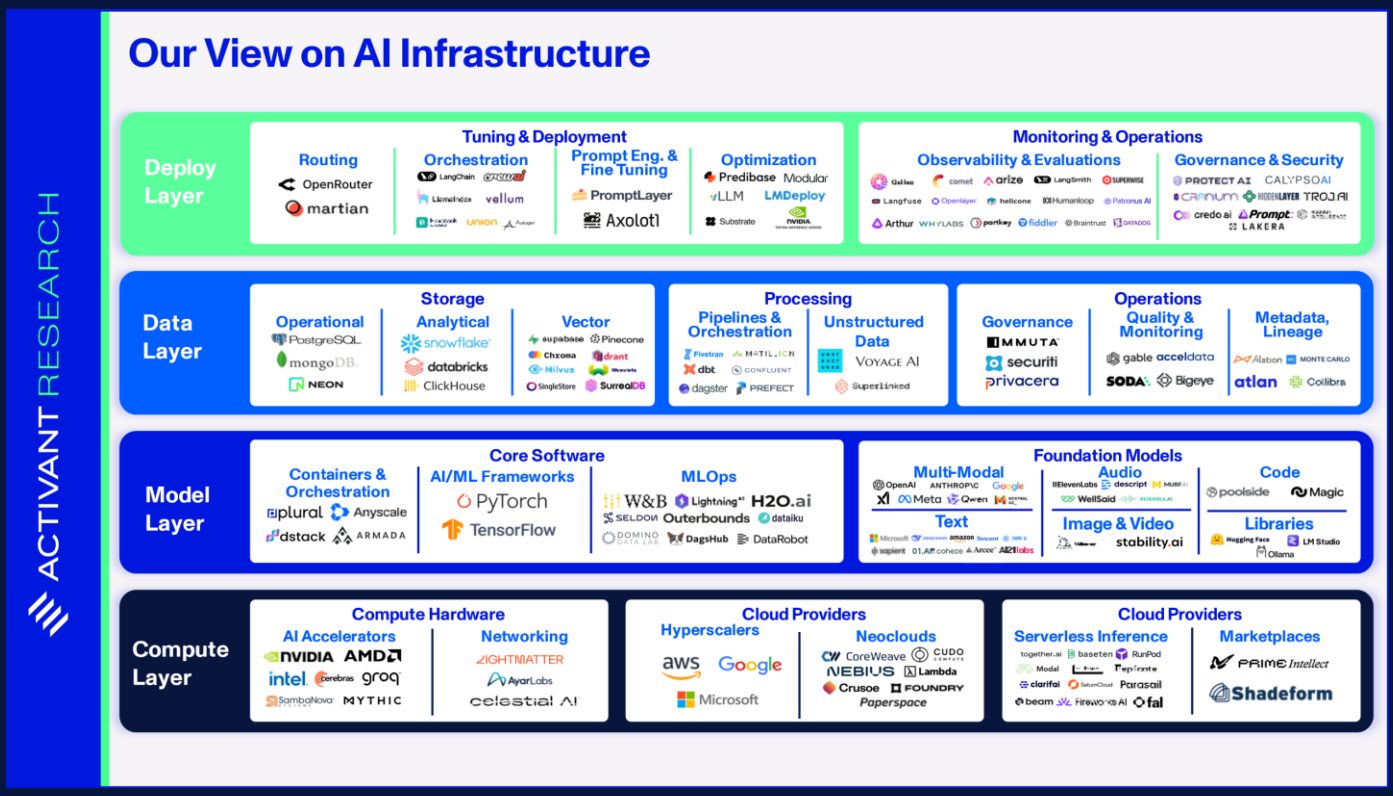
ACTIVANT RESEARCH

AI Infrastructure: Compute (1/4)

Sizing the market for AI Compute



Q1 2026

Steve Sarracino & Jono Vickery

Venture's biggest questions bubbling up

Five-trillion dollars: the estimated funding required for digital infrastructure and power underpinning the future of AI.[1] Billion-dollar deals for computing power feature in the news almost daily and datacenter campuses may soon cover the area of Manhattan. The industry is building AI infrastructure at a relentless pace. Everyone is asking the same question, "is this a bubble?".

This is arguably the most important question in the market today. If this furious build-out turned out to be speculative, "building too much, too quickly for use cases that have not arrived yet", a subsequent pop could trigger a crisis of confidence, leading to significant equity market losses and a tightness in private markets that could rival the 2022-2023 downturn. Further, without a handle on the true size of the infrastructure underpinning AI, its impossible to properly size the AI application market.

Our analysis, however, suggests the build-out is both rational and supported by highly favorable early fundamentals. What we're experiencing is a logical response to the core insight of modern AI: the more computing power you apply, the better the model (or system) performs, and that next step up in performance might unlock not just automating route work but a cure to cancer or nuclear fusion.

Further, demand for AI computing is not just theoretical. The hyperscale compute providers persistently describe demand as "significantly ahead of supply" and their own businesses as "compute starved". Companies like Google and Meta are deploying AI to billions of users and finding incremental revenue opportunities even at their $100bn+ revenue scale, almost immediately validating their headline-grabbing capex figures. Simultaneously, a new wave of AI-native startups is generating real revenue: going from $5bn to $47bn in ARR in the last year.[2]

The (general) rationality of this infrastructure buildout lays the foundation for the success of the entire AI infrastructure stack, from AI cloud compute all the way up to model observability. We believe that the AI infrastructure market is at the early stages of a structural expansion that will create an AI Cloud Computing market worth ~$450bn and support many significant businesses all the way up the stack.

AI's Mega-build: The What

Datacenters of old

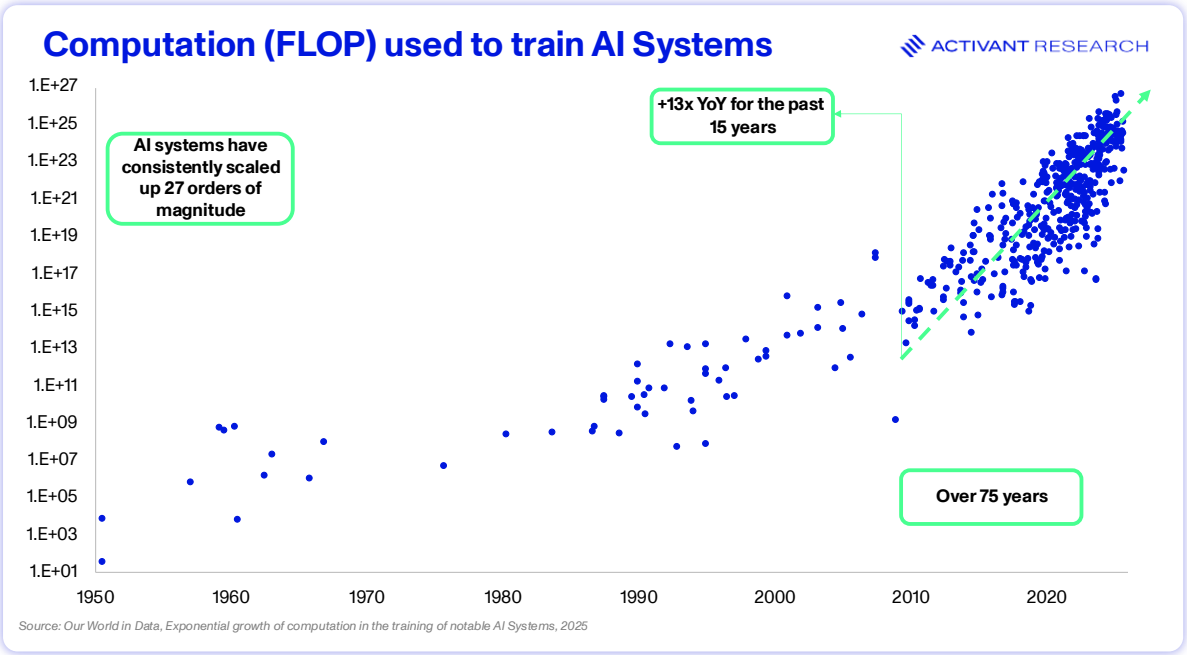
In Q3 2022, just before the release of ChatGPT, the tech industry was pretty good at building and managing datacenters. Global datacenter supply was over 40GW, enough capacity to consume the power of 32 million US homes or the entire power generation capacity of Norway.[3,4] AWS's cloud business was spinning off 26% operating margins as the largest cloud services provider on the planet[5]. Datacenter operators had become experts at the workloads that they needed to deal with, such as data storage and providing compute for deterministic systems. Many of the key innovations that shaped datacenters had become well established. For example:

1. **CPU-centricity:** There's a good reason that CPUs became the backbone of all modern computing. They are set up with a **small number of powerful cores** (typically 8 – 64), which makes them great for workloads that are sequential but complex, like running server operating systems, customer applications and complex database queries. Intel's hyperthreading, introduced in 2002 allowed CPUs to process multiple tasks simultaneously, cementing the CPU as the nucleus of the server room that could handle any task thrown at it.[6]
2. **Power efficiency:** CPU-based datacenters typically consume 5 – 15Kw of power per rack, equal to about 15 microwaves.[7] At these power densities, racks emit modest amounts of heat and through organization into hot and cold aisles, they can use standard air cooling methods, and the rack itself can be the major usage of floorspace.
3. **Hierarchical Architecture:** Traditional datacenters have been built using three layers (the hierarchy) to optimize for "north-south" traffic (in and out of the datacenter). Data flows physically up and down the hierarchy. Workloads must pass through an access layer, distribution layer and a core layer, perfectly suited to the client-server model of traditional computing. The demands for high speed networking are limited, so servers are networked with low-cost ethernet.

Effectively, years of datacenter innovation and optimization geared us towards **small, unpredictable, and heterogenous workloads associated with traditional computing**, where communication within the datacenter was far less critical. Then we had the ChatGPT moment.

Datacenter, meet AI

As we'll discuss in the next section, the primary insight driving modern AI are the scaling laws – whether it be during training, or at inference - **the more computing power you apply, the better the model performs.** This scaling law rips apart the traditional datacenter model.

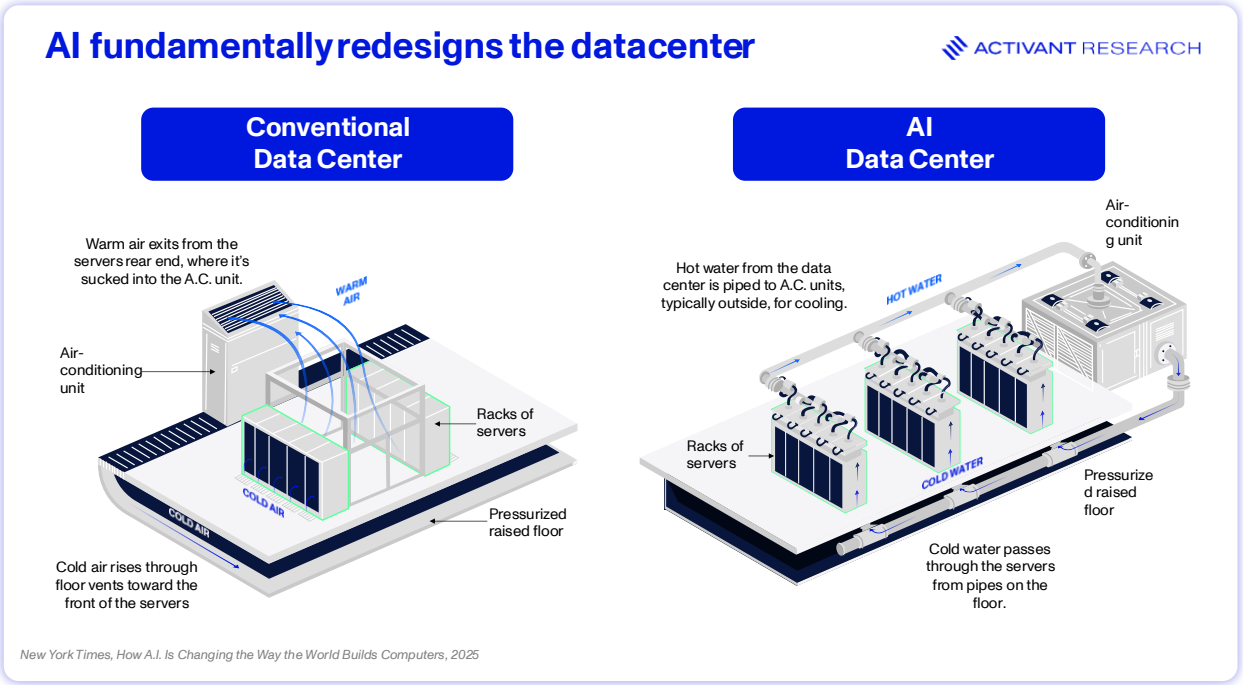


At its core, an AI model is composed of interconnected nodes, or **neurons**, organized in layers. Each connection between neurons has a corresponding **parameter** (a weight and a bias) that is adjusted during training. Leading models now have more than **100s of billions of parameters**, so training involves performing 100s of billions of calculations. However, these calculations are individually quite simple and can be performed simultaneously. So, the CPU, powerful but sequential in nature, would be a massive bottleneck to AI training. Rather, the GPU, architected with thousands of weaker cores, can perform the simple and parallelizable calculations of AI training in a fraction of the time.

These GPU-based server racks consume far more power and thus generate far more heat than CPU racks – 3x as much, and potentially ~50x as much when Nvidia's Rubin Ultra System is released.[8,9] At these power densities, air-cooled GPU server racks would melt and instead require liquid cooling. These liquid cooling systems take up significant floorspace and can't be stuffed into tightly packed CPU-heavy datacenters.

Further, training large models cannot be done on single GPUs, but rather thousands of GPUs chained together. The datacenter itself becomes the computer. These GPUs must constantly exchange data and synchronize their parameters, creating a massive volume of "east-west" (server-to-server) traffic. The traditional datacenter network, optimized for "north-south" traffic, becomes an immediate and crippling bottleneck, leaving expensive GPUs idle while they wait for data. To solve this, AI datacenters require a complete network redesign based on ultra-high-bandwidth, low-latency fabrics like NVIDIA's InfiniBand.

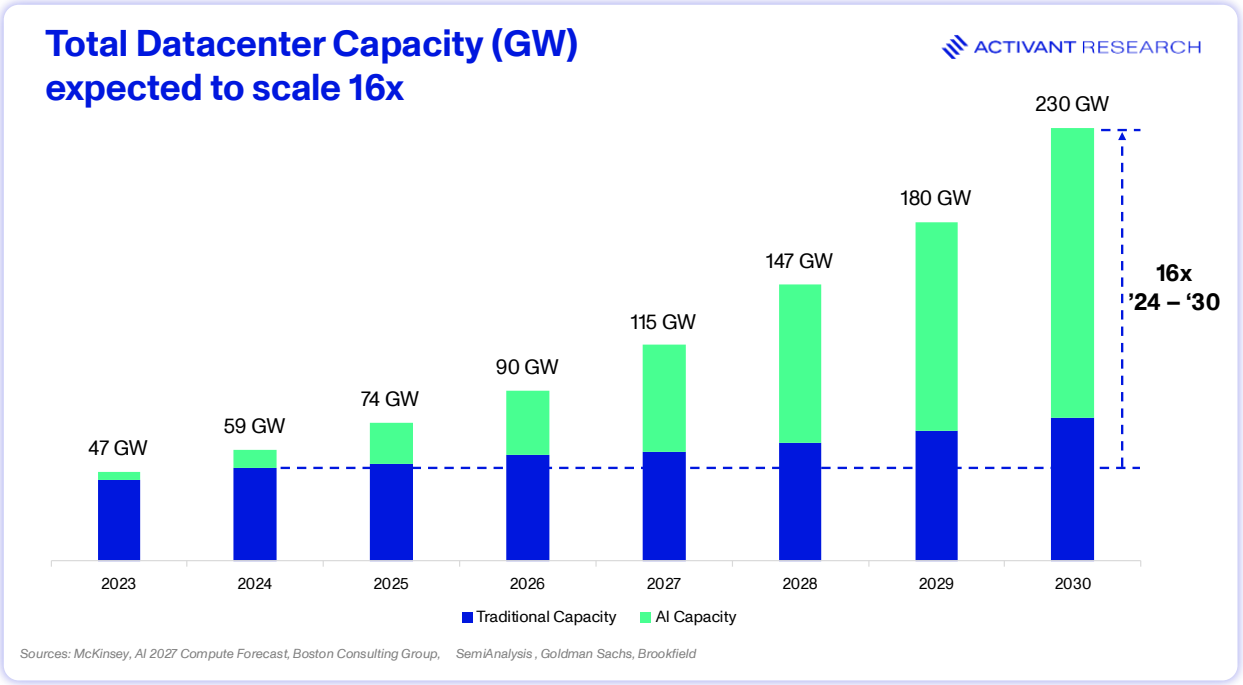
AI needs datacenters specialized for AI.



Time to invert

AI makes up only 30% of total datacenter capacity today, with close to 70% still tied to traditional workloads.[10] For AI to both continue improving and deeply penetrate the entire global economy, we see that ratio inverting. However, we clearly can't retrofit the existing datacenter infrastructure for AI. We're going to need entire datacenters that are specialized for AI – **architected around GPUs, not CPUs, liquid cooled and networked for ultra-low latency.**

The AI build that we're experiencing is about building these fundamentally new, specialized datacenters with unique demands around location, resource access, and, as the scaling laws will explain, **size**.

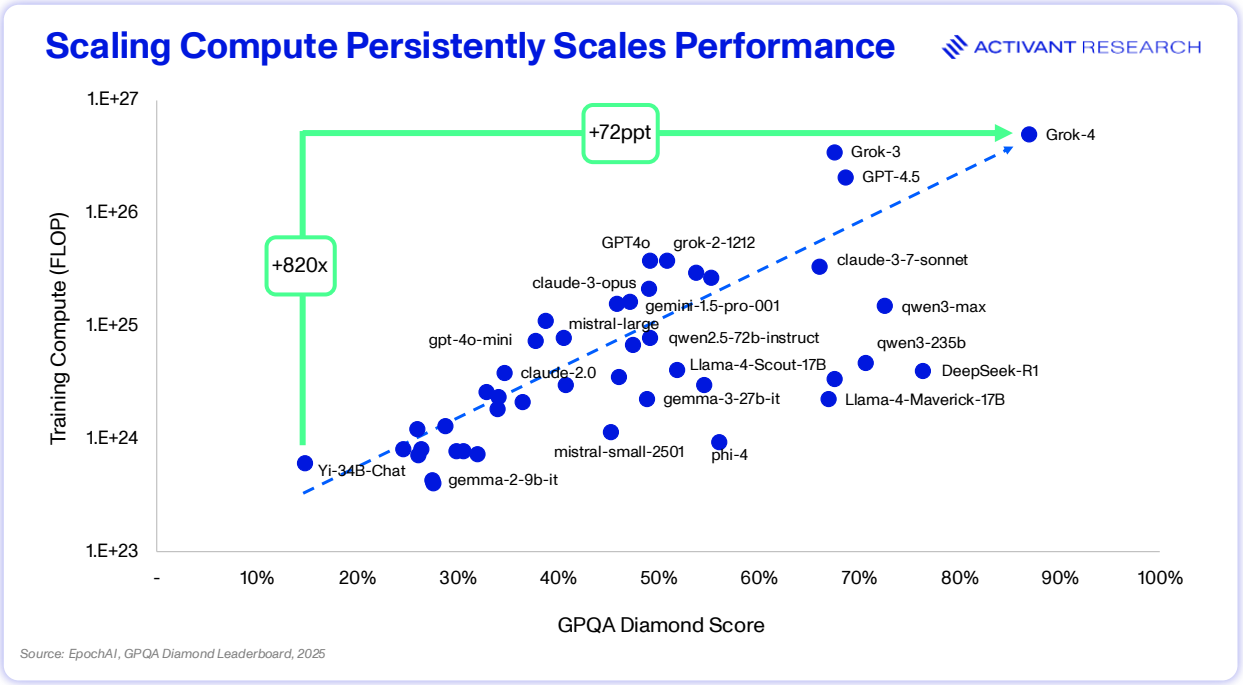


AI's Mega-Build: The Why

Bigger = Better: The Scaling Laws

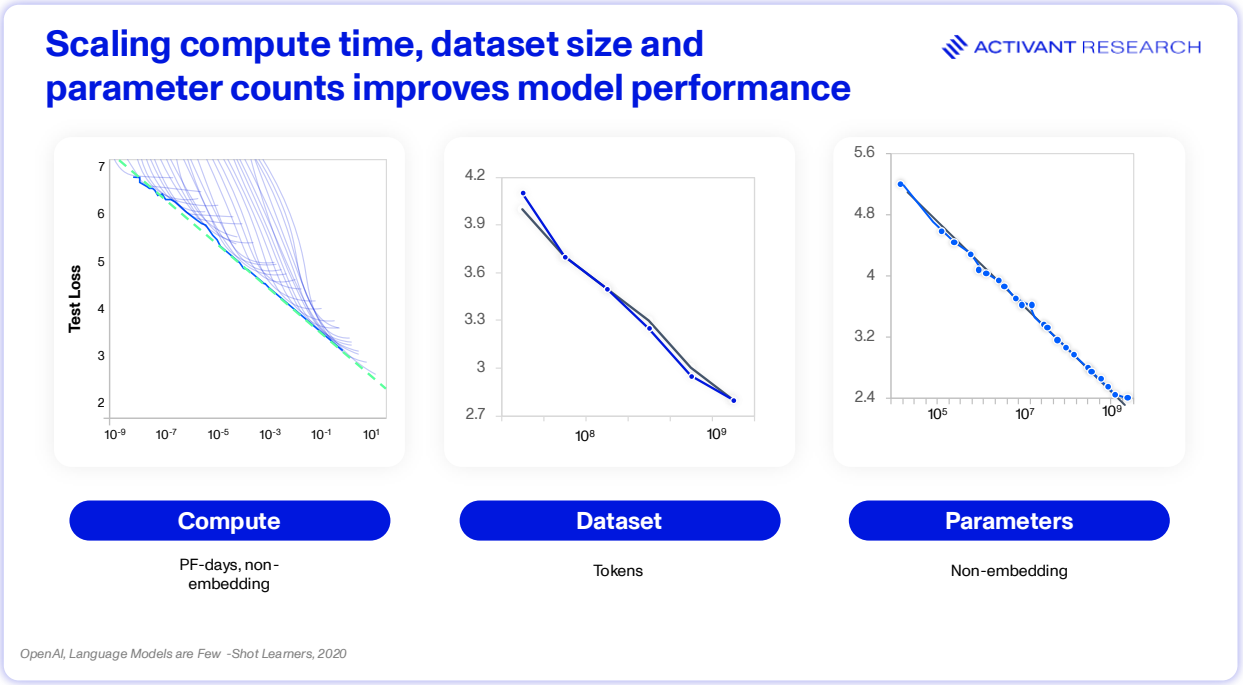
In 2017, Google quietly released the "Transformer paper", now the 7th most cited research paper of the 21st century.[11] Attention is all you need, as it was titled, was based on the insight that AI models should not use the last word to predict the next, but rather the entire sentence. The new architecture was far more parallelizable but also made models that were more *intelligent* – with the full context of the relevant text, models started to demonstrate *understanding*.

OpenAI's researchers found that by scaling an LLM up to 117 million parameters (using ~7,000 books as training data) they could train a model that not only dealt with language better, but started to solve problems in separate domains, outside of its training data. In other words, LLMs *generalized.* When OpenAI scaled GPT-2 to 1.5 billion parameters, it improved so much that the company founded on the principle of openness, was afraid to release it to the public.[12] To this day, models have demonstrated improvements for a given increase in data and compute – an observation so persistent it has been called a law – **the scaling law**.



Essentially, there are three key inputs to the scaling law:

1. **Model size:** or more technically, parameters, reflect the number of connections that the model can "learn" – more parameters equate to a deeper, more granular understanding of the training data.
2. **Data:** The size of the dataset. Providing the model with more data to make sense of improves its fundamental understanding. GPT-3 was trained on Common Crawl, an open source web-scrape database.[13]
3. **Training Time:** The length of time that the model is trained for. The longer the training run, the more the model parameters can be refined.



Scaling up compute may be capital intensive, but it is predictable and repeatable. Unfortunately, it demonstrates diminishing returns. Doubling compute does not provide a doubling of performance, rather, each **order of magnitude** increase in compute drives model performance. Note the scaling of GPQA performance above – it took an 820x increase in compute to move from a score of 15% to 87%. GPUs and algorithms enhance efficiency at ~2x per year combined, which leaves the 10x, 100x, and 1,000x scale-ups being driven by building more datacenters.[14] And that work is worth it, thanks to ***generalizability.***

Large language models were found to be able to solve problems not contained in their original training set. In October 2025, a Google model helped discover a potential new cancer therapy, generating a new, testable hypothesis not contained in its training data which was later demonstrated positively in the lab.[15] If the next leg up in model capabilities could not just write emails but cure cancer, solve nuclear fusion, and discover novel materials, then practically any capital cost becomes justifiable.

This helps to contextualize the launch of the compute scaling race – trust the scaling law, build bigger models, and reap the financial rewards when they can do economically valuable work. Training runs for models like GPT-4 and Gemini 1.0 were rumored to cost ~$50mn, and its expected that in the coming years single model training runs could cost $10 billion+.[16,17] **Model builders can't just double or triple compute resources, they need to chase 10x, 100x and beyond.**

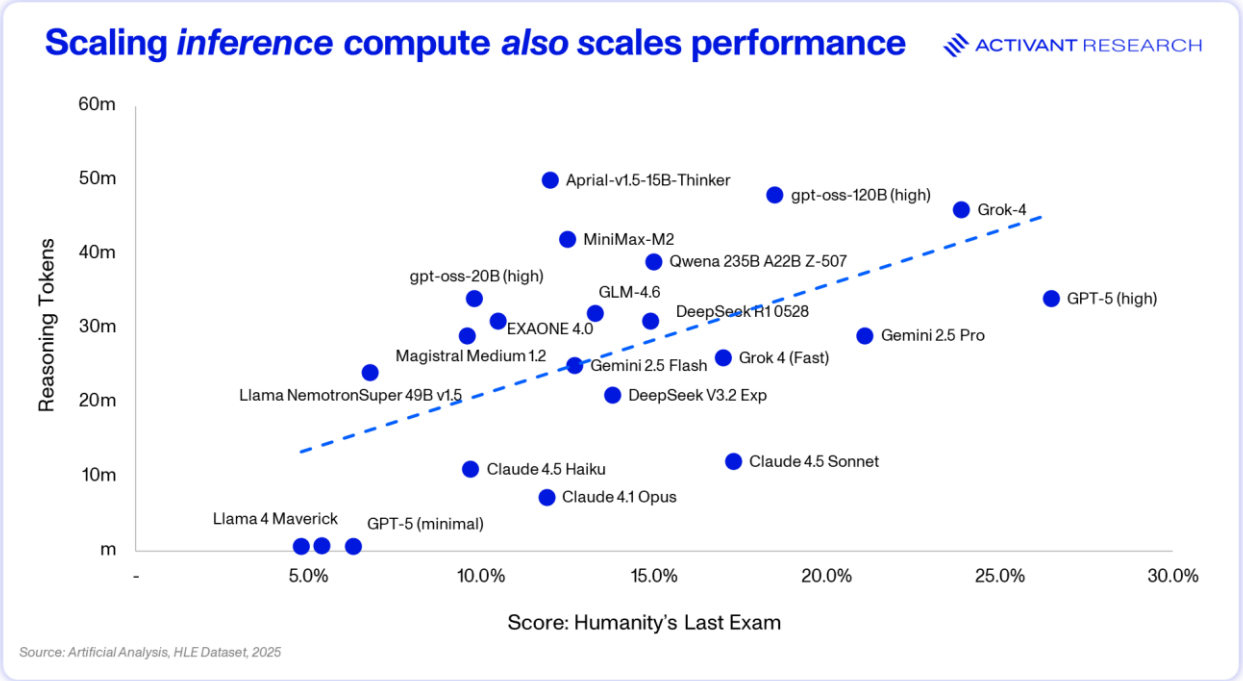
So that begs the question, why was the rumored GPT-5 training run smaller than that of GPT-4.5?

Scaling 2.0: Test-time

The paradigm we've discussed so far, using more compute at training time to improve model performance, is called **pre-training scaling**. However, all of today's state of the art models also use **test-time scaling** – they break down problems into steps, work through them using a "chain-of-thought" (CoT), and return only their final solution to the user. As Daniel Kahneman would say, you're forcing the model to use its system two, and that extra computation comes in the form of more tokens.[18]

The importance of CoT was made clear when researchers found that by simply prompting a model to "think step by step" when solving a problem, they were able to improve accuracy on a math benchmark from 18% to 79%.[19] The approach was codified in OpenAI's o1 model, which broke problems down into steps and applied CoT in the background, not visible to the user, by default. Using CoT vastly improved model performance on many benchmarks thought to be reasoning-heavy, like physics, math and formal logic, leading to this generation being referred to as ***reasoning models***.

Take Humanity's Last Exam (HLE) for example – a novel benchmark put forward after models began to routinely score over 90% at other popular benchmarks like MMLU and GPQA. HLE's questions are described as being at the frontier of human knowledge and while even the current state of the art models only score ~25%, its clear that computing more tokens for reasoning, on average, improves performance. The same scaling law holds true - **the more computing power you apply, the better the model performs**.



Thus, together with pre-training scaling, test-time scaling presses further on the demand for computational resources: Reasoning models can use up to 100x the tokens that their one-shot counterparts would for a given output.[20]

However, test-time scaling is far more efficient than pre-training scaling. We may need to 100x the size of the datacenter to train a 100x larger model, but that's not the case for scaling reasoning by 100x.

During training, calculations must pass forward and backwards through the model multiple times over, while at test-time only a forward pass takes place, once. While the latest training runs are reaching 100,000+ GPUs, inference can be done on a small 8 x GPU block or sometimes a single GPU for smaller models where memory is not a bottleneck.[21,22]

This difference between the efficiency pre-training and test-time scaling helps to explain why GPT-5 was smaller than GPT-4.5 and why, despite widespread calls for its death, pre-training scaling lives on. **Test-time scaling is just the more economically rational approach to focus on right now**. It flips upfront capex into opex and ensures that costs are only incurred with actual customer consumption. The incremental costs associated with test-time computation can be explicitly charged to the customer and smaller reasoning models are cheaper to serve - outperforming a 14x larger one-shot model, for all but the most complex problems.[23]

But that's the crucial caveat, for ***complex*** problems, the larger model still outperformed.[24] No amount of thinking can solve for problems not captured in the base model's capabilities, which is why the emergence of test-time scaling didn't point to the end of pre-training scaling, it just rewired the economic equation. While test-time scaling has accounted for most of the gains in AI performance since late 2024, pre-training scaling has held through 27 orders of magnitude since 1950 and there will undoubtedly be a return to pre-training scaling to equip models with ever greater capabilities.

Rather than simply making the model bigger, the next decade will be defined by a more nuanced approach of optimizing the balance between pre-training and test-time scaling. And at Activant, we see a third vector of scaling, not necessarily in the control of model builders, but driving the demand for compute all the same.

Scaling 3.0: Multi-agent systems

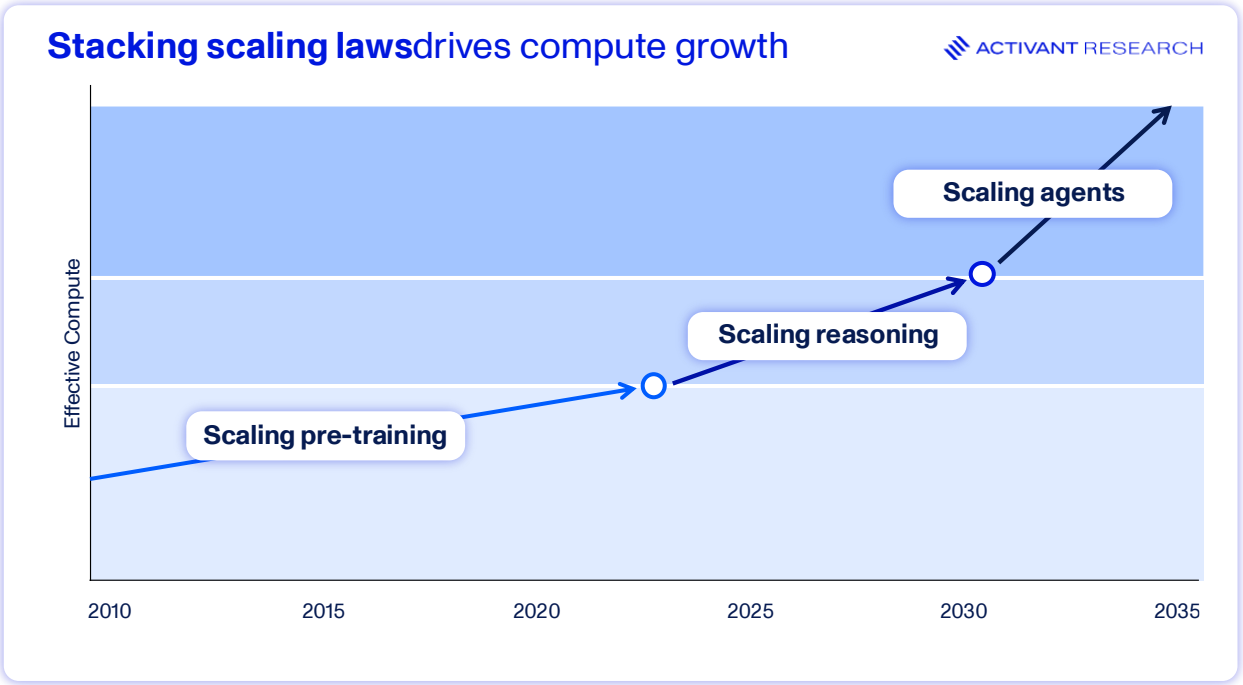
As we wrote in our research on open source AI models, we believe that AI systems will increasingly make use of multiple models to drive cost efficiency and performance. From a compute perspective, this means that a query may be routed through numerous models, each with a specific purpose, before outputting the result – **multiplying the number of tokens consumed once again.**

One special example of compound AI systems – **multi-agent systems (MAS)** take this even further. LLM performance can decline by up to 85% with extremely long context lengths, but massive context might be required for certain problems.[25] As an example, investment decisions can require review of thousands of pages of documents and consultation with numerous domain experts – a problem likely to break down in a single prompt-response turn.

This problem is best dealt with by AI in the same way as humans – with collaboration. A MAS would break the problem into units, letting one of many agents perform a unit of work using specialized context and tools and then collaborate to build the final recommendation. However, these MAS can use 15x the tokens that a non-agentic system would, which could be further compounded by using reasoning-models for each agent.[26,27]

But as in the case of scaling paradigms I and II, the extra token consumption drives performance – Anthropic's multi-agent research system outperformed a single-agent design by 90%.[28] LangChain found that the more noise they added to context, the greater the gap between single and multi-agent systems, further cementing the importance of the MAS paradigm for solving highly complex, context-heavy tasks.[29]

It's a story we are familiar with by now, to eke out that extra level of performance, **use more compute**. In the case of MAS tools like deep research, that could be up to 1,000,000x the tokens of regular one-shot models when answering a query.[30] Each scaling law provides a new vector to tackle the challenge of making AI economically valuable, and each one stacks on top of the priors. The future of highly capable AI systems will be achieved with massively scaled up based models, doing extended reasoning at test-time, and compounded into multi-agent systems. To make that possible, the system is going to have to build an extraordinary amount of new computing capacity.



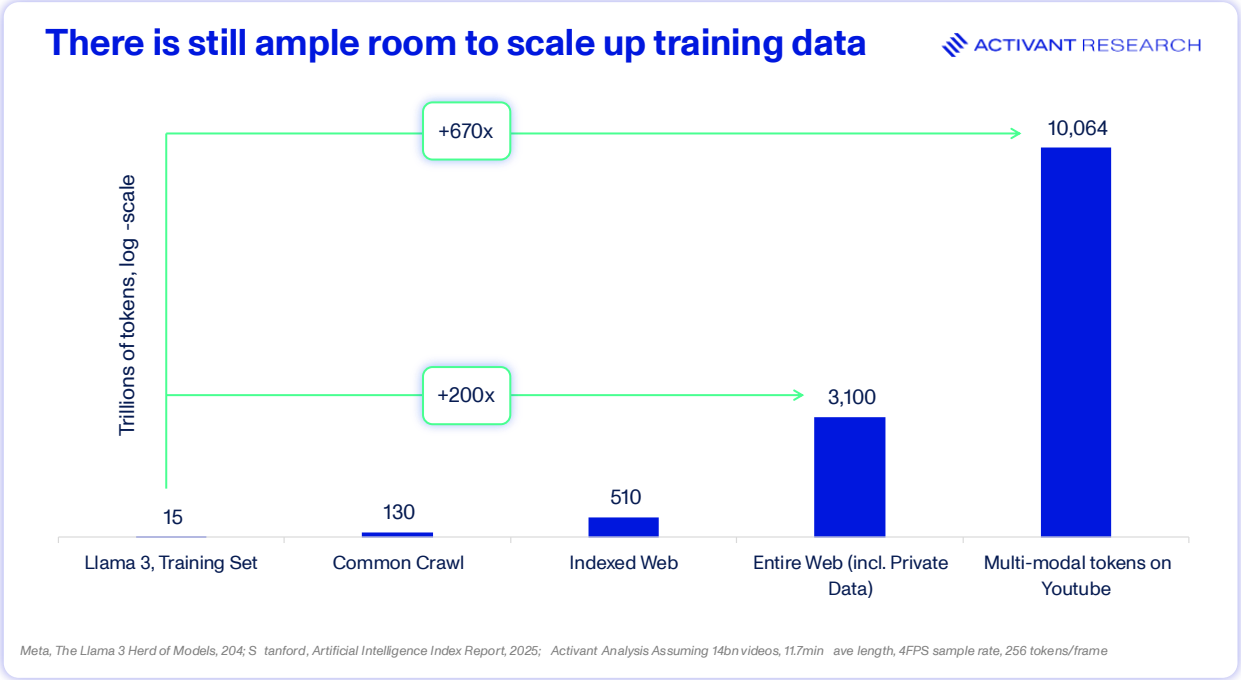
In the context of the scaling laws, chasing datacenters the size of Manhattan ***make sense*** strategically, but that doesn't mean that it makes sense practically.

AI's Mega-Build: The Brakes

As we mentioned earlier, there are three key inputs to the AI scaling law: data, model size and compute time (either at training or test-time). While researchers have been refining the ideal ratios between these inputs, its consensus that ***all*** inputs to continue scaling in lockstep. If one input becomes a bottleneck, scaling stops entirely. The AI build out depends on an ability to continue exponentially scaling data, GPUs and power consumption.

The Data Wall (of worry)

The primary concern that AI models will run out of data arises from the common claim that leading models were trained on "the entire internet". Luckily, its false. Llama 3 has the largest published dataset size at 15 trillion tokens: a selectively filtered and high quality subset of the ~130 trillion tokens available on common crawl.[31,32] When we consider the entire size of the available web, images and video, there's an opportunity for another ~200x scale up, ignoring the fact that the web is still growing exponentially.[33] Further, most estimates of the token contribution from video content assume transcription of text. However, multi-modal models can be trained on the video data itself, where there may be an extra ~10 quadrillion tokens on YouTube alone.[34]



Further, while synthetic data has been a hot but divisive topic, researchers recently found that claims of model collapse under synthetic data use were overblown, and the solution is to aggregate synthetic data with real world data, rather than replacing it.[35] Finally, recent research also indicates that we can use **repeated** training data up to 4x before its value decays.[36] There is no data wall.

GPU Availability & Fab Capacity

When the AI boom kicked off, Nvidia GPUs were absolutely the bottleneck to training and running AI models. VCs were boasting GPU access to win deals and GPU rental prices spiked at over $8/hr.[37] However, prices had stabilized in the $2 - $3 range by 2024 and have continued to decline

to just a few cents over $2.[38] GPU aggregator Shadeform lists GPU availability as "high" across Coreweave, Crusoe, and Nebius – three of the largest dedicated GPU clouds. Getting access to high end GPUs is no longer an issue and we think it will be quite straightforward for manufacturing capacity to keep pace.

Its estimated that GPU stock needs to reach 100 million H100-equivalent units by 2030 for scaling to continue.[39] We estimate that this would require an additional annual capacity of 30 – 40 million units and cumulative capex of $55bn - $70bn, only 1.5x to 2x TSMC's current annual capex and less than Nvidia's FY26E Free Cash Flow.[40] The reality is that GPUs are still a small portion of leading-edge semiconductor capacity and fabs are unlikely to be the constraint that halts scaling.

Scaling GPU capacity is ~2 years of TSMC Capex

ACTIVANT RESEARCH

	Low	High	Notes
TSMC Capacity: Pre-3nm RAMP (Q4 '21)	14.9	14.9	*3,725 12-inch wafer equiv shipments in Q4 '21, annualized*
TSMC Capacity: Current (Q3 '25)	16.3	16.3	*4,085 12-inch wafer equiv shipments in Q3'25, annualized*
Net New Capacity Added (millions of 12-inch equiv. wafers)	**1.4**	**1.4**	
(x) % Advanced Process Nodes	*80%*	*70%*	*Q4 2021 earnings call*
New Advanced Process Node Capacity	**1.2**	**1.0**	
(x) H100 Equivalent GPUs per 12-inch wafer	50	60	*800mn GPU die, with allowance for losses due to round wafer edges*
TSMC New H100 equiv capacity added, '21 - '25	**58**	**60**	
(/) Capex Incurred, Q1 '22 - Q1'25	$106 bn	$106 bn	*Advanced process node ramp started 2022, last quarters excluded for lag time to capacity*
Capex per 1million units ($'bns)	**$1.8**	**$1.8**	
(x) Required annual capacity	30	40	*2030 GPU stock of 100mn (cumulative)*
Capex Requirement	**$55 bn**	**$70 bn**	
Multiple of current annual capex	*1.5x*	*1.9x*	*Using Q3'25 Annualized*
Share of Nvidia FY26 Free Cash Flow	*57%*	*72%*	*CapIQ, 31 Jan 2026E*

In fact, Satya Nadella is on record saying that they have more GPUs in inventory than they can power, the most critical bottleneck.[41]

Power

AI's power problems start with its extreme scale, with specialized AI datacenters drawing 200MW and above, they can't be simply "plugged in". Connecting these facilities, that consume the power of a small city, requires a dedicated, high-voltage connection to the main power grid, triggering an interconnection study, a detailed analysis of how the datacenter will impact the grid. However, requesting one of these studies means "getting in the queue" and that **interconnection queue** is

now infamously long. In most cities, the interconnection queue is 2 – 3 years long and can stretch to 10 years.[42] Even with the study done, connecting to high-voltage transmission lines means building a sub-station, and wait-times for high-voltage transformers have extended to 2 - 4 years.[43]

This has led to a strategic pivot: **BYOP (Bring Your Own Power).** Major datacenter operators are bypassing the grid completely and building their own power infrastructure independent of the grid. Amazon has a direct link to the 2.5GW Susquehanna plant and Meta's 2GW Hyperion datacenter will be powered three new natural gas turbines for which they have secured capacity under a PPA.[44,45] **However, even bypassing the grid, securing ~150GW of new power in the next 5 years (30GW/year) will be a significant challenge.**

While the US added 50GW of solar alone in 2024, datacenters need base power – always on and 100% reliable power.[46] Operators are turning to nuclear but lead times ~5 years make reactors infeasible for scaling through 2030, and even Small Nuclear Reactors (SMRs), while promising, have lead times exceeding 2 years.[47] The medium term AI build will be heavily reliant on natural gas.

GE advertises that its gas turbines can be generating power in as little as two weeks, with carbon capture configurability and the IEA forecasts an "unprecedented" 300bcm expansion in LNG capacity by 2030, enough to power ~190GW of datacenters.[48] Here, the United States holds a profound strategic advantage. Thanks to the shale revolution, domestic gas prices (Henry Hub) are often 2 - 3x cheaper than in Europe or Asia, making on-site gas generation the fastest and most economical path to power.[49]

However, wait times for the gas turbines themselves now have backlogs extending to 2028.[50] It's clear that the rush to expand generation capacity is creating bottlenecks across the entire power supply chain. Luckily, GE alone expects to soon reach 20GW of gas turbine manufacturing capacity annually, with Mitsubishi and Siemens still providing further support.[51]

Datacenters need 30GW/year of new generation capacity. Gas can provide >20GW/year alone, while operators can dip into the 50GW of annual solar capacity by adding batteries, and nuclear will step in post 2030. Generation capacity can scale to support the AI datacenter build out but **none of these factors move at the speed of bits, and there is a high likelihood that individual projects see delays and disruptions due to power over the coming years.** Today's power purchase agreements (PPAs) may be tomorrow's competitive advantage.

On balance, AI scaling can continue at forecasted rates, if someone will step in to pay. What's so special about this infrastructure build out is that in so many cases, that financier is the customer.

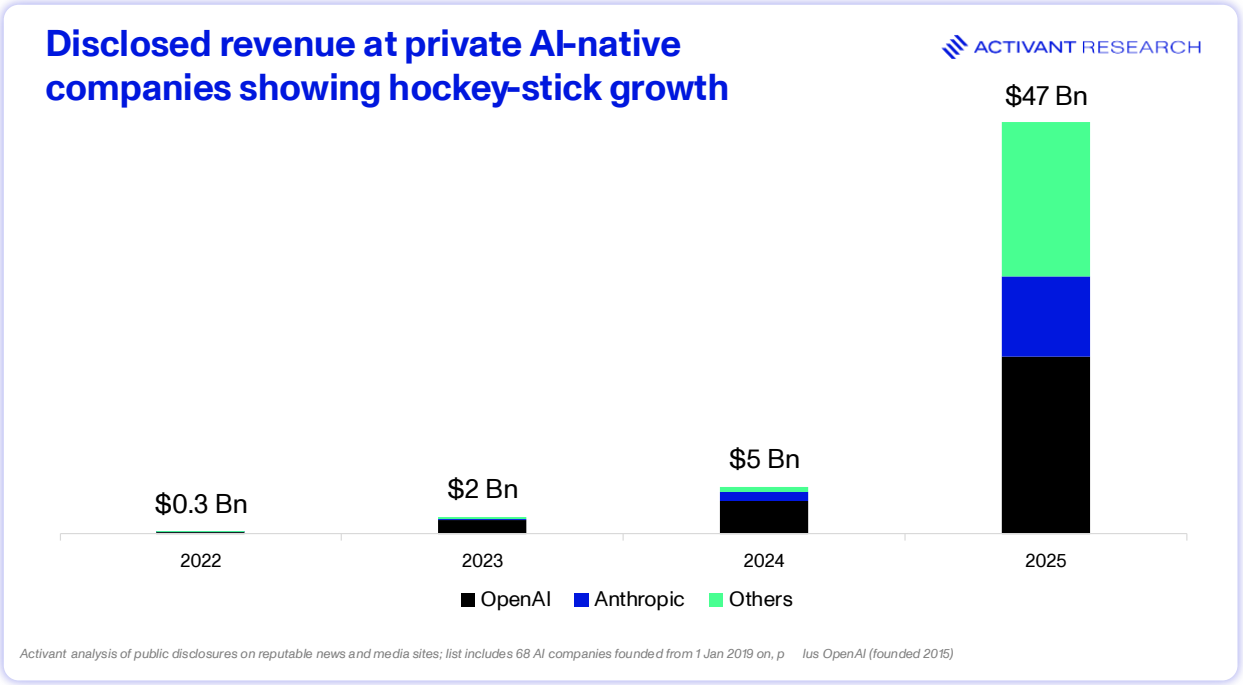
AI's Mega-Build: The Accelerator

Apply more compute, get better results. The scaling laws make it clear why operators across the value chain are excited to build capacity for AI compute and it appears that all potential bottlenecks are fundamentally surmountable. That doesn't allay fears that we're in a bubble – building too much, too quickly for use cases that have not arrived yet. For comfort there, look at the demand side of the equation.

To start, customers across the value chain are deploying AI, not just model builders. For Microsoft and Google's cloud platforms, "demand is significantly ahead of supply", Amazon's Trainium2 compute is "fully subscribed" and at Meta, they describe the business as "compute starved" after finding that they can "profitably use much more compute than [they've] been able to [deploy]."[52]

These Megacap tech companies are deploying AI at billion user+ scale, which provides almost instant payback on AI capex. Microsoft's GitHub copilot is already being used by 90% of the Fortune 500 and we estimate AI workloads on Azure at a ~$12bn run rate. Amazon expects their AI shopping assistant to deliver over $10bn in incremental sales. Google sees query growth of 10% when adding an AI overview and Meta saw ad pricing increase 10% YoY in Q3 2025 as a result of AI-driven ad performance, an incremental impact of ~$16bn annualized. Parsing AI from the core business is messy, but we estimate return on AI capex at these companies is already >10%.[53] The businesses at the center of the capex boom are healthily profitable.

Of course, it's not just Megacap tech that are benefitting. OpenAI's ARR doubled at a $13bn run rate, Anthropic expect $9bn by the end of 2025 and Cursor became the fastest SaaS company in history to reach $100 million ARR (12 months) and subsequently grew its ARR to over $500 million by June 2025.[54,55,56] AI native start-ups are scaling at a speed and scale that dwarfs the cloud SaaS wave.



This speed is not just because AI itself is so transformation, its that waves of new technology stack. AI is fundamentally changing the way that we code, write, do research and more, but these tools are being distributed over an internet already scaled over 30 years to 5 billion users.[57] Applications are being built at warp speed using not just AI but DevOps practices perfected over the past 20 years, and deployed on cloud infrastructure scaled since 2006. When companies like Google and Meta deploy AI features, they (almost) automatically have billions of users, and when the largest enterprise software companies add AI features, tens of thousands of enterprises are fast-tracked up the curve.

AI's infrastructure build out is being supported by the most successful companies in human history, funded out of positive free cash flows, seeing immediate adoption and acceptable ROICs. The build out's fundamentals are strong.

Sizing the market

The evidence laid out indicates that the strong demand for AI compute is not a temporary surge but a significant market driver, firmly rooted in the scaling laws. Applying more compute, whether through scaling pre-training, expanding test-time reasoning, or compounding models into multi-agent systems, consistently unlocks more capable and economically valuable AI.

A mega-build of specialized AI datacenters is both rational and justified. While significant bottlenecks in power and the supply chain present real challenges, they appear addressable. Critically, the build is showing highly favorable fundamentals, with significant revenues being achieved at Megacap scale, and AI natives scaling at rates only possible through stacking innovations.

We believe that the market for AI computing is set for a substantial and structural expansion.

AI Compute could be a ~$450bn industry

ACTIVANT RESEARCH

AI Cloud Computing	Dec-24	Dec-25	Dec-26	Dec-27	Dec-28	Dec-29	Dec-30
Total Datacenter Capacity, GW	59	74	90	115	147	180	230
(x) % AI Specialized	*16%*	*30%*	*38%*	*49%*	*57%*	*62%*	*67%*
AI Datacenter Capacity, GW	**10**	**22**	**34**	**57**	**84**	**111**	**154**
(/) Average Datacenter PUE	1.60	1.60	1.50	1.40	1.30	1.20	1.10
IT Equipment Power Draw, GW	**6**	**14**	**23**	**40**	**65**	**92**	**140**
(/) Server Max Power, kw	55	146	237	327	418	509	600
(x) GPU Per Server	8	103	197	292	387	481	576
Number of GPUs, mns of units	**0.9**	**9.8**	**19.0**	**36.0**	**59.7**	**87.3**	**134.3**
less: % internal training		*40%*	*36%*	*32%*	*28%*	*24%*	*20%*
Monetizable GPU units, mns	**0.9**	**5.9**	**12.2**	**24.5**	**43.0**	**66.4**	**107.5**
(x) Price per GPU Hour	$2.0	1.80	1.60	1.40	1.20	1.00	$0.8
(x) Utilization rate	*60%*	*60%*	*60%*	*60%*	*60%*	*60%*	*60%*
(x) Hours per year	8,760	8,760	8,760	8,760	8,760	8,760	8,760
Total Market Size, $'billions	**$9.2**	**$55.7**	**$102.5**	**$180.2**	**$270.9**	**$348.8**	**$451.8**
% Growth YoY		*506%*	*84%*	*76%*	*50%*	*29%*	*30%*

But establishing the scale of this new market is only the first part of the equation. The critical question remains: *who will capture the value?* Will the incumbent Hyperscalers dominate this new paradigm as they did the last, or will this foundational infrastructure shift allow new entrants, from Neoclouds to serverless inference providers, to capture the profit pools? In our next pieces, we will dive deeper into this new value chain and analyze who stands to win.

---

[1] Goldman Sachs, Powering the AI Era, 2025

[2] Activant analysis of public disclosures on reputable news and media sites; list includes 68 AI companies founded from 1 Jan 2019 onwards, plus OpenAI (founded 2015)

[3] Goldman Sachs, AI to drive 165% increase in data center power demand by 2030, 2025

[4] Energifaktanorge, Electricity Production, 2025

[5] Amazon, Amazon.com Announces Third Quarter Results, 2023

[6] Ars Technica, Introduction to Multithreading, Superthreading and Hyperthreading, 2002

[7] Magnetar, Investing in AI Infrastructure, 2024

[8] Magnetar, Investing in AI Infrastructure, 2024

[9] Nvidia, GTC 2025 Keynote, 2025

[10] Data aggregated from McKinsey, AI 2027 Compute Forecast, Boston Consulting Group, SemiAnalysis, Goldman Sachs, Brookfield

[11] Nature, Exclusive: the most-cited papers of the twenty-first century, 2025

[12] Wired, What OpenAI Really Wants, 2023

[13] Open AI, Language Models are Few-Shot Learners, 2020

[14] Leopold Aschenbrenner, Situational Awareness, 2024

[15] Google, How a Gemma model helped discover a new potential cancer therapy pathway, 2025

[16] EpochAI, How much does it cost to train frontier AI models, 2025

[17] Leopold Aschenbrenner, Situational Awareness, 2024

[18] A token is a discrete, numerical representation of a unit of raw data, such as a word, sub-word, or character, that a model can process

[19] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, Yusuke Iwasawa, Large Language Models are Zero-Shot Reasoners, 2023

[20] Melius Research, October 13 Weekly Video, 2025

[21] Elon Musk via X.com, 2024

[22] BentoML, Choosing the right GPU, 2025

[23] Google Deepmind, Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters, 2024

[24] Ibid

[25] Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, Hao Peng. ontext Length Alone Hurts LLM Performance Despite Perfect Retrieval, 2025

[26] Anthropic, How we built our multi-agent research system, 2025

[27] Melius Research, October 13 Weekly Video, 2025

[28] Anthropic, How we built our multi-agent research system, 2025

[29] LangChain, Benchmarking Multi-Agent Architectures, 2025

[30] Melius Research, October 13 Weekly Video, 2025

[31] Meta, The Llama 3 Herd of Models, 2024

[32] Stanford, Artificial Intelligence Index Report, 2025

[33] Ibid

[34] Activant Analysis: 14bn videos, 11.7min ave length, 4FPS sample rate, 256 tokens/frame.

[35] Gerstgrasser, et. al., Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data, 2024

[36] Muennighoff, et. al., Scaling Data-Constrained Language Models, 2025

[37] Cybernews, From $8 to $1 per hour: GPU rental prices are crashing, 2025

[38] Silicon Data, H100 Rental price index, 2025

[39] Epoch AI, Can AI scaling continue through 2030, 2024

[40] TSMC Q4 2021 earnings conference call, TSMC Q4 2021 – Q3 2025 quarterly earnings presentations, S&P Capital IQ, Activant Analysis

[41] BG2 w/ Brad Gerstner, All things AI w @altcap @sama & @satyanadella. A Halloween Special, 2025

[42] Brookfield, Building the Backbone of AI, 2025

[43] National Infrastructure Advisory Council, Addressing the Critical Shortage of Power Transformers to Ensure Reliability of the U.S. Grid, 2024

[44] World Nuclear News, New supply agreement expands Talen-Amazon partnership, 2025

[45] Data center dynamics, Meta signs two PPAs with Treaty Oak in Louisiana, 2025

[46] Solar Energy Industries Association, Solar Market Insight Report 2024 Year in Review, 2025

[47] Coresite, More Power! Behind-the-Meter Power Systems for Data Centers, 2025

[48] IEA, Gas 2025, 2025

[49] IEA, Natural gas prices in Europe, Asia and the United States, 2022

[50] Utility Dive, GE Vernova bullish on electrical infrastructure as turbine backlog grows, 2025

[51] ibid

[52] Microsoft, Google, Amazon, Meta, Earnings Conference Call Transcript Quarter Ended 30 Sep 2025

[53] AI ROIC is estimated as the annualized run rate of AI-incremental operating profits (40% op. margin assumed) divided by accumulated AI capex (capex in excess of core business capex)

[54] TechCrunch, OpenAI has five years to turn $13 billion into $1 trillion, 2025

[55] TechCrunch, Anthropic projects $70B in revenue by 2028: Report, 2025

[56] TechCrunch, Cursor's Anysphere nabs $9.9B valuation, soars past $500M ARR, 2025

[57] Our World in Data, Adoption of Communication Technologies, 2025