MELBOURNE MONDAY, 16 NOVEMBER 2026 MORNING BRIEFING EDITION

VOL. I · NO. 320
FREE PRESS

The Daily Signal

MORNING BRIEFING
EDITION
MONDAY 16 NOV 2026

INTELLIGENCE ON THE AI FRONTIER

Agents get smarter by forgetting, Nvidia reports its first quarter without its founder, and the company about to list is now competing with its own customers.

RESEARCH · AGENT ECONOMICS

Agents Do Better When They Remember Less

A new study cut agent context from 256K tokens to 16K. Accuracy went up, and serving throughput rose as much as sevenfold.

7×	88.8%	73%	$855B	$1.404T
THROUGHPUT GAIN AT 16K CONTEXT	SWE-BENCH VERIFIED AFTER COMPACTION	AGENT COST SPENT RE-READING CONTEXT	NVIDIA VALUE LOST IN A WEEK	TOP OF THE IPO VALUATION RANGE

Nvidia researchers tested 13 benchmarks, including SWE-bench Verified, Terminal-Bench and LongBench v2. They compared agents that keep their whole history with agents that restart at a threshold such as 16K tokens, carrying only the original instructions and a state summary. Accuracy improved. One frontier model went from 70% to 74.8% on SWE-bench Verified, and another from 84.4% to 88.8%. On H100s, the shorter window raised output throughput between 3.6× and 7× depending on the open model.

Rewriting the state after every single tool call nearly matched the full-history baseline while peak context stayed at 7,000–9,000 tokens. On long documents, storing the material as files and letting the agent grep them matched or beat full context on almost every benchmark. Most of the gain came from deciding what to read.

There are two caveats. A verbose coder collapsed from 53% to 3.6% under the 16K limit. And compaction breaks prompt caching, which raised one model's benchmark bill from $167.68 to $203.12.

CORROBORATION

Changing the Harness Changes the Bill 2.6×

A separate study of coding agents found that re-reading existing context accounts for up to 73% of cost. Agents use only 27% of what their tools return, and up to 89% of turns can be file search. Changing the harness alone moved cost by up to 2.6× on the same model and task.

TECHNIQUE

Write a Briefing, Not a Summary

The compaction that worked best reads like a handover note: the objective, established facts, relevant files, changes made, approaches that failed, work remaining and the next action. The researchers propose "semantic state space models" that pass a human-readable task state from step to step.

"Most of the gain came from deciding what to read. For agents, a big context window is a cost you pay on every turn."

DECISION MODELS

The Router Becomes a Separate Product

Decision models return a calibrated yes-or-no in one forward pass, with no generated text. Autoregressive models typically spend 30–100 tokens and 1.2–3.5 seconds on the same choice. Google’s Decisions API answers up to 1,000 questions per pass in 150–250ms for $0.10 per million input tokens. An open-weight 4B decider runs air-gapped on a 64 GB laptop in about 835ms, and coding harnesses now let you choose the decision model separately from the coding model.

One problem: out of the box these models are not calibrated. In one test, only 66% of pairs scored above 0.9 were actually relevant. A simple Platt mapping brought the Brier score to about 0.15 and lifted expected-F1 from 0.506 to 0.638.

ARCHITECTURE

Position Encoding Starts to Fragment

RoPE is still standard. It rotates query and key pairs so attention depends only on the distance between tokens, with frequencies $\theta_i = 10{,}000^{(-2i/d)}$. The split-half pairing has displaced the original interleaved one.

Partial RoPE, which rotates about 10% of features, went from a minority of models in 2025 to a majority in 2026. Some layers now use no positional encoding at all and rely on the causal mask: a 107M test model with no position encoding ranked first on length generalisation across 10 tasks. One new model returned to a T5-style learned relative bias.

PAPERS OF THE WEEK

Self-Organising Swarms and a Self-Improving Training Run

Identical copies of one small model, with no leader or assigned roles, learned to coordinate on a shared reward. With twice as many agents as in training, they solved more than 25× as many coordination problems as the best baseline, with nearly half the duplicated work. A recursive self-improvement loop ran a full pre-training: a 30B mixture-of-experts on 11.8T tokens over three months, released openly, which beats its peers in its size class.

Two more results stand out. A tactile foundation model trained on 4,983 hours of touch data cut contact error from 5.89mm to 0.16mm. Agents writing Lean-verified quantum programs reported speedups of up to 290× on quantum arithmetic.

PRACTICE

The Harness Matters More Than the Model

Several unrelated results this week say the same thing. A harness wrapped around a frozen 8B model, with no retraining, gained 64 points on average across long-context benchmarks. It beat purpose-trained methods at less than half the cost, and the harnesses were written by a frontier model. A memory system that links each stored insight back to its diagnosis and raw trajectory added 10.7 points over no memory. It also cut failures on tasks the agent could already solve from 6.5% to 4.5%.

The same result shows up in practice. On 20 companies, an analyst compared the newest and most expensive frontier model with its cheaper sibling. The cheaper model won 12, the expensive one 7, with 1 tie. Once detailed research skills constrained both, valuation work showed no meaningful difference. One team spent more than $10,000 learning the new model and found that “High” effort beat “max” on tasks that depend on judgement.

The lesson for engineering leaders: budget for the scaffolding (skills, memory, retrieval, verification) before you upgrade the model.

RELIABILITY

Health Checks That Caused the Outage

Last week’s productivity-suite outage lasted more than 24 hours and drew over 7.4 million reports. A defect from an earlier capacity upgrade surfaced after a later network change. Servers failed health checks and were removed, which overloaded the rest until they failed too. Restarted servers passed their checks, then failed again under real traffic. The usual advice: plan capacity for losing 20–30% of the fleet, and treat configuration changes like code deployments.

SUPPLY CHAIN UNCONFIRMED

A Fake Extension Reached Developer Laptops

Reports describe a stolen provider key that led to a backdoored coding-agent extension being installed on developer machines. Separately, an agent’s “don’t ask again” button is said to have leaked 1,379 secrets to public repositories. The tools developers install without thinking have become the attack surface.

SECURITY UNCONFIRMED

An Open-Source Security Audit Skill

A major edge network has reportedly open-sourced a skill in which several coding agents divide up a codebase to look for missing authentication, prompt-injection paths and APIs that expose other users’ data. How the work is split is behind a paywall.

PATTERN UNCONFIRMED

Separate the Agent That Plans From the One That Checks

Agents fail when one model plans, executes and judges its own work. The suggested design uses a reasoning-heavy planner, cheaper long-horizon workers and separate verification agents, with each job matched to a different model in the same family.

CHIPS

Nvidia Reports Without Its Founder

Nvidia shares fell nearly 12% on Thursday and as much as 19% over the week, erasing more than $855 billion in market value. The interim chief keeps the finance job and now has about 20 executives reporting to her. Tuesday's results are the first since the founder resigned. Neocloud stocks fell with Nvidia, one by about 17%, and a $13 billion acquisition of an open-model hub has not closed.

The bigger question is what stops. Nvidia had been the AI economy's lender of last resort. It guaranteed minimum revenue on neocloud capacity, put nearly $50 billion into frontier labs, and was considering backing debt for a $45 billion data center. Investors now expect it to scale back a $140 billion two-month dealmaking spree. If the "Jensen put" ends, the market tilts back toward the hyperscalers.

CAPITAL MARKETS

The IPO Pitch Is to Do Everything

The safety-focused lab will sell 825 million shares at $70–78, raising $57.75–64.35 billion at up to $1.404 trillion, plus $26 billion of non-voting stock sold to seven tech companies. Before listing it shipped spreadsheets, whiteboards, meetings and dashboards, and embedded partners' full interfaces inside its chat. Its partners now meet it as a competitor.

POWER

China Moves Data Centers to 800V DC

A Chinese wind company runs 800V DC in a working data center. Solid-state transformers step 10,000V grid AC straight down, lifting conversion efficiency to 98.5% from 94–95%, according to the company. It costs 10–15% more up front. Chinese "super racks" already draw 400–500kW, against up to 227kW for Nvidia's current flagship rack.

I · MEMORY

Compact Agent Context Without Breaking the Cache

Compaction makes agents more accurate, but it destroys the prompt cache, which is why one benchmark bill rose 21%. The fix is to structure memory around the cache: a frozen prefix (instructions and tools), a typed briefing that is appended rather than rewritten (facts, files, failures, next step), and a short tail of recent turns. Each compaction then becomes a diff the cache can survive. The patentable part is a cache-preserving compaction protocol that decides when to restart based on the expected cost of a cache miss against the drop in accuracy.

II · CONTROL PLANE

Decision Models Need Calibration Guarantees

Decision models already handle a quarter of traffic on some gateways and are about to gate refunds, security rules and agent tool calls. Yet a 0.9 score was right only 66% of the time in one test. That opens a market for calibration as a service: continuous Brier-score monitoring for each gate, re-fitting when the vendor changes the model, and a contractual service level on error rates. Whoever certifies the confidence numbers effectively underwrites autonomous systems.

III · FEEDBACK

When Forecasts Change What They Predict

Three stories this week share a pattern: air-traffic rerouting that degrades once every crew takes the same shortcut, health checks that overload the servers they protect, and agents starting to buy rides and shares. In each case, acting on a prediction changes the system being predicted. Agent platforms should model their own influence by default: estimate how many other agents will act on the same signal, and dampen or randomise the action so they don’t all move at once. The more agents there are, the more this matters.

IV · MOVE 37

Buy Open Models as Nvidia Pulls Back

The consensus says that if Nvidia stops funding open models, they wither. The contrarian view is the opposite: the patron’s exit makes open weights cheap and leaves them without an owner, just as new tools let a frontier model rewrite a small model’s weights directly for $5–10 of usage. That makes possible a made-to-order model shop: take an orphaned open model, edit its weights for one customer’s workflow, and deliver a model that is 3× faster and owned outright. The defensible piece is a signed record of every weight edit, its test results and its licence chain, which a regulator or acquirer can audit.