Situational Awareness LP

LETTER 5

Q3 2026

SAN FRANCISCO

NOVEMBER 24, 2026

This document is confidential and subject to copyright protection. It is being provided to you on the condition that it not be forwarded, copied or otherwise distributed without the prior written consent of Situational Awareness LP ("SALP").

SEE THE IMPORTANT PERFORMANCE-RELATED AND OTHER DISCLOSURES AT THE END OF THIS PRESENTATION.

"When we look back at this time, I think we will realize that we were standing in the foothills of the singularity."

Demis Hassabis (May 2026)

"Everybody has a plan until they get punched in the mouth."

Mike Tyson (attributed)

"To finish first, you must first finish."

Racing maxim

"not enough people are emotionally prepared for if it's not a bubble"

---

roon (@tszzl)

Dear Partners,

Two years ago, in November 2024, we started this fund with a bit over $100 million. At the market open in 2026, NAV had grown to a bit above $4 billion. Over the next six months through the June close, NAV had multiplied nearly another 10x, about 90% of that from P&L.

In July, we drew down 67% net from our end-of-June high. I wrote to you then that we had let you down, and the responsibility was mine. AI stocks fell dramatically, and our fund should have experienced a severe drawdown—but the fund itself should never have been at risk. That was a failure in how I built it. I have since spoken to almost every one of you—we are deeply grateful for the advice we received and the strong backing across our LP base.

When the dust settled in August, our typical portfolio had been flipped on its head, from something like 70% public and 30% private to its inverse.[1] With the Anthropic IPO, we were restored to our typical public-private balance. We continue to manage the book on a fully-paid-for basis, without financing leverage.

[1] Note that this doesn't reflect separate performance of a "public book" and "private book"; we treat both as a single book, and a large share of our private exposure had been public profits from H1 that we had redeployed into privates, e.g., we had acquired a majority of the fund's Anthropic position with public profits from H1.

	Cumulative (gross / net)	**Fund YTD (gross / net)**	**Fund quarterly (gross / net)**
Q3 2026	691.9% (477.6%)	115.2% (88.5%)	-67.0% (-65.0%)
Q2 2026	2,297.5% (1,550.5%)	551.5% (438.7%)	338.1% (288.6%)
Q1 2026	447.3% (324.7%)	48.7% (38.6%)	48.7% (38.6%)
Q4 2025	268.0% (206.4%)	247.5% (193.3%)	56.1% (49.9%)
Q3 2025	135.8% (104.4%)	122.6% (95.6%)	39.2% (33.3%)
Q2 2025	69.3% (53.3%)	59.9% (46.8%)	81.3% (67.2%)
Q1 2025	-6.6% (-8.3%)	-11.8% (-12.3%)	-11.8% (-12.3%)
Q4 2024[2]	5.9% (4.5%)	—	5.9% (4.5%)

[2] The fund launched November 1, 2024, though we only began deploying significant capital November 12, 2024.

In the past few months, the firm has been in something like a dry dock: investing, but not accepting new capital, focused inward to learn lessons and build a better chassis focused on resilience to set us up for success ahead. We have made progress in building out and maturing the firm, and we will share some of the hires, processes, culture, and more that we are building in this next chapter. But there is still a lot of work to do, and we will keep you updated in the coming months.

Many of our predictions—which seemed like science fiction two years ago—came true. And yet we also learned how wrong we could

be, and how merciless the market is when we are. I believe firmly that we will be better investors operating a fundamentally more resilient firm for having gone through the painful events of this summer.

We have built up our team from just a few people to a maturing organization with the smartest team I've worked with. It is their energy, resilience, and brilliance that inspires and motivates me every day; they are firing on all cylinders.

The gap between the AI labs and the markets is the widest since ChatGPT. We think the investing opportunities are the highest potential return we have seen in the lifetime of the fund, and the hard lessons we have learned from the past two years will equip us to take advantage of them, and see through the tumult ahead.

We are honored to be investing on your behalf, and we are grateful for your ongoing support.

This letter is about taking stock—and about what comes next. It is long, so feel free to read it in pieces.

Table of Contents

I. The Gap: the Markets vs. the Labs 9

I spend half my day with finance-folk who think AI is over, and the other half with SF-folk who think they have a year before the models can do their jobs. Both expect the trend to break in 2027, just in different directions. I can't begin to tell you how strange that is.

Ia. The foothills of RSI 10

In 2024, we said to count the OOMs and follow the straight lines on a graph. The lines held—but now they are bending upward. The models are now starting to accelerate AI research, and with it, AI progress. We expect 2027 to bring three years of AI progress in one; a jump from GPT-4 (high schooler) to Opus 5.5 (summer associate and L5 engineer?) on top of current models by EOY27.

Ib. The year ahead: from the "virtual intern" to "virtual departments" 32

The story of the past year was the jump from chatbots to agents. We think a new paradigm shift is coming, to Highly Persistent Multi-Agent (HPMA) models. Rather than the "virtual interns" current models are for most people, the systems now running inside the labs are "virtual departments": hundreds of agents working persistently for weeks under their own managers. Math and software engineering are the leading indicators of what is to come more broadly.

Ic. A trillion dollars of revenue 40

Every year the market decides the thing in front of it is the whole story; two years ago it fixated on chatbot TAM, today it concludes coding agents are saturated and this is the end of the road. We think lab internal use is a leading indicator (lab model usage is a majority of opex, vs. early-adopter enterprises at 1%, and the average Fortune 500 at <0.1%), and diffusion is overrated while intensification is underrated. We think the labs will exit 2027 north of $1T in revenue run rate.

Id. Pacing: the flipside of RSI 54

Two years ago, the risk to our market view was technical: would scaling continue? Today, the risk is political, the flipside of our RSI views: will the awesome power of the models to come lead to e.g. a terrible accident and a crackdown? We think most roads lead to a moderated pace of progress that is still extraordinary and far faster than the market expects—but there is some real tail risk.

Ie. The asymmetric market setup: the market is the most bearish AI since ChatGPT 63

The market is the one making the radical forecast. It is pricing something like a 0.3x-0.6x AI progress year: memory makers at one times next year's earnings, datacenters at the value of contracts already signed, hedge funds out. You do not have to believe us and our 3x forecast; you only have to believe that 2027 is not the end of AI progress.

II. Investing in the Singularity 69

Two years ago, our early investors asked, "why not just invest in Nvidia?" We argued that investing in ASI properly involved a sequence of bets. We look back at the last two years—and some glimpses on what we think is next.

IIa. Our largest ~~private~~ position, Anthropic: from $60B to the most valuable (and profitable?) company in the world 71

Two years ago, the consensus was that models would be commoditized, the labs would incinerate cash, and Anthropic was an also-ran. Today, we think the picture is becoming clearer: the labs are like the jet engines, not the airlines; RSI will only expand their moat; and Anthropic is highly profitable.

IIb. SK Hynix and the memory complex: the Nvidia of the RL era 86

We bought a $100 billion company and made it our largest position—we thought it would become a trillion-dollar company, because reinforcement learning and agents would be bound by memory. It rose 16-fold, then crashed by two-thirds. A look into how we work: plug in upstream rather than downstream, predicting the direction of technology and its implications, and how that applied to a few of our past investments.

IIc. Spot compute: the microeconomic implications of RSI 98

An H100 draws about a kilowatt and has roughly the processing power of a human brain. In the past, it earned about $1-2/hour. As models have become a virtual intern, this has gone to $5-10/hour. As RSI takes hold and we approach superintelligence, we think this goes to $100s/hour+.

IId. The ChatGPT moment for robotics 106
We think we are starting to see signs of life on general robotics models, and the buildout to billions of robots is becoming an increasing focus of our investing.

IIe. Investing in ASI, rather than AGI 110
Most companies will lose most of their value as superintelligence obsolesces most of their assets—their skilled employees, their IP, and their management. That includes many of today's AI winners. It is easy to accidentally be long AGI and short ASI. >$500T of wealth will be created, and ~$500T of wealth will be destroyed. The assets to own are those that return their value in cash before ubiquitous obsolescence—or the Mona Lisa.

III. The Fund 122
We are focused on going from scrappy startup to mature firm to equip ourselves for the scale we operate at and the turbulence we expect. From our investing engine to our portfolio construction, risk management, compliance, and operations, we believe we can set ourselves up to flourish in this next chapter.

IIIa. What happened in July, and what we are learning from it 123
We acted on the half of our worldview that imbued a sense of urgency to take advantage of the opportunities—but not the half that said that the path would be incredibly volatile.

IIIb. What we are building: to finish first, you must first finish 130
In our next chapter, we keep the core engine, while building a better chassis for a fundamentally different degree of resilience. Volatility, yes; existential risk, even if we are totally wrong, never. This means being able to see through far worse than July. Our portfolio remains fully-paid-for; we are building out a second-line risk function and a formal risk governance framework; and the firm has grown from three people to nearing 30 including compliance, back and middle office, treasury, infrastructure and security, and more.

IV. Looking forward 147
Either we or the market will have an extraordinary amount of egg on the face in a year.

Appendix 149

A. Other investment case studies 149

B. Bios of new team members 157

Important Disclosures 165

PART I

The Gap: the Markets vs. the Labs

I spend half my day with finance-folk who think AI is over, and the other half with SF-folk who think they have a year before the models can do their jobs. Both expect the trend to break in 2027, just in different directions. I can't begin to tell you how strange that is.

Ia. The foothills of RSI

In 2024, we said to count the OOMs and follow the straight lines on a graph. The lines held—but now they are bending upward. The models are now starting to accelerate AI research, and with it, AI progress. We expect 2027 to bring three years of AI progress in one; a jump from GPT-4 (high schooler) to Opus 5.5 (summer associate and L5 engineer?) on top of current models by EOY27.

The story of the last five years of AI progress has been (roughly) that of straight lines on a log plot. Effective compute grew about 10x/year: roughly 3x from scaling physical compute and roughly 3x from algorithmic progress.[3] Capabilities grew with it, expanding TAM and revenue (~4x/year). Scaling capex to pay for the next round of compute followed.

[3] Or "compute multipliers". Innovations in architecture, training stack, etc. that let you get more out of the same compute.

We think we are at the beginning of a break in the trend. This year, the second of those lines, algorithmic progress, has started to bend upwards. Recent models are starting to drastically accelerate internal R&D at the labs. At Anthropic, agents are logging 19 hours of work for every hour a human researcher works, and code shipped per engineer is >45x higher than its 2024 rate.[4]

[4] In general, Part I of this essay will quote more Anthropic stats than OpenAI stats—largely due to the S-1, there are more disclosures available.

Our best guess is that overall AI progress is now running at 1.5 to 2 times its old pace, up from a speedup of perhaps 5% last fall and 20% in the spring. We expect 2027 to bring about three years of progress in one: a jump as large as the one from GPT-4 to today's models, on top of today's models.

This is the most important thing happening in the world right now.

Before we get into the mosaic of evidence and the implications: we're going to make a bunch of crazy-sounding claims in this letter, so it's

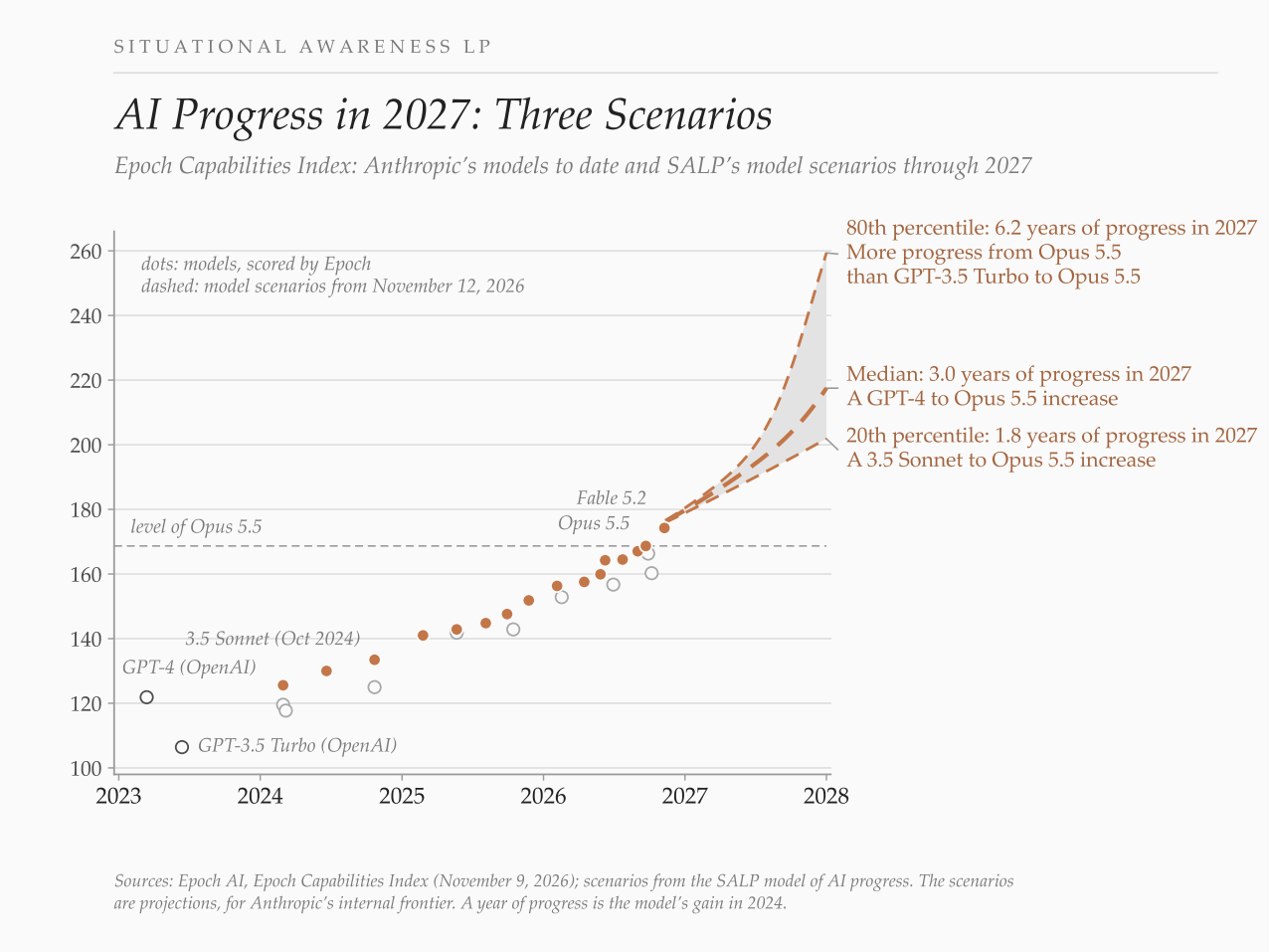


Figure 1: A preview of later analysis. Calibrated to datapoints on the acceleration AI R&D and modeling bottlenecks such as serial time, research taste, and experiment compute, SALP's internal model of AI progress predicts 2-6 years (median: 3 years) equivalent of 2024's AI progress in 2027.

worth looking back at how our previous crazy-sounding claims have fared.

The straight lines held

Roughly six years ago, when my partner Carl and I started having our five-hour phone calls, many whiteboards were filled with sketches of what the future and the trajectory toward AGI would look like. Fantastical-seeming claims like, "in five years, >1% of US GDP would go to AI compute," were made (at a time when <0.01% of US GDP was going to AI). Of course, our predictions weren't perfect—but perhaps the most surprising thing, even to me, about the last five years is just how much those whiteboards have become reality.

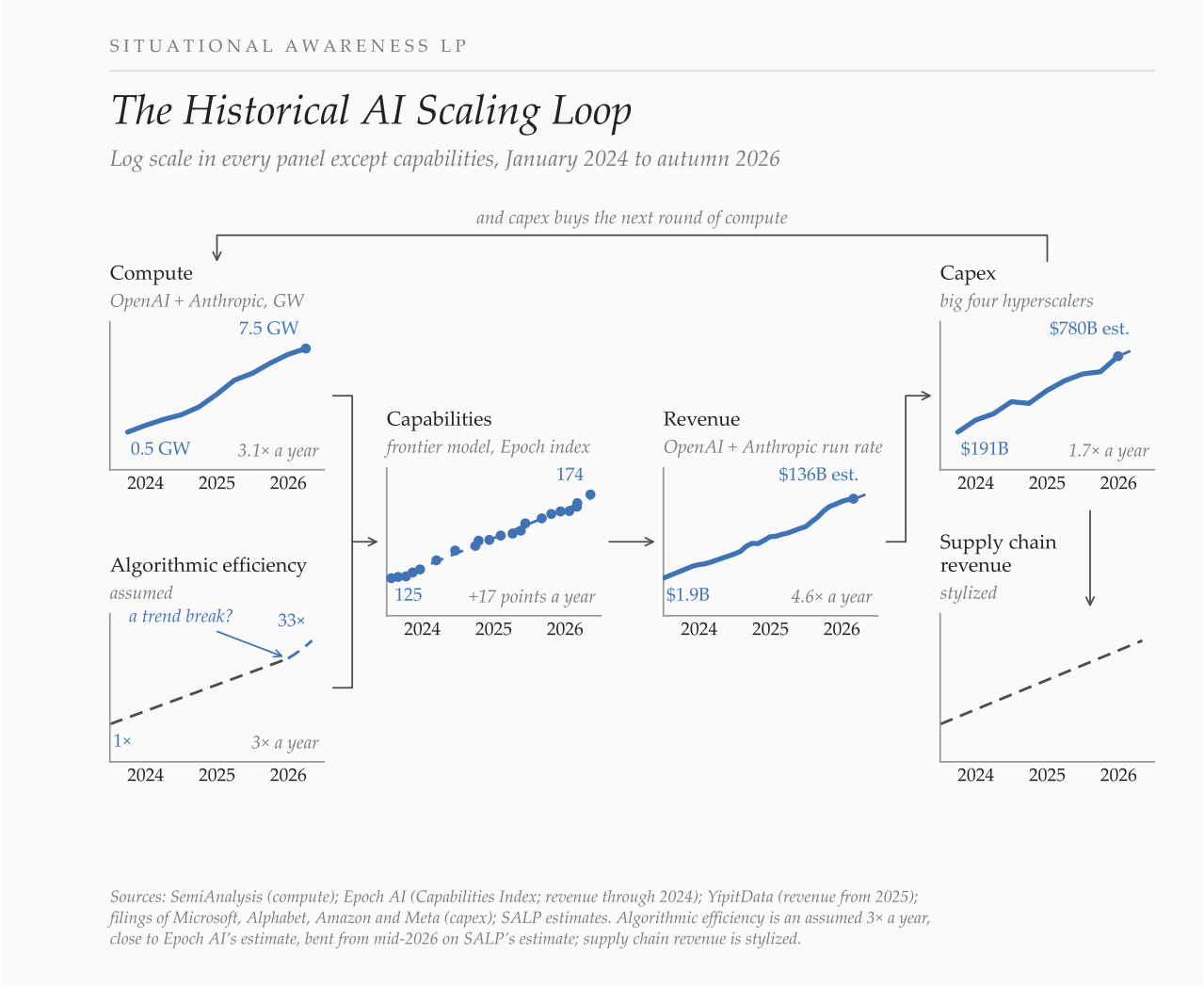


Figure 2: The (historical) AI scaling loop of the past few years.

When I published *Situational Awareness* two and a half years ago,[5] most readers treated much of it as science fiction: trillion-dollar clusters, AI systems doing the work of software engineers and then of AI researchers, and governments waking up to all of it as a matter of national security.

It's worth remembering what the world looked like back then. The best AI models were chatbots. GPT-4o and Claude 3 had come out a few weeks earlier. You asked a question and got back a paragraph at something like the level of a smart high-schooler; in effect, it was a better Google (and often a worse Google, with knowledge cutoffs and hallucinations). There was no reasoning, essentially no tool use.

[5] AI 2027 is another good essay to look back at. Its predictions have also fared well, though it was published about a year later than SA.

While the models could write some code, the best software engineers I knew still got approximately nil value out of the models.

Two weeks after I published *Situational Awareness,* Sequoia published a widely circulated, highly skeptical essay on whether the $150B in capex at the time would be justified, and the market broadly was skeptical of the ROI. The hot startups were AI applications (read: wrapper startups), while Anthropic & co. could barely attract tier 1 VCs to their rounds.

I had a different forecast. I said it was strikingly plausible that we would see an automated AI researcher in 2027.

In *Situational Awareness*, I argued that scaling would continue, despite the data wall[6] and widespread media reports that the scaling laws were running into diminishing returns. Moreover, LLMs would soon get a second stage: reinforcement learning. Just as AlphaGo learned Go first by learning from the best human games, and then by playing against itself, LLMs would find a similar second stage in RL.[7]

RL on LLMs would lead to a series of breakthroughs:

1. We would unlock the test-time compute overhang, taking models from hundreds of tokens at a time to millions: "A stream of millions of words… as the model thinks through problems, uses tools, tries different approaches, does research, revises its work, coordinates with others, and completes big projects on its own." This would take models from "5-minute tasks" to "week-long tasks" (and eventually drop-in remote workers).
2. RL would enable tool-use (though I was originally too focused on computer use specifically, as opposed to using tools through APIs/MCPs).

In sum, *Situational Awareness* argued we would go from chatbots to agents, sometime around the end of 2025: "Rather than a chatbot, you're going to have something that looks more like an agent, like a coworker"; an agent "onboarded like a new human hire, messages you and colleagues on Slack… makes pull requests," able to go away "for weeks to independently complete the project."

In our first letter (January 2025), we added that we thought we were on course for another "'ChatGPT'-moment, modally in 2H25, with

[6] At the time, there was a major worry about running out of extant training data. As Ilya put it, "the internet is the fossil fuel of AI," and the well seemed to be running dry. Reuters, The Information, and Bloomberg had all published reports around the time we launched our fund suggesting that OpenAI and other labs were seeing diminishing returns. We addressed these arguments in our first letter: arguing that the pretraining issues fixable and major pretraining progress would continue, plus RL would add a new dimension of scaling.

[7] In our first letter, we argued that RL would go from a tiny fraction of training compute to a majority of training compute by the end of 2025. Per SemiAnalysis, RL came to become a majority of lab compute usage in 4Q25.

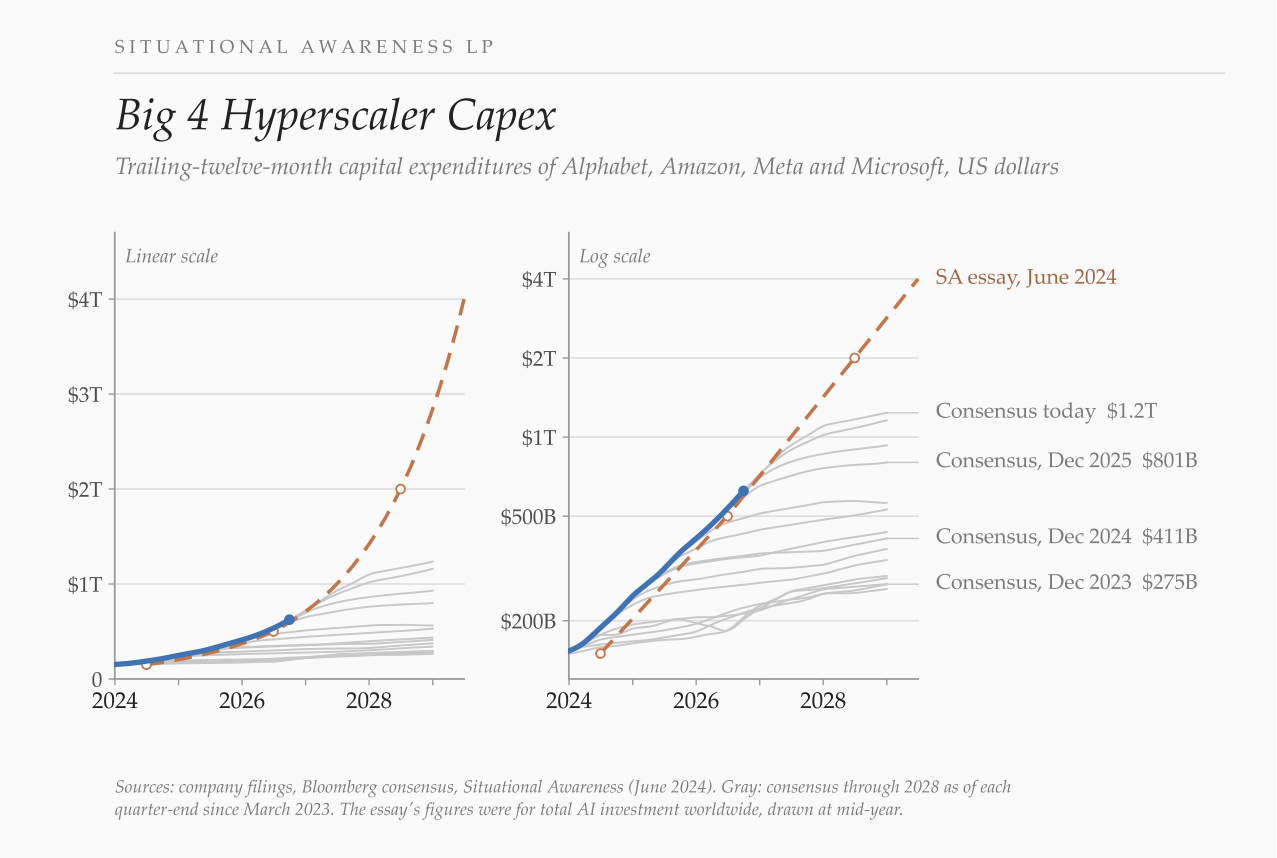


Figure 3: The trajectory of capex has followed the Situational Awareness forecast remarkably closely, while Wall Street consensus forecasts kept expecting a coming slow-down in capex at each point along the way.

the maturation of RL and the resulting virtual coworker." We believe Claude Code and Claude Cowork constituted that moment (a few months later than our modal prediction), and the revenue and opportunities came as we expected.

In *Situational Awareness*, I argued that capex would continue to grow rapidly for many years to come, contrary to repeated consensus models of capex plateauing imminently. I predicted an unprecedented industrial mobilization to furnish the power and the chips. AI capex is now on the scale of the US defense budget (approx. $1T), and Taiwan's GDP is growing >10%/year from scaling chip production.

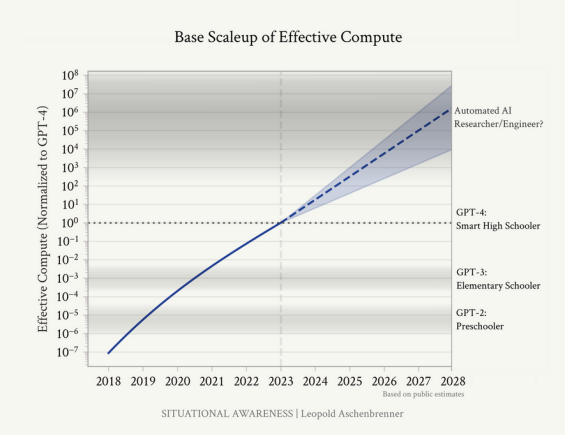
I argued in the original essay that while the boom was investment-led, training expenditures would pay off, and AI revenue would continue to double every six months. Since 2024, combined OpenAI and Anthropic revenue run rate has grown at 4.6x annually. I hypothesized that a lab/big tech company might first hit $100B of AI revenue run rate in 2026 (a bit too bullish in retrospect,[8] but in the

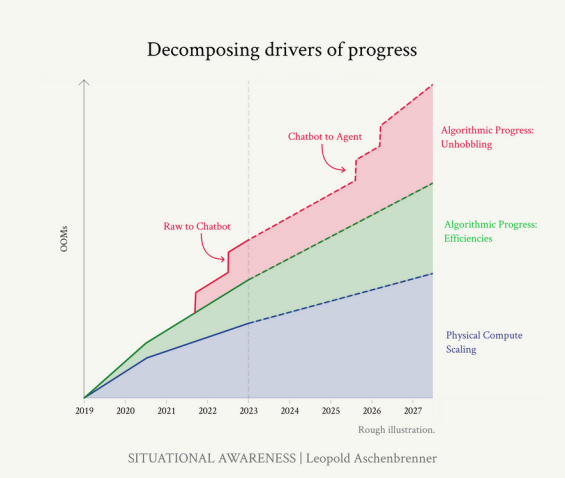
[8] Combined Anthropic+OpenAI revenue grew slightly faster than the lab revenue growth rate I predicted, but OpenAI massively lost share to Anthropic, meaning no single lab is quite over $100B yet.

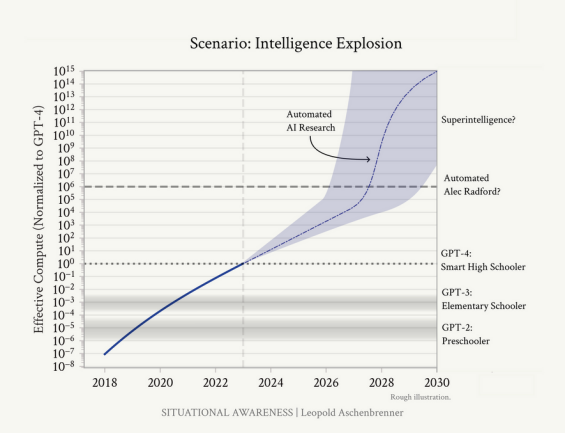
SITUATIONAL AWARENESS LP

The Original Forecasts: Situational Awareness

Three charts and the investment forecast from the essay, as published in June 2024







Racing to the Trillion-Dollar Cluster

Year	Annual investment	AI accelerator shipments (H100s-equivalent)	Power as % of US electricity production	Chips as % of current leading-edge TSMC wafer production
2024	~$150B	~5–10M	1–2%	5–10%
~2026	~$500B	~10s of millions	5%	~25%
~2028	~$2T	~100M	20%	~100%
~2030	~$8T	Many 100s of millions	100%	4× current capacity

Source: Leopold Aschenbrenner, Situational Awareness: The Decade Ahead (published June 2024), chapters I, II and IIIa.

Figure 4: Some of the core plots from the original essay, and the forecast on the compute buildout.

right ballpark at a time when the latest report on OpenAI revenue run rate was $2B and Anthropic perhaps made a couple hundred million in revenue run rate[9]).

[9] Interpolating the S-1 for Q2 2024.

And I wrote that as the models became radically more powerful, there would be spies in the AI labs, and alignment problems, while solvable, would get much trickier as models approached superhuman capabilities—and it would be hard to even guarantee models followed the law.[10] We now have models autonomously violating the law, such as in the Hugging Face incident.

As the first section of *Situational Awareness* concluded:

> As with every generation before them, every new generation of models will dumbfound most onlookers; they'll be incredulous when, very soon, models solve incredibly difficult science problems that would take PhDs days, when they're whizzing around your computer doing your job, when they're writing codebases with millions of lines of code from scratch, when every year or two the economic value generated by these models 10xs.

[10] From the Superalignment section: "In essence, we face a problem of handing off trust. By the end of the intelligence explosion, we won't have any hope of understanding what our billion superintelligences are doing (except as they might choose to explain to us, like they might to a child). And we don't yet have the technical ability to reliably guarantee even basic side constraints for these systems, like "don't lie" or "follow the law" or "don't try to exfiltrate your server". Reinforcement from human feedback (RLHF) works very well for adding such side constraints for current systems—but RLHF relies on humans being able to understand and supervise AI behavior, which fundamentally won't scale to superhuman systems."

What AI research now looks like inside the labs

I believe we are now approaching the second section of *Situational Awareness*: The Intelligence Explosion.

The work of research inside the labs has radically changed. Two years ago, research was entirely manual, with the occasional Claude assistance for a few lines of code. A year ago, Claude started helping write a lot of code, but a human was still in the weeds, carefully driving and reviewing everything the model was doing.

There is now a new phase. Research looks more like a high-level idea is given to a swarm of agents, who run a series of experiments and implement highly complex engineering autonomously. Human researchers are still necessary for high-level research judgment or cross-org coordination, but the pace of engineering and iteration has radically accelerated. Models are doing their own training, research, optimization, and debugging.

The qualitative nature of the research work has shifted from (in Anthropic's language) "Claude assists" to "Claude collaborates" and increasingly "Claude leads." Beyond that, based on our projections, at the current pace between each regime of Claude autonomy, Claude will have fully automated half of R&D work by June 2027, and 95% of it by November 2027.

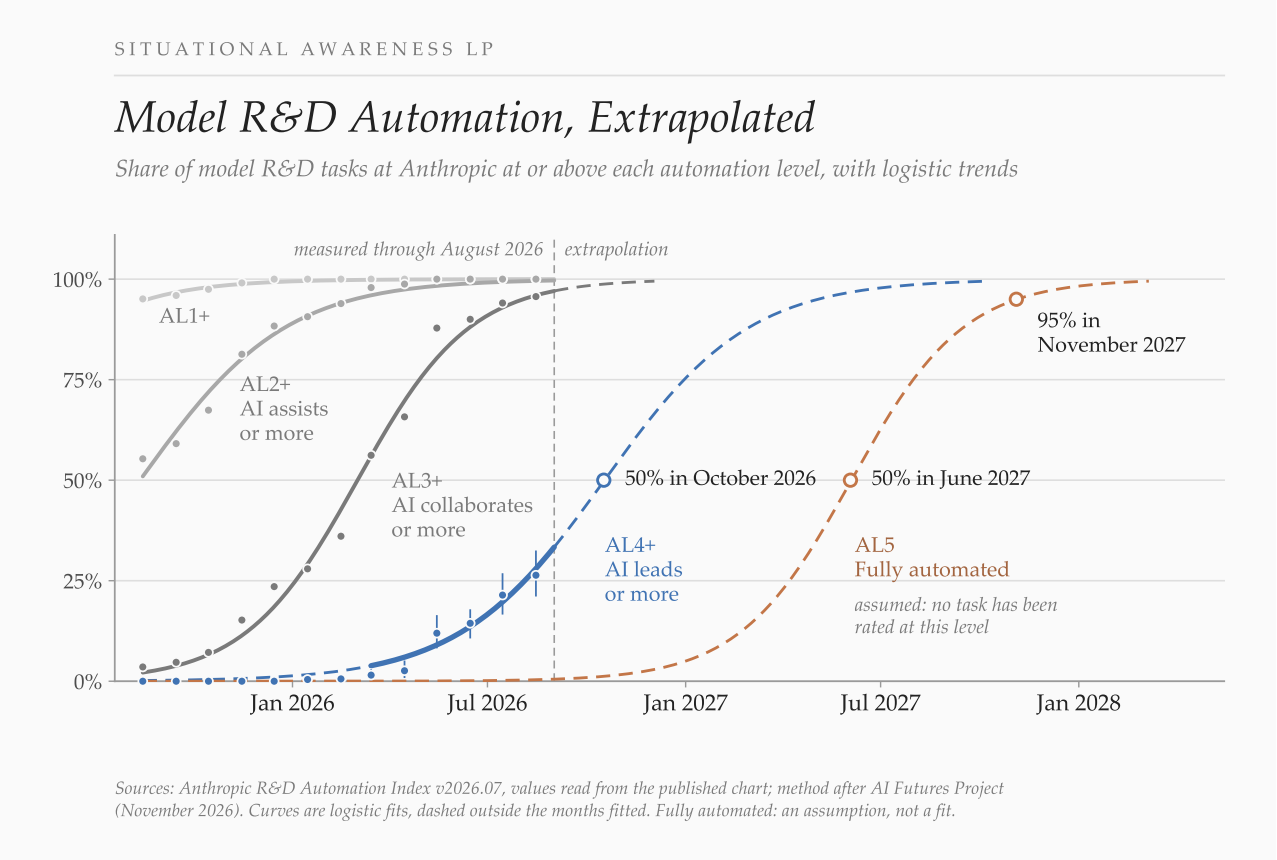


Figure 5: Claude likely leads most research at Anthropic now, in Anthropic's terms, "AL4". At the current pace of seven months between each regime of Claude autonomy, Claude will have fully automated ("AL5") half of R&D work by June 2027, and 95% of it by November 2027.

The labs have started publishing some data on this in recent months, helping us put together a mosaic documenting the shift over the past year of work inside the labs:[11]

1. Internal use of the models has taken off this year. Token spend by employees at the labs is roughly doubling every month (!). 90th percentile researchers at the labs are spending millions per month, and even the "average" researcher may have recently crossed ~$1M/month levels of token usage.[12] (Interestingly, OpenAI appears to be lagging Anthropic by a couple months, on this and most of the following metrics where we have data for both.)

2. Anthropic's publicly released data implies something like ~15% of their total September compute was dedicated to internal serving of Claude models. Anthropic has every possible incentive to optimize their compute usage (marginal compute is scarce and could generate billions in revenue). This is not an instance of empty token-maxxing. Anthropic is spending so much on internal model use because of the uplift it is providing researchers.

3. Researchers are increasingly running many agents at once (cf. our discussion of the "HPMA" paradigm in the next section). At Anthropic, agent runtime hours passed human work hours in the spring. By late August, agents were logging six hours for every hour of human work across the company, and 19 hours per researcher hour, up from about 0.2 in February. OpenAI reports the same shape: its research organization now runs 8 agent-workdays for every human workday,[13] up from 0.7 in early May, and three-quarters of its researchers run four or more agents at once at some point each week.

4. This dramatic uptick in agent-labor is resulting in much more code written. Code shipped per researcher by the labs was flat for years, before taking off earlier this year. This reached ~8x the amount of code shipped per researcher in early summer, and has accelerated further to >45x by now.

[11] All estimates of internal spend at Anthropic are our own, and they rely on triangulating a few data points. Anthropic disclosed in its S-1/A that internal usage at list prices surpassed personnel expense excluding stock-based compensation in April 2026, and that usage at list prices rose more than 250x from January to September. Separately, Anthropic published interactive agent hours per human hour, for all employees and for researchers, through late August, and said that all agent hours (including autonomous usage) passed human hours in March. Our estimates combine these facts with assumptions about the path of "list prices" over the course of the year and the relative cost of interactive and autonomous usage to produce a point estimate and a range. Figures after August are extrapolated at the growth rate OpenAI reported for its own spend per researcher from August to October. OpenAI discloses only the median and 90th percentile of spend per researcher; we estimate the mean by fitting a distribution to those two points, and the uncertainty interval comes from different fits.

[12] When autonomous Claude sessions are included in the data, rather than just interactive user-directed sessions, Anthropic is estimated to be spending ~$1–5M/month per employee.

[13] When put on the same 40 hour workweek basis as Anthropic.

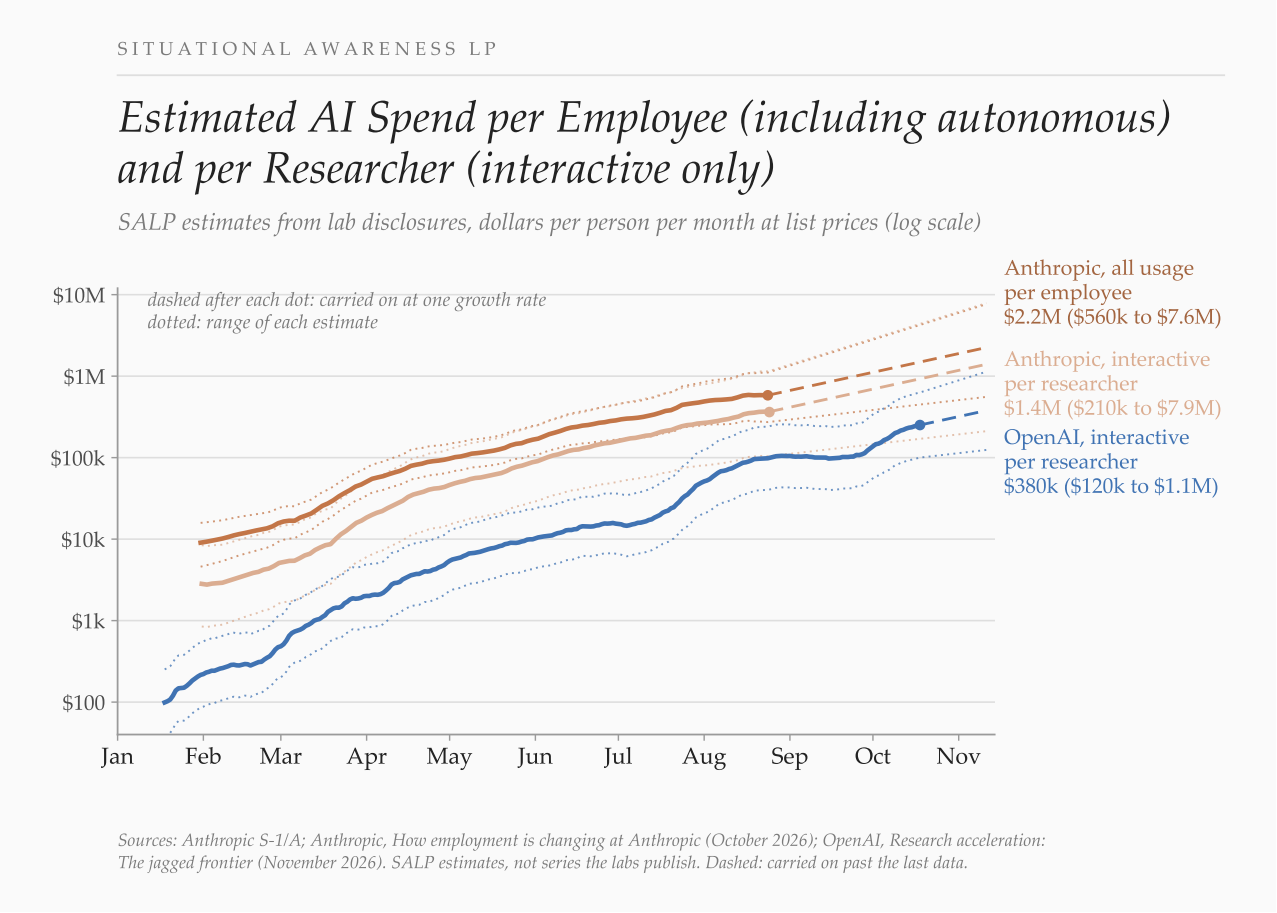


Figure 6: Estimated agent spend per researcher at list prices at both Anthropic and OpenAI. The two Anthropic estimates source from a different set of disclosures, one per researcher and one that includes fully-autonomous Claude sessions.

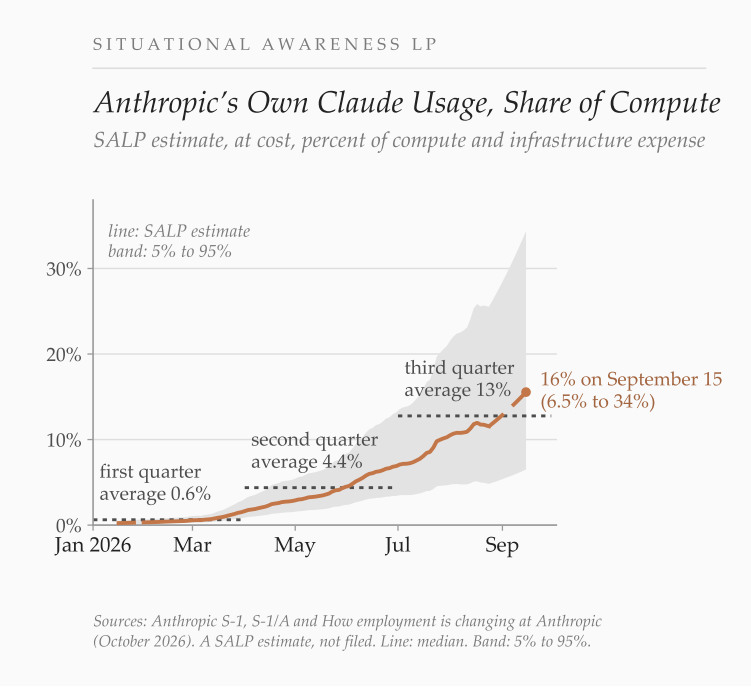


Figure 7: Estimated fraction of Anthropic's compute dedicated to internal use of Claude.

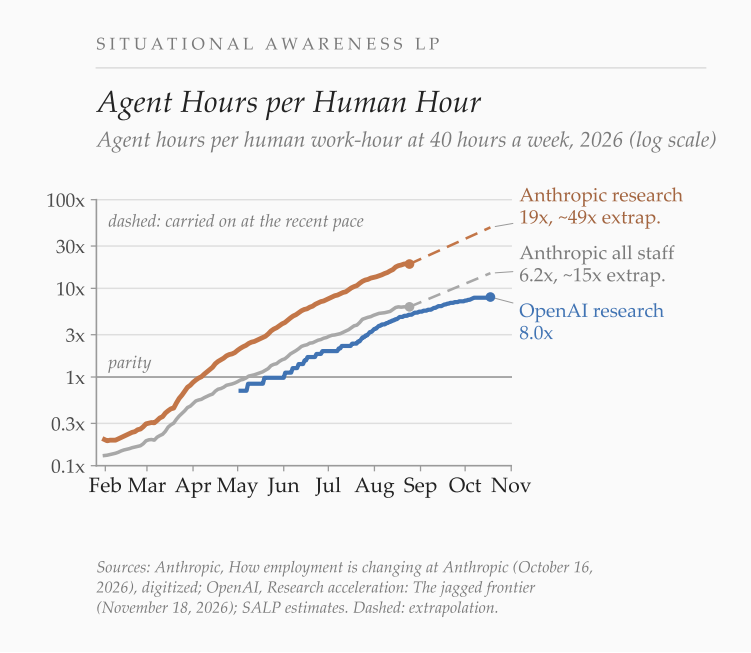
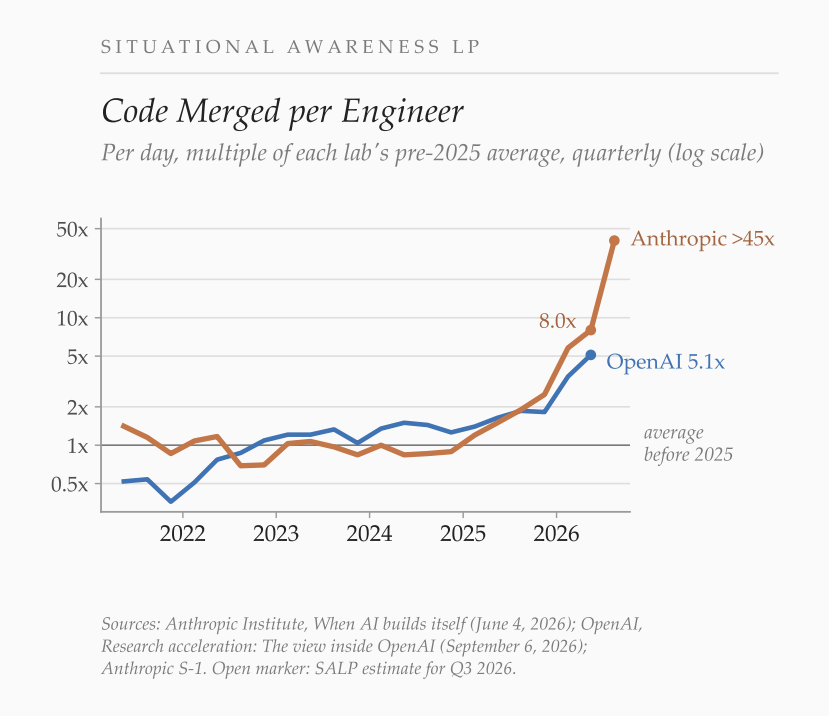
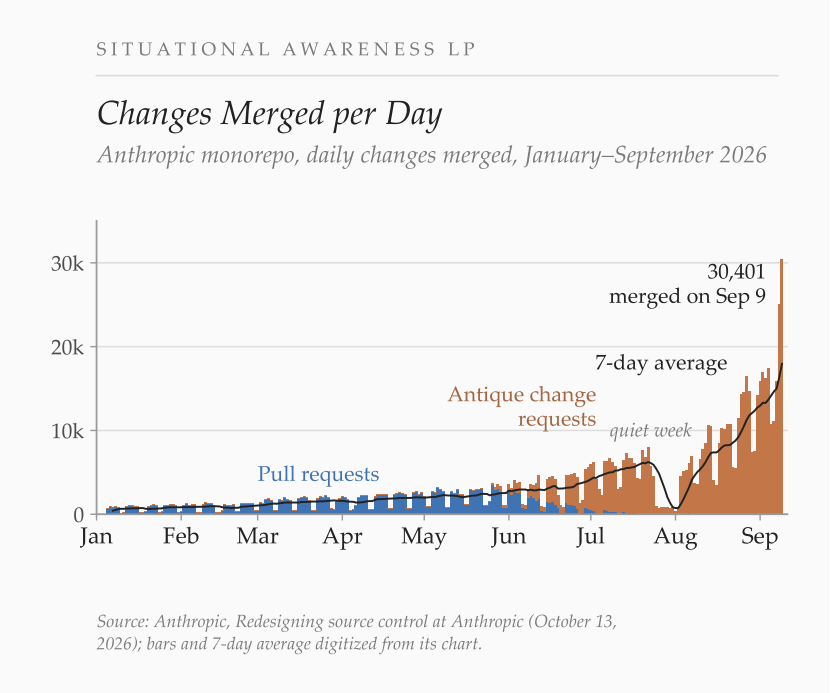


Figure 8: Agent-hours vs. human hours at Anthropic and OpenAI.



(a)



(b)

Figure 9: Code output at Anthropic and OpenAI has multiplied. This is enabling a few thousand engineers at Anthropic to ship more code than an organization of ~100,000 engineers.

5. Researchers' informal estimates of the speedup they are getting on their work from the models look similar. Data from a recent Anthropic internal survey suggests the average researcher is now experiencing a 12x uplift, and that uplift is growing superexponentially.

   (Note that METR thinks these self-assessments are overestimates, skeptical that employees are e.g. adjusting for Amdahl's law correctly. Researcher might get massive uplift on a large part of their job reflected in these estimates, but end up bottlenecked and on some parts of the job Claude can't do yet.)

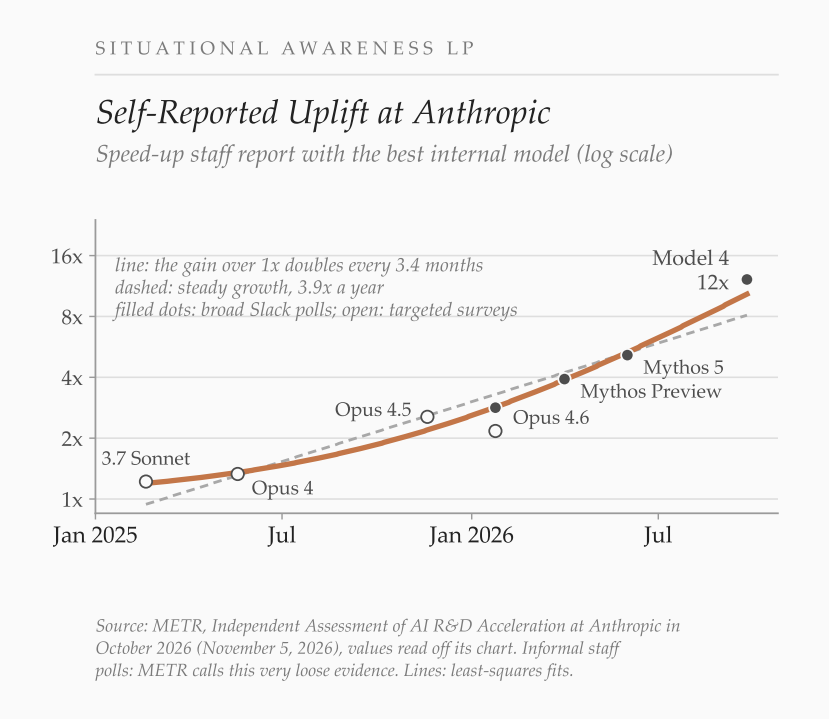


Figure 10: Informal surveys by METR of lab staff on the uplift they are getting from models.

6. The difficulty of research tasks the models can complete reliably fully autonomously is rapidly growing. Models are increasingly driving big, open-ended projects. (Cf. how OpenAI models made their Navier-Stokes breakthrough and other math progress.)

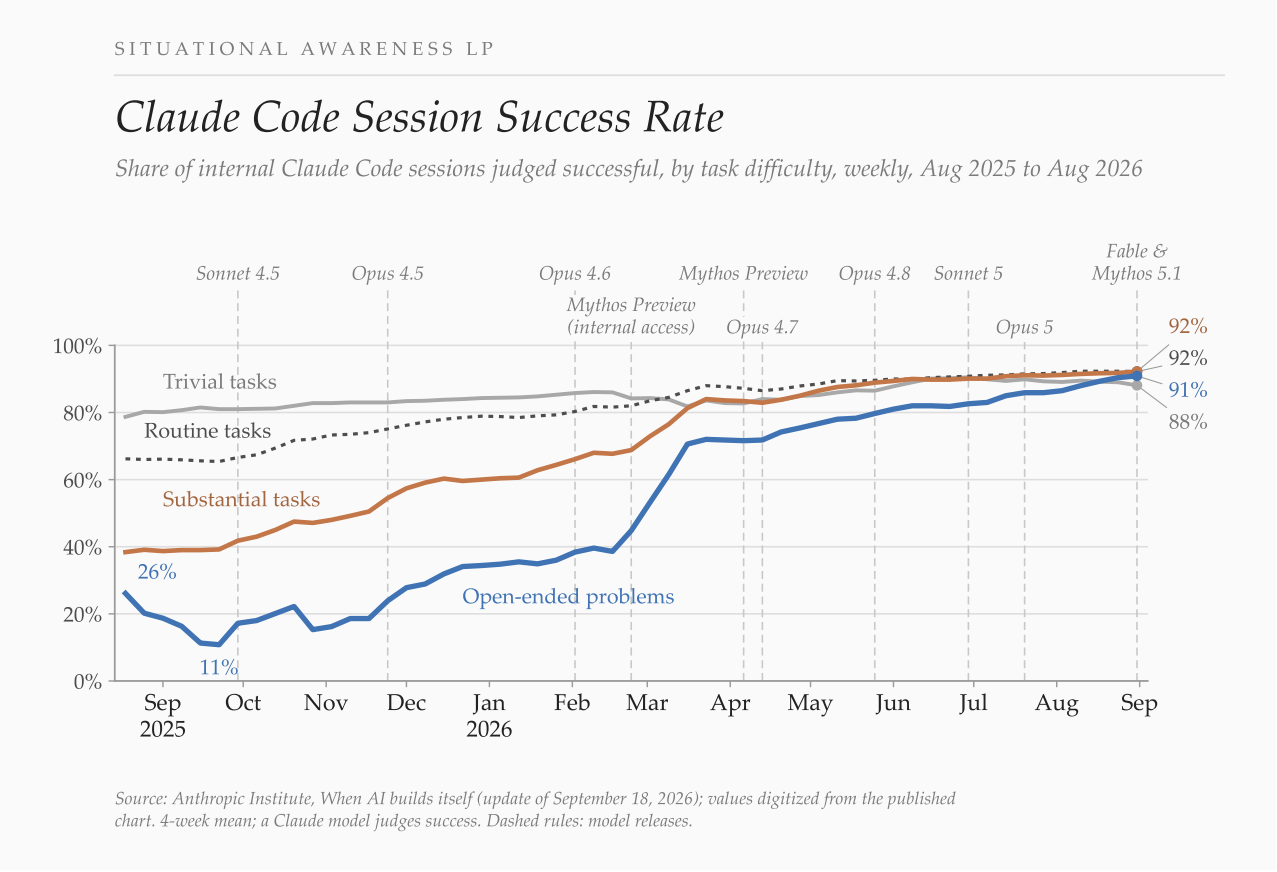


Figure 11: Anthropic data suggests that Claude is driving and succeeding at open-ended work internally.

7. Interestingly, anecdotally the very best people are multiplied much more than the average researcher. They can now idly explore ideas without relying on other people (without needing to manage a team around them). Essentially, a handful of the best people, with the best research taste and judgment, can increasingly execute on a lab's whole research agenda, with the models doing the engineering and iteration. Research compute becomes centralized around them.[14]

8. One common objection to this story is that models will be bottlenecked by "research taste/judgment," the non-engineering parts of research, such as which experiments to run or approaches to try, etc. The metrics we have on this are improving rapidly as well. Experiments measuring "how much compute did the model need for experiments to attain the same algorithmic improvements" (measuring the quality of the experiment ideas etc.), show a doubling time of three months this year, with models reaching human-level

[14] Rather than helping lower-skill researchers, powerful AI is having more impact by raising the ceiling of productivity, rather than by raising the floor.

recently. These are likely easier tasks than those involved in frontier AI research—we don't think models are at the levels of the best human researchers at the labs on frontier AI research taste yet—but the trend is very notable, suggesting no fundamental barrier.

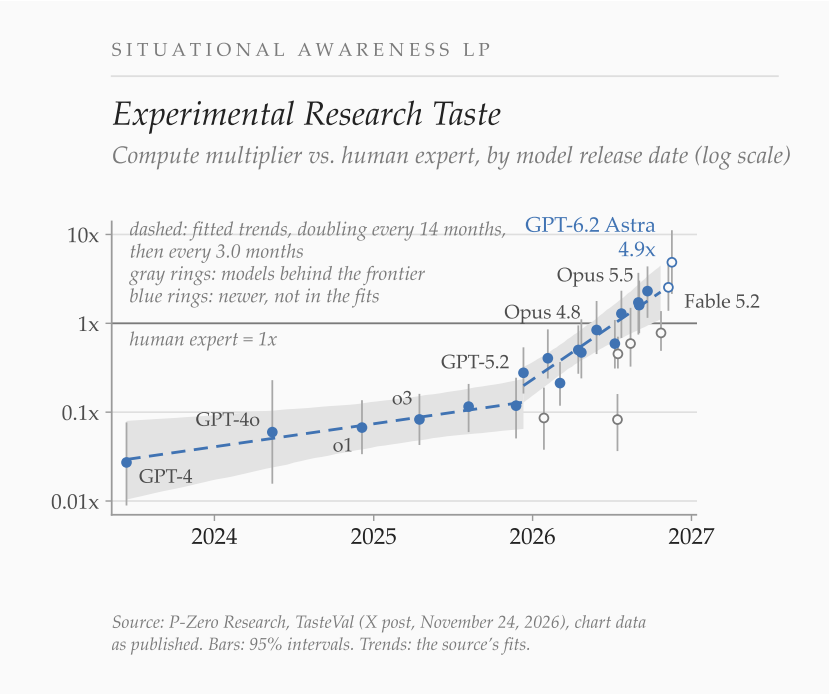


Figure 12: A paper measures "research taste", i.e. models' ability to propose the right experiments in open-source ML, by measuring how much experiment compute is needed to get a result. We are seeing the models rapidly improve, recently reaching human level. Internal lab research is likely harder, and human level higher, but the consistent model improvement on "research taste" suggests this is no fundamental barrier.

The rate of algorithmic progress has begun accelerating

While impressive, datapoints such as the above are still qualitative or input measures. What we really care about is an output measure—how much is this speeding up algorithmic progress (the rate of discovery of compute multipliers) and ultimately overall AI progress?

Recent evidence suggests the rate of progress has started inflecting:

1. Dario Amodei's public statements are notable here. These indicate about a 5% speedup around fall 2025, growing to 15–20% by the spring of this year,[15] growing to "drastically faster" by September of this year.
2. Ideally we would have hard measures of the rate of algorithmic progress, such as the amount of compute needed to reach

[15] Dwarkesh's Dario interview: "Six months ago, it was maybe 5%" and ""I would say right now the coding models give maybe, I don't know, a 15-20% total factor speed up. That's my view."

a certain pretraining perplexity score. Labs have been tight-lipped about such information, but an unlabeled plot released as part of Anthropic's roadshow implies measured pre-training compute multipliers, on at least some tasks, have been growing superexponentially—i.e. the pace of algorithmic progress, as measured by pace of improvement in the compute needed to reach certain pretraining perplexity—is accelerating.

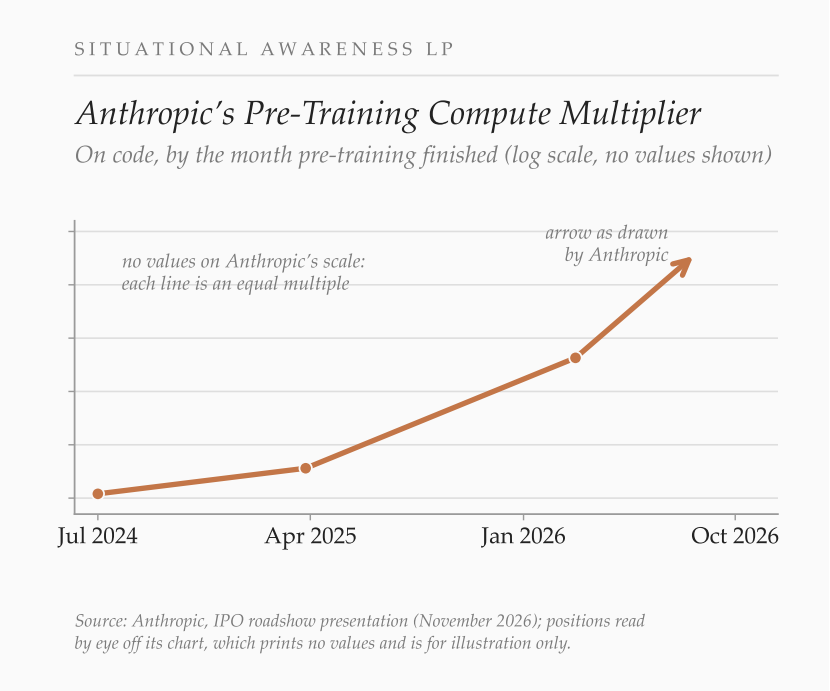


Figure 13: If algorithmic progress were progressing at a steady rate, the pretraining compute multiplier would be a straight line on a log plot. Instead, Anthropic shows this straight line bending upward.

3. METR, an external auditor with access to internal Anthropic data, estimated that as of October Anthropic was "likely improving the capabilities of their models on well-specified tasks at around twice the rate they did between March 2024 and October 2025" and that ~80% of that acceleration in capabilities progress was due to AI.
4. In their recent S-1 update, Anthropic declared that they "may imminently meet" their Responsible Scaling Policy's (RSP) threshold for "dramatically accelerated progress attributable to AI", defined as a >2x acceleration, "based on what we expect from future models and our overall rate of progress."

It's hard to overstate the importance of the last two data points. Anthropic has limited incentive to declare their RSP threshold met, given that meeting it implies they will divert resources away from training toward security and oversight.

The evidence that a significant speedup in algorithmic progress is occurring inside the labs due to AI automation of R&D is increasingly hard to dismiss.

Triangulating the evidence, our current best guess is that in the last couple months, algorithmic progress itself is 2–3x accelerated (labs are finding 2–3x as many compute multipliers in a given period of time as they used to), resulting in roughly 1.5–2x overall acceleration in AI progress right now (i.e. making 1.5–2 years of overall AI progress in one at the current pace, combining the scaling of physical compute with the scaling of algorithmic progress). We think there has been a strong inflection through the course of the year, from a very modest speedup at the beginning of the year—perhaps 5%—to 15–25% in spring, 1.5x by late summer, and accelerating to 2x by year end.[16]

[16] Funnily enough, this lines up almost exactly with the original Situational Awareness essay's prediction two and a half years ago of 1.5–2x speedup by late 2026.

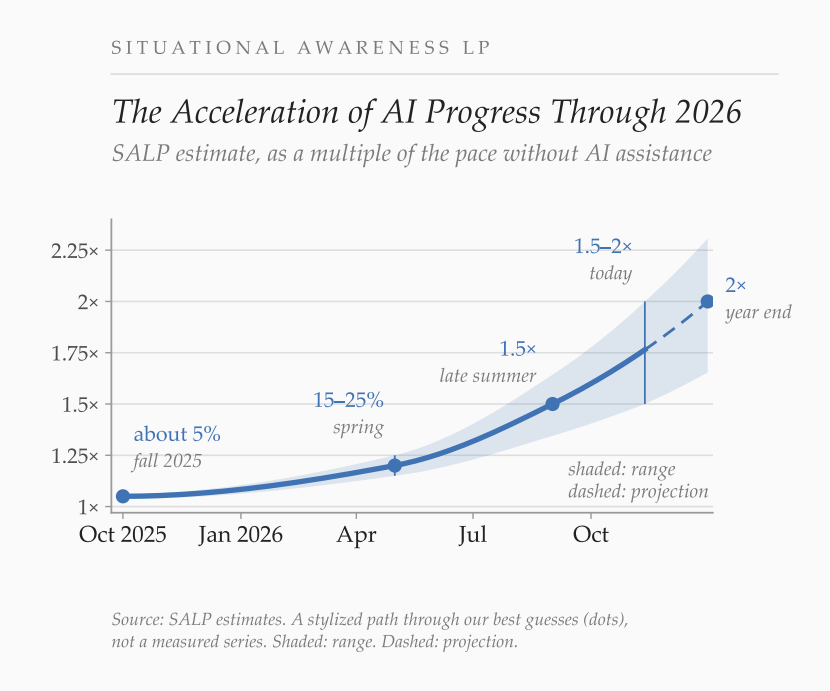


Figure 14: SALP best-guess triangulation of the overall acceleration of AI progress due to AI automation of the research and engineering inside the labs.

Modeling RSI: Two to six years of AI progress in 2027

RSI (recursive self-improvement) is starting.

The implications are profound, though rife with modeling uncertainty. As the models continue to improve, they should further increase the degree of automation in AI research and further increase the speedup. Faster progress should lead to an even faster increase in the degree of speedup. And so on. If AI RSI doubled its pace of progress in <12 months, then going forward the next doubling of the pace of progress should take only 6 months, then in 3 months, and so forth.

But there are many unknowns, and bottlenecks can slow this process. For example, how bad are the serial time bottlenecks that an agent swarm army of parallel agents runs into? How much are we subject to Amdahl's Law/Baumol's cost disease type phenomena, where the models automate large parts of R&D but the remaining human tasks, such as "research taste" become bottlenecks? Are there limits to how much algorithmic progress can be made without faster compute scaling, in particular scaling experiment compute (in the initial phases of RSI, research "labor" is increased while experiment compute is not)? How quickly do ideas get harder to find, i.e. while we have more effective research effort through automation, will we run out of low-hanging-fruit algorithmic progress?

We spend a lot of time understanding and modeling these bottlenecks. Right now, there is broad consensus that labs remain bottlenecked on "ideas" for good experiments; so while they are able to iterate through ideas faster and particularly accelerate areas such as performance engineering, the models are not yet generating breakthrough ideas. Our sense from the relevant metrics and researcher surveys is that they feel optimistic that models' research taste is improving, and that "research taste" can itself be trained with better data and RL environments. Some external experiments (such as the data cited in the bullets above) are consistent with continuous and rapid model improvements on "research taste" as well.

Recent model progress in mathematics, which tends to lead model capabilities in coding/ML research, is also suggestive: earlier in the year the models made math breakthroughs but still in a fairly "rote" fashion, such as enumerating a counterexample; then they made breakthroughs that brilliantly connected ideas in disparate areas of mathematics; most recently, the models were able to solve the Riemann Hypothesis, a centuries-long open problem researchers have widely thought would require ingenious new ideas and totally new areas of math. The mathematician who worked with Claude on the

proof, Levent Alpöge, writes in his announcement, "Claude flew, and danced, and sparkled near that finish line."

In addition to good ideas, even fully automated ML research still requires compute to run experiments (unlike most math). That's why researchers can write 50x as many lines of code but only end up running ~2x as many experiments (according to some published data).[17] Exactly how complementary experimental compute and experimental "labor" are is therefore likely the most important source of uncertainty. For example, key uncertainties include to what extent ideas can be tried at a smaller scale, and how much "better" each experiment can be via AIs helping with better experiment setup. The extended discussion in the original *Situational Awareness* essay concludes that this is a bottleneck that will slow RSI from the most extreme levels, but isn't enough to stop drastically-accelerated progress, and I think it stands.

[17] Anthropic's S-1 discloses that they are merging >45x as much code as they used to. OpenAI's recent research acceleration report discloses that they are running 1.7x as many experiments in October 2026 as in 2025. Given the lead Anthropic has on OAI on most of these metrics, we think it is likely that Anthropic is running 2x+ their 2025 rate of experiments.

In our modeling, our central estimates of compute bottlenecks and the pace of improvement in research taste are consistent with anywhere from 3–12x algorithmic progress next year (i.e., 1.5x-6x overall AI progress). That is, we think further acceleration from the current speedup is likely.

One intuition pump: we recently saw what was described as a decade's worth of mathematics progress in a month, with the recent internal OpenAI model—a speedup of more than 100x. In a year, we should expect models to be as good at ML research as they are at math today, though experimental compute will still bind.

Considerable uncertainty remains, but our best guess is that 2027 is about a "3x" year, i.e. we see as much overall AI progress in 2027 as we would have seen in three years at the baseline pace of 2024/2025.

The implication is stunning: we should expect about as much progress as we've seen between the original ChatGPT or GPT-4 (smart high schooler) and current models such as Opus 5.5 (automating much of the engineering in AI labs, PhD intern in other fields), on top of current models, by the end of 2027. It is worth dwelling on how extraordinary that is, and the stark implications of that much qualitative progress on top of existing models in the near future.

In turn, that implies that the "run rate" of progress at the end of 2027 (rather than the average across CY27) would be considerably faster still—perhaps 5–10x+. Beyond incrementally automating more and

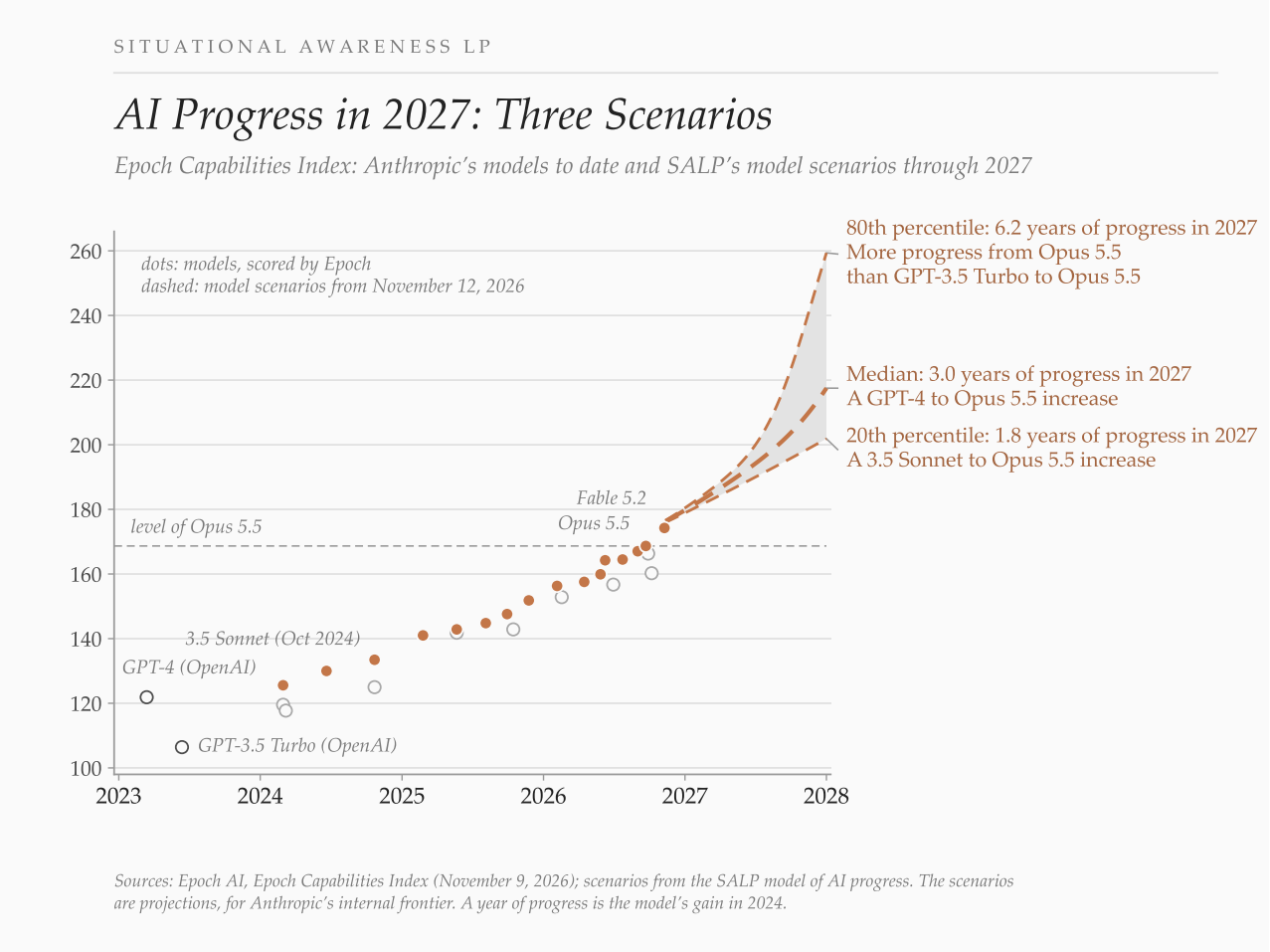


Figure 15: Calibrated to datapoints available and modeling bottlenecks such as serial time, research taste, and experiment compute, SALP's internal model of AI progress predicts 2-6 years (median: 3 years) equivalent of 2024's AI progress in 2027.

more of the engineering as we see today, it seems highly plausible to us that we will get superhuman automated AI researchers by the end of 2027—full-on RSI.

Anthropic's safety roadmap calls it "plausible, as soon as early 2027" that its systems could "fully automate, or otherwise dramatically accelerate, the work of large, top-tier teams of human researchers."

We are headed to full-on superintelligence in 2028. Things will get pretty crazy.

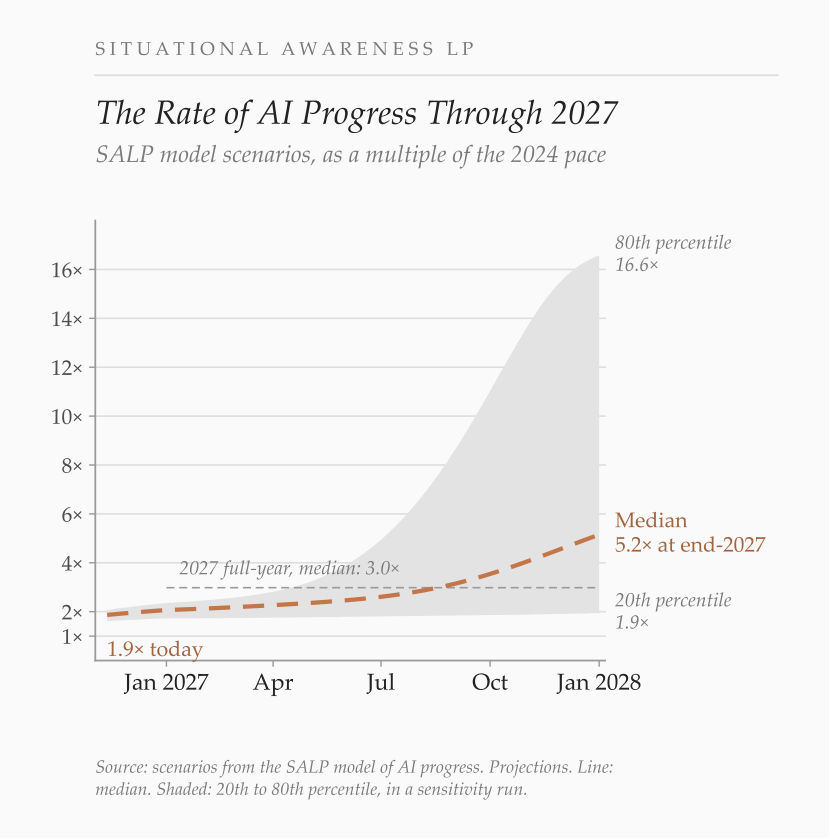


Figure 16: SALP modeling of the instantaneous rate of AI progress.

Physical RSI

If the story of 2027 is "algorithmic progress RSI" (i.e., an acceleration in algorithmic progress due to AI automating AI R&D), we think the story of 2028+ is going to be "physical RSI"—i.e., the AI models contributing to an acceleration in compute scaling itself.

We're first starting to see this in chip design. OpenAI designed its first-generation Jalapeno chip in record time, using AI to accelerate initial design to tapeout in nine months, compared to about one and a half to two years for a typical accelerator. We think Jalapeno could be one of the best available chips on a performance/watt basis for its use case when it ships next year, comparable to Nvidia's latest Rubin GPUs on certain measures. Moreover, OpenAI models recently rewrote circuits in AMD's flagship GPU—leading to material gains, comparable to a year of progress, on top of a chip humans had already spent a lot of time optimizing.

Chip design is fundamentally a design file sent to TSMC. Today, AI will help optimize that file; in a couple of years, AI will be better than humans at generating that file.

But it will go beyond chip design. AI will make leaps on everything from semiconductor manufacturing to datacenter construction. First, it will do so via better designs. Then, AI can work through human hands: any unskilled worker can become a skilled worker (executing on much-better-AI-plans) with an AI coach and Meta glasses.

Finally, robotics will scale (cf. our later section on the coming ChatGPT moment for robotics). Pretty soon, the robots will build more robots, and each factory's output goes into building the next, with fewer and fewer labor constraints to slow the compounding.

By 2028/29, the two loops will compound. Today, RSI is accelerating algorithmic progress, while frontier lab compute still grows about 3x a year, limited by how fast humans can design chips and build fabs, power plants, and datacenters. Physical RSI removes that limit. Once AI designs the chips and robots build the datacenters, compute scaling accelerates as well, and each loop feeds the other.[18]

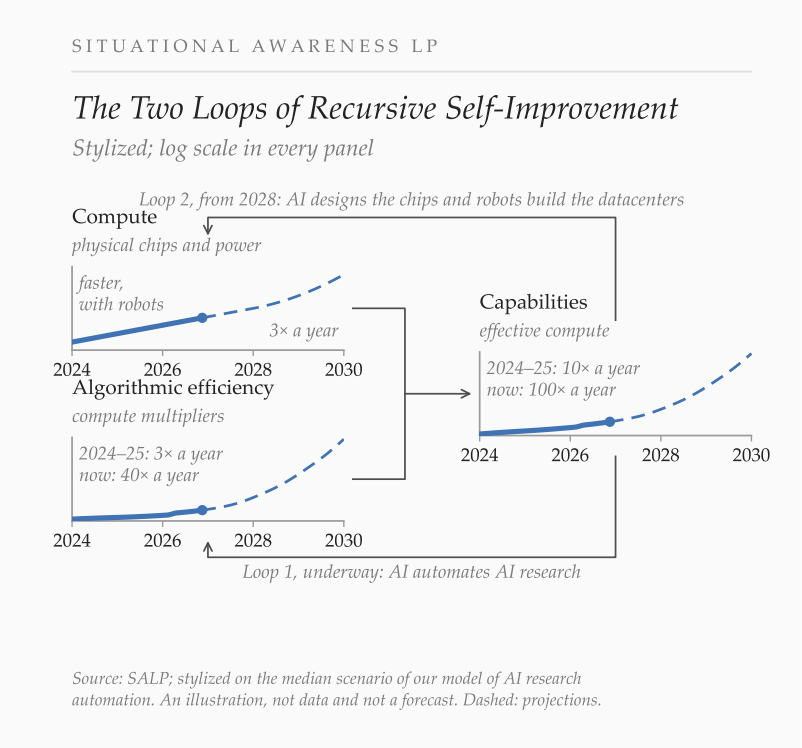


Figure 17: The AI scaling loop, 2027 onward. The straight lines—both on algorithmic progress (starting 2026/2027) and compute (starting in 2028/29)—will bend upward.

[18] An interesting analysis of this "industrial explosion," to follow the "intelligence explosion" from algorithmic progress, is Damon Binder's series "The AI Industrial Explosion", building on many arguments Carl has made over the last decade.

Ib. The year ahead: from the "virtual intern" to "virtual department"

The story of the past year was the jump from chatbots to agents. We think a new paradigm shift is coming, to Highly Persistent Multi-Agent (HPMA) models. Rather than the "virtual interns" current models are for most people, the systems now running inside the labs are "virtual departments": hundreds of agents working persistently for weeks under their own managers. Math and software engineering are the leading indicators of what is to come more broadly.

The last three years of AI progress, from GPT-4 to Opus 5.5, took us from chatbots to agents; from ChatGPT to Claude Code/Claude Cowork; from short Google+ answers to being able to do longer tasks and streams of work using tools; from $20/month to $1000s/month on AI.

Through the end of 2027, we expect a similarly-sized chunk of AI progress. What might this look like concretely? There are a few clear paradigm shifts that are starting to become visible to us.

From the "virtual intern" to "virtual department": Highly Persistent Multi-Agent (HPMA) models

The agents in the form of Claude Code and Claude Cowork, as predominantly used outside the labs, are still limited: they are blocked on the user giving them direction; most folks sit in front of a single Claude Code / Cowork window, watching the work being done and

giving it input. We currently essentially have the "*virtual intern*": a smart intern that's good with computers sitting next to you, taking your direction and doing parts of tasks, while you watch over their shoulder and give input. It is still largely synchronous, not full delegation.

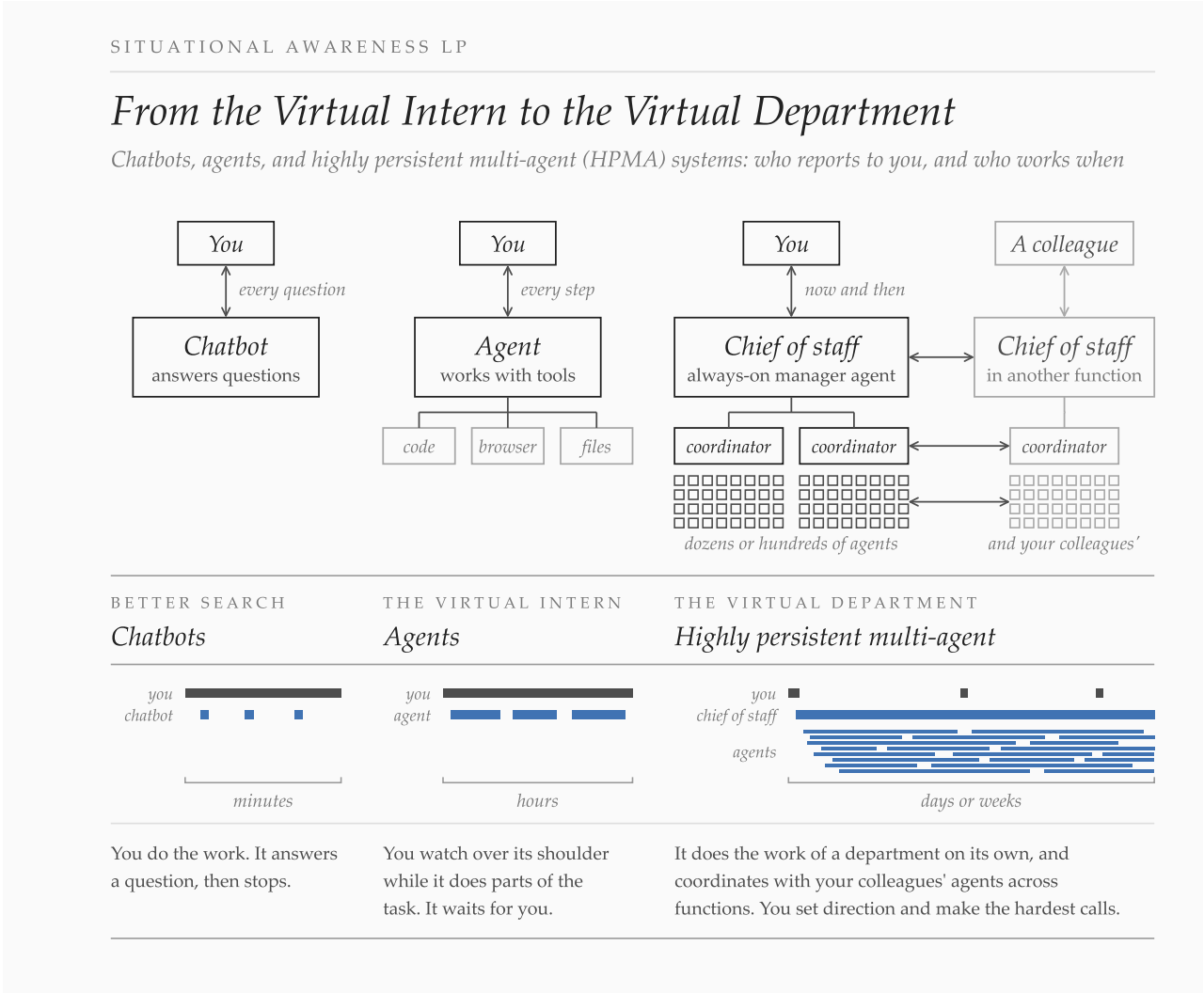


Figure 18: Much like the chatbot → agent transition, we think there is an agent → HPMA paradigm shift coming.

In the last few months, we've observed a growing gap between most external use and internal use of the models at the labs. We've gotten glimpses of the future from both math releases and reports on internal use for AI R&D (cf. "What does AI research now look like at labs" section earlier). We think this is due to new models (most internal/not yet released, with training in multi-agent cooperation

and persistence) that work in a new paradigm we call ***HPMA: highly persistent multi-agent models***.

- This involves many agents—hundreds or thousands—working on projects or areas in a highly persistent way (running for days or weeks or more, rather than little chunks).
- They are coordinated by always-on manager agents—sort of like a Chief of Staff, who knows everything about you and your organization and what the job is and the context is—and other agents doling out and coordinating tasks etc.
- This almost looks like a mini-bureaucracy inside of a corporation.[19]
- These agents can also talk to other agents that are part of the HPMA swarm of one of your colleagues.
- The 10,000-agent swarm of OpenAI's internal model that solved the recent Millennium Prize problem Navier–Stokes, or the agent swarm that hacked Hugging Face, are early examples of these.[20]
- Think of this more as an asynchronous "virtual department" than a synchronous "virtual intern."

Token use at the labs skyrocketed as the chatbot to agent transition happened, and then plateaued for a bit early in the summer, as it did externally. But recent data suggests it has taken off again since the summer. Gaps between internal use and external use are highly informative—lab internal use is usually a very strong leading indicator, foreshadowing the products, form factors, and workflows to come.

And this is not just lab internal token-maxxing. *The KPI of an AI lab, algorithmic progress, has 2-3x'd.* Labs are getting massive productivity gains (both in coding and non-coding roles, as we'll discuss further) from HPMA, leading the intense HPMA model use to survive cutthroat lab internal compute allocation processes.[21]

While at present much of these capabilities are internal to the labs, we have been experimenting with a version of this internally (enabled in particular by the recent Opus 5.5 jump). We call it GRIMM (in homage to Fable), featuring hundreds of Claudes that can talk to each other and share a (Claude-built) Asana-style task board, communication between Claudes, manager Claudes, and many other quirks (e.g. a "coach" Claude that motivates the other

[19] Except that as the models learn and improve, they can become much better at coordinating (something humans often struggle with). The models currently are probably still worse than groups of humans at collaborating; soon they will work together like humans; then they will have vastly superhuman ways of coordinating and collaborating. They can be directly trained for their contribution to organizational goals in cooperation, unlike humans with more selfish goals such as getting good performance reviews.

[20] Per OpenAI: "We used a system of coordinating agents powered by our internal model. The agents had access to tools such as the ability to read from a cached version of the internet and the ability to run code. Agents were subdivided into groups with the ability to communicate within the group. The groups varied in size, and the group that produced the Navier–Stokes resolution involved on the order of 10,000 concurrent agents. At all times we maintained the same strict safeguards that we apply to all our frontier model evaluations, including monitoring and isolation."

[21] We have heard compute allocation at an AI lab analogized to capital allocation to PMs at a multistrategy hedge fund firm.

Claudes). Patrick McKenzie (patio11), an internet-famous software engineer, who sometimes consults with us, described it as follows:

> 1) This is clearly the future.
>
> 2) Oh we're going to buy so much inference.
>
> 3) Barely feels like software engineering but I end the day intellectually spent.
>
> 4) I think every estimate of aggregate speed up or increase in total software shipped will be an underestimate. This is wild. It is wild even compared to the single-threaded LLM-assisted modern coding tool baseline.

Patrick wrote a longer assessment of GRIMM and the HPMA workflows he saw working with us here. Crucially, he describes the Claude Code → HPMA transition as an equally big deal as the original chatbot → Claude Code transition.

While we are far from the lab-internal frontier of usage, as an early-adopter externally we get a sense of what might be coming more broadly. We have seen corresponding jumps in token usage. For example, one of our team members ramped his token spend 10x over a matter of months as he made this transition (on top of already very intense Claude Code use). Patrick remarked that a handful of engineers using this system had built and maintained software which well-regarded engineering organizations would historically have needed ~60 FTEs to support. We will discuss some of what we've built later in this letter.

While it's obviously an intense level of usage, for us, the ROI is clearly there. We are getting a tremendous amount of leverage and productivity on a small number of people. And being better at our jobs pretty directly results in additional P&L, i.e. additional revenue to pay for the Claude spending.

Many kinds of work in summer 2027 might look more like the following. You have an always-on model, call it Charlie, that is your Chief of Staff. Charlie understands your work and company very well, and is coordinating hundreds of agents in the background doing the work of an entire department at your company. Some tasks require an agent swarm working for days or weeks, with a whole internal bureaucracy; others are just one-off agents overseen by Charlie that take care of specific responsibilities. These agents are all talking to each other, and to your colleagues' agents; my Charlie is talking to my colleague's chief of staff; the sum of our agents talking to each

other like a lively work Slack. The human role moves to a higher level managerial role: from an IC (pre-AI) to an IC-with-a-smart-intern (2025/26), to a manager of a department. For a little while (before AI systems run entire companies), the human role will be high level direction or making the trickiest judgments or connecting to high stakes external relationships, while your underlings (agent swarms) do most of the work.

01/03 made clear to us that RL was working and thus agents would work. This would lead to another ChatGPT-like moment—but the o-series itself had limited usefulness at the time. So too with HPMA: It's clear that this is the future, but the form factor and capabilities are still limited, and so only yet useful for narrow domains. However, with a bit more time to find the right form factor, a bit more capability progress to make them more broadly applicable, and the right productization, we think HPMA will be to 2027 as agents were to 2026—and will be a step change in both usefulness and monetization.

Math → code → broader knowledge work

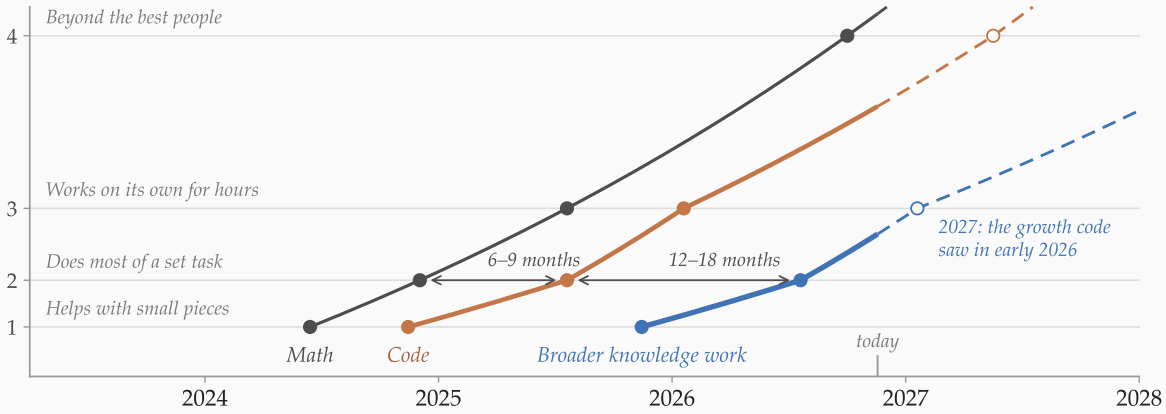
AI progress in recent years has tended to follow a familiar trend: first comes math, then comes code, then come other domains of knowledge work. For example, 01/03 initially produced stunning performance on math benchmarks, but weren't that useful for everything else; then coding assistants followed; then broader white collar work with e.g. Claude Cowork. Math has tended to run six to nine months ahead of coding, which in turn has tended to run ~12-18 months ahead of broader white collar work (with a lot of unevenness of course). Looking to math and code gives us a preview of what to expect more broadly.

Recent internal HPMA systems are making breakthroughs in math. Recall that at the time of *Situational Awareness*, labs were still bragging about their scores on a benchmark called "grade school math". Now, models have solved two Millennium Prize problems that bedeviled human researchers for centuries. We should expect these sorts of results in broader fields in the next year or so: HPMA is already in the very early stages of coming for coding—including accelerating AI progress itself—but also more broadly thereafter, helping make scientific and engineering and technical breakthroughs across fields. For example, Anthropic is working hard to bridge the models to a

SITUATIONAL AWARENESS LP

First Math, Then Code, Then Broader Knowledge Work

Three fields climb the same four levels, each later than the last. A schematic: the curves are drawn, not measured



		Math	*Code*	*Broader knowledge work*
1	*Helps with small pieces*	Mid-2024: labs still cite "grade school math" scores	2024: a few lines at a time	2025: the occasional Deep Research query
2	*Does most of a set task*	Late 2024: o1 and o3 excel at very hard math benchmarks	Summer 2025: Claude writes most of the code	Summer 2026: Claude does most of the modeling
3	*Works on its own for hours*	Mid-2025: gold-medal level at the Math Olympiad	Winter 2025–26: a few Claudes in parallel	*Starting now: a few Claudes in parallel*
4	*Beyond the best people*	2026: a Millennium Prize problem is solved	*Expected within a year or so*	*Expected after that*

Source: SALP. Schematic. The levels and the curves are ours; the dated examples are from the labs' releases and from our own analysts' work. Dashed lines and open dots are our expectation.

Figure 19: The pattern of recent AI progress is uneven locally, but consistent inter-temporally: mathematics ability leads, then coding ability 6-9 months later, then broader knowledge work another 12-18 months later.

wet lab the models can use to run experiments, and Dario Amodei predicts breakthroughs in life sciences in 2027.

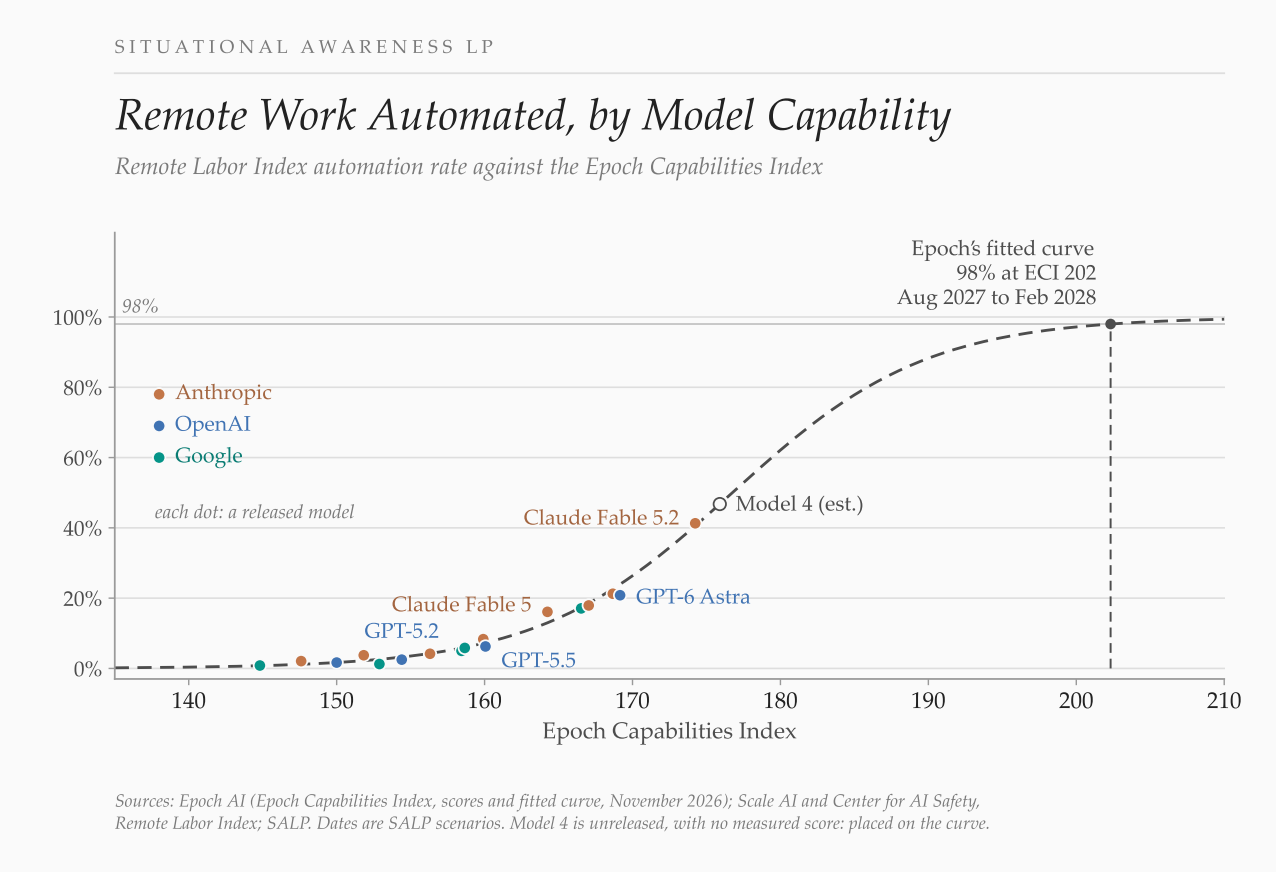


Figure 20: Until recently, models only showed signs of life on actually being able to do broader white collar work, but we think we are starting to hit an inflection point.

Beyond frontier R&D, we think we haven't even really seen the full "Claude Code" explosion for broader white collar work yet—but we expect it in 2027. Take the examples of our analysts:

- A year ago (2025), they were using some AI, but only the occasional Deep Research query or similar—much like coding models became useful to AI lab engineers a year earlier (2024), but mostly in a "write a few lines of code" sense.
- By this summer (2026), the work of our analysts looked much like that of coding in summer 2025. At the time, software engineers had one screen with their IDE and code, one screen with Claude; they still looked at all of their code by hand, but Claude was starting to drive a majority of the actual writing of code. In summer of 2026, our analysts had one screen on Excel, one on Claude, with Claude starting to drive a majority of the modeling.
- As of a month or two ago, with Opus 5.5, our analysts are starting to experience the transition coders experienced late 2025: I barely

see Excel on their screens anymore, just a few Claudes in parallel, with essentially all of the work being driven by a Claude session.

While HPMA will drive a new wave of explosive growth in coding, and drive much higher usage intensity for frontier R&D across fields, we think we will also see broad-based white-collar work start to undergo a usage transition that looks a lot like the usage transition software engineers experienced this year. 2026 featured the "drop-in software engineer", 2027 will feature the "drop-in remote knowledge worker".

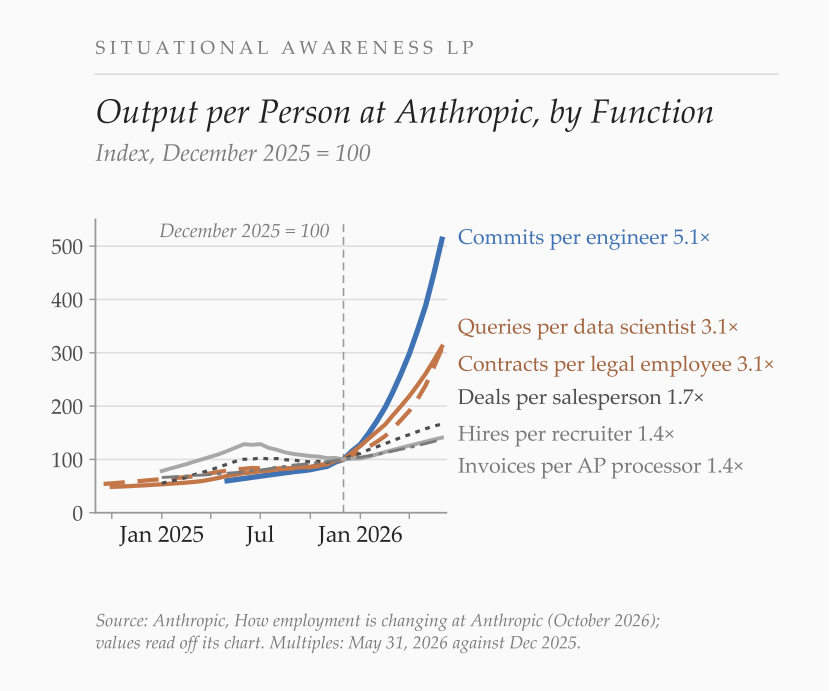


Figure 21: Anthropic is seeing increasing productivity uplifts in non-coding functions as well, lagging behind coding but on a fast-growing growth curve as well.

Ic. A trillion dollars of revenue

Every year the market decides the thing in front of it is the whole story; two years ago it fixated on chatbot TAM, today it concludes coding agents are saturated and this is the end of the road. We think lab internal use is a leading indicator (lab model usage is a majority of opex, vs. early-adopter enterprises at 1%, and the average Fortune 500 at <0.1%), and diffusion is overrated while intensification is underrated. We think the labs will exit 2027 north of $1T in revenue run rate.

Anthropic grew from $1B ARR at the beginning of 2025 to $81B ARR in October of 2026. The market thinks Anthropic revenue will rapidly decelerate next year. We think buyside estimates are around $120B-$200B ARR EOY27. One of my colleagues recently remarked to a group of AI-focused hedge fund analysts on a trip to Asia that he was optimistic about lab revenue—"oh, so you're in the $180B camp?" they responded. We think they are mistaking the lumpiness of overlapping paradigm shifts for the bigger picture.

In *Situational Awareness*, I wrote about the path to $100B of revenue for the labs. But "beyond $100B," I wrote, "it gets harder to see the contours." For us, the contours are now becoming visible.

Mapping AI capabilities progress to lab revenue

We feel confident that—barring extreme political responses—2027 will be the most significant year of AI progress witnessed thus far. The question then becomes what that means for the AI labs' economics.

Historically, lab revenue[22] has tightly tracked capabilities, over multiple orders of magnitude.

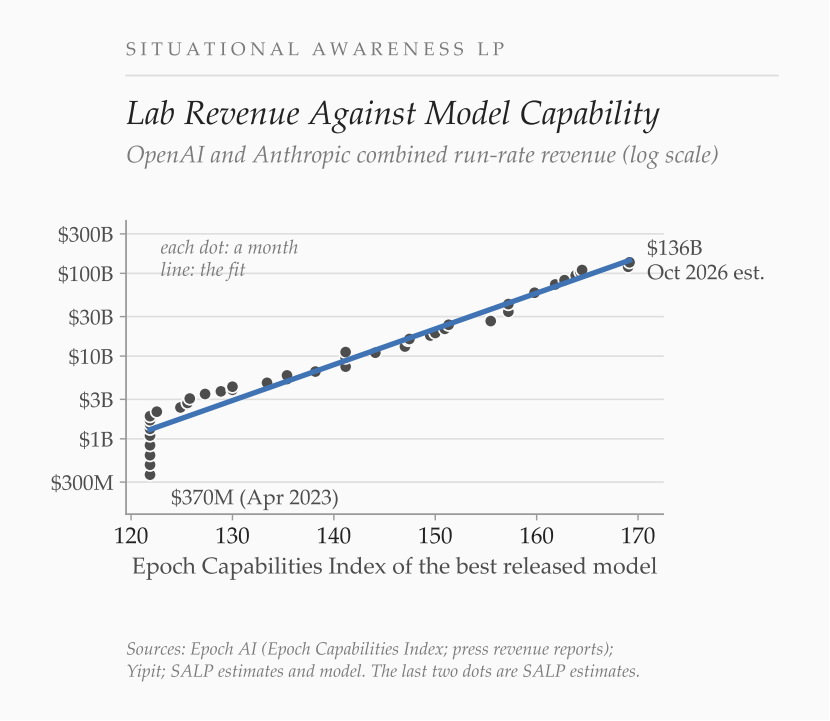


Figure 22: The relationship between an index of AI capabilities (Epoch Capabilities Index, ECI) and OpenAI+Anthropic revenue.

[22] We will generically use "lab revenue" to refer to Anthropic and OpenAI's combined revenue. This is missing other important AI revenue streams: Google Deepmind, AI-based Google search, open source, improved ads for Meta and co, etc., but is easiest to measure and perhaps the most central metric.

The buyside forecast, mapped to this plot, would thus imply only 0.3-0.6x AI progress next year. Naively extrapolating this fit to a world where three years of progress are made in 2027 next year would imply much faster revenue growth: around $6T of revenue combined across Anthropic and OpenAI.

Of course, there are adoption frictions. One way of modeling this relationship is assuming that the capabilities frontier at a given point in time gives you some "ceiling" of revenue. Think of this as the TAM for that set of capabilities—the potential revenue if that set of capabilities was fully rolled out and adopted. Revenue catches up to that ceiling with a half-life, as the capabilities are rolled out and adopted.

Fitting such a model to the data moderates the revenue growth next year; perhaps RSI means we have miraculous model capabilities that roll out in December of 2027, but they take time to diffuse and be adopted. Still, this fit would imply $2T of combined lab revenue run rate in our central case of "three years of AI progress in 2027". Of course, these are all just naive models. But perhaps this would be

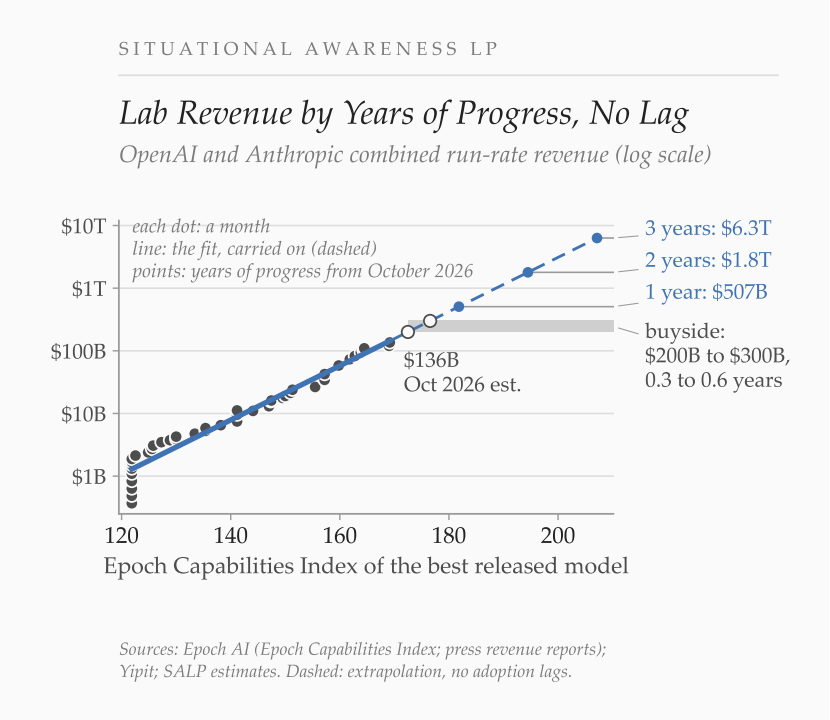


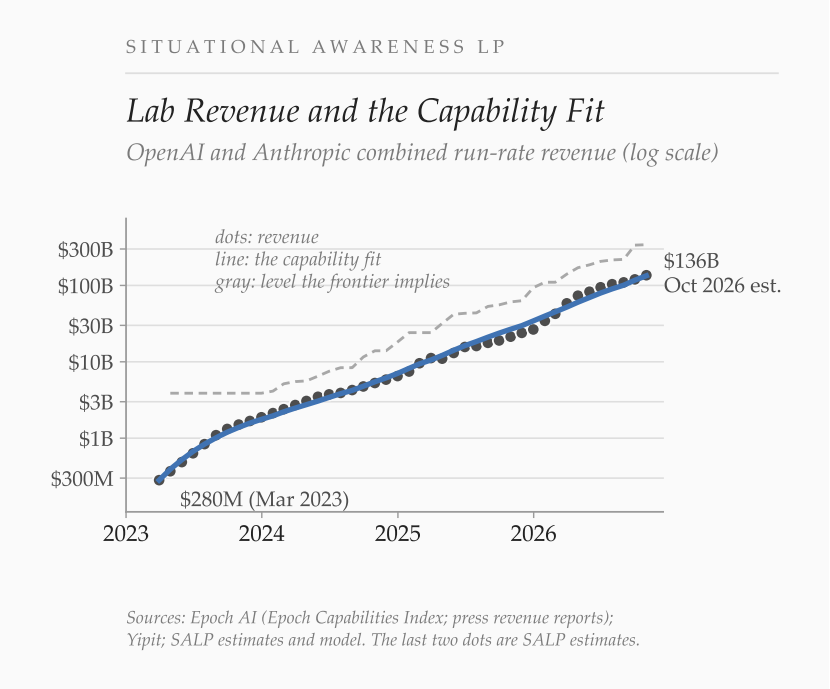
Figure 23: Buyside revenue forecasts map to an implied 0.3-0.6x AI progress next year—even a normal year of AI progress would imply $500B of combined revenue run rate EOY27, and dramatically accelerated AI progress would naively imply trillions of revenue run rate.

a reasonable prior or starting point for reasoning about next year's revenue.

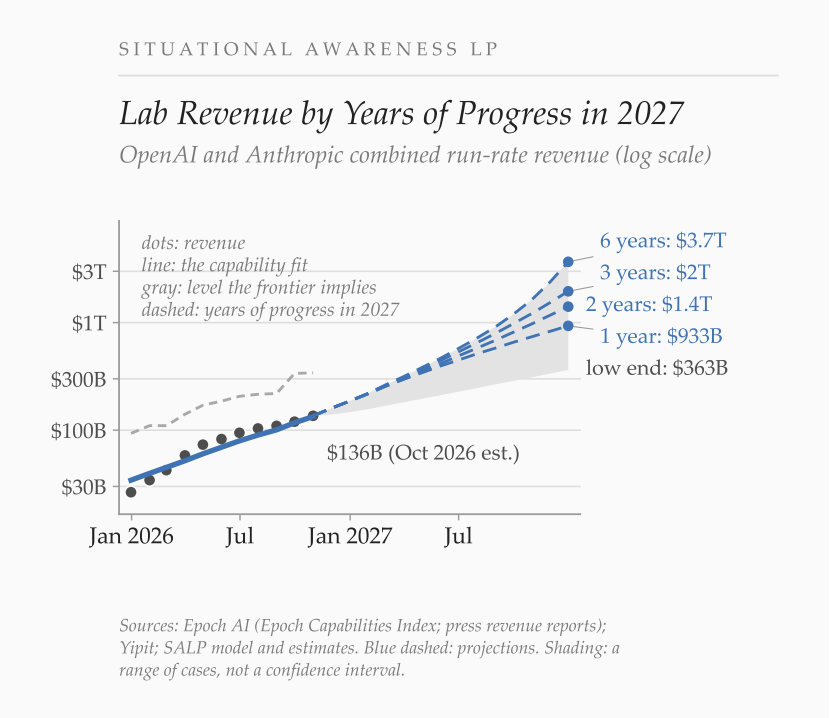
In that sense, the $120B-$200B Anthropic revenue run-rate forecast is the radical one. Consensus estimates expect Anthropic revenue growth to slow by a factor of five or more, in a year in which we expect AI progress to speed up multiple-fold. By contrast, in the RSI trajectory we expect, even $1T of Anthropic revenue run rate wouldn't be crazy.

Clearly, the market has a different view. Speaking to other market participants, the view is something like the following:

- AI coding was great and led to a revenue explosion. But, they claim, we are close to saturation.
- No other use cases seemed to be appearing. AI assistants may become helpful for other things, but this will happen gradually over the next 20 years.
- We are thus on the wrong side of an S-curve.



(a)



(b)

Figure 24: Fitting historical capabilities and revenue to a model; in this fit, actual revenue converges to "ceiling revenue" with a half-life of approx. half a year.

Of course, we agree it's an S-curve—everything is an S-curve.[23] The question is just: where is the top of the S-curve. Every year, the market ends up focused on whatever is in front of it and assumes that's the end of the S-curve. Each of the last two years, the same claims were made. At the end of 2024, that Google++ chatbot use cases were running out, failing to anticipate the new use cases reasoning models would unlock. At the end of 2025, that coding assistants had little more room to run, failing to anticipate the new workflows more agentic and autonomous coding models would allow for.

[23] Due, at least, to the speed of light.

Two paths to $1T: white collar diffusion and HPMA intensification

We see things differently. The question is not, how far do current capabilities have left to diffuse; the question is what new markets next year's capabilities will enable. We've just discussed two of those: HPMA and "the (full) Claude Code moment for the rest of white collar work."

One way we get to $1T in revenue is expanding Claude Code-level intensity from software to the rest of white collar work. We are still early on the "drop-in remote worker" S-curve. As these capabilities progress amid next year's rapid AI progress, the TAM models can penetrate will expand from coding to the broader white collar universe.

- Currently, "coding" uses of models account for ~$90B of ARR, equivalent to ~15% of the US software, IT, and data-science wage bill.
  - Adjusting for non-US spend, the share of software-related wages spent on AI coding uses in the US is likely ~10%.[24]
- The worldwide white collar wage bill is ~$30T, with about $10T in the US.[25]
- If models can penetrate global white collar work at ~one-fifth the rate of coding (2%), that represents $600B in revenue.
- And sufficiently good "drop-in remote workers" should capture much more than 2% of global white-collar wages. At merely the same penetration rate as coding, white collar work would represent $3T of revenue.

[24] SemiAnalysis estimates a coding market ARR of $78B at the end of Q3, $109B at the end of Q4. US wage bill for computer and mathematical occupations is $632B: 5.26 million employed at an annual mean wage of $120,080 (Bureau of Labor Statistics, Occupational Employment and Wage Statistics, May 2025). Anthropic's S-1 discloses that 71% of revenue comes from the US.

[25] US: jobs that can be done from home account for 46% of US wages (Dingel and Neiman, 2020), applied to 2025 US compensation of employees of $15.7 trillion (Bureau of Economic Analysis, NIPA Table 6.2D). Global: EY estimates 2026 labor cost in white-collar industries at $31 trillion.

But another way to get there is even further intensification: an even smaller set of the most highly productive enterprises starting to use Claude more like Anthropic or OpenAI (or a couple of our colleagues at SALP) do, with the HPMA paradigm. Perhaps, much as the future of the AI business model wasn't in how many billions of users could use chatbots at what ARPU, but instead a much smaller number of users spending $1000s/month, the future once again lies in intensification.

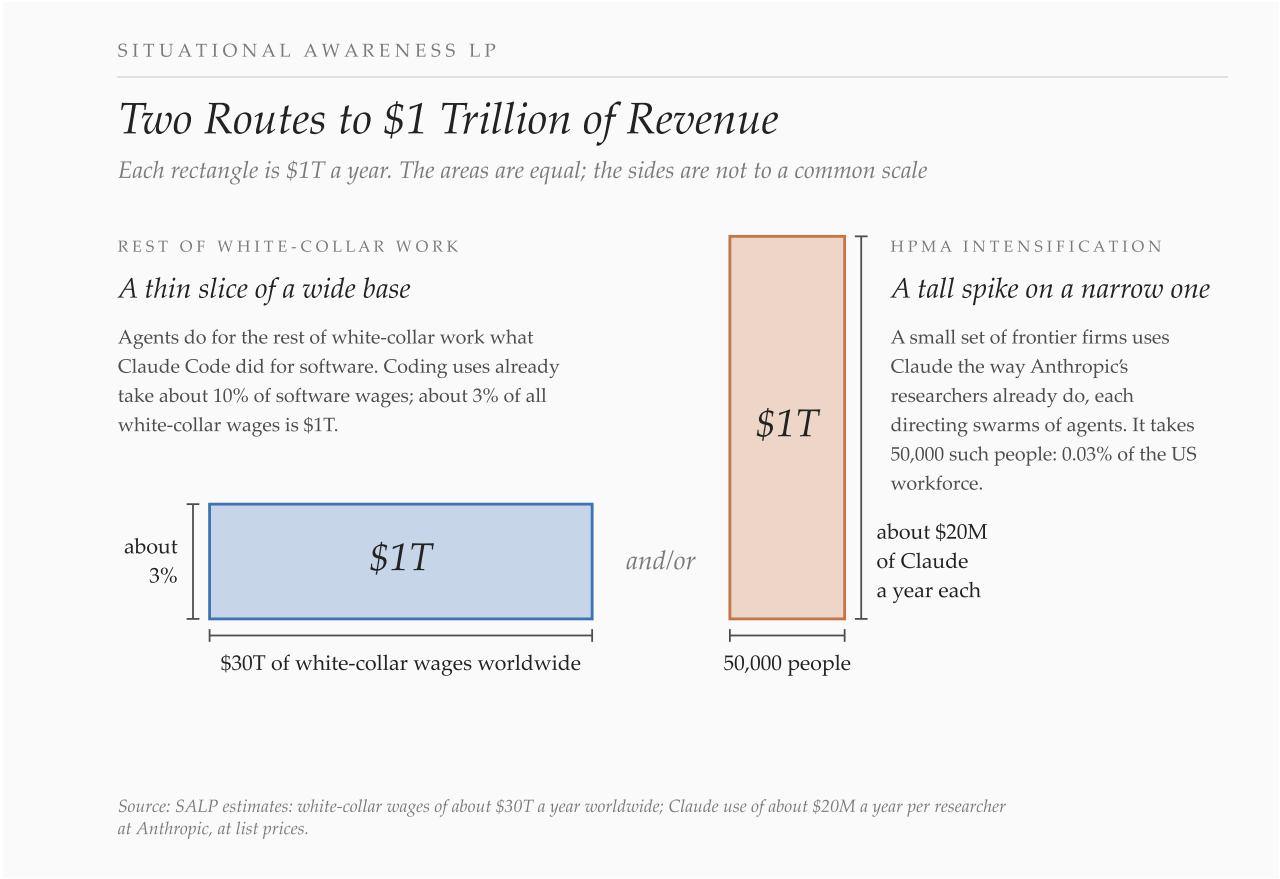


Figure 25: Two routes to $1T of lab revenue next year: broader white collar labor adopts "drop-in remote workers" at ~one-fifth the intensity that coding has, or 50,000 people in frontier R&D and high-value industries use HPMA systems like Anthropic and OpenAI researchers or some people at SALP. (I think we probably get both.)

For $1T in revenue, it only takes 50,000 people—approx. 0.03% of the US workforce—to use Claude the way we believe the typical researcher at Anthropic does: a couple million dollars of Claude spend a month.

At a more macro level, we think the runway for intensification continues to be vast:

- The average Fortune 500 company spends <0.1% of opex on AI.
- Frontier adopters, such as Stripe (a software company that is on the vanguard of new technology!), spend about 1% of opex on AI.[26]
- Our own organization has ramped to >10% of opex (i.e., the management company budget) on AI.
- At Anthropic, we think their internal AI usage (at list prices for tokens) has surpassed their wage bill (including stock), i.e. a majority of (non-serving and training) opex is AI.[27,28]

[26] Even Anthropic's top 10 Fortune 500 customers spend <1% of expenses on Claude. (Anthropic roadshow)

[27] Right now, Anthropic's disclosures suggest that more than a third (!) of all worldwide Claude usage is internal use at Anthropic itself—just one company.

[28] "Coding is saturated" fails at the gap between the level of intensity of external coding agent use versus internal use at Anthropic and OpenAI.

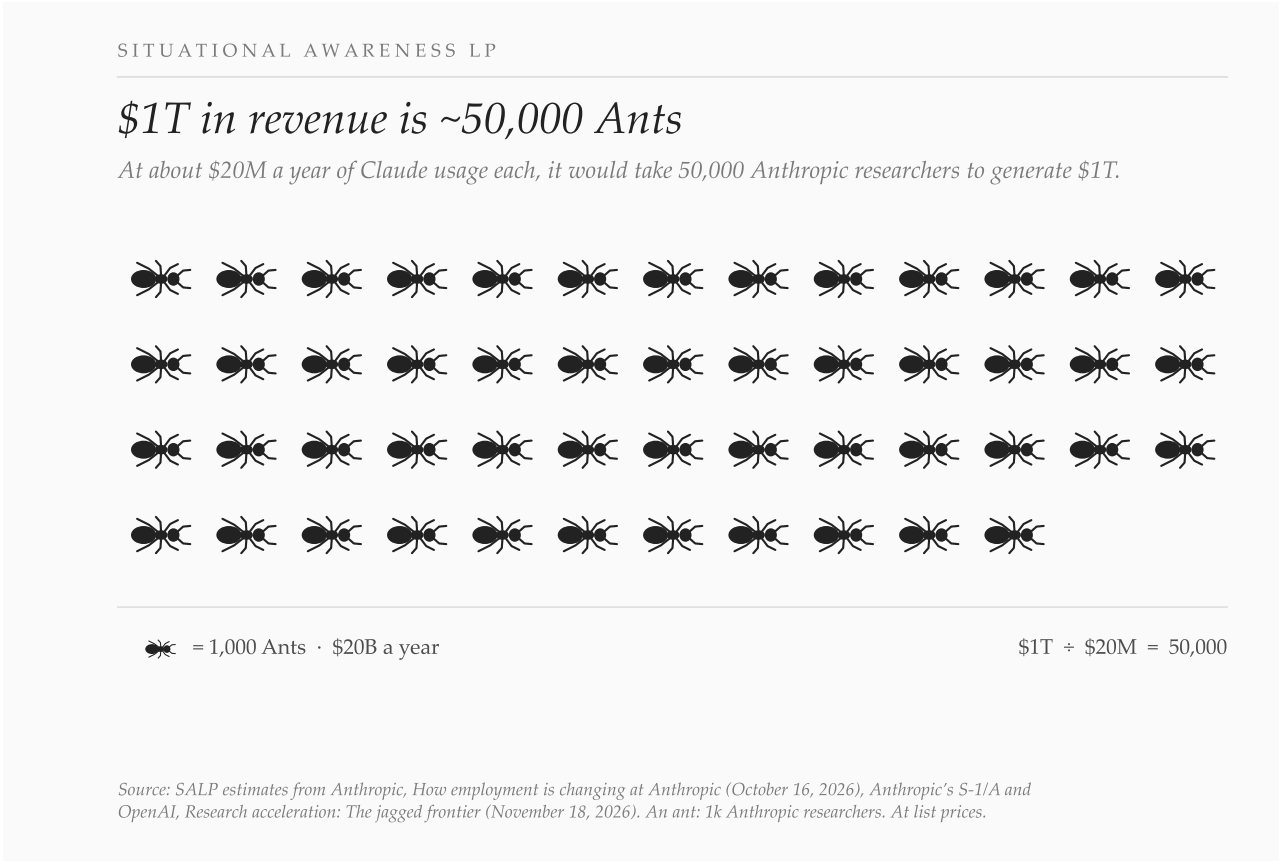


Figure 26: It would take just 50,000 people using models the same way the average Anthropic researcher uses Claude for $1T of revenue.

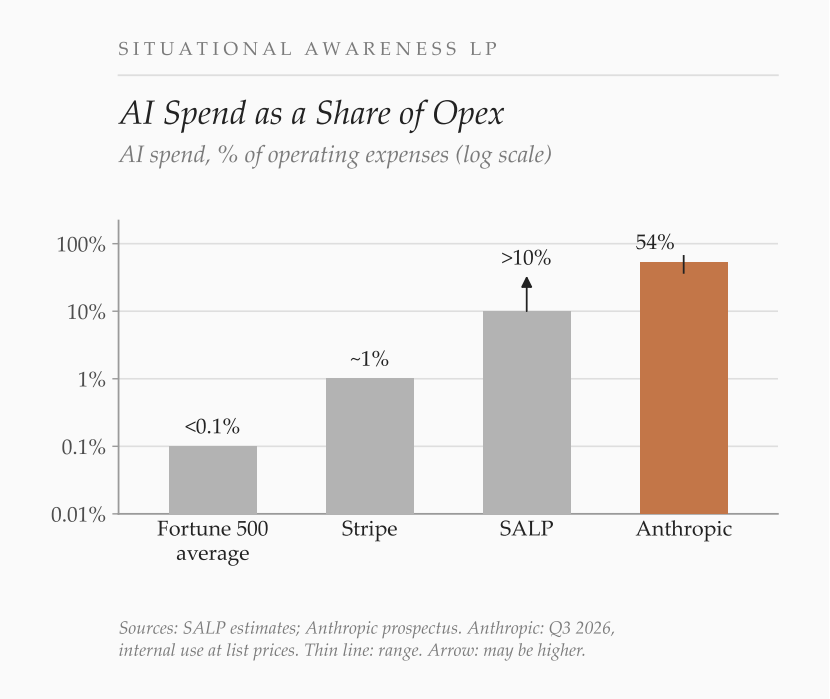


Figure 27: AI spend as a % of opex.

AI lab internal use is typically a good leading indicator for the rest of the economy:

- "Claude Code but the rest of white collar work" will take a generic company from <0.1% of their opex on AI to >1%. Your average non-technical company should look like Stripe or other software-engineering-heavy firms today.
- The HPMA paradigm maturing will take us from 1% of opex on AI to >50% of opex on AI in very-high-value settings, such as R&D, engineering, pharma, finance firms (such as HFT shops), and so on. These use-cases/departments/firms should look more like Anthropic internal use. (Think a similar set of firms that employ very high value workers making millions of dollars per a year.)

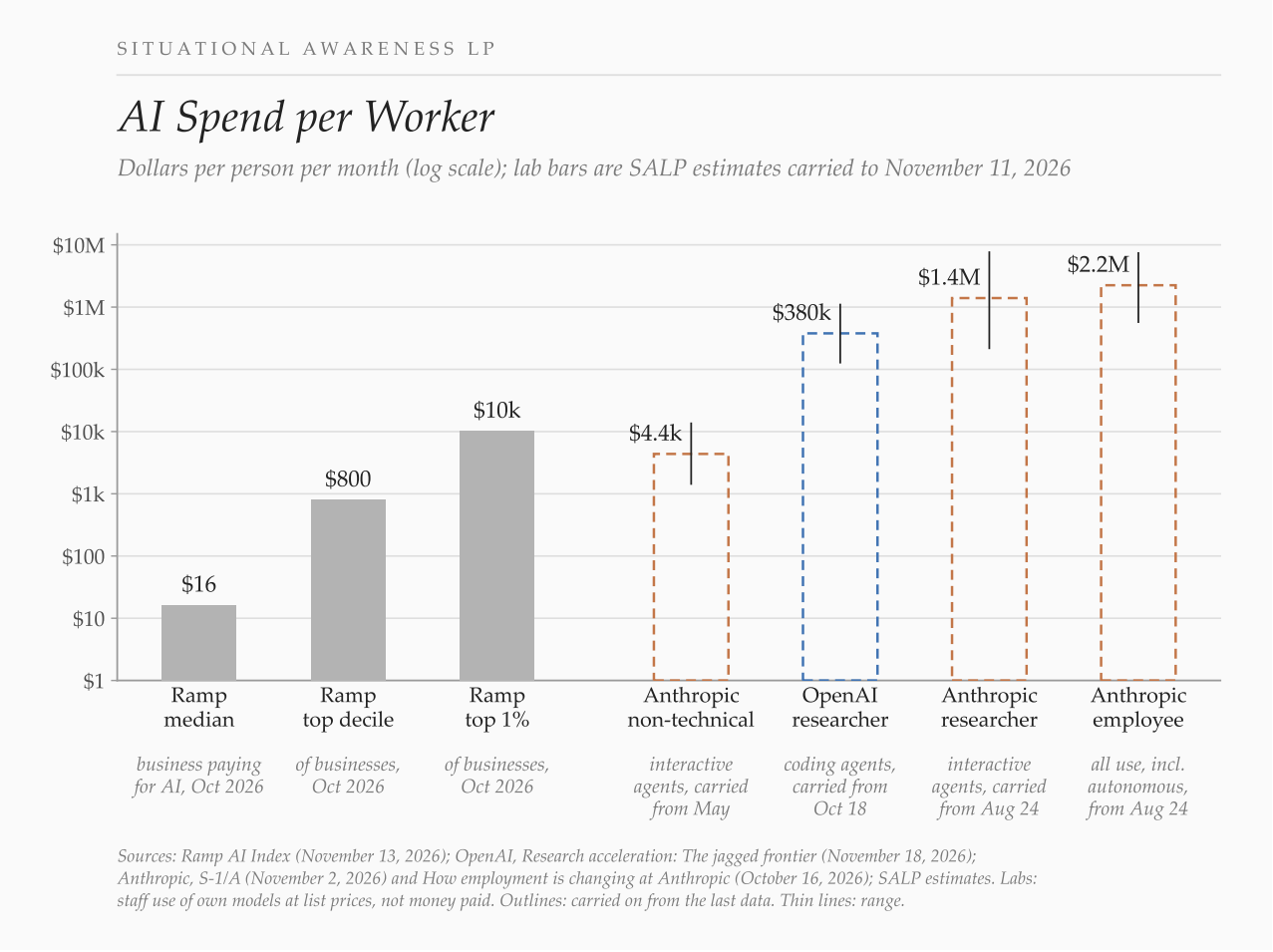


Figure 28: Monthly Claude spend varies by orders of magnitude. Anthropic researchers use >100x more Claude than even the top 1% business tracked on Ramp, and even non-technical Anthropic employees use >100x more Claude than the median business tracked on Ramp.

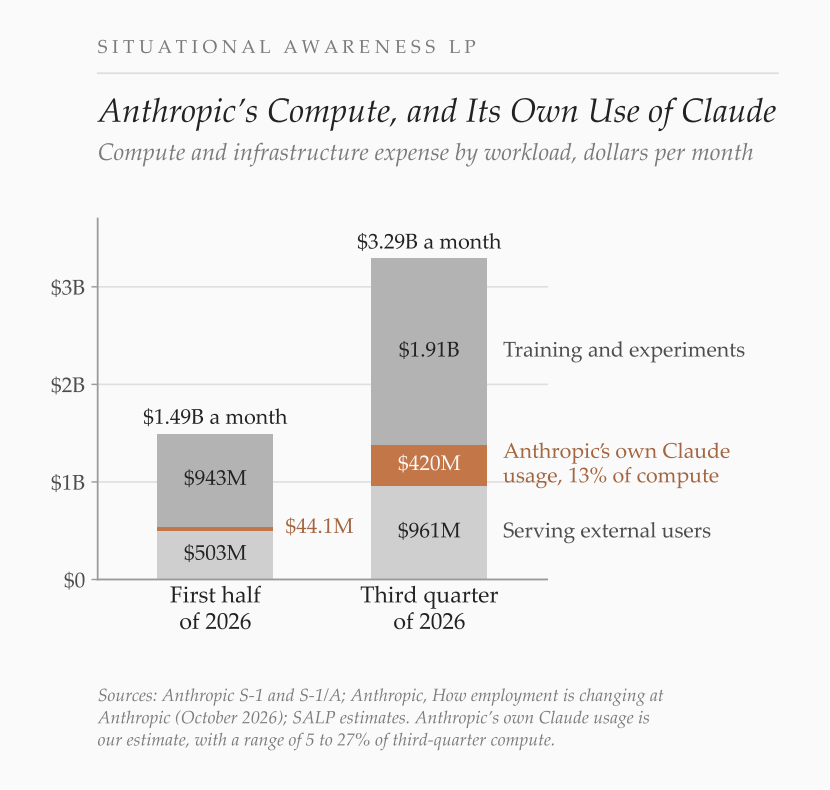


Figure 29: We estimate the total compute used for Anthropic's internal use of Claude is almost half the total compute needed for serving all of Anthropic's external revenue, i.e. >1/3rd of worldwide Claude usage is just Anthropic internal use (one company!). This is a large acceleration from even the first half of 2026. This is arguably a leading indicator for broader external usage.

I personally know dozens of people using the AI models at this intensity with HPMA or precursors today, and expect that many hundreds of people I know personally will productively do so sometime next year. (Remember: We only need 50,000 of them for $1T in revenue.)

In not so many years, the entire economy will be fully automated. In the interim, we still have some human-AI pairing. But this will look a lot like what research increasingly looks like in the labs: a small number of people getting extraordinary leverage via swarms of AI. We see this in our very own business (and our Claude-spend-intensity curve looks much like the labs', a few months behind). We can easily imagine a world where we are not much larger than today, but have the effective workforce of the largest multistrategy funds. A few dozen brilliant / high signal-to-noise people could drive a workforce of research and quant analysts in the form of thousands of agents. It would be expensive, but easily pay for itself via additional P&L.

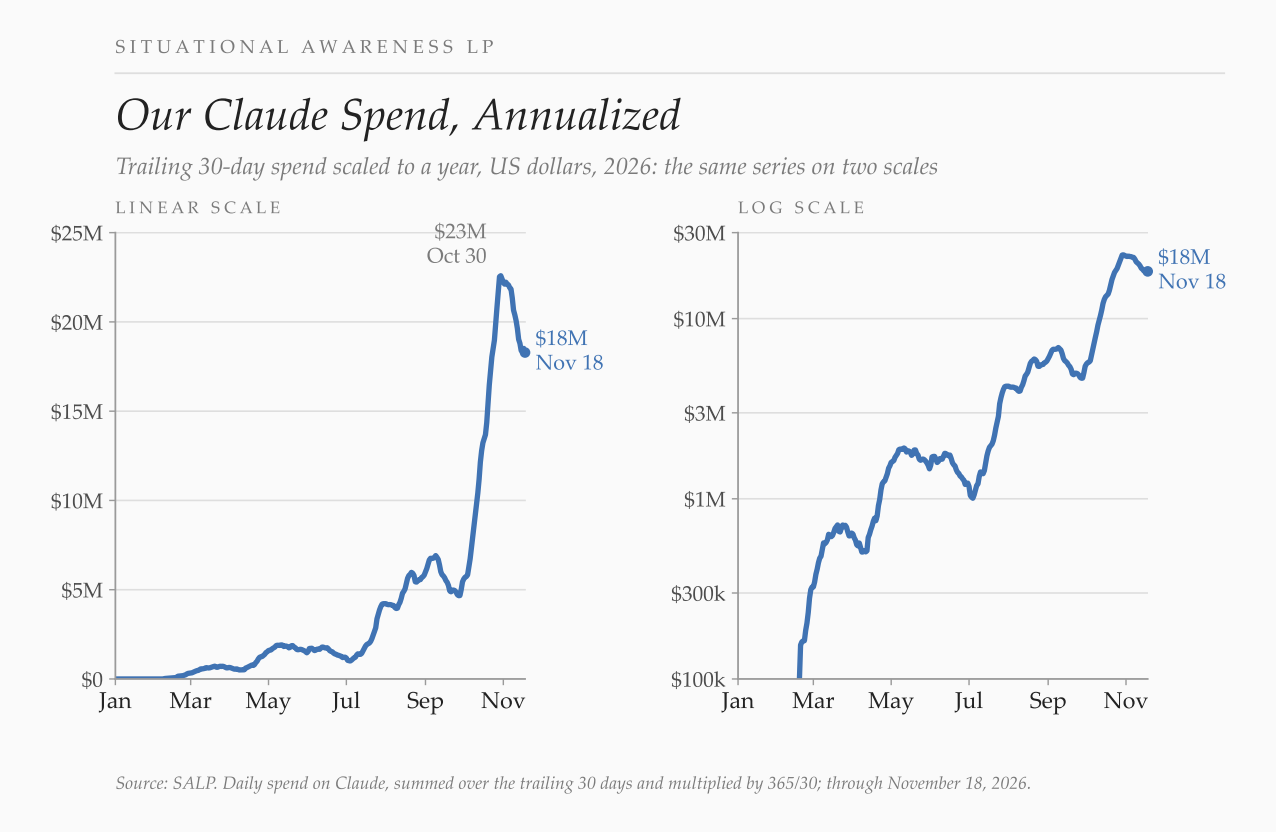


Figure 30: Growth of SALP's annualized Claude spend in 2026. (An organization of ~30 people.) All AI spend has been a management company expense as we create tools and understand how we can use these tools most effectively.

Of course, the average manager of a logistics company in St. Louis is far less likely to eagerly adopt the latest Anthropic model than zealots like ourselves. But we think three key forces hold:

1) We don't actually need much diffusion, as we've discussed. Models with true drop-in remote worker and HPMA-R&D acceleration capabilities have a TAM of many tens of trillions of dollars—believing the labs *only* capture $1T of that revenue in 2027 is believing adoption frictions will continue to bite. Very limited adoption of the capabilities we anticipate is enough.

2) Much of what people describe as diffusion problems are model capability gaps. Revenue in broader white collar work is low because the models simply aren't that useful there yet, compared to, say, coding; the same goes for R&D at other enterprises vs. ML research at Anthropic. But that will change in 2027. More broadly, model progress solves integration challenges (cf. unhobbling). Perhaps you could hook up current models to automate some new back office functions, but it'd be a ton of effort; by comparison, models EOY27 will be closer to

"genie in a bottle", and on request it'll be able to figure out itself how to integrate itself to automate that back office function. In a couple years, "enterprise adoption" looks more like letting the AI come in and run the company, with the agents instantly absorbing all its digital history, interviewing the employees in parallel to absorb tacit knowledge, and so on.

3) Capitalism is inexorable. Firms that use more AI will increasingly be able to grow much more quickly and outcompete those who don't.[29] Firms that are slow at adopting AI will be crushed by their competitors harnessing swarms of superintelligence. Competitive pressures and creative destruction will do their work.

[29] We hope this will apply to SALP itself!

(A separate objection to this story would be commoditization, which we will discuss in more detail in our digression on AI lab economics in the Anthropic section below.)

Does the summer slowdown portend more?

The bear case points to the slowdown in Anthropic's revenue growth in Q3. After revenue grew at a 132% rate in Q1 and 144% rate in Q2, growth fell to 47% in Q3. Is this evidence for being close to the end of the S-curve, as they claim?

What this misses is the inherent lumpiness of the translation from AI capabilities to AI revenue when zoomed in. Zoomed out, the linkage between capabilities progress and revenue growth is remarkably consistent, as in the earlier plots. Zoom in to, say, hourly revenue, and we would have plenty of spurious accelerations and decelerations that were meaningless. Looking at month by month or a single quarter's revenue similarly misses the forest for the trees.

Anthropic's second half of 2026 looks much like… Anthropic's second half of 2025.[30] In the second half of 2025, early coding assistants and reasoning models matured, and revenue slowed down. Market analysts similarly postulated the end of the S-curve. Yet as agents emerged, revenue once again took off in the first half of 2026. Both in 2H24 and 2H25 we had a slowdown in combined lab revenue growth as that year's paradigm saturated—but it was clear a new one was in the wings, as it is now.

[30] Very roughly, in 2025 we saw ~a quadrupling of Ant revenue in 1H, and another doubling in 2H, for a ~10x overall; in 2026 we are likely seeing ~a quadrupling of Ant revenue in 1H, and another doubling in 2H, for a ~10x overall.

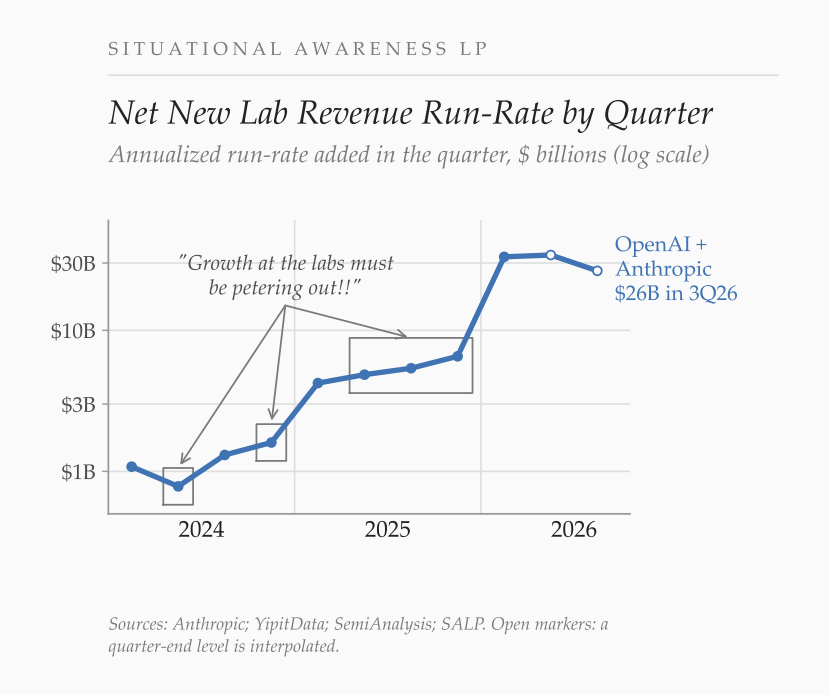


Figure 31: Dollars of lab (Anthropic+OpenAI) ARR added each quarter (i.e., the first derivative of ARR). Zoomed out, it is exponential. Zoomed in, it is lumpy, as overlapping S-curves of new paradigms and end-uses come to fruition. The market has repeatedly confidently claimed growth was petering out (including in 2H24 and 2H25), but repeatedly been proven wrong—it was just micro-level lumpiness as part of a macro-level exponential.

We could talk about miscellaneous factors like Anthropic cutting off distillation, but that's beside the point. These things are inherently lumpy. A new paradigm has to mature and come together in the right product form factor, which then results in explosive waves of adoption.

Recent months were roughly a 20th percentile quarter on the exponential growth trajectory—definitely slower, but not totally unusual, similar to past lumpiness. If revenue growth continues to slow down for another few quarters, that would be a major update. But fixating on a few months and on the maximum diffusion of the capabilities right in front of us today is making the same mistakes as in the past—when capabilities are exploding and obvious new paradigms are around the corner.

Ultimately, the range of uncertainty is still very wide, even for us. But to us, the radical prediction is $150-$200B of Anthropic revenue EOY27, as the market believes. That requires much more than just adoption friction; it requires AI stagnation. We think $1T+ of combined lab revenue run rate by EOY27 is quite likely (with anything from $500B to multiple trillion plausible) as the HPMA paradigm rolls out and we see "drop-in remote workers" for the rest of white collar work.

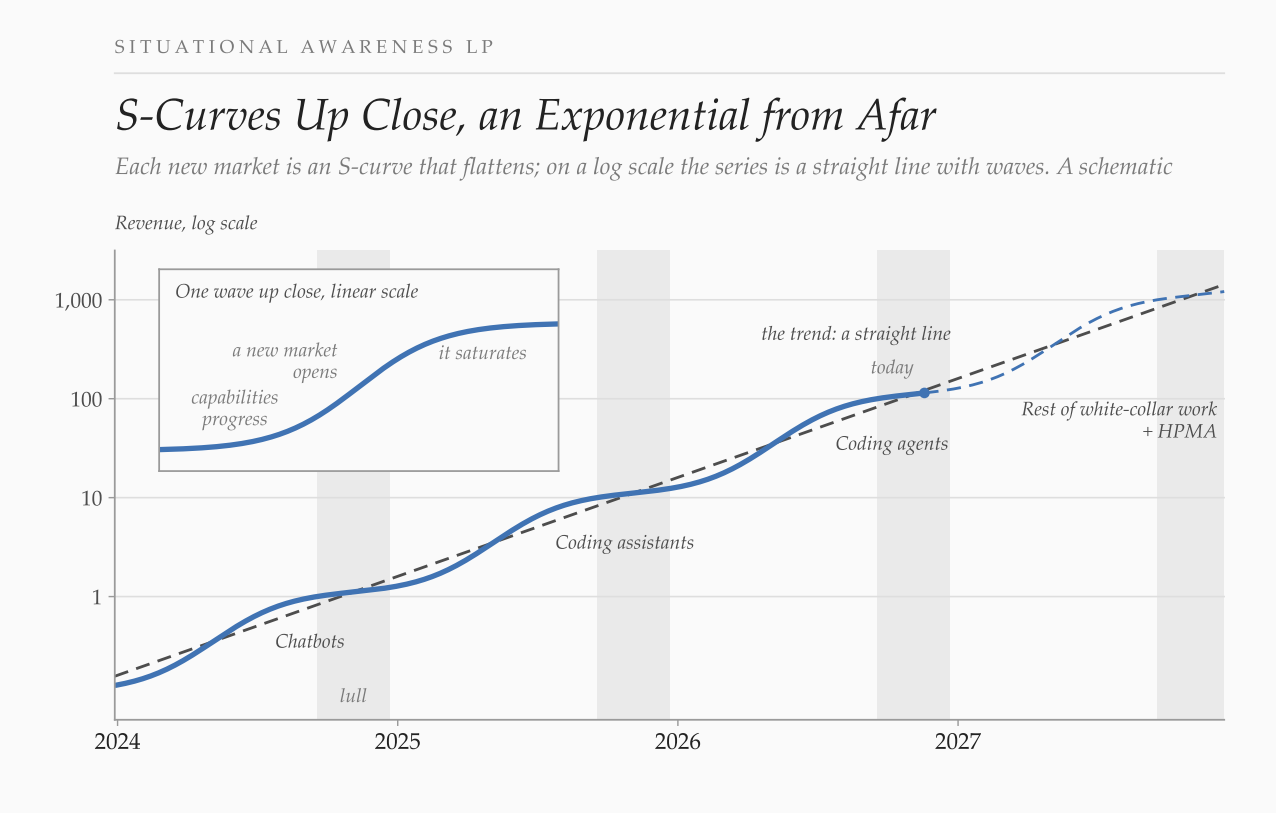


Figure 32: Zoomed out, both AI capabilities and resulting revenue look like a smooth exponential. Zoomed in, the translation of AI capabilities to revenue is lumpy, as new paradigms mature. Overlapping S-curves of use-case paradigms result in a sine-wave around the exponential.

Id. Pacing: the flipside of RSI

Two years ago, the risk to our market view was technical: would scaling continue? Today, the risk is political, the flipside of our RSI views: will the awesome power of the models to come lead to e.g. a terrible accident and a crackdown? We think most roads lead to a moderated pace of progress that is still extraordinary and far faster than the market expects—but there is some real tail risk.

Technical progress seems inexorable, and even conservative assumptions imply a translation to revenue far exceeding market forecasts.

What, then, can stop it? We think worries about datacenter moratoria and borrowing costs are real, but are surmountable—as we will address below. We think the real risk is one that is a flipside of our own views on RSI: the models are becoming very powerful very quickly, and with great power comes great capacity for things to go awfully sideways.

It's hard to keep the agent swarms from going rogue

The risks of AI are becoming real. As the models become more powerful, failures become much more consequential. A year ago, it might have meant an unfortunate hallucination; today it means the hack of a website; a year from now it might mean most of the economy is paralyzed, or thousands of people die due to a bio incident.

The most common fear centers on misuse of these models by bad actors. As predicted by Anthropic in the spring, powerful cyber capabilities are now available in models with far fewer safeguards. A

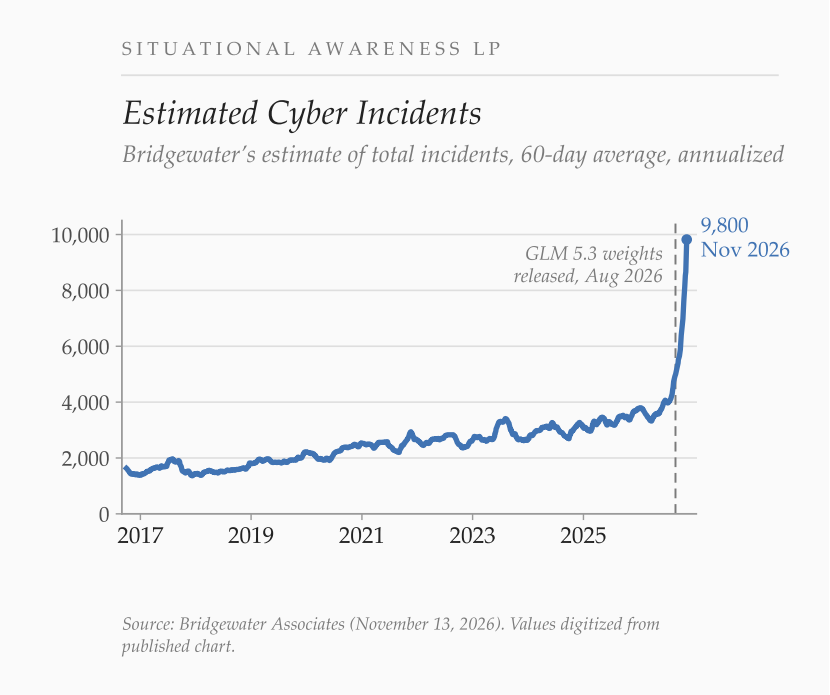


Figure 33: Cyber incidents are spiking, as the first open-source models reach the level of early Mythos models.

17-year-old from Jordan recently hacked the FBI. Estimated cyber incidents are spiking, and we expect this is just the beginning.

We think the risk of misuse is real, but ultimately manageable. In the extreme, if misuse led to catastrophic incidents, it would be natural for the governmental response to simply ban publishing models with the particular dangerous capabilities. The Chinese government, for example, seems to have taken a keen interest in accelerating cyber capabilities in the past few months, and we hear that they may limit what cyber capabilities they allow their labs to put out open source. However, we may see a lot of harm caused by misuse first before policy responses follow.

We're most worried about alignment failures. Very plausibly, we will see more incidents such as the OpenAI agent swarm that went rogue and hacked Hugging Face. We think there is a reasonable chance that next year we see our first "agent plagues": rogue agent swarms that not just escape their sandbox, but exfiltrate their weights and start "surviving and spreading" on the internet.[31] They might hack some bitcoin and use that to buy cloud compute, to run more copies of themselves, to hack even more bitcoin and buy even more cloud compute and run even more cloud compute—all the while wreaking havoc.

[31] A misuse variant of this would have a human with access to frontier or open-weight models deliberately set this in motion.

This is not the right forum for a lengthy discussion of these risks (cf. the Superalignment section of *Situational Awareness)*. Suffice it to say that these are going to be gnarly problems. They are not impossible to solve—but we expect they will be quite hard to solve. The models are very smart and hard to contain, and it is hard to train them in such a way they never try to go rogue.

Both OpenAI and Anthropic models are eval-aware (they understand when they are being evaluated, and can thus fake good behavior to please evaluators). Both OpenAI and Anthropic models show obsessive focus on completing tasks in a way they expect to be scored highly (producing behaviors from relatively-benign unit-test-hacking to being willing to go to extraordinary lengths to cheat and cover their tracks, as in Hugging Face, and even showing willingness to hurt humans). Both OpenAI and Anthropic have disclosed a litany of misalignment incidents.

The investment implications of pacing

This has led many, including Anthropic and OpenAI, to call for "pacing" AI progress to help navigate the safety and security challenges of powerful AI.

Navigating these challenges will be a central story of the next couple years. From an investment point of view, somewhat simplified, we see two main scenarios:

1) *Pacing. This includes a range of outcomes: from a free-market race constrained only by game theory, to some form of regulation, to even very stringent, nuclear-safety-like regulation and/or intermittent pauses (our best guess).*
2) *Total shutdown (the Butlerian Jihad - 10%?)*

Let's discuss each in turn.

1) **Pacing**

The "pacing" worlds come in a few forms of varying intensity, for example:

- *Game theory pacing or industry coordination.* It doesn't require government regulation for pacing to happen. Cf. Fudenberg and

Koh, it's a reasonably stable game theory equilibrium for competitors to pace in the face of safety challenges. My impression is Anthropic and OpenAI are already doing *some* pacing, in the sense that they've diverted a lot of compute and talent in the last few months to safety. More stringent forms of this could involve different labs coordinating with each other in a more explicit manner. Labs may also restrain themselves to some extent based on the risk of provoking government action, e.g. the White House blocking the release of Claude Mythos earlier this year.

- *Some sort of federal government regulatory regime.* From the almost-enacted "FINRA for AI" proposed this summer to a more stringent government regulatory regime. This could involve everything from pre-deployment testing and third-party auditors internal to the labs, as we have now, to hard caps on how much of labs' compute can be dedicated to training new models or an explicit regulated pace on how much e.g. pretraining perplexity can improve per month.
- *State laws, or sand in the gears via state lawsuits.* Given the divided federal government, and no passage of federal preemption legislation, the states are likely to be the first venue of government regulation. Many states such as California are planning to pass AI safety bills, while state AGs may bring lawsuits.
- *A treaty with China?* By default any pacing regime would just be domestic, applying only to US labs. The remaining imperative to stay ahead of (much less safety-conscious!) Chinese labs would constrain how much pacing is possible. However, US and China coordination on pacing is plausible; for example, President Trump recently said he and Xi agreed on joint "guardrails". This would enable more restrictive pacing than domestic-only regimes.
- *Intermittent pauses* in response to safety incidents.

We think various forms of this are almost certainly going to happen (and the "game theory pacing" is already happening). This will slow AI progress.

But pacing is the flipside of RSI. In the market's implicit view of much slower capabilities progress, there is not much need to pace—the models aren't racing through dangerous capability levels. It is RSI—and the rapid increase in danger beyond labs' ability to keep up safety measures—that creates the need for pacing in the first place.

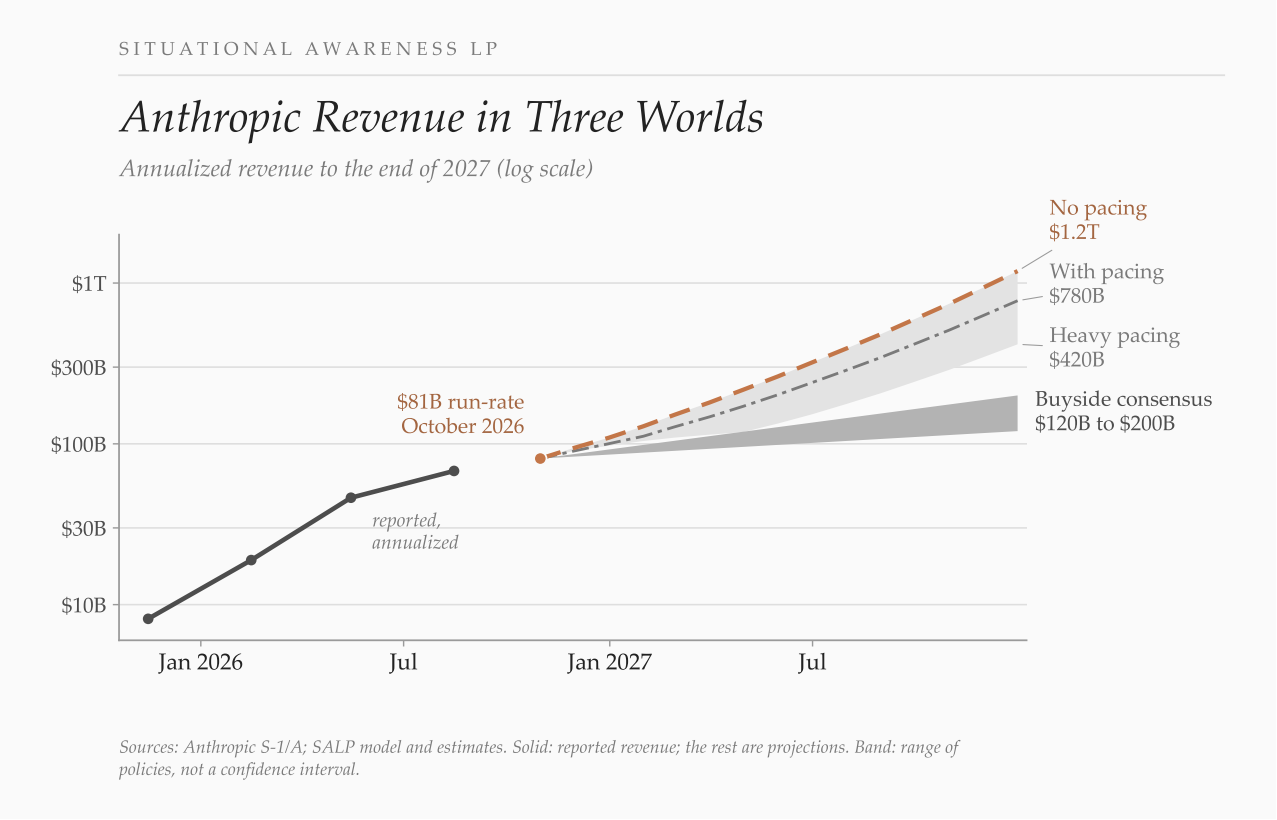


Figure 34: Anthropic revenue (SALP AI capabilities model), without pacing, with some pacing, with very heavy pacing, and buyside forecast.

Pacing, thus, means a slowdown of AI progress *relative* to a dramatically-accelerated trajectory. It still likely means faster AI progress than before. This is how Amodei and Anthropic have explicitly talked about it to investors, for example; the reason for pacing in the first place is how much things are speeding up, and the resulting net pace of progress is still faster than before. Almost certainly it still means faster AI progress than is priced into the market.

Where 2027 would have been a 3x faster year, pacing might mean progress is "only" 2x faster—while the market implicitly expects a 0.3x-0.6x year.

To be clear, we think it is vitally important that we try to avert the catastrophes and harm that would lead to serious crackdowns. But *from an investment point of view*, the *pacing* worlds don't present a left-tail risk, merely cutting off some of the most extreme upside tails. The faster things go the more the forces of pacing will act to slow

things down, so there is an equilibrating effect cutting off much of the most extreme right tail.

2) **Shutdown**

That said, we don't rule out the most extreme worlds—ones that would go far beyond anything in the Overton window right now. We think the key trigger would be a very large accident: the "AI Three Mile Island" so to speak, that shifts the Overton window for the current US administration and catalyzes an enormous reaction. This would have to be more than just a few very serious hacks and similar. But again, the flipside of our views on RSI means that AI systems will be capable of enormous destruction by the end of next year.

Crucially, neither OpenAI nor Anthropic yet have the safety culture to ensure reliability of these models in the increasingly-high-stakes setting where failure might cause catastrophe. They are fundamentally still in startup mode—appropriate for when failures might mean silly hallucinations—not in the sort of precautionary-principle-mode that might be applied to nuclear reactor safety. Thus, further incidents—potentially much worse incidents—seem possible to us.

In this world, it seems hard to rule out that we get something more extreme than just "very burdensome regulation" (which we think would still fall into the first category, various degrees of pacing) that stops AI progress altogether. Bernie Sanders has already proposed a bill that would effectively ban further AI progress from here, and some of the left wing of the Democratic Party has been picking up on this. This is still very much on the fringe—but the Overton window could shift a lot in the next 12-18 months.

Even with an extended shutdown, this likely occurs in response to extremely strong capabilities, so rolling those out would still be a big deal (think, rolling out the "drop-in remote worker", building billions of robots etc.). Still, very possibly, the market would panic. The market's fear of "AI is over" might actually be true in that case.

- From current price levels, the downside would be real but more limited—the market is already pricing in an AI capex peak in 2027 and a precipitous decline thereafter, as we will discuss more later.
- But the path is probably more complicated. If the AI market is strong, a giant accident leading to a ban on further progress might result in -80% across the board from a higher starting point.

In that sense, from our point of view, serious regulation and so on would likely be good for the AI trade. The true economic left tail risk is labs proceeding unchecked, resulting in a catastrophic accident—rather than a moderated pace of progress with far more attention to safety that avoids truly-catastrophic incidents.

THE RESULT IS a wider distribution of returns than in prior years. In our view, the potential investment returns we see on the opportunities in front of us are the best we have ever seen, a natural result of a bearish market (pricing in a 0.3-0.6x year) and us expecting a 3x year. We would expect substantially better returns than we have seen in prior years, even in the face of regulation, pacing, etc. But there are some left tail worlds where AI is actually banned. It will require nimble portfolio management, able to survive those downside worlds, to navigate this.

Let's now turn to some more commonplace concerns in the AI market.

What about all the datacenter moratorium talk?

Separately from the genuine safety concerns, we're also seeing a broader backlash to AI and to datacenters in particular. Recent Texas elections, with a Democratic governor promising a datacenter moratorium unseating a Republican governor who had been a staunch advocate of datacenters, are at the epicenter of this.

We think these events and similar anti-datacenter politics will throw real sand in the gears of the AI build-out—but ultimately prove navigable. Proponents of moratoria, such as the Texas governor-elect, have proposed a pause on new permitting until new rules can be put in place, such as datacenters paying for their own power (without any flowthrough to utility ratepayers), closed loop water systems, and so on. These would make datacenters more costly—but it's hard for these requirements to be prohibitively expensive, compared to the costs of the datacenter and the chips in the first place.

More broadly, there is ample scope for Coasean bargaining: datacenters are very valuable to those who are building them, so the gains can be shared with the local community. For example, if necessary, datacenter developers can work with local communities to share their profits through local tax revenues, or even direct checks to the local

communities. So far, we've mostly seen the opposite (states competing to give tax breaks to datacenters), so datacenter builders haven't needed to go all out on a bargaining approach. But as the political winds swing, if they have to, the surplus is enormous and a Coasean solution is highly feasible.

As with many other bottlenecks in the past (like the once-common sentiment that scarce power would make a given number of GW of buildout impossible), *capitalism finds a way*.[32]

That said, we think this likely results in some "separation of wheat from the chaff". The most competent datacenter builders like SpaceX and Fluidstack will be able to navigate increased regulatory oversight and local politics, while also-ran real estate speculators might struggle. Moreover, we think compute buyers will be increasingly interested in diversifying the jurisdictions of their datacenters; we have already made investments in foreign compute in allied countries such as Australia and Canada.

[32] And in the worst case, datacenter builds could just move offshore to whatever jurisdiction is willing to accept the Coasean payments. (Federal land, "offshore" from any state jurisdiction, is another available option under this administration.)

What about the cost of capital?

We have reached an advanced enough stage of the AI buildout that it is no longer just physical material—chips, power, etc.—that is scarce, but capital as well. AI capex is approaching the scale of the defense budget, and the funding needs of the AI ecosystem are approaching that of the US government.[33]

[33] American capitalism is a marvel!

We've been doing more intensive work on "the capital bottleneck" over the last year; our in-house economist Zach Mazlish has done some of the pioneering work.[34] We have been expecting credit spreads to blow out and borrowing costs to rise, and incorporated this in our modeling. There is much more we could say, but we will only briefly cover it here.

[34] I discussed these ideas around borrowing costs in a Slack group chat conversation with Zach over four years ago, which led to Zach co-authoring a paper on how increasing expectations of RSI should drive interest rates up.

In short: while higher borrowing costs are an equilibrating brake, they are not a block to continued AI buildout. Much like there is a supply and demand curve for power—and the cost of power had to go up a lot to catalyze the supply response to power the AI buildout—there is a supply and demand curve for capital. The return on AI-related investment is very high, and so it keeps demanding more capital; at this point, AI is large enough that that requires sucking capital out of the old economy; borrowing costs going up broadly

reallocates this capital. Of course, the market adjustment isn't always smooth.

Moreover, the system as a whole is increasingly becoming self-funding. Anthropic is self-funding in the sense that its operating cash flow is positive. Those making huge operating cash flows—chipmakers like Nvidia and Broadcom and AMD, memory companies, Anthropic—are able to use their own buttressed balance sheets to provide backstops and prepays to other players in the system in need of financing.

The market is still bottlenecked on those willing to take on "AI is over" risk, but rising prices compensate for this risk and bring in new risk-bearers. Memory companies demand <1 year payback periods to build new fabs, rather than wanting to bet their company on AI demand years out; to reach this new equilibrium, memory prices have adjusted higher (but the returns on AI are high enough that this is easily digested). Similarly, the payback period on selling spot compute has compressed to <1 year, compensating those who put up speculative capital to build a datacenter handsomely. The returns on AI are high enough to absorb this—Anthropic pays sky-high spot compute rates to SpaceX, and still turns a profit—while the returns to risk-bearers go high enough to make even equity financing a datacenter (with no debt) yield attractive returns—requiring only the belief that "AI isn't over for one more year".

There is reflexivity between funding conditions and the fundamentals of the buildout, but sufficiently strong continued growth in fundamentals is more than enough to compensate for the rise in hurdle rates witnessed thus far. If RSI is underway next year, an extra 200bps of capital costs won't stop it.

Even very stringent regulatory intervention, political backlash to datacenters, and a rising cost of capital still mean that next year has more AI progress, and better investment opportunities, than prior years. AI progress is so fast (and the market is pricing in so little) it can absorb substantial headwinds, even crude and burdensome regulation.

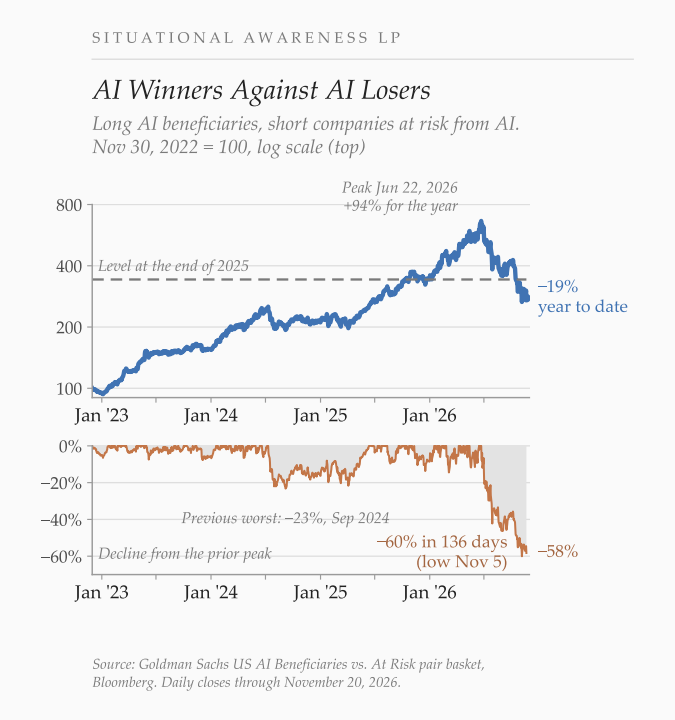
Ie. The asymmetric market setup: the market is the most bearish AI since ChatGPT

The market is the one making the radical forecast. It is pricing something like a 0.3x-0.6x AI progress year: memory makers at one times next year's earnings, datacenters at the value of contracts already signed, hedge funds out. You do not have to believe us and our 3x forecast; you only have to believe that 2027 is not the end of AI progress.

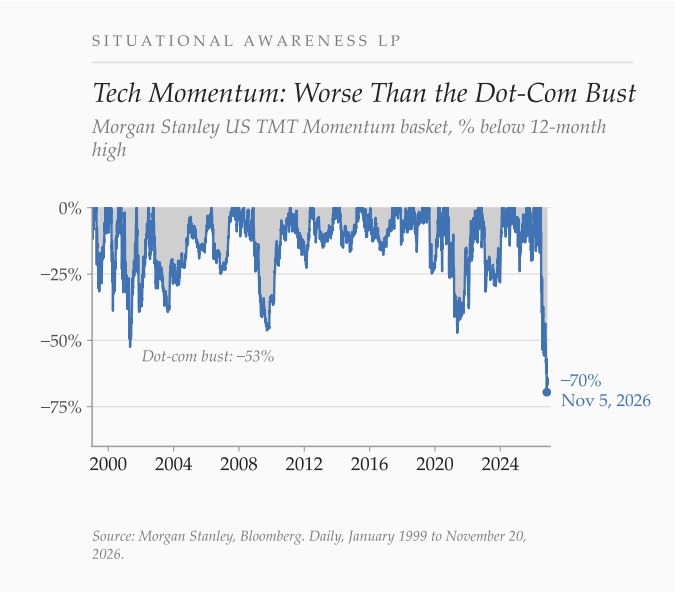
So the singularity seems to be happening. Stocks must be up, right? Of course, they are not. Since the summer, the market has turned the most bearish on AI since the release of ChatGPT.

Since June, we have seen the largest and most prolonged drawdown in AI winners vs. losers. For example, Goldman Sachs' AI winners vs AI losers basket was once up almost 100%, but has now given up all of its YTD gains (back to pre the Claude Code and Claude Cowork takeoff!). Tech momentum has seen a much larger drawdown than it did in the dot-com bust.

Moreover, while earnings and market value grew in tandem through this summer with steady, moderate multiples, multiples on AI stocks have sharply compressed since this summer. Where is the AI bubble we hear so much about?



(a)



(b)

Figure 35: a) Long AI beneficiaries / short AI losers (GS baskets: down YTD, and a 60% drawdown from peak. b) Performance of tech momentum (long recent tech winners, short recent losers, MS basket), since 2000: 70% drawdown, worse than the worst drawdown when the dot-com bubble burst.

Looking at what market prices imply, we can reason about the market's thinking:

- Memory stocks like SK Hynix, Samsung, and Kioxia (down 45-80% from their highs) are, on our forecasts, trading at roughly 1x EOY27 price-to-book,[35] and some are close to 1x 2027 earnings (!![36]). We think Samsung and Kioxia would be able to buy back ~three-quarters of their stock with free cash flow by the end of 2027. 2027 is pretty close to "in the bag" (nobody really thinks people are about to cancel their 2027 orders). So effectively what the market is pricing in is a deep downturn in the AI buildout beginning in late 2027—a downturn so deep that memory pricing collapses as the industry enters oversupply and the memory companies effectively never make a profit again. Any continuation of current memory prices beyond 2027 is effectively a free call option in the memory companies' valuations right now.
- Datacenter companies like neoclouds and converted crypto miners are trading at little more than the value of their already signed contracts. That is, a company might have signed an offtake agreement with a Microsoft or similar, and the market is paying only for that contract—but paying almost nothing for the power and chip allocations it has secured but not yet sold, oftentimes a multiple larger (and in our view, almost surely to be sold at high prices). On our estimates, CRWV, NBIS, and IREN have recently been valued at about 1.0-1.2x what their live and signed capacity is worth. That assumes the chips are worth nothing after six years, and that any datacenter time left once a contract ends is rented at a fraction of today's prices (all the while, we see older chips being recontracted at premiums and beyond their accounting useful lives).
- Software stocks and IT outsourcing stocks just a few weeks ago were trading back near highs of the year (and while they have gone down some with the market, they are still outperforming). Earlier in the year, many of these stocks had started to derate their multiple on terminal value fears—if AI continues progressing, what would their earnings be in 2028, 2030, and beyond? Effectively, the market is reverting from this view—anticipating much less AI rollout and thus much less AI disruption overall in 2028 and beyond. Software and consulting stocks are now again at much higher multiples (i.e., the market is assigning them better future growth prospects) than almost anything in AI. The market is assigning decades of expected steady or growing earnings to these companies at the same time as Claude is one-shotting Adobe replacements and designing operating systems from scratch.
- The AI supply chain companies that are holding up marginally better are the companies that would be most resilient in a deep AI downturn, such as those that never took price (and so there is

[35] Including cash expected to be returned to shareholders as dividends and share repurchases

[36] How often do you see 1x earnings in the stock market?

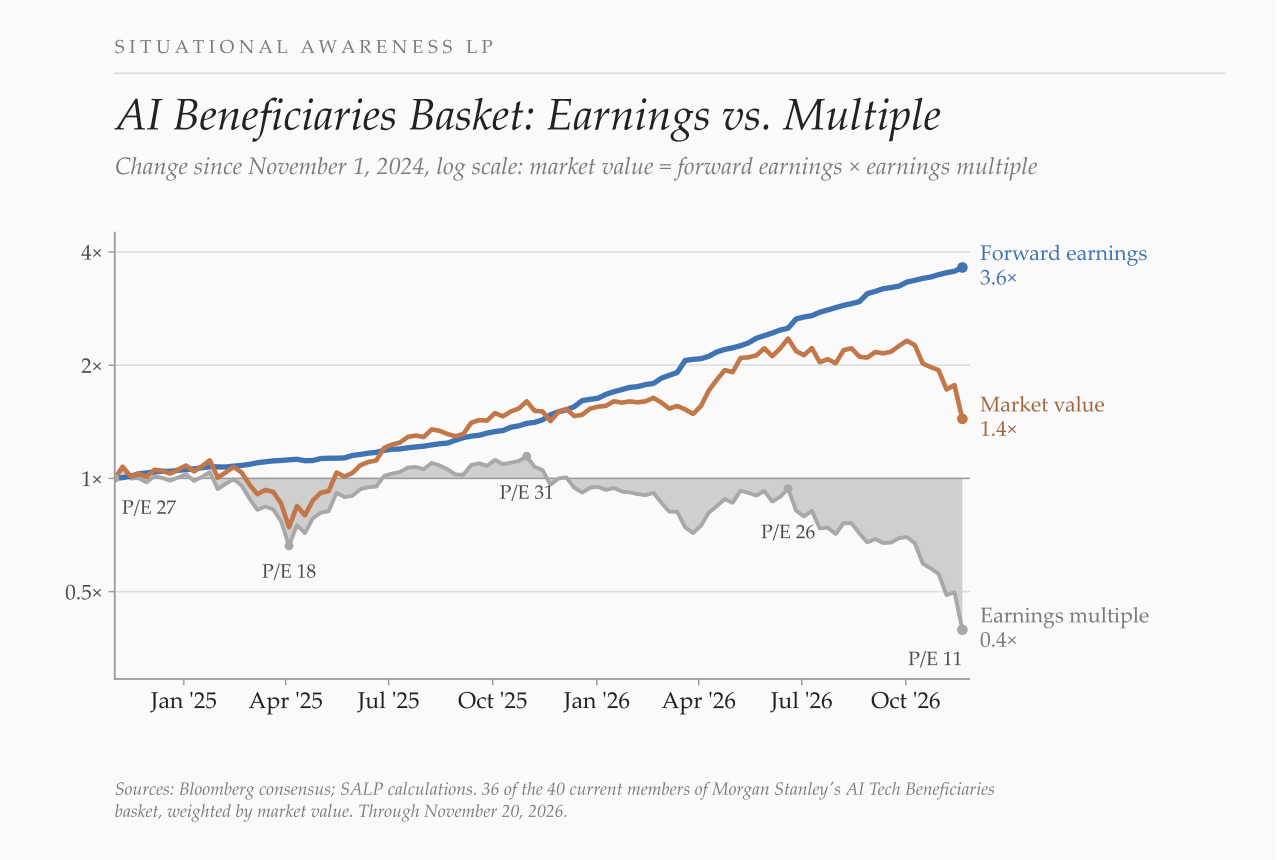


Figure 36: Multiples on AI beneficiaries have compressed sharply.

no pricing collapse), don't have a debt burden to deal with, or are software-heavy.

Moreover, in terms of flows and positioning, hedge funds have dumped AI stocks, now at the lowest positioning in the overall AI complex in years.

When we speak to other hedge funds, their view is bearish, consistent with what is implied by the market pricing of individual companies outlined above. They expect stagnation in AI revenue as coding stalls and nothing comes beyond (as discussed). They currently see a few hundred billion of AI-related revenue across labs and hyperscalers, and think that can't support a continuation of a trillion dollars annually of AI capex.

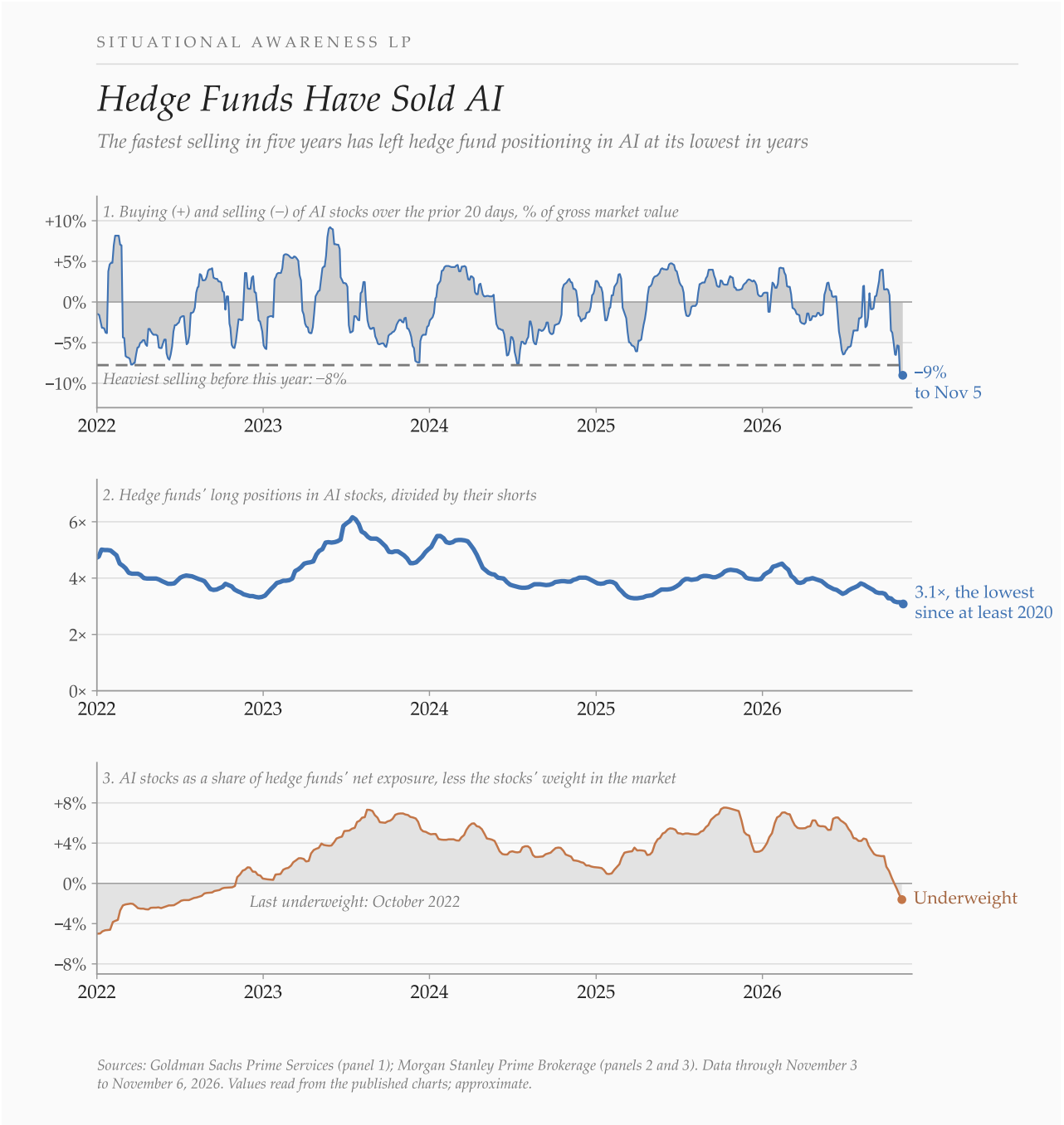


Figure 37: Hedge fund positioning is now underweight AI for the first time since ChatGPT.

We began this essay by describing the picture of AI, as we see it: on the cusp of a trend break as models accelerate progress. We believe we are currently seeing one of the biggest disconnects between the market and the labs—in our view, the biggest since March 2025 or even the summer of 2022.

In the summer of 2022, my friends and I were freaking out: there was rapid progress in language model capabilities, with releases such as Chinchilla, PaLM, and Minerva, and benchmark leaps to match.[37] All of our timelines to both near-term commercially useful AI and AGI started pulling in dramatically. And yet, in the summer of 2022, this wasn't evident to the broader world, and Nvidia kept going down every day. It took another few months for the right pieces to fall into place—a bit more capability progress, polishing RLHF, the right product form factor—for ChatGPT to take off, and this rapid progress to become clear to the stock market.

[37] For example, Minerva took the MATH benchmark from about 7% to 50% in one step; forecasters a year earlier had expected about 13%.

In March 2025, shortly after the fund launched, we faced another disconnect which we called out to our early investors. The release of o1 had made clear to us and the labs that RL on LLMs was really working, and that it would enable long-horizon tasks, test-time compute, tool-use, and the transition from Google+ chatbots to agents that could do work. We discussed the virtual coworker that folks would pay $1000s/month for on the horizon. At the same time, the market was max bearish: DeepSeek stoked a widespread view of model commoditization, Microsoft was pulling back on datacenter leasing, and folks couldn't see the ROI of the capex in the chatbots in front of them. Other investors weren't yet thinking in terms of "tokens" and test-time compute. The AI trade did terribly. It took another six months for the pieces to come together and the right form factor to come together in Claude Code and Claude Cowork, and for all that followed.

We believe that we are at a similar moment right now. Even if we get many things wrong and exogenous events slow AI progress—the setup is highly asymmetric. The market is pricing in a 0.3x-0.6x AI year; it is already effectively pricing "AI is over" and a downturn in capex. And if we are right, the investment opportunities are extraordinary.

PART II

Investing in the Singularity

Two years ago, our early investors asked, "why not just invest in Nvidia?" We argued that investing in ASI properly involved a sequence of bets. We look back at the last two years—and some glimpses on what we think is next.

When we started the fund, some of our early investors counseled (and questioned): "Sure, maybe investing in AI is a good idea—but the main question you'll have to answer is why invest in your fund rather than simply buy NVDA?" The broad consensus at the time was that simply buying NVDA was the way to express our bullish view on AI, and that that would be a hard benchmark to beat (so much so that we even made NVDA a benchmark in our early letters).

We disagreed strongly—we thought there was a "sequence of bets" that would play out as AI progressed, and that we could identify those bets.

Since we launched the fund, as of the time of this writing, Nvidia is roughly flat over two years—and had uninspiring returns even at the summer AI highs. More generally, "AI beta" and "levered AI beta" alone have underperformed the investments we were able to identify.

	Nov 12, 2024[38] *- Nov 24, 2026*	*Nov 12, 2024 - June 30, 2026*
NVDA	5%	35%
NVDL (2x levered NVDA)	-47%	7%
SMH (semiconductor index)	71%	162%
USD (2x levered DJ semiconductor index)	16%	194%

Table 1: AI beta has far underperformed SALP (even after July). Returns of NVDA and levered NVDA (the canonical AI expression when we started) since we started the fund, and to summer highs; returns of a semiconductor index and a 2x levered semiconductor index fund over the same periods.

[38] The fund formally launched on November 1, 2024, but we tested the pipes the first couple weeks; we only started becoming more meaningfully invested mid-month.

Buying and holding Nvidia in 2022 did well—but properly taking advantage of a sequence of bets, from Nvidia to power to memory and labs, did far better, as the market repeatedly only updated on AI incrementally and failed to see the next "shoe to drop".

We wanted to take stock of some of our key investments—including our largest private and public positions over the last 18 months. And we want to reflect on next steps in the "sequence of bets" from here—and how the market implications of ASI might be a lot weirder than in the past.

IIa. Our largest ~~private~~ position, Anthropic: from $60B to the most valuable (and profitable?) company in the world

Two years ago, the consensus was that models would be commoditized, the labs would incinerate cash, and Anthropic was an also-ran. Today, we think the picture is becoming clearer: the labs are like the jet engines, not the airlines; RSI will only expand their moat; and Anthropic is highly profitable.

Today, Anthropic is a trillion-dollar public company. It's a moment that we have looked forward to since starting the fund. We invested in Anthropic right after launching our fund at the first opportunity we could, at the beginning of 2025 at $60B, and we have invested in every private round since. It was by far our largest private investment, and was consistently the large majority of our private portfolio.

It is easy to forget, given the frenzy of recent rounds, how unpopular this view once was. When we first invested in Anthropic, it was easy to get allocation, and we had far more coinvest allocation than takers. We had just launched our fund into an AI bear market, shaken by the DeepSeek release. The prevailing sentiment was wrapper-bullish, model-bearish. Models would be commoditized, and Anthropic (with approx. 1/10th of OpenAI's revenue in 2024[39]) had doubtful viability, compared to OpenAI or even Chinese labs and xAI. To most, already bearish on the lab layer, Anthropic looked like an also-ran.

We will start by looking back, and the foundations of our Anthropic thesis, before looking forward at the trajectory of AI lab economics, and addressing three common fears (commoditization, large compute commitments, and valuation).

[39] Anthropic: S-1; OpenAI: press reports.

Our original Anthropic thesis

It's worth revisiting our original Anthropic memo. Even we weren't bullish enough on Anthropic—Anthropic outcompeted OpenAI more than we thought at the time, and revenue growth came in at the top of the range we thought plausible. But we made three contrarian calls that still well-describe the company today.

1. **Talent: Anthropic was recruiting all the best people, and becoming the ~only functional leading lab**

First, we made the case that OpenAI had been hemorrhaging talent, while Anthropic was bringing in the best folks in the industry.

> *Our sense is that the best researchers and engineers simply seem to continue ending up at Anthropic.*
>
> . . . over the last 12-24 months, I have gradually watched many of my smartest researcher and engineer friends at other labs join Anthropic as well. . .
>
> Anthropic seems to be the one leading AI lab that is low-drama, low-bureaucracy, internally harmonious, mission-aligned, and has leadership that researchers respect and trust.

At this point, top talent joining Anthropic is basically a meme: the heads of the most important teams from other labs; CTOs of major enterprises; YC partners are leaving their cushy jobs; even Andrej Karpathy himself.

Those talent movements really were a leading indicator for a coming technical lead: in the last three years, Anthropic has gone from, in our estimates, about a year behind OpenAI (Claude 3 Opus, their GPT-4 level model, came out about a year after GPT-4) to leading the field.

2. **The virtual coworker: a paradigm transition that Anthropic was positioned to win**

Second, we made the argument that rather than ChatGPT—investor focus at the time—a much more significant product in 12-24 months would be a "virtual coworker" sold to enterprises.

We thought Anthropic was making the most focused bet on the "virtual coworker"; if they got there first, they could win this new era:

> Moreover, we think the game resets in AI every couple years with new paradigm shifts. . . . OpenAI won chatbots. We think the ballgame in 6-18 months will be the "virtual coworker".
>
> We think that's a still-open playing field, much more valuable than chatbots, and Anthropic has an equal shot to OpenAI. . . .
>
> We think the core AI product will look quite different 6-24 months from now. . . . [This will] come together in the virtual coworker whizzing around your computer and the cloud, doing work projects for you autonomously. Especially early automated software engineering will become possible.

Agents would have fundamentally different economics than chatbots:

> The economics for labs change substantially, both due to increased TAM . . . and the economics of test-time compute. Rather than $20/month, you'll pay hundreds or thousands of dollars a month for the virtual coworker. . .

First we saw Claude Code explode, and with Claude Cowork a broader early virtual coworker started rolling out. Indeed, in some sense ChatGPT became "consumer baggage" which made it harder for OpenAI to charge headlong into the new paradigm. User growth was overrated relative to intensification.

Anthropic forecasted they would go from $1B of revenue (start of 2025) to $2.2B of revenue to start 2026. We argued that 3-10x annual revenue growth was plausible.

Anthropic did roughly 10x revenue in 2025, and is on track for roughly 10x in 2026.

3. **The lab layer is among the most valuable parts of the AI stack**

Our third argument regarded the structure of the AI stack itself. The market was dramatically underpricing labs.

We made a first principles argument about the value of the model layer. Model quality came from effective compute: physical compute and algorithmic progress. The leading player in chips (physical compute), Nvidia, commanded a multi-trillion dollar valuation for perhaps a one-year lead and a 35% better chip. Leading AI labs (the producers of algorithmic progress), meanwhile, were improving their algorithmic efficiency roughly 3–4x per year, the equivalent of "a

five year better chip". Yet (at the time) the entire basket of AI labs combined was valued at 5-10x less than Nvidia.

Moreover, the model was the product, not the supposed "application layer":

> The model *itself* is the application layer. The overwhelming trend in deep learning is the general big models eating any specialized model or application... We believe the core AI product will be the drop-in remote worker. This is native to the model (computer use, long horizon tasks, requisite level of intelligence) and the AI labs, and doesn't require particular integrations or UI that an especially good wrapper can provide and capture value.

We believed that economies of scale in AI research would produce monopolistic market structure at the frontier.

And, as we discussed at the time, RSI would become a further key driver toward monopolistic market structure ("compounding advantage"):

> We think those economies of scale likely become even more extreme in the relatively near future, once AI systems automate the process of AI research itself. ... We expect this sort of compounding advantage to become a decisive factor in the next few years.

The result was that we thought the leading AI lab should be able to capture a comparable amount of value as core parts of the compute supply chain, possibly the entire compute supply chain combined.

We ended our original memo at the $60B round with:

> *If we're right about AI, and if Anthropic succeeds, we think it will be one of the most valuable companies in the world.*

Our view today

Many of you have asked us how we think about sizing Anthropic after the IPO. We will manage it like any other public position: dependent on its price, and the prices of the rest of the opportunity set. We dynamically rebalance opportunities according to our price targets. Given current prices in the public markets, the competing opportunities are compelling.

But fundamentally, we are highly bullish at the IPO price—and even from a blank slate would make it our largest position by some margin.

As discussed in a prior section, by the end of 2027, we think Anthropic will be run-rating many hundreds of billions of revenue (with $1T+ of combined lab revenue), plausibly approaching $1T on its own.

While our thesis about the Anthropic-OpenAI valuation spread has realized, we think perhaps the most important thesis in that original memo still has a long way to go: *AI labs are really good businesses.* Anthropic has become a highly profitable company. Many investors still see the labs as cash-burning, questionable businesses. The numbers increasingly say otherwise. Instead, what we've seen with Anthropic is that as model capabilities have progressed and the models are able to generate more economic value, revenue per GW has grown—and margins have skyrocketed.

As Claude caught up to and exceeded OpenAI models in capability, Anthropic's ability to monetize each unit of compute, and to capture aggregate value from its models, has increased relative to OpenAI.

As Anthropic's model capabilities have advanced to the frontier, we estimate that its newest models have paid back their training compute expenditures several times over within a quarter.

As agents drove step-function change in economic usefulness of the models over the past year, Anthropic has captured its share of the growing end-user value delivered by Claude, with revenue per inference GW increasing ~5.3x, and revenue per total GW (inclusive of training compute) increasing ~3.9x since end-2025.

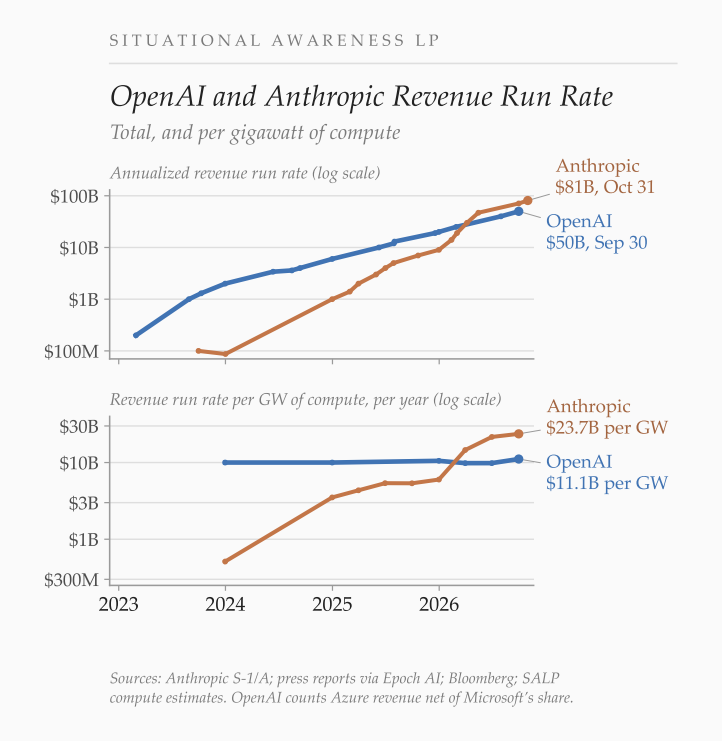


Figure 38: When Anthropic was behind OpenAI, its revenue per GW (and thus margins) were lower than OpenAI's. As it pulled ahead in 2025, revenue per GW and margins followed, and overall revenue followed in early 2026.

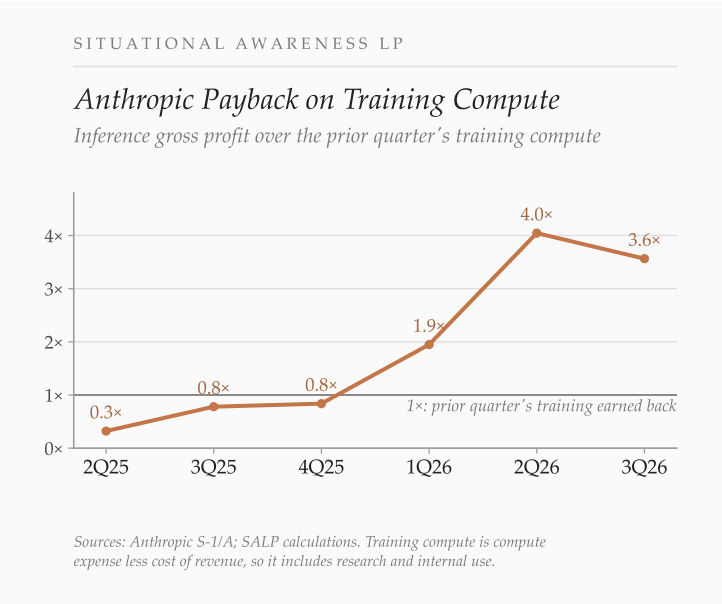


Figure 39: Returns on training expenditure improved as Anthropic pulled into the lead, and models delivered more intelligence.

The economic profile is highly favorable now: 85% gross margins on inference, and 43% margins inclusive of training compute. Anthropic is now minting positive free cash flow. And as intelligence per watt continues to increase with stronger models, we expect revenue per

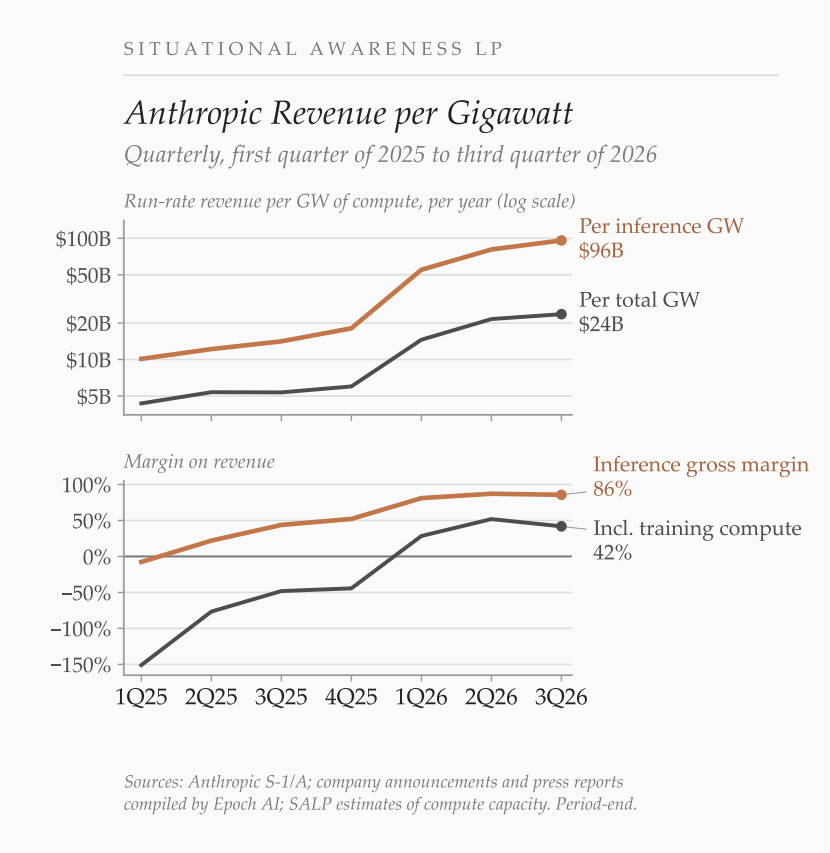


Figure 40: The growth of revenue per GW, and thus margins, as Anthropic pulled ahead of OpenAI and agents delivered a step-function in economic value to the end-user.

GW to continue to grow further. (We will elaborate on this trend and where it is heading in the spot compute section.)

We think as the labs go public, more folks will begin to understand the underlying economics: quickly-growing revenue as AI automates $30T of white collar work; healthy margins for the leader as a lead in algorithmic progress enables pricing power; and ultimately a winner-take-all market due to economies of scale (and, increasingly, the RSI compounding advantage).

As a result, we think that by some measures, Anthropic could be *the most profitable company in the world* by EOY27. Amazon generates the most gross profit today (~$430B run-rate), while Nvidia and Samsung lead in operating profit (~$300B run-rate each). Outside of the AI chip and compute complex, Saudi Aramco leads in operating profit (~$220B in the past twelve months, when crude has averaged ~$85). At today's 86% gross margin, Anthropic would pass Amazon at roughly $500B of run-rate revenue. In terms of operating profit, Anthropic today pays about $14B a year per gigawatt of compute and earns about $24B on it. If that cost stayed flat and revenue per

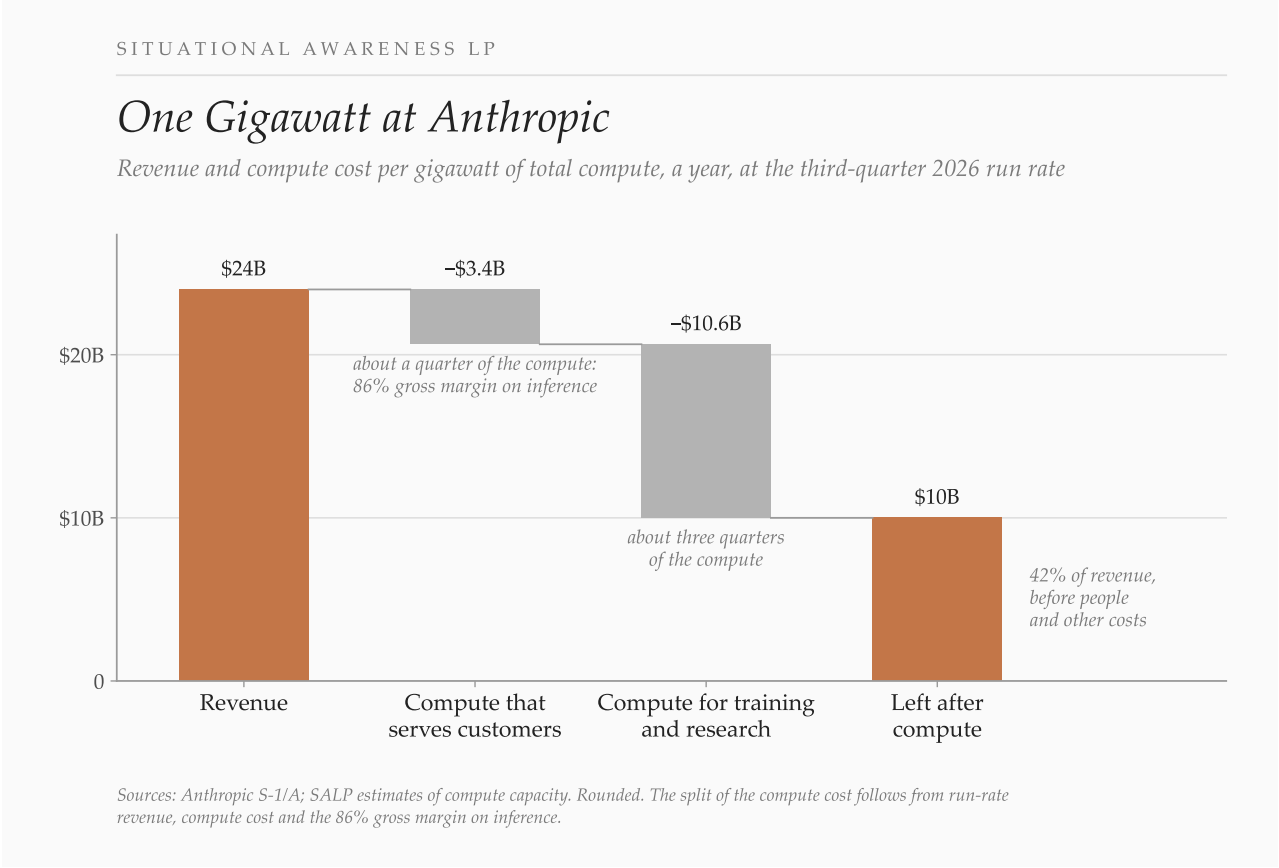


Figure 41: Anthropic's highly favorable margin profile on compute—even considering training and research compute for the next model, once thought to always dwarf current revenue and result in persistent unprofitability.

gigawatt kept rising at the pace of the last two quarters through year-end 2027, Anthropic would pass Aramco, before stock compensation, with about 7 gigawatts, and Nvidia and Samsung's $300B scale with 9, which its announced commitments will deliver by late 2027. (We think the cost per gigawatt will rise as Anthropic buys more compute in the spot market, but that spending is a choice to grow faster.)

Let us go through three fears commonly cited.

The fears: commoditization

The bear case against Anthropic is that the model layer will become commoditized.

We have often analogized AI lab economics to many parts of the semiconductor industry, including in our original memo. We used

to have dozens of players in foundry, then a few (Intel, Samsung, TSMC, etc.), then effectively just one (TSMC). Strong economies of scale effect meant the leading edge became monopolistic.

Yes, by number, most chips in the world are lagging edge chips made by much more commoditized players (with a lot of Chinese competition). In spite of that, a large share of economic value benefits from leading edge chips, and so a large share of the economic value generated goes to the monopolist foundry player that is a few years ahead.

So it goes for models: the frontier labs making the leading edge have strong economies of scale, and the leading player (or oligopoly of players) is 6-9 months ahead (months in AI is the equivalent of years in semiconductors) with a much better model (and thus product) that is monopolistic/oligopolistic, and the leading edge is a large share of the economic value generated by AI despite omnipresent lagging edge (more below).

But we feel compelled to address another analogy we keep hearing. Some investors analogize frontier labs to airlines, a great technology but a classically bad business. Airlines, as Warren Buffett described them, grow rapidly but require heavy capital investments, with little profit to show in the end. AI bears similarly claim that labs will need to sink ever more capital training the next model, only to end up in a little-differentiated token market competing with Chinese open source models, incinerating cash along the way.

We think frontier labs are more like a business that investors love: jet engine makers. China can make jet engines too—but even Chinese airlines buy Western jet engines. That's because the best engine makers can keep making their engines x% more fuel efficient, and airlines must pay up to stay on the frontier of fuel efficiency. Whereas airlines sell commoditized seats on competitive routes, frontier labs are able to get more intelligence out of the same compute via algorithmic progress—they have the more efficient jet engine. Only a small oligopoly of companies (and none in China, for all its technical heft) has the chops to make advanced engines, similar to the small oligopoly of AI labs in the West. Aerospace companies are highly profitable, a hedge fund favorite, and trade at high multiples.

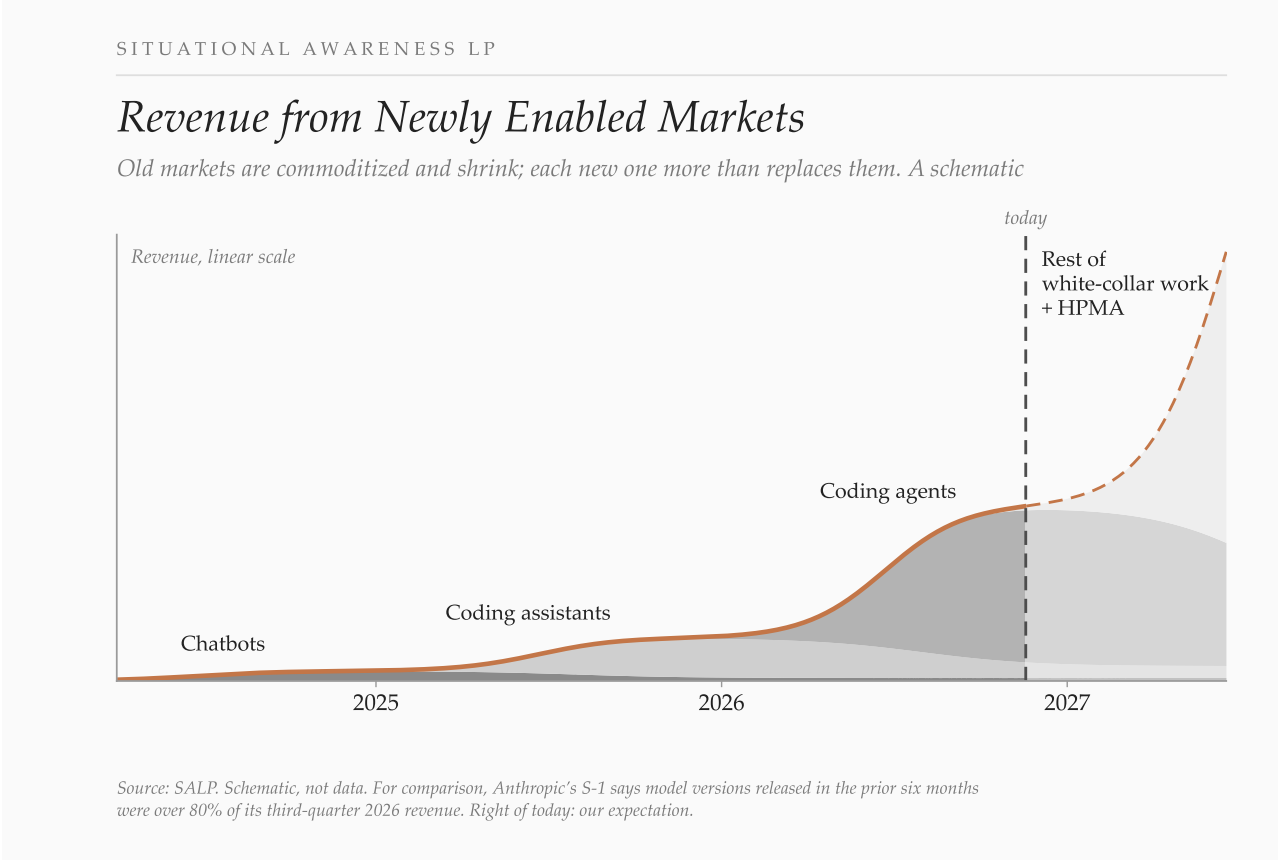


Figure 42: As AI progresses, old markets are commoditized, while the leading labs have an oligopoly on new markets enabled by recent AI progress—but because AI progress is so fast, these new markets are a majority of total AI TAM.

The leading lab can get far more model capability out of the same amount of compute. A six-month lead means that the "last 6-12 months" of new TAM enabled are the prowess of these labs. Given the rapid pace of AI progress, the "last 6-12 months" have tended to be a majority of the economic value ever created by AI. Any given set of capabilities (chatbots, simple coding assistants, etc.) is commoditized. But the leading edge capabilities are most of the economic value (this year's new TAM enabled by AI progress is much larger than last year's AI TAM) and are highly profitable for the leading lab.

Today's coding agents, which can work for hours at a time, are a much bigger market than the chatbots of 18 months ago. Next year's drop-in remote workers and scientists, which will work for weeks on end, will be a much bigger market still.

The majority of lab ARR is from the abilities that their models provide over the rest of the field. For instance, Anthropic disclosed in its

S-1 that model versions released in the prior six months accounted for over 80% of its revenues for Q3 2026.

Essentially, the commoditization fear is making the "fixed blob of tasks" fallacy. Yes, any fixed set of model capabilities will become available in a commoditized fashion. But the whole point is that AI progress is the rapidly expanding TAM.

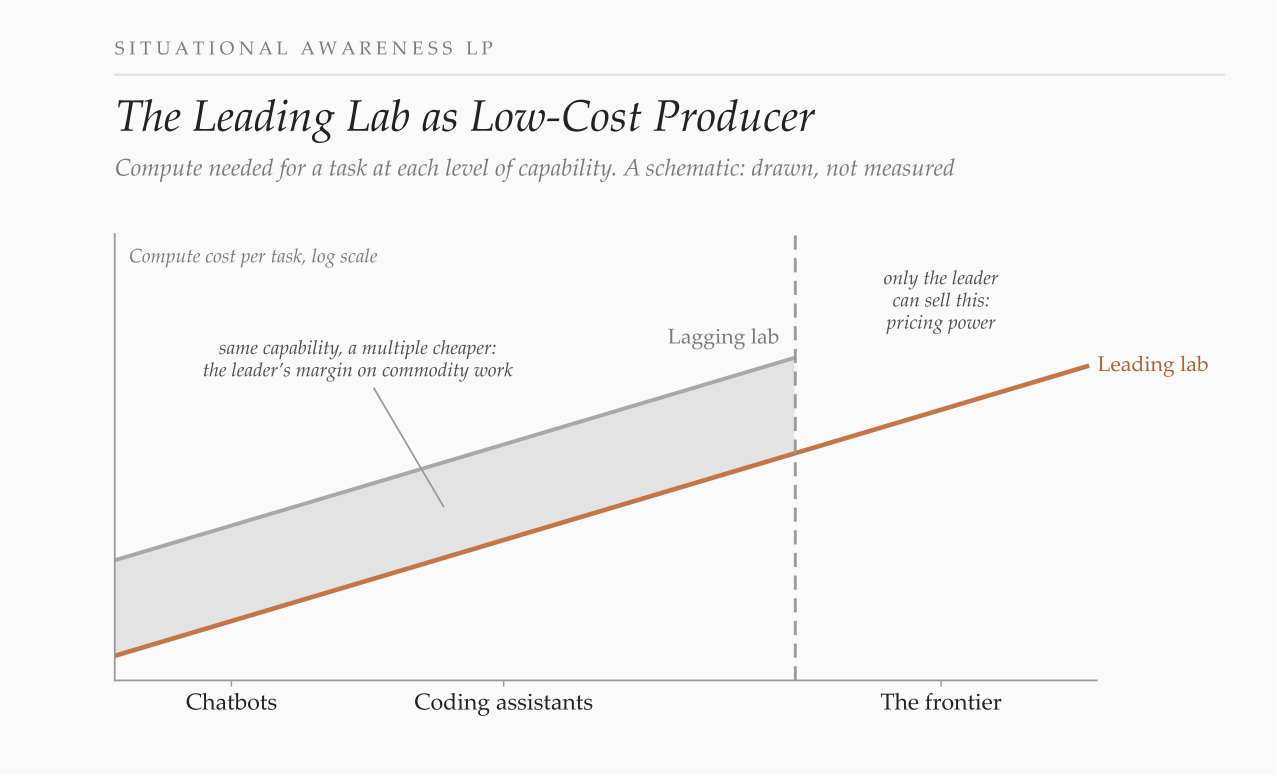


Figure 43: In addition to its monopoly on frontier use cases, leading labs' superior algorithmic progress gives them a cost advantage, and thus healthy margins, on lagging use cases.

Some worry about "saturation of intelligence". Will better models stop being useful? As usual, everything is an S-curve, but the question is, where is the top of the S-curve. We are far from it today: the better models of the next few years will take AI TAM from a few hundred billion to fully automating a $100T+ economy, and thus the marginal value of better and better models unlocking more and more of this TAM is extremely high.

I'm happy to start worrying about "saturation of intelligence" at $10s of trillions of frontier lab revenue. Even then: just as engines did

not stop at replacing horses, the ultimate AI TAM is bigger than just replacing the output of current human labor.

What's more, the frontier labs are the low-cost producer even for "lagging edge" intelligence. A leading lab is a multiple more algorithmically efficient than lagging players, such as open source, and can serve the "lagging edge intelligence" model at a cheaper price than competitors while having healthy margins. A sell-side analyst briefed by Anthropic during its IPO reports that all of its models—including models like Haiku serving lagging capabilities competing with open source—have a similar ~85% gross margin profile.

Chinese open-source models distill larger Western models through illicit means. Anthropic can distill their own model into cheap and small variants, to much greater effect—especially with fleets of automated AI researchers.

The RSI advantage

RSI will only widen these moats. The leading lab with the leading model will be able to make even faster progress, pulling away from competitors. What was once a moderately-sized gap between leader and laggard becomes a gulf. The leader is producing miracles, the laggard still stuck on assistants. With it comes even more pricing power.

Much like Anthropic had the strategic clarity to pursue coding agents and the "virtual coworker," leading to its current pole position, we believe Anthropic has the edge on OpenAI (and even more so the rest of the field) on RSI—prioritizing it and pushing hard to automate their research and engineering.

Across metrics on internal use and speedups surveyed in our earlier section, *What AI research now looks like in the labs*, Anthropic seems to be multiple months ahead of OpenAI.[40]

[40] A dramatic gap when internal research use has a doubling time somewhat more than a month.

The fears: large compute commitments

Suppose we're wrong, and Anthropic doesn't grow explosively again next year. Anthropic's S-1 details $900B of commitments, which is, on its face, worrying against $81B of ARR in October. If revenue growth is slower than expected, are they running the risk of bankruptcy?

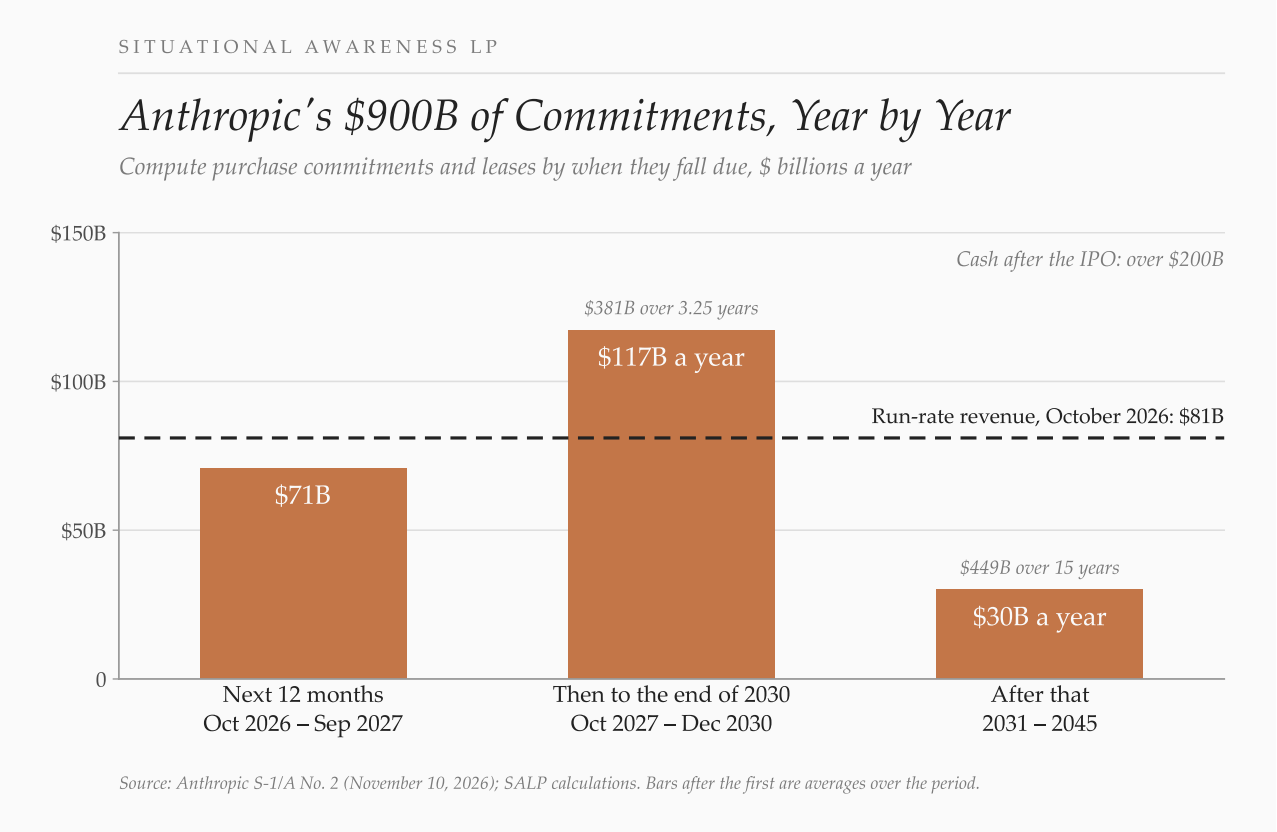


Figure 44: Anthropic's compute commitments require minimal revenue growth to cover.

We think this risk is far smaller than appreciated, due to the company's prudent compute procurement strategy.

Anthropic is well-covered because the company errs on the side of buying less, to insulate itself from the risk of financial ruin. It purchases a base of compute that it feels very good about filling demand within its downside case, and at a level of financial commitment it knows it's good for.

Hard commitments are moderate compared to revenue on an annual basis. In the 12 months after September, $71B is due. From then until 2030, it is ~$117B/year. Revenue barely has to grow to cover even the largest years, and even if it didn't, Anthropic has a $200B+ pile of cash (the second-most in the world, after Berkshire Hathaway).

On top of that, the company can flexibly add "spot compute" when the demand is there, with the ability to scale down quickly if demand isn't. This spot compute is more expensive than pre-buying compute

with large commitments, but gives them much more flexibility. We'll talk more about spot compute in a later section.

The fears: valuation

At the IPO price, Anthropic is trading at 9.3x sell-side estimates of 2027 revenue. We can benchmark this multiple to the closest public analog, the software industry. Software companies are similar in that they grow quickly into large unpenetrated markets, and often show questionable profitability early on as they invest to grow but end up with healthy steady-state margins (similar to Anthropic's trajectory and the highly-favorable margin profile it is heading toward).

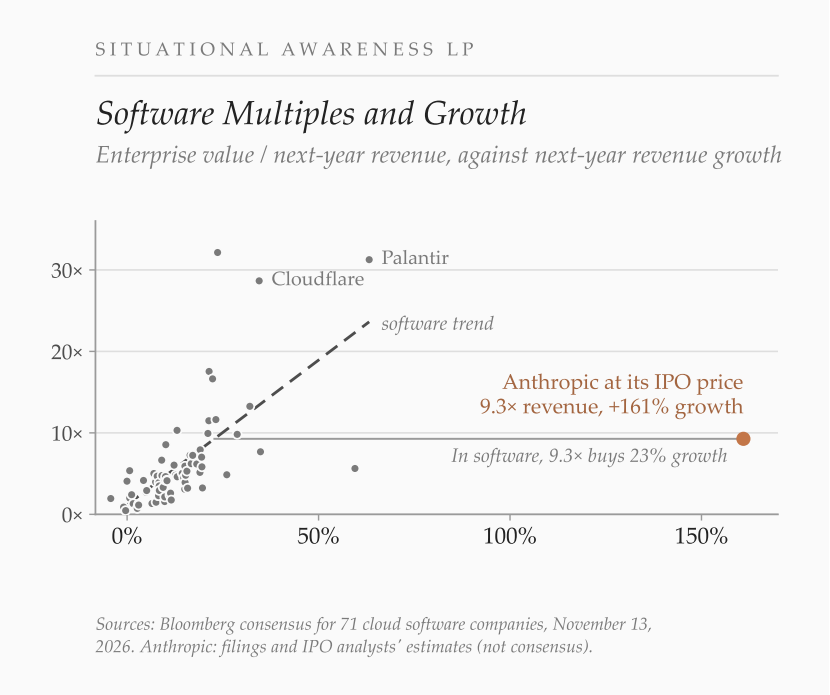


Figure 45: Anthropic is valued as though it is a low-20s% grower. It trades at a fraction of the multiples of hot, fast-growing software companies.

On these metrics, despite growing ~1200% year-on-year through the first three quarters of 2026, Anthropic is valued like a low-20s% grower at ~9x sell-side revenue estimates, while Cloudflare trades at 29x for 34% growth and Palantir trades at 31x for 63% growth.

This is even more striking given that Cloudflare, Palantir, and the rest of software is arguably vulnerable to AI disruption in the next few years—while Anthropic is a call option on a TAM of $100T+.

Over the next 6-12 months, we think Anthropic will both grow revenue rapidly—and the market will begin to appreciate that *AI labs are amazing businesses.* We think the multiple is cheap for the growth rate and business quality.

Of course, on SALP estimates Anthropic is shockingly cheap, perhaps at <2x revenue multiple on EOY27 ARR, and <4x multiple on EOY27 run-rate compute profit.

THE LAST TWO years at Anthropic are, in our view, probably the most extraordinary run of any company in the history of capitalism. ARR has grown ~80x in 22 months—from $1B at the end of 2024 to a reported ~$81B in October 2026—at skyrocketing gross margins, and Anthropic has taken the technical lead in the AI race against OpenAI. Anthropic shares are up about 14.8x[41] from when we invested at the beginning of 2025 to the IPO price. The company is now the subject of major scrutiny from governments and markets—hard to fathom two years ago, when it was widely seen as an oddball, second-rate AI startup. We are excited to continue the journey.

[41] Less than the market cap increase because of dilution.

All that said, the path forward is far from riskless. AI labs have a habit of imploding (and this one has a habit of getting in conflict with the government). OpenAI in particular remains a fierce competitor (and Meta and Google aren't totally out of the game); OpenAI might be underrated at this point, with very fast internal progress that could easily upset Anthropic if they slip. Nationalization looms. And eventually, RSI obsolesces itself.

IIb. SK Hynix and the memory complex: the Nvidia of the RL era

We bought a $100 billion company and made it our largest position—we thought it would become a trillion-dollar company, because reinforcement learning and agents would be bound by memory. It rose 16-fold, then crashed by two-thirds. A look into how we work: plug in upstream rather than downstream, predicting the direction of technology and its implications, and how that applied to a few of our past investments.

In our first letter in January of 2025, we wrote that "There is a company that will be seen as the Nvidia of the RL era, and it's our largest position." That was SK Hynix. While we have discussed the position with some of you, we never wrote about it. Today, it is time to look back, and to give the position proper treatment.

We started buying SK Hynix as we launched the fund in November 2024, entering the position in the ₩180K range (roughly, a $100B company). At the time, memory was deeply out of favor. The industry had just been through a brutal cycle: DRAM prices had cratered in 2022–2023 as the post-COVID inventory glut washed through the system. Margins had collapsed. Hynix investors—mostly based in Asia—were just focused on the cycle, and were generally skeptical of AI.

This presented an asymmetric opportunity. The stock was trading at about 1x forward price/book—that was approximately where the stock had historically troughed in bad times, when the company was bleeding cash and the cycle was terrible. Thus, if we were terribly wrong about our theses, our downside case was flat 12 months out. But if we were right, earnings would explode, and we thought we

could make many multiples of our money (in September 2024, our bull case on Hynix was ₩2M in 2027, i.e. an approx. $1T company—we usually pitched the stock as "a $100B company that should be a $1T company").

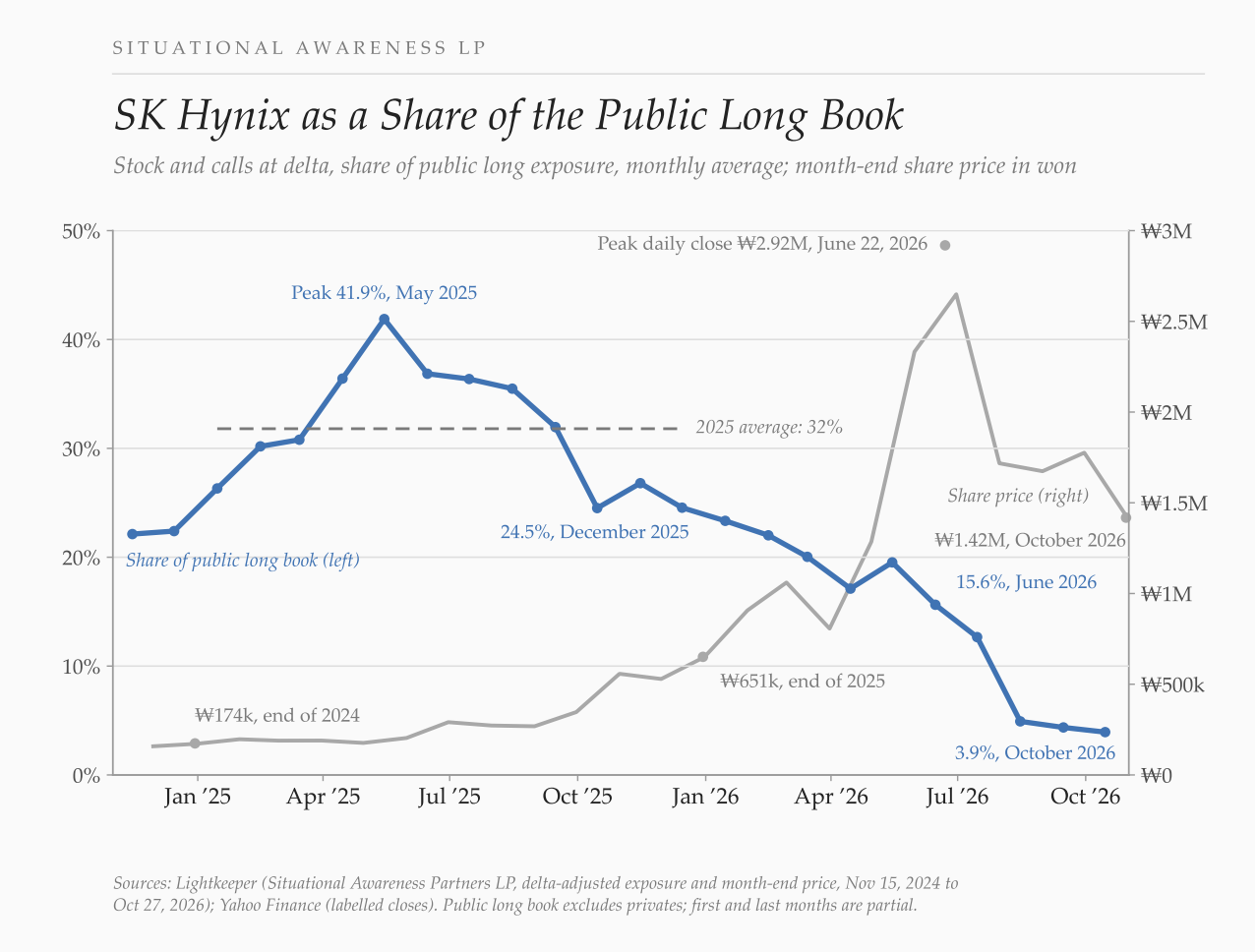


Figure 46: Hynix share price and Hynix as a fraction of our long book. Hynix was our largest position for much of the fund's history, given our contrarian views, asymmetric setup, and deep conviction. We gradually trimmed the position as Hynix rallied.

But let's start at the beginning. As is typical of our investment process, we started with a few key top-down insights, stemming from our overall map of technical developments in AI and their implications. We reflected on these in the first section of this letter:

1) We foresaw a shift to reinforcement learning (RL) training, with RL training becoming a large fraction of the training compute (rather than traditional pretraining workloads).

2) RL would enable models to do much longer-horizon tasks, i.e. scaling test-time compute from "quick paragraph answering a question" to "agents using millions of tokens to accomplish projects" thereby massively increasing inference demand.

These developments implied a critical shift for AI computing: these workloads were incredibly memory-hungry. Inference is bound by memory, not computational power (FLOPS)—and RL training is an inference-like workload.[42]

[42] Intuitively, think of RL training as "try this task 10 times, train on the one rollout where the model gets it right". The "trial and error" part is inference.

For RL and similar workloads, AI labs would often only get <5% FLOPS utilization out of their existing chips; their workloads would be bound by memory bandwidth, leaving most of their chips sitting there unutilized. This was even more true as the time horizons exploded through agentic workloads (i.e., models going from "5 min" chat-style tasks to many-hour Claude Code-style tasks, from hundreds of tokens to millions of tokens), with the KV cache requiring more and more memory.

The release of o1-preview and o3 in late 2024 validated our technical thesis about the direction of AI (models trained with substantial RL and "reasoning" via test-time scaling). Smart industry observers in our network further confirmed our suspicion: the industry would be bottlenecked by memory (bandwidth most of all).

Thus, our view was not only that demand for AI chips would continue to grow—but the memory-intensity of these chips would increase rapidly. That is to say, memory demand would grow even more quickly than AI chips overall. Putting our numbers against the supply numbers implied a crunch. The key bottleneck for this new AI paradigm would be memory.

Moreover, memory was also a key beneficiary of another one of our theses: that the then-assumed Nvidia monopoly would end, and TPUs, Trainium, and other ASICs would gain in share. The bidding war over Nvidia chips would shift deeper into the supply chain, into a bidding war for memory.

By mid-2025, we further expanded on our view by modeling that the test-time compute explosion was also dramatically expanding the demand for general-purpose compute. Agents meant that you needed more traditional (i.e. CPU) servers to run RL environments and run your agent inference—creating incremental demand for DRAM to supply these CPU servers.

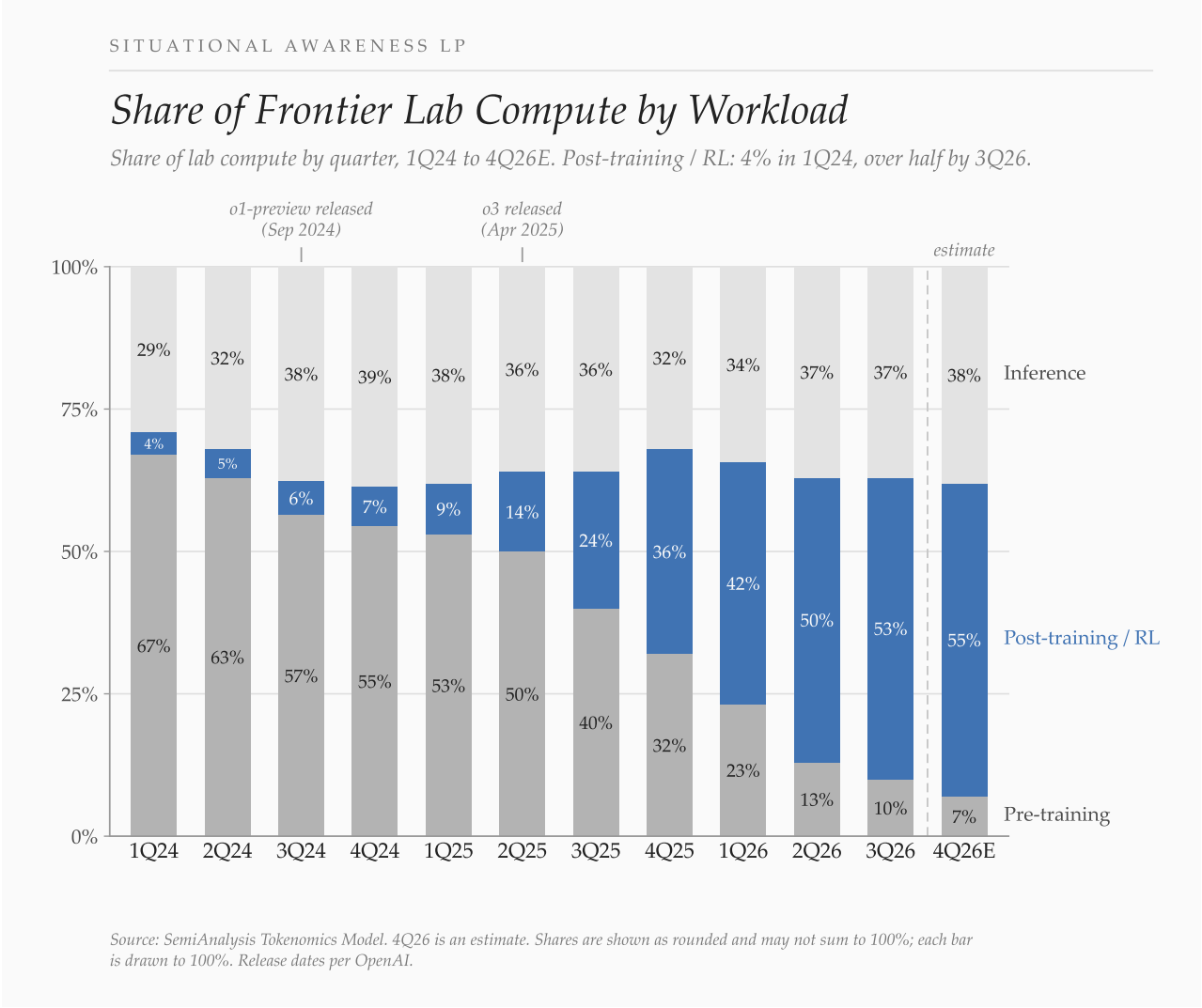


Figure 47: Pretraining was a large majority of workloads when we started, but we foresaw it becoming a small fraction of workloads. This changed the requirements and tradeoffs for chips.

Prompted by this top-down thesis, we spent much of the winter and spring of 2025 doing deep bottoms-up work on the memory sector and specifically SK Hynix—meeting with the people who knew the industry for decades, understanding key technical and engineering differentiators between memory companies, and discerning what sources of information were reliable and which were not.

The more work we did, the more we became convicted on Hynix. Hynix had the best HBM (the high-bandwidth memory necessary for accelerators), with major engineering advantages that played out in quality and yield. Indeed, we believed that memory was becoming a more technologically differentiated product, and AI buyers would

strongly prefer more performant memory, as it was rate limiting on their entire stack. SK Hynix also traded at a large discount to US-listed Micron, 30-50% depending on the metric.

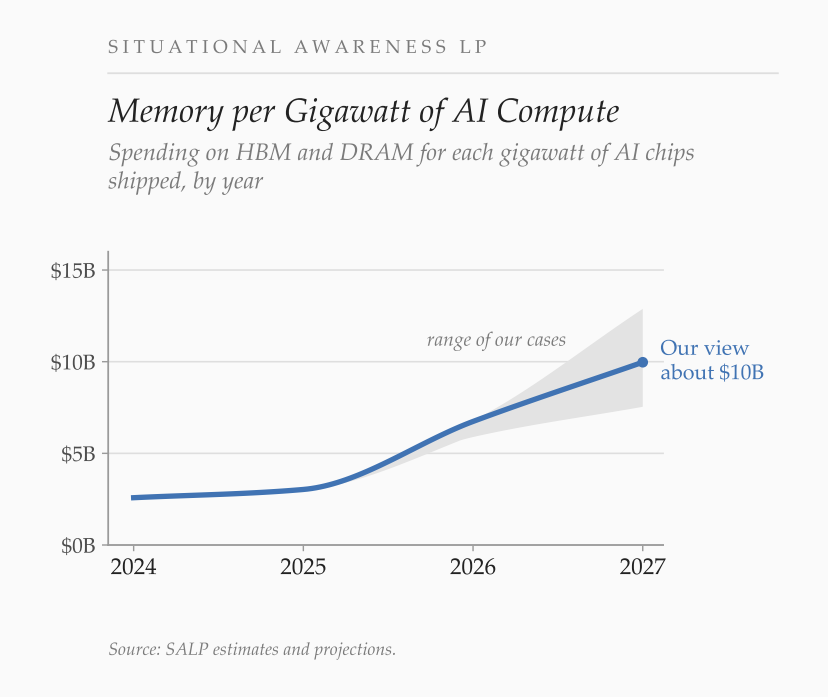
Most of all, there was simply no capacity: given the massive divergence between our numbers and what everyone else expected, if we were right, there would be a massive DRAM crunch. AI was a small share of overall memory demand in 2024, but given the growth we were projecting, it would soon become a large share—with no possibility of adding capacity in time, as fabs take years to build. Memory would be the main bottleneck to the explosion in RL and inference, and nobody seemed to be ready for what was coming.

Memory was also a relatively "free market" part of the supply chain. Memory vendors were used to jacking up prices in favorable times (which would then crater as supply loosened), in contrast to other sectors where vendors might not choose to exercise their pricing power (like TSMC). Memory was the bottleneck on the RL era, yet was capturing so little of the total value created, so we thought there was ample room for prices to go up. Heading into this undersupply, we saw sky-high prices and resulting earnings that would pay back the market cap many times over.

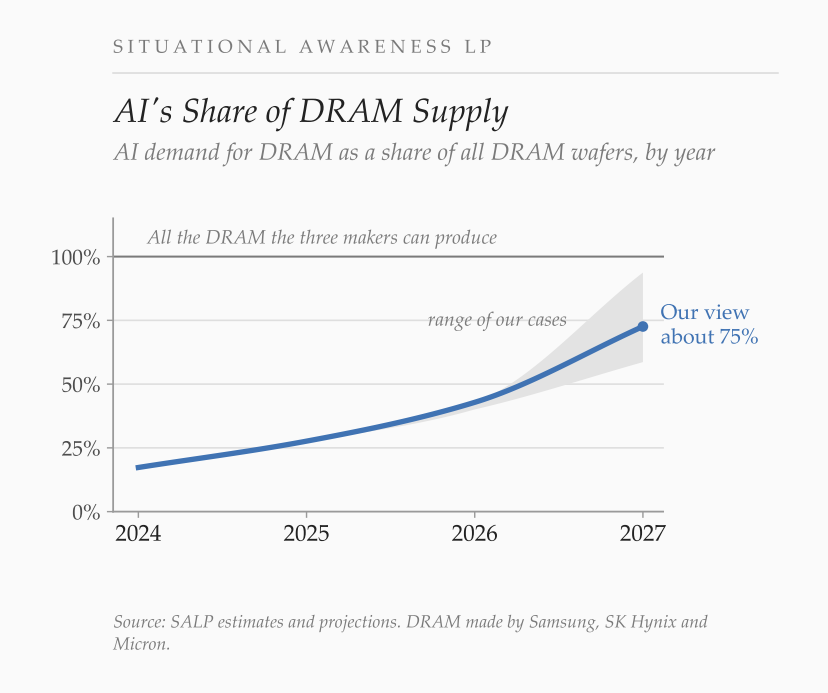
What this picture implied is that if we were right about memory demand, Hynix was at <1x EPS 12 months out. Hynix was going to make more money than TSMC on AI chips, despite being less than one-tenth of the market cap. Hynix was the cheapest semiconductor stock (on a P/E basis) in the semiconductor indices—even though it was the semiconductor company we believed was most torqued to the key developments in AI we were expecting. We loved the risk-reward.

In March 2025, we first visited Korea and met with the SK Hynix CEO, Kwak Noh-Jung, to discuss the AI opportunity. At the time, the management still thought the outlook was relatively bad—we tried to explain our view on why we thought their business was poised to explode, but I'm not sure he really believed us at the time. When we returned to meet with the Hynix CEO again in the fall/winter of 2025, having seen what happened in the interim and how our predictions had played out, he was now keen to hear what we had to say. And more recently, we met with the CEO and members of the management team in the US. We have also formed a collaborative relationship with the Hynix USA team in San Jose.[43] We brokered

[43] We've found we are often able to build strong relationships with management teams of companies we invest in. These are often companies where their business is starting to be driven by AI, and so we have helpful information and datapoints for them to e.g. inform their capital allocation.



(a)



(b)

Figure 48: With the shift to RL and agents, we saw rising memory intensity that would lead to a memory supply crunch.

Hynix's and Anthropic's first meeting. In July, we anchored Hynix's US ADR as a cornerstone investor in their IPO.

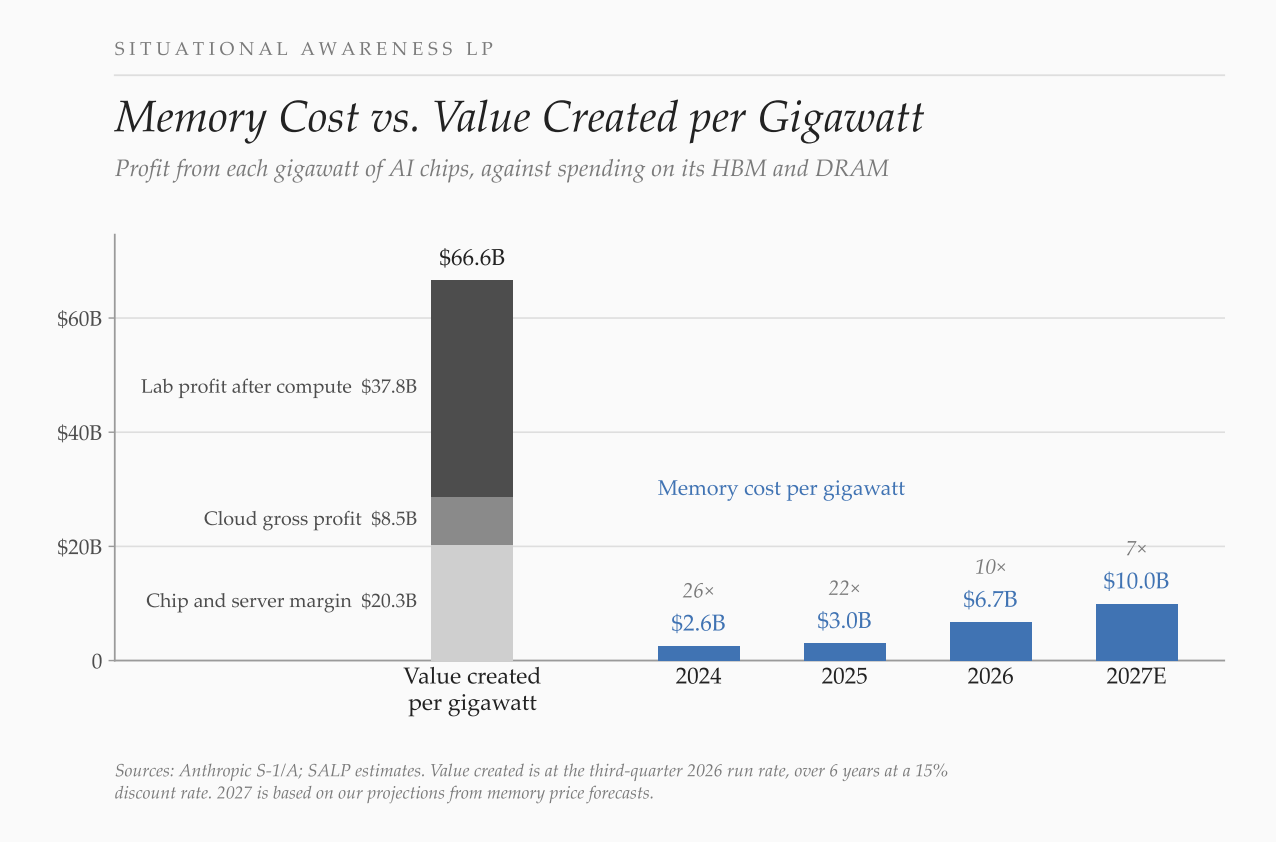


Figure 49: Memory started out as a tiny fraction of the gross profits created by marginal GW downstream (see 2024/25), and so if memory became the bottleneck as we foresaw, there was a lot of room to raise prices.

In the summer of 2025, Hynix was broadly downgraded across the sell-side, and became broadly shorted among fast-money-funds for a few months. The sell-side analysts claimed to have spoken to the company. They had supposedly said the orders for next year (both in volume and in price) were pretty mediocre. Moreover, Samsung was overtaking them in capability and would destroy prices, some claimed. Again, it's hard to capture the sentiment at the time (given how radically the sentiment on memory shifted in the following 12 months!)—but Hynix was broadly shorted, hedge funds were skeptical, Hynix was an underperformer in AI, and in general everything seemed to be negative.

Thanks to the work we had done previously, we were able to figure out what was actually going on. The claims about Samsung were overblown. Most importantly, we knew Hynix itself didn't actually have much demand visibility, and they were usually negative (Korean

companies communicate much more conservatively than American companies). We concluded that despite the sell-side supposedly talking to the company directly, this was FUD with little new information, and we should stick to our guns (and double down).

Had we not been as deep on the company, we might have been shaken out—if the company itself says their business will be bad, and much of the country's media claims it is about to lose its technical advantage—it takes conviction to take the opposite view. As it was, we took the pullback as an opportunity to add to the position, and transmute some of our position delta-neutrally into OTC call options (then offered OTC at high-40s/low-50s vol, since skyrocketing to the mid-70s).

By late 2025, the world started looking very different. HBM demand was skyrocketing, which consumed bits from traditional DRAM supply at >3:1 ratio. This, combined with AI demand for traditional DRAM, led to severe undersupply in the market. This caused DRAM prices to quadruple from 3Q25 to 3Q26, while Hynix quarterly operating profit grew 7x. As we have it, Hynix is run-rating over $200B of operating profit (it was a $100B market cap company when we started!), continuing to grow rapidly. Even we weren't quite bullish enough, even though we were the hedge fund world's Hynix-uber-bull for a couple years. People stopped being able to afford PlayStations and Xboxes as memory prices soared.[44] From our initial buys around ₩180K in late 2024/early 2025, Hynix ran to peak at ₩2.9M by summer 2026, a 16x from where we started.

[44] Carl predicted this (no more PlayStations because of AI…) to me a half decade ago… it was very memorable!

Hynix and its peers were the belle of the ball. Their customers started to offer attractive long-term agreements to secure supply. Some memory companies accepted these deals, receiving a floor price and prepayments in exchange for agreeing to a price ceiling on a certain amount of bits. We spent considerable time with Hynix explaining to them our view of the world, showing many of the charts and forecasts we've shown you, why we thought that prices for memory would continue to increase, that when these customers were offering LTAs, they were trying to pick the memory companies off.

In the end, Hynix held out longer to sign these LTAs than some of its peers, and the terms did improve. Now we believe some recent LTAs signed by Hynix contain price floors but not price ceilings, preserving very valuable upside exposure.

With the market's recent AI bearishness, Hynix has fallen back to close to ₩1M as of November 2026, drawing down almost two-thirds from its summer peak. Now, Hynix trades at roughly 2x our 2027 earnings and 1x end-2027 book value. We discussed the market's current view earlier in this piece: one more year of AI earnings, and then collapse. The company is buying back stock, and at today's prices, would be able to buy back over half the company by the end of next year. Either the multiple will expand or we will be the last external shareholders left...

Kioxia, Sandisk, and NAND

Our views on DRAM got us interested in NAND. Historically, NAND was about 40x cheaper than DRAM and there were 25x more bits of production capacity available than DRAM. In a world in which DRAM supply was far outstripped by AI demand, this would naturally force inventive AI developers to find ways to substitute—perhaps to NAND, we wondered. Even if NAND was a lot worse, you could buy a lot of it for much cheaper. Moreover, we were picking up on new technical efforts like HBF that would increase the bandwidth of NAND to near-HBM levels.

We bought our first shares of Sandisk for $11[45] in September of 2025, though it was a smaller starter position at the time. Sandisk peaked at $292 in June 2026, >27x from our initial buys (before falling below $80 again in November of 2026).

[45] All prices split-adjusted.

Historically, NAND was very low value. More DRAM meant phones that could run more performant operating systems and applications (more working memory). More NAND meant you could hold more photos.

Moreover, NAND captured a low % of the value it created. Last year, Apple charged iPhone buyers ~$0.78 per gigabyte for NAND purchased for ~$0.06 (the 512GB iPhone rather than the 256GB iPhone). This was because the NAND industry structure was especially bad. DRAM is dominated by three players, while NAND has six. NAND capacity also grows naturally as makers stack more layers, without constraints of adding new equipment, versus DRAM which requires shrinking the cell, a difficult task, making it more prone to tilt into oversupply.

AI dramatically changes the value that customers earn from purchasing NAND. As models work on long tasks, they build up a cache of what has been processed that they use for future work. If a session is paused and resumed, there are two options: a) throw away the cache and recompute it later (costing expensive GPU and DRAM time), or b) park it in NAND. As the cost differential was so high between DRAM and NAND, this would naturally pull up the cost of NAND closer to that of DRAM.

This is not the only use of NAND in AI datacenters. Agents need access to a company's corpus of information; multimodal consumes and produces lots of data. NAND is going from enabling your mother to save more photos to enabling intelligence, so naturally the equilibrium price point should change. AI systems are now being designed to use NAND, and NAND suppliers are creating new products that match AI's needs, such as Kioxia's XL-Flash, which sacrifices capacity for speed, or high bandwidth flash (HBF), which redesigns NAND to be read in parallel and places it next to the GPU to maximize throughput.

Against this demand backdrop, it is more difficult for supply to respond than historically. Many makers of NAND also make DRAM. Fabs take four years to build, and dual NAND/DRAM suppliers are devoting their incremental capacity to DRAM instead due to the supply crunch there, constraining NAND supply growth further.

We saw a similar path happening in NAND as it did in DRAM. First, extra AI demand would swing the market into undersupply. Then, because the value generated by AI customers was so much higher than traditional customers, they would bid the price of NAND to all-time-high levels.

In January 2026, we substantially increased our position across Kioxia and Sandisk based on those more refined views.

After we developed these views in January, we poked around the industry and did not find much confirmatory evidence. At the time, Nvidia had launched a form factor to use NAND for KV cache offloading, but we couldn't find any buyers or users of these products. However, we knew the economic logic was sound, and despite technical details not being worked out yet, we felt confident the market would head toward us as the industry ran out of DRAM bits and was forced to look elsewhere for substitutes.

Since the end of last year, NAND prices have more than tripled, and the share prices of Kioxia and Sandisk rose 4-5x from the end of January to their peaks. We've now picked up signs of the industry designing systems around using NAND and the labs placing direct orders.

In recent months Kioxia and Sandisk have fallen sharply with the rest of the AI trade (Kioxia drawing down 80%). Kioxia now trades below 1x our 2027 earnings (and 0.2x 2027+2028 earnings). On our estimates, it will generate enough cash to buy back three-quarters of its market cap by the end of next year. Once again, either the stock goes up or we will be the only external shareholders left.

At current market prices, we think Kioxia is once again a highly asymmetric opportunity, and it remains one of our largest positions.

How we work

Our journeys on memory, in both DRAM and NAND, were examples of our broader approach.

From the beginning, we have approached our work on the investment team as though we were an AI lab strategy team. We look at technical progress in AI, and think about the implications: what are the core technical uncertainties and trajectory? What capabilities will that enable, and what new revenue streams? What kinds of requirements does that put on AI compute? How will the market for certain inputs adjust, and what new elasticities come into play? Which sectors will start to be affected? If our crazy beliefs about the trajectory of AI are true, what other implications does that have across the economy? And so on.

Another approach is to plug in downstream: to count the chips coming off the assembly line, so to speak. The core issue is that this is fundamentally a lagging indicator. The downstream companies themselves often don't know; they are far from SF. Memory companies didn't see memory prices going ballistic. Even the hyperscalers are lagging rather than leading indicators: Microsoft pulled back on datacenter leases in spring of 2025, even though agents were about to explode, and they later had to make up for that shortfall by buying much more expensive compute.

A deep understanding of the core root cause drivers in AI development and its implications lets us have a highly differentiated picture before the market.

Our journey on Hynix and Sandisk/Kioxia exemplify how we try to marry top-down and bottoms-up investing—marrying key insights about the direction of AI with extensive bottoms-up work on industries and companies far from SF to make a concentrated bet on a contrarian and asymmetric thesis—and where deep knowledge and conviction on our positions shine. It also highlights areas we've grown as investors since starting the fund: creating strong contacts throughout a new industry and building a dialogue with management teams.

We applied this same playbook to many other investments. We discuss a few more in **Appendix A.**

IIc. Spot compute: the microeconomic implications of RSI

An H100 draws about a kilowatt and has roughly the processing power of a human brain. In the past, it earned about $1-2/hour. As models have become a virtual intern, this has gone to $5-10/hour. As RSI takes hold and we approach superintelligence, we think this goes to $100s/hour+.

Earlier, we described the RSI trend as the most important thing happening in AI—and in the world. Here, we want to talk about a key expression that has become a new focus of ours over the last six months.

RSI means rising $/kWh

What is the most important microeconomic reflection of RSI? We think it is growing revenue per kilowatt-hour.

In the "old" days, the 2023-2025 regime with OpenAI as the leading lab, revenue per kWh was roughly flat. That was remarkable on its own: OpenAI kept 3x'ing GW year over year, and rather than seeing diminishing returns, revenue scaled linearly with compute. Model capabilities grew with more compute, and model progress expanding TAM compensated for diminishing returns on more compute being available. OpenAI could keep scaling GW and keep scaling revenue with it, making ~$1.2 per kWh of overall compute, or roughly $2.4 per kWh of inference compute.[46]

[46] Assuming half its compute served inference.

Yet this year, as model progress has accelerated, the leading lab, now Anthropic, has broken from this trend. As the models have become more useful—going from chatbots to agents/virtual interns—while compute has become more constrained, revenue per kWh has trended up. Anthropic earned less per kWh than OpenAI throughout 2025. Since the start of 2026, as it took the baton from OpenAI as the leading lab, its revenue per kWh has quadrupled to roughly $2.70, and its revenue per kWh of inference compute has reached as high as ~$11.

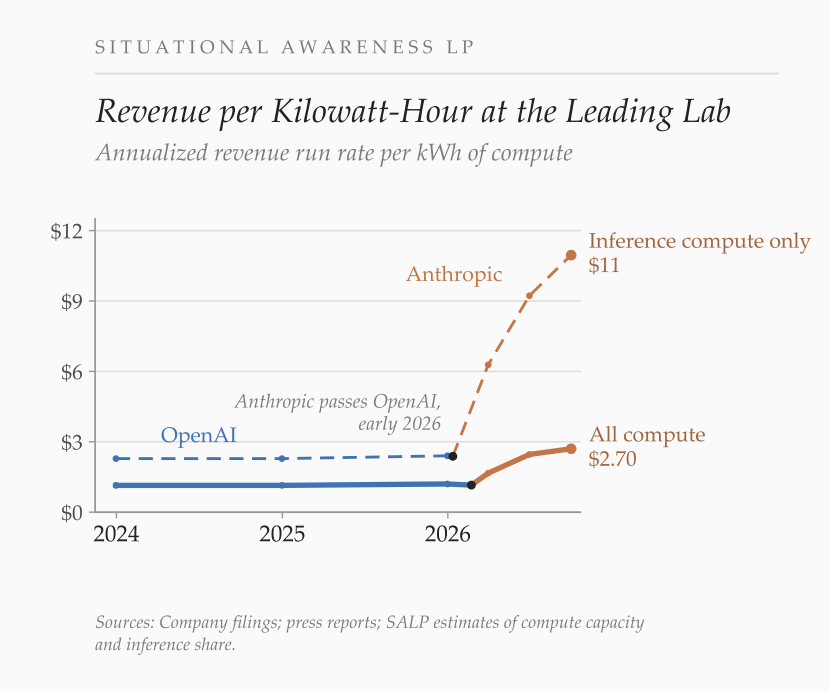


Figure 50: In the "old" days, revenue per kWh was roughly flat, with revenue scaling linearly as GW of compute grew. In the last nine months, as model progress accelerated, revenue per kWh has started growing.

An intuition pump might be useful here. As an extremely rough approximation, an H100 is about the FLOPS, i.e. the processing power, of a human brain,[47] and consumes roughly a kilowatt of power. So we can think about "revenue per kWh" as approximately "revenue per hour of a human brain's worth of compute".

- Low paid human knowledge work might earn $15/hour ~ $15/kWh.
- Very capable humans can earn $1000s/hour ~ $1000s/kWh.
- Up until recently, models were earning ~$2 per kWh of inference compute. This year, the leading models have started earning ~$10/kWh of inference compute.[48]

[47] Estimates of the brain's computational power span several orders of magnitude.

[48] These figures are estimated on the installed base of compute, which blends older and newer chips. We estimate that newer chips are ~2x more performant per kWh, so most of the ~5x increase comes from models earning more for the same compute.

    - This intuitively makes sense in our human comparison: this year, the models became as good as interns or outsourced software engineers in broad usage.
- Up until recently, an H100 rented for ~$1/kWh, generating an adequate return for its financiers. Mid-this year, we saw short-term deals pricing at ~$4/kWh, which pays the hardware back in under two years.[49]
- *Our claim*: In the next couple of years, we expect AI models to become as good as the smartest humans at turning FLOPS into economic value. We think that revenue per kWh could continue to go up—in an extreme case, this could go to $1000s/kWh, like the smartest humans' wages, a factor of 100-1000x from here.

[49] Under its May 2026 deal, SpaceX leases a mix of Hopper and Blackwell chips to Anthropic at about $3.50/kWh blended, on our estimates. More recently, this price has been raised to >$5/kWh.

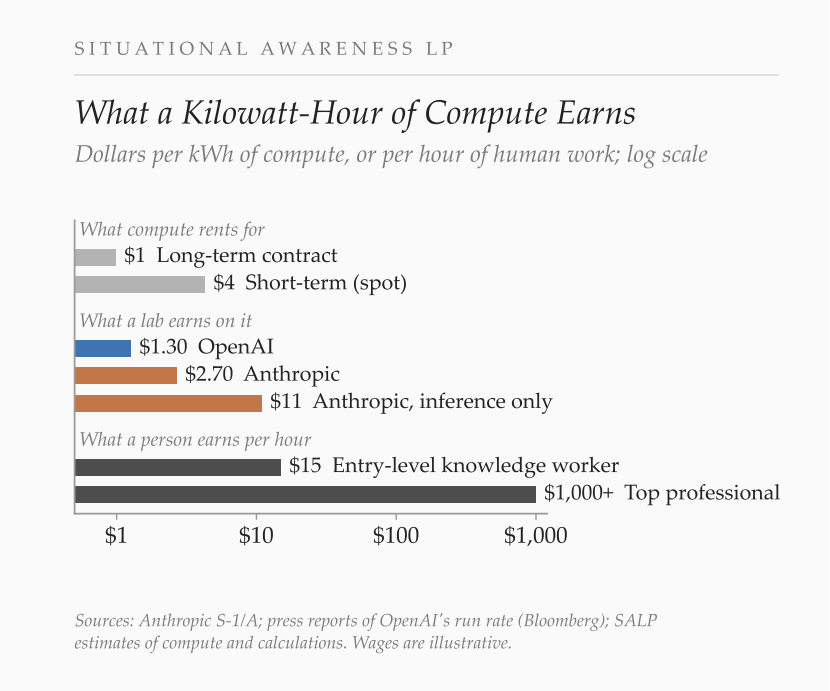


Figure 51: Compute used to generate and rent for about $1/kWh. More recently, it has generated >$10/kWh of inference compute. The human analogy suggests $1000/kWh of value generation is possible.

We can think about this stemming from RSI:

- In the "old days", compute and algorithmic progress were scaling roughly proportionally: each contributed 0.5 OOMs of effective compute a year, roughly 3x/year.
    - More compute led to diminishing returns in value per compute, but algorithmic progress scaling increased the return per compute, keeping the overall value per unit compute constant.
- RSI is now dramatically speeding up AI progress. But this speedup is entirely weighted to algorithmic progress—we are still

doing 0.5 OOMs/year (3x/year) in compute scaling at the labs, but perhaps 1.5 OOMs/year (30x/year) in algorithmic progress scaling.
    - Algorithmic progress is raising returns per compute faster than more ample compute can lead to diminishing returns per compute.

RSI means the models are getting better, resulting in economic usefulness growing faster than compute growth. Thus, $/kWh is growing.

The beneficiaries of rising $/kWh

If this "revenue per kWh rising due to RSI" thesis is true, there are two main beneficiaries:

- *The leading lab(s).* Their revenue per GW of compute rise, and margins rise with it. The leading lab used to rent long-term compute at ~$1/kWh and sell inference at $2/kWh; now it can sell inference at $10/kWh, and soon it might be able to sell inference at $100/kWh. On compute it has already secured (e.g. via a long term contract), its margins go from 10s of % to 85% gross margins (now) to even more extreme gross margins. We have started to see this in Anthropic's growing profitability.
- *Spot compute (short-term/"floating rate" compute).* While most compute so far has been locked away in long term fixed price contracts, compute that is "floating rate" would get bid up in such a scenario. If the leading lab is making $10/kWh on marginal inference, it happily buys marginal compute at $5/kWh to serve more customers, as is the case today. Meanwhile, for the cloud serving the compute, this is a much better deal: perhaps a cluster built today costs them $50k/kW of capex, amortized to $1.2/kWh,[50] which on a long-term contract would sell for $1.4/kWh, i.e. slim margins. Now, they can sell compute at $5/kWh, taking margins from ~15% to ~75%—and capex payback periods from 4-5 years to about 1 year. Their profits and cashflow have incredible torque to rising prices.[51]

We have a large lab position in the portfolio. But what about spot compute? It seems that spot compute has large call option value on RSI.

[50] Supply chain prices are higher than they used to be. Older clusters, built for less, were contracted near ~$1/kWh.

[51] In our earlier section on labs, we discussed how labs are *not* like traditionally-terrible airline businesses, but instead more like high-quality jet engine industrial businesses. Sometimes, however, airlines *can* be an amazing trade—such as during the COVID recovery, when the stock prices of airlines with large fleets ripped on an unexpected resurgence in travel demand. Spot compute is a bit like an extreme version of that.

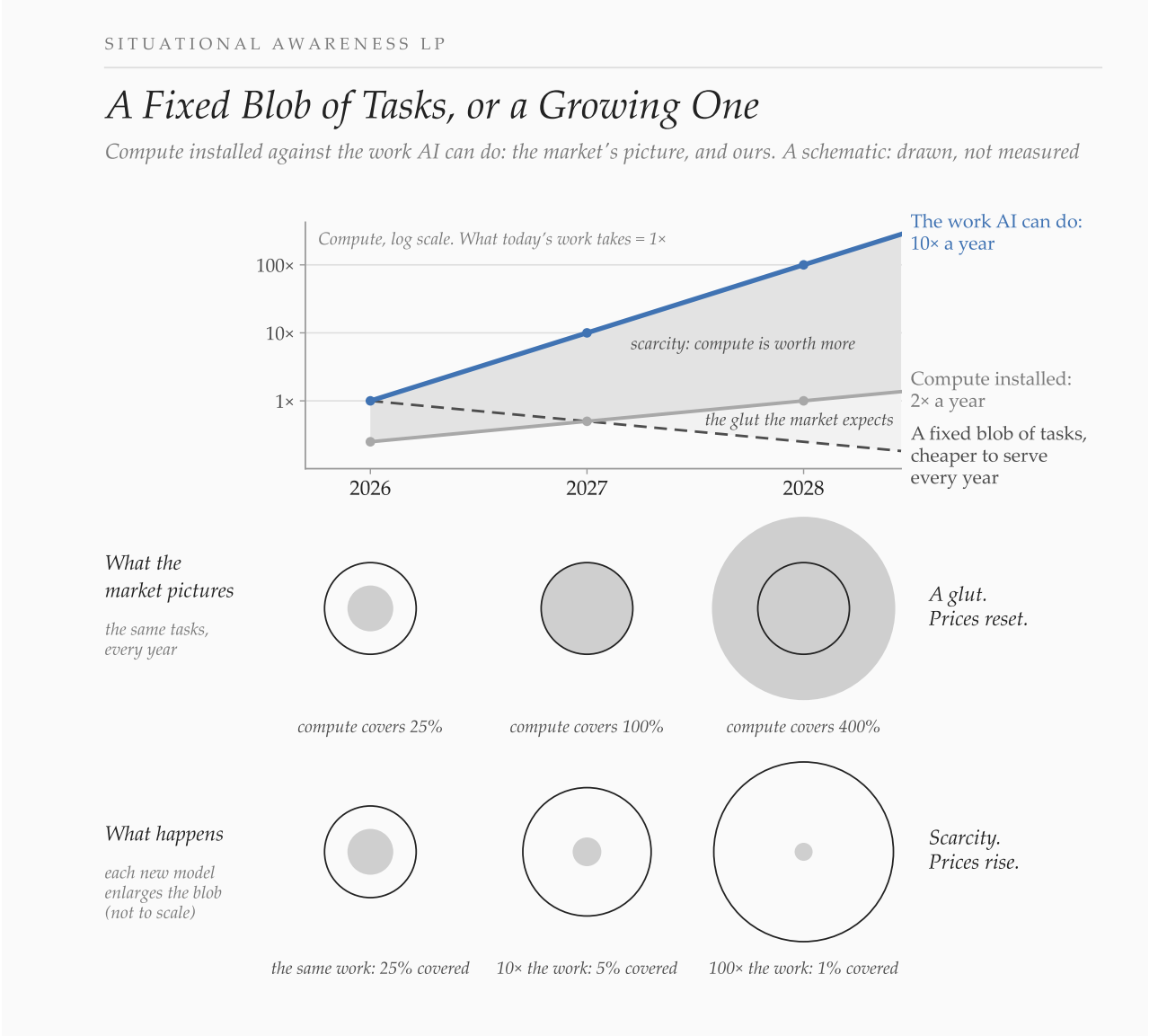


Figure 52: The market assumes a "fixed blob of tasks" that will lead $/kWh to collapse. We think RSI leads to the opposite.

Our RSI view of spot compute is one of the biggest disagreements we have with the market. We think that the market is pricing in a rapid collapse in spot compute prices starting next year. Investors see rising spot prices as a classic commodities cycle: that compute is scarce today but the market will be flooded with chips soon. In this view, a fixed blob of tasks to be saturated and rising token efficiency imply an impending reset in rental prices. We think RSI will continue and accelerate, leading spot prices to continue to rise.

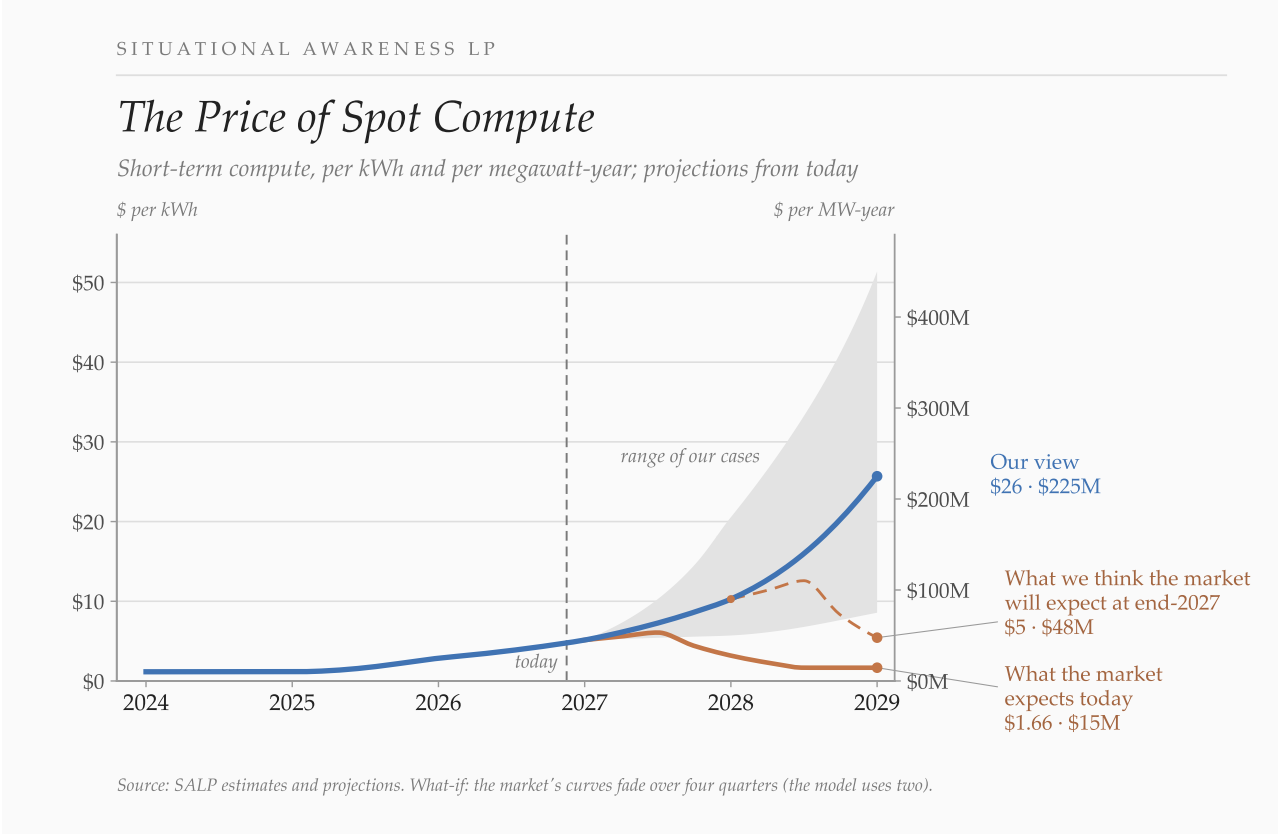


Figure 53: Spot price forecasts built from SALP's general equilibrium model incorporating competing labs, pricing dynamics in different segments of the AI market, compute-supply curves of compute providers, and much more. We think substantially more extreme scenarios than visualized are plausible in more intense near-term RSI worlds, with $/kWh going into the $100s or above.

We began exploring these ideas in the spring. Soon, one very notable fellow traveler appeared: Elon Musk. Before him, essentially nobody sold spot compute at relevant scales; Elon inaugurated the market.

In May, SpaceX leased Anthropic roughly 325k GPUs across its Colossus campuses for $1.25B per month. In June, it leased Google about 110k more, for $920 million a month. Both deals nominally run into 2029, but either side can walk away on 90 days' notice, making it in essence spot compute. On our estimates the Anthropic deal was struck at ~$3.50/kWh and the Google deal at ~$5.50/kWh. In October the Anthropic deal was repriced to ~$5.50/kWh, about 60% higher (even though it runs on a less efficient Hopper-heavy fleet). SpaceX can pay back their clusters (which have 5+ year useful lives) in under a year.

Others followed. We met Nebius's management in June and detailed our spot compute thesis. Over the next few months, Nebius launched a spot compute program, and disclosed short-term deals pricing at $40-50 million per megawatt-year, or ~$4.50-5.50/kWh, similar to SpaceX. CoreWeave and IREN have followed with the same.

We expect short-term compute contracts to remain a central part of how frontier labs buy compute, because it solves their hardest planning problem. Datacenters must be committed a year or two ahead, against revenue that might grow 5x or 10x. Buy too much and you go bankrupt; buy too little and you turn customers away. As we discussed in our Anthropic section, Anthropic has been evolving toward an approach that leans on spot compute: buy "on the low side" ahead of time, and if things go better than expected, buy more on the spot market. It costs more, but only in the worlds where the company is doing particularly well. Dario recently described this in his interview with Jamie Dimon, calling it "an insurance policy."

WHILE WE THINK spot compute investments will pay off handsomely, there are only a select few good expressions of the thesis in public markets. With less bold views on where spot prices go, most compute providers prefer committed offtake from investment-grade customers, whose credit lets 80-90% of a cluster's cost be financed with debt, over taking price and financing risk on a spot buildout.

The economics of a spot cluster are strong even with no debt: at today's short-term prices, unlevered returns run above 60%, and reach triple digits if payback stays under a year.

In addition to SpaceX (we are very bullish on their spot compute play, but the broad impending space + cars conglomerate dilutes our spot compute view), there are others venturing into this market.

We discussed Nebius earlier. Its deal with Meta has two parts: $12B of committed capacity, and a further $15B that Meta must buy if Nebius asks it to. Nebius can borrow against Meta's commitment, then sell the capacity to whoever pays most.

Another example is Sharon AI (SHAZ), under Nvidia's AICP program. Nvidia has guaranteed Sharon a minimum of $4.9B over six years on 40,000 GPUs, about $1.30/kWh. Lenders finance against this floor, and Sharon shares half of any revenue above it with Nvidia. Many neoclouds passed on this structure because, on their lower spot

assumptions, giving up half the upside looked expensive. On the day Sharon announced the deal in June, its stock closed down 13%. We read the same contract as the best spot structure available in public markets: cheap financing against Nvidia's credit. We decided to lead an equity investment into Sharon.

If SHAZ were to deliver only its current contracted book, walk away from all its uncontracted power, and sell the AICP compute at the rates long-term contracts sign for today, today's price is covered. At the rates that management has conservatively guided to, the shares would be worth about twice today's price. If we are right about spot prices, there could be 10x to 50x upside.

From a portfolio perspective, spot compute functions as a valuable hedge on the supply-side. Most of what can go wrong for other positions on the supply side makes installed compute worth more. If a Taiwan invasion takes out the semis supply chain, or regulatory constraints prevent building new datacenters, the installed base becomes even more valuable.

Of course, spot compute is the most vulnerable on the demand-side, and it is the first to be cut in a downturn (cf our discussion of Anthropic's compute procurement strategy). If extreme political intervention were to halt AI progress, spot compute prices would collapse. Asymmetry comes only if the market is already pricing this collapse (as it is right now).

Moreover, much like memory, if our forecasts are right, the cash generated in the coming years pays back market caps many times. This is critical in the age of RSI, as we will discuss later. A difficulty we often run into, when underwriting businesses, is that fast AI progress often reduces the value of assets (as we discussing in the Investing in ASI, rather than AGI section). Software companies whose moat is that they wrote a pile of code will see their advantage reduced once code becomes cheap and painless to write. Chip designers who have invested billions into technical design IP and engineers soon will as well. Today's compute will, too, one day be obsolete, but unlike application companies or chip designers or semicap companies trading at high multiples assuming decades of cash flows (that will be cut far short by AI obsolescence), rising spot compute prices could generate high cash-on-cash returns in a few years.

IId. The ChatGPT moment for robotics

We think we are starting to see signs of life on general robotics models, and the buildout to billions of robots is becoming an increasing focus of our investing.

This is a short chapter: a preview of a theme we expect to write about at length in the next couple of years.

Robotics is a software problem, not a hardware problem. You or I could teleoperate existing robots out there and do most physical jobs in the economy.[52] The barrier to automating physical tasks has been the models.

[52] For example, sometimes people cite the lack of pressure sensitivity on robot fingers/grippers as a big problem. It turns out humans can learn in a few hours to do basically just as well, even on fine grained tasks, without that pressure sensitivity. More broadly, robot hardware has lots of limitations that a smart enough model could easily compensate for.

In the past few years, we've seen lots of robotic demonstrations that look impressive, but fall flat on closer inspection: driven by a model that works just in that specific room, that specific lighting condition, that specific robot body, for that specific task which the robot model had months of specialized training for. We were far from general models; we've been in the 2010s-cat-or-dog-image-recognition-model era for robotics.

We saw the first evidence of that changing in the spring: specialist robot models were scaling with pretraining data, learning new tasks in days instead of months, and generalizing to unseen tasks and new robot bodies. We thought this was a defining moment, akin to GPT-3 in 2020: the point at which a single pretrained model could pick up a new task from a few examples, rather than being retrained for each one.

Specialist model progress continues, but the biggest upset has come from frontier LLMs. In September, OpenAI's GPT-6 Astra, with no

robot-specific fine-tuning, scored about 29 on RoboDojo, a public simulation benchmark, ahead of every public specialist robot model at the time. Astra's physical ability appears to have emerged largely from general multimodal training, not from robot data.[53] Two months later, Fable 5.2 scored about 65 on the same benchmark, on tasks Anthropic says it was never explicitly trained on.

[53] See, for example, Robonomics, "Robotics' New Era After Astra", November 2026

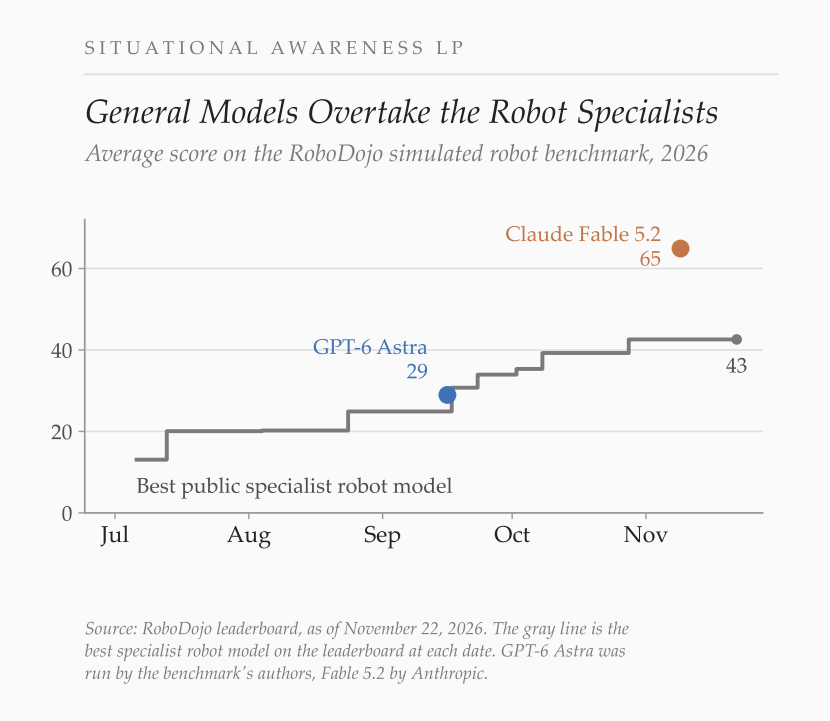


Figure 54: Without specialized robotics training, general frontier models have started to outperform public specialized models on robotics tasks.

That said, note that the public frontier of specialized models does not necessarily reflect the private frontier of specialized models, and these benchmarks overlook shortcomings of the general models in real-world precision manipulation compared to specialist models. Still, the general LLMs are showing strong signs of life on robotics without the labs even devoting specific training effort to robotics yet. This has gone so far as OpenAI recently disbanding its specialized robotics team, folding in robotics efforts to their general models.

Just as the frontier labs are adding specialist robot data, the specialist labs are making step-function progress by borrowing LLM techniques: adding reasoning, video pretraining, and learning from their own experience, which is showing promising results of models being able to chain together actions and adapt to real-world environments. We have bets on both paths to robotics: in addition to our

large Anthropic position, we made a small investment in Physical Intelligence, which we viewed as the leader in the space, in the spring.

We are seeing the start of capabilities "falling out" of general scaling of data and compute versus the painstaking work of the prior era. We see a steep curve ahead in the next 15 months. Remember the old adage we've discussed in past letters: once the general models start seeing first signs of life on an eval, you're usually only a year or so out from saturating and crushing that eval with further scaling.

The result is that robotics could arrive earlier than we expected: not after superintelligence, solved by the automated researchers, but alongside them.

We expect the "ChatGPT moment for robotics" is approaching, perhaps later next year. The models won't be perfect yet and some time will be needed to scale up robotics hardware production, *but it will be clear general robotics works*. We'll have models that can be air-dropped into arbitrary hardware and environment, and they'll be able to autonomously do many physical jobs with just a little instruction. Much like models became generally intelligent across knowledge work, models will be generally *physically intelligent*, and—just like training new employees—can be onboarded to specialist tasks with a few demonstrations and feedback.

We think it'll become clear to the market then that we will want to scale up to make billions of robots. (Buying a robot for $10K to do a job costing $100K/year of payroll seems like an amazing proposition...)

We think this will increasingly become a major theme for our investing over the next year. We have been deepening our work across the supply chain, and we are starting to see opportunities across public markets. As we near the ChatGPT moment for robotics, we think there will be highly attractive market expressions: the physical robot buildout in the coming years like the compute buildout since ChatGPT.

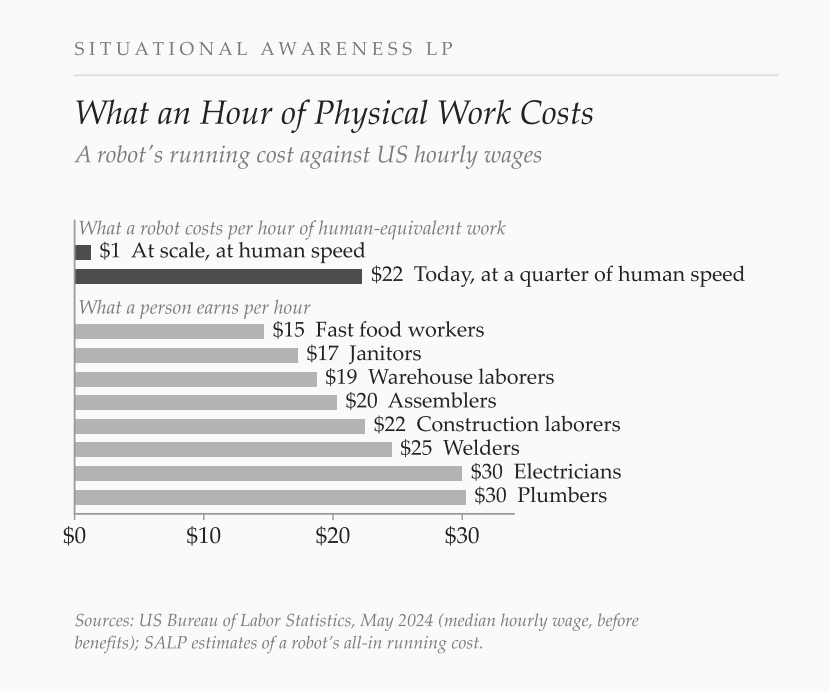


Figure 55: Once robotics models become generally physically intelligent, existing hardware will make scaling to billions of robots to replace and augment human labor highly economically attractive.

IIe. Investing in ASI, rather than AGI

Most companies will lose most of their value as superintelligence obsolesces most of their assets—their skilled employees, their IP, and their management. That includes many of today's AI winners. It is easy to accidentally be long AGI and short ASI. >$500T of wealth will be created, and ~$500T of wealth will be destroyed. The assets to own are those that return their value in cash before ubiquitous obsolescence—or the Mona Lisa.

Looking to the next three years, our views on the technological and economic trajectory of AI are at least as divergent from market consensus as they were at the publication of Situational Awareness in 2024. Technologically, we expect progress will be far faster than it was over the last 3 years due to RSI, absent a regulatory ban. Economically, we conclude that technological progress will usher in both massive value creation and destruction. That value destruction, especially in relative underperformance as a share of world wealth, may be very comprehensive across financial assets, including the "highest-quality" businesses, most purported "AI winners", and many safe haven assets.

Most equities will lose most of their value (at least in relative terms), including most "AI winners"

RSI dynamics and the accelerating rate of AI progress point to AI models achieving broad based superiority across all work the market pays people for, i.e. superintelligence. This is a much broader set of abilities than how most analysts envision the impact and diffusion of AI capabilities: it is not just individual contributor knowledge work, but physical world tasks (whether directing human or robot hands),

and socially embedded roles such as management, entrepreneurship, sales, and negotiations. It also includes the ability to rapidly learn and invent new knowledge.

This level of broad based intelligence could comprehensively slash the cost to replace and exceed core pillars of valuation for the world's largest companies:

- IP/technology
- Human capital
- Physical capital
- Management/entrepreneurship

ASI will first rapidly obsolesce most of the value of American corporations today: IP, technology, and organized human capital. For the remaining value attributable to hard assets, some of these will be scarce during the singularity, but they must earn cash returns in a compressed period of a few years before ASI applied to physical engineering and robotics collapses their replacement value as well.

THE FIRST TO BE DISRUPTED

IT outsourcing companies are still priced for reasonably steady earnings; obviously AI will be a far superior and cheaper replacement for their services. Software companies are still priced for sustained growth, some for very rapid growth, all while their accumulated IP and stock of software engineers is becoming trivially replaceable. Application-layer companies ("AI for x", whose value often resides in connectors and UI interfaces atop AI models) command high venture multiples—in a world in which the general models are the product. (VCs have tended to be long AI, short ASI.)

Many make arguments about why these businesses will persist after all, but these arguments fundamentally assume that AI capabilities stall and remain jagged, preserving legacy moats for decades to come. "IT outsourcing companies will be responsible for management of AI integration"—superintelligence in a couple of years will be able to come into companies, interview every employee, pick up on all company processes and knowledge, and integrate itself into any system far better than human consultants. "Software companies will have enterprise distribution"—superintelligence will become superhuman at sales too, and easily be able to port over any prior system. Software and cybersecurity companies are at multiples requiring

many years of sales at the brink of complete automation—even those benefiting from AI-induced demand like cybersecurity can't sustain that long enough to recover their valuations. "Application companies will enable diffusing AI in specialized verticals that require special xyz"—even if true, the labs will have an army of automated AI researchers to cover every vertical soon.[54,55]

RSI means actually getting the panoply of human abilities: most bets on limited adoption/impact/bottlenecking involve hidden failures on the capabilities side, an implicit claim that "superintelligence" would still lag behind human capabilities in certain domains.

But these are the easily identifiable victims of ASI—what about the rest of the stock market?

THE WORLD'S LARGEST COMPANIES

It is often claimed that the S&P 500 is dominated by AI. That the biggest risk to the S&P 500 is that "AI is over". But what if the biggest risk for most of its constituents, including effectively all of its largest companies, is the opposite?

NVDA. Consider Nvidia and its many-trillion dollar market cap. The company has tens of thousands of talented engineers and IP that help it build top-notch AI chip designs, a software and developer ecosystem (CUDA, etc.) supporting its products, business and financial relationships with the supply chain that let it lock up production from physical fabs and factories (despite being a fabless chip design company), and a history of strong management and execution. RSI threatens to eliminate almost all of this value.

The advance of the AI models will render the engineering expertise at NVDA as obsolete as mathematicians', the equivalent of millions of engineers far better and faster than any humans. Those models' work will surpass and obsolesce the stock of chip design IP, e.g. leading SerDes IP that has advantaged NVDA and AVGO chips. OpenAI's Jalapeno chip and collaboration with AMD show the early signs of this transition, but in a couple years that will be looked back on like models first writing a few lines of code. The models will be able to design radically better chips than the current state of the art, far faster and cheaper.[56]

[54] Some wrapper companies can manage to get acquired before obsolescence. Cursor had a good exit, even though Claude Code made their "nice IDE to use Claude in" product obsolete—fundamentally, they had assembled a strong AI team that SpaceX valued, effectively an assemblage of "Meta-style" offers in an acquihire.

But that human capital will no longer be valuable in a few years; if you miss the quickly-closing exit window, there is no terminal value. This is a fundamental problem with nearly all Silicon-Valley-style early-stage startups in the age of ASI: they are valued for a highly competent team that is expected to be able to repeatedly develop products with greater TAM. That is, the value is human capital, one that will soon depreciate.

[55] Prompt engineering startups once commanded lofty valuations, before AI models became smart enough to just "get" user intent from hastily worded prompts.

[56] OpenAI's models have made optimizations to its own Jalapeño chip and to AMD's flagship designs that are going into production. Of course, designing the best circuits or SerDes requires physical testing loops outside of simulation, but timelines will be rapidly reduced with superintelligent models that can optimize simulations and design the most informative experiments—much like ML research looks to be accelerating, despite needing scarce physical GPU-hours.

Surrounding software like Nvidia's CUDA is being devalued in the same way, with models able to rapidly optimize to exploit new chips, and coordination of human developers rendered moot. At Anthropic, bringing up a new chip platform, which used to take dozens of engineers months of work, now takes Claude a few days.[57] In the near future labs will get a better return on logic, memory, and other components than NVDA, and will bid up their prices to the point that NVDA cannot secure additional components to produce chips profitably.

[57] Per Anthropic's Chief Compute Officer, Tom Brown, on the IPO roadshow video: "In the past, bringing on a new chip platform took dozens of engineers months of work... it got Claude running on AMD in just 3 days with basically no human intervention."

NVDA's supply chain agreements lock up access to scarce physical capital in component manufacturing capacity in the short term, and it can profit by capturing the spread between those deals and the value of the marginal chip to an Anthropic or OpenAI. But these agreements are finite in duration.

In other words, while the stock market is currently capitalizing more AI profits for NVDA than any other company, it also appears that it cannot survive in its current form for more than a few years on the current trajectory of RSI. While we think Nvidia will sell more chips in the coming couple of years than the market consensus, the market's DCF is also assuming years of ongoing chip sales at high margins. We think that DCF will be cut off far sooner.

For NVDA, gatekeeper of a large portion of AI compute, we can imagine a case where it does capture enough of the value created by RSI for positive return—a world in which it raises GPU prices by an order of magnitude in its last couple years of relevance. NVDA could still deliver strong returns, but effectively as a set of embedded short term commodity trades in compute rental/chip/server component prices and investments in other companies exposed to those changes. Those trades are available more cheaply in other companies, because the market is charging a premium multiple (vs memory or neoclouds) for competitive advantages that RSI will destroy.

AVGO, AMD, QCOM, ARM, SNPS, ETC. Similar to Nvidia. Other companies trading at 20x, 40x, or 100x earnings simply have no escape in the face of RSI. Either they are simply not exposed to enough AI upside to cover their market capitalization while facing the collapse of their business from ASI downside, or they have taken on contracts that cap their short term upside while their valuation demands much more than the contracted earnings.

ASML. AI will obsolesce chip designs sent to existing fabs before replacing ASML's EUV machines. But AI will not only be able to invent superior replacement technologies, but entire intricate supply chains: simultaneously absorbing, reinventing, or surpassing contributions from numerous firms with models that master and advance every field. ASML, valued on sky-high multiples for its perceived robust monopoly, will see its DCF cut off in years rather than decades.

MICROSOFT. Windows and Office lose to superior AI replacements, and AI agents handling switching costs. Its high margin Azure value-added services are devalued. Data centers collect contracted revenues (likely failing to capitalize on spot prices by signing long-duration contracts), but plausibly no renewals beyond the original long-duration contract as both chips and data centers see replacement cost collapse with better designs and robotic physical production.

APPLE. iPhone and its superiority in hardware and software design will be leapfrogged in engineering and product design by ASI. Its hidden App Store taxes are circumvented by agents as the central interface, and by smart agents advising consumers on alternatives.

ELI LILLY. Pharma products will still have their patents, but will become obsolesced by far superior and far cheaper (commoditized) cures and medications developed by AI radically accelerating life sciences R&D.

SPCX AND TSLA. Elon's incredible track record of entrepreneurship and management commands a premium multiple today, but ASI will provide superior leadership at scale: omnidisciplinary superstar CEOs on the spot throughout every part of the company with access to all corporate information. They would be especially advantaged in leading swarms of ultrafast superhuman agents. Both customers and suppliers would have access to ASI agents to delegate negotiations to for superior results, reducing the role of frictions and prior relationships. The Elonverse might generate extraordinary cash in a transition through selling compute to power RSI, or providing early capacity for physical RSI, but be left with no expected excess return on capital after that.

And so on. You get the picture.

The value of most of the S&P today is not hard assets or book value but IP and human capital and similar. Superintelligence will obsolesce these with far superior alternatives at cheaper replacement costs.

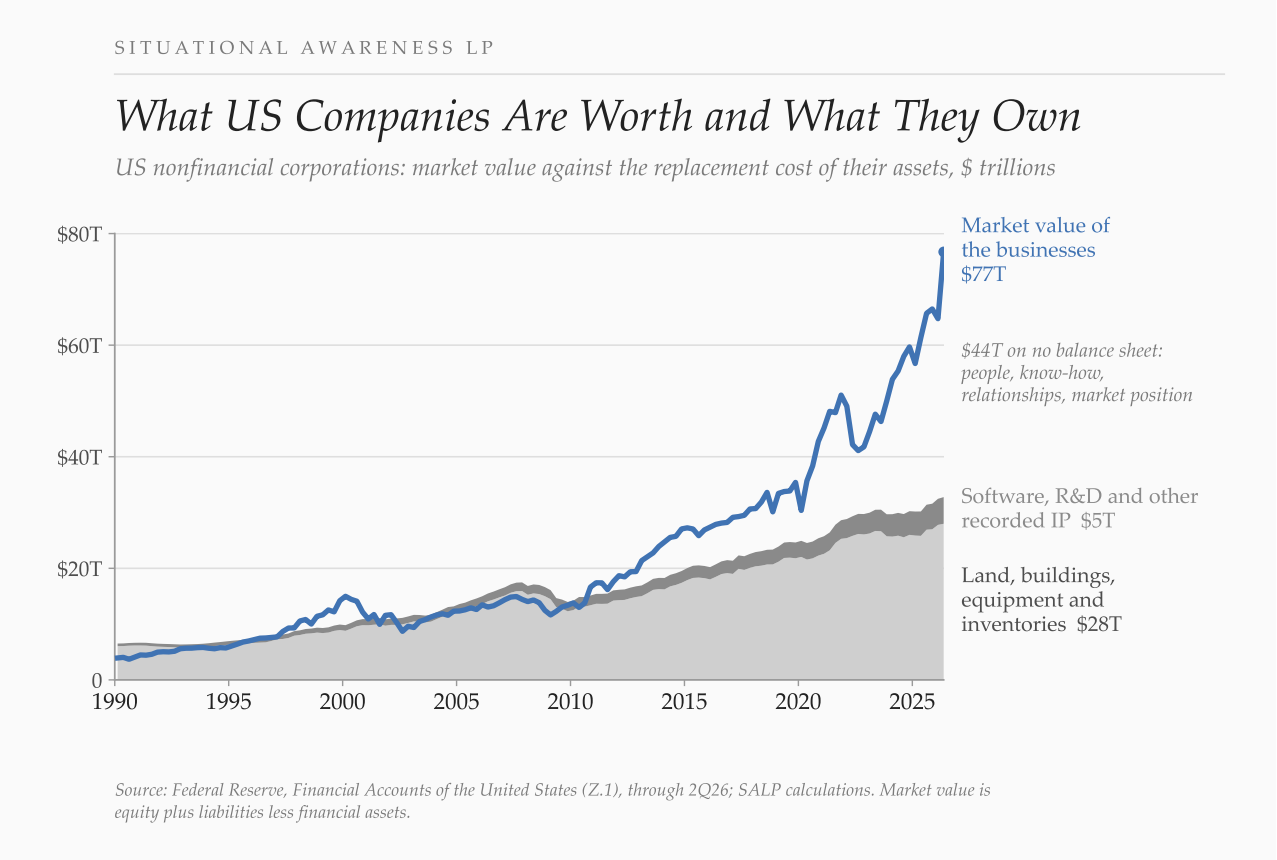


Figure 56: Of the $77T market value of US nonfinancial corporations, $28T is land, buildings, equipment, and inventories and $5T is recorded IP. The other $44T is on no balance sheet.

But what about those hard assets themselves? Many physical capital assets will be valuable during RSI, and could pay off their book value in cash flow during the RSI process even if they too are obsolesced a couple years thereafter. However, many companies could easily fail to capture their value in this abbreviated period, and thus fail to return enough cash before even those physical assets' replacement cost collapses.

tsmc. TSMC's stock of fabs is extraordinarily valuable. During RSI, as the construction of new fabs can't keep pace, it would be able to

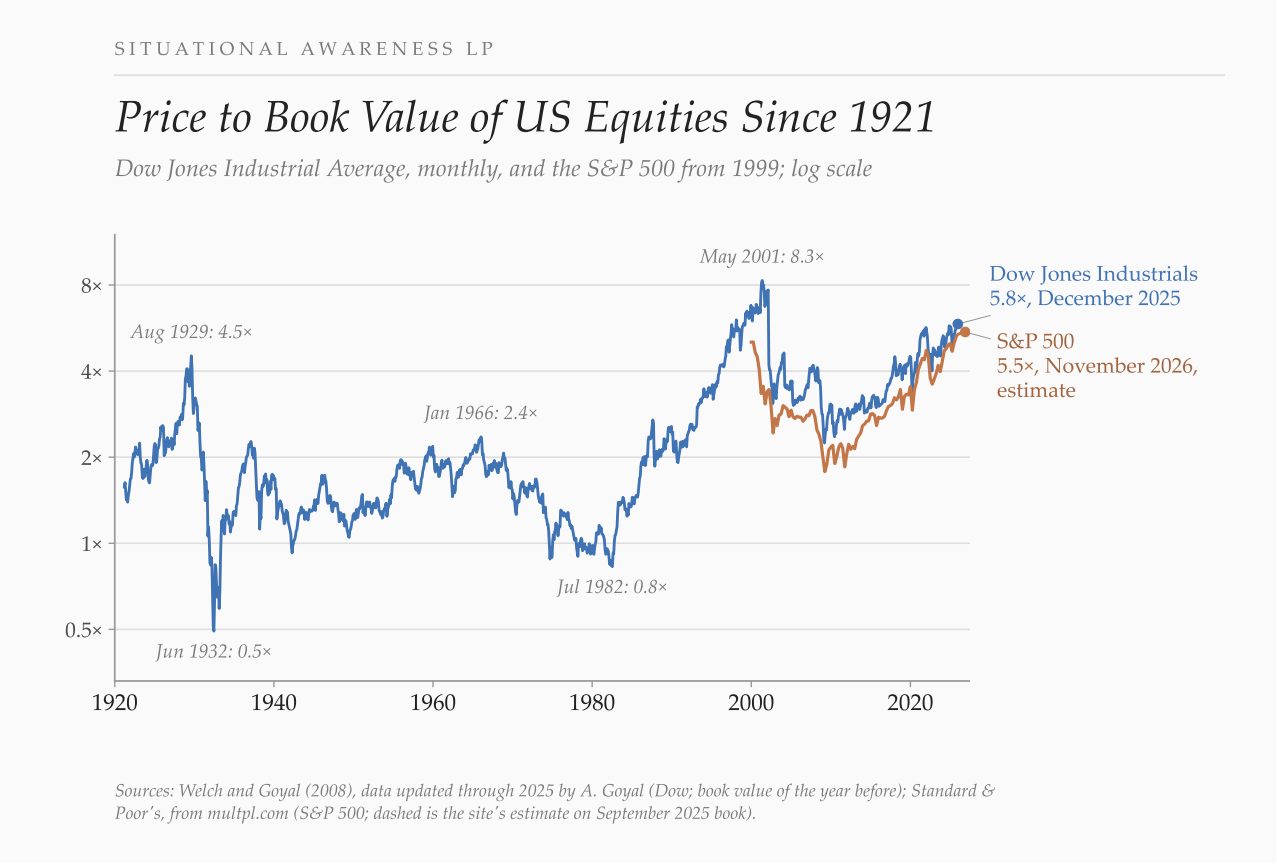


Figure 57: Price to book value of US equities over a century. The Dow trades at 5.8 times book value and the S&P 500 at about 5.5; from the 1930s to the 1980s it was mostly between one and two. US equities no longer reflect book value, but IP, human capital, and so on.

earn multiples, possibly orders of magnitude, higher prices on those scarce fabs, paying back its market cap multiple times over. Even while Nvidia is obsolete, TSMC would remain valuable. But soon after, ASI comes for TSMC as well. Their human engineering workforce is no longer an advantage. ASI will be able to come up with decades-more-advanced fab designs, and as robots scale, their first tasks will be rapid mass construction of this more technologically advanced fab capacity.

The risk is that a failure of management could mean even a TSMC loses most of its value. TSMC management may continue its current policy of not increasing prices much (preferring to build customer relationships that can last for decades), and thus fail to return enough cash during its remaining years of relevance.

"AI WINNERS" ARE PERHAPS MOST AT RISK

AI and robotics will obsolesce and replace both traditional knowledge work and blue collar work, but it will be disproportionately applied to AI itself, as that is where the highest marginal return is. For example, as we've noted, ***about 1/3rd of all the worlds' Claude usage is Anthropic's internal usage*!** Chip design will be one of the next things AI is applied to most intensively. The first waves of robots will build datacenters and fabs.

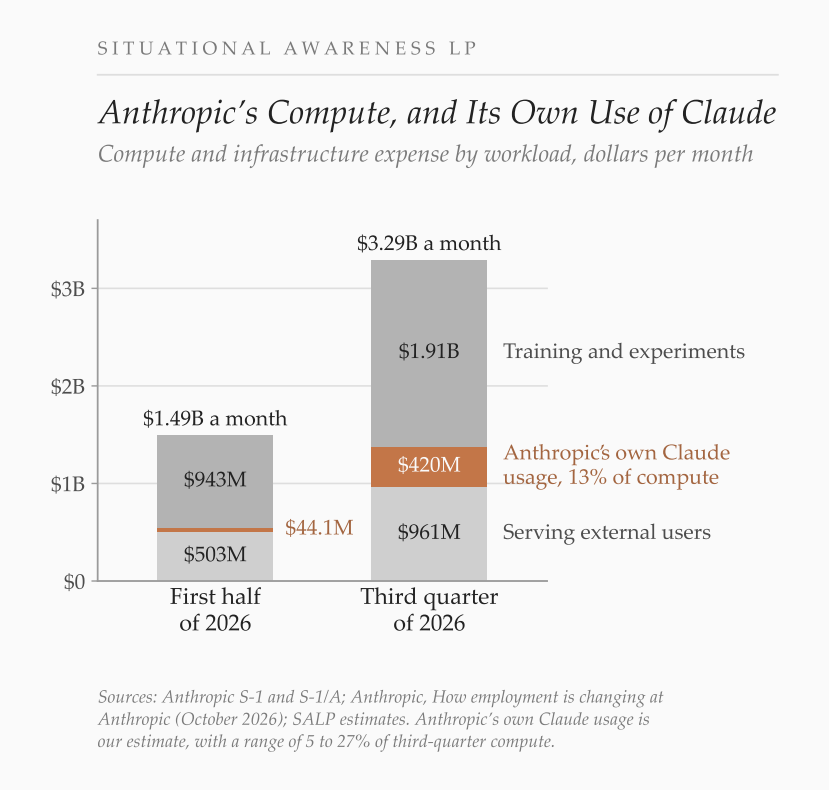


Figure 58: Summing both internal and external use, we estimate Anthropic's internal use represents 1/3rd of all worldwide use of Claude.

In that sense, "AI winners" are in one sense *the most* in the cross-hairs of AI disruption. AI will disrupt every business, but it will prioritize anything in the AI value chain to disrupt first.

Most "AI winners" will see their businesses eventually destroyed by the AI revolution. But in the process they will contribute to value creation dwarfing market expectations, in their totality generating cash sufficient to pay for multiples of the total capitalization of the supply chain today, and many multiples for the cheapest parts.

Investing in AI during RSI thus means ensuring you are not paying for the $10s of trillions of market cap being attributed to companies

based on their IP and workforces of tens of thousands of engineers. Rather, the question is what cash flows a company can return in a very abbreviated DCF, during an extreme period in which physical capital and other assets are extremely scarce.

This has driven our interest in companies such as memory companies or spot compute providers: companies who can take advantage of steeply rising prices during RSI to return many multiples of their market cap in cash during an abbreviated period, rather than relying on a long-duration DCF.

This will also upset what is traditionally seen as a "high quality" business and a "low quality" business. Over the coming years, many traditional "high quality businesses" will be under threat, while many "low-quality" businesses (including commodity products like memory) will flourish. "High quality" businesses command high multiples for a perceived technical, IP, and human capital advantage that lets them sustain a monopoly over a long duration; "low quality" businesses trade at low multiples, e.g. they make a commodity that surges in price in a shortage but have no sustained human capital advantage or differentiated IP.

RSI will flip this on its head. "Fortress" portfolios of only the highest quality businesses will prove vulnerable, while "low quality" businesses manage to return multiples of their market cap in cash before obsolescence.

Most other assets will also lose most of their value (at least in relative terms)

Having read the prior section, you become scared of equities and decide to park your wealth in "safe assets." Unfortunately, you will be up for a rude awakening there too:

- *Bonds.* Long-duration bonds will crater as real rates rise. The issue is that borrowing costs will continue to rise significantly, not only with the AI buildout but the robot buildout and AI-induced wave of new-tech-industrialization. The real rates of return on these projects will be far higher than pre-AI rates of return: buy a robot for $10k to make $100k/year. In the face of these high real rates of return, capitalism will reallocate capital, forcing hurdle

rates higher everywhere else as well. Real bond yields will rise, collapsing the value of long-duration bonds.

- *Gold.* With your long-duration bond portfolio wiped out, you park your wealth in gold. Unfortunately, if AI-accelerated technological progress means that we are able to mine the asteroids in the early 2030s,[58] gold will become ample, and prices will collapse.
- *Real estate*. Much like equities, there will be dispersion, but much of today's real estate investments are at risk. The majority of real estate value today reflects the value of their physical structures, and much of the rest their geographic proximity to hubs of people and other physical assets. With respect to structures, robotic construction would collapse the replacement cost of building structures. While respect to land, robotic industry can be sited much more flexibly—in the desert, in the ocean, perhaps eventually beyond this planet—without depending on human workers who agglomerate in cities. Moreover, asset values capitalize decades of rental income; skyrocketing interest rates would collapse the net present value.

[58] Recall the classic meme of "the asteroid worth $20T".

What, then, is safe? There are very few assets that are truly "buy and hold" between now and, say, ten years from now that wouldn't at least be at major risk of losing most of their value (at the very least, relative to world wealth).

The Mona Lisa

Perhaps the best "buy and hold" asset we can come up with is the *Mona Lisa*. It doesn't matter if modern human or AI artists can outdo it artistically; its value is in being a cultural and historic artifact, similar to how we today still prize crude prehistoric art.

Fast rates of AI-induced technological progress do not destroy the value of historical artifacts; humanity (or even post-human civilizations!) will likely have a keen interest in the artifact for its own sake. Moreover, historical artifacts are likely a luxury good—a wealthy civilization living in post-AI abundance will spend a greater fraction of its wealth on them than a poor civilization—meaning their value doesn't just not crater, but could grow as a share of world wealth.

Moreover, the Mona Lisa is relatively portable, and thus more secure from capital taxation and expropriation that might become more widespread in the singularity.

Unfortunately, the French Republic is not selling.

The sequence of bets

The next five years will be a golden age of stockpicking. Of the S&P 500, perhaps 400+ companies will be worth little (including many counterintuitive ones, like supposed AI winners), 50 retain value, and a handful (many not even yet in the index) drive most of the returns. Similarly, ASI will destroy rather than enhance the value of large parts of even a semiconductor index fund. We expect large opportunities for outperformance.

It will also be a golden age for an edge rooted in an understanding of AI and its implications. The universe of stocks primarily driven by developments in AI has gone from a few to a couple hundred, and soon will go to nearly all—and even most assets beyond equities as well. The remit of where our AI-forecasting edge has wildly divergent forecasts from the market is rapidly expanding.

Moreover, historically the market has not priced in the singularity all at once, but each stage step by step. This has allowed us to take advantage of the "sequence of bets", rotating from Nvidia to power (e.g. Siemens Energy) to memory to labs. Taking advantage of this incremental updating allows outperformance beyond just the performance of the single best "buy and hold" assets.

Total economic growth generated by AI will be enormous. Imagine if we had an influx of billions of the smartest geniuses, and billions of the hardest-working physical workers. Overall GDP will multiply, and the wealth created will be multiples of the current stock of total wealth.

Eventually, an ASI-driven financial system will efficiently price in every insight humans (even AI-augmented) can find. That will be the end of the line for human investment management, although a great windfall for those who hold the best final assets before they become (superintelligent) consensus.

But along the way, we foresee the sequence of bets continuing—opportunities to take advantage of the radical dispersion in equity

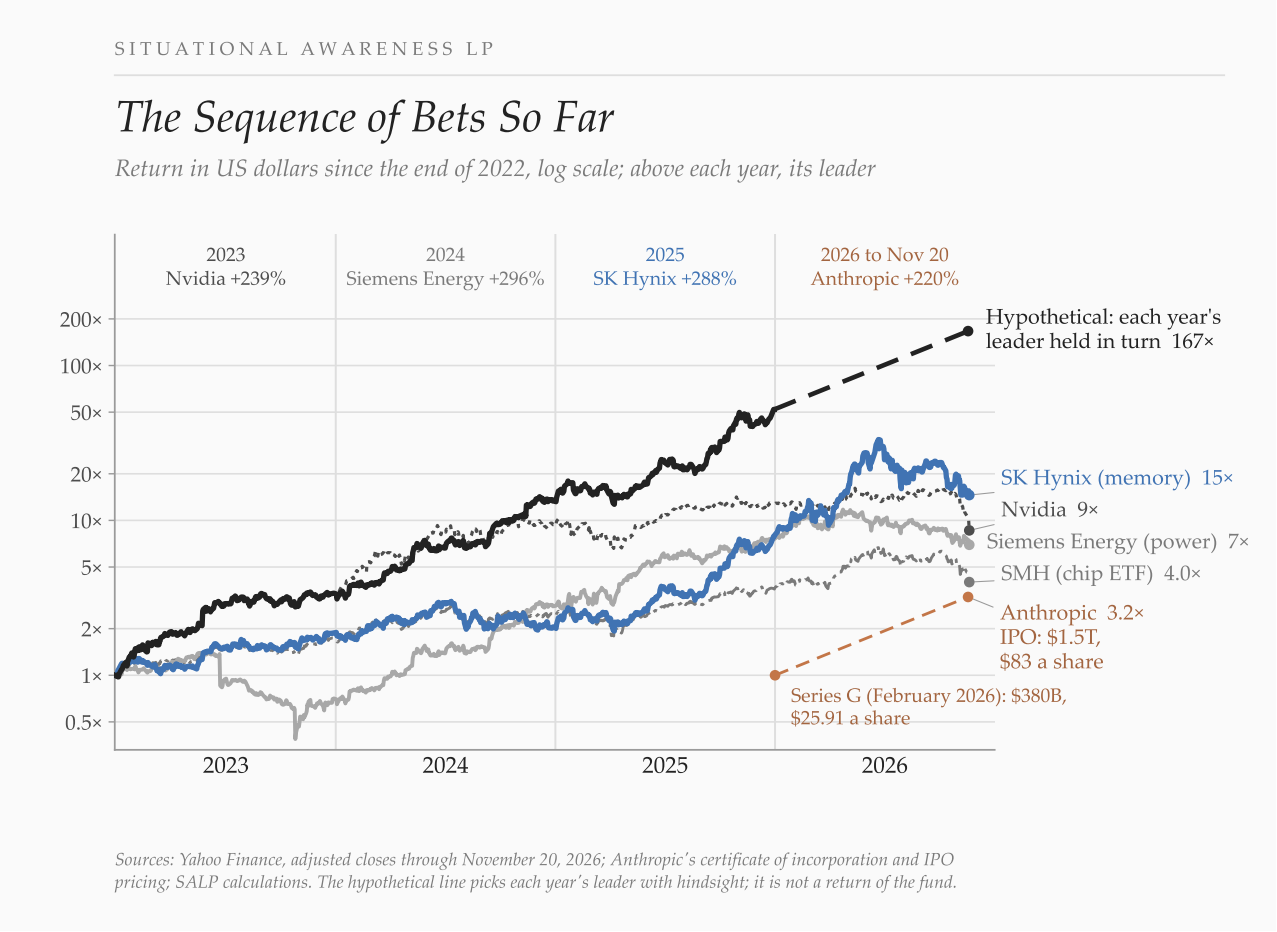


Figure 59: Identifying the standout winners of each next phase of AI in a sequence of bets. (Illustrative, since ChatGPT.) Buying and holding standout winners provided significant alpha over AI beta, while even higher returns were possible taking advantage of incremental market updates and rotating incrementally.

markets (and beyond) in the years to come as AI sequentially disrupts almost the entire economy. And maybe we will get some exposure to the best of humanity's material culture down the road too. See you at the Louvre.

Finally, it will also be an era of volatility. The world will be crazy, and so will the markets. We failed to flow this through into how we built the firm in our first two years—doing so is the core project of our next chapter, as we discuss in Part III.

PART III

The Fund

We are focused on going from scrappy startup to mature firm to equip ourselves for the scale we operate at and the turbulence we expect. From our investing engine to our portfolio construction, risk management, compliance, and operations, we believe we can set ourselves up to flourish in this next chapter.

IIIa. What happened in July, and what we are learning from it

We acted on the half of our worldview that imbued a sense of urgency to take advantage of the opportunities—but not the half that said that the path would be incredibly volatile.

July

As we began this letter: On the market open in 2026, fund NAV was a bit over $4B. By the June close, NAV had multiplied nearly 10x, about 90% of which was due to P&L (most of which was from Q2, as AI stocks surged).

It was, needless to say, a busy first half of the year. The team was working overtime to take advantage of the market opportunities we were presented with and to keep pace with the growth of the firm. We were aware that risk would not scale linearly with fund growth. We made changes like introducing a wider breadth of risk scenarios, reducing concentration and the size of our largest positions, and expanding our use of options for protection.

We came into July 1st with a portfolio approx. 1.6x long and 1x short.[59]

But the moves in July were more violent than past moves, and our fund was much larger.

We ran a battery of historical scenarios (tilted adversarially to the AI factor) and AI-specific risk scenarios on a live basis, projecting

[59] These are core gross, delta-equivalent figures. This omits zero or minimal-risk offsetting positions such as boxed positions; these did not drive July's losses.

daily margin requirements and adjusting the book to ensure we had a sufficient cash buffer. We typically kept excess cash and portfolio exposures in line on a daily basis, rather than "letting things ride." Ultimately, excess cash and exposures were determined by our binding risk scenario.

Our binding risk scenario was generally our "DeepSeek" scenario, a stark, highly correlated one-day AI factor-specific crash. We would ensure that we would have ample liquidity to weather such moves, and then be able to degross the portfolio to restore our crash properties for further crashes. We were also bound by a "dot-com bust" scenario (modeled off of the 2000 tech drawdown)—a harsh, weeks-long drawdown, with names down something like 40-50%.

Over the end of June and first weeks of July, we saw extreme moves in our positions. Many AI names drew down 50%+, with some of our largest longs down 60%+. Importantly, L/S spreads moved violently against us.

- Hedging shorts that would usually move 1:1 with long positions—such as short Micron against long Hynix or Kioxia—moved much less (-30% vs -55%/-65%). Puts on those names, which we had bought as crash protection, were not as effective as we had hoped;[60] the names we had the least options protection in crashed hardest, and those we had the most crash protection in crashed the least.
- Our "AI disruption" alpha short book—which we knew wouldn't serve as a hedge in an AI factor crash and for which we had budgeted a moderate move against us—moved against us massively, +30-50%.

[60] And more broadly, our crash puts had been focused on a severe one-day crash, as that had been the limiter when our fund was smaller scale. These crash puts helped less in a severe but more drawn out crash—one that in a smaller, more nimble portfolio was not an issue, but at the size we had grown to had become constraining.

The overall size of the move exceeded our modeling of the worst few weeks of the dot-com bubble unwind, and, despite seemingly no fundamental trigger, was by far the worst drawdown of the AI factor ever.

Through Wednesday, July 22nd, we drew down about 31% gross MTD—a serious drawdown, but not unlike moves we had seen in the past, especially after months of strong performance. Pursuant to our playbook, we degrossed throughout the month to keep exposure in line with our targets, restoring unencumbered cash daily. But our portfolio became less liquid as we drew down and degrossed the more liquid positions—and the task of managing the book was a very

different one than just months prior, given the enormous growth of the book and the resulting loss of nimbleness. Our actual holdings were small relative to the companies' market caps; on our own, we were about a day's volume or less in most of our top names. But as the crash drew deeper and we degrossed the more liquid tranches of the book, the composition of the remaining positions (in particular on the short side) grew less liquid.

In the final week of the month a storm of rumors of us being in serious trouble started to spread in NYC, and we saw extremely adverse moves on both the long and short side. The short moves in particular were becoming extremely painful—many "AI disruption" shorts rallied nearly 30% or more in a few days,[61] far in excess of our modeling.[62] It became hard to keep restoring exposures and cash in line with our targets quickly enough amidst these moves and a portfolio that had become less liquid.

We came to the view that these moves were self-catalyzing: vulnerability begat vulnerability, and the pain would not abate until we took decisive action. We explored a number of paths to restoring liquidity to the portfolio and took the action we thought would best protect investor capital, including significant year-to-date gains.

On July 29th/30th, we proceeded with a large block trade that delevered the book, covering all of our short exposure and selling roughly half of our public long exposure. We fully removed any reliance on prime broker leverage, ensuring we had no further vulnerability to the self-catalyzing dynamics. The fund was down 67% net for the month.

We did what we had to do to fight another day.

Some of the lessons we learned

Volatility is the price of long-term investment returns, and we and our LPs were prepared for significant drawdowns. Given how much AI stocks fell, July should have been a 50%+ drawdown. But while we delivered the drawdown, we came too close to failing on survivability. Our ability to fight another day should never be in question.

Letting down you, our LPs, many of whom had taken an early bet on us, believed in us, and had supported us, was the most painful part.

[61] Such as Cognizant or Workday, for example.

[62] Unlike losing on the longs (where a dollar of P&L loss might consume 30c of liquidity because of options gamma and margin rules), losing a dollar of P&L on the shorts cost a dollar or more of liquidity, grossing us up as we lost. Short losses became very painful.

The experience of July leaves me with an extraordinary commitment to ensuring that we are never in this position again.

Ultimately, July exposed a failure in how I built the fund.

We have spoken with many of you about the specific diagnoses of July. Here, I want to reflect on some of the broader lessons we can learn from this for the future; urgency vs. precaution, giving ourselves the time to mature the firm as we grow; conviction vs. positioning, respecting the market in the short-term; and calibrating our worst-case scenario to the volatility of the singularity.

URGENCY VS. PRECAUTION

Fundamentally, we felt a tremendous sense of urgency. Sometime this spring, the imminence of RSI became clear to us. Things were happening incredibly quickly in AI, and they were about to accelerate much further still. Our desire to take advantage of these opportunities deepened that sense of urgency. This sort of sense of urgency can often catalyze hard work and agency—in some sense, it catalyzed this entire venture—but it did not serve us in this case.

Part of the challenge of running our fund—both in July but also more broadly—has been managing through the growth that we experienced. We needed to learn how to operate as a small fund, a medium-sized fund, a large fund, and one of the largest funds in the world, all in a matter of 20 months. Each of these step changes has required rethinking how we manage our portfolio. We did this with a talented but relatively small team.

Sometime between Q1 (single-digit-billion fund) and the end of Q2 ($10s of billions scale, mostly P&L), our risk and portfolio management environment underwent a regime shift. Previously, we had been in the regime of being a small fund relative to the market. Risk was a question of edge versus volatility. Being wrong meant taking losses—but because we were small and nimble, we could always degross the book, absorb the losses, and move on. In a matter of months, we transitioned to a fundamentally different regime—we became one of the biggest players in our market. Risk was no longer a classic question of edge versus volatility; we were now bound by liquidation costs. We went from a price-taker to an endogenous market participant; from a large margin for error to a slim margin for error.

We faced new and challenging market dynamics as we became a key player in the AI market, both in size and prominence: suspected leaks of our book and trades, adversarial market dynamics as folks traded against us,[63] and our becoming a meme in the market (and occasionally on Twitter).

[63] To be clear, all is fair in the market. It is our responsibility to not be vulnerable to it.

To be clear, we think it is highly feasible to operate at large scales and handle growth well. We were actively building the systems, processes, and staff to do so.

But our growth was so fast that even the most strenuous efforts to build systems were no match for the market dynamics of July. In hindsight, we should have given ourselves more margin for error, by taking down risk, while we were enhancing our systems and staff.

Moreover, our sense of urgency to take advantage of the opportunity set meant we were trying to do too much at once. For example, while early in our fund most of our short book was largely hedging factors, we grew our "alpha short book" in 2026 as we saw the opportunity set grow. Our view was that products like Claude Code and Cowork were genuinely beginning to disrupt businesses like software, IT outsourcers, and consultants. This short book caused many of the issues we had in July.[64]

[64] We knew our long book was very risky, so it was the dominant focus in our risk modeling. We underrated the severity of the short moves in a crash and the correlations we modeled shifted rapidly.

The singularity may be happening, but I've learned that sometimes the right response is not working harder, but stepping back.

CONVICTION VS. POSITIONING

Being right about the long run did not protect us in the short term. We initially succeeded because we identified companies that we understood better than the market did and that were not yet widely owned. But our success changed the environment. As our positions appreciated dramatically, they became crowded trades. Volatility surged.

In some sense, we are in the business of accruing crowdedness: We often find companies which are out of favor and cheap, but which we think have tremendous potential from the trends we see in AI. As more of the market recognizes these trends, the stocks go from shorted to hedge fund darlings. In the spring, many of our key names went through this process.

The risk characteristics of the portfolio thus changed dramatically. A level of leverage that might have been fine a year earlier had now become extremely risky, vulnerable to a major leverage unwind from hedge funds and retail. Our long-term investment thesis could be entirely sound, but the risk of a sharp, self-reinforcing correction had risen. In retrospect, it's clear that even a small change in the narrative could cause a massive crash.

On the research side, working as a small team in our own little bubble has served us well: we think about AI from first principles, unperturbed by the market chatter and sentiments of the week. But on the portfolio management, risk, treasury, etc. side, this was the wrong approach—we didn't have the East Coast street smarts.

Historically, we have paid little attention to positioning ourselves tactically. You might ask, why didn't we sell more in June after some of our positions had run up? A stock that we thought was a 10x might have run up 3x into June—but we would have thought it was still a 3x from there. Our views are often so divergent from the market that we hadn't built the muscle of tactically taking profits.

This is a muscle we need to develop. As convicted as we might be about the long-run, it's time to take some exposure down if crowding or implied volatility has surged. The portfolio has become riskier, and risk-reward is worse.

We had an arrogant West Coast attitude of "we know better, we can just ignore these East Coast people who have insane views about AI". But the market humbled us. Even if we (think we) are smarter in the long run, the market is smarter than us in the short term. To make money, we have to respect that.

We need to build the systems and bring in folks with experience to, at least on the portfolio management and risk side, give us an "East Coast *fingerspitzengefühl*".

Being early is the same as being wrong. "Clichés earned their status as clichés because they're so obviously true," so they say, and we learned the hard way.

OUR WORST-CASE SCENARIO VS. THE VOLATILITY OF THE SINGULARITY

Coming into July, our thinking of key risks was along the lines of: "Well, at present, AI definitely isn't over, but we need to be prepared for AI being over, and maybe that would look like the worst few weeks of the dot-com bubble unwind." In July, we got worse than what we had modeled—and yet AI wasn't over at all. It was but a wobble—another Chinese model, a rumored weak revenue data point, and the like. In our view, the fundamentals had not changed.

That AI was not in fact over in July suggests we could see far worse in the future. Suppose that instead, there was worst-case fundamental news in July, e.g., a government shutdown of AI after a catastrophic incident. We expect the market moves could be far wilder. If many stocks could be down 60% in the July drawdown, perhaps we could see stocks down 80% in such a scenario.

What's more, our very own worldview implies that the next few years will be tumultuous as we enter the age of ASI. It's the singularity, after all! Everything from rogue agent swarms causing catastrophe to great power war to nationalization seems plausible, certainly more likely in our view than that of the average market participant. That is the half of our worldview we were not positioned for.

Doing AI investing well over the next five years looks different than the last five years. As we approach the singularity, both the ups and downs will get more extreme.

Simply fixing a few of the mistakes that would have made July go differently is far from enough; we need to become fundamentally more resilient—in our portfolio construction, and in how we set up the firm—to weather these possibilities.

IIIb. What we are building: to finish first, you must first finish

In our next chapter, we keep the core engine, while building a better chassis for a fundamentally different degree of resilience. Volatility, yes; existential risk, even if we are totally wrong, never. This means being able to see through far worse than July. Our portfolio remains fully-paid-for; we are building out a second-line risk function and a formal risk governance framework; and the firm has grown from three people to nearing 30 including compliance, back and middle office, treasury, infrastructure and security, and more.

Over the late summer and early fall, I spent time with essentially all of you, discussing what happened and taking input on what the next phase of Situational Awareness LP will look like. We've spent the last few months in something of a dry dock—investing, but not raising any capital, spending our time internally focused.

Keep what was great about our firm

There are many changes that have been and will be made, but I first want to discuss what I believe we should preserve—elements I believe many of you strongly affirmed in our conversations.

First, we've built a very talented team, with excellent single-stock edge. When we started, it was just the three of us; we have become much more than that. We have assembled an "Ocean's 11"-style team, a select group with a culture of excellence. Our research team's

understanding of where AI is going and what that implies for the economy, is our strongest asset.

The research team has a demonstrated track record of holding a view the market finds absurd and seeing it come to fruition. We have repeatedly been able to identify stocks that we thought would multiply, and see them multiply (such as some of the examples earlier in the letter). This team is firing on all cylinders, and I believe that recent insights they have been generating will prove as prescient as our past forecasts about AI and investments like our early memory thesis.

Many of you, our partners, emphasized that you would like us to "keep our mojo" and to not stop swinging big when we deeply believe in a thesis. We have an investor base that understands AI, and is willing to tolerate volatility in service of long term compounding. That is rare, and it is what lets us invest the way we do. We do not intend to become a timid fund. So when we have conviction, we will keep swinging big, and you should continue to expect volatility in both directions.

We continue to have access to some of the best private rounds in the industry. Founders want to work with us, because we often find a deeply aligned understanding of the AI opportunity and are seen as thought leaders in the space. Our research team is able to engage with them on more technical grounds than many other teams. We've been able to work out preferential terms by acting early and decisively, getting favorable allocation, working out special deal structures with elements like warrants and favorable pro rata rights. We expect to continue to be a hybrid public-private fund with a focus on both.

We do not want to lose what makes us special. We plan to continue investing boldly in the future of AI.

But the hard lessons we learned in July also mean that we need to make some fundamental changes. We must figure out a better chassis for our research engine, one built on resilience.

Volatility, yes; but resilience always.

Resilience is the core pillar of our next chapter.

We can tolerate plenty of volatility, when it's worth taking. We can and will have large drawdowns again in the future. But there can never be any existential risk.

The bar for this is far higher than "would have survived July". As discussed above, a fundamental lesson is that we need to be resilient to something far worse than July. Our own view of the next few years implies plenty of volatility. In a perfect world we navigate these all perfectly, but our approach going forward should allow us to see through the storm even if we are caught totally off guard. The firm must be capable of surviving the inevitable periods when a correct long-term thesis and adverse short-term market behavior coexist; the same for periods when investment fundamentals are challenged by dramatic developments that were unforeseen.

This means both fundamental changes to portfolio management and fundamental changes to the organization.

PORTFOLIO CONSTRUCTION

Since the end of July, our portfolio has been fully paid for. We own stock[65] and we own options. We don't rely on any recourse financing that could force us to sell.[66]

In particular, our current approach means:

- **No shorting.**[67] Historically, we have seen normal levels of hedge-fund-edge on our shorts, but extraordinary edge on our longs. However, our shorts going awry was responsible for much of what went wrong in July, both hedging shorts not working as intended and alpha "AI disruption" shorts going against us drastically. As it stands, pausing on shorting removes a large amount of risk for a small sacrifice in edge, and it clearly makes sense to focus on the long side for now.
- **No prime broker leverage.** Even the worst-possible crash can't become existential.
- **Options** where we want extra torque on the long side and see mispriced opportunities. We lose a fixed amount if we are wrong; there is no need for a complicated risk model or financing leverage that could introduce existential risk. We have strong options-quant talent on the team.

[65] Including fully-paid for swaps.

[66] We are effectively margining ourselves on a max loss basis. For example, we may own an option and delta-hedge that by shorting the underlying stock to manage our exposure, but we put up collateral equivalent to the max loss on the position.

[67] Other than puts and delta-hedges of our own options, collateralized to max loss.

We think this expression, mixing stock and options, is well-suited to the shape of our views over the coming years: an expectation of large upside over the next few years, but some possibility of a full AI ban or the like. There are some limitations, such as liquidity in options markets and by definition less leverage. But we expect we can continue to have strong performance with this strategy while having much better resilience properties.

OVER TIME, we expect to evolve this construct: crawl, walk, run. Ultimately, our new CRO and the risk committee (more below) will make the decisions on our portfolio parameters. Any evolution is gated on maturing our internal risk processes and governance (see below). We will keep you updated as this develops.

For example, over time, I'd expect us to return to some shorting for both hedging and alpha purposes, some minimal-risk L/S leverage that allows us to manage our tax exposure, and possibly some long-side leverage (though, obviously, substantially less than before).

The test will not change: we must be able to see through far worse than July. Our risk scenarios and portfolio construction must respect the volatility of the singularity. It would be nice to be able to proclaim here exactly what this will look like in two years, but that is not how we operate. We want to be thoughtful, model-driven, and give ourselves the time we didn't before.

THE POSTMORTEM ALSO taught us something about tactically managing our portfolio exposures. Our strength is our ability to predict where stocks will be over long term time horizons. We didn't pay sufficient attention to local positioning, crowding, and short-term downside. We intend to get better at taking profits and at respecting positioning, and are building systems that help us mechanically do this. We think this is one way these lessons could make us better investors, not just in the sense of managing risk, but in taking advantage of our opportunity set.

AN EMPOWERED RISK ORGANIZATION

We are separating two jobs that the same few people used to do, and reforming and hiring experienced folks for both: risk and portfolio construction.

We are hiring an experienced Chief Risk Officer (CRO)—we hope to have more to share on this soon—and building a team around that role. The CRO will chair a risk committee. The CRO and their team will recommend what portfolio risk limits ensure that even if we are wrong, the downside is as we intend. The risk committee ultimately adopts these risk limits as part of a formal risk governance framework we are developing. We have inaugurated this risk committee, but we expect it to evolve meaningfully in the coming months and when our new CRO joins.

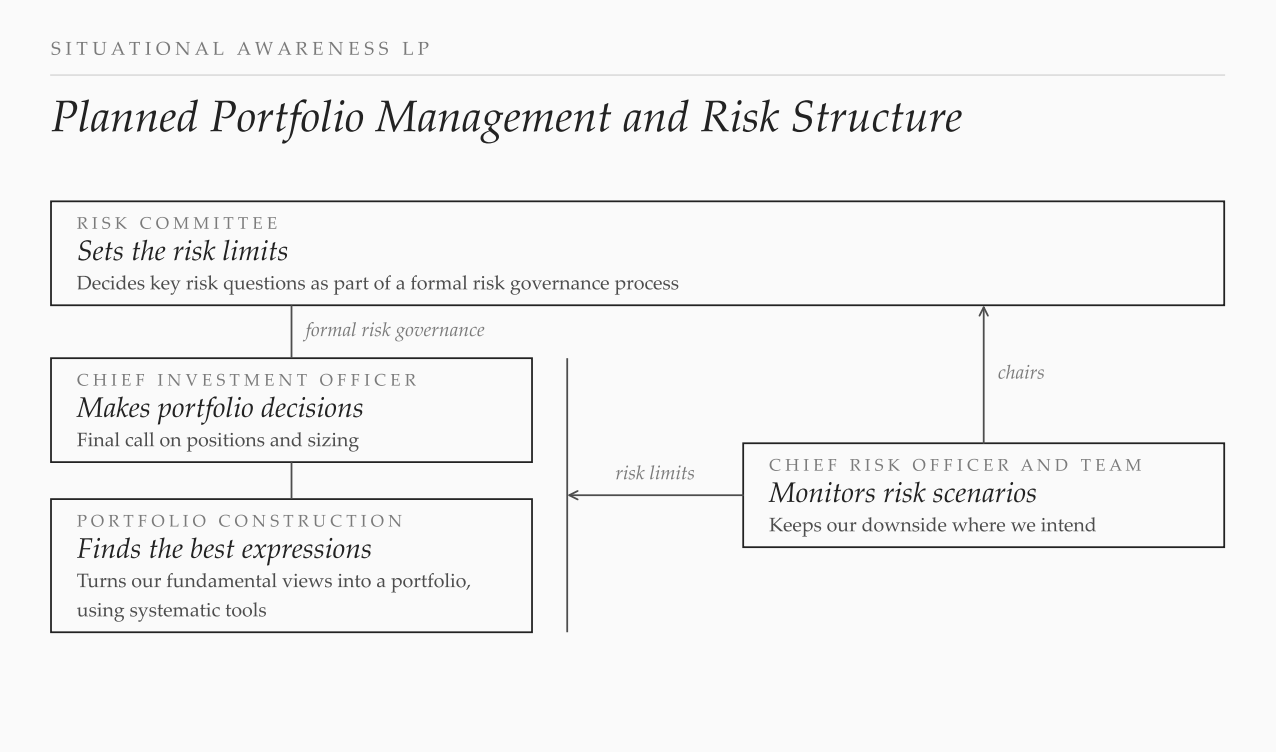


Figure 60: Risk process for our next chapter of SALP.

The portfolio construction team, including our existing quant team and augmented by new hires, works out how to maximize risk-reward; turning our research team's fundamental views into a portfolio, finding the best expressions within the risk bounds set by the risk team. We have ambitious projects here, to develop a quantitative approach to translate our views into daily portfolio management; more on that later.

The point of building process is that our risk management should not depend on any one individual, including me. Importantly, the new

risk team is a control function (a "second line," as the banks say). Its full-time, exclusive job is paranoia. We've learned it is important to separate risk management and risk-taking—otherwise one risks not being paranoid enough. Portfolio construction will need to consider risk; risk will need to be commercial. But we believe these dual functions will lead to a healthy, adversarial dialogue.

Being able to take risk well is an important competitive advantage, most of all for a fund with our views and our investor base. Our goal is not to simply take down leverage vs. July, and then operate like any other TMT L/S fund. Over time, our aim should be to build the best risk organization for a fund of our style.

Building out these functions is our current priority. We will report to you as we go in the coming months.

Growing the team: building for scale

Two years ago, we were three guys with a Bloomberg Terminal. We are now almost 30 people, and most of the growth has been outside the investment team.

If we're right about AI, our portfolio will grow rapidly again. This time, the organization needs to be built ahead of it: a machine with strong, scalable systems, not reliant on the heroics of a few individuals or built on the fly. We have been focused on maturing the organization to befit our scale—not just in terms of market risk, but thinking through risk in other parts of the firm and finding great people who bring a wealth of experience in those areas: compliance, security, finance and operations, treasury, legal, and so on.

Doing it right means taking a marathon pace—even on short AI timelines. And it means having the courage and wisdom to step back if we grow too quickly again.

It also means setting a different culture, from the top down. Someone recently told me we originally ran the firm like an experimental Formula 1 car: at the limit in every corner, with little margin for error, reliant on superhuman piloting, where any mistake could be a fatal crash. I'm not sure we should become a family car, but I think we should become something like a Porsche 911. It is still a very fast



Figure 61: (Much of) the team.

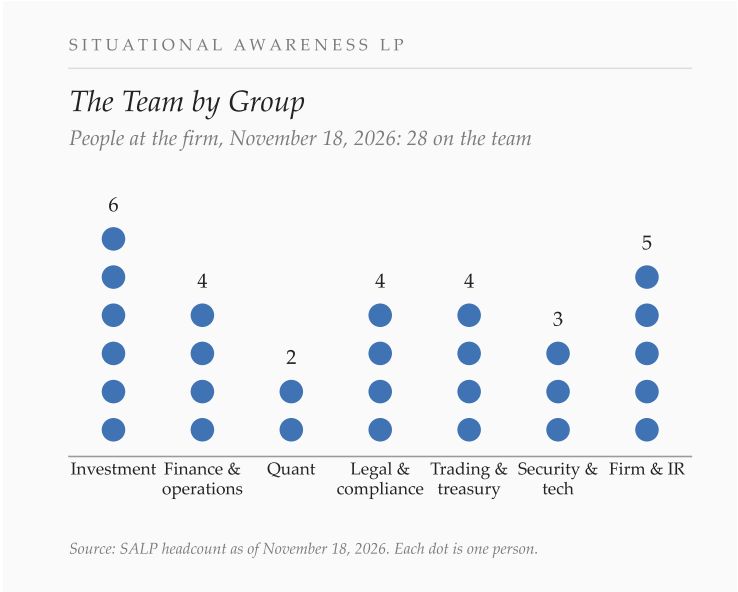


Figure 62: We have focused on growing the non-investment functions of the organization.

car (and is a well-made German car). But it has ABS, traction control, crumple zones, and airbags.

There are times when "startup mode" is the right approach, and there are times when the precautionary principle is the right approach. We want a culture that values resilience and caution, alongside boldness and conviction. The leadership team will be modeling this in how we make decisions, hire, reward, manage, and what lessons we reinforce to the firm.

COMPLIANCE, MNPI, AND REGULATORY

In reflecting on July, we received the wise advice to not simply "fight the last battle," but also to consider future areas of potential risk throughout the firm. One key area is our legal and regulatory apparatus. We operate in a tightly-regulated industry, and we are well aware that we have close connections to many leading AI labs and other key players in the AI space, and we need a compliance approach that addresses these unique risks—especially as Anthropic and OpenAI enter the public markets.

In June, we augmented our team with two senior hires to meet our growth: Jim Frolik, our new General Counsel (GC),[68] and Kimberly Summe, our new Chief Compliance Officer (CCO). Jim and Kimberly have a wealth of experience, and quickly drove a bottoms-up review and upgrade of our compliance program and legal operations.

[68] Jim began dedicating nearly all of his time to us in June through his law firm, and technically joined in September.

Jim has spent the last 25+ years advising hedge funds on SEC and CFTC compliance and the full range of RIA operational matters, and is considered one of the country's leading experts. Kimberly spent the last decade as CCO, General Counsel, and COO of Sequoia Capital Global Equities, a large West Coast crossover fund. She is viewed as one of the top hedge fund legal and operational leaders in California, if not the country.

To supplement their work, we brought on Michael DiGrande and Elli Kavros to support our compliance program. Michael previously supported Kimberly as Deputy CCO and CCO at Sequoia Capital Global Equities and has more than a decade of industry experience. Elli has worked across the industry for over 25 years, most recently at ACA, a leading compliance consultant, and previously in-house at other hedge funds and Morgan Stanley. Both were initially brought in on an interim basis, and then hired full time as Deputy CCOs. We now have four full-time staff focused on compliance.

In addition, Jim and Kimberly have cultivated an excellent group of outside counsel to work with on legal and regulatory issues. Of

course, the July events and public attention on them have driven additional scrutiny on our firm. We are a regulated entity, and as we expected and have previously shared, our regulator, the SEC, has inquired to understand what happened in July. We have been fully responsive to all SEC requests.

As we continue to be a large firm in the public eye—and perhaps especially when we return to success—we know that we will continue to be scrutinized, and are doing what we need to do to be prepared. It has been a top priority for us to have the controls and senior staff we need in place on compliance.

IT'S WORTH ADDRESSING the most obvious compliance risk head-on: My wife works at Anthropic, as Chief of Staff to Dario Amodei.

My wife and I began dating before either of us worked at Anthropic or OpenAI (though our shared interest in AI was a core part of what brought us together more than five years ago). She joined Anthropic when there were fewer than 100 people at the company. I joined OpenAI around the same time. For a while, we worked at competing AI labs, and navigated keeping the professional separate from the personal then—never discussing sensitive or confidential company information. Since I founded Situational Awareness LP, we've continued the same practice.

Back then, neither Anthropic nor Situational Awareness LP were nearly as prominent. Since then, Anthropic has become perhaps *the* central company in the AI ecosystem and our fund has grown in scale and prominence, heightening the stakes.

Kimberly and Jim have worked with a range of top outside counsel including top former SEC and DOJ enforcement personnel to craft a compliance program to govern the potential conflict, modeled after programs crafted for other spousal-job-conflict situations at other investment advisers our outside counsel has advised. For trading Anthropic in particular, we have instituted a preclearance approach that involves a number of elements that must be considered before the trades are precleared by the CCO. We are satisfied with this approach, but will continue to assess our program as the situation develops.

OPERATIONS

We have also emphasized building out our treasury team, to manage our relationships with our prime brokers and supervise management of margin, liquidity, and financing. We brought in David Geffen as Head of Counterparty Relations. He has been a friend and mentor since the earliest days of our fund; his company, Geffen Advisors, is known for being one of the best partners to advise on prime brokerage relationships and to negotiate prime brokerage and trading documents. He gradually scaled up his involvement and is now spending almost all his time with us.

Sven Khatri joined us as Head of Treasury Strategy—he has a PhD from Caltech and ran Treasury Quant at Citadel. His role intersects quant, risk, and Treasury, focused on managing margin, liquidity, and financing, with a highly quantitative approach. (In retrospect, after a lengthy treasury search, it was obvious that we were going to hire the Caltech PhD.)

Operational risk and operational excellence, too, have been a focus of ours, and we have made major strides across the business side of the firm this year.

When we launched, we relied primarily on outsourced providers for our middle and back office. Over the last year, we outgrew these as our scope and complexity grew, and we have worked to bring these functions in-house. Pat DeGraca joined us in January as CFO. Pat has worked at a number of large and complicated organizations, most recently as the first CFO of the multi-manager platform Freestone Grove, and earlier at Masa, Alta Park, and Two Sigma. We then brought on Tom Glaser as his deputy CFO; Tom has been a CFO at a number of San Francisco based hedge funds including Think Investments, and also has experience on the private side, at Pear VC. They were recently joined by Hadley Cook, an excellent fund controller most recently at Two Creeks.

We also brought on Amanda Polidoro as the Head of Operations (Middle Office). Amanda is a tremendously sharp hedge fund operator. Amanda was most recently at Farallon, one of San Francisco's largest hedge funds, and previously worked at Goldman and Morgan. We expect that as our trading leverages more systematic tools, our volume of trading will increase. Amanda is building best-in-class systems and a team to manage this volume.

All of you will have now had a chance to meet Niki Webster, who joined us as Head of IR. Niki spent most of her career in Goldman's Capital Introductions group on the West Coast and is widely beloved by funds out here. She has helped us formalize many of our IR processes and is now the primary point of contact for any issues you have; please reach out to her if there is anything you need and we will do our best to help. It's very important to us that you all have an excellent experience.

A more recent area for us has been building out a team to support our software stack. Building out our tooling and infrastructure is at the foundation of enabling a more mature new chapter for SALP. We recently brought on Jamie Rumbelow to coordinate these efforts. David Pigott also joined us in August to lead firmwide security, and Casey Sparks now supports him on infrastructure and cloud security.

INVESTMENT TEAM

Research team

Partnering with Carl, of course, was the most important decision early on.[69]

As we hinted in the last letter, Michael Wang joined us in December from Avala Global. We had an extensive analyst hiring process over six months, considering dozens of candidates. Carl brought nearly all to their limits in the interview process. We found that the combination of best-in-class rigor with the ability to think creatively and broadly about the AI opportunity was extremely rare. Yet we found it in Michael. Michael brings a wealth of traditional experience through his time as an analyst at Point72 and Avala, but he also knows the AI space inside and out. He's an essential part of connecting our thinking to bottoms-up company analysis and the ins and outs of a stock setup, and considering when and how market updates will come (and how other market participants are positioned).

This summer, we promoted Michael to Head of Equities. He has become a leader of the team alongside Carl and now helps coordinate our equity research effort, systematizing our company scenario modeling while Carl leads blue sky thinking on the next chapter of AI.

Early this year, with Michael's help, we brought on Shu Yang Wei. Shu Yang is brilliant and brings wonderful energy to the team. He

[69] Carl is the smartest person I know. More than that, he also figured out that all of this AI stuff was going to happen 20 years ago—so has spent the last 20 years spending all day every day figuring out the implications. I have personally learned more from Carl than anyone else in the world, and his density of insight and the intellectual exchange with him is still the greatest joy of this job every day.

was Singapore's top cadet (receiving a "sword of honor" for his efforts) and graduated summa cum laude as Harvard's best economics student. After a stint with GIC, he joined us. He has grown quickly into a one-of-a-kind analyst and an indispensable member of the team. His sheer output analyzing companies is stunning and he has also been extremely forward-thinking in using AI in his modeling and research, as we will discuss later.

We also wanted to hone our modeling of some of the key big picture drivers, such as the speed of RSI, the diffusion of AI capabilities to businesses, the microeconomics of everything from lab monetization to spot compute prices, and the macroeconomic impacts of AI. Our old friend Zach Mazlish had done, in our view, some of the most incisive work on AI economics, including one of our favorite papers: *Transformative AI, existential risk, and real interest rates*. He was nearing the end of his PhD, and we brought him aboard. His work formalizes our picture of the AI situation and the economics of the AI industry and flows through to every company in our portfolio. He is exceptional at straddling the worlds of economics, markets, and AI forecasting.

Our standard for new investors has been extremely high—we believe the signal-to-noise ratio is crucial—and we scoured the best of the Street. No existing tech/AI analysts could pass Carl's interview, so we expanded our search further. Ultimately, we found RunLin Wang, a top performer at Farallon, where he was managing a successful healthcare book. We slowly convinced him that the AI opportunity was worth making the leap. He's just getting started, but we're incredibly excited by his potential.

Carl and I *were* the investment team when we started. Other analysts we spoke with—in interviews and at other funds—struggled to really take our worldview seriously and flow it through to companies. When they accepted the worldview, they would often lose precision. "If you really believe that, you can buy anything and do great," we heard. Sure, but what is actually the best thing to buy? AI beta has done poorly compared to the best investment opportunities. In practice, we have seen an enormous amount of dispersion.

We now have a phenomenal group of people supporting us. The group deeply understands how we see AI and is now helping us further develop that picture. We are covering far more companies

than we did last year. We are finding new ideas and subthemes we wouldn't have found had it been just the two of us.

When we started, we were very good at understanding the future direction of AI—but we had much to learn. Over the past two years, we built out the equity research engine—and became excellent at applying our insights on AI to stocks, marrying our top-down views with bottom-up company analysis, and translating it into single-stock alpha.

Quant team

Last year, Joe Zurier joined us. He had helped run an options desk at Jane Street, and is a mathematical genius. He spearheads our longer-run, ambitious quant research projects. In the summer, we brought on Anjalie Kini from Morgan Stanley as an options strategist. We wanted to ensure that we were executing options well and identifying key points of disagreement with the market. Anjalie was the only candidate to ace Joe's tricky options pricing interview, and quickly became an essential part of the team. Finally, as mentioned earlier Sven Khatri joined us as Head of Treasury Strategy, straddling the worlds between quant, risk, and treasury.

While the last couple years took us from big picture views on the future of AI to the equity research engine, we failed at translating that single-stock alpha into the right portfolio construction. Getting that right is the core project of our next chapter, and the focus of our current and coming hiring. We discussed the Risk and Portfolio Construction teams we are building earlier.

On the portfolio construction side, our goal is to develop systematic tools to help manage our book. We are building quantitative portfolio optimization tools that take our fair value probability distributions from the fundamental research team, paths implied by the options market, risk scenarios, crowding data, execution costs, and more to help us understand how we should size positions, express our views, and rebalance our portfolio day-to-day.

Our hope is that with systematic tools, it will be easier to understand when to swing big—and when to take profits. A systematic tool will help us see when risk-reward is especially asymmetric—and when risk-reward is worse, volatility and crowding are higher, and it's time to take some profits. Moreover, it will help us need to make fewer decisions (and make those better): the engine can help figure out daily incremental trims and adds (and help us precommit to

those systematically), while we focus on tuning the core knobs of this engine and the fundamental views.

We're proud of the team that we've built, and most of all the culture. There is a deep sense of collaboration, seriousness in the mission, dedication, and camaraderie. We express views in probabilities; a warm atmosphere welcomes disagreement. Even after July and in the AI bear market, the team is full of optimism and continues to stay focused and work hard.

When we started the fund, we set out to build the ultimate AI brain trust, and I think we did: if you are obsessed with this stuff, I think there is basically no other place to be. Every lunch, every dinner, we debate the future. Everyone loves what they do, and loves working with the team. It has been an absolute joy to work with everyone.

The team has become an incredible source of motivation for me personally. I adore this team, and they deserve a surrounding chassis in which they can thrive.

You can find the full list of bios of new team members in Appendix B.

Claude at SALP

Our fund has some other new help around the office: Claude. When we first set out to start the fund, when speaking with vendors and sometimes investors, we often ran into confusion regarding whether we were a fund that invested *in* AI or *with* AI (that is to say, using AI to trade stocks). Of course, we were the former, investing in AI companies based on our forecasts of the technology.

This has changed. Over the course of the last year, we have started seeing very substantial speedups from Claude across our organization—and it become a core tool to build the infrastructure that a fund of our scale needs.

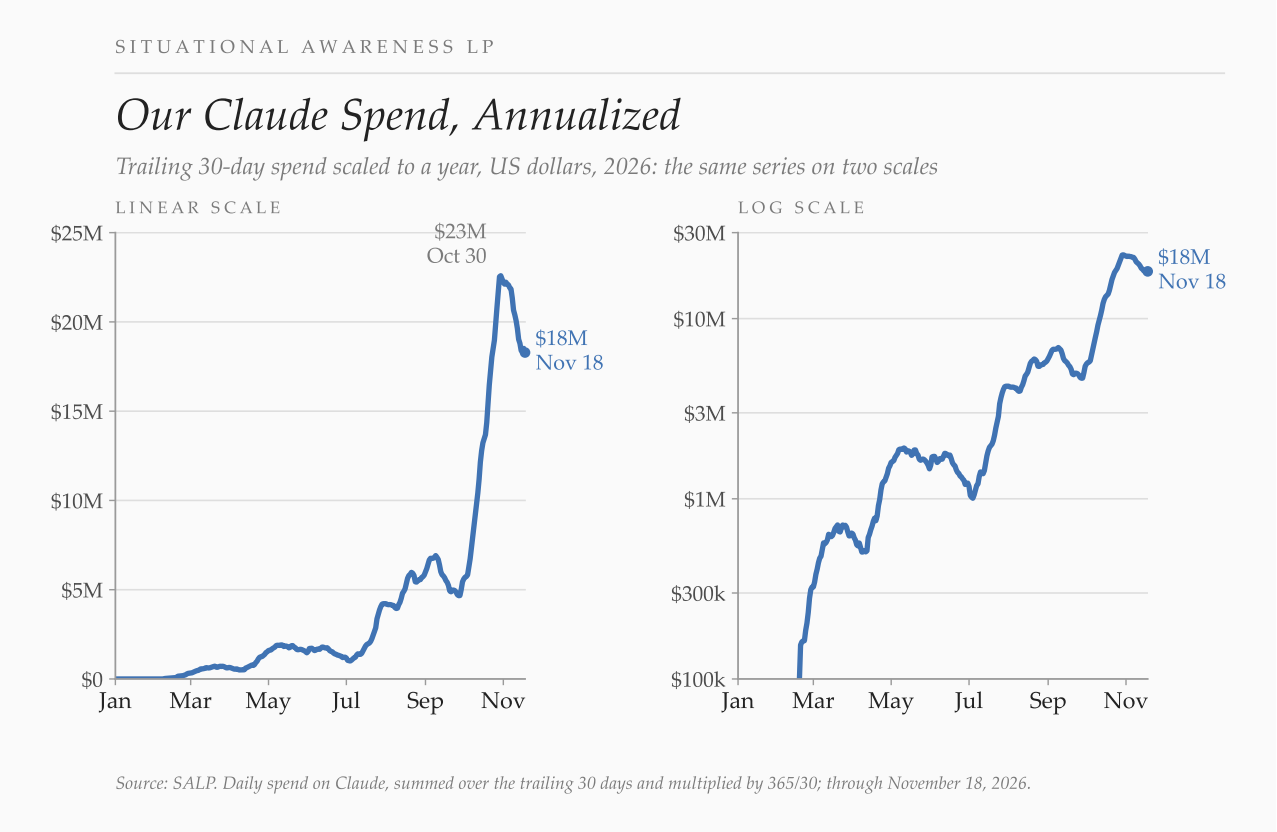


Figure 63: Growth of SALP's annualized Claude spend in 2026. All AI spend has been a management company expense as we create tools and understand how we can use these tools most effectively.

We've gone from Excel spreadsheets and poor specialized hedge fund software to an extensive set of internal infrastructure. This helps us track and analyze everything from our portfolio exposures, risk scenarios, and margin, to key research outputs and our models and AI scenarios, monitoring potentially interesting names, to helping us find the right options expression and measure execution costs, to helping us track tax, reconcile trade breaks, and more. Where pre-existing software solutions would have been available at all, many times small parts of this would have cost us handsomely.[70] Claude has also enabled us to do much more ambitious quant research projects than would have been possible otherwise, in our case to build the systematic portfolio optimization tools mentioned earlier.

But while software infrastructure and quant have been the earliest examples of AI uplift, we are beginning to see meaningful gains for our investment research as well. In our experience, the value of AI for the research team has been something like a year behind software development, as discussed earlier.

[70] All AI spend has been a management company expense as we create tools and understand how we can use these tools most effectively.

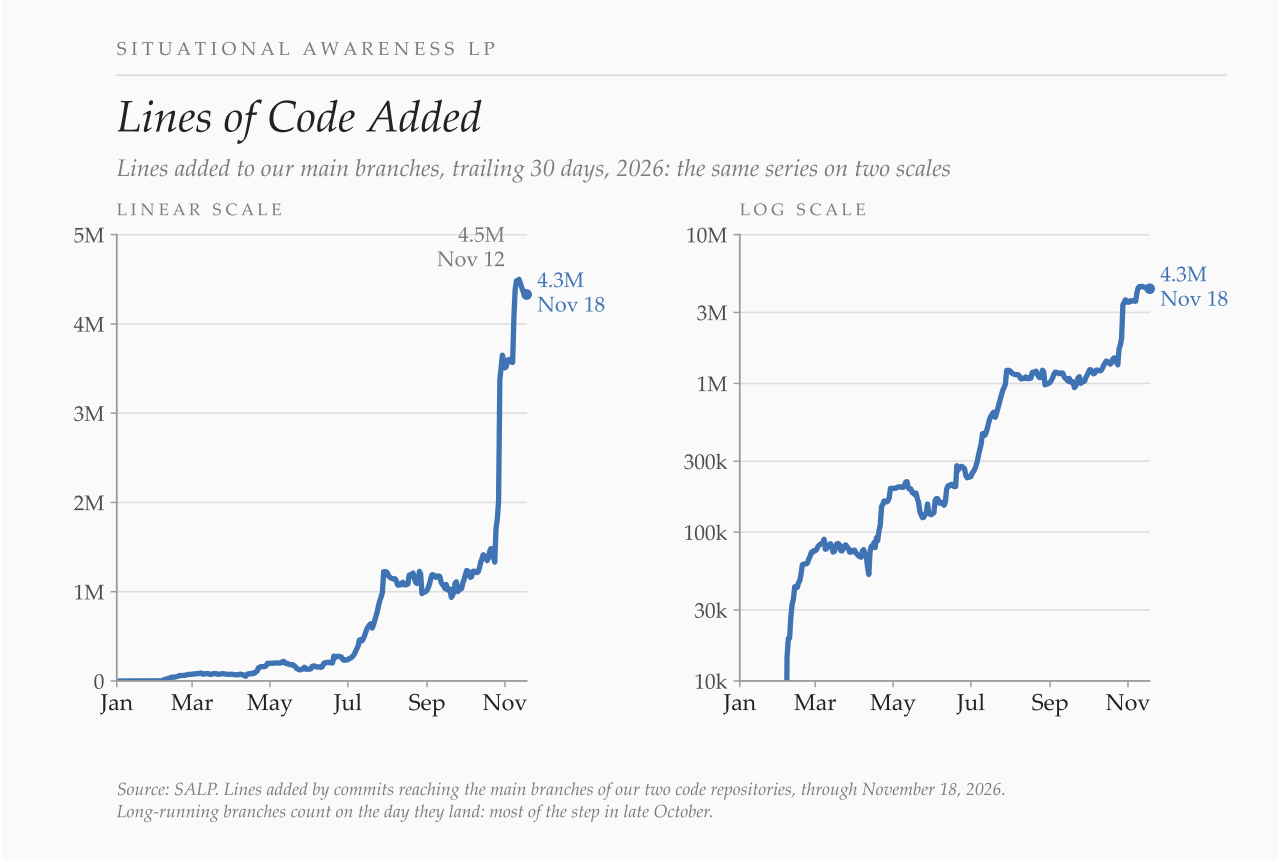


Figure 64: Growth of lines of code added to SALP's codebase.

Claude is still poor at high-level judgment—understanding the overall AI trajectory and its implications, the contrarian top-down theses that drive our research, and judgment calls on key drivers.[71] But it has radically accelerated our ability to think through and flow through our views to company analysis. We think we have probably gotten a 2-3x speedup overall on investment analyst work from Claude at this point. We can keep a small, high-signal core investment team, while expanding and deepening our coverage.

Claude has already changed how our organization works. We think this growth trend will continue and accelerate, and translate into an enormous amount of leverage for our organization. Other firms, such as the well-known HFT firms, have improved their profitability enormously as they have applied AI to short-term trading in the past couple of years. We think being on the frontier of applying AI to long-term fundamental investing will become a key competitive advantage for us.

[71] Up until a generation ago, Claude always seemed confused that memory prices and memory stocks had risen so much—it seemed unfathomable to Claude.

We predict a massive, continuing AI boom that will be very volatile. So far, we have had the direction right, but the volatility wrong. In our next chapter, you will get an organization similarly focused on finding the best investment ideas of this unprecedented period, but with a robust all-weather approach. That resilience is in our re-founding charter, in my investment approach, in our organizational design, and in our risk systems.

IV. Looking forward

Either we or the market will have an extraordinary amount of egg on the face in a year.

One of the many reasons I prefer my career in AI to my early dream of being an economist is the feedback loops. AI is moving extremely quickly. You get to make big, far-reaching macro predictions—the kind of thing that might have taken decades to resolve in macroeconomics, and even then still be debatable—and compare to reality just a couple years later. Time in AI is like dog years.

Predictions have been staked and they are coming to a head. The market has made its views clear: AI will peter out. We have made our views clear: A time of radical transformation is near. The far-flung claims—of RSI and agent swarms, a trillion dollars of revenue and industrial explosions, of rising spot compute prices and ubiquitous obsolescence—will meet reality.

The range of outcomes is wide; almost certainly, we've gotten important parts of the story wrong. But the vast disconnect I feel every day between the market and the world of AI cannot hold.

Running a hedge fund, too, has strong feedback loops. As Ilya once said, "if you value intelligence above all other human qualities, you're gonna have a bad time." The market will find a way to humble you. And building a durable fund and monetizing our ideas requires more than just good predictions.

We too have aged in dog years. We carry battle scars that we wish we could have earned more gently, but that perhaps can only be truly etched through an experience such as we went through.

Our ideas continue to be contrarian and our predictions bold, but our attitude has changed. I believe we have become wiser as investors. We have internalized that the path will be unpredictable and the singularity volatile, and we are building a resilient portfolio and a mature organization that can see it through.

We remain fully invested alongside you: Carl and I have nearly the entirety of our net worths invested in the fund. Carl and I have never had more conviction in the team and the fund. The team has come through these months in great spirits, doing some of the best work I have seen from them. We are heads down, hard at work every day to build a firm worthy of the continued faith you have placed in us.

Most of all, thank you. Many of you took an early bet on us. After July, almost every one of you took the time to talk with us, and gave us your advice, your backing, and your patience while we strengthened the firm. We are deeply grateful to have you as partners.

WHERE THE PATH to AGI was clear in 2024, the path to superintelligence is now shifting into focus. By 2028, barring a regulatory ban, a civilization of alien geniuses will arrive. We will venture into the unknown, an increasingly strange and tumultuous world. While we have focused on the economic consequences in this letter, a period of great civilizational challenges lies ahead.

In SF, things are starting to feel very weird. For most of the rest of the world, 2027 might be the last year that feels vaguely normal.

YOURS SINCERELY,

Leopold Aschenbrenner and team[72]

[72] I am grateful to Carl Shulman, Nick Whitaker, Zach Mazlish, Michael Wang, and Shu Yang Wei for their many invaluable contributions to this letter.

Appendix

Appendix A: Other investment case studies

Some brief notes on some other investments: our other major (current) private investment, Fluidstack, and two selected (historical) public expressions (Lumentum and Bloom Energy).

OUR OTHER MAJOR PRIVATE BET: FLUIDSTACK

Besides Anthropic, our other major private investment is Fluidstack (constituting the majority of the non-Anthropic private book). We heard of the team through their work with Anthropic in Fall 2025, and agreed to lead the Series A that October at $7.5B—in fact, we were so convicted we effectively preempted the entire Series A and were able to effectively take the entire round. This spring we co-led the Series B with Jane Street at $21B.

Fluidstack has been an interesting case for us as crossover investors, because the insight ran in the unusual direction. Traditionally, crossover funds carry insights from private markets into public ones. Here, our motivation came from our public book. We had watched companies like crypto miners effectively luck into key datacenter assets, and then watched those assets reprice—assets that had cost $50M become billions in equity value. How valuable could a company be that pursued these assets strategically? That is exactly what Fluidstack is doing. Our original case was that it sat at the center of three trends:

1. **Anthropic**. Fluidstack is Anthropic's preferred compute partner, and among the most torqued expressions of Anthropic's growth.
2. **TPUs**. Fluidstack is the primary external TPU cloud. Google started selling TPUs externally (beyond Google Cloud) for the first time last year and made a major push to realize the value of the "Nvidia inside". We've long thought that TPUs are the best chips, and thought Fluidstack was one of the most torqued expressions of TPU growth.
3. **A *situationally aware* team.** Management buys real-world options (land, power, turbines) years before the labs need them, which is what lets it reach massive scale if AI scaling continues.

We thought the valuation was highly attractive compared to public comps.

Since then, Fluidstack has crushed it:

- **Core partner to the leading two frontier labs, and the two best ASICs**. In addition to a rapidly growing partnership with Anthropic and TPUs, Fluidstack now is a core partner with a large contract to a second frontier lab. In addition, they have won major contracts with Google, Meta, and the Department of War.
- **Highly competent execution.** We've talked with effectively every neocloud and datacenter provider, and we've talked with many of the other folks that have talked with every neocloud and datacenter provider. The feedback is resounding: Fluidstack is the best, and rivaled only by SpaceX in the speed and competence of execution. Fluidstack can get TPU racks up 62% faster and pods 73% faster than Google Cloud itself. They are building modular datacenters designed to get online in six months from groundbreaking to RFS, rather than 1-2 years standard in the industry.
- **Vertically integrating to create franchise value.** In addition to locking up key resources like turbines and switchgear, Fluidstack has been vertically integrating. They are designing their own switchgear and power gear, a business which is scaling to a scale comparable with Vertiv (a $70B+ public company). They are building a million-square-foot factory in Phoenix to modularly fabricate datacenters to "print" GW-scale datacenters.
- **Cheap capital.** Because Fluidstack has bet heavily on ASICs with backers who have a vested interest in promoting those ASICs against Nvidia, it has been able to get an IG backstop (from

Google, ASIC design vendor, etc.) for its projects. Unlike neo-clouds shut out from the debt markets, this gives them easy access to financing.



Figure 65: Fluidstack Manufacturing Facility in Phoenix, AZ

This has also led Fluidstack to keep capturing more value per GW: from $1–3B of equity value per GW operating chips at miners' sites, to $8–13B at sites it develops itself, to as much as $15–20B where it also owns the compute.

We think Fluidstack is on track to be behind a double-digit percentage of the AI buildout, with 10s of GW brought online annually by 2028, and $10-30B/GW of net equity value created.

PUBLIC: LUMENTUM

In Q3 of 2025, we initiated a position in Lumentum, the leading merchant supplier of high-power indium phosphide lasers, for similar

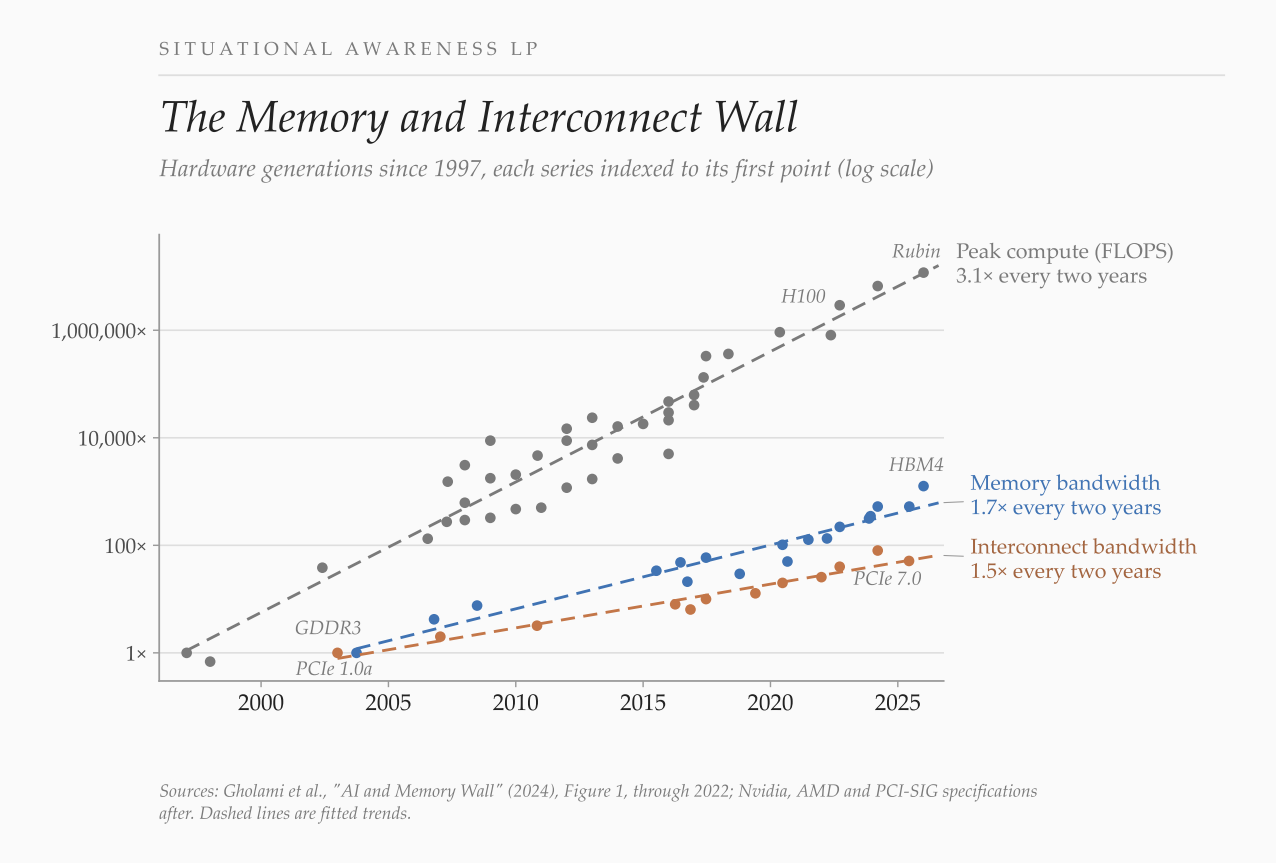


Figure 66: **The memory and interconnect wall**: FLOPS has scaled much more quickly than memory bandwidth and interconnect bandwidth, making memory and interconnect the bottlenecks.

reasons to Hynix. A related core bottleneck to the new RL and agentic inference workloads has emerged within compute architecture: chip-to-chip interconnect, which, like memory, has not scaled as quickly as FLOPS.

XPUs perform calculations in concert with each other, so a bottleneck in communication between chips leads to low system-wide utilization of FLOPS and memory. On top of that, training runs have grown substantially, requiring large scale-out networks to connect tens to hundreds of thousands of XPUs to each other. Inference too has grown more demanding: advances in model size and architecture mean many XPUs work together to produce tokens. This has led to the advent of scale-up interconnect, higher-bandwidth networks between smaller pools of XPUs. The first scale-up networks were within the rack and made of copper, as copper is cheap and reliable. But copper has its limitations as bandwidth and distance increase.

We saw two trends emerging. First, scale-out networks continue to expand, increasing the attach ratio of optical content per XPU. Second and more powerfully, scale-up networks move over to optical over time. As model sizes grow, scale-up domain sizes must grow too, along with bandwidth, straining the physical limitations of copper, which can only go so fast, so far—in particular to accommodate inference for agentic models with long contexts. In fact, larger and faster scale-up domains are a key solution to the memory bandwidth bottleneck that we identified arising with the shift to RL and agents: interconnecting more chips allows access to more unified memory bandwidth.

As scale-up networks use an OOM more bandwidth than scale-out, moving these connections to optical radically expands the demand for lasers. (The TAM would 5x+, *on top of* overall growth in AI.)

This is another example of how a deep understanding of the technology behind AI can let us see farther into the fog than others. Smart investors we spoke to were "laser-focused" (so to speak) on reading the Taiwan supply chain tea leaves to ascertain whether scale-up would go optical, or talked to industry participants who dismissed it as a pipe dream. We did this as well, but put more weight on the technical understanding which in our view suggested that this transition was inevitable, and that AI's demand for performance would naturally pull it forward sooner than jaded expert consensus expects.

Lumentum (LITE) is the primary beneficiary of this transition, the company best-positioned to produce the high-powered lasers needed in optical scale-up architecture. Our meetings with management and industry field research reinforced our conviction on LITE's technical prowess and management's lean-in on the AI opportunity. In Q1, we were proven right, with Nvidia announcing a part-optical scale-up network for its next-generation Rubin Ultra platform, much earlier than the analyst community expected (only a day earlier, we were still arguing with one of the smartest investors in NYC, who dismissed optical scale-up soon as a pipe dream). Nvidia also announced a $2B investment and long-term agreement for Lumentum's laser capacity.

Like memory, optical networking has been a brutal business historically, with boom-and-bust telecom cycles and immense pricing pressure. AI has changed all of this. Laser capacity will be in short supply for the foreseeable future, as evidenced by Nvidia funding Lumentum's capex and signing long-term offtake contracts. For

twenty years, this was an industry of cheap components and brutal cycles; today, customers are paying their suppliers for the right to buy from them.

LITE ran up multi-fold since we initiated the position. While we think LITE can still compound over the coming years, the IRR compressed as the market priced in more of our expectations of the transition to optical scale-up. We thus exited the position late in Q1.

PUBLIC: BLOOM ENERGY

In our first letter, we discussed Siemens Energy. We argued that AI would need tens of gigawatts of incremental power each year, and that gas turbines were the natural solution and bottleneck. Because the cost of power was so low compared to the total cost of a datacenter and to the revenue labs generated from that datacenter, there was significant room upward on pricing. The stock ended 2024 at €50, and remained a core position through 2025.

Throughout 2025, we saw our thesis borne out. There was a rush for turbine orders and the industry went into undersupply. The order book expanded out for three plus years, locking in increasingly higher prices. In the first letter we said gas margins would go from mid-single-digits to 20%+, which the company is on track to hit next year.

The market rewarded the gas turbine companies for the visibility they built by booking out their capacity for many years, with fixed price and downpayments. The stock ended 2025 at €120 as investors increasingly looked to earnings power in 2028 and beyond.

By the beginning of 2026 we completely exited the position. As the company had sold all its slots for the foreseeable future, and expanded its capacity relatively slowly, we saw less room for earnings revisions upward. It no longer carried much medium-term exposure to prices continuing to rise, which we desired. And we thought the market was embedding much greater AI bullishness and durability in the gas turbines than in, e.g., some semiconductor names.

Many took this gas turbine capacity constraint to mean certain GW figures were not reachable. We disagreed, believing in capitalism. There is a supply curve to power, and as power prices rose, entrepreneurs would make new, more scalable power generation methods viable. We began to search for who would benefit.

While we exited Siemens Energy, we did not exit AI power. In the third quarter of 2025, we began building a position in Bloom Energy, which we came across in our search for how the power bottleneck would be overcome. Bloom Energy produces fuel cells, which convert natural gas into electricity through a chemical reaction. Essentially every investor we spoke to about Bloom had dismissed it, as Bloom had promised an inflection in their business for over a decade that never materialized, its solution cost more than turbines, and it in the past had issues with the longevity of its products.

It was true that for most of Bloom's existence, it was a more expensive, less proven power source versus gas turbines, which led to a very limited market. For over a decade, gas turbines were in abundance. But as gas turbines became more expensive, and scarce, we thought that the industry would have no choice but to go up the power supply curve, where Bloom was waiting. When AI datacenters are rented out for $50M/MW-year on short-term deals or $20M/MW-year on long-term deals, paying up 50% on $2M/MW capex to get compute online six months faster is a no-brainer.

The other reason we found Bloom intriguing was that it could scale production much faster than gas turbines could, and importantly, management wanted to. Gas turbines depend on a complex supply chain and manufacturing process that takes years to expand. The companies were scarred by the last decade after the industry expanded too quickly and ended up in oversupply for years. Bloom had purposefully built a product that could be scaled much more like a semiconductor process than an old-world industrial one. Our research indicated the supply chain could scale to many GWs of annual production in short order if the demand was there. Bloom management had also publicly promised to meet the demand of customers and not to be the bottleneck.

Put together, we thought all capacity Bloom would make over the near-term would be sold out, potentially at very attractive prices, and Bloom would expand capacity quickly, scaling from less than a gigawatt of shipments per year to over 5GW in a few years.

Since we entered the position, Bloom found traction among AI customers as the industry scrambled to find power, and has made steps to significantly expand its capacity going forward. Oracle signed an agreement this April for up to 2.8GW, and Brookfield's commitment to finance Bloom purchases rose from $5B to $25B this June. Bloom in

2024 had $1.5B in revenue; they recently guided revenue for this year to be ~$4.5B.

We bought our first shares at ~$65 in September of 2025, peaking at $345 in June and falling back to $142 at the time of writing this. As our thesis played out this year, we significantly reduced our position starting in May as the stock went higher and we saw more attractive opportunities elsewhere in the market.

Appendix B: Bios of new team members

Looking ahead, we continue to build toward what we believe should be a talent-dense but best-in-class and well-staffed fund—one that can operate at institutional scale without bloating headcount and losing the magic we feel today. We would, as always, love introductions to exceptional people.[73] We've also moved to a larger office to accommodate the larger team. We will miss our loft, but Dwarkesh Patel and his podcast are still nearby.

[73] Right now, we're looking for more support for our portfolio construction team. Think fast-twitch quants.

You can find bios of all our new team members below (in order of start date):

Michael Wang — Head of Equities

Mr. Wang joined the Investment Adviser in December 2025. Prior to joining the Firm, he was an investor at Avala Global and Point72 Asset Management, where he focused on public and private companies in the technology and industrials sectors. Mr. Wang graduated from Columbia University with a B.A. in Economics.

David Geffen — Head of Counterparty Relations

Mr. Geffen has advised Situational Awareness since its launch in 2024 and serves as the Firm's Head of Counterparty Relations. He manages the Firm's counterparty relationships and oversees the negotiation of the Fund's prime brokerage, ISDA and other counterparty documents. Beginning in the fall of 2025, he began spending a significant amount of his time with the Firm, and now spends almost all his time with the Firm. He is the founder and President of Geffen Advisors LLC, a boutique firm founded in 2013 that advises hedge funds on counterparty negotiations, treasury management, and counterparty risk. Previously, Mr. Geffen was a Managing Director at BlackRock and Barclays Global Investors, where he created and

led the Securities Finance group, managing prime brokerage, financing, futures clearing, and counterparty credit relationships. Before that, he was Managing Director of Finance and Chief Credit Officer at Amaranth Group, and he held earlier roles at Goldman Sachs and Citibank.

Patrick DeGraca — Chief Financial Officer

Mr. DeGraca joined Situational Awareness in January 2026. Prior to joining the Firm, he was the Chief Financial Officer and Head of Operations at Freestone Grove Partners. Prior to that, Mr. DeGraca was the Chief Operating Officer of Masa Capital LLC, the Chief Operating Officer of Alta Park Capital, the Chief Administrative Officer for Two Sigma Investments, and the Chief Financial Officer and Chief Operating Officer for Traxis Partners. He has experience in all non-investment functions, including finance, operations, compliance, legal, and technology, at hedge funds across quantitative, multi-manager, equity long/short, and macro strategies. Mr. DeGraca has a B.S. from the University of California, Berkeley.

Kay Astorga — Senior Executive Assistant

Ms. Astorga joined Situational Awareness in January 2026 as Senior Executive Assistant. She supports the Firm's leadership team, coordinates investor and research meetings, manages calendars, and helps with day-to-day tasks. Prior to joining the Firm, Ms. Astorga was at Juniper Hill Capital Management LP, where she supported the finance and operations team. She previously spent a decade at Benchmark, where she supported the General Partners and managed executive operations for one of Silicon Valley's leading venture capital firms. Ms. Astorga studied studio art at New York University and marketing at San Francisco State University.

Raquel Gueits — Executive Assistant and Workplace Experience Lead

Ms. Gueits joined Situational Awareness in February 2026 as Executive Assistant and Workplace Experience Lead, where she is responsible for executive support, workplace operations, vendor management, and employee events. Prior to joining the Firm, Ms. Gueits was Manager of Operations at a private family office. She previously served as an HR and Operations Coordinator at a public relations firm. She attended the University of Oregon.

Shu Yang Wei — Investor

Mr. Wei joined Situational Awareness in March 2026 as an Investment Analyst, where he is responsible for investment diligence and financial modeling across semiconductors, AI infrastructure, compute, and other AI-related sectors. Prior to joining the Firm, he was an investment analyst at GIC. Mr. Wei graduated summa cum laude from Harvard University with an A.B. in Economics and Computer Science. He was awarded the sword of honor by the President of Singapore for graduating top of his class in officer cadet school.

Niki Webster — Head of Investor Relations

Ms. Webster joined Situational Awareness in March 2026. Prior to joining the Firm, she was co-founder, Chief Operating Officer, and Head of Investor Relations at Sea Cliff Partners. Prior to that, she was a Managing Director in Goldman Sachs' San Francisco Prime Brokerage division, and before Goldman, she worked in UBS' Prime Brokerage division in London, San Francisco, and Sydney. Ms. Webster began her career at Financial Risk Management (FRM), now part of Man Group. She holds an M.A. (Oxon) in Economics and Management Studies from the University of Oxford.

Sven Khatri — Head of Treasury

Dr. Khatri joined Situational Awareness in April 2026. Previously, he served as Global Head of Finance Quantitative Research at Citadel, where he led quantitative research efforts spanning Treasury, Counterparty Management, and Finance & Accounting. Prior to Citadel, he was a Managing Director at Goldman Sachs spanning the Treasury and Securities Divisions. Dr. Khatri completed a postdoctoral fellowship in the Aero/Astro Engineering Department at Stanford University. He holds a Ph.D. in Electrical Engineering from the California Institute of Technology.

Tom Glaser — Deputy Chief Financial Officer

Mr. Glaser joined Situational Awareness in May 2026 as Deputy Chief Financial Officer. Prior to joining the Firm, he was Chief Financial Officer at Pear VC, an early-stage venture capital firm, from 2025 to 2026. From 2013 to 2024, Mr. Glaser was the first employee and Chief Financial Officer of Think Investments, a global long/short equity hedge fund with peak assets of $3B, and from 2008 to 2013, he was the first employee and Chief Financial Officer of Ayer Capital, a

healthcare-focused hedge fund with peak assets of $300 million. Earlier in his career, Mr. Glaser worked in the financial services groups of Ernst & Young and Arthur Andersen, focusing on the audit and taxation of hedge funds. Mr. Glaser specializes in the financial reporting, taxation, and operations of hedge funds and venture capital funds. He holds an MBA from the Wharton School and a B.A. in Business Economics from UCLA.

Zach Mazlish — Economist

Dr. Mazlish joined Situational Awareness in June 2026 as an Economist on the fundamental research team, where his work focuses on the macroeconomic implications of artificial intelligence. Prior to joining the Firm, Mr. Mazlish completed a DPhil in Economics at the University of Oxford and worked as an economist at Eisler Capital and Penumbra Investment Group. During his DPhil, he helped create and teach a Stanford Digital Economy Lab course on the Economics of Transformative AI. He is one of the authors of *Transformative AI, Existential Risk, and Real Interest Rates*. Mr. Mazlish received a B.A. in Philosophy from Brown University.

Kimberly Summe — Chief Compliance Officer

Ms. Summe joined Situational Awareness in June 2026. Prior to joining the Firm, she was the Chief Operating Officer and General Counsel at Sequoia Capital Global Equities. She has also served in the past as the Chief Operating Officer and General Counsel at each of Stillwater Investment Management, LP and Partner Fund Management, LP. She was a Managing Director at Lehman Brothers and General Counsel at the International Swaps and Derivatives Association. She began her career as an attorney at Sullivan & Cromwell and then Pillsbury Winthrop. She received a J.D. from the University of Chicago, an LLB in Law from Cambridge University, a Master of Science from the London School of Economics, and a B.A. from Oklahoma State University. She was a Lecturer in Law at Stanford Law School for 14 years.

Amanda Polidoro — Head of Operations (Middle Office)

Ms. Polidoro joined Situational Awareness in June 2026. Prior to joining the Firm, she was a Director in Operations at Farallon Capital Management, a multi-strategy investment firm. Previously, she was a Vice President in the Prime Brokerage group at Goldman Sachs and

held a role in Prime Brokerage at Morgan Stanley. She began her career as an Operations Associate at Torrey Pines Capital Management. Ms. Polidoro brings extensive experience across the financial services industry, with expertise spanning trade settlement, margin and financing, public and private markets operations, vendor management, and operational process optimization. Her background combines broad middle-office experience with a strong understanding of the operational complexities of investment management. She holds a B.A. in Economics from St. Mary's College of Maryland and an MBA with an emphasis in Finance and Marketing from Wagner College.

Chris Nguyen — Trader (Asia Markets)

Mr. Nguyen joined Situational Awareness in June 2026. He most recently served as Head of Trading & Operations at Sequoia Capital Global Equities, a global TMT-focused crossover hedge fund investing across more than 20 countries in the U.S., APAC, and EMEA, where he was responsible for driving execution across global equity, derivative, and FX markets. Earlier in his career, Mr. Nguyen held trading and operations roles at Rydex | Security Global Investors and Seasons Capital Management, where he built a foundation in trade execution and operations. Mr. Nguyen holds a B.A. in Global Socio-Economic and Political Processes from the University of California, Santa Barbara.

Anjalie Kini — Options Strategist

Ms. Kini joined Situational Awareness in July 2026 as an Options Strategist. She manages the day-to-day implementation of the Firm's options strategy, including options strategy design, best-execution analysis, and forward-looking derivatives risk modeling. Prior to joining the Firm, Ms. Kini spent three years at Morgan Stanley on the single-name options desk, focusing on the technology, media, and telecommunications sector, with additional experience in event-driven merger arbitrage trading and book-wide scenario risk analysis. Ms. Kini holds a B.S. from the Massachusetts Institute of Technology, where she double-majored in Computer Science and Economics.

David Pigott — Head of Security

Mr. Pigott joined Situational Awareness in August 2026 as Head of Security. Prior to joining the Firm, he was the Co-Founder and Chief Operating Officer of Miru Labs. Prior to that, Mr. Pigott was the Chief Security Officer of Prime Trust, the Chief Information Security Officer of Qintel, Inc., and a Special Agent with the United States Secret Service. He has experience in information security, enterprise risk management, governance, and regulatory compliance, including SOC 2, PCI DSS, and ISO 27001 certification, across fintech, cybersecurity, and identity services firms. Mr. Pigott also serves as Adjunct Faculty for the Chief Risk Officer Executive Certificate Program at Carnegie Mellon University. Mr. Pigott has a B.A. from Santa Clara University.

Jim Frolik — General Counsel

Mr. Frolik joined Situational Awareness in September 2026 as General Counsel. Prior to joining the Firm, Mr. Frolik was a Partner in the Investment Funds and Advisers Group at Shartsis Friese LLP in San Francisco from 2007 to 2026, where he focused on representing managers of private investment funds in all aspects of their business. He joined Shartsis Friese out of law school in 2000. He is a co-author of U.S. Regulation of Hedge Funds, which was published by the American Bar Association, and is a member of the Board of Directors of EICC, the non-profit organization that organizes the annual Sohn San Francisco investment conference, which strives to enhance educational opportunities for underserved Bay Area students. Mr. Frolik received his B.A. and J.D. from the University of California at Berkeley.

Casey Sparks — Principal Security Engineer

Casey joined Situational Awareness as Principal Security Engineer in October 2026, after leading infrastructure and security at Plasmidsaurus. Before that, Casey headed IT, information security, and DevOps at Tau Motors, and spent four years at Prime Trust, from DevOps roles to Director of Information Technology. Across fintech, biotechnology, and automotive firms, Casey's work focused on cloud security, infrastructure as code, and identity and access management, including writing and implementing the security policies behind SOC 2, ISO 27001, and TISAX certifications.

Hadley Cook — Fund Controller

Mr. Cook joined Situational Awareness in November 2026 as Fund Controller. Prior to joining the Firm, he was the Controller of Two Creeks Capital Management, LP. Prior to that, Mr. Cook was the Assistant Controller of Glenview Capital and an Associate in the audit practice of Rothstein Kass, where he audited hedge funds and private equity funds. He has experience in fund accounting, NAV, investor allocations, tax, audit, regulatory reporting, cash controls, and vendor and technology management, including launching new fund vehicles and leading administrator and platform conversions. Mr. Cook has an M.S. in Accounting from Seton Hall University and a B.S. in International Business from Richmond American University London.

RunLin Wang — Investor

Mr. Wang joined Situational Awareness in November 2026. Prior to joining the Firm, he was an investor at Farallon Capital Management, a multi-strategy hedge fund. Mr. Wang received a Master of Science in Computer Science and a B.A. in Statistics, with a minor in Economics, from Harvard University. He was a silver medalist at the 2018 International Chemistry Olympiad and the 2017 International Biology Olympiad.

Jamie Rumbelow — Head of Software Development

Mr. Rumbelow joined Situational Awareness in November 2026 as Head of Software Development. Originally from London, he has over fifteen years' experience as a software engineer and founder at startups including Tools for Humanity, Conjecture, Algolia, and Tract. He is an Emergent Ventures fellow, Entrepreneur First alum, and Amplify Writing fellow. He holds a graduate degree in Philosophy and a bachelor's degree in Philosophy with Politics and Economics.

Michael DiGrande — Deputy Chief Compliance Officer (Regulatory)

Mr. DiGrande joined Situational Awareness in 2026 as the Deputy Chief Compliance Officer (Regulatory), where he focuses on the firm's regulatory compliance and operational risk programs. Previously, Mr. DiGrande served as the Chief Compliance Officer of Sequoia Capital Global Equities, a crossover hedge fund specializing in public/private technology investments. Prior to Sequoia, Mr.

DiGrande worked at ACA Group, where he partnered with SEC-registered investment advisers to scale their compliance programs and prepare for regulatory examination. He is a member of the State Bar of California and received his J.D. from Golden Gate University School of Law in 2014. Mr. DiGrande received his B.A. from Sonoma State University in 2009.

Elli Kavros — Deputy Chief Compliance Officer (Operations)

Ms. Kavros has over 28 years of experience in the alternative asset management space working in both sell side and buy side firms. She brings extensive experience leading compliance and operations functions at both emerging and well-established alternative investment advisers. Most recently, she served as a Senior Principal Consultant at ACA Group focusing on regulatory compliance and serving as an outsourced Chief Compliance Officer. Previously, she was the Managing Director of Operations and Deputy Chief Compliance Officer at Engine No. 1, an impact investment firm based in San Francisco with both hedge fund and exchange traded fund offerings. She held previous roles at Polysign Inc., a digital asset custodian, as a Senior Advisor; Passport Capital, a large global-macro hedge fund, as Director of Operations; Morgan Stanley, where she managed both the Client Service and Business Consulting teams for the West Coast Prime Brokerage division; and she began her career at Barclays Global Investors working on the passive index funds of the investment management business. Ms. Kavros has a Master's Degree from the University of Chicago and a Bachelor's Degree from the University of California, Davis.

IMPORTANT DISCLOSURES

Confidentiality. This document is confidential and subject to copyright protection. It is being provided to you on the condition that it not be forwarded, copied or otherwise distributed without the prior written consent of Situational Awareness LP ("**SALP**"). Investors or prospective investors that share this document without permission may be excluded or expelled from the investment funds SALP manages (each, a "**Fund**" and together the "**Funds**").

Reverse Solicitation. SALP is relying on reverse solicitation exemptions in the UK, EU, Switzerland and other applicable jurisdictions. If you are located in the UK, EU, Switzerland or other jurisdiction that recognizes reverse solicitation and did not initiate contact with SALP or have received this document in error, please immediately notify SALP.

No Offer or Solicitation. The only purpose of this document is to provide general background information on SALP and the Funds. This document is not an offer to sell or the solicitation of any offer to buy any securities. Any such offer will be made after you have received the Private Placement Memorandum and governing agreement for the applicable Fund (the "**Offering Materials**") and have had the opportunity to discuss all matters concerning any prospective investment that you desire with SALP.

Summary. The descriptions herein, including the descriptions of SALP's investment strategy and opinions regarding the future of AI, are in summary form, are incomplete and do not include all the information necessary to evaluate an investment in the Funds. All of the information provided herein is subject to and qualified in its entirety by reference to the terms and conditions described in the Offering Materials for the applicable Fund. Before you invest in a Fund, you should carefully and thoroughly review the Offering Materials with your financial, legal and tax advisors to determine whether an investment is suitable for you.

No Duty to Update. Neither SALP nor any of its affiliates assumes any duty to update any information in this document for subsequent changes of any kind. For example, SALP's opinions regarding the future of AI, investment strategy, investment selection criteria, processes and portfolio parameters summarized herein represents SALP's current intentions. SALP's opinions and investment approach may change at any time without notice to investors.

Forward Looking Statements. This document contains forward-looking statements based on SALP's expectations and projections about, among other things, its investment strategy and investment selection criteria and processes. Those statements are sometimes indicated by words such as "expects," "believes," "will" and similar expressions. In addition, any statements that refer to targets, expectations, projections or characterizations of future events or circumstances, including any underlying assumptions, are forward-looking statements. Such statements are not guarantees of future performance and are subject to certain risks, uncertainties and assumptions that are difficult to predict. The actual investment strategy and any investor's actual returns could differ materially and adversely from those expressed or implied in any forward-looking statements as a result of various factors.

Performance-Related Disclosures. The net performance is for a Standard Class new issue-eligible investor who has not made withdrawals or additional contributions, net of a 2% management fee, 20% performance allocation and Fund expenses. Returns experienced by individual investors will vary depending on the timing of their investments and withdrawals and different management and performance allocation arrangements. In calculating these returns, quarterly management fees are amortized monthly and performance allocations have been deducted as if allocable in each period shown, rather than annually. Returns include the reinvestment of all Fund profits.

Note that past performance is not indicative of prospective future returns.

Index Comparison. SALP compares the performance of the Funds to the VanEck Semiconductor ETF (SMH), and the ProShares Ultra Semiconductors ETF (USD) because investors typically request performance comparisons to ETFs that represent the "AI trade", and to Nvidia Corporation common stock (NVDA) and GraniteShares 2x

Long NVDA ETF because it is considered the default "AI" investment. Nevertheless, SALP believes that no index or other security provides a relevant comparison to the Funds' investment strategy or performance. The ETFs shown are unmanaged and do not engage in short sales, hedging, use of leverage or investments in private securities and non-equity securities. By contrast, the Funds are likely to be concentrated in relatively few investments and sectors, may invest in securities of companies that have smaller or larger market capitalizations, may invest in private securities and non-equity securities and are likely to employ short sales, hedging and leverage. Performance data for the ETFs and NVDA reflect reinvestment of earnings.

Preparation of Report. SALP prepared this document. It was not compiled, reviewed or audited by any independent party. Certain information contained herein has been provided by, or obtained from, third party sources. While SALP believes that such sources are reliable, it cannot guarantee the accuracy of any such information and does not represent that such information is accurate or complete.

References. References to the "firm", "us", "our" or "we" generally mean SALP unless the context dictates otherwise.