obtain the Power They Need.’ *Data Center Knowlegde*, May 1. https://www.datacenterknowledge.com/energy-power-supply/data-centers-bypassing-the-grid-to-obtain-the-power-they-need

Rogers, K. 2025. ‘AI Data Center Boom Has to Contend With Realities of Tough Labor Market.’ CNBC, September 30.

Roubini, N. 2025. ‘Tech Trumps Tariffs: Why US Exceptionalism Will Last.’ *The Financial Times*, November 25.

Saini, M. and N. Nishant. 2025. ‘Wall Street Heavyweights Flag Risk of Pullback in Equity Markets.’ *Reuters*, November 14.

Saplakoglu, Y. 2024. ‘How AI Revolutionized Protein Science, but Didn’t End It.’ *QuantaMagazine*, June 26. https://www.quantamagazine.org/how-ai-revolutionized-protein-science-but-didnt-end-it-20240626/

Schmidt, E. and S. Xu. 2025. ‘Silicon Valley is drifting out of touch with the rest of America.’ *The New York Times*, August 19. https://www.nytimes.com/2025/08/19/opinion/artificial-general intelligence-superintelligence.html

Seetharaman, D. and K. Hu. 2025. ‘Altman Touts Trillion-Dollar AI Vision As OpenAI Restructures To Chase Scale.’ *Reuters*, October 29.

Sen, A. 2025. ‘US Hedge Funds Trim Stakes in ‘Magnificent Seven’ Stocks in Third Quarter.’ *Reuters*, November 15.

Shaw, D.J.. 2024. ‘Roubini Warns of ‘Secular Stagflation’ Era Ahead.’ Etf.com, December 16. https://www.etf.com/sections/news/roubini-warns-secular-stagflation-era-ahead

Shiller, R. 2025. Shiller Database. Link: https://shillerdata.com/

Shojaee, P. *et al*. 2025. ‘The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity.’ Apple Co. June. https://machinelearning.apple.com/research/illusion-of-thinking
https://doi.org/10.70777/si.v2i6.15919

Shumailov, I., Shumaylov, Z., Zhao, Y. *et al.* 2024. ‘AI Models Collapse When Trained on Recursively Generated Data.’ *Nature* **631**, 755–759. https://doi.org/10.1038/s41586-024-07566-y

Sløk, T. 2025a. ‘Similar Contribution to GDP Growth From Consumer Spending and Data Center Investments.’ *The Daily Spark*, Apollo Global Management. August 18. https://www.apolloacademy.com/similar-contribution-to-gdp-growth-from-consumer-spending-and data-center-investments/

Sløk, T. 2025b. ‘AI Adoption Rate Trending Down for Large Companies.’ *The Daily Spark*, Apollo Global Management. September 7. https://www.apolloacademy.com/ai-adoption-rate-trending down-for-large-companies/

Smith, I., E. Herbert and R. Rees. 2025. ‘Fund Managers Warn AI Investment Boom Has Gone Too Far.’ *The Financial Times*, November 18.

Smith, Y. 2025. ‘Yet More AI Bubble Worries, Now on Debt Side, with Wall Street Journal Featuring AI Datacenter Borrowing “Frenzy”.’ *Naked Capitalism*, November 17.

Spark, Z. 2025. ‘The LLM Cost Paradox: How “Cheaper” AI Models Are Breaking Budgets.’ *iKangai.com*, August 21.

Stanford University. 2025. *The 2025 AI Index Report*. Human-Centered Artificial Intelligence. https://hai.stanford.edu/ai-index/2025-ai-index-report

Storm, S. 2025. ‘The AI Bubble and the U.S. Economy: How Long Do ‘Hallucinations’ Last?’ *INET Working Paper 240*. New York: Institute for New Economic Thinking.

Szyzka, E. 2025. ‘Future AI Bills of $100k/Yr Per Dev.’ *Kilo Code Blog*, August 8. https://blog.kilo.ai/p/future-ai-spend-100k-per-dev?ref=wheresyoured.at

Tangermann, V. 2025. ‘Top Economist Warns That AI Data Center Investments Are “Digital Lettuce” That’s Already Starting to Wilt.’ *Futurism*, November 21.

Tauman Kalai, A., O. Nachum, S.S. Vempala and E. Zhang. 2025. ‘Why Language Models Hallucinate.’ September 4, https://arxiv.org/abs/2509.04664

Techcrunch. 2025. ‘ChatGPT’s Mobile App Is Seeing Slowing Download Growth and Daily Use, Analysis Shows.’ October 17. https://techcrunch.com/2025/10/17/chatgpts-mobile-app-is-seeing-slowing-download-growth-and-daily-use-analysis-shows/

The Small Cap Strategist. 2025. ‘J.P. Morgan's $7 Trillion AI Warning.’ Substack. Link: https://tscsw.substack.com/p/jp-morgans-7-trillion-ai-warning

Thornhill, J. 2025. ‘Brace for a crash before the golden age of AI.’ *The Financial Times*, August 21. https://www.ft.com/content/a76f238d-5543-4c01-9419-52aaf352dc23

Torene, S. 2025. ‘Understanding the Impact of Increasing LLM Context Windows.’ Meibel, April 24. https://www.meibel.ai/post/understanding-the-impact-of-increasing-llm-context-windows

Tracy, M. 2025, ‘Moody's Flags Risk in Oracle's $300 Billion of Recently Signed AI Contracts.’ *Reuters*, September 18.

Tunguz, T. 2025. ‘OpenAI's $1 Trillion Infrastructure Spend.’ Link: https://tomtunguz.com/openai-hardware-spending-2025-2035/#fn:5

Vafa, K. *et al*. 2025. ‘What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models.’ https://doi.org/10.48550/arXiv.2507.06952

Wall Street Journal. 2025. ‘OpenAI CFO Would Support Federal Backstop for Chip Investments.’ November 5. Link: https://www.wsj.com/video/openai-cfo-would-support-federal-backstop-for-chip-investments/4F6C864C-7332-448B-A9B4-66C321E60FE7

Weil, J. 2025. ‘Is the Flurry of Circular AI Deals a Win-Win—or Sign of a Bubble?’ *The Wall Street Journal*, October 22.

Wigglesworth, R. 2025a. ‘How Many ‘Bragawatts’ Have the Hyperscalers Announced So Far?.’ *The Financial Times*, November 4.

Wigglesworth, R. 2025b. ‘“The Global Data Centre and AI Build-Out Will Be an Extraordinary and Sustained Capital Markets Event”. *The Financial Times*, November 11.

Wirz, M. 2025. ‘Three AI Megadeals Are Breaking New Ground on Wall Street.’ *The Wall Street Journal*, November 11.

Wirz, M. and P. Rudegeair. 2025. ‘Wall Street Blows Past Bubble Worries to Supercharge AI Spending Frenzy.’ *The Wall Street Journal*, November 16.

Williams, S. 2024. ‘The Stock Market Has Crossed This Threshold 6 Times Since 1871 - and History Couldn't Be Clearer What Comes Next.’ *The Motley Fool*, December 15.

Woollacott, E. 2025. ‘AI Chatbots Are Feeding You More False Information Than Ever.’ *Forbes*, September 5. https://www.forbes.com/sites/emmawoollacott/2025/09/05/ai-chatbots-are-feeding you-more-false-information-than-ever/

WSJ Bold Names Podcast. 2025. ‘The World’s Tech Giants Are Running Out of Power. This CEO Plans to Deliver.’ November 14.

Yoon, J. 2025. ‘What If the AI Race Isn’t About Chips At All? ’ *The Financial Times*, November 12.

Zhao, C. *et al*. 2025. ‘Is Chain-of-Thought Reasoning of LLMs a Mirage? A Data Distribution Lens.’ https://doi.org/10.48550/arXiv.2508.01191

Zitron, E. 2024. ‘The Subprime AI Crisis.’ *Where's Your Ed At?* Substack. September 16. https://www.wheresyoured.at/subprimeai/

Zitron, E. 2025a. ‘Big Tech Needs $2 Trillion In AI Revenue By 2030 or They Wasted Their Capex.’ *Where's Your Ed At?* October 31. Link: https://www.wheresyoured.at/big-tech-2tr/

Zitron, E. 2025b. ‘Exclusive: Here's How Much OpenAI Spends On Inference and Its Revenue Share With Microsoft.’ *Where's Your Ed At?* November 12. Link: https://www.wheresyoured.at/oai_docs/