INVESTMENT PERSPECTIVE • MAY 2026

Conviction Without Concentration

Underwriting the AI thesis on evidence rather than narrative, and constructing exposure that holds up across a wide range of futures.

Blind pessimism, absent deep work, is laziness disguised as sophistication. The bear case on AI gets coded as the rigorous, adult-in-the-room position. The bull case gets dismissed as naïve, promotional, or overly speculative. That framing has the prestige hierarchy exactly backward. The work, the actual underwriting of growth, unit economics, supply, demand, adoption, capability, and price, is what separates analysis from posture. A reflexive bear and a reflexive bull are doing the same thing, which is using a directional label as a substitute for the work.

What follows is not an argument for any particular investment thesis. It is an argument for the discipline of building theses from evidence rather than narrative. AI provides the cleanest current test of this principle. The piece walks through six pieces of evidence, ends on what investors are actually being asked to pay for access, and then lays out how all of this should inform thoughtful portfolio construction. Not as a bet on AI, but as an example of how careful underwriting creates asymmetric opportunities while managing concentration risk.

I. The Growth Is Real, and It Is Unprecedented

Start with Anthropic, because the numbers are almost difficult to believe even when you have verified them. Anthropic's annualized revenue went from roughly $1 billion at the start of 2025, to $5 billion by August 2025, to $9 billion by year-end, to $45 billion by May 2026. That is a thirty-fold increase in sixteen months. For context, Salesforce, one of the most successful enterprise software companies of the past two decades, took roughly twenty years to reach $30 billion in annual revenue. Anthropic did it in under three years from a standing start and reached a $45 billion run-rate only one month later.

The quality of that revenue matters as much as the magnitude. Roughly 80% of Anthropic's revenue comes from enterprise customers. Over 1,000 of those customers spend more than $1 million annually, a figure that has doubled in under two months. Eight of the Fortune 10 are now Claude customers. This is not a consumer hype curve. It is a procurement curve from the largest, most cost-conscious purchasing organizations in the world, paying seven-figure annual contracts for capabilities they did not believe existed eighteen months ago.

OpenAI, the larger competitor, is on a parallel trajectory. Reports indicate roughly $24 billion in annualized revenue as of April 2026, up from $6 billion the prior year, with enterprise revenue on track to reach parity with consumer by year-end. Two companies, both growing at rates that have no precedent in software history, neither slowing. They are accelerating.

II. Unit Economics Are Improving Even as Absolute Losses Grow

This is the pillar most commonly mischaracterized in both directions, so I want to be very careful with the data.

The bear headline is correct on its own terms: OpenAI's absolute losses are growing, not shrinking. Internal projections show losses growing from roughly $5 billion in 2024 to $14 billion in 2026, with break-even not expected until 2029 or later. Stated that way, the picture looks alarming.

The deeper picture is more interesting. OpenAI's compute margin on paid products roughly doubled from 35% in early 2024 to approximately 70% by late 2025. GPT-5's per-call gross margin sits near 48%. The unit economics of inference, the cost of actually serving a query once the model exists, have been improving dramatically. What is consuming the cash is not deteriorating economics, it is deliberate, aggressive reinvestment into training the next generation of models and locking up compute capacity.

Investors (including Yours Truly in the past) often believe that each subsequent model from the labs consumes more compute and is less efficient. As the models get better and more sophisticated, outsiders wrongly liken it to going from a Toyota Camry to a Lamborghini Revuelto and assume it's less efficient. Except it's the other way around. More advanced models are proving far more efficient. Yes, they are costly to create. But once released, they and subsequent models (assuming scaling laws hold) improve the unit economics.

This is quite similar to the Amazon 1999 playbook, and it has been misread before. A business with rapidly improving unit economics and a deliberate decision to reinvest gross profit into capacity expansion looks identical, on the income statement, to a business with broken economics burning cash to stay alive. The two look the same. They are not the same. The difference shows up in the gross margin trend and the underlying cost-per-token curve, both of which have been moving in the right direction for two years.

Anthropic offers the cleaner version of this argument. The company passed OpenAI on annualized revenue in April while spending roughly four times less on model training, in large part because they are prioritizing cost-per-token compute efficiency, which is the central variable an evidence-based underwriter of this space should track. If frontier-model economics were structurally broken, you would not see one competitor surpass another at a fraction of the cost. You would see both burning equally and racing toward the same wall.

III. Compute Demand Is So Strong That Older Chips Are Appreciating

This is the data point I find hardest to dismiss, and the one that should trouble the bears most.

The standard bear argument requires GPU rental prices to fall as supply catches up to demand, because that is what commodity hardware does over time. Older generations get cheaper as newer ones arrive. The H100, Nvidia's two-generation-old training chip, should be the canonical example: launched in 2022, surpassed by Blackwell, theoretically headed for the bargain bin.

The opposite is happening. According to SemiAnalysis, the one-year H100 rental contract index rose from $1.70 per GPU-hour in October 2025 to $2.35 by March 2026, a 40% increase on a chip that has been on the

market for three years. Hyperscaler on-demand H100 capacity is effectively sold out across all major providers. Available Blackwell capacity through August-September 2026 is already fully booked. Renters are subletting their clusters like Monaco apartments during the Grand Prix or Green Bay homes in walking distance to Lambeau on a gameday weekend (you get the point).

If we were in a bubble, this is the price signal you would not expect to see. Bubble compute markets show falling rental rates as new supply arrives and speculative demand evaporates. What we have is the inverse: a structurally constrained market where even outdated and depreciating hardware is appreciating because real economic demand exceeds the physical ability to supply it.

This is not survey data or projected demand. This is what real companies are paying right now to access compute they cannot otherwise obtain.

IV. We Are Early. Genuinely Early.

The narrative that AI adoption is saturated is one of the more puzzling claims in the current discourse, because every credible source on enterprise adoption says the opposite. Few people and enterprises have adopted AI yet.

McKinsey's 2025 State of AI report found that only 1% of organizations consider their AI deployments mature. Roughly two-thirds of respondents say their organizations have not yet begun scaling AI across the enterprise. S&P Global puts the share of enterprises with any AI agent in production at 31%. Gartner forecasts that 40% of enterprise applications will embed task-specific AI agents by the end of 2026, up from less than 5% in 2025, a multiple, in a single year, that adoption curves rarely produce.

The pattern across every major research firm (McKinsey, Gartner, IDC, Forrester, BCG, Bain, Deloitte) is consistent. Enormous experimentation, very little maturity, ROI concentrated in a small group of leaders that are pulling away from the rest. The median enterprise's monthly LLM bill is growing 7x year-over-year, and the spend is still tiny relative to where it is going.

If this were the dot-com cycle, we would be somewhere around 1996 or 1997. The infrastructure is being built, the first wave of enterprise contracts is being signed, the leaders are establishing themselves, and the broad enterprise economy has barely begun to integrate the technology into actual workflows. The bear case requires us to be near saturation. The data says we are near the starting line.

V. Productivity Gains Are Now Measurable

Capability is the leg of the AI debate where the goalposts have moved most aggressively. Two years ago the question was whether models could write coherent paragraphs. One year ago it was whether they could write working code. Today it is whether they can run autonomous multi-hour engineering workflows.

Stanford's 2026 AI Index documents the pace. Frontier models gained 30 percentage points in a single year on Humanity's Last Exam, a benchmark explicitly designed to be difficult and to favor human experts. On

standard coding benchmarks, models went from solving roughly 60% of tasks to nearly 100% at a human-expert level. Evaluations designed to last for years are saturating in months.

More importantly, the capability gains are translating into measurable economic value. Field studies cited in the AI Index show productivity gains of 14% in customer service and 26% in software development. Bain's forward model projects knowledge-worker productivity gains expanding from 7-9% in early 2026 to 14-19% by year-end 2027. The estimated annual value of generative AI tools to U.S. users alone reached $172 billion by early 2026, with the median value per user tripling within a year.

The capability is real, the deployments are real, and the productivity is real. None of this is hypothetical anymore. The bear case requires the gains to plateau, the deployments to disappoint, and the productivity to fail to materialize. None of those have happened, and the trend has been moving in the opposite direction every quarter for two years.

VI. And Here Is What We Actually Pay for It

Everything in investing is downstream of value, and this is the point that deserves the most space because it is the place where the bubble narrative collides hardest with the actual prices relative to valuation metrics.

A bubble, by definition, requires extreme valuations. The market is supposed to be paying ruinous prices for the assets in question, in anticipation of future cash flows that cannot possibly justify the current cost. That is the textbook definition. It is what happened to Cisco in 2000 at 130x earnings. It is what happened to the Nifty Fifty in 1972 at 50-60x. It is what is supposed to be happening now.

Here is what is actually happening. As of mid-May 2026:

Microsoft trades at roughly 20-21x forward earnings, with 18% revenue growth (Azure grew 40%) and 30% earnings growth.

Nvidia trades at roughly 26x forward earnings, with 65% revenue growth and 71% gross margins.

Amazon trades at roughly 30x forward earnings, with AWS growing 28% year-over-year.

Broadcom trades at roughly 32x forward earnings, with a PEG ratio below 1 and revenue growth over 52%.

Now compare those multiples to the companies the same market considers safe precisely because they do not touch AI:

Costco trades at roughly 46x forward earnings, growing earnings in the high single digits.

Casey's General Stores, a Midwestern gas station and pizza chain, trades at roughly 45x forward earnings.

Walmart trades at roughly 45x forward earnings.

The market is currently asking us to pay a higher multiple for incremental cases of soft drinks sold at a Midwestern convenience store than for shares of the company designing the silicon at the foundation of the

most consequential technology transition of the decade. We are being asked to pay a higher multiple for a warehouse club than for the hyperscaler whose cloud infrastructure half the Fortune 500 runs on.

You can disagree with everything written in pillars one through five and still find this valuation asymmetry difficult to defend. The numbers do not look like a bubble. They look like a market that has talked itself into pessimism on the most important technology since the internet while quietly paying premium prices for the most defensive, slowest-growing businesses it can find.

This is not what bubble pricing looks like. This is what the opposite of bubble pricing looks like, fear pricing, in the names that should be priced for growth.

VII. Conviction Is Not Concentration

Everything above should be read as analysis, not as a portfolio prescription. A portfolio constructed with discipline is not designed to maximize exposure to a single thesis, even one held with this much conviction. It is designed to maximize the probability of compounding capital reliably across a wide range of futures, including ones where the underlying thesis turns out to be partially wrong.

There is a distinction between conviction in a thesis and concentration in a thesis. Those are different things, and most investors conflate them. A manager who is bullish on AI and runs a 60% AI-pure-play book is making a single bet. A manager who is bullish on AI and owns it through quality businesses with multiple ways to win is making a portfolio. The first is a directional call dressed up as research. The second is what “optimization over maximization” and the genuine pursuit of asymmetric outcomes actually looks like in practice.

In practice, this means exposure to the AI thesis should be layered, not stacked. The first layer is a deliberately modest sleeve of pure-play exposure. Names whose multiples depend, in some meaningful way, on the AI capex cycle continuing — Nvidia, Broadcom, and Micron are obvious candidates. The case for holding them rests on a belief that the cycle has further to run and that, even here, the entry multiples are reasonable relative to growth: Nvidia at roughly 26x forward earnings against 65% revenue growth, Broadcom with a PEG ratio below one, Micron working through a cyclical recovery that stands on its own. The sleeve should be sized to participate meaningfully in the upside without letting it dictate outcomes if the cycle disappoints.

The second layer, and the far larger one, is the set of businesses that win regardless and win more with AI. Amazon and Microsoft are the cleanest examples. These are hyperscale franchises with durable, profitable businesses outside of AI: AWS’s non-AI workloads, Microsoft’s productivity and gaming franchises, Amazon’s retail and advertising engines. The allocator is not paying full price for the AI growth at either name; the allocator is paying a reasonable price for the base business and treating the AI contribution as optionality. As noted in Pillar VI, these are the same companies trading at material discounts to consumer staples despite better unit economics, stronger balance sheets, and faster growth.

The third layer is the part of the portfolio where AI is essentially irrelevant to the underwriting. Aircraft lessors like AerCap and well-run regional bank franchises like First Citizens BancShares are illustrative. The investment case for businesses of this type is built around capital allocation discipline, durable contracted cash flows, and operating models that compound book value through cycles. If AI productivity gains accrue to these operations over time, that is a marginal tailwind. If they do not, the investment case is wholly intact. This is where a serious allocator should spend a disproportionate share of underwriting time, because these are the positions that determine whether the portfolio survives any individual thesis being wrong, including this one.

The pessimism critique that opened this piece was about the absence of work. The portfolio architecture above is what the presence of work looks like. Reflexive bears get to be "right eventually" by sitting out, and pay for it in compounding opportunity cost. Reflexive bulls get to be right when the cycle cooperates and badly wrong when it does not. The third path, the one that requires the underwriting, is to build a book where each position holds up on its own, and where any macro thesis, including a strong one, is a tailwind rather than a load-bearing wall.

Confidence in a thesis is not the same as concentration around it. The work that produces the first should also discipline the second.

About Erenda Capital Management

Erenda Capital Management LLC is a registered investment adviser. Registration does not imply a certain level of skill or training. For complete information about advisory services, fee schedules, and material risks, please refer to our Form ADV Part 2A, available upon request by contacting info@erendacapital.com.