

THE AI VALUE CHAIN

ANTHROPIC IS WINNING EVERY BATTLE AND LOSING THE WAR

The Four-Year Window After the Quality Gap Closes

NABIL EL MAHYAOUI

Anthropic Is Winning Every Battle and Losing the War

Anthropic Is Winning Every Battle and Losing the War

The Four-Year Window After the Quality Gap Closes

Nabil El Mahyaoui

An essay on infrastructure, prestige, and the recomposition of the AI value chain

Copyright © 2026 Nabil El Mahyaoui.
All rights reserved.

Contents

1 The Chilean Coast **1**

1.1 The Karlsruhe Laboratory 3

1.2 The Four-Year Window 5

1.3 The Argument Nobody Wanted to Hear 7

2 The Prestige Trap **9**

2.1 The Numbers That Should Reassure . . 11

2.2 The Axis That Stops Mattering. 12

2.3 The Diagnostic Question 15

3 The Quartz Reprise **18**

3.1 DeepSeek V4 and Day Zero 20

3.2 The Toyota Fleet 23

3.3 The Software Lock 26

4 Asphyxiation by Success **29**

4.1 The Postmortem 30

4.2 Mythos in the Wild 33

4.3 The 2027 Tranche 35

4.4 Money Cannot Compress Physics 37

5 The Tenant, One Floor Up **40**

5.1 The Verticals 41

5.2 Anthropic Owns the Model. Just the Model. 45

5.3 The Mensch Diagnosis 48

5.4 Beautiful When Scarce, Fragile When Cheap 50

6 Capex Eats Opex 52

6.1 The Capital Transfer 53

6.2 The Markets Stop Reacting 56

6.3 The Career Map 58

6.4 The Geometry of the Recomposition . . 60

7 Epilogue: The Four-Year Window 63

7.1 What Anthropic Could Do 64

7.2 What They Almost Certainly Will Do . . 67

7.3 The Diagnostic for the Rest of Us 68

7.4 Closing the Coast 71

References 75

1

The Chilean Coast

In 1900, the most strategically valuable resource on Earth was bird droppings.

Chile produced two-thirds of the world's fertiliser. The deposits along its northern coast supplied over eighty percent of global fixed nitrogen. The trade was so profitable, and the geological windfall so concentrated, that the British Empire structured shipping policy around access to the islands. Germany built

its agricultural projections on the assumption of stable Chilean exports. The United States, Peru, and Bolivia fought a war—the War of the Pacific—over who would tax the guano-rich coast and the saltpetre interior.

If you held nitrate at the start of the twentieth century, you held the food supply of Europe and the explosive supply of every modern army. There was no substitute. The chemistry of life required nitrogen, and the only commercially viable nitrogen on Earth came from the slow accumulation of seabird excrement on a few square kilometres of South American rock.

It was the most prestigious commodity on the planet. And by 1934 it was effectively worthless.

1.1 The Karlsruhe Laboratory

In July 1909, in a laboratory at the Technische Hochschule in Karlsruhe, Fritz Haber demonstrated a tabletop apparatus that pulled nitrogen out of ordinary atmospheric air, combined it with hydrogen under pressure and at temperature, and produced ammonia at a rate of about a cup per hour.

A cup. Per hour. From thin air.

Carl Bosch, the BASF engineer who turned Haber's bench experiment into industrial reality, had a working pilot plant by 1913. Within four years of that pilot plant, the Chilean nitrate trade entered terminal decline. The collapse was not immediate, because no collapse ever is. The exporters had inventory, contracts, and political weight. They also had a real argument: synthetic ammonia was, by every clas-

sical metric of fertiliser quality, inferior to the natural product. Lower trace mineral content. Less complete nutrient profile. The Karlsruhe stuff was a chemical reduction of the rich, biologically aged guano of the Atacama coast.

They were correct.

It did not matter.

By 1925, Chilean production had fallen from its prewar peaks to 2.5 million tons. By 1934, that figure was 800,000 tons. The workforce on the coast collapsed from 60,000 to roughly 14,000. Prices fell from $45 a ton to $19. The most prestigious commodity of 1900 became, within a single human generation, a regional commodity sold at a discount to whoever still had infrastructure to ship it.

1.2 The Four-Year Window

There is a specific structural shape to a recomposition like this one, and it is worth naming precisely, because it recurs.

There is a moment when the new technology becomes physically possible. There is a later moment when the old industry collapses. The window between those two events is short. In the nitrate case, it was about four years from Bosch's pilot plant to the moment the trajectory was visibly inevitable. The Chilean exporters did not know they had four years. They thought they had twenty, or fifty, or forever, because the prestige product had been the prestige product for as long as any of them had been alive.

Inside that window, the incumbents tend to do one specific thing: they argue. They write

white papers about why the new product is inferior. They commission studies showing that under controlled conditions the natural fertiliser still outperforms. They lobby for regulation. They convince themselves that quality, in the end, will reassert itself.

It does not. Quality is a real axis. It is just not the axis that determines the outcome of a commoditisation. In a recomposition, value migrates from the prestige axis to the surrounding infrastructure—production cost, supply security, distribution scale, integration into the broader industrial stack. The Haber–Bosch process did not win because synthetic ammonia became higher quality than guano. It won because nitrogen, which had been geographically scarce, suddenly became atmospherically abundant. The Atacama could not compete with the sky.

1.3 The Argument Nobody Wanted to Hear

Here is the part of the story that survives.

The Chilean exporters spent their four-year window explaining that synthetic was inferior. They were not lying, and they were not even mistaken about the chemistry. They were making the only argument available to a producer whose entire economic logic depended on the prestige axis remaining the load-bearing axis.

Once the underlying axis shifts—once the question changes from *which fertiliser is best* to *which fertiliser scales*—the prestige argument becomes a sentence the market politely ignores. It is not refuted. It is not even debated. It is simply no longer the question being decided.

This is the analogy I want you to hold in your mind for the rest of this book. Because right now, in April 2026, an AI laboratory headquartered in San Francisco is making the prestige argument about its own product—accurately, eloquently, and with full justification. Their model is the smartest model. Their reasoning is the deepest reasoning. Their enterprise customers are paying premium prices because the quality is genuinely real.

By every metric the AI industry currently celebrates, Anthropic is winning.

I think they are going to lose.

And I think the reason is exactly what happened on the Chilean coast.

2

The Prestige Trap

On 7 April 2026, Anthropic announced that its annualised revenue run rate had crossed $30 billion. Four months earlier, at the close of 2025, that figure had stood at roughly $9 billion. They had tripled in a single quarter.

In the same week, they passed OpenAI in revenue for the first time in their history. The Ramp data on enterprise spending, which I read every Monday morning, showed that

roughly seventy percent of new enterprise AI contracts in the United States were going to Anthropic. The number of customers paying more than a million dollars a year for Claude had doubled in eight weeks, from around 500 to over 1,000. Their orchestration protocol, the Model Context Protocol, had crossed ninety million installations and had become the de facto standard that every other vendor was forced to support.

Claude Code, their developer product, hit one billion in annualised revenue within six months of public launch. By April it was the reference product for serious software work. Not the favourite. The reference. The thing engineering teams compare every other tool against.

And the model itself is, by my own daily use and by every honest benchmark, the best

general-purpose reasoning model on the market.

2.1 The Numbers That Should Reassure

If you are an Anthropic employee, an Anthropic investor, or an Anthropic enterprise customer, you read those numbers and you sleep well. The narrative is closed. The product is the best, the customers love it, the revenue is compounding, and the enterprise segment—the high-margin, low-churn, contractually sticky segment—is consolidating around Claude with a velocity nobody anticipated.

This is the position the Chilean exporters held in 1925. In absolute terms, in dollar terms, in market-share-of-the-prestige-axis terms, they were doing fine. The decline was not yet vis-

ible in the export ledgers. The order books were full. The customers, when surveyed, expressed loyalty and satisfaction. The product was, demonstrably, superior.

A diagnostic that only measures the prestige axis cannot, by construction, see a recomposition coming. That is the entire trap. If your dashboard is built around the question *are we still the highest-quality producer*, the answer will remain *yes* until the quarter the answer no longer matters.

2.2 The Axis That Stops Mattering

The relevant question, in any commoditisation, is not whether you are the best. It is whether the axis you lead on is the axis the market still pays a premium for.

For Anthropic, that axis is reasoning quality—and specifically, the kind of multi-step, long-context, agentic reasoning that distinguishes a frontier model from a competent one. Their bet, from the founders' first public talks onward, has been that this axis matters more than anything else, that it justifies premium pricing, and that the gap between the best and second-best will remain economically decisive.

That bet has paid off through 2026 Q1. It paid off because reasoning quality, in a world of nascent agentic workflows and high-stakes enterprise deployments, was genuinely the bottleneck. When the model fails, the agent fails. When the agent fails, the customer churns. So the customer paid the premium.

But three things are now visibly closing the gap.

The first is that DeepSeek V4, released on 24 April, ships at roughly six months behind the frontier on most benchmarks while pricing itself at one-tenth of the cost. We will examine this in detail in the next chapter. For the moment, hold this: a six-month lag at one-tenth of the price is not a competitor. It is a different category of product entering the same market.

The second is that Anthropic itself is rationing capacity. Their own customers are getting a measurably degraded model, not because Anthropic chose to ship a worse model, but because their infrastructure cannot serve the demand they have already won. We will examine that incident in chapter four.

The third is that the data substrate the frontier-quality argument depends on is itself running out. The 2026 Stanford AI Index documents what the field has begun calling *peak data*: the

exhaustion of high-quality human-generated training corpora. When the next quality jump requires data the laboratories no longer have, the differentiation that justifies the premium pricing begins to compress structurally.

Each of these three forces, taken alone, is manageable. Taken together, they are the four-year window.

2.3 The Diagnostic Question

The question every decider should be asking, and almost none of them are, is this:

> *If model quality commoditises in the next twenty-four months, which of my AI vendors still has a defensible position?*

The honest answer is uncomfortable. The vendor with the best model has the weakest structural position, because their entire premium depends on a quality gap that is closing. The vendors with weaker models but stronger infrastructure—Google most obviously, but also the vertically integrated Chinese stack—get stronger as the gap closes, because their value was never anchored to the prestige axis in the first place.

This is the prestige trap. Anthropic is winning the 2026 quarter on the only axis their cost structure permits them to win on. They are losing the 2027 recomposition on the axes they do not control.

Nvidia is the cautionary mirror. On 29 October 2025, Nvidia became the first public company in history to cross a five-trillion-dollar market capitalisation, having taken just three

months to add the trillion that separated four from five. The CEO of the most prestigiously valued company on Earth then went on a podcast and explained, on the record, that he was worried about precisely this scenario—that Chinese laboratories would optimise around an alternative software stack and bypass the moat his company had spent twenty years building.

Forty-eight hours after that interview, DeepSeek shipped V4.

When the most expensively valued beneficiary of the current arrangement is the one warning about its end, you are looking at a recomposition. The question is no longer whether it happens. The question is which of the laboratories survives the window.

3

The Quartz Reprise

In December 1969, Seiko introduced the Astron, the world's first quartz wristwatch.

It was not as accurate, in absolute terms, as a chronometer-certified Swiss mechanical movement. It was, on the prestige axis, the inferior product. It cost a tenth of a comparable Swiss timepiece. It was produced by stamped electronics on assembly lines, not by master watchmakers tuning balance wheels by hand.

It carried no patrimony, no inheritability, no cultural weight.

It was also more than precise enough for anyone whose actual purpose was reading the time.

By 1978, Switzerland's share of the global watch market had collapsed from over fifty percent in the 1960s to twenty-four percent. Swiss watchmaking employment fell from 90,000 in 1970 to 47,000 in 1980. The number of registered Swiss watchmakers fell from 1,600 in 1970 to 600 by 1983. Mechanical watch exports collapsed from 40 million units in 1973 to 3 million units a decade later. A ninety-two percent collapse in the volume segment, while Seiko alone shipped 25 million quartz units in 1980.

The Swiss kept the luxury market. They kept Patek Philippe, Vacheron Constantin, the high-prestige tier where the watch is jewellery rather than a timekeeping instrument. They lost everything else.

The lesson was not about quality. The lesson was about volume and price, mediated by the discovery that for most of the demand curve, *good enough* was sufficient.

3.1 DeepSeek V4 and Day Zero

On 24 April 2026, DeepSeek released V4: a 1.6-trillion-parameter mixture-of-experts model with 49 billion active parameters per inference, alongside a lighter V4-Flash variant of 284 billion parameters with 13 billion active. Both ship with a one-million-token context window. Both are open-weight.

That is the headline most coverage stopped at. The headline is not the news.

The news is that DeepSeek V4 is the first frontier-class model architected from the ground up to run natively on Huawei silicon. Not ported. Not adapted. Co-developed. The kernels were rewritten in CANN—Compute Architecture for Neural Networks, Huawei's direct equivalent of Nvidia's CUDA—in co-ordination with Huawei engineering teams over a multi-month delay that DeepSeek explicitly took to align the launch with hardware readiness.

Huawei confirmed the same day that its full Atlas Ascend SuperPod product line, from the 910C through the new 950 series, supported V4 natively at day zero. Reuters and Fortune both reported that AMD and Nvidia had been

deliberately excluded from the prelaunch pipeline.

This is not a Chinese laboratory reacting to American sanctions. This is a coordinated state-scale industrial play, executed with the patience and compartmentalisation that the Chinese system specifically rewards. The model and the silicon were designed for each other. The pricing reflects the integration: V4-Pro at $1.74 per million input tokens, $3.48 per million output. V4-Flash at fourteen cents per million input. A launch promotion dropped V4-Pro input pricing to roughly three and a half cents per million through 5 May. A representative inference workload that costs roughly $35 on a frontier OpenAI or Anthropic API costs roughly $5 on V4-Pro. Eighty-five percent compression.

DeepSeek has already committed to a further price reduction in Q2, contingent on the second-half scale-up of Huawei Ascend 950 production.

3.2 The Toyota Fleet

The Western reaction to DeepSeek V4 has fixated on whether the Huawei chips are as fast as Nvidia's. They are not. The Ascend 910C delivers about sixty percent of an H100 on inference workloads, with roughly 320 teraflops of FP16 throughput against the H100's reference capacity. On training, the gap is wider. By 2027, the projected per-chip gap is something like seventeen to one in Nvidia's favour.

The Western analysis stops there and concludes the Chinese stack cannot win. That conclusion is the analytical equivalent of arguing

that Chilean guano had a richer mineral profile than synthetic ammonia. Correct, irrelevant.

The actual economic question is different. A 910C runs roughly seven to ten thousand dollars per unit. An H100 runs about thirty thousand. Even adjusting for the sixty-percent performance ratio, the cost per token served on inference is forty to fifty percent lower on the Huawei stack. And inference, in 2026, is the bottleneck.

Training is a planned, asynchronous, off-peak operation. You can lease, queue, schedule, and amortise it. Inference is the load that runs continuously, whose demand explodes the moment a customer deploys an agent that thinks for hours, calls tools, retries, branches, and consumes tokens by the hundred million. Agentic search alone consumes orders of mag-

nitude more compute per query than classical search. Anthropic, OpenAI, and every serious enterprise deployment are now squarely inference-bound.

In an inference-bound world, the binding constraint is not peak capability. It is fleet availability at predictable cost. Ten reliable chips beat two best-in-class chips that are perpetually back-ordered. The taxi company with a fleet of dependable Toyota Corollas, none of them the fastest car on the road, beats the taxi company with two Ferraris that spend half the week at the garage.

The Swiss watchmakers in 1972 understood the prestige axis better than anyone in the world. They lost anyway, because the question was no longer *which movement is most precise* but *which producer can put a working timepiece*

on every wrist on Earth at a price ordinary people can pay.

DeepSeek and Huawei are now executing that exact play in the AI stack.

3.3 The Software Lock

The deepest layer of this story is not the chip. It is CUDA.

CUDA is the programming environment that Nvidia GPUs speak. Without CUDA, an Nvidia chip is a paperweight. With it, a chip is the substrate of the entire global machine-learning workforce. Nvidia released CUDA, free, in 2006—a decade before deep learning made GPUs the most contested commercial substrate on Earth. By the time the field exploded, every researcher, every PhD student,

every framework, every major model had been written for CUDA. The lock-in was not at the silicon layer. It was in the brains of a million developers.

That is the moat the five-trillion-dollar valuation is built on.

DeepSeek V4 is the first frontier-class proof that the moat is breachable. CANN is not a port of CUDA. It is a parallel software ecosystem, with its own kernels, its own compilation pipeline, its own scheduling primitives, optimised end-to-end with the underlying Huawei silicon. The model that runs on it was not adapted to it. It was born on it.

If the next generation of Chinese laboratories adopts CANN as their default, and if open-weight models become the substrate of enterprise inference outside the American sphere,

then the moat that justifies Nvidia's market capitalisation begins to leak. The CEO of Nvidia has said as much, on the record, the week before V4 shipped.

The Swiss did not lose to a better mechanical movement. They lost to a category that by-passed the question entirely.

4

Asphyxiation by Success

There is a specific way that a company can be destroyed by its own dominance, and Anthropic is currently demonstrating it.

The pattern goes like this. You build the best product in the category. The market consolidates around you faster than your infrastructure can scale. To preserve the appearance of capacity, you begin to silently degrade the product. The degradation is rational from an

operations standpoint and catastrophic from a reputational one. By the time you have built the infrastructure to serve the demand at the original quality, the customer has already seen what your product looks like under stress, and the trust that justified the premium pricing has structurally weakened.

That is the Anthropic story for the last six weeks.

4.1 The Postmortem

On 23 April 2026, Anthropic published an engineering postmortem on the Claude Code performance degradation that users had been complaining about for over a month. The postmortem confirmed three separate, overlapping issues:

On 4 March, the default reasoning effort for Claude Code had been quietly reduced from *high* to *medium*. The change was not announced. Users observed that the model had become measurably less thoughtful on multi-step problems. Anthropic later acknowledged that the change had been a wrong tradeoff, made to reduce latency.

On 26 March, a bug shipped that caused Claude Code to silently discard portions of its own reasoning history mid-session. The model became forgetful and repetitive within long agentic workflows. Users blamed themselves, then blamed prompt drift, then eventually blamed Anthropic.

On 16 April, a system-prompt instruction was added that capped intermediate reasoning to twenty-five words between tool calls. The cap

measurably hurt coding quality and was reverted four days later.

The bugs were not isolated. They are the visible surface of a single underlying pressure: the company does not have the GPU capacity to serve its own success at the quality its customers paid for. Each of the three changes was, at the operational level, an attempt to reduce per-request memory and compute footprint. Each was, at the user level, a silent quality degradation. The official line is a sequence of accidents. The structural read is a system in distress.

Dario Amodei said it himself, on the record, the week before the postmortem: compute is now the principal constraint on the company's growth.

4.2 Mythos in the Wild

The Mythos episode reads as a darker variation on the same theme.

Mythos is Anthropic's internal cybersecurity-capable model. The name surfaced publicly in April when Anthropic restricted access to it through a programme called Project Glasswing, releasing it on 10 April only to a closed consortium of partners. The official rationale was safety: the model was too powerful to deploy publicly without controlled-environment guardrails.

The internal rationale was infrastructure. A model with the kind of agentic depth Mythos was reportedly engineered for would, at any meaningful adoption rate, immediately overwhelm the inference capacity the company is already rationing for Claude. Restricting ac-

cess to a curated set of partners is, among other things, a way to cap concurrent inference load. Safety was the visible justification. Capacity was the load-bearing one.

Within days of the Project Glasswing launch, a Discord group obtained unauthorised access to Mythos. They had not breached Anthropic's primary infrastructure. They had used a data leak from Mercor, an AI-training contractor that worked with Anthropic, to reconstruct Anthropic's internal model-naming conventions. One member of the group was a third-party Anthropic contractor with peripheral knowledge of the deployment URL pattern. They guessed.

They guessed correctly. The group has had continuous access since 10 April and at the time of this writing was using the model for

benign tasks while waiting to see whether Anthropic could close the access path.

The story is not the breach itself. The story is what it reveals: that the most capable model in Anthropic's possession was so constrained at the infrastructure layer that it had to be hidden behind URL obscurity rather than served, and that its security perimeter was thin enough to fall to a guessable URL. The prestige laboratory cannot deploy its prestige product, because the prestige product would melt the substrate.

4.3 The 2027 Tranche

The infrastructure response is in motion. It is also, by physical necessity, late.

On 7 April, Anthropic, Google, and Broadcom announced the largest compute deal in Anthropic's history: 3.5 gigawatts of next-generation Google TPU capacity, supplied through Broadcom's networking and silicon partnership, scheduled to begin coming online in 2027. That tranche is on top of an additional 1 gigawatt scheduled for 2026 under the existing Google Cloud agreement. Broadcom committed to designing and supplying the underlying TPU silicon through 2031.

Read carefully, the deal is a confession. Anthropic does not have the compute it needs in 2026. The capital commitment to fix that condition is enormous and the operational lead time is twelve to twenty-four months. As one analyst put it: capital cannot compress physics.

Twelve to twenty-four months is a long time in this industry. It is, almost exactly, the four-year window of the Chilean coast collapsed by half. By the time the 2027 tranche delivers, the question will not be whether Claude is the best model. The question will be whether the segment that Claude was the best model for still pays a premium for being the best.

4.4 Money Cannot Compress Physics

The capital is not the problem. Amazon has invested twenty-five billion dollars in Anthropic. Alphabet has invested forty. Anthropic has signed compute commitments stretching to 2031. The cash exists.

The constraint is the calendar.

A datacentre takes roughly eighteen months to build out from broken ground to GPU racks under load. The TPU silicon Broadcom is designing for the 2027 tranche has tape-out, fabrication, validation, packaging, integration, and qualification cycles that cannot be parallelised past a certain bound. The grid interconnect has to be planned with regional utilities. The cooling has to be installed. The networking fabric has to be commissioned. None of these steps respects the urgency of a Q2 quality complaint.

Meanwhile, the demand that the infrastructure is being built to serve is doubling roughly every six to nine months. The capacity arriving in 2027 was sized for a 2025 understanding of demand. Even if the build executes flawlessly, it arrives undersized.

This is the precise window OpenAI is now exploiting with GPT-5.5, repositioning itself toward agentic execution where Anthropic was strongest, while Anthropic's enterprise customers begin quietly noting that the service quality they pay for is not the service quality they receive. The prestige laboratory is asphyxiating on its own success, and the suffocation is loudest exactly where it can least afford to be heard: inside the enterprise contracts that drove the $30 billion run rate in the first place.

5

The Tenant, One Floor Up

The previous essay in this series, on platform dependency, ended on a phrase that I want to pick up directly. It said: *the tenant syndrome does not disappear, it descends one floor in the stack.*

The wrappers—Cursor, Perplexity, the agent platforms—are the tenants of the model providers. The model providers, it turns out, are themselves the tenants of the hyperscalers. And in 2026, this is no longer an abstract

structural observation. It is the most powerful single predictor of which laboratories survive the recomposition that has just begun.

Look at the four protagonists of the current cycle and one pattern dominates everything else: vertical integration.

5.1 The Verticals

DeepSeek is integrated with Huawei. Silicon, software stack, model architecture, kernel-level optimisation—designed for each other from the foundation. There is no rent. The Chinese state has explicitly subsidised the construction of an end-to-end alternative to the American AI supply chain, and that alternative is now operational. According to the Center for Strategic and International Studies, Huawei has the capacity to ship close to

750,000 Ascend chips per year through 2026, with SMIC scheduled to push that baseline higher. The supply is sovereign, the price is below market, and the export ceiling for the Chinese sphere of influence has been removed.

Google is the second vertical. Their integration is older, deeper, and quieter. They design their own TPUs. They fabricate them through Broadcom's silicon partnership. They run the Google Cloud that hosts them, the Search and YouTube and Gmail estates that generate their training data, and the Android plus Chrome distribution that delivers the resulting products to roughly three billion daily users. On 22 April 2026, at Google Cloud Next, Google announced its eighth-generation TPU split into two specialised chips: the TPU 8t for training and the TPU 8i for inference. The train-

ing chip delivers 2.8x the performance of the previous generation at the same price. The inference chip delivers 80% better performance per dollar. Both ship at twice the perf-per-watt of the seventh-generation Ironwood. Google revealed during the same event that 75% of Google Cloud customers using AI now run on Google's own products, and that their model serving infrastructure handles more than 16 billion tokens a minute.

Notice what Google did during this entire cycle. They did not enter the public benchmark war. They did not chase headlines about reasoning depth. They quietly built the most fully integrated AI value chain on Earth and rented the capacity to their own competitors—including Anthropic, who is now signing 3.5-gigawatt deals on Broadcom-fabricated, Google-designed TPUs, hosted on

Google Cloud, financed by an ownership stake Google itself holds in Anthropic.

Then look at the non-verticals.

OpenAI is structurally a tenant. They depend on Nvidia silicon. They depend on Microsoft Azure capacity. They have re-positioned, this April, toward execution and the agentic Codex environment, leveraging the only durable distribution they own: Microsoft's installed base. That is a real asset. It is not vertical integration. The day Microsoft decides to push its own model harder, the OpenAI distribution channel narrows.

Anthropic is the most exposed of all. They rent compute from Google's TPUs, AWS's Trainium, and Nvidia's GPUs. They depend on the goodwill, the pricing, and the supply queue of every layer below them. They have

brilliant model research, the strongest enterprise traction in the industry, the de facto orchestration standard, and the highest-trust brand in safety-conscious procurement. They own none of the floors below the model floor.

5.2 Anthropic Owns the Model. Just the Model.

This is a beautiful position when intelligence is scarce. It is a fragile position when intelligence becomes cheap.

When a frontier model is the binding constraint on every interesting workload, owning the best model is sufficient. The hyperscalers will rent to you on favourable terms because your model demand justifies their capacity build-out. The wrappers will pay you premium rates because they cannot deliver their

product without yours. The enterprise buyers will absorb price increases because the alternative is shipping a worse product to their own customers.

When a frontier model becomes one of several adequate models—when V4 ships at six months behind at one-tenth the cost, when Gemini's enterprise platform offers the deepest integration, when GPT-5.5's agentic execution closes the gap that Claude opened—then the leverage of owning only the model collapses. The hyperscalers begin to negotiate harder. The wrappers begin to multi-source. The enterprise buyers begin to ask why they are paying twice as much for a marginal advantage that no longer dominates the workflow.

The 2026 Stanford AI Index documents this commoditisation in motion. Total AI compute

capacity has grown more than threefold every year since 2022. Eighty of the ninety-five most notable models released in 2025 shipped without disclosing their training code. The Foundation Model Transparency Index dropped from 58 to 40 points in a single year. When the laboratories stop telling you what is in their models, it is because the answer has become a competitive secret. When something becomes a competitive secret, it has become commercially valuable. And what has become commercially valuable in 2026 is not model architecture—it is the data pipeline and the inference infrastructure that surround the model.

Anthropic is winning the model layer. The model layer is precisely the layer that is, structurally, becoming one input among several.

5.3 The Mensch Diagnosis

Arthur Mensch, the CEO of Mistral, stood in front of an auditorium of École Polytechnique students last year and said, in plain language, that AI technology was going to commoditise. He said it not as a prediction but as a description of what was already underway. He pointed out, with the particular bluntness of someone whose own company is structurally exposed to the trend, that his own generation of engineers had successfully commoditised their own profession.

The observation is uncomfortable inside the industry, because the people who run the laboratories are the people who built the prestige axis. They are personally and economically invested in the proposition that frontier in-

telligence remains scarce, expensive, and centralised.

It will not. The economics of inference, the open-weight trajectory of the Chinese stack, the arrival of Google's split-purpose TPU eighth generation, and the data exhaustion documented in the Stanford report all point in the same direction. The cost of intelligence collapses. The differentiation migrates to whoever owns the data, the silicon, and the distribution.

Anthropic is on none of those three axes.

5.4 Beautiful When Scarce, Fragile When Cheap

There is a question worth asking precisely. Why did Google not enter the public benchmark war?

Because Google understands the structure of the recomposition better than any other actor in the industry. The benchmark war is the prestige axis. Whoever owns the prestige axis, in a recomposition, owns the layer that is about to lose its premium. Google is willing to ship Gemini at a quality slightly below frontier, because frontier is not where their profit will come from. Their profit will come from selling capacity to the laboratories that are competing on frontier, while their own integrated stack absorbs the volume of inference workloads from below.

Google sells pelles. The metaphor is overused, but it remains accurate. Anthropic is digging.

This is why I think Anthropic's position is paradoxical rather than triumphant. Every metric the industry currently celebrates—revenue, enterprise penetration, model quality, brand trust—is the metric Anthropic optimises for. Every metric that determines who survives the next two years—supply security, data ownership, inference cost structure, distribution leverage—is a metric Anthropic does not own. They have rented every floor of the building they built their reputation on.

When the tenant on the top floor sells the most prestigious apartment in the city, the rent is still owed to whoever owns the building.

6

Capex Eats Opex

There is a number you should hold in your mind before reading the rest of this chapter.

In 2026, the four largest American hyperscalers—Amazon, Alphabet, Microsoft, Meta—will spend close to seven hundred billion dollars in capital expenditure. Roughly seventy-five percent of that, about $450 billion, will go directly into AI infrastructure: datacentres, GPUs, custom silicon, networking fabric, cool-

ing, and the grid interconnect required to feed it all. Including the next tier of hyperscalers, the figure pushes past $750 billion. Some forecasters have it crossing a trillion within twelve months.

This is the largest single capital reallocation in the history of the technology industry. And it is being financed, in significant part, by a deliberate transfer of resources from the labour line of the income statement to the silicon line.

6.1 The Capital Transfer

In April 2026, Meta announced that it would lay off roughly 8,000 people—ten percent of its workforce—effective 20 May. It also withdrew approximately 6,000 open positions. The official language used the word *efficiency*. The actual line item in the company's projections

is unmistakable: Meta's capital expenditure for 2026 has been guided to between $115 billion and $135 billion, up from $72 billion in 2025. The doubling has to be financed somewhere. It is being financed, in part, out of the salaries of the engineers, marketers, and middle managers who built the company that now requires the capital.

In the same week, Microsoft announced its first-ever buyout programme, targeting employees at the level of senior director and below whose age and tenure summed to seventy or more. Snap announced layoffs of sixteen percent of its workforce, with CEO Evan Spiegel explicitly citing AI-driven efficiencies. Nike cut 1,400. GoPro cut 25%. By the end of the third week of April, more than 96,000 tech employees had been laid off in 2026 alone, across roughly 250 separate corporate actions.

This is not a recession. The companies announcing the cuts are reporting record revenue, record margins, and record capital plans. They are not retrenching in response to bad news. They are reallocating capital toward an infrastructure build-out that they expect to compound for the next decade.

The income statement is being rewritten in front of us. Capex eats opex. People are now a category of cost that, increasingly, can be substituted for silicon at favourable terms. The substitution is not metaphorical. Each role eliminated finances, with mathematical specificity, a fraction of a server rack in Texas, Virginia, or northern Sweden. The rack does not take holidays. It does not produce children, request medical certificates, or unionise. It runs at full capacity, twenty-four hours a day,

against a unit cost that compresses with every new generation of inference optimisation.

6.2 The Markets Stop Reacting

The clearest signal that this has become structural rather than episodic is the response of the equity markets.

A decade ago, an announcement of mass layoffs by a major technology company would have produced a speculative pop in the stock as investors anticipated margin expansion. In April 2026, Meta dropped two percent on its 8,000-person announcement. Microsoft dropped four percent on its buyout programme. The market treated the news not as a positive surprise but as a routine line item, already priced in, already understood as the new operating norm.

What has changed is the shared understanding among investors, executives, and labour markets that this is no longer the cyclical retrenchment phase of a tech bear market. It is the steady-state operating posture of the AI-infrastructure era. The companies are not laying people off because they are in distress. They are laying people off because the economics of running a hyperscale technology business now favour continuous substitution of headcount with infrastructure, indefinitely. The 2026 layoffs are not the bottom. They are a midpoint.

Once a transformation of this magnitude becomes priced in, it stops generating volatility. The absence of a spike on the news is, in fact, the strongest evidence that the new regime has been accepted as inevitable.

6.3 The Career Map

The implications for individual careers, in plain language, are unkind.

Roles that consist of standardised, high-volume work—first-line support, basic data analysis, recurring reporting, routine quality assurance, junior HR, the entire stratum of middle management whose function is to relay decisions made above to people who execute them below—are now in direct competition with inference at $1.74 per million input tokens. DeepSeek V4 is not as smart as Claude. It does not need to be. For the standardised, high-volume operations that consume the bulk of corporate labour cost, it is sufficient. And the price of sufficient just collapsed by eighty-five percent.

The roles that remain defensible are not the roles that execute decision trees. They are the roles that design the decision trees, that govern the agentic systems that execute them, that hold accountability when those systems drift, and that maintain the human relationships—trust, judgement, negotiation, taste—that no current frontier model is remotely positioned to absorb.

The architecture of an enterprise's agentic stack is now a defensible, scarce, and high-leverage skill. The execution of any task within that stack, increasingly, is not. If your daily work is the execution of a documented procedure, the procedure can be automated. If your daily work is the design and supervision of the system that executes the procedure, you are operating at the layer the substitution cannot yet reach.

This is not a permanent settlement. The line between *designs the system* and *operates within the system* will continue to migrate upward. But for the next 24 months, the line is approximately legible. If you can identify which side of it you are on, you can act on the diagnosis.

If you cannot, the diagnosis acts on you.

6.4 The Geometry of the Recomposition

The capex transfer is not happening in a vacuum. It is the financial mechanism by which the AI recomposition is being executed. Every dollar shifted from payroll to silicon is, simultaneously, a dollar that finances the inference infrastructure that makes the next layer of substitution economically viable. The transfer is self-reinforcing.

Anthropic's predicament, examined across the previous chapters, is the prestige-tier symptom of the same structural force. Anthropic is paying ten-figure compute bills to the hyperscalers, who are funding their TPU programmes with capital reallocated from their own labour lines. The hyperscalers' silicon then shows up, eighteen months later, as additional inference capacity that compresses the per-token price the laboratories can charge—which forces those laboratories to ration the model quality their customers experience, in order to preserve the margins that justify their valuations.

It is a closed loop. Each of the actors inside it is behaving rationally on their own terms. Each is contributing, by paying their own bills on time, to the recomposition that will bypass several of them within twenty-four months.

The Chilean exporters, in 1925, were also behaving rationally. They were shipping nitrate to the customers who still wanted nitrate, at the prices the market still bore, with the workforce they still employed. The system that was about to make them obsolete was being financed, in part, by the orders they were still fulfilling.

This is what a recomposition looks like from inside it. Everyone is paid. Everyone is performing. Everyone is, very politely, financing the substrate that is about to bypass them.

7

Epilogue: The Four-Year Window

The Chilean nitrate exporters had four years between Bosch's pilot plant and the moment the trajectory of their industry became visibly inevitable. They spent those four years explaining that synthetic ammonia was inferior. They were correct. It did not matter.

Anthropic has roughly the same window. They are spending it explaining that their model is smarter. They are also correct.

It will not, by itself, matter.

7.1 What Anthropic Could Do

Let me be honest about the strongest counter-case to everything I have written. It is not that Claude will remain the smartest model. That premise will compress whether Anthropic likes it or not. The strongest counter-case is that Anthropic is not trying to win the prestige-quality axis at all. They are trying to win the high-stakes-deployment axis, where reasoning depth, alignment, and trustworthiness matter more than benchmark scores or token cost.

If they are right about that, they have a defensible position that does not require them to own silicon. The argument runs like this: in the workflows that actually generate enterprise revenue—regulated industries, legal review, financial advisory, agentic deployment in critical operations—the marginal cost of model error is enormous, the marginal value of reliability is structural, and the customer will pay a multiple of the cheap-inference price for a model they can trust to behave correctly when nobody is watching.

This is not a stupid bet. It is the bet that built the $30 billion run rate. And there is real evidence that it works at the high end. Roughly 1,000 enterprise customers paying more than $1 million a year is not an accident. It is a market segment that has, after due-diligence cycles measured in months, concluded that

Claude is the model they trust with their most consequential workloads.

The question is whether that segment is the entire market or a slice of it that becomes proportionally smaller as inference commoditises everything around it. If the slice holds, Anthropic survives the recomposition by retreating into the high-trust luxury tier—the equivalent of Patek Philippe and Vacheron Constantin in the post-quartz Swiss industry, who are still profitable five decades later because the buyers of mechanical complications were never the buyers of timekeeping.

That is a real exit. It is also a much smaller company than the $30 billion run rate currently implies.

7.2 What They Almost Certainly Will Do

The honest read of corporate behaviour is that companies in Anthropic's position rarely choose the small, defensible, luxury-tier path while the large, exposed, prestige path is still profitable. The pull of the current numbers is too strong. The investors are calibrated to the current trajectory. The internal incentive structure rewards capturing the next quarter's enterprise pipeline, not contracting into a defensible niche.

So Anthropic will almost certainly continue to do what every winning incumbent does inside a recomposition: invest more, scale faster, defend the prestige-quality axis, and explain, accurately and at length, why the cheap alternative is inferior. The Project Glasswing release of Mythos was, in miniature, exactly this

pattern. So was the 3.5-gigawatt Broadcom deal. Both are correct moves on the prestige axis. Neither addresses the structural question of whether the prestige axis remains the load-bearing axis.

Inside Anthropic there are, almost certainly, people who see this clearly. There are also commercial and capital-structure pressures that make the alternative very difficult to execute on the timeline available. This is how recompositions kill companies that nobody, internally, ever doubted.

7.3 The Diagnostic for the Rest of Us

For the reader who is not running a frontier laboratory but is making technology choices on behalf of a team, an enterprise, or a career, the practical question is the same one I posed

in the previous essay. Which layer of the value chain are you betting on?

If your strategy depends on a single American frontier laboratory continuing to dominate model quality at premium prices, you are on the prestige axis. That axis is structurally compressing. If your strategy depends on the hyperscaler infrastructure that the prestige laboratories rent, you are one floor closer to durable. If your strategy depends on data substrate—proprietary corpora, customer interaction logs, instrumented deployment telemetry—that no model provider can replicate, you are operating on the axis that the Stanford report identified as the new locus of differentiation.

The 2026 deciders should be asking three questions in order. Where is my data substrate, and is it actually proprietary? Where is my in-

ference capacity, and is the cost trajectory on it under my control or someone else's? And finally—only finally—which model do I run? If you ask the third question first, you are optimising the layer that is about to commoditise. If you ask the first two questions first, you are positioning for the recomposition that has already begun.

For the European reader specifically, the exposure is doubled. European enterprises run American wrappers on American models on American clouds, three layers of dependency, with the middle layer—the model—in the most structurally fragile position of all. The Digital Markets Act regulates app stores and search engines. It does not yet regulate the model layer or the inference layer. That is where the leverage now sits, and it sits with

operators who answer to neither European regulators nor European customers.

Arthur Mensch said, in front of his own students, that his generation of engineers had successfully commoditised their own profession. The same engineers can choose, deliberately and with eyes open, where in the recomposed stack they want to position the next decade of their work. Or they can keep optimising their CV for the layer that the recomposition is about to bypass.

7.4 Closing the Coast

I keep returning, in my own thinking, to a single image from the Chilean coast.

It is 1925. The export ledgers in Antofagasta still show a profitable year. The freighters are

loading nitrate for European fertiliser markets at $45 a ton. The miners on the shore are still employed, still paid, still extracting from the same deposits that their fathers worked. The buyers in Hamburg and Liverpool are still placing the orders. The system is intact.

Inside the BASF plants in Ludwigshafen, the new Haber–Bosch reactors are running continuously, twenty-four hours a day, pulling nitrogen out of the air at a unit cost that the Chilean producers cannot match and would not believe if they were told. The trajectory is decided. The collapse is mathematical. None of the actors on the ledger sheets in Antofagasta have integrated the trajectory into their plans, because the ledgers have not yet caught up to the chemistry.

The four-year window is the window between the moment the chemistry decides and the moment the ledgers admit it.

Anthropic is currently posting the best ledgers of any AI laboratory on the planet. The chemistry, in a different building, has already decided.

Whether the company chooses to integrate the chemistry into its strategy, or spends its window arguing the prestige axis to a market that has stopped listening, will determine whether the $30 billion run rate is the foundation of the next era or the visible peak of a position that was already, structurally, beginning to erode.

The Chilean exporters, in 1925, did not know they had four years. Anthropic, in 2026, has the advantage of being able to read the history.

That advantage only matters if they choose to use it.

References & Further Reading

Anthropic (2026). "An update on recent Claude Code quality reports." Engineering postmortem, 23 April.
The official confession that the prestige product was being silently degraded for six weeks while the company explained nothing to the customers paying premium prices for it.

Anthropic, Google, Broadcom (2026). 3.5-gigawatt TPU capacity announcement, 7 April.
Read the SEC filing rather than the press release.

The 2027 delivery date is the load-bearing detail; everything else is mood lighting.

DeepSeek (2026). V4 technical report and Huawei Ascend co-launch, 24 April.
The first frontier-class model architected from the ground up to bypass the CUDA ecosystem. The pricing schedule is a strategy document disguised as a billing page.

Stanford HAI (2026). *The 2026 AI Index Report.*
The annual reckoning. This year's editions document the data exhaustion, the transparency collapse, and the 30-fold compute expansion since 2021. Read the chapter on training data carefully.

Smil, V. (2001). *Enriching the Earth: Fritz Haber, Carl Bosch, and the Transformation of World Food Production*. MIT Press.
The definitive account of the nitrogen recomposi-

tion. Worth reading in full. The economics of the Atacama collapse are documented chapter by chapter.

Christensen, C. M. (1997). *The Innovator's Dilemma*. Harvard Business Review Press.
The foundational text on why the most rationally managed incumbents lose recompositions. The Anthropic chapter has not yet been written, but the template is here.

Seiko Museum Ginza (ongoing). *The Quartz Crisis and Recovery of Swiss Watches.*
Primary source from the winning side of the 1969–1980 recomposition. Read it next to Swiss industry-association histories from the same period for the full stereoscopic view.

Mensch, A. (2025). Address at École Polytechnique.
The CEO of a frontier laboratory telling engineer-

ing students that his generation has commoditised their own profession. Worth re-watching annually.

Epoch AI (2024–2026). *Hyperscaler capex trends and the trajectory toward peak data.*
Independent quantification of compute and data exhaustion curves. The most clear-eyed numerical work on the recomposition currently in print.

Thompson, B. (ongoing). *Stratechery*.
The standing reference for platform-economics analysis. The recurring essays on the AI value chain are essential.

Reuters & Fortune (2026). Coverage of the DeepSeek V4–Huawei prelaunch coordination.
The reporting that confirmed the deliberate exclusion of AMD and Nvidia from the prelaunch pipeline. The detail that turns the launch from a product release into an industrial-policy event.

Center for Strategic and International Studies (2026). Reports on Huawei Ascend production capacity and the supply chain consequences of US export controls.
The dataset behind the 750,000 chips per year figure. Also the most reliable source on the unintended consequences of the sanctions regime.

About the Author

Nabil El Mahyaoui

https://nabil-el-mahyaoui.com/

https://www.linkedin.com/in/elmahyaoui