AI prices are falling but firms are spending more

Growing use of open-weight models is changing the economics; instead of relying entirely on a frontier model, firms can use different models for different tasks.; smaller or open-weight models can handle some workloads at a lower cost while more expensive frontier models can be used for tasks that require greater capability; cheaper tokens are encouraging widespread usage of AI

EXPLAINER

Balasubramanyam C.P.
CHENNAI

Artificial intelligence (AI) is turning sharply cheaper to run on a per-token basis but the amount firms are spending on the same is rising as they move from experiments and chatbots to more complex, multi-step applications.

At a briefing, Coforge management said over the previous seven months, the unit cost of tokens to deliver the same level of AI had fallen about 100-fold while token consumption surged about 8,000-fold.

Coforge also said open-weight models accounted for about 35% of token use on the platforms it was observing compared with about 11-12% in early 2025.

There are broadly two types of models. Closed or proprietary models such as OpenAI's GPT, Anthropic's Claude and Google's Gemini, keep trained parameters under the developer's control and are typically accessed through a hosted service or API.

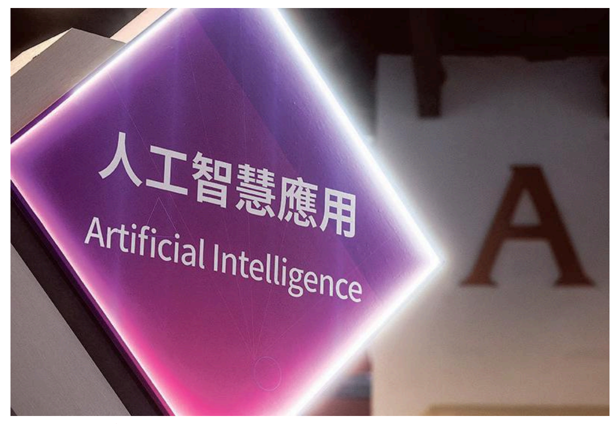
Open-weight models such as Meta's Llama make trained parameters available for organisations to download and run themselves subject to the model's licence. Open-weight need not necessarily mean fully open-source. Their training data, training process and other components may be undisclosed.

Coforge also said the right way to assess AI economics was "cost per unit task" viz. the cost incurred to finish a defined piece of work and value it creates.

Turning cheaper?

Stanford's 2025 AI Index, an annual report tracking major developments and trends in AI, found the cost of querying a model with GPT-3.5-level performance on a standard AI knowledge and reasoning benchmark fell from $20 per million tokens in November 2022 to $0.07 per million tokens by October 2024, a reduction of more than 280 times in about 18 months. Drawing on an analysis by Epoch AI, it found LLM inference prices had fallen between nine and 900 times per year depending on the task.

The OECD (Organization for Economic Co-operation and Development) found a similar trend and said its July analysis estimated quality-adjusted prices for text-to-text AI models fell by almost 80% between January 2024 and April 2026.



Cost-benefit: Cheaper models are making more applications viable while agents are making individual applications more intensive. AFP

The nature of the work being given to AI is also changing . A chatbot might receive a question, process it and return an answer. An AI agent can break a task into several steps, reason, call external tools/systems, inspect the results, revise approach and continue till the task is done. Each additional step can involve more model calls/tokens.

The OECD notes AI agents consume substantially more tokens per task and can raise model-use intensity by order of magnitude. As a result, falling prices/token do not translate into lower costs for complex AI applications.

Gartner described this as the "Inference Paradox". Its August 2026 analysis says routing a task to an agentic reasoning model can increase provider inference costs by at least five times compared with a basic chatbot interaction, and often by much more as complexity rises. Gartner sees inference cost per agentic workflow to rise over fivefold through 2028.

Changing the equation

A 2026 working paper by researchers from the University of Michigan, Stanford University, All Hands AI, Google DeepMind, Microsoft AI and MIT found agentic coding tasks consumed about 1,000 times more tokens than the code-reasoning and code-chat tasks used for comparison. It also found runs of the same task could differ by as much as 30 times in total token consumption.

This is showing up in enterprise spending. McKinsey's May 2026 Enterprise AI FinOps survey, covering 75 respondents across five industries, found 93% reported exceeding AI budgets. McKinsey also cited Menlo Ventures data showing enterprise spending on large language models roughly tripled over the 12 months to the end of 2025.

The growing use of open-weight models is changing the economics. Instead of relying entirely on a frontier model, firms can use different models for different tasks. Smaller or open-weight models can handle some workloads at lower cost while more expensive frontier models can be used for tasks that require greater capability.

Coforge said its clients were increasingly moving towards a mix of models. It also identified seven billion tokens of consumption as a "tipping point" in its experience at which clients start considering whether to consume and operate more of the AI stack themselves rather than simply subscribe to a frontier model. It said that threshold could fall as open-weight and sovereign AI develop further. OECD research, also shows competition and alternative models contributed to falling AI prices.

Measuring AI costs

For firms, a useful measure is economics of the complete task like model calls, tokens, computing resources and human oversight needed to produce a result set against the value of that result.

Gartner's research also points towards measuring token use against business value and economics of the overall workflow rather than treating token prices as proxy for AI. The result is a rapidly changing cost structure, like cheaper models are making more applications viable while agents are making individual applications more computationally intensive.

NTPC drops ₹413 cr. BESS deal as contractor fails on obligations

The Hindu Bureau
NEW DELHI

State-owned power producer NTPC Ltd. said it terminated the contract awarded to GR Infraprojects Ltd. (GRIL) for building the 400 Megawatt-hour (MWh) battery energy storage system (BESS) project at its Mouda Super Thermal Power Station (Maharashtra) with immediate effect citing the latter's inability to meet contractual obligations and meet required project progress.

The project – with a contract value of ₹413.37 crore - awarded in March– was scheduled to be completed in 15 months. Explaining the cause for terminating the contract, NTPC stated, "The project witnessed significant delay in critical activities. A contractual notice was issued to GRIL calling for immediate corrective action, but GRIL failed to undertake the required remedial measures." NTPC added it encashed about ₹91 crore from bank guarantees.

MSEDCL uses AI, ML to detect power theft

The Hindu Bureau
MUMBAI

Maharashtra State Electricity Distribution Company Ltd. (MSEDCL) said it cracked down on power theft across the State by leveraging Artificial Intelligence (AI) 'Super Intelligence' module.

Combining state-wide physical inspections by 190 "flying squads" with AI-driven predictive analysis, MSEDCL said it exposed 27,964 instances of power theft worth ₹66.28 crore in 3 months. Criminal cases were registered in 289 instances. MSEDCL said it adopted AI and Machine Learning (ML) to predict power theft accurately.

FinMin, bank CEOs meet

Press Trust of India
NEW DELHI

The Finance Ministry has called for a meeting of CEOs of public sector banks and regional rural banks on Monday on steps to ensure continuity of banking operations and essential customer services during the 3-day strike of the United Forum of Bank Unions from September 28.

Hormuz: EIL, Saudi, UAE mull oil infra

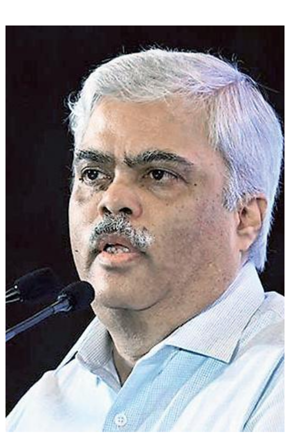
Press Trust of India
NEW DELHI

State-run Engineers India Ltd (EIL) is in early-stage talks with Saudi Arabia and the UAE to provide consultancy and engineering services for oil and gas infrastructure aimed at reducing the Gulf producers' dependence on the Strait of Hormuz, severely disrupted by the Iran conflict.

Saudi Arabia and the UAE are planning about $1 billion spend on pipelines, storage units, oil terminals and infrastructure to provide alternative routes for moving crude and petroleum products to international markets, CMD Atul Gupta told reporters.

CWC chief flags 'malpractices' in phytosanitary tag

Balasubramanyam C.P.
CHENNAI


Santosh Sinha

Central Warehousing Corporation (CWC) MD Santosh Sinha raised concerns over what he described as "malpractices" in phytosanitary certification at ports while declining to name the entities involved.

Speaking to reporters on the sidelines of an event, Mr. Sinha was asked about his reference to malpractices in certification.

He said there had been "a certain result which was not to our liking" but did not elaborate on the incident or identify any port, agency or official.

Mr. Sinha raised the issue while explaining CWC's plans to expand its role beyond conventional warehousing and into integrated logistics for agricultural exports. He cited the rejection of Indian rice at Chinese ports and mango exports by Japan as examples of the challenges Indian exporters could face when consignments fail to meet norms in destination markets.

Asked how CWC could address the issue, Mr. Sinha said: "We have the technical expertise with us."

He said CWC was also upgrading its technical expertise, pointing to its experience in preserving food commodities. CWC currently stores about 50-60 lakh tonne foodgrains and, according to Mr. Sinha, less than 0.5% is lost during storage.

Phytosanitary norms

Mr. Sinha said the issue of export rejection was closely linked to phytosanitary requirements.

He said India would encounter more such challenges as exports rise under free trade agreements.

"This is actually the sign of a growing economy. Because you are now going in for free trade agreements, you will start exporting, you will encounter challenges, you will have to meet them," he said.

In his address, Mr. Sinha said CWC could add a layer of phytosanitary treatment and preservation to the export supply chain to help ensure agricultural produce was not rejected at destination ports.

CWC is seeking to expand its role in agricultural logistics by bringing together storage, preservation, grading, sorting, treatment, cold storage and transportation with connectivity to ports.

CWC was targeting capital expenditure of about ₹1,000 crore this year, similar to the previous year, Mr. Sinha said.

Parallelly, CWC was planning its first hyperloop freight trial at its CFS Logistics Park at JNPT, Mumbai, with a ₹23-crore order placed with IIT Madras-incubated TuTr Hyperloop.