Intelligence Got Cheap. Trust Didn't

Dear Friends,

Our fund is one month old. Before anything else, thank you for trusting us with your savings. The title of this letter tells you where I believe the world is heading, intelligence *(the machine kind)* is getting cheaper by the week. Trust is not. We intend to be on the right side of that divide, both in what we own and in how we deal with you.

If you looked only at stock prices this past month, you would conclude that nothing happened. Underneath, plenty did:

- The US-Iran war restarted. Iran struck three oil tankers in the Strait of Hormuz on July 6–7, and by July 8 the truce was over.
- Oil rode a roller coaster. Brent crude slipped below $70 a barrel on July 1, roughly where it stood before the conflict began. Then snapped back toward $80 within two weeks when the truce collapsed. All this just months after an April peak above $120.
- The Russia-Ukraine war worsened. July was the deadliest month for Ukrainian civilians in more than four years.
- Chinese AI labs released three models in three weeks that are nearly as capable as the best American ones, at one-third to one-fifth the price.[1]

We wrote in January that the world is a messy place and always will be. You cannot predict the storms, you can build a ship that handles them. Preparation matters far more than prediction.

Most of this letter is about AI, what is actually going on, and where I believe the market has the story wrong. But first, some housekeeping.

Performance update

The fund began operations on 2nd July 2026. We will start reporting performance from the next quarterly letter onwards, once we have a full quarter behind us. You will, however, be receiving a monthly Statement of Account (SOA), so you know the value of your holdings.

[1] CNBC, "Chinese AI models are gaining ground with U.S. companies," July 7, 2026; Axios, July 16, 2026 (Kimi K3); HPCwire, July 21, 2026. Moonshot prices Kimi K3 at $3/$15 per million input/output tokens, versus Claude Fable 5 at $10/$50.

Portfolio update (till 5th August, 2026)

We do not own a “portfolio” in the way most funds use the word. We own a collection of individually chosen businesses, each picked from the bottom up, each standing on its own quality, its own growth, and its own reason to be re-valued by the market. Most fall in the ₹1,000-10,000 crore market-cap range.

Here is my honest expectation, and I would rather set it now than apologize for it later. If we do our job well, our returns will come in lumps, not in a steady monthly drip. A stock we own can sit still for a year and then move sharply in three months when the market changes its mind about the business. I can usually see the reasons a re-rating ought to happen (we happily suffer from an abundance of such ideas) but I can almost never tell you when it will. I do not control the mood of the market, and I do not try to. **We would much rather have a lumpy result that averages well than a smooth one that averages poorly.**

We currently hold 14 stocks.

Two are one-year trades. One is expected to benefit from a fast-developing bottleneck. The other has been unfairly punished due to prolonged seasonal factors, making it very attractively priced at current levels. Neither is a business of the highest quality. Ordinary trouble can find them, as it eventually finds every ordinary business, so they are sized small and watched closely, as trades should be.

The other twelve are investment holdings. They span, Power T&D, Construction Equipment, Auto Components, Ratings/Certification, Commercial Vehicles, Two-Wheelers Components, OEM/ODM, Healthcare, and Software.

The above sector split does not really mean much to me. We did not buy sectors. We bought twelve specific businesses, each is what we believe a high-quality business at attractive price.

AI: Reading the Five-Layer Cake

The simplest way we know to think about the AI industry is as a cake with five layers. From the bottom up:

1. **Land and power**- the real estate, electricity, and cooling that data centres need.
2. **Capital goods and silicon**- the chips, transformers, switchgear, and machinery.
3. **Models**- the “brains” themselves: ChatGPT, Claude, Gemini, Grok and their fast-rising Chinese cousins.
4. **The harness**- the wrapping around a brain that turns it into a dependable worker. This is the most important idea in the letter, more on it shortly.
5. **Applications**- the finished products that you and I actually use.

Value is settling very differently across these five layers, and the past two months have been unusually revealing.

Layers one and two- still scarce

Whatever drama happens above, the bottom of the cake remains in short supply. The five largest American technology companies plan to pour up to roughly $690 billion into AI infrastructure in 2026, nearly double last year. Powered land, grid connections, transformers, switchgear, cooling, all of it is needed no matter which AI lab "wins." This is the sturdiest, least fashion-dependent layer of the cake.

The drama this quarter was one layer up.

Layer three- the brains are becoming a commodity

Two things happened at once.

The American labs slowed down. A June executive order now requires a government review before the most powerful new models are released (GPT-5.6 went first to a government-approved group).[2] Caution is the new default in America.

The Chinese labs sped up. Three top-class models in three weeks, most of them "open-weight", meaning any company can download the brain and run it on its own computers, much as you might buy a generator rather than depend on the power company. And they are priced at a fraction of the American offerings.

Put the two together and the conclusion writes itself, **the cost of intelligence is going to fall, materially and for a long while.** When something becomes cheap and interchangeable, we call it a commodity.

One more development is worth understanding, because it explains a paradox you will read about in the papers. Modern AI "agents" no longer answer once and stop. They check their own work, consult sources, and retry until the answer holds together, exactly the way a careful analyst re-works a number until it reconciles. AI is billed roughly by the word ("tokens"), and one of these self-checking agents uses about 15 times the tokens of a simple chat. So while the price per word is collapsing, the words used per job are exploding.

Four observations from the past two months strengthen my conviction on move towards Open-weight-

- **Privacy matters more, not less.** Companies increasingly want the intelligence brought to their data, not their data shipped out to someone else's intelligence. Models a firm can run on its own machines win on exactly this point.

---

[2]Executive Order of June 2, 2026 established a pre-release review process for covered frontier models (TechTarget, June 5, 2026). GPT-5.6 was released first to a government-approved cohort on June 25, 2026.

- **A government can now switch off a brain.** On June 12, a US export-control directive took Claude's top model (Fable) offline worldwide. The first time a deployed frontier model has been shut off by government order. Access returned on July 1, with new safety controls bolted on.[3] **Whether a model is available to you is now partly a national-security decision made in Washington.** We flagged this direction last quarter, within three months, the precedent travelled from selective release to outright removal.
- **The returned Fable product feels weaker.** This is our own experience rather than a formal benchmark. Since the July 1 restoration, our team has quietly shifted much of its work to an alternative models.
- **The Chinese models are winning real business.** One recent Chinese model scored within a whisker of the best American model on a closely watched test, at roughly a fifth of the price. American companies now route over 30% of their AI usage through Chinese open models, up from about 4.5% a year ago.[4]

Layer four- the harness, where the value goes

If the brains are becoming a commodity, where does the advantage move? To the layer just above them, which the industry calls the "harness."

Think of it this way. **A model without a harness is a brilliant employee with amnesia**, **every morning is his first day at work.** He does not know which supplier to trust, when to escalate a problem, or what "good" looks like in your shop. All of that lives in the heads of your senior people. The harness is the checklists, the connections to your systems, and above all the permanent memory, that lets a firm write its hard-earned know-how down and wire it into the machine's daily work.

Last quarter we wrote that judgment is not downloaded, it is earned. The harness is how a firm bottles what it has earned. **Firms that build a culture of writing down what they know will get far more from these machines than firms that do not.**

Building a good harness takes two rare things at the same time:

- capable engineers, and
- business people with the clarity and taste to shape the tool around the actual work.

The industry learned the taste part the expensive way. The first instinct of 2026 was brute force, burn more computing and call it productivity. The practice earned a nickname, "tokenmaxxing". Uber burned through its full-year AI budget in four months and capped engineers at $1,500 a month, and

[3]Anthropic statement, June 12, 2026; US Commerce/BIS export-control directive lifted June 30; access restored globally July 1 with a new safety classifier. First instance of a deployed frontier model being suspended by government order.

[4]OpenRouter/Vercel data via CNBC: US companies now route over 30% of open-router tokens through Chinese open-weight models, up from ~4.5% a year ago; GLM 5.2 landed within a percentage point of Opus 4.8 on a closely watched agentic benchmark at roughly a fifth of the cost.

Meta shut down its internal usage leaderboard within days of launching it.[5] What replaces brute force is context, giving the machine the right information, not the most information.

One trap deserves special mention. To save money, most systems compress their stored memory into short summaries, then work from the summaries. A recent study found that after heavy compression, a system could recall only 40% of the facts it had filed away, the other 60% were simply gone.[6] Imagine an assistant who summarizes your files and then throws away the originals. **Right information, complete information, only then a nuanced answer.**

Two further points about this layer.

- First, it holds the crown jewels, the firm's permissions, its memory, its private know-how, so security matters here more than anywhere else.
- Second, complete dependence on a single, fully integrated AI provider is, in my view, a serious risk. The June shutdown was a live demonstration of why. A well-built harness makes the brains swappable underneath it. And that is precisely why, as the brains commoditize, the harness is where the value should accrue.

Layer five: applications, finally a great time to build

With capable intelligence now cheap and swappable underneath, it is finally a great time to be building AI products. The model companies see this too. A price war is on. Flagship model prices are already down more than 90% from their 2023 levels, steeper cuts are being weighed, and both large American labs are preparing for public listings, where profit margins will be scrutinised.[7] As margins on raw intelligence shrink, I expect the labs to defend themselves by moving up the cake into applications, **competing with the very builders who rent their brains or by simply billing more words for the same work.**

So expect a wave of new AI products, for consumers and for businesses. In this layer the game will be won on distribution and retention. Notice, too, that the billing unit is already shifting from "token used" to "job done" (for example- $0.99 per resolved customer query, $2 per conversation, and so on.[8]) Every completed job generates data that makes the next job better, so usage and quality compound each other.

---

[5]Fortune, May 28, 2026: "Tokenmaxxing is dead." Uber exhausted its full-year 2026 AI budget in roughly four months and capped spending at $1,500 per tool per month; Meta shut its internal token-consumption leaderboard within days of launch.

[6]"Facts as First-Class Objects: Knowledge Objects for Persistent LLM Memory" (arXiv, 2026): after ~37x compaction of 2,000 stored facts, the agent recalled only 40%; the other 60% were lost outright.

[7]WSJ, June 10–11, 2026, via follow-on reporting: OpenAI weighing steep API token price cuts; flagship-token prices down over 90% since GPT-4's 2023 launch pricing; both large labs preparing public listings where gross-margin narratives matter.

[8]Intercom Fin at $0.99 per outcome (Salesforce agreed to acquire Fin for ~$3.6bn in June 2026); Salesforce Agentforce at $2 per conversation plus per-action credits; Zendesk at ~$1.50 per verified resolution; HubSpot at ~$0.50 per resolved conversation.

What cheap, tireless intelligence changes

For the first time, **we have intelligence that is cheap and has, effectively, infinite stamina**[9].
I am not yet ready to hand any machine full discretion over anything critical, and I doubt I will be soon. But these agents have already become an extra junior member of our team. Some of what changes-

- **Diligence quality improves.** A machine can read and map the financial statements of a joint venture of a subsidiary of a target company, the kind of nested disclosure a human process can miss.
- **Work that never made economic sense, now does.** Tasks too tedious, too voluminous, or too low-yield per hour for a person can now be pursued.
- **The tails get covered.** The long tail of documents, customers, and questions, previously not worth anyone's time is now serviceable.
- **Cross-checking gets automated.** An agent can search, verify, and search again until independent sources agree. The slow grind of triangulation that is the heart of diligence. The investor Li Lu describes this job as investigative journalism. The machines can now do legwork.

A small live example

Our regulator, SEBI, now requires every regulated entity's investor-facing websites to meet accessibility standards. Vendor quotes to fix the Suraag Capital site ran to about INR 80,000. Instead, I sat down with an AI coding agent and rewrote the site myself, audit-ready in two hours, without our tech team touching the code. The auditor's "fully compliant" certificate arrived within four days.

An entire class of technology work, small scripts, integrations, routine fixes has commoditized to the point where the workflow simply can move in-house. We are seeing this movie at scale, as IT services work moves from third-party vendors to captive GCCs.

It is tempting to stretch this logic to every software company. The market has done exactly that. I believe the market is wrong.

Where I believe the market is wrong

The market's default narrative has become that AI kills Software companies. My view is closer to the opposite, **Software businesses with non-tech moats are winners in AI.**

Consider the enterprise software firm serving a heavily regulated industry. Ask why it wins deals, and the honest answer is, relationships, reputation, and brand first. Then product second.

---

[9]Andrej Karpathy, X post, Jan 26, 2026, on AI as a tireless junior partner: intelligence with "infinite stamina," at rapidly falling cost.

- Its real asset is understanding: its customers' unwritten rules, exceptions, and habits, accumulated over years and living in the heads of its product, technology, and sales teams. **A lean new competitor can copy the code. It cannot shortcut the years of listening.**
- Contrary to popular opinion, I do not see their profit margins shrinking either. These are large, critical, highly regulated workflows that go live only after every human approval is in place. Vendors compete on trust, change management, and years of experience. A cheaper tool does not buy any of the three.
- In fact, these same firms now get AI-powered engineering teams of their own, faster and leaner, so the same headcount ships more. They gain the efficiency without inheriting a competitor who can undercut them.

  Intelligence got cheap. Trust didn't.

Closing Thoughts

Wars re-ignite, crude whipsaws, models get banned and un-banned, all in a month in which, on the surface, *"nothing much happened"* to stock prices.

Beyond all of this, our work continues quietly. We remain deeply focused on businesses and long-term learning. I feel fortunate to spend my days pursuing this craft and to get a little better at it each day.

बस लगे हुए हैं। (We're steadily at work)

Best Regards,

Rishabh
Founder & CEO, Suraag Capital
rishabh@suraagcapital.com | +91-9818924000