Asilo Argo – Portfolio Managers Quarterly Letter Q1 / 2026

Q1 / 2026: A QUARTER OF THREE ANNOUNCEMENTS

INTRODUCTION

Q1 / 2026 was a quarter defined by three announcements — announcements that at first glance appeared unrelated, but whose combined impact irreversibly altered the logic of AI adoption.

We saw it coming.

Markets did not. Instead they defaulted to a familiar pattern: panicked at efficiency gains, compared valuations to dot-com peaks, and treated historical records as ceilings for what was possible. Each of these reactions made sense through an old-world lens. Each was, in our view, more or less wrong.

In our Q4 / 2025 letter we wrote about Jevons' paradox — how lowering the cost of compute does not dampen demand but explodes it. This quarter we received confirmation of an equally important truth: technology does not cross the chasm to markets alone. Alliances do.

Geoffrey Moore said the mainstream customer does not buy technology. They buy a solution. This quarter the most powerful actors began doing exactly that — assembling the whole product, in a way markets do not yet know how to price. Our job is to hunt precisely these misunderstandings. This quarter there were plenty.

It was 1999, and Microsoft faced a problem. They had built the dominant operating system for personal computers, and they were now determined to conquer the enterprise. Windows NT was capable. Exchange was powerful. The technology worked. But Microsoft knew that the product alone was not enough.

Selling software to a multinational corporation meant that someone also had to design the architecture, manage the migration from legacy systems, train the workforce, and take responsibility when something broke at two in the morning. Microsoft could write the code. They could not be in ten thousand data centers simultaneously. Customers who wanted Microsoft did not just want Microsoft, they wanted a complete solution: the software, the implementation, the ongoing support, the accountability. In other words, in the eyes of the customer Microsoft did not provide a whole product.

To win, Microsoft forged an alliance. In 2000, together with Accenture, they created Avanade — a joint venture staffed by engineers who could sit inside client organizations, implement Microsoft technology, and deliver outcomes rather than licenses. The alliance completed the product. Within a decade, Avanade had deployed Microsoft infrastructure across the largest organizations on earth, and Microsoft's enterprise position was effectively unassailable. The technology had not changed. The wholeness had.

THE CHASM AND ITS CURE

Geoffrey Moore described this problem thirty years ago in *Crossing the Chasm*. The gap that kills most technology companies is not the gap between bad products and good ones. It is the gap between a product that works in the hands of early adopters — who will tolerate incompleteness, integrate the missing pieces themselves, and evangelize despite the friction — and a product that works for the mainstream customer, who will not.

The mainstream customer does not buy technology. They buy solutions to problems. If the technology solves only part of the problem, they will not buy it regardless of how good that part is. This is the chasm: the distance between what the core product delivers and what the customer actually needs. Companies that cross it successfully almost never do so by improving the core product alone. They do it by assembling alliances, partners who supply the missing components, who complete the whole product, who transform a technology into a solution that the mainstream market can adopt without heroism.

This pattern has repeated across every major technology transition. The question for any investor watching a new technology emerge is not only whether the core product is good. It is whether the whole product is being assembled — and by whom, and how fast.

Q1 / 2026: THE ALLIANCE WAVE ARRIVES

In the first quarter of this year, the answer arrived with unusual clarity, and unusual speed. In a single week of March, three announcements landed that appeared unrelated.

1) Jeff Bezos was reported to be raising $100 billion to acquire manufacturing companies and deploy AI across them.

2) OpenAI entered advanced negotiations with Advent International, Bain Capital, Brookfield, and TPG for a joint venture.

3) Anthropic simultaneously opened parallel discussions with Blackstone, Hellman & Friedman, and Permira. The AI labs were not raising capital. They were assembling whole products.

The scale deserves a moment of pause. The seven firms named in these deals — *so far* — collectively manage approximately $2.7 trillion in assets. That is roughly eight times the entire Helsinki Stock Exchange. It is a portion of a global private equity industry that controls $13 trillion in total, meaning what was announced this quarter is only the prelude, not the whole symphony. A single alliance with four PE firms is the equivalent of opening hundreds of enterprise relationships simultaneously, relationships where the PE firm's own financial incentive is now aligned with AI deployment succeeding.

The Bezos move completes the picture. Project Prometheus — the AI startup Bezos co-founded in late 2025 — builds models to understand and simulate the physical world: materials, prototyping, pre-production engineering. The wholeness gap in physical AI is acute. You cannot sell AI for aerospace or chip manufacturing without domain expertise, trusted relationships with engineering teams, and years of embedded deployment. Bezos is not buying factories as a financial exercise. He is buying the whole product completion that no software vendor can provide alone. The portfolio companies become simultaneously the implementation layer, the proof cases, and the distribution channel. Microsoft needed Accenture. Prometheus needs the factory floor.

WHY THIS WILL WORK

The alliance model succeeds when the partners can show each other something real.

AI labs no longer need elaborate pitch decks to command the attention of private equity; they only need one data point. Just fourteen months ago, Anthropic's annualized revenue sat at $1 billion. Today, it has skyrocketed to $14 billion. OpenAI crossed $20 billion in annualized revenue in 2025, growing from $2 billion in two years. Brad Gerstner, on the All-In podcast this March, called it "the splitting-the-atom moment for AI revenue." This rate of growth has no precedent in the history of business. Not in software. Not in consumer technology. Not anywhere. The product is not in evaluation. It is in production, at the highest levels of the global economy. The initial force has been achieved. The snowball is rolling down a steep slope: it gathers speed from its own mass and swallows everything in its path.

Then they present the Goldman Sachs case:

Six months of embedded engineers. Two specific workflows selected precisely because they had resisted automation for decades: trade and transaction accounting, and client onboarding.

Both combine document parsing, regulatory judgment, and exception handling in ways that earlier generations of software could not navigate. Goldman's CIO Marco Argenti described what the testing revealed: the same reasoning capability that writes code also navigates KYC databases and reconciles transactions. Goldman was surprised by the breadth. They should not have been, but their surprise is commercially useful, because it means the opportunity sitting inside the average PE portfolio company is larger than that company currently believes.

PE firms are not paid primarily to grow revenue. They are paid to expand EBITDA, compress cycle times, and create the operational leverage that justifies exit multiples. This alliance provides them with the ultimate tool for margin expansion. The JV structure is not primarily a capital raising mechanism. It is a deployment vehicle. A way to take the "Goldman playbook" and run it simultaneously across hundreds of portfolio companies, with the PE firm's own financial incentive aligned with every deployment succeeding. For the AI labs, each deployment creates a customer whose switching cost compounds with every workflow integrated, every process rebuilt, every model embedded deeper into how the business actually runs.

The match is structural, not circumstantial. The AI labs bring the technology and the proof points. The private equity firms bring the portfolio companies and the mandate to drive value creation. Bezos brings access to the physical world that neither side can reach alone. Each party supplies exactly what the others lack — together, they solve one another's wholeness problem.

This is no coincidence. It is the chasm being crossed, but not in the way Geoffrey Moore imagined. Moore likened the challenge to Normandy in 1944: a brutal, heroic landing under fire, where survival itself was uncertain. This moment feels different. It is less D-Day and more the shift from Tehran in 1943 to Yalta in 1945 — when the Big Three sat down together and began to partition the postwar order. What we are witnessing is not a desperate assault, but the quiet formation of a new industrial alliance that will reshape the global economy.

A DIFFERENT MODEL FOR A DIFFERENT MOMENT

What you have just read is the power of a latticework of mental models. An analysis rooted in historical pattern recognition. A framework sold in millions of copies, field-tested across decades of technology transitions. A mental model that yields concrete, actionable insight — not despite its qualitative nature, but because of it.

Think of what the whole product framework implies for demand. Every enterprise that deploys an AI agent becomes a persistent, recurring consumer of inference compute, memory bandwidth, and networking capacity. Switching costs accumulate rapidly: a model embedded in Goldman's trade reconciliation workflow is not replaced on a quarterly procurement cycle. The alliances, once formed, do not just accelerate adoption. They make adoption durable. The demand consequence is structural, not cyclical, and it compounds.

One might assume this is how the broader market reads the situation. After all, *Crossing the Chasm* is not an obscure academic paper. It has been read by every management consultant, every venture capitalist, every strategy executive in the technology industry for thirty years. But this is not how the market is reading it. A significant portion of institutional analysis relies exclusively on backward-looking quantitative metrics. Hard numbers. Verifiable data. Scalable models. Methods that produce arguments that sound precise and therefore compelling.

Let us examine what those arguments actually look like.

At the peak of the dot-com bubble in 2000, Cisco Systems reached a market capitalization of $555 billion — the highest ever recorded at that time for any company on planet earth. For years afterward, that number became a shorthand for irrational exuberance. Any company approaching that valuation was, by this logic, approaching the scene of the crime. The argument sounds rigorous: compare current valuations to the historical peak of the greatest bubble in living memory and conclude that proximity to that number signals danger.

The problem is that today there are dozens of companies with market capitalizations that dwarf Cisco's dot-com peak. Nvidia alone has crossed $4 trillion. The number that once defined the outer boundary of human financial imagination is now a mid-table figure in the S&P 500. The argument was not wrong because the arithmetic was faulty. It was wrong because it treated a historical extreme as a permanent ceiling rather than a temporary milestone in a continuing expansion.

The same error appears in growth rate analysis. How fast can a company grow? The quantitative analyst looks at the fastest historical examples and treats them as limits. Netflix reached 100 million subscribers in ten years — that ceiling defined the upper bound of what consumer technology adoption could achieve. Then ChatGPT reached 100 million users in 60 days, faster than any product in history. Then OpenClaw became the fastest-growing open source project in the history of software, exceeding what Linux achieved in 30 years in a matter of weeks. Think of that difference, 30 years, and a matter of weeks! Each time the model was calibrated to the previous record, the next development rendered the calibration obsolete. The limit was not the technology. It was the analyst's sample size.

The same pattern appears in the spread of open platforms. Linux was the canonical example of open-source adoption — a decades-long journey from hobbyist project to enterprise infrastructure standard. That journey set the mental model for how long open platforms take to cross the chasm. OpenClaw crossed it in weeks. Jensen Huang, at GTC in March 2026, called it the fastest-growing open source project in human history and declared that every company needs an OpenClaw strategy — the same framing that every company once needed a Linux strategy, compressed from decades into a single earnings cycle.

THESE INDIVIDUAL ERRORS OF CALIBRATION COMPOUND WHEN THEY ARE AGGREGATED INTO INSTITUTIONAL MODELS

The Bank of England and the IMF have drawn explicit comparisons to the dot-com bubble, citing valuation multiples and debt levels as evidence of irrational exuberance. Goldman Sachs analysts documented hyperscalers taking on $121 billion in new debt in a single year — a 300% increase from historical norms — and framed it as overextension. Morgan Stanley estimates $3 trillion in AI infrastructure spend through 2028, half funded by debt, and presents the number as a warning. Deutsche Bank's Jim Reid projects $140 billion in cumulative losses for OpenAI by 2029 and concludes the economics are unsustainable.

The reality, however, is underbuild, not overbuild.

During Q1 of this year, the capex guidance from the five largest hyperscalers grew by roughly $150 billion in a single earnings cycle. This is not a number that stems from overestimation, it is a number that stems from underestimation.

An even starker signal came from AWS CEO Matt Garman in February: AWS has never retired a single A100 server. Not one.

Think of it through a car analogy. Jensen Huang promised "his current car", the Blackwell GPU, would do "5 liters of gasoline per 100 kilometers". SemiAnalysis (the specialized research and analysis firm focused on the semiconductor and AI industries) took it for a test drive and measured only 2.5 liters/100 km. Underpromise, overdeliver! Now, picture a fleet of cars. The new cars do 2.5 liters per hundred. The oldest "Datsuns" (i.e. the Nvidia A100 GPU) from 2020, now with updated software, still do over a hundred liters/100 km — compared to the updated Blackwell chips. Microsoft's CFO said this quarter that GPUs sit idle in warehouses for lack of electricity. And still, not a single Datsun has been scrapped while we are experiencing an energy bottleneck! And at this point we have not yet even considered NVIDIA's upcoming Vera Rubin chip, arriving later in 2026 — at which point the Datsun's comparable consumption would be in the range of several hundred liters.

That is not a bubble. That is what happens when industry kept underbuilding for years and is now scrambling to close the gap. Operators run six-year-old hardware at full capacity because there simply isn't enough infrastructure to replace it — not because the new technology doesn't exist, but because they underinvested in the power, cooling and facilities needed to deploy it.

THE RETAIL ECHO CHAMBER: PRECISION WITHOUT UNDERSTANDING

If institutional models suffer from a failure of calibration, the retail market suffers from a failure of context. Information that is technically true is frequently used to reach conclusions that are logically impossible.

Take, for example, the recent reaction to Google Research's "TurboQuant" announcement. Within hours, social media was flooded with the narrative that the "AI bubble" had popped (see next page) because a new compression algorithm reduced KV-cache memory requirements.

To the uncritical eye, the logic seems sound: if Google can shrink memory usage by 6X, we need 6X fewer chips. If we need fewer chips, the "memory stocks" must crash. It is a linear, clean, and entirely incorrect deduction.

Ejaaz @cryptopunk7213 · Mar 25

wow google might've popped the ai bubble, memory stocks down massively today:

their new algorithm shrinks an AI model's memory by 6X WITHOUT reducing it's intelligence making it 8x faster with the SAME # of GPUs:

if this works - we don't need as many GPUs to train AI

- kv-cache is basically a model's short term memory. it gets massive pretty quickly = larger, slower, expensive ai

- google's algo compresses it to just 3-bits with ZERO loss in accuracy (usually models are like 32-bit)

the combined market cap of micron and sandisk is $527 billion and im not even factoring in SK hynix and samsung

ai has driven up memory prices by 500%+ over the last few months - if google's algo scales then this might crash.

Google Research @GoogleResearch · Mar 24

Introducing TurboQuant: Our new compression algorithm that reduces LLM key-value cache memory by at least 6x and delivers up to 8x speedup, all with zero accuracy loss, redefining AI efficiency. Read the blog to learn how it achieves these results: ...

274 602 8.2K 1.6M

This specific "panic" relies on three fundamental errors of calibration:

1. **The KV-Cache Fallacy:** The KV-cache (Key-Value cache) is not the "total memory" of an AI system; it is a specific, dynamic subset of memory used during inference. Reducing the footprint of the cache is an incredible feat of engineering, but it does not eliminate the massive HBM (High Bandwidth Memory) requirements needed for the model weights themselves. Shrinking the "short-term memory" does not suddenly make the "brain" smaller.

2. **Jevons Paradox:** This is the most consistent blind spot in technology analysis. When you make a resource more efficient, you do not decrease total consumption; you *increase* it. By making AI 8X faster and 6X more memory-efficient, Google hasn't reduced the need for GPUs—they have just made it economically viable for ten times as many companies to deploy models. Efficiency lowers the barrier to entry, which explodes the total addressable market.

3. **Categorical Confusion:** The post cites SanDisk as a "memory stock" casualty. SanDisk (owned by Western Digital) primarily produces NAND flash for consumer devices and SSDs. The AI boom is driven by **HBM (High Bandwidth Memory)** and **DDR5**—produced by SK Hynix, Micron, and Samsung. Punishing SanDisk because of a breakthrough in LLM inference logic is like selling Ford stock because someone invented a more efficient jet engine.

And there are more where these came from.

We don't fear these headlines. We welcome them. Everyone sees the same numbers; we see what they mean. We have the ability to tell a "Datsun" from a "Blackwell" — and a cache optimization from a demand collapse.

When the market reacts to "TurboQuant" by selling off the entire semiconductor supply chain, it creates a massive disconnect between price and reality. The retail investor sees a "pop"; we see an efficiency gain that will lead to the next 100 million users.

We hunt for misunderstandings. And of misunderstandings, there are plenty. We're just getting started.

Thank you for your trust,

Henri Blomster & Ernst Grönblom
Portfolio Managers of Asilo Argo