N47

N47.COM

A SIX-PART REPORT

Intelligence on Tap

The economics, architecture, and future of AI, and what must change for ubiquitous, on-demand intelligence to become a sustainable, long-term reality.

Lak Ananth
GLOBAL MANAGING PARTNER, N47

Jonathan Goldberg
FOUNDER, D2D ADVISORY

Six parts. One thesis.

A NOTE FROM THE AUTHORS

We are entering the era of Intelligence on Tap.

The premise of this series is simple: ubiquitous, on-demand AI is not a question of if but of how. The technology works. The demand is real. The capital is flowing. The harder question is what must change in the economics, architecture, and value chain for Intelligence on Tap to become a sustainable, long-term reality.

We bring complementary lenses. Lak is the Global Managing Partner of **N47**, a product-first venture firm backing companies across the AI stack, from infrastructure to applications. Jonathan is the founder of **D2D Advisory**, a proprietary research firm that dissects technology economics for institutional investors after years on the sell side covering semiconductors. Together, we combine builder-side pattern recognition with rigorous financial decomposition.

The Founding Parallel

Every transformative technology reaches a moment where its initial infrastructure cannot support its ultimate ambition. Google Search could not have existed on the enterprise technology stack of its era: proprietary servers, commercial databases licensed at $47,500 per processor, networking equipment with 65% gross margins.

Google's founders didn't question whether search was real. They reinvented the stack: commodity hardware designed to fail, software that assumed failure and routed around it, warehouse-scale architectures optimized for aggregate cost. The result was a 20× cost advantage that made ad-supported search viable.

The Opportunity Ahead

AI sits at an analogous moment. AI queries cost 1–10¢ to serve versus 0.2–0.3¢ for search, while generating comparable revenue per interaction. The dominant chip supplier extracts 73–75% gross margins. Hardware becomes obsolete in 2–4 years but is depreciated over 5–6. These are engineering problems, not existential ones.

This series takes a hard-nosed look at every layer of the AI cost structure, tests whether current economics match the value being created, and maps the specific shifts required to make Intelligence on Tap a durable reality.

PART 1: FROM SEARCH ENGINE TO INTELLIGENCE UTILITY

The Economics of Intelligence on Tap

What does it actually cost to deliver AI on demand, and what must the unit economics look like for it to last?

LAK ANANTH & JONATHAN GOLDBERG

The Economics of Intelligence on Tap

From search engine to intelligence utility

In 1999, Larry Page and Sergey Brin had a problem. Google was processing 3.5 million searches a day on hardware that could barely keep up. The company was burning through servers, running out of money, and facing a fundamental question that had nothing to do with whether search was valuable. Everybody knew search was valuable. The question was whether anybody could afford to deliver it.

Their answer was to reinvent the stack. Instead of buying expensive enterprise servers from Sun or IBM, they built custom racks from commodity PCs - cheap, failure-prone machines held together by software that made the whole cluster reliable. The result was roughly a 20x cost advantage over anyone running traditional infrastructure. That cost advantage didn't just make Google profitable. It made ad-supported search, a product given away for free to billions of users, one of the most lucrative business models in the history of technology.

Twenty-five years later, we are watching the same movie play out with artificial intelligence. The technology works. Demand is exploding. And the unit economics don't yet support the business model that the market requires. The question, once again, is not whether AI is real. The question is whether anyone can afford to deliver it at the scale the world wants to consume it.

History says they can. History also says the path from "this works" to "this is a utility" is an engineering problem, not a science problem, and the companies that solve it will define the next era of technology.

The 1,000x Cost Collapse

The single most important number in AI economics is this: the cost of generating a token of AI output has fallen roughly 1,000x in three years.

When OpenAI first made GPT-3 available via API in late 2021, the cost was approximately $60 per million tokens. By late 2024, equivalent-capability inference through models like GPT-3.5 Turbo had dropped to roughly $0.07 per million tokens. That's a 1,000-fold reduction, according to data from Epoch AI and a16z's "LLMflation" research.

For context, Moore's Law delivered a 2x improvement in compute price-performance roughly every two years. LLM inference costs are declining at approximately 10x per year, about seven times faster than Moore's Law at its peak. No comparable technology has ever dropped in cost this rapidly.

The decline has been relentless across every major model provider:

MODEL	DATE	BLENDED COST/ M TOKENS	CHANGE
GPT-3 (Davinci)	Nov 2022	$20.00	Baseline
GPT-3.5 Turbo	Early 2023	$0.70	-96%
GPT-4	Mar 2023	$36.00	(Frontier)
GPT-4 Turbo	Nov 2023	$14.00	-61%
GPT-4o	May 2024	$4.00	-71%
GPT-4o Mini	Jul 2024	$0.24	-94%

Source: OpenAI API pricing; blended = 80% input / 20% output per Andrew Ng methodology

A clear pattern has emerged: roughly 18 months after a frontier model launches, equivalent capability is available at commodity prices. GPT-4's performance level went from $36 per million tokens to under $1 through open-source alternatives in about 18 months. What was state-of-the-art last year is a commodity today.

The Google Precedent

Google's financial history provides a precise template for understanding where AI goes next. The parallels are not metaphorical. They are structural.

When Google went public in 2004, it generated $3.2 billion in revenue from roughly 73 billion searches per year. It spent $1.5 billion on traffic acquisition costs, the payments to partners like AOL, Firefox, and later Apple that directed users to Google's search engine. TAC consumed 45.7% of revenue.

Over the next twenty years, a remarkable thing happened. Revenue grew 110x, from $3.2 billion to $350 billion. But traffic acquisition costs grew only 38x, from $1.5 billion to $55.6 billion. As a share of revenue, TAC dropped from 45.7% to 15.9%.

YEAR	REVENUE	COT OF REV.	TAC	TAC % REV.	GROSS MARGIN
2004	$3.2B	$1.5B	$1.5B	45.7%	54.0%
2008	$21.8B	$8.6B	$6.6B	30.2%	60.4%
2012	$50.2B	$17.2B	$11.0B	21.8%	65.7%
2016	$90.3B	$29.0B	$17.0B	18.8%	67.9%
2020	$182.5B	$71.0B	$32.8%	18.0%	61.1%
2024	$350.0B	$146.3B	$55.6B	15.9%	58.2%
2025	$403.3B	~$163B	~$60B	~14.9%	~59.6%

Source: Alphabet 10-K filings; 2025 estimated from quarterly reports

Meanwhile, search volume grew from approximately 73 billion queries per year in 2004 to an estimated 6 trillion in 2025, an 81x increase, but revenue grew 110x. Google didn't just process more queries; it extracted more value from each one, because scale economics drove down the marginal cost of serving a query

while advertising monetization improved. The underlying dynamic was straightforward. Google's cost per query fell from roughly 0.5–1.0 cents at IPO to 0.03–0.1 cents by 2022, a 10–30x decline, while revenue per query rose from about 0.5 cents to 1.6 cents. Costs went down, revenue per unit went up, volume exploded. That is the playbook.

> Google didn't make search cheaper. Google reinvented the stack so search could exist at all.

Where AI Economics Stand Today

AI's unit economics in 2026 look almost exactly like Google's in 2001: the technology works, users are flooding in, and the companies providing it are hemorrhaging money.

OpenAI generated $3.7 billion in revenue in 2024 but lost roughly $5 billion. Its compute spend alone, $5 billion between training, inference, and research, exceeded total revenue. Sam Altman publicly confirmed that OpenAI loses money on ChatGPT Pro, its $200-per-month power-user subscription. The daily cost of running ChatGPT is estimated at $700,000 to $3 million, depending on the source.

The structural gap between AI inference and Google Search is real, but it's narrowing:

METRIC	GOOGLE SEARCH	AI INFERENCE (2026)	AI TARGET
Revenue per query	~1.6¢	~1–4¢	3–5¢
Cost per query	0.03–0.1¢	0.36–10¢+	<0.5¢
Operating margin	50%+	Negative to ~20%	50%+
GPU / Server utilization	70–80%	30–50%	65–75%
Capex intensity	~$50B / yr	~$13B (OpenAI)	Declining ratio

Sources: SemiAnalysis, Alphabet 10-K, industry estimates

Three structural factors explain this gap. First, AI generates rather than retrieves. A search query triggers a lookup in a pre-built index; an AI query requires a forward pass through a neural network with billions of parameters, consuming orders of magnitude more compute. Second, large models require dedicated memory allocation even when idle, which crushes utilization rates. Third, subscription pricing creates heavy-tail usage patterns where power users consume 100x the compute of average users, making per-user economics wildly uneven.

But each of these is an engineering operational constraint, not a fundamental limitation. And we can already see them being solved.

The Demand Explosion

While costs are collapsing, demand is growing at rates that would have seemed absurd two years ago.

ChatGPT now handles over 1 billion queries per day, with 800–900 million weekly active users. OpenAI's API processes 2.2 billion daily requests. OpenRouter, a multi-provider inference platform, went from processing 10 trillion tokens per year to over 100 trillion tokens per year between 2024 and 2025, a 10x increase in twelve months. The platform now handles more than 1 trillion tokens per day.

ChatGPT message volume grew 8x year-over-year, per OpenAI's 2025 State of Enterprise AI report. That's not just more users. It's more intensive usage per user. Average prompt length has quadrupled, from roughly 1,500 tokens per API request in early 2024 to 6,000 tokens by late 2025, as developers send more complex instructions and larger context windows become standard.

The application layer tells the same story. GitHub Copilot grew from 5 million to 20 million users in a year, with 90% of Fortune 100 companies now using it. Cursor became the fastest-growing SaaS product in history, growing from $100 million to $1.2 billion in annual recurring revenue in a single year. Reasoning models, which didn't exist before December 2024, now account for more than 50% of all token consumption on multi-provider platforms.

1B+

ChatGPT queries handled per day

10x

Increase in tokens routed through Open-Router, 2024 thru 2025

8x

YoY growth in ChatGPT message volume

Jevons Paradox

Despite a 1,000x reduction in the cost per token, organizations saw their AI spending increase by 320% year-over-year, according to OpenAI's enterprise data. They didn't pocket the savings. They consumed vastly more. After DeepSeek demonstrated dramatically cheaper inference, Satya Nadella publicly invoked the Jevons Paradox: "As AI gets more efficient and accessible, we will see its use skyrocket, turning it into a commodity we just can't get enough of."

This pattern is the defining feature of every technology that becomes a utility:

TECHNOLOGY	PRICE DECLINE	VOLUME INCREASE	NET MARKET EFFECT
Electricity (1930–1980)	Price stabilized	Consumption grew 12x	Massive market expansion
Internet bandwidth (2005–2020)	~100x cheaper	~100x more traffic	Total spend grew
Google search (2004–2024)	Cost / query fell 30x	Volume grew 68x	Revenue grew 110x
AI tokens (2002–2025)	1,000x cheaper	Growing 3–8x / year	Inference market: $106B

Sources: EIA, Cisco, Alphabet 10-K, Epoch AI, Markets and Markets

Google's own history makes this concrete. Between 2004 and 2024, the cost per search query fell approximately 30x. But search volume grew 81x, and revenue grew 110x. The volume curve didn't just match the price curve. It overwhelmed it. Revenue grew faster than either cost fell or volume rose, because cheaper delivery made the product accessible to use cases that were previously uneconomical.

The Scissors: Training vs. Inference

Underneath the headline numbers, a structural divergence is reshaping AI's economics in a way that directly parallels Google's infrastructure evolution.

Training costs, the upfront investment to build a model, are accelerating exponentially. The original Transformer architecture cost approximately $900 to train in 2017. GPT-3 cost $4.6 million in 2020. GPT-4 cost $78–100 million. Estimates for GPT-5 range from $500 million to $2.5 billion. Epoch AI data shows training costs growing at 3.5x per year. Anthropic CEO Dario Amodei and team pointed this out while they were at OpenAI.

Meanwhile, inference costs, the marginal cost of serving each query, are collapsing at 10x per year. This is the scissors dynamic: the fixed cost of creating intelligence rises while the variable cost of delivering it falls. This creates a business model that looks a lot like Google's: enormous upfront capital expenditure amortized across billions of transactions at trivial marginal cost.

MODEL	YEAR	TRAINING COST	INCREASE
Transformer	2017	~$900	Baseline
GPT-3	2020	$4.6M	5,100x
GPT-4	2022	$78–100M	17–21x
Gemini Ultra	2023	~$191M	~2x
Llama 3 (405B)	2024	$200–500M	~2x
GPT-5	2025	$500M–2.5B	5–10x

Sources: LambdaLabs, Stanford AI Index 2024, Epoch AI, TechRadar

The hardware cost curve reinforces this trajectory. GPU cloud pricing for H100s has fallen from roughly $4 per hour in early 2024 to $1.38 per hour in March 2026, a 65% decline, driven by overcapacity as NVIDIA's Blackwell generation ramps production. B200 GPUs, unavailable a year ago, now start at $2.40 per hour. Each GPU generation delivers approximately 2x the price-performance of the prior generation, and the cloud market's competitive dynamics accelerate the pass-through to end users.

Late 2025 marked a structural milestone: for the first time, inference surpassed training in data center revenue. By 2030, analysts project inference will account for 65–70% of all AI compute spending, up from roughly 25% in 2023. AI is transitioning from the "build" phase to the "serve" phase, exactly the trajectory that turned Google's infrastructure investment into a cash-generating utility.

Where the Value Accrues

If the Google parallel holds, then the question for investors is not whether AI becomes an affordable utility. It's who captures the value as it does.

Google's lesson is that the company that solves the infrastructure problem captures the economics. Google didn't win search because it had the best algorithm (though it did). It won because it rebuilt the server stack so completely that no competitor could match its unit economics. AltaVista had search. Yahoo had search. They didn't have the infrastructure economics to sustain it at the scale the internet demanded.

Today's AI landscape presents the same opportunity. The inference market is projected to grow from $106 billion in 2025 to $255 billion by 2030, a 17–19% CAGR. But the companies that will capture that value disproportionately are the ones solving the hardest engineering constraints: utilization efficiency, model serving optimization, and hardware-software co-design that drives marginal cost below the revenue line.

The signals are already visible. Custom silicon programs at Google (TPUs), Amazon (Trainium), and Microsoft (Maia) echo Google's original commodity-server strategy. Open-source models like Meta's Llama family are commoditizing the model layer, shifting value to the infrastructure and application layers, just as the proliferation of web content shifted value from content creation to search and distribution. Inference optimization startups are pursuing the same kind of fundamental cost reduction that Google's early server team achieved.

The AI industry's $5–9 trillion infrastructure buildout through 2030 is not a bubble. It is the same kind of capital cycle that built the electric grid, the telephone network, and the cloud. The returns will not go to everyone who spends the capital. They will go to the companies that spend it most efficiently: the ones who reinvent the stack. AI is going through a similar process to those historical examples, albeit in a way that it is still not entirely clear where the value capture ultimately lands.

PART 2: DECOMPOSING THE AI COST STACK

Where Every Dollar Goes

Decomposing the AI cost stack — chips, talent, data, energy — and the specific leverage points where compression unlocks Intelligence on Tap at scale.

LAK ANANTH & JONATHAN GOLDBERG

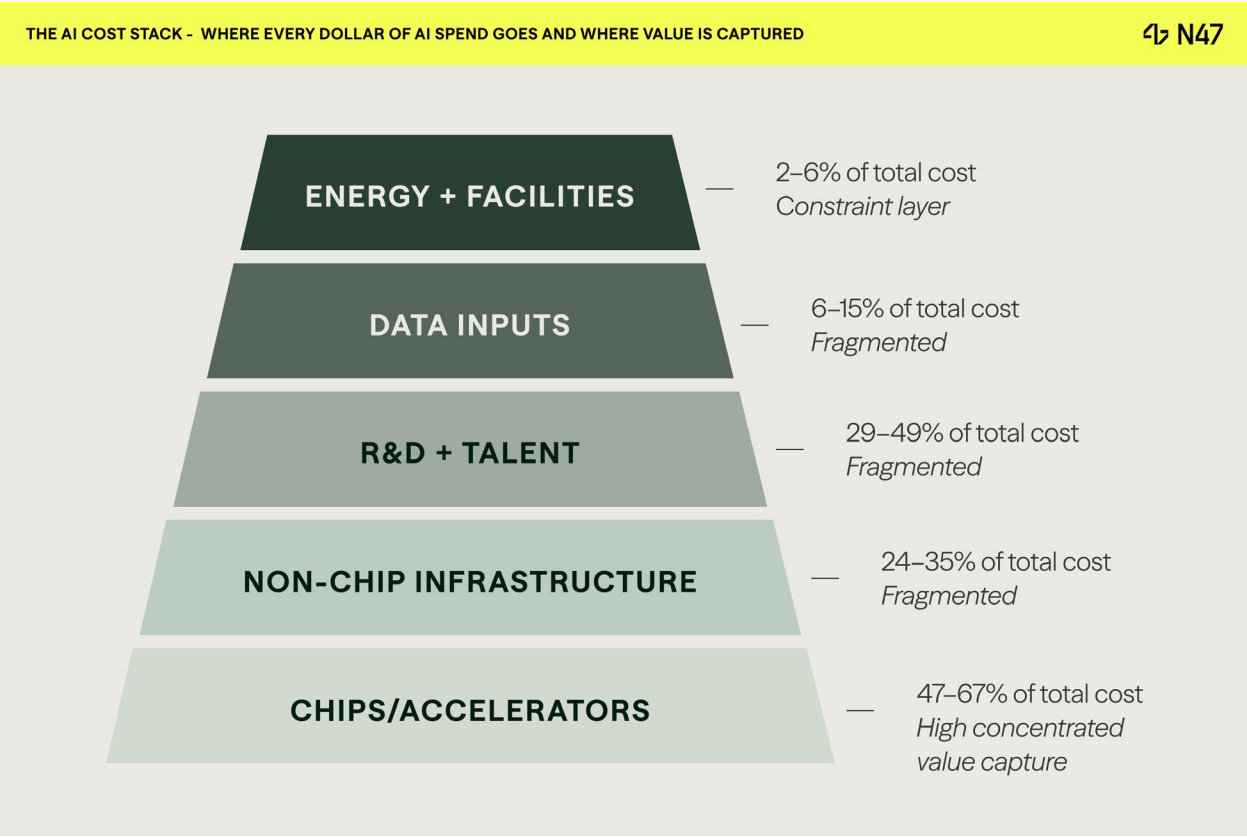
Where Every Dollar Goes

Decomposing the AI cost stack

When people describe AI as expensive, they usually gesture at GPUs. The GPU explanation is technically accurate and analytically shallow. It is the equivalent of explaining why airlines are unprofitable by pointing at jet fuel. Correct, incomplete, and not particularly useful for anyone trying to understand what changes.

Most analyses treat AI costs as a monolith: compute is expensive, things will get cheaper, wait and see. That framing misses the structure of the problem entirely. The AI cost stack has at least five distinct layers, each with its own economics, its own leverage points, and its own set of beneficiaries extracting margin. Breaking open the black box reveals something important: the obstacles to Intelligence on Tap are real, specific, and — with one or two exceptions — solvable.

Part 1 of this series established why solving them matters. The trajectory of inference costs, falling 1,000x in three years, suggests the economics of AI delivery are moving in the right direction at an unprecedented pace. But trajectory is not destiny. Understanding where every dollar currently flows is the prerequisite for identifying where compression will happen next.



Layer One: Chips and Accelerators (47–67% of Training Cost)

The GPU explanation for AI costs is incomplete, but it is not wrong. Chips and accelerators represent the largest single layer in the training cost stack, consuming 47 to 67 cents of every training dollar depending on workload and model architecture.

The dominant supplier in this layer is capturing margins that have no parallel in modern technology history. NVIDIA's data center GPU business operates at 73–75% gross margins, exceeding the most notorious technology monopolies of prior eras. Oracle databases at peak dominance reached 65% gross margins. Cisco networking equipment, at the height of the late 1990s infrastructure buildout, operated in the same range. NVIDIA has surpassed them. The acceleration of revenue, profits in aggregate and per employee at NVIDIA are stunning and are unprecedented in our modern era of competitive capitalism.

What makes this margin durable, for now, is not the silicon itself. CUDA, NVIDIA's parallel computing platform, has been installed and learned by approximately 4.5 million developers over fifteen years. The lock-in is not hardware alone. It is software, workflow, and accumulated expertise. A developer moving from NVIDIA to an alternative accelerator does not just swap chips. They retrain their team, rewrite their code, and absorb significant transition risk at a time when time is priceless, and capital is abundant to the players. That friction is worth more to NVIDIA than any physical moat in its supply chain.

The path through this layer is already visible. Custom silicon programs at Google (TPUs), Amazon (Trainium), and Microsoft (Maia) have demonstrated 50–70% lower training costs for specific, targeted workloads. These are not theoretical cost reductions. They are in production. The strategy mirrors Google's original commodity-server approach precisely: sacrifice flexibility for efficiency on the tasks you run at scale, and extract enormous cost advantages as a result. The question for the rest of the industry is not whether custom silicon works. It is whether any given organization runs sufficient volume to justify the development cost.

Layer Two: Non-Chip Infrastructure (24–35%)

The second layer is less visible but nearly as significant. Interconnects, servers, memory, and networking consume 24 to 35 cents of every infrastructure dollar, and this layer has its own margin extraction problem.

NVIDIA's NVLink and NVSwitch interconnect ecosystem — the high-speed fabric required to link GPUs within and across servers — adds $3,000 to $5,000 per GPU in interconnect hardware alone. High-bandwidth memory, the specialized memory that feeds GPU compute at speeds standard DRAM cannot match, costs $300 to $600 per chip. Every component in the AI server stack commands premium pricing because the demand is inelastic: there is no equivalent substitute for running frontier models at production scale.

The compounding effect matters here. Each layer in the stack carries its own margin. A data center

operator buying NVIDIA GPUs, NVIDIA interconnects, specialized memory, and custom networking equipment is paying premium pricing at every step of the assembly. The fully loaded cost of a GPU cluster is substantially higher than GPU list prices suggest. This is the AI equivalent of what happened with enterprise software stacks in the 1990s, when the server cost was just the beginning of what Oracle, Sun, and middleware vendors extracted from each deployment.

Layer Three: R&D and Talent (29–49%)

This is the number that surprises most people analyzing AI unit economics.

Researcher compensation rivals hardware as a cost driver in the AI stack. Top AI researchers earn a median total compensation of approximately $875,000, according to industry surveys. Elite researchers — the handful capable of independently advancing the state of the art in training efficiency or model architecture — command packages ranging from $1 million to $20 million. A frontier training effort requires a team of 50 to 200 people. Personnel costs for a single major training run, including salaries, infrastructure engineering, and research operations, run $150 million to $250 million.

The R&D-to-revenue ratio for leading AI labs runs 60–150%. Mature technology companies typically spend 15–20% of revenue on research and development. AI labs are spending multiples of their total revenue on research, before accounting for compute costs. This is not a sign of dysfunction. It is the correct behavior for a technology in its formative phase, where the research output has not yet been fully monetized. But it is a cost component that does not appear in most discussions of AI economics, and it must.

The talent layer also has a different compression curve than chips. GPU prices fall as manufacturing scales and competition emerges. Researcher salaries do not follow the same trajectory. As the number of organizations competing to hire elite AI talent continues to grow, this layer may prove stickier than any other in the stack.

Layer Four: Data (Emerging but Structural)

The free data era is over.

Through approximately 2022, frontier AI models were trained substantially on internet data scraped at low or zero marginal cost. The legal and commercial landscape has shifted decisively. Known data licensing deals exceeded $800 million in aggregate in 2024, with more announced regularly. Reddit's content licensing arrangement with Google runs approximately $60 million per year. Arrangements with major news organizations, academic publishers, and other content owners are establishing market rates that will only increase as AI companies compete for training data with demonstrable quality advantages.

Frontier labs are also paying credentialed experts directly to generate new training data from scratch. Scale AI and Mercor have built billion-dollar businesses recruiting lawyers, doctors, teachers, investment bankers, senior developers, and math olympiad winners at rates of $200 per hour and more for the

reasoning-heavy datasets that post-training now demands.

Beyond licensing costs, copyright exposure adds billions of dollars in potential liability to the balance sheets of AI labs actively litigating their data practices. The litigation risk alone is altering procurement behavior. Several labs have shifted meaningfully toward synthetic data generation, which carries its own computational cost but eliminates legal exposure. Either way, the days of training foundation models on essentially free internet data are behind the industry. Data has moved from a near-zero-cost input to a structural line item with its own set of market dynamics.

Layer Five: Energy (Small Today, Strategic Tomorrow)

Energy represents only 2–6% of training costs at current prices — a relatively modest share of the total stack. The forward picture is different.

AI-optimized data center facilities cost $20 million to $40 million per megawatt to build, compared to $7 million to $12 million for traditional data center construction. Power consumption per rack has increased roughly 10x over five years as compute density has grown. The power demands of frontier model training are doubling approximately every year, with projections of 4 to 16 gigawatts of dedicated AI compute capacity required by 2030.

Energy is not yet a cost problem. It is becoming a bottleneck problem. The limiting factor in several hyperscaler expansion plans is not capital, not silicon availability, and not talent. It is the speed at which power infrastructure can be permitted, built, and connected. This constraint does not appear prominently in cost-per-token calculations today, but it is already shaping where data centers get built, how quickly capacity expands, and which providers can serve latency-sensitive workloads from which geographies. Energy is the layer that starts small and ends strategic.

LAYER	SHARE OF COST	DOMINANT MARGIN CAPTURER	COMPRESSION PATH
Chips and Accelerators	47-67%	NVIDIA (73-75% GM)	Custom silicon (Google, Amazon, Microsoft)
Non-Chip Infrastructure	24-35%	NVIDIA ecosystem, memory suppliers	Open interconnect standards, commodity HBM
R&D and Talent	29-49%	Concentrated talent market	Open-source model development, distillation
Data	Growing	Content owners, litigation risk	Synthetic data generation
Energy	2–6% training	Power infrastructure constraints	Geographic diversification, efficiency gains

Sources: SemiAnalysis, LambdaLabs, a16z infrastructure analysis; percentages vary by model type and workload

The Amortization Problem

Cutting across all five layers is a structural distortion in how AI infrastructure costs get reported.

GPU hardware becomes economically obsolete in 2 to 4 years as each new generation delivers roughly

2x the price-performance of its predecessor. Companies depreciate that hardware over 5 to 6 years using accounting conventions designed for physical infrastructure that wears out slowly. The result is a systematic understatement of true infrastructure costs in reported financial results.

This is not an academic observation. It directly affects unit economics comparisons. A data center that depreciated its H100 fleet over 6 years and then faces replacement pressure from B200 availability in year 3 is carrying assets on its balance sheet at values that do not reflect their remaining economic utility. The true cost of AI infrastructure is higher than depreciation-adjusted numbers suggest — and any analysis that ignores this mismatch will reach systematically optimistic conclusions about the pace of margin improvement.

The resolution is architectural separation: dedicated training hardware on accelerated replacement cycles, inference hardware on longer ones. Training workloads drive the frontier and require the latest silicon. Inference workloads operate on more mature, price-stable hardware and can tolerate longer replacement cycles. Managing these fleets separately, with different replacement cadences and different depreciation treatment, is a meaningful lever for improving both economics and the accuracy of how those economics are reported.

Where the Leverage Is

Breaking open the AI cost stack reveals a consistent pattern. Each layer has near-term defenders: NVIDIA's CUDA ecosystem, specialized memory suppliers, hyperscaler talent markets, incumbent data holders. But each layer also has a credible compression path.

> The AI cost stack is not a monolith. It is a set of discrete engineering problems, each with identifiable leverage points, progressing at different speeds.

The Google precedent matters here as well. Google did not solve its infrastructure cost problem in a single move. It addressed commodity servers first, then custom networking, then eventually custom silicon with TPUs — a process that took nearly a decade. AI is following the same sequence, compressed into a shorter timeframe by competitive pressure and the pace of underlying hardware improvement.

The most consequential compression path operates on the work itself. DeepSeek-V3 trained a 671-billion-parameter mixture-of-experts model for $5.6 million in 2.79 million GPU hours, less than 10% of Llama 3.1 405B's compute. Sparse activation means only 37 billion of those parameters fire per token at inference, and distillation packs reasoning into smaller models, so savings compound on every query served, not just the training run.

Custom silicon for the chip layer. Open interconnect standards and next-generation memory for non-chip infrastructure. Open-source model development (Meta's Llama family being the most visible example

that amortizes talent costs across the industry rather than concentrating them at individual labs). Synthetic data generation as a substitute for licensed content. Geographic diversification and purpose-built efficiency gains for energy. None of these solutions arrives simultaneously. None arrives without friction. But the structure of the problem is now clear.

The companies that understand exactly where their dollars are going are the ones positioned to compress costs where compression is possible, and to avoid spending capital chasing leverage that does not exist. That clarity is what separates an engineering problem from an open question.

Part 1 of this series established that AI's cost trajectory looks like Google's: falling fast, with volume growth that will eventually outrun price decline. Part 2's answer is that the cost stack, when decomposed, looks like Google's as well: multiple discrete layers, each solvable, none solved all at once. The companies that worked through those layers methodically — Google with servers, then networking, then silicon — built the most valuable infrastructure businesses in history.

Intelligence on Tap requires working through the same layers in AI. The stack is visible. The leverage points are identifiable. The only question is sequence and speed.

SOURCES AND DATA NOTES

Cost stack percentages
Industry estimates based on LambdaLabs, SemiAnalysis, and a16z infrastructure analysis; percentages vary by model type and infrastructure configuration.

NVIDIA margins
NVIDIA 10-K and quarterly earnings filings; CUDA developer count from NVIDIA public disclosures.

Custom silicon cost reductions
Google TPU performance data from Google Research; Amazon Trainium from AWS re:Invent disclosures; Microsoft Maia from public announcements.

Interconnect and memory costs
SemiAnalysis GPU server teardown analysis; HBM pricing from memory industry surveys.

Researcher compensation
Levels.fyi AI researcher salary surveys; company filings and public disclosures.

Data licensing
Licensing deal aggregates from industry reporting; Reddit/Google licensing terms from SEC filings.

Energy costs
McKinsey data center construction cost estimates; Uptime Institute power density surveys; IEA AI energy consumption projections.

DeepSeek-V3

https://www.deeplearning.ai/the-batch/deepseek-v3-redefines-llm-performance-and-cost-efficiency
https://arxiv.org/html/2412.19437v1
https://deepwiki.com/deepseek-ai/DeepSeek-V3/3.3-mixture-of-experts-(moe)

PART 3: WHO OWNS YOUR INTELLIGENCE?

The Architecture Wars

Both the vertical architecture model and the distributed stack will survive. They won't both win.

LAK ANANTH & JONATHAN GOLDBERG

The Architecture Wars

Who owns your intelligence?

When you joined the internet in 1994, you probably didn't join the internet. You joined AOL. AOL gave you access, email, chat, news, and entertainment. In exchange, it owned the experience, the data, the integrations, and the workflow. For millions of users, AOL and the internet were the same thing until they weren't. CompuServe ran the same model. Prodigy too. The walled garden was the dominant architecture for online information until a different model (open protocols, user-controlled data, decentralized infrastructure) made the walled garden's economics unsustainable and its user proposition obviously inferior.

The AI industry is working through the same architectural question today, and the outcome is not yet settled. Two competing models are emerging. The choices made over the next three years will determine where value in AI accrues, who captures it, and whether Intelligence on Tap becomes a durable utility or a concentrated service controlled by a small number of actors.

The Vertical Model: Convenience by Design

The first architecture is the one the dominant AI labs have built by default. Call it the vertical model: the lab trains the model, hosts the model, stores the conversation history, builds the integrations, designs the interface, and controls the workflow. The user brings a subscription and a problem. The lab provides everything else.

This architecture has real advantages, and dismissing them would be a mistake. Training frontier models requires capital that only the vertical model can currently justify — OpenAI's 2025 compute spend ran approximately $8.5 billion. The operational complexity of serving billions of queries reliably, routing between model tiers, managing context windows, and maintaining sub-second response times is genuinely substantial. The vertical labs have solved these problems. That is not a small thing.

The enterprise pitch is equally coherent. A company deploying ChatGPT Enterprise or Claude for Work gets a managed service: one vendor relationship, one security review, one compliance framework, guaranteed uptime. For a legal firm that needs to know exactly where client data goes, or a hospital system navigating HIPAA, the vertical model's clear chain of custody is a genuine feature, not merely a convenience.

The economics also mirror a well-understood historical pattern. Oracle in the 1990s, SAP in the 2000s, Salesforce in the 2010s — enterprise technology incumbents have always extracted the highest margins from customers willing to pay for integration simplicity and vendor accountability. There is a durable market for this. It will not disappear.

The limitation is scale. The vertical model, as currently structured, cannot support Intelligence on Tap at planetary scale without remaining prohibitively expensive. Parts 1 and 2 of this series established the unit economics gap: AI queries cost 1 to 10 cents to serve while generating 1 to 4 cents in revenue. That gap closes as costs fall, but it closes faster in a competitive, distributed market than in a concentrated, vertically integrated one. The pre-Google enterprise server stack had 65–75% margins. It also had no path to Google-scale unit economics. The vertical model faces the same structural ceiling.

The Distributed Stack: Intelligence as Utility

The second architecture is being assembled from below, by developers and users who want something different: intelligence that works the way electricity does. Available on demand. Provided by a competitive market. Not tied to any single supplier's data strategy or business model.

The evidence that this architecture is gaining traction is not theoretical. It is measured in GitHub stars and monthly downloads.

OpenClaw, created by Austrian developer Peter Steinberger and open-sourced in November 2025, hit a viral inflection in late January 2026 that produced numbers no enterprise software launch could match: 145,000 GitHub stars and 20,000 forks within two weeks. Millions of installs across macOS, Windows, and Linux. More than 3,000 community-built skills on ClawHub, OpenClaw's open marketplace. Integration with every major messaging platform — WhatsApp, Telegram, Slack, Discord, Signal, iMessage, Teams — built not by Steinberger but by community contributors over a matter of weeks.

The architecture that produced these numbers is worth examining carefully. OpenClaw is a harness, not a model: users choose between cloud models (Claude, GPT-4) or fully local models through Ollama. The user's data stays on the user's machine. The workflow is the user's to design. The intelligence layer is a competitive input, priced by the market and substitutable based on cost, capability, or trust.

Even the security vulnerabilities discovered in early OpenClaw deployments illustrate the scale of demand rather than undermining it. SecurityScorecard found 42,900 OpenClaw instances exposed on the internet, with 15,200 vulnerable to remote code execution. Software does not accumulate security researchers' attention at that scale unless it has first accumulated users at that scale. The demand came before the hardening because the demand was that strong.

OpenClaw is the most dramatic data point, but not an isolated one. Ollama, which allows users to run open-weight models locally, has accumulated millions of downloads and a developer community that has built hundreds of integrations. LM Studio follows the same pattern. Anthropic's Claude Cowork extends the distributed model to file and task management workflows, allowing intelligence to operate on user-controlled data without that data ever passing through a centralized lab. Meta's Llama model family, released as open weights, has been downloaded hundreds of millions of times and deployed by organizations that have no intention of routing their AI through any single vendor's API.

The pattern across these products is consistent: users want intelligence available on tap, but they want to hold the data and control the workflow. The lab provides the model. The user owns the harness.

What the Adoption Data Says

The interesting analytical question is not which architecture is better in the abstract. It is what the adoption patterns tell us about where the market is actually going.

The vertical model has the larger current revenue base. OpenAI, Anthropic, and Google are generating substantial subscription and API revenue from the vertical architecture. Enterprise adoption of managed AI services is growing. The vertical model is not going away.

But the distributed stack is growing faster at the edges, and the edges are where technology transitions begin. The open-source software movement did not initially threaten Oracle's enterprise revenue. It threatened the web server market, a space Oracle did not take seriously, and eventually undermined the economic logic of proprietary software at every layer of the stack. Red Hat, Apache, MySQL, Linux: each started at the margin and moved toward the center.

The distributed AI stack is following a similar trajectory. Coding assistants and developer tools, where the distributed model is strongest, now represent more than 50% of all token consumption on multi-provider platforms. Developers are the most price-sensitive, integration-aware segment of the AI market. They are also the ones who build the enterprise applications that run on top of whatever infrastructure wins. The architectural choices developers make now will shape what enterprise buyers can access later.

And the labs see the same number. If half of all tokens flow through software development, the application wrapping that workflow is the most valuable real estate in the stack. OpenAI shipped Codex. Anthropic built Claude Code. Google followed with Gemini CLI. Each routes around Cursor and GitHub. Claude Design now aims at Figma, Canva, Adobe, and Lovable. The vertical model is no longer just selling tokens. It is moving up the stack to own the applications that consume them.

DIMENSION	VERTICAL MODEL	DISTRIBUTED STACK
Unit economics	Higher margin per seat; constrained by cost structure	Lower margin per query; expands by volume
Data control	Lab holds conversation history and usage data	User controls data; model is a stateless service
Model choice	Single vendor's model lineup	Competitive across labs, open weights, local models
Enterprise fit	Strong: managed service, compliance, accountability	Growing: requires more integration work
Developer adoption	Significant via API	Dominant in tooling and open-source ecosystem
Cost trajectory	Improves with scale; constrained by concentration	Compresses faster under competition

Sources: a16z/OpenRouter 100T Token Study; OpenAI State of Enterprise AI 2025; SecurityScorecard; industry estimates

The Economic Case for Each Architecture

The long-run economic argument for the vertical model rests on capability and trust. Frontier model training requires capital that only concentrated investment can support. The top AI labs are spending $500 million to $2.5 billion per training run and deploying those models across billions of queries. The investment can only be justified if the model provider captures a meaningful share of the value delivered. Without the vertical model's margin structure, the incentive to fund frontier research collapses.

> The vertical model funds the research. The distributed stack benefits from it. The long-run question is whether that arrangement remains stable as open-weight models approach frontier capability.

The long-run economic argument for the distributed stack rests on three structural advantages. First, competition: a market where users can route queries across multiple providers creates price pressure that a vertical model cannot generate internally. The 1,000x cost decline in inference documented in Part 1 has been driven substantially by this competition. Second, data alignment: the most valuable AI applications require access to proprietary data that many organizations are not willing to centralize at a third-party lab. The distributed model makes data-local intelligence practical. Third, the breadth of the development community: when Meta releases Llama weights, every organization that builds on those weights benefits from Meta's training investment without contributing to Meta's revenue — a form of R&D socialization that creates cost floors constraining what any vertical provider can charge.

The AOL Question

The internet parallel is not perfect, but it is instructive. AOL had real advantages in 1995: curated content, reliable access, customer support, and a business model that worked. The open web had worse content, slower speeds, and no support. But the open web had something AOL could not match: an architecture that anyone could build on, at a cost that kept falling.

By 2002, AOL's subscriber count had peaked. By 2006, the walled garden was a historical artifact — not because the open web was better in every dimension (it wasn't, not for years), but because its cost structure and architectural flexibility made it the inevitable foundation for everything that came next.

The AI labs are not AOL. Their technical capabilities are genuinely superior to any open-weight alternative today, and the capital barriers to competing are substantially higher than they were for web development in 1996. But the distributed stack is demonstrating the same fundamental property: it aligns incentives between builders and users in a way that the vertical model cannot match at scale. Developers who route around a single provider's pricing are not disloyal; they are responding rationally to architecture, exactly as web developers responded to the economics of open protocols.

Both Architectures Will Survive. They Won't Both Win.

The most honest conclusion is that both architectures will coexist for a long time. The history of technology offers many examples of parallel ecosystems that maintain some form of equilibrated balance, at least for a time – Windows and Mac, iPhone and Android. The vertical model will retain strong positions in regulated industries, large enterprise deployments, and use cases where integration simplicity justifies a premium. The distributed stack will expand fastest at the developer layer, in cost-sensitive markets, and in any application where data sovereignty is non-negotiable.

The historical pattern across comparable transitions — proprietary versus open-source software, walled gardens versus the open web, mainframe timesharing versus personal computing — is that distributed architectures eventually capture the larger share of long-run value by expanding the total market. The pipes carry more water when they are cheap and open. The question for every investor and builder in this space is the same one that mattered in 1999 and in 2006: are you positioned for the architecture that maximizes margin today, or the one that maximizes volume tomorrow?

Part 4 of this series examines how the foundation labs are responding to this architectural pressure: by segmenting tokens into tiers of thinking depth, accuracy, and speed, in what may be the vertical model's most durable competitive position.

SOURCES AND DATA NOTES

OpenClaw adoption data
GitHub public repository statistics; SecurityScorecard, 'OpenClaw Security Analysis,' January 2026.

Developer tool token consumption
a16z/OpenRouter, 'State of AI 2025' (100 Trillion Token Study).

OpenAI compute spend
OpenAI financial disclosures and investor reporting; The Information, 'OpenAI 2025 Financials.'

Enterprise AI adoption
OpenAI, 'State of Enterprise AI 2025'; Bain, 'AI's Trillion-Dollar Opportunity 2024.'

Pre-Google enterprise stack margins
IDC data center cost surveys; SemiAnalysis historical analysis.

Meta Llama download data
Meta AI blog and developer conference disclosures, 2025.

PART 4: THE TOKEN TRAP

Tokens Thinking, and Speed

The defining commercial misalignment of the current AI era.

LAK ANANTH & JONATHAN GOLDBERG

Tokens, Thinking, and Speed

The token trap

What is a token worth?

The question sounds simple. The answer is the source of the most significant structural misalignment in AI's commercial model.

A token spent on casual conversation — 'summarize this email' — costs the same compute as a token spent on clinical differential diagnosis. The inference hardware does not distinguish between them. Neither does the pricing model. But the value delivered differs by orders of magnitude. A summarized email saves a few seconds. A well-reasoned diagnosis changes a patient's outcome.

This mismatch sits at the center of a problem the AI industry has not yet solved: how to price intelligence in a way that reflects the value it creates rather than the volume it consumes. Two problems hide inside this one. The supply side has a routing problem: matching each query to the cheapest model that can handle it. The demand side has a segmentation problem: matching each customer to a price that reflects what the answer is worth in their hands. The industry has barely started on the first and almost entirely ignored the second.

The answer matters more than it appears. Get it right, and AI pricing aligns incentives between labs and users in a way that sustains the economics of Intelligence on Tap. Get it wrong, and the current models will systematically undercharge for the most valuable interactions and overcharge for the least, slowing adoption exactly where adoption matters most.

The Reasoning Paradox

The clearest illustration of the pricing problem is the reasoning model.

Until 2024, AI models generated output autoregressively: one token at a time, as fast as the hardware allowed. Thinking and generating were the same process. The result was fast responses that could be confidently wrong on problems requiring multi-step logic.

The reasoning models introduced by OpenAI (o1, o3), Anthropic (extended thinking in Claude), and Google (Gemini with deep research) work differently. Before generating a visible response, they produce internal 'thinking' tokens — chains of reasoning that explore the problem, check assumptions, and identify errors before anything appears in the output. The response is shorter and more accurate because it is preceded by reasoning the user never sees, but the hardware still fully processes.

The results are striking. On mathematical benchmarks, complex coding problems, and multi-step reasoning tasks, reasoning models substantially outperform their non-reasoning counterparts. The cost profile is equally striking. A simple factual query might consume 500 tokens. A complex reasoning task on

the same model can consume 50,000 — a 100x difference in compute for a single interaction. At frontier reasoning model pricing, that one interaction costs several dollars. The same query routed to a fast, cheap model costs a fraction of a cent.

This is not an anomaly to be corrected. It is a signal about how different AI tasks actually are from each other, and why treating all tokens as equivalent units has become the defining pricing error of the current AI era.

The Tier Explosion

The foundation labs have recognized the heterogeneity problem and responded with a proliferation of model tiers designed to match capability to cost. The market now has at least four distinct levels, each with its own price point and appropriate use case.

TIER	REPRESENTATIVE MODELS	PRICE (INPUT $/M TOKENS)	BEST FOR
Budget	GPT-4o Mini, Claude Haiku, Gemini 1.5 Flash	$0.07-$0.25	Classification, summarization, simple Q&A, extraction
Mid	GPT-4o, Claude Sonnet	$2.50-$3.00	Complex instruction-following, nuanced writing, moderate reasoning
Frontier	Claude Opus, Gemini Ultra	$15.00+	Tasks requiring the full capability of current best models
Reasoning	o1, o3, Claude extended thinking, o1-pro	$15-$150+ (output)	Multi-step logic, math, complex coding, high-stakes analysis

Sources: OpenAI, Anthropic, Google API pricing pages; blended estimates based on 80/20 input/output split

The tier structure is the right strategic response to query heterogeneity. It is also, in practice, applied clumsily. Most applications today route every query to one model — the one the developer originally integrated — rather than selecting dynamically by query type. A customer service chatbot handles simple returns questions and escalated billing disputes through the same model, overpaying for simple interactions and potentially underserving complex ones. The tiers exist. Systematic use of them does not.

The tier structure addresses heterogeneity on the supply side. The demand side has tiers of its own, and they are uncorrelated with the model tiers above. A consumer trivia app and a clinical decision support tool can route the same query to the same model and the lab will receive identical revenue from both. The customer captures wildly different values. The lab captures none of the differences.

The Routing Problem

Matching each query to the right model tier in real time is what the AI industry calls the routing problem. It is one of the highest-leverage unsolved problems in the entire stack.

Done well, intelligent routing reduces effective inference cost by 5 to 10x without any reduction in output quality, by directing simple queries to cheap models and complex ones to capable models. The efficiency gain follows directly from the pricing differential between tiers. An application that currently routes all

queries through Claude Sonnet at $3 per million tokens, and could route 80% of its volume through Claude Haiku at $0.25 per million tokens without quality loss, is paying roughly 3x more than it needs to. At enterprise query volumes, that gap is measured in millions of dollars annually.

Martian, one of a small number of startups working on the routing layer, has built infrastructure designed to classify query complexity in real time and dispatch to the appropriate model. The approach requires solving a hard sub-problem: assessing the difficulty of a query before routing it, using a classifier cheap enough that its cost does not eliminate the savings. Unify.ai is pursuing the same space from a slightly different angle, abstracting across model providers and optimizing for cost, quality, and latency simultaneously.

The challenge is not unlike the one Google solved with PageRank — the value lies in the ranking and routing function, not in the content being served. The company that makes intelligent routing reliable and deployable at scale will have built one of the most structurally important positions in the supply side of the AI stack: a layer that captures margin from every AI interaction regardless of which model answers it. Routing optimizes what the lab earns per query. It does not determine what the customer can be charged. That problem sits one layer up.

The Efficiency Levers Beyond Routing

Model routing is the highest-leverage lever, but not the only one. A set of inference optimization techniques has demonstrated efficiency gains of 90% or more for specific tasks, and they are increasingly composable.

Distillation trains a smaller, cheaper model to replicate the outputs of a larger frontier model for a defined task. The result is a model that costs a fraction of the original to serve but performs comparably on the specific distribution of queries it was trained on — a technique that works best when the task is narrow and the query distribution predictable, conditions that apply to most enterprise AI deployments.

Quantization reduces the numerical precision of model weights, typically from 16-bit to 8-bit or 4-bit representations. Quality loss is minimal for most tasks while memory requirements fall by 50–75%, enabling larger models to run on cheaper hardware or smaller models to run on edge devices entirely.

Speculative decoding uses a small, fast model to generate draft responses that a large, accurate model then verifies and corrects. Because most draft tokens require only minor correction, the large model does a fraction of the work it would if generating from scratch — delivering near-frontier output quality at substantially lower latency and compute cost.

None of these techniques is exotic. All are in production use at leading AI companies. What is not yet standard is applying them systematically across an application's full query distribution, which requires exactly the kind of query classification and routing infrastructure that the market is still building.

Two Pricing Models, Two Wrong Answers

The efficiency tools exist. The tier structure exists. The routing problem is identified. And yet the two dominant pricing models in the AI market create systematic incentives against using any of them well.

Flat-rate subscriptions create what OpenAI has acknowledged publicly: gym membership economics. The pricing model assumes average usage. Heavy users consume 10 to 100 times average compute. OpenAI has confirmed it loses money on ChatGPT Pro at $200 per month for its most capable models. The subscription model pits user incentives (consume as much as possible) directly against lab economics (cost is variable and scales with usage). One side of that equation wins at the expense of the other.

Per-token API pricing solves the margin problem but creates a different misalignment. A developer building a medical diagnosis tool pays the same per-token rate as one building a trivia chatbot. The pricing model cannot distinguish between a token that helped a physician catch a drug interaction and a token that answered a question about celebrity birthdays. Both are priced identically. The high-value use case captures no premium and subsidizes nothing.

> The token is not the unit of value. The task is closer. The customer and the moment are closer still. AI is pricing the unit the hardware counts, not the unit the buyer experiences.

The direction the market must move is toward outcome-aligned pricing: tiered by the capability deployed, metered by the complexity of the interaction, and ultimately calibrated against the value of the task completed. This is how professional services are priced. It is how usage-based SaaS has evolved. It is the only model that can simultaneously sustain the economics of frontier AI development and price access broadly enough to support diffusion into healthcare, legal, education, and financial services — the industries where AI's most durable value will be created.

The Segmentation Layer

Routing addresses the supply side of the pricing problem. Segmentation addresses the demand side. They are different problems, and the AI industry has barely begun to recognize the second.

Every prior wave of enterprise software has been priced this way. Bloomberg charges roughly $30,000 per terminal per year for what is, at the technical layer, a data feed and a chat interface. The compute cost is trivial. The pricing reflects a single fact: the buyer is a trader, and a thirty-second information edge is worth a fortune. Veeva charges life sciences companies a multiple of generic CRM pricing for substantially the same primitives, because the segment values FDA-compliant workflow differently than a real estate brokerage does. ServiceTitan does the same for the trades. The capability underneath is commoditizing. The segment redefines the price.

The same pattern is now emerging in AI, and almost none of it lives at the model layer or the routing layer. Harvey prices by the legal seat, calibrated against the partner-hour saved. Abridge prices by clinical encounter, calibrated to documentation throughput. Hippocratic prices by nursing role and shift coverage. None of these companies meters tokens to its customers. They meter the unit the customer actually buys — an hour, an encounter, a shift, a covered patient. The token cost paid upstream to the lab is an input cost, not a customer-facing price. The margin between what is paid per token and what is charged per outcome is where the segmentation surplus lives.

> Routing optimizes which capability gets used. Segmentation determines what that capability is worth.

This is the layer missing from most discussions of AI economics. The lab sells capability by the token. The router optimizes which capability gets used. Neither captures what the same capability is worth in the hands of a cardiologist versus a content marketer. That difference is the entire commercial opportunity, and it is being captured today by the companies that own a customer segment and wrap the underlying model with the workflow, accountability, and trust the segment requires. Routing is leverage. Segmentation is destiny.

The Token Trap

The AI industry is caught in a trap of its own construction. It built pricing around the token because the token is what the hardware counts. It is a convenient unit of measurement that has almost no relationship to the unit of value — the task completed, the decision supported, the outcome improved.

The path out runs through both layers. Routing solves the supply side, matching each query to the cheapest model that can answer it. Segmentation solves the demand side, matching each customer to a price that reflects what the answer is worth. Routing earns its margin on volume. Segmentation earns its margin on value. Neither alone closes the gap between what the hardware counts and what the buyer experiences. Together they do.

Two structural positions emerge from this. The first is the router. The company that classifies queries by complexity and dispatches them to the right model at scale will sit between every user and every model, on every interaction, permanently. PageRank did not create web content; it organized the surplus. Routing does not create AI capability; it organizes the surplus of deploying it.

The second position is the segment owner. In every prior platform shift, the vertical application that owned a customer segment captured more economic value than the horizontal infrastructure beneath it. The application that owns the surgeon, the lawyer, or the underwriter will likely capture more of AI's surplus than any router ever does. The router is leverage. The segment is the customer.

The token is cheap and getting cheaper. The task is closer to value, but still not the unit. The customer is. The infrastructure that prices intelligence by what it is worth in the hands of the buyer, not by what the hardware counts, is the most valuable thing in AI that has not yet been built. Two layers, not one. The routing layer is being built now. The segmentation layer is where the next decade of AI value will be captured.

SOURCES AND DATA NOTES

Model pricing
OpenAI, Anthropic, and Google API pricing pages; blended estimates follow Andrew Ng methodology (80% input / 20% output).

Reasoning model token consumption
OpenAI o1 and o3 technical documentation; Anthropic extended thinking model card; community benchmarks on reasoning token budgets.

ChatGPT Pro economics
Sam Altman public statements on ChatGPT Pro margin; The Information reporting on OpenAI 2024-2025 financials.

Routing efficiency gains
Martian and Unify.ai technical documentation; academic literature on LLM routing and cascade inference.

Inference optimization techniques
Hugging Face documentation on quantization and speculative decoding; academic papers on RAG and knowledge retrieval; industry benchmarks on distillation efficiency.

PART 5: THE SEARCH FOR NEW INTELLIGENCE

Beyond Transformers

Transformers won't be the final architecture. The winning bet is inference-layer flexibility, versus training-layer scale.

LAK ANANTH & JONATHAN GOLDBERG

Beyond Transformers

The Search for New Intelligence

In 1985, the dominant logic transistor technology was bipolar. Bipolar transistors were fast, proven, and backed by decades of engineering investment. The companies that had built the deepest bipolar expertise — Fairchild, National Semiconductor, and analog specialists across the industry — were not making a foolish bet. They were betting on the best available technology, supported by manufacturing scale and a concentration of talent that had no peers.

They were also betting on the wrong architecture. CMOS transistors had existed for years but were initially slower and less capable than bipolar. What they had was a different set of economic properties: a fraction of the power consumption, dramatically better scaling with each manufacturing generation, and a cost structure that let them proliferate into applications bipolar could not reach. Within a decade, CMOS dominated. The companies that had led the bipolar era were not the ones that won the CMOS era.

The AI industry could be at an analogous moment. Transformers, the architecture underlying every major language model deployed at scale today, are genuinely extraordinary. They may also not be the final form of intelligence. The open question, one that carries significant capital implications, is not whether alternatives will emerge, but which ones, when, and what that means for the infrastructure currently being built to serve a transformer-centric world.

What Transformers Do Brilliantly

It would be a mistake to underestimate the transformer architecture in the process of arguing it will eventually be supplanted. The core mechanism — attention, which allows the model to weigh the relevance of every token in a sequence against every other — has proven remarkably general. Language understanding, code generation, image recognition, protein structure prediction, mathematical reasoning: transformers have produced state-of-the-art results across domains that researchers did not initially expect the architecture to handle.

The scaling properties have been equally extraordinary. As compute and data increase, transformer performance improves in ways that have been largely predictable. This predictability is what allowed OpenAI, Anthropic, and Google to justify spending hundreds of millions of dollars on individual training runs: the returns are estimable. The architecture has been refined by thousands of researchers over nearly a decade, with optimizations at every layer of the stack.

Transformers earned their dominance. The question is whether dominance is permanent.

The Cracks

The transformer's central limitation is also the source of its strength. Attention compares every token against every other token in a sequence, which means compute requirements scale quadratically with sequence length. Double the context window and attention compute quadruples. This is manageable at 4,000 tokens. It becomes expensive at 100,000 tokens and genuinely prohibitive at million-token contexts, which is exactly the kind of long-horizon reasoning that the most valuable AI applications will eventually require.

A second limitation, less discussed but equally consequential, is that transformers are statistical pattern matchers. They predict the most probable next token based on patterns in training data. For language tasks, this is extraordinarily powerful — human-generated text encodes vast amounts of implicit structure that the model learns to replicate. For tasks requiring verifiable correctness, mathematical proof, legal analysis, clinical diagnosis, and financial modeling, 'statistically plausible' and 'provably right' are different things. Transformers excel at the former. They struggle with the latter.

A third limitation matters most for embodied applications. Transformers learn the structure of language by training on text. They do not learn the structure of causality by acting in environments. For autonomous driving, robotics, simulation, and any domain where a model must predict what the world will do in response to what an agent does, the transformer is not the right primitive. The relevant primitive is action-conditioned, not token-conditioned.

These are not product flaws that can be patched in the next model release. They are architectural properties. And they are driving serious research and investment in alternative approaches that are already past the theoretical stage.

The Alternatives Taking Shape

Five architectural directions have moved from research papers into production deployments.

State-space models: Where transformers process sequences by comparing every token to every other, state-space models process sequences more like a signal filter: they maintain a compressed memory state that updates as new tokens arrive, rather than reprocessing the full sequence for each new input. The practical result is linear rather than quadratic scaling with sequence length, a fundamental efficiency advantage for long-context tasks. The Mamba architecture, developed by researchers at Carnegie Mellon and Princeton, demonstrated competitive performance against transformers at a fraction of the compute for long sequences. Cartesia, one of several startups commercializing state-space approaches, is already deploying the technology in production speech and language applications.

Mixture of experts: Rather than activating an entire model for every query, Mixture of Experts architectures maintain many specialized sub-networks and route each query to only the relevant subset. A model might contain a trillion parameters total but activate only 20 billion for any given query, delivering near-frontier quality at dramatically lower per-query compute. This is not a future direction — it is in production at scale today. Google's Gemini and Mistral's Mixtral are both MoE architectures.

DeepSeek's models, which drew significant market attention in early 2025, are also MoE-based, and the cost advantages they demonstrated contributed directly to investor reassessment of AI infrastructure economics.

Neurosymbolic systems: Transformers understand language; they struggle to prove things. Neurosymbolic architectures combine neural pattern recognition with structured logical reasoning, producing systems that can both understand a problem and verify a solution against formal rules. For domains where 'statistically plausible' is insufficient, such as legal analysis, clinical decision support, scientific research, and financial modeling, neurosymbolic approaches offer something pure transformers cannot: auditable reasoning. DeepMind's AlphaProof, which solved International Mathematical Olympiad problems by combining language model reasoning with formal proof verification, represents the current research frontier. Commercial applications remain early-stage, but the trajectory is clear, and the addressable market is large.

Diffusion language models: Transformers generate text autoregressively: one token at a time, left to right, each output conditioned on all previous outputs. Diffusion models take a different approach: they start with noise and iteratively refine toward a coherent output, a process that can be parallelized more efficiently for certain tasks. Inception Labs' Mercury demonstrated a diffusion language model that processes text in parallel rather than sequentially, with speed advantages over autoregressive models for latency-sensitive applications. The architecture is earlier-stage than MoE or state-space models but represents a live direction with commercial investment behind it.

World models: Where the four architectures above all predict the next token in a sequence, world models predict the next state of an environment, conditioned on an action. The model takes in observations and produces a plausible next observation given what an agent did. This is the architecture beneath modern self-driving simulation (Wayve's GAIA-1 and GAIA-2), interactive 3D environments (Google DeepMind's Genie 3, Decart's Oasis), and robotics policies trained 'in dreams' before deployment to real hardware (Meta's V-JEPA 2, Comma.ai's openpilot policy). Two architectural families exist inside this category. Generative world models, pursued by Wayve, General Intuition, Runway, and Decart, produce visible playable futures in pixels or 3D scenes. Latent world models, including Yann LeCun's JEPA work at AMI Labs and the philosophical descendants of DeepMind's MuZero, predict in compressed abstract space, throwing away unpredictable visual detail in exchange for efficiency. The category is real, and the capital behind it suggests serious investors believe the path to embodied AI runs through architectures that look very little like transformers.

ARCHITECTURE	CORE ADVANTAGE	CORE LIMITATION	PRODUCTION STATUS
Transformer	General-purpose, well-optimized, predictable scaling	Quadratic compute with context length; statistical, not provable	Dominant at scale
State-space models (Mamba)	Linear scaling with sequence length; efficient for long context	Less flexible than transformers for general tasks	Early production (Cartesia, others)
Mixture of Experts	Near-frontier quality at fraction of per-query compute	Routing complexity; requires large parameter counts	Production at scale (Gemini, Mixtral, DeepSeek)
Neurosymbolic	Verifiable, auditable reasoning for high-stakes domains	Narrower task coverage; harder to scale	Research / early commercial
Diffusion LLMs	Parallelizable generation; speed advantages at inference	Architectural maturity; quality gap on complex tasks	Early commercial (Mercury)
World Models	Action-conditioned simulation; learns causality from observation; fixed-cost compute over stochastic environments	Earlier-stage for general tasks; latent variants are harder to evaluate; not a replacement for language tasks	Early commercial in autonomous driving, gaming, and robotics

Sources: Carnegie Mellon / Princeton Mamba research; Google Gemini technical report; Mistral AI; DeepSeek technical report; DeepMind AlphaProof; Inception Labs Mercury

The Infrastructure Implication

Each of these alternatives has a different compute profile than the transformer. State-space models prioritize memory bandwidth over floating-point operations. Mixture of Experts requires efficient routing infrastructure and large total parameter capacity. Neurosymbolic systems need tight integration between neural and symbolic compute. Diffusion models benefit from different parallelization patterns than autoregressive generation.

World models have a different profile again. Generative world models are heavy on diffusion and visual decoding, with real-time interactive inference requirements that current inference stacks weren't designed for. Decart, for example, is moving from Nvidia GPUs to Etched's transformer-specific ASICs because of latency demands. Latent world models like JEPA push compute toward representation learning rather than pixel generation. Neither workload looks much like the matrix multiplication patterns the current GPU buildout is optimized for.

This is the infrastructure implication that current capital allocation has not fully absorbed. The GPU clusters being built today are optimized for the matrix multiplication operations that transformer attention requires. They are transformer infrastructure. As alternative architectures move from research into production, the compute platforms serving them will need to adapt, and the adaptation will not be purely software.

The companies positioned for this transition are those building model-agnostic inference infrastructure. Groq, whose custom inference chips were designed around throughput and latency rather than training-scale brute force, can serve multiple model architectures efficiently. Cerebras, with its wafer-scale chip architecture, targets workloads that prioritize memory bandwidth over raw floating-point compute — a profile better suited to several of the emerging architectures. The open-source inference engine vLLM, now serving requests for dozens of model types, represents the software layer of the same flexibility bet.

> The training clusters of today are the mainframes of tomorrow. The inference layer is the cloud.

The thesis follows directly: the companies that build model-agnostic inference infrastructure will capture more durable value than those betting a single architectural paradigm will persist. Just as the internet's value accrued to the protocol layer — TCP/IP, HTTP — rather than to any single hardware configuration, AI's enduring value will accrue to the flexible inference layer that can serve any model, any architecture, any workload.

Transformers Are the Internal Combustion Engine

The bipolar-to-CMOS transition is instructive precisely because CMOS did not win by being better than bipolar at everything bipolar could do. For years, it wasn't. It won because it was good enough at most things and radically cheaper at scale. That combination of sufficient capability and superior economics made CMOS the foundation of every computing generation that followed, including the GPU architectures that now power the transformer era.

Transformers are the internal combustion engine of AI: transformative, dominant, and not the final form. The alternatives emerging today are not better than transformers at everything transformers can do. At least, not yet. State-space models sacrifice some of transformers' flexibility for long-context efficiency. MoE models require sophisticated routing infrastructure. Neurosymbolic systems cover a narrower task surface. Diffusion models are earlier in their development curve. None is a complete replacement. And some alternatives are competing to be a better transformer; world models are competing to be a different kind of model entirely.

But each addresses a specific limitation that transformers carry as an architectural property, not a product bug. And each is advancing along a cost-and-capability curve that has historically tended to surprise those who assumed the current dominant form would persist. The companies with the deepest transformer expertise — OpenAI, Anthropic, Google — are not standing still; all three have active research into alternative approaches and hybrid architectures. They are not assuming permanence either.

What This Means for Investors and Builders

Four practical conclusions follow from this analysis.

First, evaluate training and inference investments on different timelines. The frontier training workload will remain transformer-centric for several years — the returns on transformer scaling are still positive, and the talent and tooling are concentrated there. But the inference layer, which will serve an architecturally diverse model ecosystem as the field fragments, requires different infrastructure planning. A training cluster optimized for one architecture is capital equipment for a specific workload.

A flexible inference platform is a durable business.

Second, the Mixture of Experts transition is already underway and its capital implications are regularly underestimated. MoE's per-query compute reduction does not simply make existing AI cheaper, it changes the economics of which applications become viable. Use cases that were uneconomical at transformer inference costs may cross into viability as MoE deployment scales. That is both an opportunity and a risk for existing AI infrastructure investments premised on current cost structures.

Third, the neurosymbolic direction represents the most underappreciated long-term bet. The domains where verifiable correctness matters — law, medicine, science, finance — are the domains where AI can deliver the most durable economic value and command the highest revenue per interaction. A medical AI that is statistically plausible is a liability. One that can show its reasoning and verify it against clinical evidence is a product. The architectural shift required to build that product is not incremental. It is a different approach to intelligence, and the timeline to commercial viability is measured in years, not decades.

Fourth, the embodied AI category is being built on a different foundation entirely. The capital flowing to world models is not a bet on a faster transformer. It is a bet that robotics, autonomous vehicles, and physical-world automation require action-conditioned models that transformer infrastructure cannot serve well. For capital allocators with exposure to physical-world AI or to the inference layer that will serve it, this is the architectural question that matters most.

The AI infrastructure buildout underway is real and necessary. The question for capital allocators is not whether to invest, it is whether the infrastructure being built today is optimized for the architecture of the next three years or the next ten. History is consistent on this point: the companies that assumed their dominant architecture was permanent did not lead the transition that followed. The ones that built for flexibility captured the value. The winning bet is not the best engine. It is the most flexible chassis.

DATA AND SOURCE NOTES

Bipolar-to-CMOS transition
IEEE Solid-State Circuits Society historical records; Caltech and Stanford semiconductor research archives.

Mamba / state-space models
Gu and Dao, 'Mamba: Linear-Time Sequence Modeling with Selective State Spaces,' Carnegie Mellon / Princeton, 2023; Mamba-2 follow-on research, 2024.

Mixture of experts
Google Gemini technical report; Mistral AI Mixtral model card; DeepSeek-V2 and DeepSeek-V3 technical reports, 2024-2025.

Neurosymbolic / AlphaProof
DeepMind, 'AI achieves silver-medal standard solving International Mathematical Olympiad problems,' July 2024.

Diffusion language models
Inception Labs, Mercury technical documentation and benchmarks, 2025.

Inference infrastructure
Groq and Cerebras public technical disclosures; vLLM project documentation (UC Berkeley Sky Computing Lab).

World Models
Ha and Schmidhuber, "World Models," 2018; Wayve GAIA-1 (2023) and GAIA-2 (2025) technical reports; Meta V-JEPA 2 paper, 2025; Google DeepMind Genie 3 documentation, 2025; LeCun, "A Path Towards Autonomous Machine Intelligence," 2022.

PART 6: THE VALUE INVERSION

From Concentration to Diffusion

The companies that build the pipes rarely own the water. The question for every AI investor: are you funding pipes or water?

LAK ANANTH & JONATHAN GOLDBERG

From Concentration to Diffusion

The Value Inversion

In 1879, the Erie Railroad went bankrupt for the second time. The track was laid, the freight was moving, and the capital that built it was gone. A few hundred miles away, John D. Rockefeller did not own a mile of track. He used the rails, negotiated the rates, and built a national distribution business on infrastructure other people had paid for. By 1900, his fortune was the largest in the country and the Erie was on its third bankruptcy.

This is the pattern at the end of every infrastructure phase. The builders create the platform. The operators capture the value.

Almost everything people call 'AI adoption' today is concentrated in two places.

More than 50% of all token consumption on multi-provider platforms flows to software programming tasks. Content generation — drafting, summarizing, editing — accounts for most of the remainder. When industry surveys report that 92% of Fortune 500 companies are using AI, they are reporting on access, not depth. They are reporting on beachheads, not diffusion.

The concentration runs deeper than application categories. On the supply side, a small number of chip suppliers extract the highest margins in the stack. A handful of foundation labs control frontier model development. A concentrated set of hyperscaler customers account for the majority of AI infrastructure spending. By almost every metric — revenue, compute consumption, development investment — AI value capture in 2026 sits at the narrow end of a pyramid that is supposed to widen into a utility available to everyone.

This is normal. Every technology in its infrastructure-build phase looks like this. The commercial internet in 1997 was used primarily by people with enough technical sophistication to navigate it, for a narrow set of applications that worked on early dial-up connections. Semiconductors in 1980 were deployed by defense contractors, mainframe manufacturers, and a small number of technology companies. Electricity in 1900 powered factories and wealthy urban households, not farms, hospitals, or rural small businesses. Less than 10% of organizations today have successfully scaled AI agents in any business function, a figure that would look familiar to anyone who tracked enterprise software adoption in 1996 or 1997.

The concentration is a starting condition, not a destination. The question for investors is not whether diffusion happens; history is unambiguous that it does. The harder question, and the one this piece is about, is the value inversion that follows. The question is which industries cross the adoption threshold next, what those thresholds actually require, and where the value accrues when diffusion is complete.

The Five Thresholds

Each of the industries on the cusp of significant AI adoption has a specific set of requirements that current AI does not yet fully meet. The threshold analysis matters more than the technology timeline: these sectors will not adopt in the order that AI becomes capable, but in the order that cost, reliability, and regulatory conditions align simultaneously.

Healthcare is the most frequently cited next wave and the most structurally complex. Diagnostic assistance, clinical documentation, and drug discovery acceleration represent enormous addressable value. The FDA approved more than 500 AI-enabled medical devices through 2023, and diagnostic AI has demonstrated radiologist-level accuracy in controlled imaging studies. The threshold requirements are strict on three dimensions at once: inference costs must fall further to make per-patient economics viable at primary care scale, reliability standards require consistent performance rather than probabilistic output, and regulatory frameworks for AI-assisted clinical decisions are still being written. The opportunity is large, and the timeline is real. The adoption path runs through procurement committees, liability frameworks, and compliance infrastructure, not just engineering.

Legal services present a structurally different version of the same problem. Document review, contract analysis, and regulatory compliance are tasks where AI already demonstrates measurable value on benchmarks. The adoption constraint is auditability: a model that confidently cites a case that does not exist is not an assistive tool; it is a malpractice risk. There is also a structural friction specific to the legal industry that most technology forecasts underweight: the billable-hour model means that efficiency gains from AI compress revenue for the firms adopting it. Adoption will be driven by clients demanding lower costs and by new entrants structuring their practices around AI from the start, not by incumbent firms voluntarily reducing their own billing capacity.

Education has the most technically straightforward case and the most institutionally complicated one. Personalized tutoring at scale — a capable, responsive tutor available to every student, not only those whose families can afford one — is achievable with current AI capability. Early deployments have demonstrated real learning gains in controlled settings. The threshold requirements here are not technical: they are data privacy frameworks for minors, procurement processes in systems that move slowly, and the cultural question of how AI tools interact with teaching as a profession. The institutional change management is the constraint, not the technology.

Manufacturing and logistics may be the sector where adoption has already crossed the threshold in a way that has not generated proportional press coverage. Predictive maintenance, supply chain optimization, and quality control are use cases where integration with existing operational technology is challenging but tractable, the value per interaction is measurable and immediate, and the regulatory exposure is low compared to healthcare or legal. The next layer is real-time inference at the factory edge, which requires latency and cost reductions already underway. This is less a question of whether AI enters manufacturing than how quickly it reaches the long tail of mid-market manufacturers who cannot afford bespoke implementations.

Financial services sit across the threshold in some applications and short of it in others. Fraud detection and risk modeling are deployed at scale in large institutions today — these are not future applications. The next layer is customer advisory and credit analysis, where regulatory frameworks require explainability that current models provide inconsistently. Regulation B's adverse action notice requirements in consumer credit and fiduciary standards in investment advice create compliance floors that AI systems must clear before deployment. As in legal, adoption of the harder use cases will be driven more by competitive pressure from new entrants than by incumbent institutions voluntarily disrupting their own processes.

INDUSTRY	LEAD USE CASES	PRIMARY THRESHOLD	ADOPTION DRIVER
Healthcare	Diagnostic assist, clinical docs, drug discovery	Cost + auditability + regulatory approval	Regulatory clarity, cost declines
Legal	Document review, contract analysis, compliance	Auditability; billable-hour structural friction	Client demand, new entrant competition
Education	Personalized tutoring, adaptive curriculum	Institutional change management, data privacy	Evidence of learning gains, budget pressure
Manufacturing	Predictive maintenance, quality control, logistics	OT integration, edge inference latency	Measurable ROI, supply chain pressure
Financial Services	Fraud detection (deployed), advisory (emerging)	Regulatory explainability requirements	Competitive pressure from fintechs

Sources: FDA AI-enabled device approvals; Bain, 'AI's Trillion-Dollar Opportunity'; McKinsey Global Survey on AI adoption; industry analyst estimates

Three Patterns, One Conclusion

History offers three independent templates for how technology value migrates during and after a diffusion phase. They converge on the same answer.

The railroad parallel is the most instructive for investors because it is the most counterintuitive. Between 1865 and 1890, the United States built approximately 150,000 miles of railroad track in the largest infrastructure investment in American history to that point, and also one of the greatest capital destruction events in American financial history. Most of the major railroad companies went through bankruptcy at least once. The Erie Railroad went bankrupt twice. Capital was consumed at a rate that shocked even the investors who had funded it. And yet the railroads created something more valuable than anything the railroad operators captured: they created the conditions for an entirely different tier of wealth. Standard Oil used railroad logistics to build national distribution infrastructure no competitor could replicate. Sears used the railroad network to reach rural customers no retailer had previously been able to serve. Carnegie Steel used railroad demand and railroad cost structures to achieve dominance in an industry that became global. The infrastructure built the platform. It did not capture the value that ran on top of it.

The semiconductor transition offers a second template with a shorter and more precisely documented timeline. The shift from bipolar to CMOS transistors reshaped the industry over approximately fifteen years, from the mid-1980s to 2000. The companies that led the bipolar era — Fairchild, National Semiconductor — were not the ones that won CMOS. Intel built the microprocessor. TSMC built the foundry model. But the largest fortunes of the personal computing era were built not by Intel or TSMC, but

by the companies that used cheap CMOS compute to build software franchises that have proven more durable than any chip architecture: Microsoft, Oracle, Cisco. Later, Google and Apple. The chip transition was the prerequisite. The software layer was the prize.

The internet adoption curve provides the third template, and its most important lesson is about the relationship between technology availability and value capture. E-commerce was commercially available in 1994. It took fifteen years to reach 15% of U.S. retail. Enterprise SaaS companies had working products by 1999. It took a decade for SaaS to materially displace on-premises software. At each stage, the first-wave winners of the transition were not the same as the second-wave winners. Portals and early search engines dominated the first phase of internet adoption. Google, Amazon, Apple, and Meta defined the second phase. The companies positioned to capture durable value in the second wave were either not yet dominant or did not yet exist at the start of the adoption curve. The investment implication is direct: the timing of portfolio weight shifts matters as much as the direction of the technology.

The Value Inversion

Three historical patterns, three independent lines of evidence, one conclusion: infrastructure buildouts enable application-layer dominance. The cost stack, the architecture choice, the routing layer, and the inference-flexibility thesis from the prior pieces in this series all describe the mechanics of how the inversion happens. The companies that build the pipes rarely own the water.

> The companies that build the pipes rarely own the water. The question for every AI investor: are you funding pipes or water?

The AI transition to commodity economics is not yet complete. Based on the trajectory of inference cost declines and the competitive dynamics among foundation labs, open-weight models, and custom silicon programs, the transition to commodity-level pricing is likely to be completed by 2027-2028. That timeline is consistent with Google's own infrastructure reinvention, which required approximately five to seven years from the late 1990s to its full expression in 2006.

When that transition completes, the value stack inverts. Chips, models, and inference compute become competitive inputs rather than sources of monopoly margin. The prize shifts to the application layer: the companies that used cheap, reliable, widely available AI to build durable positions in industries that are only beginning to adopt it. The companies that will look, from 2030, the way Standard Oil looked from 1900, or Microsoft looked from 2000, or Google looked from 2010.

This is not an argument against infrastructure investment. The railroads were necessary. The CMOS fabs were necessary. The hyperscaler GPU clusters are necessary. None of the application-layer value in any of these historical cycles would have materialized without the infrastructure buildout that preceded it. But the investment question is not only whether to fund AI infrastructure, it is whether the portfolio is positioned for the value inversion that follows.

The concentration visible in AI today — in chips, in model development, in paying customers — is the signature of an infrastructure phase. It is the railroad era, the bipolar transistor era, the dial-up internet era. It is the phase that creates the platform. The diffusion phase that follows is the one that creates the lasting fortunes, in industries and applications that are currently underestimated precisely because they have not yet crossed their adoption thresholds.

Intelligence on Tap will be a reality when a doctor in a rural clinic, a small business owner without an IT department, and a student in an underfunded school district all have access to the same quality of intelligence that a Fortune 500 company deploys today.

The technology to deliver that is already in development. The economics are moving in the right direction at an unprecedented pace. The value question is who builds what runs on top of it.

Every technology cycle asks the same question. The answer, across railroads and semiconductors and the internet, has been consistent. The question for every investor in this space is the one that mattered in 1880, and in 1990, and in 2000: are you funding the pipes, or the water?

DATA AND SOURCE NOTES

AI usage concentration
a16z/OpenRouter, 'State of AI 2025' (100 Trillion Token Study); OpenAI, 'State of Enterprise AI 2025.'

Enterprise AI scaling
Bain, 'AI's Trillion-Dollar Opportunity 2024' (less than 10% of organizations have scaled AI agents in any function).

Healthcare AI
FDA AI-enabled medical device approvals database; Stanford HAI AI Index 2025.

Railroad history
Chandler, 'The Visible Hand: The Managerial Revolution in American Business'; U.S. Census Bureau railroad mileage data 1865-1890.

Semiconductor transition
Flamm, 'Mismanaged Trade? Strategic Policy and the Semiconductor Industry'; Intel and TSMC historical filings.

Internet adoption curves
U.S. Census Bureau e-commerce retail data; Gartner enterprise SaaS adoption surveys.

AI commodity economics timeline
Epoch AI inference cost trend data; Google Alphabet 10-K historical filings.

About the authors

Lak Ananth

GLOBAL MANAGING PARTNER, N47

Lak is the Global Managing Partner of N47, based in Silicon Valley, where he invests across applications, infrastructure, and systems. A lifelong builder — from electronics tinkerer to dotcom-era engineer — he has spent his career partnering with founders at companies including Verkada, Tractian, VAST Data, Cohesity, Aurasell, Meraki, and Thoughtspot. His investment lens is Product First: he looks for builders with the taste, judgment, and execution to make entire categories feel obsolete the moment users experience the new way.

Jonathan Goldberg

FOUNDER, D2D ADVISORY

Jonathan is the founder of D2DAdvisory a proprietary research firm, and a former sell-side semiconductor analyst. He brings the financial analyst's rigor: decomposing cost structures, stress-testing unit economics, and applying historical technology cycle models to AI's trajectory.