Ali Ihsan Nergiz

A MANIFESTO · NOVEMBER 2026

Compute is all that matters

There is no moat. The only reality is how much compute allocation you have and how efficiently you're able to use it.

A 30-minute read, 73% human-written (Pangram), with graphics and earnings-report analysis by AI. Despite my big bet on AI in this report, clear human writing and communication is the only thing to keep human (for now).

MAIN TAKES

01

Why it's not a bubble

Amazon, Alphabet, Microsoft and Meta spent $202bn on capex in three months, while AWS earned a 38.7% operating margin. The sell-off is a temporary cash problem, not falling demand: companies are cutting staff and borrowing to buy compute. Used H100s cost 16% more than a year ago, so the hardware is holding its value. 01 →

02

There is no cap on the demand

Weekly tokens on OpenRouter rose 33-fold in a year. The price of intelligence has fallen, but the money we put into models keeps rising, because what they can do has risen faster. Computer work is under 6% of US wages yet 36% of Claude's use, so most of the demand is still locked. 02 →

03

Human attention is not the limiter anymore

Agent traffic passed human traffic on OpenRouter in February 2026. One person can follow only three to five agent sessions, so work is moving to swarms that run without a human in the loop. OpenAI put 10,000 agents on one proposed proof, and compute is the only limit left. 03 →

04

When the scaling is going to stop?

Anthropic's compute spend per dollar of revenue fell from $2.51 to $0.58. The length of task AI agents can finish doubles about every four months, and models now help build the next generation. So far, each step makes models more useful and cheaper to serve, which feeds more demand. 04 →

05

There is literally not enough compute

Anthropic bought $5.0bn of short-term compute at a premium in one quarter, and Microsoft and TSMC say demand exceeds supply. Getting chips online is its own limit, from gas turbines to transformers. The shortage is spreading to memory and CPUs: Apple calls memory pricing "unprecedented". 05 →

06

There is no moat

Open models now carry about 70% of OpenRouter's tokens and trail the frontier by about four months. Years of human effort are turning into one prompt, and Anthropic's own filing warns that model know-how may spread. What's left is how much compute you have and how efficiently you use it. 06 →

07

In both cases, spending goes to compute

Revolut serves about 2,000 engineers from an open model on 56 B200s, for a tenth of its old bill. Companies with enough users and sensitive data will run models themselves, on rented GPUs. AI product companies will do the same or spend more, and either way the money goes to compute. 07 →

08

There are very different ways of being part of this

Nothing is more important than being part of the compute build-out. Anthropic alone has committed $895bn to compute and infrastructure. You can be part of it through energy, chips, data centres, clouds, labs, AI companies, or by owning the hardware yourself. 08 →

One prompt away

I quit my job as a medical doctor in the middle of 2023. Seeing how fast AI was moving, I realised this is going to be the most important technology of our lifetime. I hadn't realised how big it was going to be.

The first thing I built was a relatively complicated AI product at the time. It took three months of talking to ChatGPT, copy-pasting code and trying to understand things, and I had to work with a couple of engineers.

I built my own startup, I became an engineer, and now I'm a founding engineer at an AI startup. I'm now certain that there is no task that isn't one prompt away, as long as you have enough compute.

10× my contribution since I joined in January 2026, with the biggest jump in September

For each pull request, the smaller unit of work an engineer produces, I now get several AI models working on the same thing from code implementation to generating on demand videos about it. This was unthinkable two years ago.

While I contribute as much as 10 engineers at work, my old startup runs on agents, very profitably, with not much human involvement. I also made my own mobile game in a couple of weeks of free time; according to Claude, it would have taken a small games studio a couple of years and a couple of million dollars.

It isn't only me

How much of the work machines now do

>90%
of Anthropic's code contributions
written by Claude, September 2026

>45×
code per engineer per day
at Anthropic, compared with 2024

>40%
of GitHub pull requests
started by an agent, September 2026

34.3bn
commits on GitHub in a year
up from 12.9bn the year before

Source: Anthropic amended S-1, 2 Nov 2026; GitHub Octoverse 2026.

01

Why it's not a bubble

Right now there's been a big sell-off in the Nasdaq over the last week through the earnings reports. People believe capital expenditure may be limited, and that the companies might face the realities of the economy. Interesting contrast is looking at companies spending and what they are betting on. They are reducing the free cash flow, laying off people and trying to get any money they can to spend more on compute.

People out, compute in

~8,000

jobs cut at Meta

April 2026; Zuckerberg named compute and people as its two big costs

60,000

Microsoft staff eligible for buyouts

26 October 2026, after 4,800 job cuts in July

$170bn

Microsoft's 2026 capex guidance

given two days after the buyout offer

21%

of US job cuts cite AI

120,136 cuts, January to September 2026

Source: Reuters via Fox 9; Microsoft; Reuters via CNA; Microsoft Q1 FY2027 call; Challenger, Gray & Christmas.

Contrary to the bearish belief, markets are not pricing a lower demand. What we see is a temporary cash problem, mostly because companies spend more while revenues actually grow. This volatility is a good period to spend more on compute.

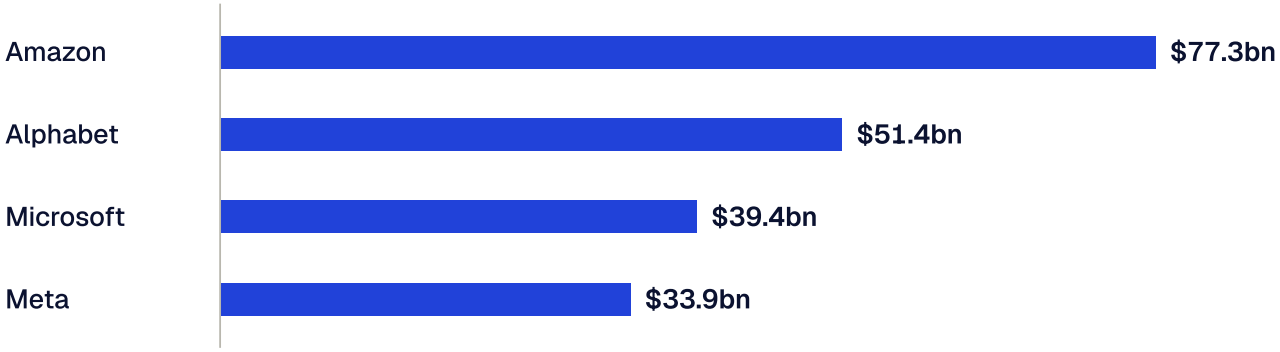
This turbulence shows the stark reality of the world. Big companies are willing to take a big hit, as we've seen with Google, because they know that spending on compute is the only thing that matters in the long term. Memory and compute are going to get even more expensive from now on. The winners of the future are going to be the people who own these chips, and the people who are able to do work on them. That's the pure reality.

THE HIT GOOGLE IS TAKING

Google's search revenue still grew 12% in the third quarter. The hit is in its cash: free cash flow went to –$9.1bn, it raised $12.4bn by selling shares and $28.7bn by selling bonds, and losses at the Alphabet level, mainly shared AI research, grew from $4.5bn to $7.7bn a quarter (Alphabet Q3 2026 release). Its shares fell 12% the day after the results.

Four companies spent $202bn on capital expenditure in three months

Cash capital expenditure, quarter to September 2026, $bn



Source: Amazon 8-K; Alphabet Q3 release; Microsoft Q1 FY2027; Meta Q3 results.

Cloud margins are high while Amazon's free cash flow is negative

38.7%	+48%	$677bn	–$42.4bn
AWS operating margin	**Azure growth**	**Microsoft's commercial backlog**	**Amazon's free cash flow**
Q3 2026, on $46.0bn of revenue, up 39%	year on year, quarter to September 2026	up 73%, about 2.3 years on average	over the twelve months to September 2026

Source: Amazon Q3 2026 release; Microsoft Q1 FY2027. Backlog is remaining performance obligations.

Sticky inflation and owning the hardware

The world has been experiencing what inflation means since COVID. Everyone hoped that inflation would go down. It's not going down. The part of the market that is not related to AI and compute is bearish, while AI growth seems to be the only thing that matters. The investment thesis has been very strong with AI compute and inference, and it is going to be even stronger regardless of how the market turns.

In this scenario, you want to own hardware. This can be through physical ownership or long term commitment but owning the hardware is owning the means of work. In this environment, investing in AI compute infrastructure is the only thing that will make money. And being in this chain will be even more important as what models are able to do gets increasingly better.

STICKY INFLATION

Inflation is above the 2% targets almost everywhere: 3.7% in the US, 3.5% in the UK and 3.8% in the euro area in September, and the IMF expects 4.7% globally this year. Much of it is energy, and core inflation is lower than a year ago in the US and the UK. Meanwhile US producer prices for computer storage rose 87% in a year (BLS via FRED), and Micron and SK hynix expect memory to stay short into 2027 and beyond.

This asset class is becoming the only thing that holds, and will hold, value for many people.

On the personal level I'm often cautious about committing money and stayed away from any speculative investments even when they made sense. Now the only thing I'm trying to do is get credit and buy more compute. The hardware you buy appreciates because demand is incredibly high. You can turn the equipment you buy into money by running AI models on it while the underlying asset appreciates in value.

What used hardware and rents did in 2026

~2×

launch price for a used Mac Studio

M3 Ultra 512GB, eBay sales, Sept–Oct 2026, after Apple withdrew it

+16%

used H100 price in a year

to August 2026; still about 40% of its 2023 price

+20%

on renewed GPU contracts

Oracle, on renewed and resold GPU capacity

+38%

one-year H100 rent

October 2025 to March 2026

Source: MacPro-LA eBay tally; American Compute; Oracle Q1 FY2027; SemiAnalysis.

02

There is no cap on the demand

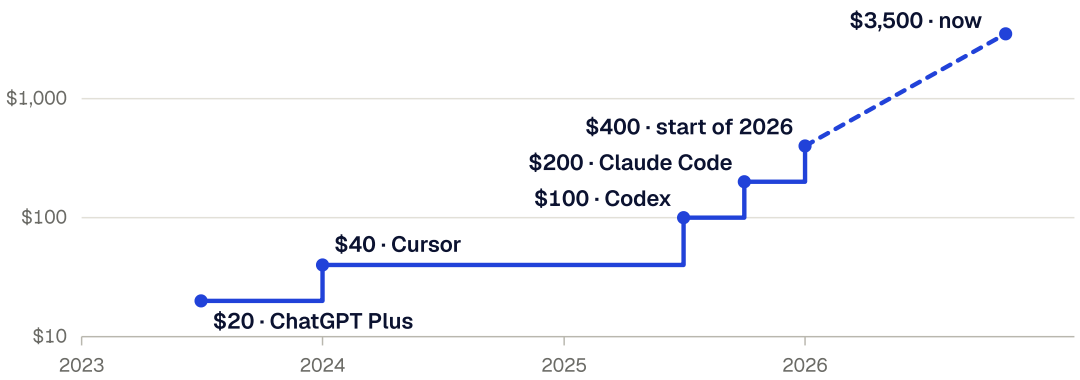
In 2023, I was spending twenty dollars on ChatGPT subscription, and one subscription was enough. There wasn't much of a limit to how much you could do. Not just because OpenAI was generous, but also the value you could get from a model was not that much. You were on one chat, trying to do one thing, in slow mode.

Then I started using Cursor, which had usage limits, and I moved to the forty-dollar plan. When Codex came out I was paying a hundred dollars, around summer 2025, and that was magic. Then the two-hundred-dollar Claude Code plan seemed very attractive, and at the start of this year I was using two Claude subscriptions, around four hundred dollars.

Right now, between my job and personal projects, I spend $3,500 a month on AI subscriptions. If you include the GPUs I use from Modal and other cloud services, it is well beyond $4,000. I spend a couple of billion tokens a day across multiple Codex and Claude accounts.

My monthly AI subscriptions went from $20 to $3,500

US dollars a month, log scale. With GPUs and other cloud services, it is well beyond $4,000



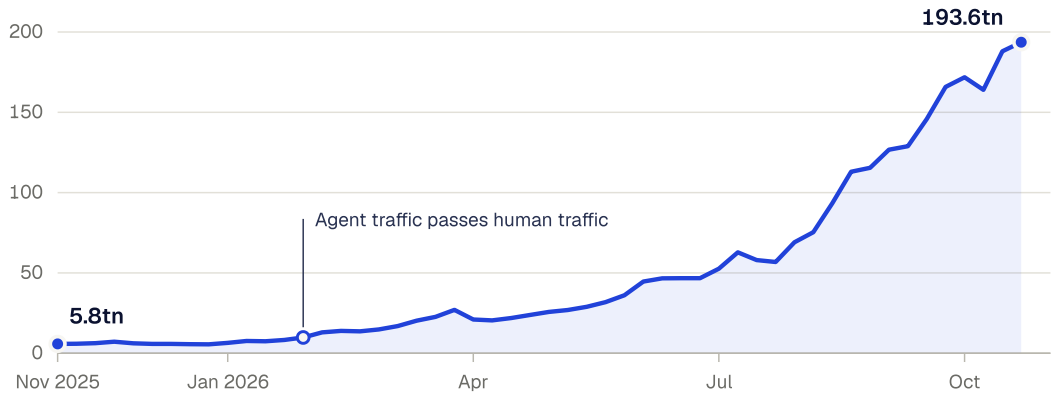
Source: my own spending, from memory. Dates are approximate; the 2026 rise isn't dated month by month, so it is drawn dashed.

When I did the jump from 40 dollars to 100 dollars, it was a big decision. But 400 to 4000 was not. The models are now so good, and the value that they bring is so high.

The price of intelligence has been reduced, but the money we put on models increased, because what models can achieve increased much more significantly.

Weekly tokens on OpenRouter rose 33-fold in a year

Tokens processed per week, trillions



Source: OpenRouter rankings; OpenRouter on agent traffic. Complete weeks, 3 Nov 2025 to 1 Nov 2026.

Models being much better means they spend more tokens, and even if they are more token efficient, they will spend more time on the task. People's spending is going up.

> **There is no world where we will not consume the endless compute. It is just the natural way things are going to progress.**

Better models do more work per task

3×

the agentic turns per task

Sonnet 5 against Sonnet 4.6, with 40% more output tokens

330×

Google's monthly tokens processed, in two years

9.7 trillion (May 2024) to 3.2 quadrillion (May 2026)

~900%

Anthropic's net revenue retention

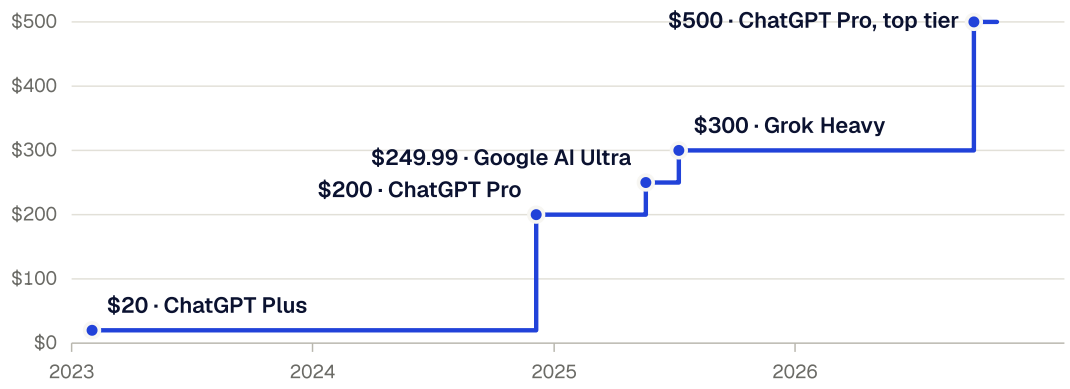
September 2026

Source: Artificial Analysis, 30 Jun 2026; Google I/O 2026; Anthropic amended S-1.

It's becoming a norm to pay for higher tiers as an individual.

The most expensive individual AI plan went from $20 to $500 a month

Highest list price of a one-person monthly plan, US dollars



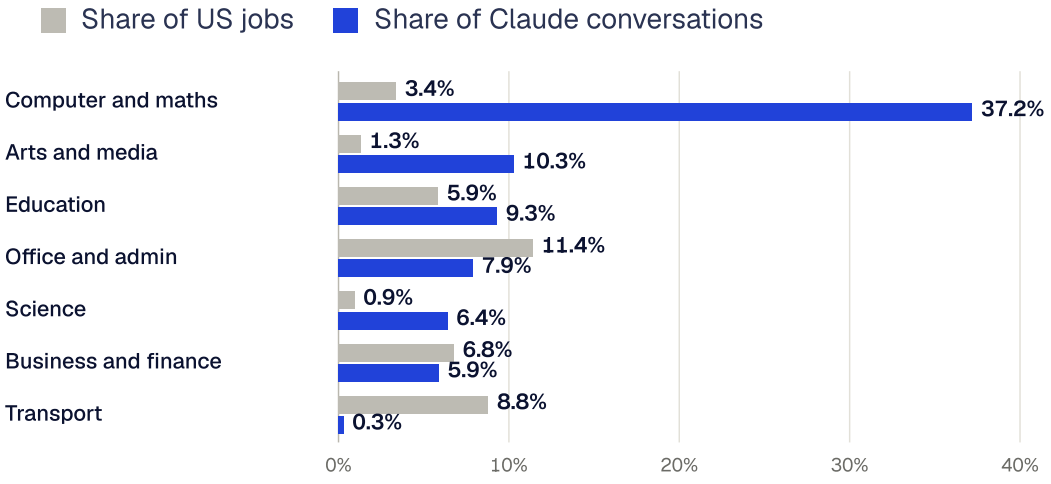
Source: Anthropic Max (TechCrunch); xAI (TechCrunch); Google AI plans; ChatGPT pricing.

The demand that is still locked

AI token spending is mostly limited to computer tasks so far. If we take this example and apply it to other sectors, it's very easy to see how much demand is still locked.

Computer jobs are 3% of US work but 37% of Claude conversations

Share of Claude conversations and of US jobs, by occupation group, %



Source: Anthropic Economic Index, Feb 2025; BLS occupational employment, May 2025. Groups shown are those the index reports. In its September 2025 update, computer and maths was 36% of Claude conversations and a little less than half of API traffic.

Computer and maths jobs are 3.4% of US workers and 5.8% of the $10.8tn US wage bill. Management, healthcare, business and finance, office work and sales each pay more in wages. Yet computer and maths tasks are 36% of Claude conversations and a little less than half of its API traffic (Anthropic Economic Index, September 2025). Across all models on OpenRouter, programming went from about 11% of tokens

in early 2025 to more than half by the end of the year (OpenRouter, State of AI). If every kind of work used AI as heavily, for each dollar of wages, as computer work does now, usage would be about six times what it is today, before coding grows any further.

02

Prohibitively expensive

As a species, we've been resource constrained all of our lives. Maybe we look back at how people lived a hundred or two hundred years ago and think, how did they do it? But what we did was create more resources to make the world bigger.

Going through medical school the whole training was really built on resource constraint thinking. When you think about a patient, you always consider doing simpler things, not always because that's the best for the patient, but because you're resource constrained. You can't get ten different consultants to examine a patient, ten different diagnostic methods to collect data about them, or continuous monitoring by people who are top of their field.

Even the referral system: you go to your GP, you wait for months for a specialist, and you're on this cycle. God forbid you have a chronic condition. Then you will see how limited modern medicine actually is.

Modern medicine is quite good at making sure you don't die from simple things like microbes, because we have antibiotics, or from very traumatic accidents, because we have surgery. But most of the chronic conditions, we just manage them, quote unquote, and often very poorly.

What scarcity looks like in medicine

7.4m
waiting for planned care
NHS England, August 2026

12.2 weeks
median wait
up from 5.2 weeks in March 2010

843,419
waiting for trauma and orthopaedics
the largest list, where spine problems sit

26.7 hours
a day, for one clinician
to give all recommended care to a list of 2,500 patients (US study)

Source: NHS England referral-to-treatment statistics, August 2026; Porter et al.

Some things are still financially expensive, so we are not doing them. Very fast inference, for one. Because of this constrained world of compute, providers have to batch requests across thousands of users, so you don't get an instant response but a middle speed, whatever is best for the provider. OpenAI started offering fast and then ultra-fast serving, and Anthropic offers fast serving, but for many people it's prohibitively expensive.

When you look at how OpenAI went after the Navier–Stokes equation, they used swarms of agents. It wasn't easy. Once compute is abundant, and there is a clear story around spending that compute resulting in a very positive result, you would be happy to pay for that.

And then there's robotics. It's very early days, but you can very easily imagine billions of robots, a couple of robots per person, doing industrial work and household work, and how much compute they will require to operate.

Right now there are lots of very expensive services people buy: a very expensive surgeon, a very expensive personal service, transport, whatever makes their life better and more convenient. In the future all those tasks will be done by AI, physical or software, or AI will accelerate them, and you will just pay more money.

Dario Amodei expects that when AI makes a change to a molecule worth tens of millions of dollars, it will be paid for the value of the task, not only by the token. If it was a human that created that drug and licensed it to a pharma company, the human or the team would get the licence money. So getting the compute is so important. It's like grabbing land, and it's not just about becoming another inference company, but about the possibilities of having the compute.

PAYING FOR THE OUTCOME ALREADY EXISTS

Intercom charges $0.99 for each conversation its agent resolves (pricing); Zendesk charges $1.50 per automated resolution (Zendesk). Amodei's example is in his Dwarkesh Patel interview.

03

Human attention is not the limiter anymore

We are moving away from one API call, one Responses API request, to agents. Karpathy called this the decade of agents. We had the year of agents, but the way agents worked was still a human prompting the agent, with some going back and forth.

Then we had the goal feature from Codex, where you define the end goal and agents work toward it, maybe together with other agents. What's happening right now is that you don't want a human to slow down the agents. So you have managed agent runtimes, or super-agents, as the DeepSeek Harness team calls them, and you just run agent swarms.

From one call to swarms

One call

A single API request

a prompt in, an answer out

2025

Agents

a person prompts, the agent works, back and forth

Apr 2026

Goals and managed runtimes

Codex /goal sets the end state; Claude Managed Agents run in the cloud

Nov 2026

Super-agents

a proposed design: always on, driven by events, dispatching swarms of workers

Source: Karpathy on Dwarkesh Patel, Oct 2025; OpenAI Codex docs; Claude Managed Agents; DeepSeek Harness team, 1 Nov 2026.

We have cracked the point where human intelligence, human attention, is not the limiter anymore for agents to do work. Right now it's all about how much compute you can get. OpenAI put 10,000 agents on one proposed proof. Ten thousand today, hundreds of thousands tomorrow.

THE DEEPSEEK HARNESS TEAM'S SUPER-AGENT

Because one person can follow only three to five agent sessions, the team's proposed super-agent starts likely work in the background without being asked. In its preliminary test it finished the slowest tasks in about half the time of a single agent session, using two to four times the tokens.

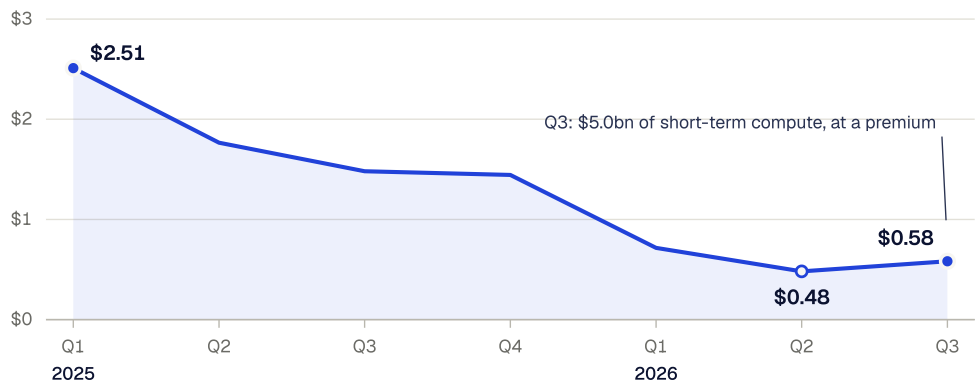
04

When the scaling is going to stop?

I often wondered where the scaling is going to stop, and where the demand is going to stop. Models were helpful but the economics around serving models were not always very clear.

Anthropic's compute spend per $1 of revenue fell from $2.51 to $0.58

All compute and infrastructure, including training, $ per $1 of revenue, quarterly



Source: Anthropic amended S-1, 2 Nov 2026. Unaudited quarterly figures.

What we are seeing is models that are able to improve themselves and the next generation of models, while making inference much more economic. What about demand?

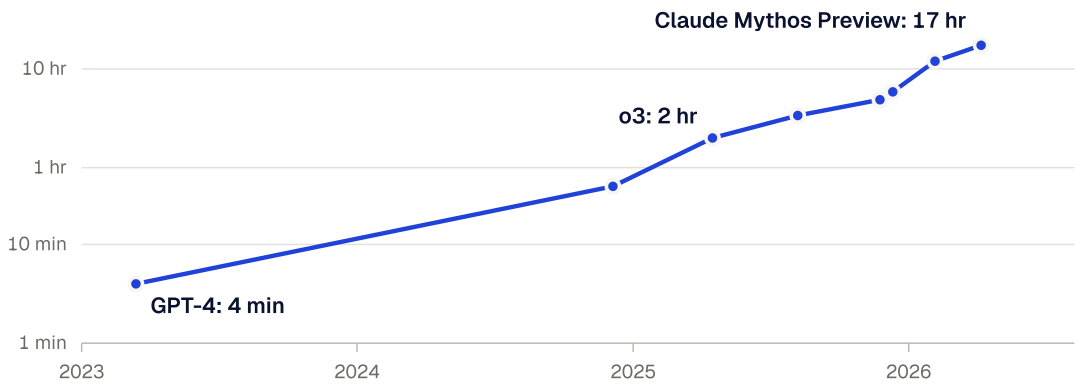
FROM ANTHROPIC'S IPO FILING

Anthropic calls it a "recursive dynamic": each version of Claude assists in building the next. Claude now runs experiments, generates and processes training data, automates evaluations and safety audits, and writes and reviews code for the models that follow. The company says its models are contributing to the pace of its own research.

It has committed about $895.4bn to compute and infrastructure, about $447.5bn of it before 2031.

The length of task AI agents can finish doubles about every four months

Human working time of software tasks models complete half the time, frontier models, log scale



Source: METR time horizons. Doubling time about 129 days since 2023 (METR's fit). METR treats estimates above 16 hours as unreliable.

Up to two months ago, I wasn't this sure how much this can scale. Then I was able to experience one Codex thread managing tens of other threads, and it became much more clear: once you have a good system in place, what models can achieve is significantly different.

05

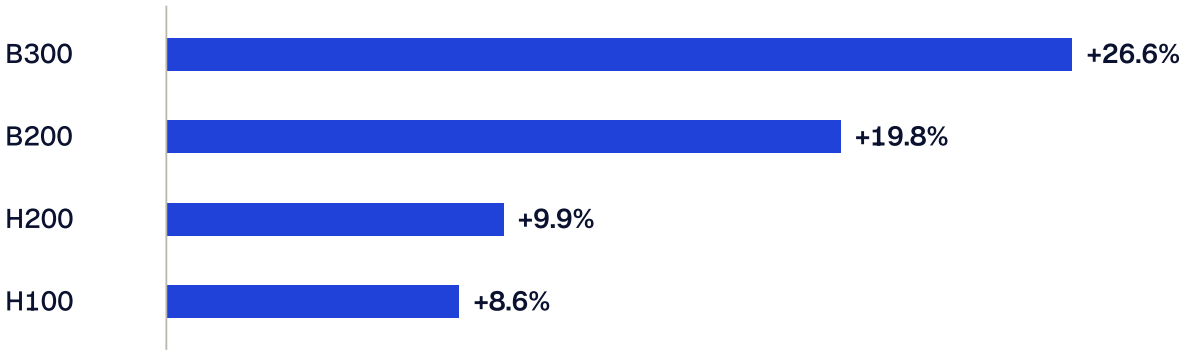
There is literally not enough compute

Try to get a cluster of nodes right now: you will see how painful that is, and how expensive that is. There is literally not enough.

Modal, one of the best inference providers, just announced it is increasing its GPU prices. In the last three to six months I've been constantly increasing my spend on Modal, but it's coming to a point where it's going to be more expensive, and more prohibitive for some.

Modal is raising prices most on its newest GPUs

Change in on-demand price per GPU-hour from 1 December 2026



Source: Modal customer notice, 2 Nov 2026; current rates.

DEMAND EXCEEDS SUPPLY, IN THE COMPANIES' OWN WORDS

> **"A compute-constrained environment where customer demand exceeds available compute."**
>
> Anthropic, amended S-1, 2 Nov 2026

> **"Customer demand continues to exceed available supply."**
>
> Amy Hood, Microsoft CFO, 28 Oct 2026

> **Demand for TSMC's chips still exceeds supply, and the gap is still large.**
>
> C.C. Wei, TSMC chief executive, 15 Oct 2026 (paraphrase)

As the big labs grow, they find that the compute required by training is also increasing rapidly. There is always a big fight over whether to use the chips to serve customers or for research. If there was 10× more compute at OpenAI, the researchers would be more than happy to take it.

ANTHROPIC, IN ITS OWN WORDS

Most of Anthropic's compute is contracted 12 to 18 months ahead. When customer demand or research needs exceed that, it buys short-term compute that costs significantly more. In the third quarter it bought $5.0bn of it, and its operating result swung from a $974m profit in the second quarter to a $567m loss, which the filing attributes mainly to that compute. It expects its use of short-term compute to keep rising. (amended S-1)

There's the aspect of getting new compute, the chips, and then the other aspect of getting those chips online, and there is a big limiter around that.

GETTING CHIPS ONLINE

GE Vernova's gas-turbine orders and reservations reached 122 GW at the end of the third quarter, and large transformers take three years or more to deliver (GE Vernova Q3 2026; Core Automation).

Compute islands

Being able to get the compute is very important, and the moment you're able to sort out the infra, the demand is there. Distributed training is possible, and everyone is trying to scrape whatever compute they can find.

TRAINING ACROSS COMPUTE ISLANDS

On 2 November Core Automation reported training a 32-billion-parameter model across two of them, 160 GPUs in the US and 160 in Asia, about 9,000 miles apart, with slightly better results than standard training (Core Automation).

It's not just GPU

And it's not just GPU. Multi-agent workflows expose the limitations we have with CPU and memory.

Something very interesting has happened over the last two months: because of how much AI I'm using, my computer has started to be insufficient, and it's a MacBook with an M4 Pro and 24GB of RAM. It ran out of memory. If the number of agent tokens I use goes up tenfold, I need a machine ten times as big.

MEMORY IS THE NEXT BOTTLENECK

Apple's chief executive, John Ternus, called memory pricing "unprecedented" on 2 November and said it will keep rising beyond December. Apple's manufacturing purchase obligations, which cover more than memory, rose by $33bn between June and September (earnings call). Apple withdrew the 512GB Mac Studio M3 Ultra in March 2026 (MacDailyNews). Servers are heading from one CPU per eight GPUs to one per two as agent workloads grow, and DRAM contract prices rose 13–18% in the third quarter (TrendForce).

It's not just NVIDIA

Apple shows how to connect multiple Mac Studios and serve state-of-the-art open-source models across them, which is what I'm going to do. I've ordered a £24,000 Mac Studio M5 Ultra with 512GB of memory, because of how much RAM I'm going to need if my agent usage keeps increasing at this speed.

Sovereign compute

Right now, most of this build is happening in the US. But everyone is realising that building the compute, and having that compute sovereign, is way more important. When a US government directive on foreign access made Anthropic switch off its Fable model for a while, it told everyone that as much as there is a build in the US, that build is going to happen anywhere and everywhere else.

For now the chips still go to the US: it took 78% of new AI chips in the first half of 2026 (Epoch AI).

06

There is no moat

We talk about moat, and where the moat lies at the moment when looking at the market. We've seen in the last couple of weeks how closed-source software is reverse engineered, and how games are getting mashed and merged with AI. Years of "human" effort turned into one prompt.

THIS YEAR

- REA, "Reverse Engineer Anything", went up on GitHub.
- PhotoCraft is rebuilding Photoshop, and Adobe's shares are down 28% this year.
- SkyCraft merged Minecraft and Skyrim.

If you have spent all the years gaining a stronghold, you might think that you have distribution. But if someone comes in with unlimited compute to create the 10× better version of your product, and creates the distribution, the go-to-market engine, with models, what's left?

I've gone through this. It was not that easy to create good software with AI. You were proud of what you've created and people would appreciate the craft. Right now, everything is generated by AI without too much of a bug, and once you have a good system in place, it recursively improves itself.

> **There is no moat. No moat for model providers, no moat for the application layer, no moat in the middle. The only reality is compute and how efficiently you're able to use that compute.**

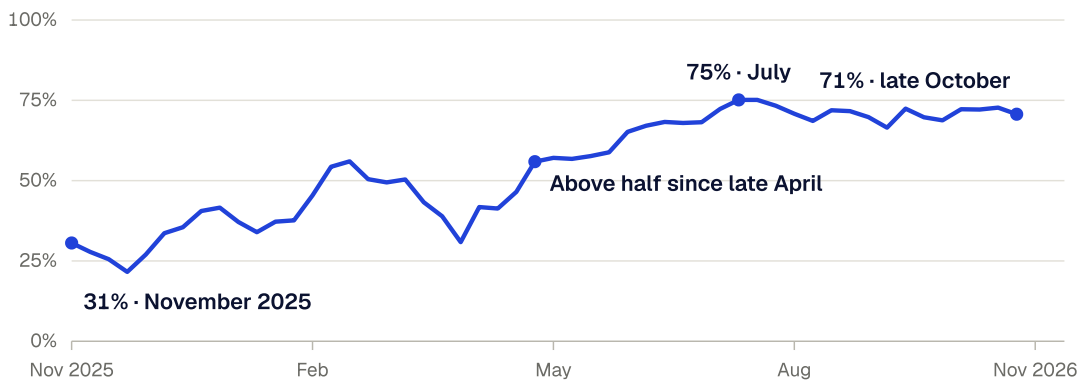
Anthropic says much the same in its own filing. As AI models are increasingly used to develop other AI models, it warns, the know-how to build frontier models may spread, including through distillation. Competition "could increasingly depend on access to large-scale compute", and compute on economical

terms could become "the primary constraint" (amended S-1).

Open-weight models already trail the frontier by about four months (Epoch AI, May 2026). On OpenRouter, closed models' share of spending fell from 89% to 73% between January 2025 and September 2026, and spending on Moonshot, Z.ai and DeepSeek models each grew at least tenfold between January and August 2026 (OpenRouter). Revolut serves about 2,000 engineers with the open GLM 5.3 model (07).

Open models now carry most of the tokens on OpenRouter

Share of weekly tokens from open-weight models, %



Source: OpenRouter rankings, open and closed weights. Weeks from 3 November 2025 to 1 November 2026, by OpenRouter's own classification. OpenRouter is not the whole market.

A STORY I ADMIRE

Applied Compute went from zero to $100m in annualised revenue run rate less than 18 months after it was founded, and has just raised $350m at a $3.25bn valuation. It sells training, inference and continual improvement on one platform to customers including Microsoft, Shopify, Harvey and DoorDash (Applied Compute, 2 Nov 2026).

07

In both cases, spending goes to compute

And for enterprises: right now, someone is running an application and you are paying money to use part of that application. Then you realise you should actually run it yourself, because you have enough users in-house, you have sensitive data you can't risk, and open-source models are good enough for you. It's like

deciding to run your application in the cloud: you use a cloud service to make life easier rather than going on-prem, and that's what's going to happen. You pay Modal, you pay Nebius, and that means the GPU pricing is never going to go down.

Running your own model still means renting compute

How a company's AI spending moves from an application to GPUs

1

You pay for part of an application

someone else runs it, and you pay to use it

2

You outgrow it

enough users in-house, sensitive data you can't risk, open models good enough

3

You run the model yourself

on a cloud service such as Modal or Nebius, rather than on-prem

4

Spending goes to compute

more demand for rented GPUs, so GPU pricing doesn't go down

Source: my own reasoning. Revolut's numbers follow.

Revolut serves about 2,000 engineers from an open model, for a tenth of its old bill

Revolut's in-house inference, as reported by its head of AI operations

56

B200 GPUs

running the open GLM 5.3 model

~2,000

engineers served

across the company

$0.30

per million blended tokens

input and output together

1/10

of its previous invoice

what it had paid Anthropic and OpenAI

Source: AI Engineer talk, 13 Oct 2026. Meanwhile rents for the newest GPUs are rising: Modal's B300 goes up 26.6% on 1 December (05).

The way people's spending is going up is not going to hold for very long. AI product companies are going to have an awakening, both in their serving and unit economics and in how lean they can stay with smaller engineering teams. Then they will have two choices: set up their own infrastructure to serve local models, which Revolut seems to be doing, or spend more. In both cases, spending goes to compute.

08

There are very different ways of being part of this

I worked at a medical device company that was creating hardware, and I currently work at an AI company that is selling agents. What I'm realising is that AI agents are going to be the future of workers. But selling agents is going to be a sales business. As we've seen with Harvey, the legal AI company, and with Revolut, they're going to utilise the compute, and everyone is going to utilise the compute. Similar to how you have

the IT team in a hospital, you are going to have the inference and AI team in the hospital, but they're not necessarily going to have their chips in the hospital. They're going to use chips somewhere, whether that's OpenAI, Anthropic or a smaller provider.

I think the engineering effort is getting less and less important in building agents and all these AI product companies. It's going to be sales. And when you look at OpenAI and Anthropic, although they are great and they are training better models, the only moat they have is compute, and compute seems to be the only thing that matters.

Ways of being part of this, from energy to owning the hardware

One figure for each, from the section linked

122 GW
Energy
GE Vernova's gas-turbine orders and reservations · 05

+87%
Chips and memory
US producer prices for computer storage, in a year · 01

$202bn
Data centres
Big Tech capex in three months · 01

$677bn
Clouds and providers
Microsoft's commercial backlog · 01

$895.4bn
Labs
Anthropic's compute commitments, $447.5bn of it before 2031 · 04

$100m
AI product companies
Applied Compute's run rate, under 18 months after founding · 06

56 B200s
The inference and AI team
Revolut's open model, for about 2,000 engineers · 07

~2×
Owning the hardware
launch price, for a used 512GB Mac Studio · 01

Source: as given in each linked section.

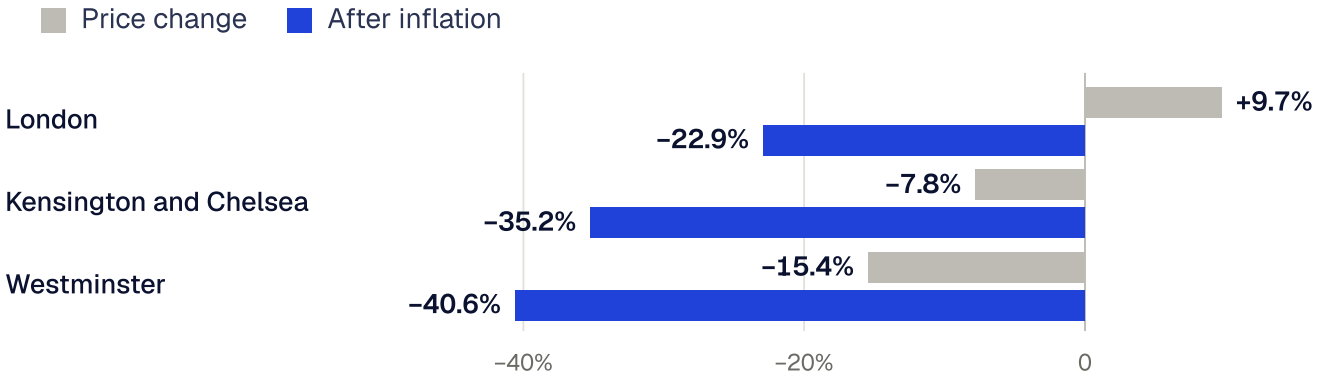
When the music stops

When the music stops. When RSI becomes possible, when the agent swarms are doing a thousand years of human tasks in seconds, where do you want to find yourself?

When you look at London house prices in the last ten years the value of the property significantly reduced once you adjust for inflation. You're putting millions into a property that goes down in value, while you probably spend lots of money to maintain it.

After inflation, Westminster property lost 41% of its value in ten years

Change in average property price, August 2016 to August 2026



Source: UK House Price Index, August 2026; ONS consumer price index. Mix-adjusted average prices for all property types; latest month provisional; excludes rent, maintenance and other costs.

Do you want to be the owner of chips, the worker of the future? Or do you want to hold on to assets that are not moving, and that are not going to mean anything when AI agents, a physical swarm of robots, are able to build a house in days?

> **At the end of the day, compute is going to be all that matters, because there is no moat.**

Disclosure. I've ordered a £24,000 Mac Studio M5 Ultra with 512GB of memory, to set up a small "data centre" at home that can run open-source models. At the start of November I applied for a few loans.

How this was written. Pangram 4.1 rates this text 73% human-written. The parts it flags as AI are mostly AI paraphrasing certain findings from the research.

Sources

Company figures are as the companies report them. Charts are drawn directly from the data linked under each one.

NOTES ON THE DATA

- OpenRouter figures (02, 03, 06) measure traffic on OpenRouter, not the whole market. My couple of billion tokens a day (02) include cache reads, which make up most agentic tokens (about 90% on OpenRouter).
- Locked demand (02): Claude's usage leans further towards coding than consumer ChatGPT use, where coding is a small share of messages. The six-times figure assumes every occupation group uses AI as heavily for each dollar of wages as computer and maths work does now (36% of conversations against 5.8% of wages).
- Inflation figures (01) are the latest on 3 November 2026. The euro-area flash (4 November), the Bank of England decision (5 November) and US October CPI (10 November) will update them. Storage-device producer prices are the closest official index for memory; there is no DRAM-specific series.
- Backlogs (01) are each company's remaining performance obligations; they are defined differently and shouldn't be added together. Cloud segment margins include work that isn't AI.
- Revolut (07): the talk did not say who owns the GPUs.
- The plan ladder (02) shows the most expensive individual plan; Google cut its top plan from $250 to $200 in May 2026.
- Adobe's share price (06) is the change from the end of 2025 to 2 November 2026 (Yahoo Finance). Most of the fall came before PhotoCraft appeared at the end of September.
- The open-model share (06) uses OpenRouter's own open and closed classification. Its weekly totals differ slightly from the main ranking series, so it is used only as a share.
- The DeepSeek Harness test (03) is preliminary and indicative, as its authors say. Anthropic's quarterly figures (04, 05) are unaudited.

COMPANIES AND FILINGS

- Anthropic, amended S-1 registration statement, 2 November 2026
- Applied Compute, $100 million run rate and fundraise, 2 November 2026
- Micron, fiscal Q4 2026 prepared remarks, 30 September 2026
- SK hynix, Q3 2026 results, 28 October 2026
- Apple, fiscal Q4 2026 earnings call transcript, 2 November 2026
- Microsoft, Q1 FY2027 earnings, 28 October 2026
- Amazon, 10-Q for the quarter to 30 September 2026

DATA AND RESEARCH

- OpenRouter rankings, including open and closed weights, and The Token State of Play
- Epoch AI, the open–closed capability gap
- Anthropic Economic Index, February 2025 and September 2025; OpenRouter, State of AI, December 2025; BLS occupational employment and wages, May 2025

- Inflation: BLS CPI, ONS CPI, Eurostat HICP, TurkStat CPI, IMF World Economic Outlook, October 2026; producer prices via FRED (storage devices, semiconductors)
- METR, time horizons
- Epoch AI, data centres and AI chip imports
- Artificial Analysis, Sonnet 5 agentic cost
- GitHub Octoverse 2026
- UK House Price Index, August 2026
- NHS England referral-to-treatment statistics, August 2026

AGENTS

- DeepSeek Harness team, "Exploring super-agent design", 1 November 2026
- Anthropic, how we built our multi-agent research system
- OpenAI, Navier–Stokes
- Andrej Karpathy on the Dwarkesh Podcast, October 2025

INFRASTRUCTURE

- Core Automation, "The hitchhiker's guide to training on compute islands", 2 November 2026
- GE Vernova, Q3 2026 results

The phrase "when the music stops" echoes the film Margin Call (2011).