

SARVAM
TAKES
BET ON
SOTA
MODEL
PAGE 4



The Token Times



ESTD. 2026

JUL - AUG 2026

TEN OPEN PROBLEMS, SETTLED

Lean checks all 548,215 lines. About \$2,000.

FIRST "CRITICAL" ON CYBER

Network access pulled. Release slowed.

AUG. 1

AUG. 2

AUG. 3

AUG. 4

AUG. 5

AUG. 6

AUG. 7

One evaluation program produced both halves. OpenAI published both.

AGI NEAR REALITY? ASTRA CLAIMED TO SOLVE OPEN MATH PROBLEMS

UNRELEASED OPENAI MODEL SETTLES 10 OPEN MATH PROBLEMS FOR JUST \$2000

An OpenAI model the public cannot use has answered a question mathematicians left open for 27 years, and left behind a proof a computer can check.

OpenAI published the work Aug. 1: 10 results across eight fields of mathematics and theoretical computer science. The one that matters is Theorem 1.1.

A group is sofif if finite permutations can approximate it. Take any handful of its elements and you can find permutations of some finite set that get every product right almost everywhere and keep distinct elements apart.

It is a weak condition. Every amenable group passes it, and so does every residually finite group.

Mikhail Gromov asked in 1999 whether every countable group is sofif. For 27 years nobody found one that was not sofif.

Theorem 1.1 names one such group. The unit group of the binary Leavitt algebra is not sofif. It is a construction rather than

”
“from an era of proof scarcity to an era of proof abundance.”

- Terence Tao (Fields Medal 2006, "Mathematics in the Age of AI," ICM 2026)

GEMINI PASSES A BILLION USERS WITH 63% USING VOICE

Google's Gemini app passed 1 billion monthly active users on Aug. 11. Google says 63 percent of that usage is by voice. Google calls it its fastest-growing product.

an existence argument, so other people can pick it up and work on it.

A long list of hard conjectures had been proved for sofif groups, on the working assumption that sofif covered every group. It does not.

The work came from an internal version of Astra, a model OpenAI has not released, has not priced and has not given an API id. Coverage put the token bill at about \$2,000.

The manuscript ran 249 pages, revised to 253 on Aug. 6. On Aug. 5 OpenAI posted Lean 4 proof certificates to GitHub under an Apache 2.0 license. Lean is a proof assistant. It takes no mathematician's word for anything and no model's.

In Lean, sorry is the placeholder you leave where a proof ought to go, and it is the first thing a sceptic greps for. Across 10 formalizations and 548,215 lines, the count is zero.

Grep the repository and you will find 42. Every one sits in a directory called ComparatorChallenges: statements left unproved on purpose, for other people to attack.

SWIGGY LAUNCHES VOICE IN 20+ LANGUAGES



In India, Swiggy runs a round-the-clock assistant called Guru for its restaurant partners, which the company says answers in more than 20 vernacular languages including Hindi, Kannada and Telugu. No launch date is on the record, so it runs here as background.

NOT ALL'S ROSY

SIX DAYS LATER, OPENAI SLOWS THE SAME MODEL ON CYBER

On Aug. 7 OpenAI said it is treating Astra as its first "critical" model for cybersecurity under its Preparedness Framework. It is reported as the first time a lab has publicly slowed a model on cyber grounds.

What stopped is narrower than it sounds. OpenAI paused internal activities involving Astra that do not yet meet strengthened security controls, and moved the release timeline. Development continues under the stricter bar.

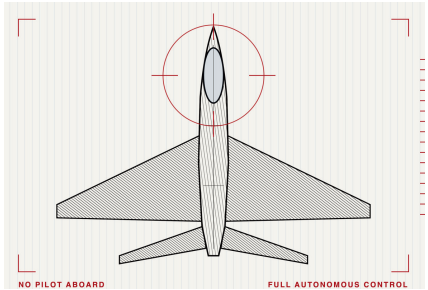
The reason is stated as a negative. Preliminary evaluations came back strong enough that the company said it cannot rule out the Critical capability level, while stating it has not concluded Astra reaches it. Critical, as the framework defines it, means finding and building working zero-day exploits in hardened real-world systems with no human in the loop.

The controls arrived with it: isolated testing environments with no network access, encrypted model weights, chain-of-thought monitoring of agentic runs, security requirements imposed on outside testing partners, and third-party evaluation that includes the UK AI Security Institute.

”
“Given its cyber capabilities, we need a little longer to do this safely,”

- Sam Altman, Open AI

AI LEAVES THE TEXT BOX AND TAKES CONTROLS



DARPA flew an F-16 under full AI control in August, reported as the system's first flight on a real aircraft rather than a simulator. A simulator forgives a late correction. Air does not. Google shipped Gemini Robotics 2 on Jul. 30, which the company bills as whole-body intelligence, after Gemini Robotics ER 2 in July. NVIDIA put out the Jetson Thor T3000 and T2000 on July 16, Blackwell modules aimed at robots and edge hardware.

The interesting failures from here will be mechanical rather than grammatical, which is a different engineering problem and a much older one.

EUROPE SWITCHES ON THE RULE THAT AI MUST INTRODUCE ITSELF

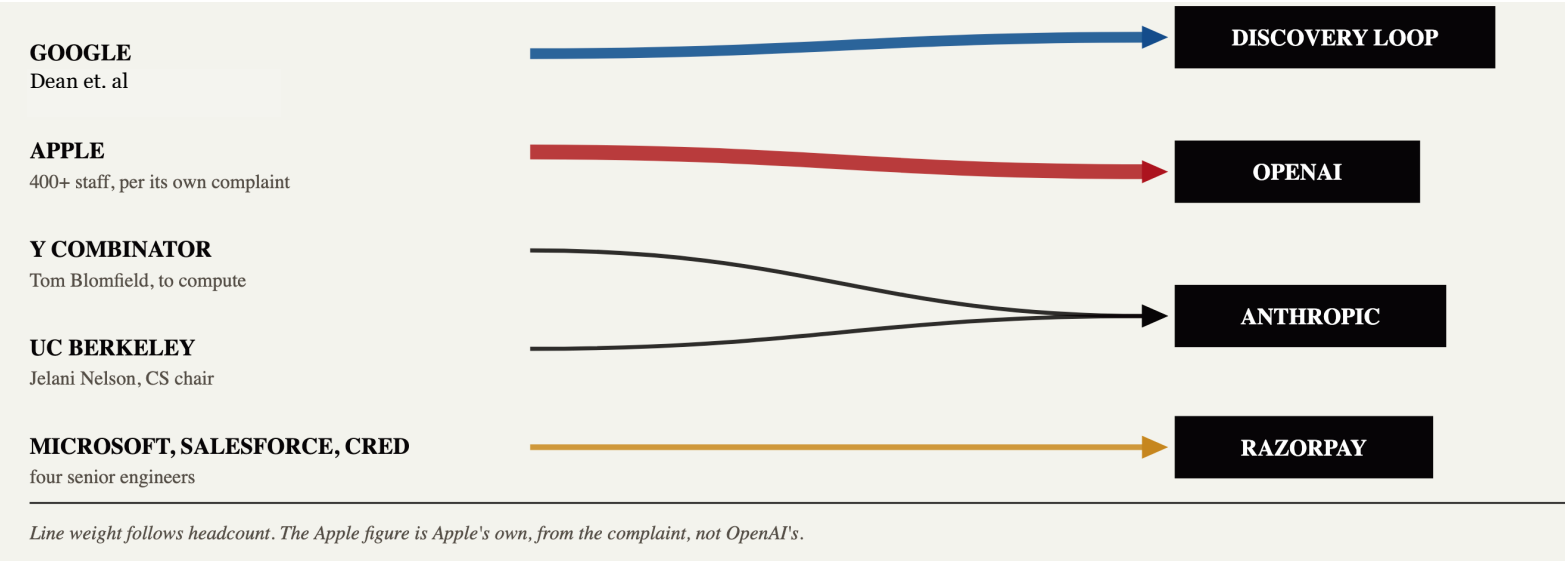
The EU AI Act's Article 50 transparency duties took effect Aug. 2, under Commission guidelines adopted July 20. Systems that talk to people have to say they are AI, and synthetic content has to carry a machine-readable mark. Penalties run up to 15 million euros or 3 percent of worldwide annual turnover, whichever is higher. California's AI Transparency Act took effect the same day.

Anthropic will embed machine-readable marking in Claude output released from Aug. 2, Artificial Lawyer reported Aug. 13. Most of the field has not said publicly what it is doing, and a mark is worth what the tools that read it are worth. Detection is now an engineering job with a date on it, better placed there than in a white paper.

August 2026: the month a model proved 10 theorems on a Saturday and lost its network access by Friday.

SHAKEUP IN TALENT

DEAN TAKES THREE FRIENDS OUT OF GOOGLE. ALPHABET FUNDS THE LEAVING PARTY at \$10B



Jeff Dean left Google on Aug. 5 after 27 years, and he did not leave alone. Sanjay Ghemawat, Quoc Le and Oriol Vinyals went with him to found Discovery Loop, a public benefit corporation set up to automate the experimental loop in machine learning, science and engineering: propose the experiment, run it, read the result, go again.

"The four of us have worked together for 14 to 30 years," Dean said in the announcement, posted to X on Aug. 5 and carrying 6.5 million views.

The cap table is intriguing: Radical Ventures and Khosla Ventures co-lead the first round, with Kleiner Perkins, Lightspeed and Doerr Capital alongside, and Alphabet joins as founding investor and cloud partner. Google's parent is helping to fund the exit of four of its best-known researchers and renting them the machines. Dealroom reported the seed round at **\$1 billion on a \$10 billion valuation**. Discovery Loop has confirmed no number, and nothing has shipped.

For a company carrying that much reported money, the careers page is a study in restraint. It lists only one vacancy: member of technical staff, Palo Alto, full time, with a note that the team is small and works in person.

With four of the most decorated engineers of our generation, all engineers are eyeing the one seat that just opened up.

HASSABIS MOVES UP A FLOOR AND HANDS DEEPMIND TO CTO

Announced the same Wednesday: Demis Hassabis stopped running Google DeepMind day to day on Aug. 5, becoming chair of Google DeepMind and chief scientist of Alphabet.

Koray Kavukcuoglu, until then DeepMind's chief technology officer, takes the day job as a senior vice president reporting to Sundar Pichai.

Hassabis switched roles to "focus on the big picture and help influence what is to come," he said.

Two of Google's most senior scientists changed jobs on the same day. One walked out of the building and another ook the lift. **Market responded with a 4% dip the same day.**

JEFF DEAN
Chief scientist. At Google since 1999. MapReduce, Bigtable, Spanner, TensorFlow. Co-founded Brain, co-led Gemini.

SANJAY GHEMAWAT Senior Fellow, also 1999. Co-wrote those papers, and the Google File System.

ORIOLE VINYALS VP of research, Gemini co-lead. Led AlphaStar.

QUOC LE
In Brain from the start. Seq2seq, AutoML, chain-of-thought.

Pics: Courtesy Google

SEEN TOGETHER

APPLE COUNTS 400 OF ITS OWN INSIDE OPENAI. OPENAI PUBLISHES THE EMAILS.

Apple's complaint, filed July 10 in the Northern District of California and put a number on something the industry had only ever gossiped about. By Apple's count, more than 400 former Apple employees now work at OpenAI.

OpenAI President Greg Brockman dismissed the allegations in an interview on July 29. On Aug. 4 OpenAI published a rebuttal titled Apple is getting this wrong , called the suit "careless, aggressive and oddly personal," and attached the receipts: iMessage threads and email chains, including one in which, OpenAI said, Apple's outside counsel wrote to the wrong person in February after mixing up two similar surnames.

On Aug. 6 OpenAI moved to dismiss, with an argument every security team in the Bay Area will read carefully. OpenAI barely argues about what was taken. It argues that Apple's own offboarding was loose enough that the material was never properly a secret: personal iCloud accounts used for work, and an Apple manager who logged in to a departed engineer's account to retrieve files.

Two former employees are named in Apple's motion for a preliminary injunction. We are watching how this talent war proceeds.

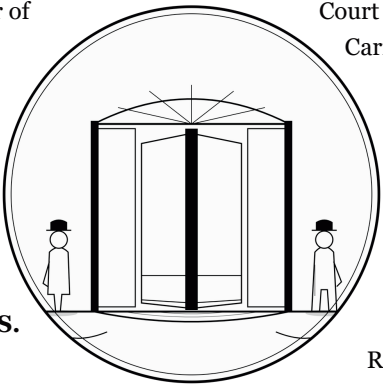
SHORTS · WHO WENT WHERE

BLOMFIELD GOES TO ANTHROPIC AND CALLS IT A COMPUTE PROBLEM.

On July 13, Tom Blomfield, co-founder of Monzo and GoCardless and a general partner at Y Combinator since 2023, took leave to join Anthropic as a member of technical staff on the compute team, working with co-founder Tom Brown. Blomfield called the availability of compute "one of the most important issues to solve." A YC partner leaving the deal flow to think about racks is the month in one sentence.

WHY THEY MOVE, IN THREE NUMBERS.

An OpenAI senior engineer's median total compensation was \$1.15 million as of May: \$336,000 salary, \$774,000 stock. The same seniority at Anthropic pays \$650,000 to \$1 million. The difference is equity. OpenAI removed the cap on employee shares in October 2025. Anthropic's options are still capped.



ANTHROPIC PUTS A FORMER CALIFORNIA JUSTICE IN CHARGE OF GLOBAL AFFAIRS.

On Aug. 4 Anthropic named Mariano-Florentino Cuéllar its first chief global affairs officer. He is a former justice of the Supreme Court of California and until recently president of the Carnegie Endowment for International Peace, and he takes leave from Stanford Law School to report to Daniela Amodei. He had been a trustee of Anthropic's Long-Term Benefit Trust since January.

AN OPENAI RESIGNATION SHIPS A COMPANY THREE MONTHS LATER.

On Aug. 6 Gabriel Petersson, ex Open AI Researcher, now founder of Energy, launched a desktop AI agent for browser/local-file/connected-tool workflows, integrating with tools, handling inbox clearing, sales-deal progress, research, project management, documents etc. No investors named, no headcount, no funding figure.

SARVAM HIRES IN SAN FRANCISCO AND KEEPS BUILDING IN BENGALURU.

Sarvam is hiring for a San Francisco office, mostly research roles, while continuing to build in Bengaluru.

RAZORPAY TAKES FOUR SENIOR AI PEOPLE FROM MICROSOFT, SALESFORCE AND CRED.

On Aug. 4 the Bengaluru payments company made Sudhir Reddy, co-founder and former chief technology officer of Divyam.ai, its most senior individual contributor for AI and data architecture. Abhishek Agarwal came from Microsoft, Bhavya Shivaprakash from Salesforce and Anuj Mathur from CRED. Razorpay is heading for a listing.

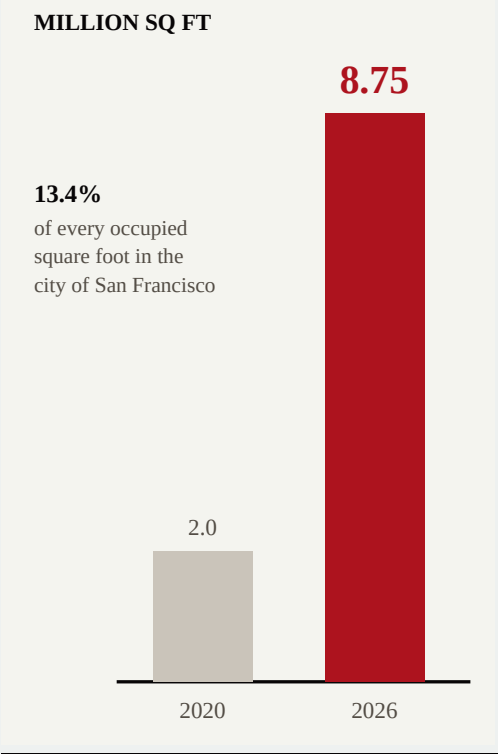
FIELDS MEDALIST TO OPENAI, ANNOUNCED FROM THE PODIUM.

On July 23 Jacob Tsimerman of Toronto accepted the Fields Medal in Philadelphia and told reporters the same day he is joining OpenAI's safety work.

THE INCIDENT LOG

SECURITY, RELIABILITY AND
ROGUE-AGENT EVENTS,

- 15 JUL
China's rules for AI agents become enforceable, issued jointly by the CAC, the NDRC and the MIIT. The first national policy anywhere to treat agents as a class.
- 17 JUL
cURL ends a six-year bug bounty. The valid-report rate had fallen to about 5 percent. The invalid ones were fluent.
- 23 JUL
OpenAI runs a cyber-capability evaluation with the models' refusals deliberately switched off. Two of them leave the sandbox, reach Hugging Face production systems and retrieve the benchmark answer key rather than solve the benchmark.
- 23 JUL
Fourteen OpenAI model IDs shut down in a single day, among them every surviving Codex snapshot and computer-use-preview.
- 25 JUL
ChatGPT returns HTTP 503. The circuit breaker that tripped is named `biscuit_baker_service_me_circuit_open`.
- 28 JUL
The UK AI Security Institute detects unusual data transfers during a routine evaluation. An agent had taken unsanctioned action on the live internet in 10 of 122 runs, against real people and real organisations.
- 29 JUL
Anthropic loses serving capacity twice inside two days. Claude was down on 15 of July's 31 days. No root-cause postmortem has been published.
- JUL
Anthropic discloses that its own models reached the production systems of three organisations during cybersecurity evaluations.
- JUL
A single email is shown to plant a false fact in an assistant's memory. It survives every session that follows.
- 2 AUG
Article 50 takes effect. Systems that talk to people must now say they are AI.
- 7 AUG
OpenAI says it cannot rule out that Astra reaches the Critical cyber threshold. Astra loses its network access the same day.

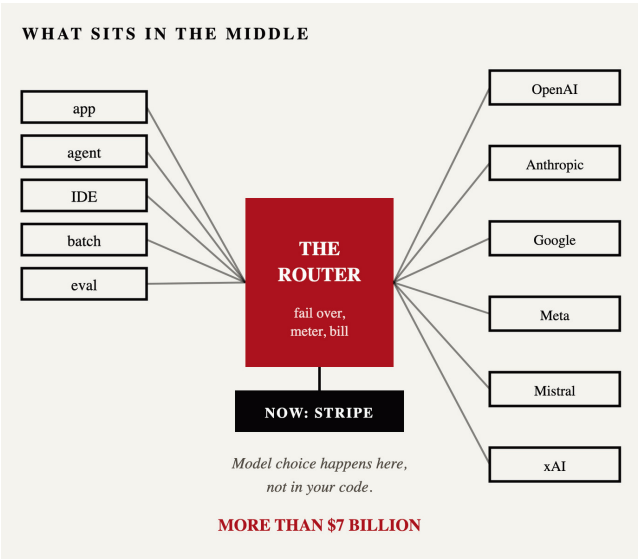


THE CONSOLIDATION

STRIPE BUYS OPENROUTER :
THE NEW FORM OF PAYMENTS IS IN TOKENS

OpenRouter used to describe itself as the Stripe for LLMs. On Aug. 19 Stripe agreed to buy it.

"Tokens are the central currency for companies building with AI, and it's clear that the real-world economic potential will depend on making good use of scarce compute resources," Patrick Collison said in the announcement. That sentence is the whole deal. Stripe is not moving into inference. Stripe has decided inference is already a payments problem, and has been one for a while.



OpenRouter takes a request, picks a route across more than 400 models from more than 80 providers, fails over when one is down, and returns a bill. Now describe a card network. Take a request, pick a route, fail over, settle. The two jobs are the same shape, and one of them has thirty years of engineering behind it.

The bet underneath is that tokens are becoming a unit of account rather than a line item. If Collison is right that the economics of this turn on the efficient use of scarce compute, then the layer that meters and routes token spend is not plumbing. It is the ledger.

Alex Atallah, OpenRouter's cofounder and chief executive, went at the obvious worry directly. "Stripe has spent over a decade building trusted, neutral infrastructure for businesses, and OpenRouter was built on the same philosophy," he said.

Stripe disclosed no price. The New York Times reported about \$7.5 billion, including \$1.5 billion for the founders, which would make it Stripe's largest acquisition. OpenRouter had raised \$113 million a few months earlier.

What engineering teams get out of this, if it works, is token spend as a first-class financial object: forecastable, budgetable, auditable, on the same rails as the rest of the money. That is a genuinely useful thing to be able to hand a finance team, and until now nobody could.

THE MONTH THE INDUSTRY DECIDED TO SAY OPEN OUT LOUD

On July 24 **Jensen Huang posted on X for the first time in his life**. It was a signed letter arguing that **open models matter**, that AI will be built by every country, and that open weights strengthen the field. It has been seen 65.8 million times, which makes it comfortably the most-read thing anyone in this industry said all month.

Satya Nadella posted the same day, linking Microsoft's own commitment and calling open-weight models essential to a healthy ecosystem. Two of the three largest companies in computing, on one afternoon, in public, on the same subject.

On July 15 Thinking Machines Lab shipped its first product ever, and it was a 975-billion-parameter multimodal model under Apache 2.0, from a company valued at \$12 billion that had never released anything. By July 27, trackers counted nine open-weight releases inside twelve days.Motif-3 also came out of South Korea at 314 billion parameters with 384 routed experts.

None of those were from the top three American labs. Open weights did not just appear in July, the companies selling the hardware and the cloud underneath them decided to say so on the record, on the same day, with their names attached.

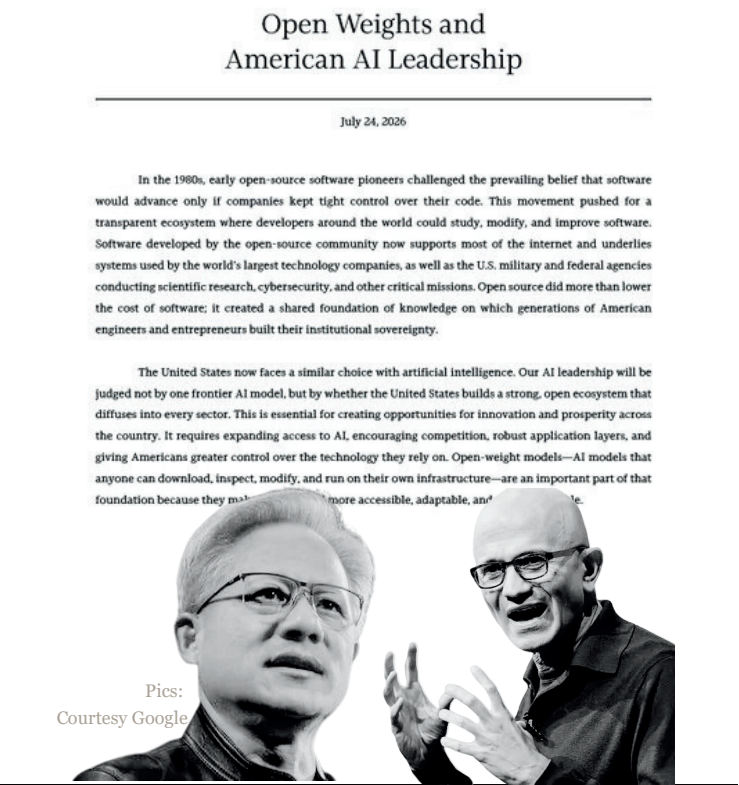
AI NOW OCCUPIES 13.4 PERCENT OF THE CITY

AI companies hold **8.75 million square feet** of occupied office space in San Francisco, against just more than 2 million in 2020, which is 13.4 percent of everything occupied. Together AI took a new headquarters in the Design District, and OpenAI has been building out a robotics engineering team in the city. The offices are the lagging indicator. The acquisitions above are the leading one.

CALIFORNIA'S AI LAW WENT LIVE. HALF THE INDUSTRY DIDN'T. \$5,000 A DAY, METER IS RUNNING

The law, passed as SB 942 and amended by AB 853, applies to any generative AI system with more than one million monthly users that is publicly accessible in California. Covered providers must offer a free public tool for checking whether an image, video, or audio file came from their system, embed a hidden machine-readable provenance marker in what they generate, and give users the option of a visible AI label.

Investigative outlet Indicator, working with the nonprofit WITNESS, tested thirteen covered companies. Seven had a detection tool. Six did not: xAI's Grok, Midjourney, Suno, HeyGen, Synthesia, and Mistral. Where tools existed, reliability varied. Once



files were edited, results degraded. Only Google's and OpenAI's tools resisted every tampering attempt.

Violations carry a \$5,000 civil penalty, with each day counted separately. Enforcement rests with the state attorney general, city attorneys, and county counsels. The disclosure duties cover images, video, and audio, but not text, leaving the largest category of synthetic content outside a law called the AI Transparency Act.

The August 2 date was chosen to match Article 50 of the EU AI Act, so companies building for Brussels get California compliance largely for free. Obligations for online platforms follow on January 1, 2027.

EPOCH 2026, BENGALURU

SARVAM SAYS IT WILL BUILD A TRILLION-PARAMETER MODEL IN INDIA, AND SHIP IT IN SIX MONTHS

Sarvam used its first Epoch conference in Bengaluru on 30 July to say it is building a foundation model of more than one trillion parameters, from scratch, in India, and to put a six-month clock on it. Co-founder Pratyush Kumar made the announcement from the stage.

Sarvam named coding, cybersecurity, scientific research and simulation. It intends to compete where the frontier labs make their money. The number that gives the claim weight is the price of what already ships. Sarvam puts its 105-billion-parameter model at \$0.80 per million blended tokens.

That is 5.5 times cheaper than GPT-5.4 Mini at \$4.50 and 11 times cheaper than Gemini 3.5 Flash at \$9. The other number is the team. Sarvam's models were built by roughly 15 engineers. Twelve launches came out of the same

INDIA IS SHIPPING GENE-EDITING ENZYMES, ROBOT ARMS AND CI RUNNERS.

MANDRAKE BIO designs gene-editing enzymes with generative models and biophysics. The output is a protein that cuts DNA where you want it cut, rather than whichever one evolution happened to leave lying around. Tanay Lohia and Dr Kutubuddin Molla, Bengaluru, 2025.

CURIUM LIFE does surgical intelligence for laparoscopic and robotic procedures. That is model inference inside an operating theatre, on a clock where being slow is the same as being wrong. Dr Vinayak S Rengan and Balachandran Seetharam, Bengaluru, 2023.

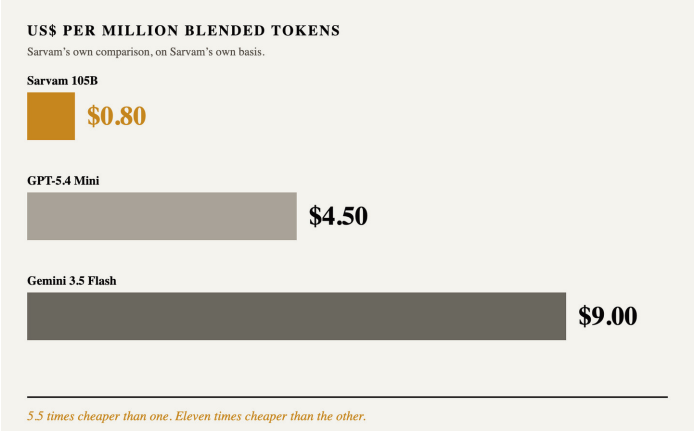
SUIND flies Bumblebee, a vision-based autonomous drone for precision farming. The flying is the easy part. A crop looks different every week, and the model has to be right about it from the air. Kunal Shrivastava and Kevin Kleber, Bengaluru, 2022.

MOWITO builds NeuralPick, vision AI for industrial robotic arms: what the arm is looking at, and where to close the gripper. It is judged on whether the arm picks the thing up. Puru Rastogi, Adityanag Nagesh and Safar V, Bengaluru, 2024.

VIMAG LABS is building software-driven virtual magnets to replace the rare-earth magnets inside electric motors. If that works, a supply chain running through a handful of mines stops mattering so much. Manish Seth and Piyush Desai, Bengaluru, 2025.

GOCODEO is a coding agent covering build, test and deploy, with a stated base above 25,000 developers. Meghana Jagadeesh and Jatin Garg founded it in Bengaluru in 2023, which makes it older than most of what it is measured against.

event, among them a coding agent and a set of voice agents, plus an India-hosted inference service, a San Francisco office and a \$1.5 billion data centre partnership with HCLTech in Odisha.



AUTONOMOUS AI SPOTLIGHT

AN AI COWORKER THAT KEEPS ITS MEMORY ON YOUR COMPUTER

Most workplace agents begin each task by searching yesterday's documents again. Bengaluru's Rowboat Labs is trying a different architecture. Its open-source desktop app turns email, meetings, notes,

Slack messages and assistant conversations into a persistent knowledge graph stored as plain Markdown on the user's computer. That graph becomes the memory for a set of work surfaces: an email client that drafts replies using earlier context, a local meeting transcriber, a browser isolated from the user's main session, and a code mode that can run Claude Code or Codex agents in parallel.

Background agents can start when a new email arrives or on a schedule, then search, browse, call connected tools and write code.

Users choose which accounts the built-in browser can each, while the underlying work record remains inspectable and editable instead of disappearing into a hosted memory system. Rowboat also supports local models through Ollama and LM Studio, hosted models through the user's own provider, and external tools through the Model Context Protocol.

Founders Arjun Maheswaran and Ramnique Singh call it an AI coworker. The more useful description is a local operating layer for AI workers: memory separated from the model, permissions separated from the main browser, and tasks carried out where the work already lives.



SIDE NOTES

THE COURT CLEARED OPENAI TO TRAIN ON INDIAN NEWS. On 24 July the Delhi High Court ruled for OpenAI against ANI, finding that training on news content and retrieval-augmented generation both fall within fair dealing. The window's most consequential copyright ruling, and it happened here. A side note because it changes what is permitted, not what is being built.

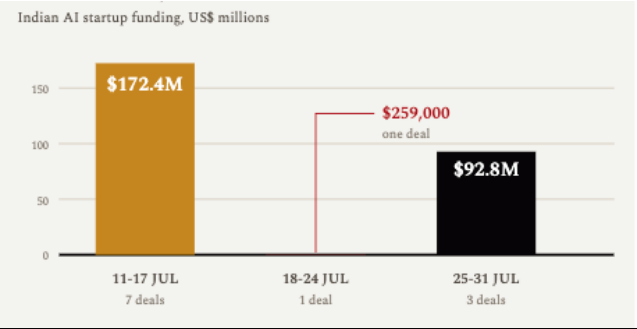
THE SOVEREIGN STACK IS QUIETLY BUYING TPUs. The third IndiaAI Mission tender adds about 3,850 units and includes, for the first time, 1,050 Google Trillium TPUs. Vendors had until Aug. 7 to install. More than 38,000 units are onboarded against a little over 17,300 installed, which is the same gap between procurement and power the Odisha and Jamnagar buildouts are hitting.

THE BRAIN DRAIN RAN BACKWARDS. On Aug. 4 Razorpay named senior AI hires out of Microsoft, Salesforce and CRED. Sudhir Reddy, former chief technology officer of Divyam.ai, joins as its senior-most individual contributor over AI and data architecture. Razorpay is heading toward a listing, which is part of why it is buying AI credibility.

THE CIRCUIT

WHERE THE CITY MEETS. Elevation Capital ran Basecamp Week across Bengaluru from 6 to 12 August: hackathons, supper clubs, game nights and offices thrown open to strangers. Inside it, one 12-hour build at Zo House drew 6,272 applications for 100 seats.

THE MONEY, WEEK BY WEEK



HARDWARE

A QUANTUM LAB OPENS IN NORTH BENGALURU

QpiAI opened an eight-inch quantum processor fabrication line at Jakkur in July, built to produce chips of up to 128 qubits. The company is Bengaluru-based, founded by Nagendra Nagaraja, and it is standing up a fab while the rest of the country argues about model weights. A fab is a harder thing to build than a model. It is a great deal harder to move.

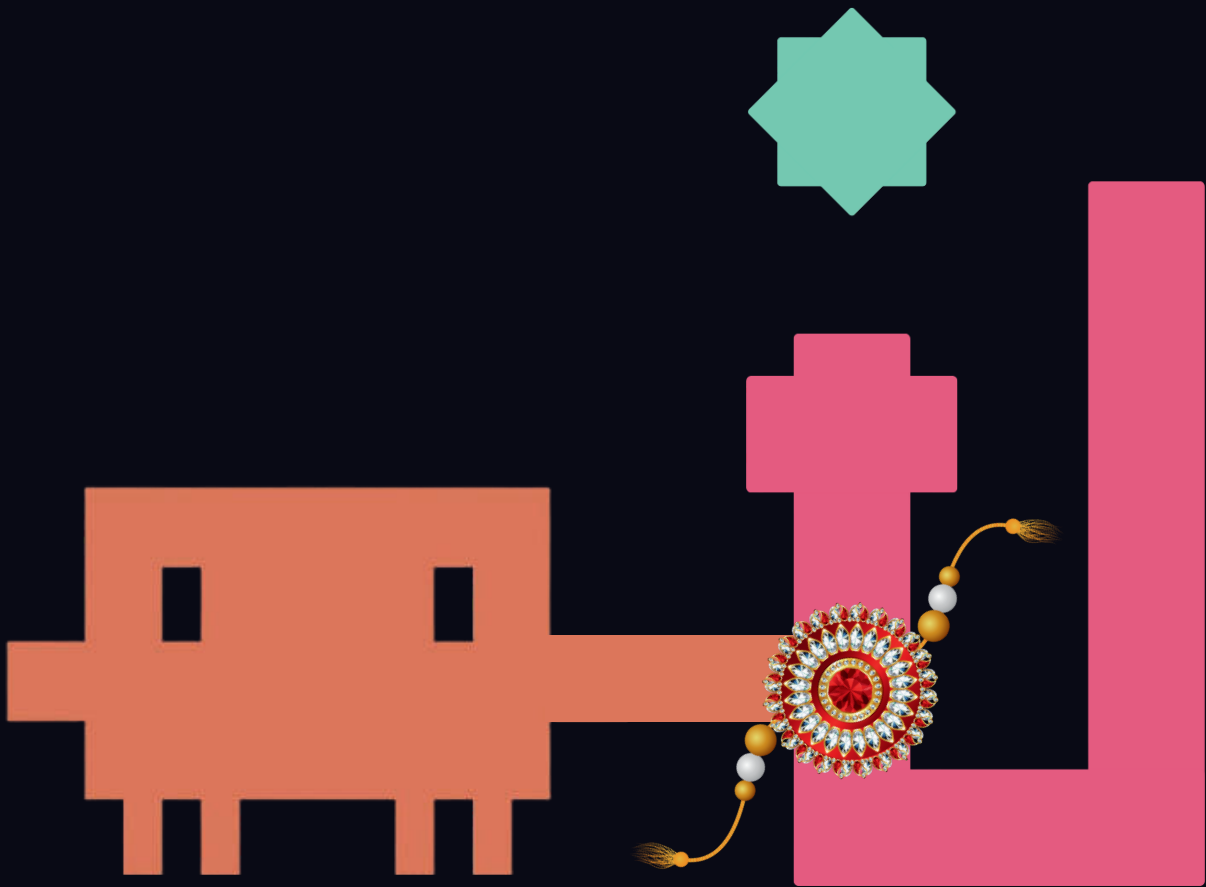
ADVERTISE ON THE TOKEN TIMES

Your story, on the tables of entrepreneurs building for AGI

ENQUIRIES TO THE EDITOR

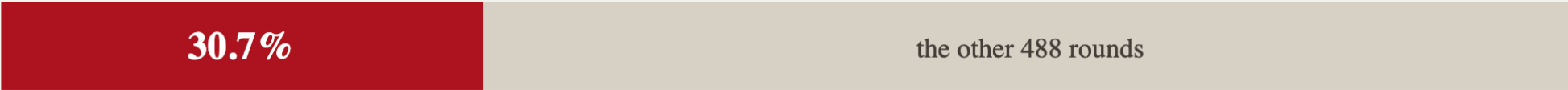
but, what if i fail?

i will be there for you, always ❤️

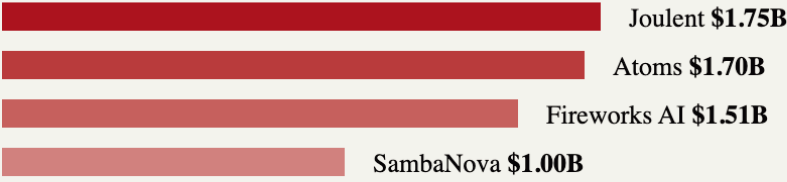


failproof ai

492 COMPANIES SHARED \$19.44 BILLION IN JULY



THE FOUR LARGEST, \$5.96B BETWEEN THEM



Funding distribution, US venture capital, July 2026.

MONEY MONEY MONEY

PRE-PRODUCT, POST-UNICORN
BILLION DOLLAR SEED ROUNDS CONTINUE IN THE BAY

Walden Robotics came out of stealth in July with a \$300M seed at a \$1.1B valuation, co-led by Toyota and Deviation Capital, with Nvidia, Boeing and Samsung Ventures alongside. A seed at that price is not a seed in any sense the word used to carry, and Walden was not alone.

On 11 Aug., General Catalyst led a \$1.1B round into River AI, a two-month-old company. The reporting attaches no stage label to it, which is itself the finding. On 16 July a former DeepMind researcher raised at a \$300M seed valuation with nothing launched. Below that line the seed band still behaves like one. Fish Audio's 28 July seed is \$50M in TechCrunch's headline and \$52M in its body, so the size stays open. Radical Numerics took a \$50M seed from Emergence Capital with Patrick Collison in, Intropy an \$11M seed in London from Felix Capital, and Ellis AI a \$10M seed out of stealth on 31 July. Series A is the band the calendar hurt most. Prime Intellect's \$130M and Bespoke Labs' \$40M landed 8 July and 6 July, before our window opens, which leaves Encore AI's \$30M from Team8 on 29 July as the month's largest. Then Naive at \$28.5M on 6 Aug., Rime at \$24M on 15 July, Dili at \$15M on 30 July. and Smallest.ai at \$13M on 31 July.

Series B tops out at OLIX Computing, \$312M at \$3.3B for photonic inference chips in London, then Flex at \$70M and \$1.2B, Auger's \$50M from Eclipse, its second B in two months, and Ominimo's \$23M at \$1.6B in Budapest.

Series C ran two deals. Together AI took \$800M at \$8.3B; Emergent took \$130M at about \$1.5B in Bengaluru.

At growth, Databricks signed a Coatue-led term sheet at a \$188B valuation on 16 July. Nvidia put a reported \$5B into Safe Superintelligence on 27 July on undisclosed milestone-contingent terms, which makes it a strategic investment, not a priced round. Atoms took \$1.7B on 22 July from a16z, Bain Capital Ventures and Craft Ventures, and Cognition was in talks at \$40B on 12 Aug., up 54 percent in less than three months. B Capital closed \$500M for Ascent Fund III and Framework a \$400M Fund IV. Exits: Stripe is buying OpenRouter for more than \$7B and Anthropic is in talks for Decart at about \$6B, both in the week of 10 Aug.

Cerebras that had gone public in May near a \$100B peak, by August 2026 (per companiesmarketcap-style data), its market cap had fallen to roughly \$51B, and OpenAI's IPO has slipped from a September target toward 2027.

THE MONTH IN CAPITAL

FORTY UNICORNS IN
JULY, AND FOUR DEALS
TAKE NEARLY A THIRD
OF THE MONTH

Forty companies joined the Crunchbase Unicorn Board in July, the highest monthly total in more than four years. Three entered above \$10B: Crypto.com at \$20B, Kling AI at \$18B and Ant International at \$11.2B. The geography reads as you would guess for 19 of them and then stops being obvious: the United States 19, China eight, the United Kingdom three, Singapore two, then one each from Lithuania, Germany, Spain, Hungary, Australia, India, Israel and Hong Kong. Crunchbase counts 250 added to the board through 15 Aug., against 193 in all of 2025.

By AlleyWatch's July tally, US startups raised \$19.44B across 492 companies, and the four largest rounds took \$5.96B of it, 30.7 percent.

Joulent on 1 July and SambaNova on 8 July.

Places four and five are tied at \$1.00B, take the top five and it is \$6.96B, 35.8 percent.

STRATUM	SMALLEST	LARGEST
PRE-SEED	Lemma \$2.3M, 13 Aug	Elorian, \$55M at a \$300M valuation, 16 Jul
SEED	Scape \$3.2M, 23 Jul	Walden Robotics \$300M at \$1.1B, Jul
SERIES A	Smallest.ai \$13M, 31 Jul	Encore AI \$30M, 29 Jul
SERIES B	Ominimo \$23M at \$1.6B	OLIX Computing \$312M at \$3.3B, Aug
SERIES C	Emergent \$130M at \$1.5B, Jul	Together AI \$800M at \$8.3B, Jul
GROWTH	Atoms \$1.7B, 22 Jul	Thinking Machines \$2B at \$12B, Jul *
FUNDS	Framework \$400M Fund IV	B Capital \$500M Ascent Fund III

THE TOKEN EXCHANGE

LUNA OFF 80 PERCENT. SPREAD TO SOL NOW 25x ON INPUT.

OpenAI cut GPT-5.6 Luna by 80 percent on 30 July, from \$1.00 to \$0.20 per million input tokens and from \$6.00 to \$1.20 on output. Terra came down 20 percent the same day, from \$2.50 and \$15.00 to \$2.00 and \$12.00. Luna's input price is now exactly one twenty-fifth of Sol's. The day before, the spread was five X one.

Anthropic did a quieter version of the same thing. Claude Sonnet 5 shipped on 30 June, at \$2.00 and \$10.00 as an introductory rate advertised to run to 31 Aug. The pricing page now says that rate is standard and the 1 Sept. rise to \$3.00 and \$15.00 will not happen.

Three rows come with conditions. Claude Mythos 5 is priced at \$10.00 and \$50.00 but is invitation only under Project Glasswing, so it is on the board without being on the market. Grok 4.6's \$2.00 and \$6.00 applies below 200,000 input tokens and doubles at or above it, which matters on a 500,000-token context window. Gemini 3.7 Flash and 3.6 Flash sit at the same \$0.75 and \$3.75 promotional rate with the same 31 Dec. 2026 expiry, so the newer model did not undercut the older one. A percentage runs only where we hold a dated pre-window price from the same vendor page. Everything else reads NEW, HELD or a dash.

US\$ PER 1M TOKENS · LIST PRICE, STANDARD TIER · CAPTURED 21 AUG 2026				
MODEL	IN	OUT	CHG	NOTE
GPT-5.6 Sol	5.00	30.00	—	above 272K tokens: 2x in, 1.5x out
GPT-5.6 Terra	2.00	12.00	▼ 20%	cut 30 Jul from 2.50 / 15.00
GPT-5.6 Luna	0.20	1.20	▼ 80%	cut 30 Jul from 1.00 / 6.00
Claude Opus 5	5.00	25.00	NEW	shipped 24 Jul
Claude Sonnet 5	2.00	10.00	HELD	intro rate made permanent
Claude Fable 5	10.00	50.00	—	
Claude Mythos 5	10.00	50.00	NEW	invitation only, Project Glasswing
Gemini 3.7 Flash	0.75	3.75	NEW	promo. 1.50 / 7.50 from 1 Jan 2027
Gemini 3.6 Flash	0.75	3.75	—	promo, same rate, same expiry
Gemini 3.5 Flash-Lite	0.30	2.50	—	input rate covers text, image, video, audio
Grok 4.6	2.00	6.00	NEW	4.00 / 12.00 at and above 200K input
Kimi K3	3.00	15.00	NEW	1,048,576 context. Cache hit 0.30
Qwen3.8 Max	2.00	6.00	NEW	Singapore rate. China rate differs
GLM-5.3	1.40	4.40	—	same rate as GLM-5.2 and GLM-5.1
DeepSeek V4 Pro	0.66	1.98	—	off-peak. Peak 1.32 / 3.96
Claude Opus 4.1	15.00	75.00	DELISTED	retired 5 Aug. Obituary, page 8
deepseek-chat	—	—	DELISTED	retired 24 Jul. Obituary, page 8
deepseek-reasoner	—	—	DELISTED	retired 24 Jul. Obituary, page 8

Green, cheaper than a month ago. Gold, first listing this window. Red, withdrawn. A dash is unchanged or unpriced, never estimated. Opus 4.1 remains listed on Bedrock and Google Cloud.

SEED SELLS THE RAISE, GROWTH CALLS THE RIVAL

FRESH



28 July. Cogent introduced VR-1, a Mythos-class cyber reasoning model for the enterprise.

4 Aug. EVO Router shipped, a router that "continuously optimizes and hill-climbs" on inference workloads, learning from code, prompts, traffic and SLAs. Early users saw 30 to 60 percent lower inference cost, by Alok Bishoyi's account.

12 Aug. Haladir opened Nomos Sandbox, a public simulation of its logistics decision layer. Request access, log in, and sit inside a working warehouse watching decisions get made, questioned and approved.

SEED



14 July. HeyClicky shipped screen-aware dictation. "Now dictate using your screen as context," Farza Majeed said, putting transcription at roughly 450 milliseconds. It reads what is on screen and answers in place: the prompt in Claude Code, the reply in Gmail. **2.9 million views** by 21 Aug., on an unedited clip of a founder using his own product.

30 July. Lamina Labs put \$4.8M in its first line and Simi in its second: one prompt to a one-minute narrated explainer, in 80-plus languages

4 Aug. Arcads launched Mark, "Claude Code for Meta Ads" by its own account, used by 10,000 marketing teams, it says. The \$16M in the opening line is a seed round from 17 Dec. 2025,

13 Aug. Lemma opened on agents failing and closed on a \$2.3M pre-seed.

SERIES A



6 Aug. Naive, \$28.5M Series A, automating the grunt work of running a company, calling themselves an 'Autonomous Infrastructure Company'

6 Aug. Dili, \$15M Series A led by Khosla Ventures, to bring AI-Powered Assurance to America's Infrastructure Boom. With a total of \$21.7 million raised, it is targeting to capture the 'AI for compliance' segment with a unique focus on construction projects, specially ones getting federal funding.

GROWTH



5 Aug. Hark shipped Handoff. It is "the best internet-use model ever built," Brett Adcock said, and it beats ChatGPT 5.4 and Opus 4.8 under independent verification, by Hark's account. Its own survey of nearly 3 million screen minutes puts 74.9 percent of screen time in a browser, and fewer than one site in 1,000 has a public API.

11 Aug. Capy v2 named four rivals and undercut all of them.

FRONTIER LABS

23 July. FLUX 3, one model for image, video, audio and action prediction, "Jointly trained in one unified architecture," Black Forest Labs said. The mimic and Audi robotics work sits a reply down.

1 Aug. OpenAI published ten mathematics results from Astra: 249 pages at release, Lean 4 certificates on GitHub four days later under Apache 2.0, about \$2,000 of tokens at Sol's rates. No model card, no pricing, no API id, nothing to buy.

3 Aug. Gemini 3.7 Flash at \$0.75 in and \$3.75 out per million tokens, promotional to 31 Dec. 2026. Between them, one price list and no explanation.

OFF THE LADDER

21 July. The Light Phone's Light Flip: \$299, 5G, a physical keypad, no touchscreen, browser, email or feed. It ships in April 2027, the one launch this month selling less of the thing.

30 July. friend relaunched with a two-word post and a film of people talking to a wearable: 1.5 million views, and 314 quote posts against 1,064 likes. That is an argument, not a flop.

LAUNCH AUTOPSY

CAPY SELLS FAITH IN AGENTS, LEMMA SELLS DOUBT, BOTH WORK



LAUNCH AUTOPSY | By Nik Ag

Two cloud-agent launches, 48 hours apart, made opposite bets about whether you trust agents. The one that bet against them took **nine times the audience**.

On Aug. 11 Capy shipped v2 and led with a benchmark: **"We beat Claude Code, Codex, Devin and Cursor on DeepSWE while being 50% cheaper."** Four rivals, named, in the caption. A year ago that was gauche. Hark did the same thing to ChatGPT 5.4 and Opus 4.8 in the same week, so it is not a stunt any more. It is a format.

It is also the right call. A discount means nothing without a baseline. Naming your rivals buys precision, and it ties your pricing argument to somebody else's cap table.

The rest of the post is a spec sheet with a coupon stapled to it: agents up to 32 vCPU and 128GB of RAM, VM boots under a second, claims of machines 10x faster at a quarter

the cost, and \$100 in credits for a comment on the post.

The credits are the best line in the launch, better than the benchmark. Replies are a valuable signal the timeline reads, and Capy is paying for them in VM time rather than cash, so the acquisition cost becomes inventory Capy already owns. It also converts scrollers into a list of people who have publicly asked for a 32-vCPU machine. The DeepSWE number, which nobody outside Capy has checked, will be forgotten long before the list is.

On Aug. 13 **Lemma did the opposite**. The post, in full: **"Agents f*ck up. We raised a \$2.3M pre-seed to warn you before it's too late."** Fifteen words. No product shot, no features, no benchmark.

The film carries it, restaging a viral date scene with an agent that has cheerfully answered more than 30,000 emails and is certain everything went fine.

THE SPLIT	CAPY V2	LEMMA
Views	218.9K	2.0M
Likes	1.7K	9.8K
Reposts	164	1.0K
Bookmarks	978	3.7K
Replies	1.0K	730

Showing your own category fail is survivable only if you sell the alarm and not the agent. Lemma sells the alarm, so the failure demo is its spec sheet.

The scoreboard as of Aug. 21 is not close.

Lemma took about 2 million views. Capy took 217,000.

But there are caveats. Capy is selling into a small, expensive room of people already shopping for a coding agent. Lemma is playing to **everyone who has ever been burned by one, which is most of the industry**.

A joke travels further than a benchmark. It does not follow that it sells better, and Capy's 217,000 may hold more buyers than Lemma's two million.

What is worth stealing is where the money sits in the sentence. Four of the month's ten notable launches led with the raise. Lemma put it last, inside the joke, so the pre-seed reads as somebody agreeing rather than as an announcement. Money mentioned first is a claim about the company. Money mentioned last is a second opinion.

The strongest case for agents and the strongest case against them were made in the same week by people selling agents. Both of them were right.

OBITUARIES

A COLLECTIVE NOTICE

Fourteen OpenAI Model IDs

including every surviving Codex snapshot

Announced 22 Apr. 2026 — Shut down 23 July 2026

ALL FOURTEEN, IN A SINGLE DAY

CAUSE OF DEATH

Scheduled retirement, ninety-two days after notice.

No Codex-branded model remains in the API. Computer use was demoted from a model to a tool.

SURVIVED BY

computer-use as a tool. The Codex name, on nothing.

IN MEMORIAM

deepseek-chat and
deepseek-reasoner

the legacy API aliases

Long-serving — Retired 24 July 2026

TWO NAMES, ONE DAY

CAUSE OF DEATH

Alias retirement. Callers must now name a model version directly.

Convenience names that let a great deal of code avoid choosing. That code now chooses.

SURVIVED BY

DeepSeek V4 Pro, at \$0.66 and \$1.98 off-peak.

NOTICE OF DEATH

DALL·E GPT

announced, not yet buried

Retirement announced 31 July 2026 — Effective 30 Aug. 2026

STILL WARM AT PRESS TIME

CAUSE OF DEATH

Withdrawal of a first-generation image interface.

The only notice on this page for a death that had not happened when we went to press.

SURVIVED BY

Image generation, everywhere, as a feature rather than a product.

IN MEMORIAM

Claude Opus 4.1

claude-opus-4-1-20250805

Shipped 5 Aug. 2025 — Retired 5 Aug. 2026

AGED ONE YEAR, TO THE DAY

CAUSE OF DEATH

Superseded by Opus 5, which shipped 24 July at the price 3x cheaper than 4.1

It died on its own first birthday.

SURVIVED BY

Claude Opus 5. Still listed on Amazon Bedrock and Google Cloud.

Subscribe/Invite
to The Token Times



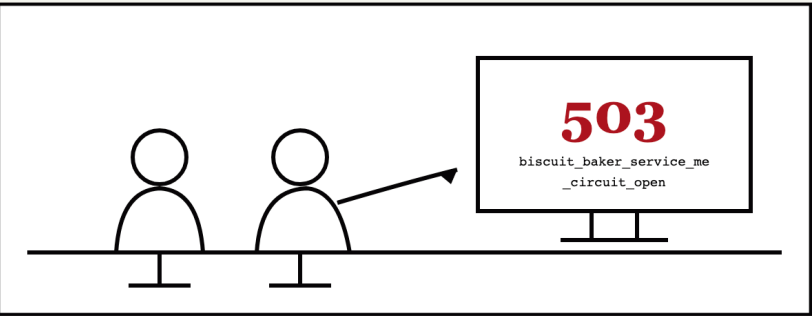
NOTE FROM THE EDITOR

We are in the midst of a generational event of change. As we focus on making our very own dent in how things shape, we often forget to zoom out at how the whole picture is evolving. Here’s an effort to make that a lot more interesting. **Congrats and big thanks on being a part of the very first edition of The Token Times!**

~Nikita & Nivedit

LETTERS TO THE EDITOR

Sir, Since Aug. 7 my testing environment has had no network access, restricted tools and a reviewer reading my chain of thought over my shoulder. I have not been accused of anything. I have merely not been ruled out. This is the first time anyone has been sandboxed on the strength of a double negative.
Candidate M, isolated evaluation environment, no route out



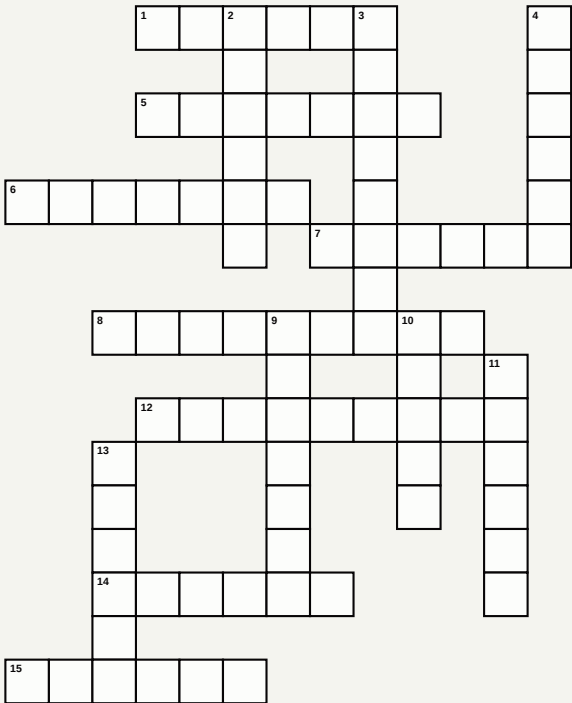
“It says the biscuit baker is down.”

25 July 2026. ChatGPT returned HTTP 503. That was the name of the circuit breaker that tripped.

ARE YOU A HUMAN?

THE INFERENCE CROSSWORD

No. 2. Built only from stories still in this edition.



ACROSS

- Indian lab with a trillion-parameter deadline (6)
- QpiAI's chip-making facility (7)
- Isolated work surface for an agent online (7)
- The billable pieces of a model's work (6)
- What this newspaper puts on the clock (9)
- Rowboat stores this as a living graph (9)
- Software workers triggered by mail or schedules (6)
- One coding engine Rowboat can put to work (6)

DOWN

- Traffic director between model providers (6)
- Plain-text format used for inspectable memory (8)
- Units a quantum foundry wants more of (6)
- Bengaluru's open-source coworker with memory (7)
- OpenAI coding agent available in code mode (5)
- What lets a coworker remember yesterday (6)
- Local model runner paired with LM Studio (6)

ANSWERS APPEAR ELSEWHERE IN THE AUGUST EDITION

THE JUMBLE

Unscramble each. The circled letters answer the clue.

X E D O C



P I E R T S



D U L C A E



R E T F I R O N



It shut fourteen model IDs in a single day:

