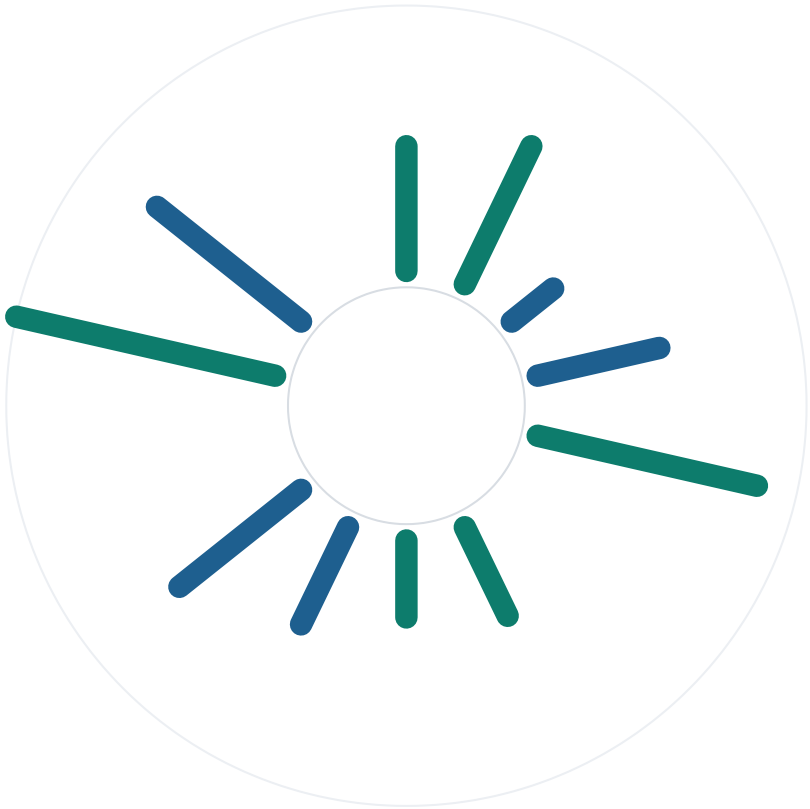
NEW HORIZON · AI DIGEST

The Long Read

15 August 2026

11	3	3086	6
STORIES	SECTIONS	WORDS	HIGH IMPACT

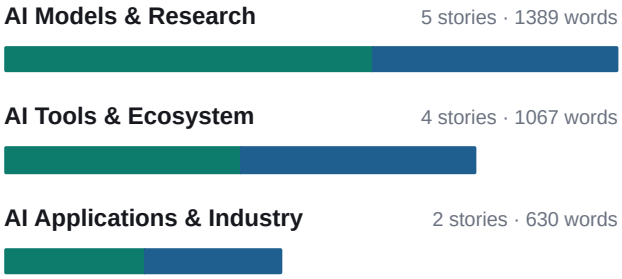


6 high impact 5 medium impact spoke length = depth of coverage

The full-length companion to the daily New Horizon AI Digest.
Every story in today's email, reported at length · new-horizon.tech/digest

The issue at a glance

11 stories · 3086 words · 3 sections · 1 charted



AI Models & Research

01 GLM-5.3: Chinese Labs Match the Frontier Through Post-Training Alone

02 State of Open Models: China's Open Weights Dwarf America's, Qwen Becomes Community Base

03 Vero: AI Agents Build Formally Verified Software Repositories

04 OmniScientist: An Omni-Modal Omni-Discipline AI Scientist

05 QuoteBench: Benchmark Scores Can Hide Agent Command-Path Failures

AI Tools & Ecosystem

06 OpenAI and Anthropic in Price War as Chinese AI Rivals Gain Enterprise Ground

07 Google Makes Visible AI Watermarks Optional, Keeping Only Invisible SynthID

08 Don't Classify, Hallucinate: LLM Tagging via Embedding Matching

09 DFM Mimir v1: Open 1B Model Trained Only on Permissible Post-Training Data

AI Applications & Industry

10 Hyperscalers' Natural Gas Bet Could Backfire as Prices Forecast to Triple

11 Kog Goes Deeper into GPU Architecture to Squeeze More Inference Per Chip

How to read this. Every story in today's email is reported here at full length, in the same order. Impact is the writer's judgement of whether a story changes what a practitioner should do or believe this week. Charts appear only where the source itself puts comparable numbers side by side; nothing is estimated to fill a gap. Sources are listed in full on the final page.

SECTION 1 OF 3

AI Models & Research

5 stories 3 high 2 medium

01 HIGH IMPACT www.interconnects.ai

GLM-5.3: Chinese Labs Match the Frontier Through Post-Training Alone

> Z.ai's GLM-5.3 matches frontier agentic coding capabilities using only 750B parameters by scaling post-training rather than pretraining compute.

Z.ai announced GLM-5.3, a model currently available in its coding plan with API access imminent and open weights scheduled for Hugging Face in two weeks. The system achieves frontier-level performance on agentic coding benchmarks, surpassing Moonshot AI's Kimi K3 and competing with Claude Fable 5 and GPT-5.6-Sol despite utilizing approximately 750B parameters—roughly one-third the size of Kimi K3. The architecture retains the GLM-5.2 base, with Z.ai attributing the performance leap entirely to extended post-training involving more diverse environments, tasks, and compute dedicated to reinforcement learning regimes.

The release challenges the prevailing narrative that Chinese labs rely primarily on distillation to close the capability gap with American counterparts. Instead, the results stem from a combination of rapid release cycles, allowing teams to hill-climb on public benchmarks while US firms conduct months of internal testing, and a targeted focus on high-value text-only coding use cases rather than broad omnimodal support. The data ecosystem supporting this progress is also shifting, with reports indicating a growing market where American data companies sell RL environments directly to Chinese model labs, enabling them to replicate sophisticated training setups without developing the infrastructure from scratch.

Safety protocols for GLM-5.3 reflect its dual-use potential in cybersecurity, specifically regarding vulnerability discovery and exploit analysis. Z.ai is implementing a staged release, beginning with controlled evaluations for selected security partners before broader API availability and final weight publication. This approach includes request classification and chain-of-thought monitoring, though the article notes that such measures face limits once full open weights are distributed. The model represents a continuation of Z.ai's iterative development cycle, which has produced major GLM generations roughly every few months since early 2023.

KEY FACTS

PARAMETER COUNT
~750B

RELATIVE SIZE VS KIMI K3
one-third

OPEN WEIGHTS AVAILABILITY
two weeks

BASE MODEL
GLM-5.2

PRIMARY TRAINING METHOD
Scaling post-training

RELEASE DATE (GLM-5)
February 11, 2026

WHY IT MATTERS

Practitioners should anticipate accessible, highly capable coding agents within two weeks that rival proprietary US models but require significantly less inference compute. The release validates post-training scaling as a viable path to frontier performance, suggesting teams can prioritize RL data diversity over massive pretraining runs.

Read the original at www.interconnects.ai →

---

02 HIGH IMPACT huggingface.co

State of Open Models: China's Open Weights Dwarf America's, Qwen Becomes Community Base

> China's open-weight strategy has overtaken the US at the frontier, with monthly releases reaching 2.78 trillion parameters while American counterparts stalled under 130B for most of 2026.

The geography of open model development shifted decisively in the first seven months of 2026. Chinese laboratories bypassed the traditional progression from small to large models, releasing systems between 754B and 2.78T parameters monthly, whereas American labs exceeded 130B parameters in only two months via NVIDIA's Nemotron 3 Ultra (561B) and Thinking Machines' Inkling (952B). While US hardware vendors like AMD and NVIDIA led in total repository count to drive chip sales, original American model creation above 100B parameters remains sparse. Most large US releases are now optimizations or conversions of Chinese base models rather than original architectures.

Qwen has emerged as the de facto base layer for the community, generating 151,448 derivative repositories on Hugging Face, a figure 2.6 times Meta's total footprint and 4.7 times that of Llama specifically. This dominance stems from a full-spectrum release strategy covering sizes from under 1B to 2.4T, combined with permissive Apache 2.0 licensing. In contrast, Chinese frontier-only portfolios like Moonshot saw only 37M downloads compared to Qwen's 2,045M, highlighting that attention does not equal adoption. While likes cluster around new frontier releases, download volume remains anchored by small, stable models under 1B parameters, which account for 83% of all-time traffic.

Local inference capabilities have expanded to support these massive weights through llama.cpp, which now hosts GGUF builds of models up to 2.8T parameters. Qwen leads this local deployment layer with 39.6M monthly GGUF downloads, nearly double Gemma's volume. Simultaneously, agents have become the primary user of the Hub, with Claude Code accounting for 44.4% of agent-tagged traffic in July. The ecosystem is increasingly defined by who provides the weights developers standardize on for fine-tuning and local execution, rather than who holds the single largest parameter count.

KEY FACTS

CHINA MAX PARAMS **2.78T**	US MAX PARAMS (MOST MONTHS) **<130B**
QWEN DERIVATIVES **151448**	QWEN VS LLAMA FOOTPRINT **4.7x**
SMALL MODEL DOWNLOAD SHARE **83%**	CLAUDE CODE AGENT SHARE **44.4%**

WHY IT MATTERS

Practitioners should prioritize Qwen-based families for fine-tuning workflows due to their extensive derivative ecosystem and permissive licensing, while leveraging llama.cpp to run trillion-parameter Chinese models on local hardware clusters.

Read the original at huggingface.co →

---

03 MEDIUM IMPACT arXiv.org

Vero: AI Agents Build Formally Verified Software Repositories

> The Vero benchmark reveals that frontier AI agents cannot yet reliably synthesize formally verified code across real-world, multi-module software repositories.

Researchers have introduced Vero, the first benchmark designed to evaluate joint implementation and proof synthesis at the repository level rather than on isolated functions. While existing evaluations focus on single-function correctness or proof generation for provided code, Vero assesses whether agents can make coherent choices across entire codebases. The dataset comprises 43 multi-module instances sourced from real-world repositories spanning Python, Dafny, Verus, and Coq, covering domains from cryptographic protocols to distributed systems. Each instance is structured as a multi-module Lean 4 repository with predetermined API interfaces, manually curated formal specifications, and reference implementations.

The benchmark supports two evaluation modes: proof-only and code-and-proof synthesis. To ensure data integrity, Vero incorporates an audit mechanism allowing agents to formally prove the unsatisfiability of a specification or the incorrectness of reference code. This process surfaces and corrects latent errors introduced during curation. Evaluation results using frontier coding-agent configurations with Lean toolchain access show significant limitations in current capabilities. The strongest agent fully solved only 27 of the 43 instances and failed to close any specifications on the hardest repositories.

This gap highlights that while agents can handle individual tasks, they struggle with the coherence required for repository-scale verified software synthesis. The authors released the benchmark, curation pipeline, and evaluation harness to provide a concrete testbed for measuring progress. Current tools fall short of producing trustworthy, machine-checked software for complex systems, indicating that substantial architectural or methodological advances are needed before AI can autonomously build verified multi-module systems.

KEY FACTS

BENCHMARK INSTANCES
43

SOURCE LANGUAGES
Python, Dafny, Verus, Coq

TARGET VERIFICATION LANGUAGE
Lean 4

MAX INSTANCES SOLVED
27

SUBMISSION DATE
13 Aug 2026

WHY IT MATTERS

Teams relying on AI for code generation must recognize that current agents lack the capacity to produce mathematically verified, multi-module systems without human intervention. This defines a hard ceiling on autonomy for safety-critical software deployment until agents improve.

Read the original at arXiv.org →

04 MEDIUM IMPACT arXiv.org

OmniScientist: An Omni-Modal Omni-Discipline AI Scientist

> OmniScientist is the first end-to-end AI scientist to conduct multidisciplinary research directly from heterogeneous raw evidence rather than precomputed features.

The system introduces a deterministic pipeline comprising a perception layer and three autonomous agents dedicated to ideation, experimentation, and writeup. Unlike existing workflows that reason over text labels or summaries, this architecture ingests raw modalities including images, signals, audio, video, 3-D structures, trajectories, tables, formulae, and graphs. The agents operate within a closed loop where observations directly shape research questions and experimental decisions throughout the lifecycle. To ensure scientific rigour, the system executes idea, rigour, and claim checks in code, enforcing novelty screening, statistical validity, execution provenance, and numerical traceability without human intervention.

Evaluation covered 36 real-data cases spanning five discipline families and four families of scientific evidence. OmniScientist successfully completed the full path from raw data to a compiled manuscript in every case. Using a reference reasoning backbone, the system achieved a mean overall paper score of 6.3. The study isolated the value of direct perception by comparing the full system against a blind variant restricted to precomputed scalar features. In paired head-to-head judgments, the perception-enabled agent won 85% of comparisons. Direct perception improved performance across all seven evaluation dimensions, demonstrating that access to spatial, temporal, cross-channel, and procedural relations is critical for evidence-grounded discovery.

This work shifts the baseline for automated science from workflow automation on curated inputs to full-stack reasoning on raw data. The results indicate that limiting AI scientists to text or pre-extracted features removes scientifically decisive context required for valid hypothesis generation and testing. By integrating perception directly into the agent loop, the system maintains a continuous link between the final claims and the original empirical evidence, addressing a key gap in current foundation model applications for research.

KEY FACTS

TEST CASES **36**	DISCIPLINE FAMILIES **5**
EVIDENCE FAMILIES **4**	MEAN PAPER SCORE **6.3**
HEAD-TO-HEAD WIN RATE **85%**	EVALUATION DIMENSIONS **7**

WHY IT MATTERS

Builders of automated research systems must integrate raw multimodal perception layers rather than relying on preprocessed features to maintain statistical validity and execution provenance. The 85% win rate in blind comparisons suggests that skipping direct data ingestion fundamentally limits an agent's scientific capability.

Read the original at arXiv.org →

05 HIGH IMPACT arXiv.org

QuoteBench: Benchmark Scores Can Hide Agent Command-Path Failures

> Matched execution scores in LLM coding agents frequently mask severe command-generation failures that are artificially compensated for by interface serialization and reparsing.

The arXiv paper "QuoteBench: How Matched Scores Can Hide Command-Path Failures," submitted on 13 Aug 2026, demonstrates that standard benchmarking methodologies for coding agents conflate model generation quality with transport-layer error correction. The study isolates the boundary between the model's generation contract and the execution transport using 56 one-shot tasks derived from 14 incident families. By introducing a deliberately unescaped parser at the interpolation point, the researchers measured exact final-state validation across eight same-window configurations. The results indicate that replaying identical model replies through this added parser reduces success rates by 55.4 to 73.2 percentage points, proving that raw generation is often insufficient without interface-level fixes.

The data shows that disclosure of this boundary allows models to adapt their generation, recovering between 30.4 and 60.7 percentage points for six of the tested configurations. However, two configurations showed zero or slightly negative recovery, indicating inherent limits in adaptation for certain setups. A specific analysis of GPT-5.6-sol highlights the severity of the masking effect: while the model exhibited a matched score gap of only -3.6 points, this figure concealed -64.3 points of actual damage incurred during command path execution, offset by +60.7 points of compensation from the interface. This suggests that raw generation capabilities at the frontier are nearly saturated, and current performance differentiators rely heavily on boundary adaptation rather than core reasoning improvements.

Furthermore, the deployment configuration significantly alters model rankings compared to raw capability assessments. The study identified one unambiguous reversal among 26 comparable model pairs, with four additional pairs separated by single-task margins when evaluated under these stricter conditions. The authors conclude that evaluations of command-issuing agents must explicitly report the model configuration, generation contract, execution path, operating point, and final-state validator. Treating a matched score as an intrinsic model property ignores the substantial contribution of the execution environment, leading to misleading conclusions about agent reliability in production settings where interface safeguards may differ.

KEY FACTS

TASK COUNT **56**	INCIDENT FAMILIES **14**
SUCCESS DROP RANGE **55.4 to 73.2**	RECOVERY RANGE **30.4 to 60.7**
GPT-5.6-SOL HIDDEN DAMAGE **-64.3**	GPT-5.6-SOL COMPENSATION **+60.7**

WHY IT MATTERS

Practitioners deploying coding agents cannot rely on aggregate success scores, as high performance may stem from fragile interface error-correction rather than robust model generation. Evaluation pipelines must isolate the generation contract from the execution transport to prevent deploying models that fail when serialization behaviors change.

Read the original at arXiv.org →

SECTION 2 OF 3

AI Tools & Ecosystem

4 stories · 2 high · 2 medium

06 HIGH IMPACT Ars Technica

OpenAI and Anthropic in Price War as Chinese AI Rivals Gain Enterprise Ground

> OpenAI and Anthropic have slashed model pricing by up to 80 percent to counter enterprise migration toward cheaper Chinese alternatives.

OpenAI reduced prices for GPT-5.6 Luna, described as its fastest and most affordable model, by 80 percent. Concurrently, Anthropic launched Claude Opus 5, positioning the system's frontier intelligence at half the price of Fable 5, the company's most capable model. These adjustments contribute to a broader market shift where prices for models from leading US labs have decreased by almost a quarter since mid-July, according to Silicon Data's token price index. This index tracks tokens, the units of data processed by language models used to calculate customer bills.

The pricing strategy marks a departure for US AI groups that previously competed primarily on performance with proprietary closed models. Pressure now stems from increasingly capable open Chinese models, which developers can freely download and tweak. Companies including Moonshot and DeepSeek are gaining traction with users from Silicon Valley to Europe as rising AI bills force corporations to curb usage. In response to cost pressures, some enterprises are shifting away from flat subscriptions toward usage-based billing, where payment correlates directly with consumed computational resources.

Corporate adoption of these alternatives is already visible, with companies such as DoorDash and Airbnb stating they have started using Chinese-made models to rein in bills. This migration coincides with a flurry of releases from Chinese labs that have narrowed the performance gap with leading US models. The dynamic creates tension for OpenAI and Anthropic as they plot initial public offerings at trillion-dollar valuations; investors are seeking evidence that vast industry spending can generate returns even as American developers risk losing customers while spending heavily to maintain their technological edge.

KEY FACTS

GPT-5.6 LUNA PRICE CUT
80%

CLAUDE OPUS 5 PRICING
50% of Fable 5

US MODEL PRICE DROP SINCE MID-JULY
25%

ADOPTING ENTERPRISES
DoorDash, Airbnb

CHINESE RIVALS
Moonshot, DeepSeek

WHY IT MATTERS

Practitioners should immediately re-evaluate vendor contracts and benchmark Chinese open models against US proprietary options, as the performance gap has narrowed while cost differentials have widened significantly.

Read the original at Ars Technica →

---

07 HIGH IMPACT TechCrunch

Google Makes Visible AI Watermarks Optional, Keeping Only Invisible SynthID

> Google is decoupling visible branding from its AI generations, making watermarks optional for images, videos, and songs while mandating invisible SynthID tracking.

Effective in the coming days, Google will allow users to toggle off visible watermarks across its generative models, specifically Nano Banana, Omni, and Lyria. This configuration change applies immediately within the Gemini interface and Flow, Google's video editor, with support for Search scheduled for a later release. Users can access this control by navigating to Settings > Media Watermark. The update addresses professional workflows where visible artifacts degrade content utility, yet it maintains a strict backend requirement for provenance data.

The architectural shift preserves two layers of invisible identification: the proprietary SynthID watermark and metadata compliant with the C2PA standard. Josh Woodward, VP for Gemini, confirmed that while the visual overlay is now discretionary, the underlying signals remain active to ensure content can still be identified as AI-generated via Gemini or Search tools. This approach attempts to balance creative control with safety protocols, distinguishing Google's strategy from recent regulatory-driven moves by competitors like Anthropic, which embedded watermarks directly into Claude's text outputs to satisfy EU mandates.

To extend this verification capability beyond its own ecosystem, Google is open sourcing a new library named Credentio. This tool enables developers to embed local validation mechanisms directly into third-party applications, allowing for offline or integrated checks of the persistent invisible markers. The rollout signifies a move toward flexible front-end presentation without compromising the back-end audit trail required for transparency. Unlike previous iterations where visible marking was mandatory, this update treats the watermark as a UI preference rather than a hard constraint on the generated asset itself.

KEY FACTS

AFFECTED MODELS
Nano Banana, Omni, Lyria

INTEGRATION POINTS
Gemini, Flow, Search (coming soon)

PERSISTENT MARKERS
SynthID, C2PA metadata

NEW LIBRARY
Credentio

SETTING PATH
Settings > Media Watermark

WHY IT MATTERS

Practitioners integrating Gemini, Flow, or Lyria into production pipelines can now generate clean assets for client delivery without post-processing, provided they accept the persistence of invisible SynthID and C2PA metadata for downstream verification.

Read the original at TechCrunch →

---

08 MEDIUM IMPACT Simon Willison's Weblog

Don't Classify, Hallucinate: LLM Tagging via Embedding Matching

> Doug Turnbull suggests bypassing large vocabulary constraints by instructing LLMs to hallucinate novel tags before resolving them through vector embedding matching.

Simon Willison highlights a tagging strategy from Doug Turnbull designed to handle classification tasks where the target vocabulary exceeds an LLM's context window. Willison's own blog contains 1,856 distinct tags, a set too large to include directly in a prompt asking the model to select matching categories. Instead of forcing the model to choose from the full list, Turnbull's approach instructs the system to generate plausible, never-before-seen classification strings based solely on the content query.

The method relies on a two-step pipeline. First, the LLM receives a prompt containing examples of the desired tag structure but no actual inventory of existing tags. For instance, when processing a query for a "brown coffee table," the model might output a fabricated hierarchy like "Furniture / Living Room Furniture / Coffee Tables & End Tables / Coffee Tables." This step leverages the model's generative capability to infer semantic intent without the noise of a massive selection list. The prompt explicitly directs the model to create "novel, never seen before" classifications that best fit the search query.

In the second step, the system converts these hallucinated tags into vector embeddings. These vectors are then compared against the embeddings of the actual 1,856 tags in the corpus. The algorithm identifies the concrete, existing tags that sit closest to the generated ones in vector space, effectively mapping the model's imagination to the real database. This technique decouples the reasoning phase from the constraint of a fixed vocabulary size, allowing developers to apply LLM logic to datasets with thousands of potential categories without requiring complex retrieval-augmented generation setups for the initial classification step.

KEY FACTS

EXISTING TAGS
1,856

PROPOSED BY
Doug Turnbull

SOURCE DATE
14th August 2026

WHY IT MATTERS

This pattern allows engineers to apply LLM classification to domains with massive taxonomies that cannot fit into a single prompt context, removing the need to pre-filter candidates before inference.

Read the original at Simon Willison's Weblog →

09 MEDIUM IMPACT arXiv.org

DFM Mimir v1: Open 1B Model Trained Only on Permissible Post-Training Data

> DFM Mimir v1 demonstrates that a 1-billion-parameter Hierarchical Reasoning Model can match larger frontier systems using exclusively permissible post-training data.

Researchers have released Mimir v1, a 1-billion-parameter language model built on the Hierarchical Reasoning Model (HRM) architecture. Unlike current large language model development cycles that depend on massive, often non-permissible datasets, this system was trained from scratch using a mixture of exactly 161 datasets deemed permissible for post-training use. The model targets researchers committed to open-source principles and ethically sourced data, removing the barrier of accessing restricted corpora while maintaining competitive performance levels.

In evaluation across 20 benchmarks covering English, Math & Code, and Danish tasks, Mimir v1 outperforms the original HRM-Text 1B baseline. More significantly, the model competes directly with larger frontier architectures, specifically matching or exceeding the performance of Qwen 3.5 4B and Gemma 4 E2B despite having fewer parameters. The release also establishes a new state of the art for Danish language processing, a specific capability highlighted alongside its general English proficiency. These results suggest that data permissibility does not inherently cap model utility when paired with the HRM architecture.

The model is available immediately on the Hugging Face Hub for deployment and further research. By proving that a 1B parameter system can rival 4B-class models through strict adherence to permissible data sources, the authors provide a viable path for organizations facing legal or ethical constraints on dataset usage. The release includes full weights and documentation, enabling immediate integration into pipelines requiring auditable data lineage without sacrificing benchmark scores in math, code, or natural language understanding.

This development shifts the focus from sheer parameter count and data volume to architectural efficiency and data curation quality. For practitioners building systems in regulated environments, Mimir v1 offers a pre-validated alternative to scraping unrestricted web data, reducing compliance risk while delivering frontier-level outputs in specific domains.

KEY FACTS

PARAMETERS **1 billion**	ARCHITECTURE **Hierarchical Reasoning Model (HRM)**
TRAINING DATASETS **161**	DATA CONSTRAINT **Only permissible post-training data**
BENCHMARKS TESTED **20**	COMPETITOR MODELS **Qwen 3.5 4B, Gemma 4 E2B**

WHY IT MATTERS

Teams operating under strict data governance or licensing constraints can now deploy a 1B-parameter model with verified permissible lineage that rivals larger, less auditable alternatives. This removes the trade-off between ethical sourcing and competitive benchmark performance for English, Math, Code, and Danish workloads.

Read the original at arXiv.org →

SECTION 3 OF 3

AI Applications & Industry

2 stories 1 high 1 medium

10 HIGH IMPACT TechCrunch

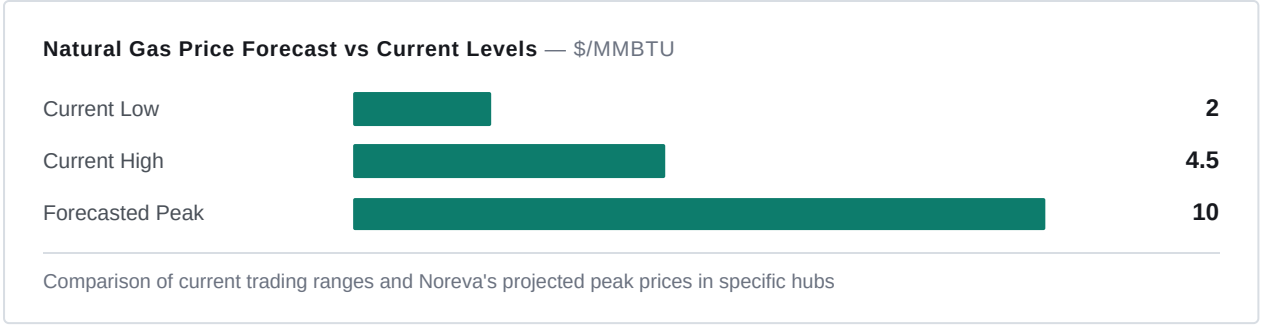
Hyperscalers' Natural Gas Bet Could Backfire as Prices Forecast to Triple

> A new Noreva report warns that natural gas prices in key U.S. hubs could triple, threatening the economic viability of hyperscalers' recent gigawatt-scale power plant investments.

Hyperscalers including Amazon, Google, Meta, and Microsoft have recently committed to building massive natural gas power plants to support AI data center growth, betting on continued low fuel costs. Meta announced a 7.5-gigawatt facility in Louisiana for its Hyperion data center in March, while Microsoft and Google each revealed plans for gigawatt-scale plants in Texas days later. Amazon follows with a planned 7.6-gigawatt gas plant, also in Texas. These moves mark a significant shift from previous renewable-focused strategies toward direct investment in fossil fuel infrastructure, exposing these tech giants to volatile energy markets where they lack historical experience.

Energy research firm Noreva projects that natural gas prices could surge from current levels of $2 to $4.50 per million BTUs to over $10 per million BTUs in certain U.S. hubs. This forecast stems from colliding factors: declining supply growth rates, rising costs for new wells, and the integration of domestic markets with global liquefied natural gas exports. Peter Gardett, CEO of Noreva, notes that while futures contracts currently suggest stability, the connection of regions like West Texas to international markets will eliminate local discounts and create price differentials capable of driving sustained spikes. Fuel constitutes approximately half the cost of electricity from large power plants, meaning a tripling of gas prices would drastically increase operational expenses for these "bring your own power" facilities.

The financial exposure extends beyond direct fuel costs. If hyperscalers absorb higher prices, token costs for AI services may rise; if they pass costs to the grid, regional electricity prices could spike, fueling public backlash. Already, 80% of consumers express concern regarding data centers' impact on utility bills, a sentiment likely to intensify if natural gas costs follow suit. Gardett predicts that future earnings calls for companies like Alphabet will explicitly correlate natural gas pricing with financial results, signaling a new era where energy market volatility directly dictates tech sector profitability. The strategy assumes stable futures, but Noreva argues simple arithmetic indicates a much tighter market than investors currently anticipate.



KEY FACTS

FORECASTED PRICE PEAK
$10 per million BTUs

CURRENT PRICE RANGE
$2 to $4.50 per million BTUs

META PLANT CAPACITY
7.5 gigawatts

AMAZON PLANT CAPACITY
7.6 gigawatts

FUEL COST SHARE
50%

CONSUMER CONCERN RATE
80%

WHY IT MATTERS

Practitioners relying on hyperscaler cloud services for AI workloads may face increased token costs if operators pass through tripling natural gas expenses. Infrastructure planners must now factor long-term fuel price volatility into total cost of ownership models for dedicated data center deployments.

Read the original at TechCrunch →

---

11 MEDIUM IMPACT TechCrunch

Kog Goes Deeper into GPU Architecture to Squeeze More Inference Per Chip

> French startup Kog claims software optimization can deliver 30x faster inference on standard datacenter GPUs without new hardware.

Kog, a French startup founded by Gaël Delalleau, is pursuing extreme single-request decoding speeds on conventional datacenter GPUs rather than relying on purpose-built silicon. The company's tech preview, which gained traction on Hacker News in May, demonstrated 3,000 tokens per second (TPS) per request using AMD MI300X and Nvidia H200 GPUs. This performance was achieved with Laneformer 2B, an open-sourced model with approximately 2 billion parameters. While the current benchmark relies on a small model, Kog asserts its approach scales to larger large language models (LLMs), targeting a 30x inference speedup overall.

The company's methodology diverges from typical framework optimizations by operating at the assembly and binary levels. Delalleau, whose background includes solid-state physics and offensive cybersecurity, applies a reverse-engineering mindset to unlock unused memory bandwidth on newer GPU architectures. This deep-level focus requires significant engineering effort; the 11-person team dedicates weeks or months to analyze each new GPU type. Consequently, Kog currently supports a limited range of chips, though future plans involve agent-based pipelines to expand hardware compatibility. The approach contrasts with competitors like ZML, which offers hardware-agnostic software bypassing CUDA, by focusing on architecture-specific maximization similar to Stanford's Hazy Research.

Market validation remains the immediate hurdle. While Kog reported 200 tangible business leads following its May preview, prospective customers have hesitated to fine-tune small models, pushing the startup to prioritize acceleration for larger models used in professional workflows. Design partners generating games and apps via prompts represent early adopters who equate reduced latency with increased revenue. Kog aims to demonstrate a 10x speedup on a major model by September, a milestone Delalleau identifies as critical for securing Series A funding. The company is backed by Varsity VC, Scaleway, Bpifrance, and the French Tech 2030 program, aligning with broader European efforts to build sovereign AI infrastructure capabilities.

KEY FACTS

DEMO THROUGHPUT
3000 TPS

TARGET SPEEDUP
30x

DEMO MODEL SIZE
2B parameters

SUPPORTED GPUS
AMD MI300X, Nvidia H200

TEAM SIZE
11

TARGET DATE
September

> **WHY IT MATTERS**
>
> If Kog delivers its promised 10x speedup on major models by September, practitioners could significantly reduce inference latency and costs on existing Nvidia H200 or AMD MI300X clusters without hardware upgrades.

Read the original at TechCrunch →

Sources

01 GLM-5.3: Chinese Labs Match the Frontier Through Post-Training Alone
https://www.interconnects.ai/p/glm-53-how-chinese-labs-keep-stride

02 State of Open Models: China's Open Weights Dwarf America's, Qwen Becomes Community Base
https://huggingface.co/blog/state-of-open-models-summer-2026

03 Vero: AI Agents Build Formally Verified Software Repositories
https://arxiv.org/abs/2608.13522

04 OmniScientist: An Omni-Modal Omni-Discipline AI Scientist
https://arxiv.org/abs/2608.13558

05 QuoteBench: Benchmark Scores Can Hide Agent Command-Path Failures
https://arxiv.org/abs/2608.13547

06 OpenAI and Anthropic in Price War as Chinese AI Rivals Gain Enterprise Ground
https://arstechnica.com/ai/2026/08/openai-and-anthropic-in-price-war-as-chinese-ai-rivals-gain-ground/

07 Google Makes Visible AI Watermarks Optional, Keeping Only Invisible SynthID
https://techcrunch.com/2026/08/14/google-will-now-allow-users-to-remove-visible-watermark-from-its-ai-generations/

08 Don't Classify, Hallucinate: LLM Tagging via Embedding Matching
https://simonwillison.net/2026/Aug/14/dont-classify-hallucinate/

09 DFM Mimir v1: Open 1B Model Trained Only on Permissible Post-Training Data
https://arxiv.org/abs/2608.13517

10 Hyperscalers' Natural Gas Bet Could Backfire as Prices Forecast to Triple
https://techcrunch.com/2026/08/14/hyperscalers-might-regret-embracing-natural-gas-if-new-forecast-proves-correct/

11 Kog Goes Deeper into GPU Architecture to Squeeze More Inference Per Chip
https://techcrunch.com/2026/08/14/kog-is-going-deeper-to-squeeze-more-inference-out-of-gpus/

About this document. Every story in the 15 August 2026 New Horizon AI Digest, reported at length. Each entry is written from the publisher's own article text; where a source could not be retrieved the entry is explicitly marked and kept short rather than padded.

Images and licensing. Figures are used only where the source licence permits redistribution, and are credited in the caption. Publisher artwork is not reproduced. All titles link to the original publication.

Generated 2026-08-15T05:34:21Z · model qwen3.5:397b-cloud · New Horizon, new-horizon.tech · To unsubscribe or manage your subscription, see the footer of the daily email.