SHRIRAM Mutual Fund

GOALS ANEK, MUTUAL FUND EK


# THE AI BUBBLE DEBATE

## A UNIT-ECONOMICS LENS

In our Q1 FY27 commentary we promised a separate look at the AI capital-expenditure question. The purpose of the note is not to pick the winner of the model race but to answer the single question our investors keep putting to us. Is the trillion-dollar AI build-out a rational investment or a historic misallocation of capital. Our central contention is a simplifying one that the entire "bubble or not" debate collapses, into one question of unit economics. Does the revenue thrown off by the compute earn an adequate return on the capital sunk into building it, before that capital depreciates? Everything else is just headline numbers.

## The capex, and the return it must earn

The scale is without precedent. The four large hyperscalers committed ~$1.08 trillion of capital over 2021–25, and 2026 guidance alone is ~$725bn, up ~77% year-on-year. Goldman Sachs now estimates a ~$5.3 trillion cumulative capex over FY2025–30. Roughly 60% of an AI data centre's cost is GPU (Graphics Processing Unit), but the true binding constraint is power. A single AI-grade gigawatt costs $20–50bn and draws the electricity of ~750,000 homes.

The return hurdle is where the debate exists. Independent frameworks triangulate a required ~$600–650bn of new annual revenue just to earn a 10% return on what is being built (J.P. Morgan, Sequoia), rising to Bain's ~$2 trillion by 2030 estimated against today's generously-defined AI revenue of ~$50–150bn, a shortfall of ~4× to 13×. Crucially, this is not a solvency risk as in the dot-com or telecom busts. The build-out is largely self-funded by the most cash-generative companies on earth. The real risk is return on capital. The possibility of years of depressed ROIC (Return on Invested Capital) and possible write-downs if the returns cannot justify the massive capex.

## Who is spending, and how they aim to earn it back

Before turning to the unit economics it is worth seeing who is actually writing the cheques and how each of them intends to get the money back, because the routes differ more than the shared capex headline suggests. Most of the big spenders are landlords of one kind or another, renting out compute. The striking exception is Meta, which sells no cloud at all and instead earns its return internally, by pointing better AI at its own advertising machine.

| Company | Primary way it monetizes AI spend | Primary focus |
|---|---|---|
| **Microsoft** | Azure AI capacity, plus Copilot seats bundled into Microsoft 365 and enterprise software | Rented compute + software subscriptions |
| **Amazon (AWS)** | Renting AI infrastructure with in-house Trainium chips and GPU instances | Renting raw and managed compute |
| **Alphabet (Google)** | Google Cloud / Vertex rental, plus AI woven into Search and Ads | Cloud rental + its own ad engine |
| **Meta** | Better AI feeds ad targeting with better ranking and engagement | Internal return via advertising. Just announced foray into cloud |
| **Oracle (OCI)** | Leasing dedicated AI capacity on long, take-or-pay contracts (e.g. OpenAI / Stargate) | Renting compute to a few large labs |

**The two frontier labs take opposite bets.** Both are circling trillion-dollar valuations but with opposite financial DNA. On a run-rate basis Anthropic (~$47bn, May 2026) has now overtaken OpenAI (~$25bn, early 2026), yet OpenAI has pre-committed on a far larger scale. It plans ~$1.4T of infrastructure across Stargate, Oracle, Nvidia, AMD, Broadcom and AWS. The bet being that owning the biggest compute footprint is itself the moat. Anthropic has taken the opposite bet where it stayed lighter on capex (~$380bn of compute deals), lean on a higher-gross-margin enterprise-and-API mix, and letting the capital efficiency compound. Anthropic now guides to operating break-even by 2027–28 against OpenAI's 2029–30. Both are credible strategies, but they are not the same risk.

## Two ways to monetize the GPU

To see why the whole debate reduces to unit economics, we model a single $10bn AI cluster and see two different ways to monetize it. That $10bn buys roughly 170,000 top-end GPUs drawing about 243 MW at full tilt. The chips are only part of the bill. About 60% goes on the GPUs and the high-speed networking that stitches them together. The rest goes into the building shell and racks, power wiring and electrical gear, backup generators, cooling and HVAC, and insurance. Once it is built, running it costs comparatively small with the single largest running cost by far is electricity. This is precisely why power availability, not GPU supply, is the constraint that ends up gating the whole thing. With the hardware fixed, the only real choice is what you sell off the top of it, and there are two ways.

**Rent the GPU (the landlord model).** The simplest option is to be a landlord. Rent the chips out by the hour and let your tenants worry about what to run on them. At today's scarcity premium the newest GPUs fetch around $4.50 per chip-hour, and a full cluster pays its $10bn back in roughly three years. What makes the landlord model actually work, though, is a constant refresh cycle sitting on top of a healthy second-hand market. You keep buying the newest chips, because those command the premium as they are what tenants want for training and serving the best models. Crucially, the chips you retire do not become scrap. They hold a decent salvage value because they can be cascaded down to cheaper, low-complexity 'inference' work and keep earning at a lower rate. So the fleet runs on two tiers at once. The latest GPU rented at a premium for top-tier workloads, the older GPU rented at a discount for simpler workloads. That resale floor and the extension of usable life of GPUs is the whole ballgame. The rental premium earned today steps down gradually as chips age rather than falling off a cliff. If retired GPUs cannot be economically rented for low complexity tasks, the landlord economics would collapse.

**Sell the tokens (own the model).** The other option is not to rent the chips out at all, but to run your own AI model on them and sell what it produces via tokens. This can bring in several times more per chip, but it only works under tight conditions. You need to own a top-tier model that customers will pay a premium for, rather than serving a commoditized open-source model or a last-generation mid-tier one

that anyone can run cheaply. Utilization on these top-end clusters is often surprisingly low. The cluster holds capacity for peak demand and for the highest-value queries. This monetization strategy leans on pricing power and not on volume. The pricing power lasts only as long as your model is clearly better than what is freely available. The instant open-weight models catch up and become 'good enough', a large slice of low-complexity work drifts away to them, because few will pay a premium for capability they do not need. This has therefore resulted in a treadmill where frontier labs keep shipping newer and materially better models at an ever-quickening cadence, steer as much traffic as possible onto your highest-capability, highest-priced models. The idea is preserve a visible daylight in capability over the open-weight alternatives. The premium survives only as long as that difference in model capability does.

How long that capability difference can be defended is itself an open question. While the gap between top tier models from frontier labs and open weight models has narrowed recently, we think that the trend is about to reverse for two primary reasons. The first is anti-distillation. The frontier labs are increasingly building in measures that make it harder to train a cheap copycat simply by learning from a top model's outputs which has historically been the fastest shortcut for the chasers. The second is hardware access. The very best chips are subject to tightening export controls and allocation, so the groups behind open-weight models may struggle to get enough top-end compute to train at the frontier at all. If either of these works, the 'good enough' open weight models could stall a generation or two behind and the premium tier keeps its daylight for longer. If not then we expect a severe compression in the valuations of frontier labs.

It is important to highlight one caveat sits underneath the entire token calculation that none of it includes the cost of building the models in the first place. The billions spent on research, training runs and scarce talent to produce each new frontier model sit completely outside the cluster's $10bn price tag. A token business can look extraordinarily profitable on the hardware alone and still lose money once the bill for staying on the model treadmill is counted.

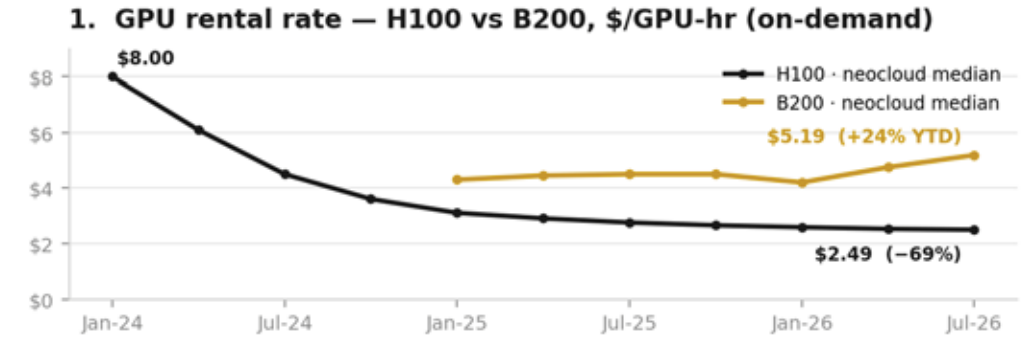
## So which model is better?

While there is no clear answer, understanding of the unit economics helps us in reducing the number of monitorables to a handful of things. Renting of chips/compute or infrastructure-as-a-model can look lucrative if GPU utilization is kept high and a scarcity premium makes to earn higher and for longer than what is assumed today. The token-as-a-service model looks significantly better if the demand for top tier model tokens accelerates faster than how quickly costs for mid-tier model falls. This is exactly why we track the three live signals below rather than settle on a single verdict.

## Collapsing the dashboard to three numbers

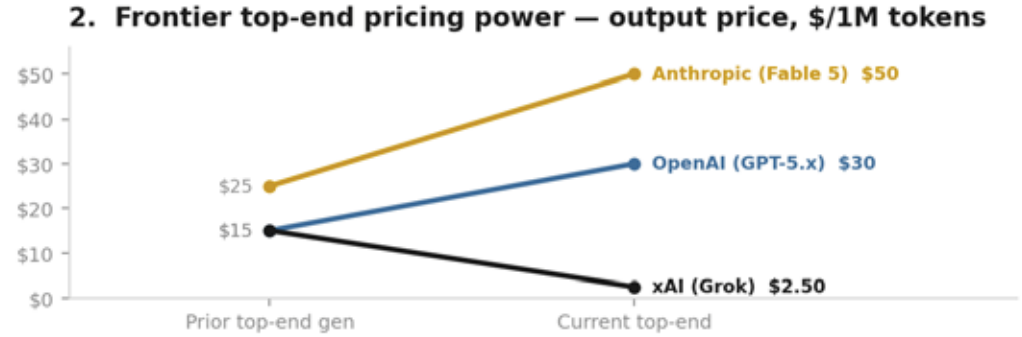
Everything above including the capex headlines, the private valuations and the balance-sheet strategies reduces to essentially 3 monitorable. The unit economics of the entire AI spend hinges on these 3 variables. Two of them are on the supply side of intelligence and one is the price of intelligence itself.

**1. GPU rental rate.** The price of the "rent compute" model, and therefore the ROIC of the whole infrastructure layer. Today the H100 neocloud median is ~$2.4/GPU-hr. Down from ~$8 at its Jan-2024 launch as 300-plus new GPU clouds flooded supply, while the B200 sits at ~$5.2, up 24% YTD. Whether rental rates settle at the inference floor at about $1.25/GPU-hr or crash through it (oversupply, misallocation). Rentals for NVIDA Rubin, in H2-2026, is the next inflection.



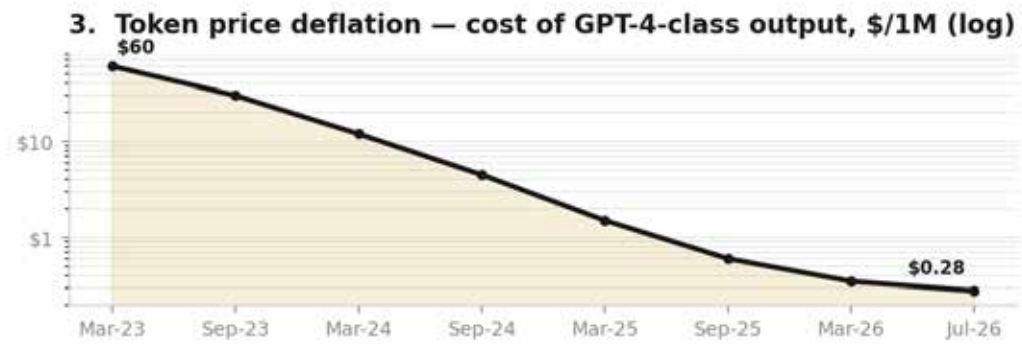

*Source: Silicon Data GPU price indices; public provider listings (CoreWeave, Lambda, RunPod, Vast.ai). Shriram AMC Research.*

**2. Frontier top-end model pricing power.** Whether the model owner can hold price on its best model. Whether frontier intelligence is a franchise or a commodity. Currently top-of-stack pricing is diverging. OpenAI raised GPT-5.5 output to ~$30/M (a bet that agentic workloads justify a premium). Anthropic pushed its ceiling higher still, pricing the new Mythos-tier Fable 5 at $10/$50 (double Opus 4.8, and above GPT-5.5 on output), and xAI is slashing (Grok output down ~83% in twelve months). Stable and rising prices signals a durable franchise and underwrites the token-selling economics. A broad capitulation would signal commoditization of intelligence and severely undermine Tokens-As-a-service model.




*Source: Anthropic, OpenAI, xAI official pricing pages; Fable 5 / Mythos 5 list $10/$50. Prior-gen est. (E). Shriram AMC Research.*

**3. Token price deflation.** The rate at which the value of a unit of intelligence falls. The hurdle that volume must clear. Cheap-tier output tokens have deflated ~90% over 24 months, and the cost of a given capability halves roughly every 12–18 months on a Wright's-law curve. This cut both ways as it is bullish for application-layer margins and demand elasticity, bearish for per-query differentiation. The capex thesis survives if and only if volume growth outpaces per-token price collapse. As of today, it does by a wide margin.




*Source: a16z "LLMflation"; company pricing pages. Illustrative, log scale. Shriram AMC Research.*

## So, is it a bubble?

### Our verdict

On the evidence today, it is not a solvency bubble since the spend is overwhelmingly self-funded by the most cash-generative companies on earth. Hence, what distinguishes this cycle from the dot-com and telecom busts. It is instead a return-on-capital question, and that question is still genuinely open. The unit economics work today at the cluster level under both business models, with the ranking between renting GPU and selling tokens set by refresh cadence, utilization and the path of token demand rather than settled in advance. The risk is that price deflation outruns volume growth, or that a winner-takes-most outcome strands the capital of the other who sunk in massive amounts in capex. We are not obliged to call the outcome. We are obliged to watch the three prices above and, in the meantime, to position our Indian portfolios where the answer matters least.

## Implications for India: Power overweight, selective capex beneficiaries

India has no listed frontier lab and no hyperscaler. Our only direct exposure to the debate above runs through Indian IT, which we treat as a demand-side risk and have addressed separately in the main commentary. Where India does have investable, listed exposure is Power. Power is one layer of the AI stack that is physically constrained, does not deflate, and is indifferent to who wins the model race. The binding constraint on the entire build-out is not GPU but electricity. Whatever the resolution of the unit-economics debate, whichever lab wins, however fast tokens deflate, electricity will get consumed. That is the intellectual basis of our standing overweight on the Power theme, with exposure spanning generation, transmission, transmission EPC, power financiers, diesel-genset manufacturers and the wires & cables ecosystem. For India, the AI capex cycle is first and foremost a power-demand story.

In Capital Goods where we remain underweight at the headline level, we are selectively buying the capex beneficiaries. The “picks and shovels” of the physical build-out that read through to Indian manufacturers, namely the data-centre power and electrical-infrastructure chain (switchgear, transformers, grid, cooling and cabling). These are names levered to the physical, gigawatt-scale roll-out rather than to the token economics layered on top of it.

While the positioning is not entirely agnostic to the bubble question, it is immune to which frontier lab comes out on top and in what form the tokens are consumed. If the unit economics prove out, demand compounds and so does the expected capex. If demand for tokens disappoints, the self-funded build-out still likely continues, but in a smaller way for some years anyway. It is the corner of the AI trade where India can be long the physical infrastructure while being a bystander everywhere else.

**Prateek Nigudkar**
**Sr. Fund Manager**
**Shriram Mutual Fund**

Data compiled by Shriram AMC Research from company filings and pricing pages, Silicon Data GPU indices, and sell-side estimates (Goldman Sachs, J.P. Morgan, Sequoia, Bain). All figures USD unless noted; as of early July 2026.