TROY TECHNICAL

AI & MACHINE LEARNING

WEEKLY REPORT

November 14 - November 20, 2026 | Issue 2026-W46

Deployable systems, permissions, infrastructure and exit readiness

11	10
ANALYZED ARTICLES	PRIORITY THEMES

Issued 2026-11-20 | English edition | fixed approved cover artwork



AI & Machine Learning: the boundary moved from isolated claims to qualification evidence

11 English articles were reviewed independently from the Japanese edition. Ten themes are retained for management decisions.

SIGNAL 01 | MODELS & DATA

Fujitsu Japan AI: Clinical trial matching and 2027 roadmap

Fujitsu Japan, the National Cancer Center Hospital East, and Paradigm demonstrated an AI-driven system that reduced the patient review pool for clinical trials by 89%.

SIGNAL 02 | COMPUTE & DEVICES

NVIDIA, IBM, and Fermilab: Quantum-Classical Codesign Platform

NVIDIA, IBM, and Fermi National Accelerator Laboratory have launched a modular quantum-classical codesign platform to advance error-correction and scientific applications.

SIGNAL 03 | INFRASTRUCTURE

OpenAI Chip Design: PPA Gains for Jalapeño and AMD Hardware

OpenAI models have demonstrated the ability to improve Power, Performance, and Area (PPA) in production chip designs, including a 13% area reduction in Jalapeño's vector unit.

SIGNAL 04 | DEPLOYMENT

NVIDIA Vera Rubin: MLPerf Training v6.1 and Endpoints Results

NVIDIA's Vera Rubin platform achieved leading performance across MLPerf Training v6.1 and MLPerf Endpoints benchmarks, including a 10x generational gain in throughput per watt for agentic AI.

DECISION

Focus this week's management attention on deployable systems, permissions, infrastructure and exit readiness. Move only when the linked evidence can be reproduced under your own operating conditions.

THREE MANAGEMENT MOVES

1. Buyers: benchmark representative workflows and define permission, audit and shutdown requirements.
2. Suppliers: translate model or hardware capability into workload-level latency, energy, reliability and lifecycle evidence.
3. Executives: track integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

EVIDENCE DISCIPLINE

FACT states what the collected English source reports. BUSINESS READ and ACTION are editorial interpretations. UNKNOWN lists information that the source file does not establish. A linked article is not treated as independent confirmation of every claim.

Numbers and evidence anchors that require conditions

89%
A02

SOURCE-BOUND CONDITION
Fujitsu Japan, the National Cancer Center Hospital East, and Paradigm demonstrated an AI-driven system that reduced the patient review pool for clinical trials by 89%.

000x
A09

SOURCE-BOUND CONDITION
The initiative includes a new decoding algorithm, TCBP-JOBS, which accelerates quantum error-correction research by up to 13,000x.

13%
A10

SOURCE-BOUND CONDITION
OpenAI models have demonstrated the ability to improve Power, Performance, and Area (PPA) in production chip designs, including a 13% area reduction in Jalapeño's vector unit.

10x
A07

SOURCE-BOUND CONDITION
NVIDIA's Vera Rubin platform achieved leading performance across MLPerf Training v6.1 and MLPerf Endpoints benchmarks, including a 10x generational gain in throughput per watt for agentic AI.

5.5 w
A06

SOURCE-BOUND CONDITION
These models significantly reduce costs compared to GPT-6 and Claude Opus 5.5 while maintaining high quality for enterprise coding, security, and image editing.

HOW TO USE THESE NUMBERS
Each value is shown with the condition stated in the collected English article. Do not compare values across different test methods, device sizes, operating temperatures or commercial stages without normalizing those conditions.

Where the value chain moved

Each stage combines one leading theme and one supporting theme. The report treats handoffs as evidence transfers, not decorative arrows.

1

Models & data

P01/P06

Fujitsu Japan, the National Cancer Center Hospital East, and Paradigm demonstrated an AI-driven system that reduced the.

2

Compute & devices

P02/P07

NVIDIA, IBM, and Fermi National Accelerator Laboratory have launched a modular quantum-classical codesign platform to advance.

3

Infrastructure

P03/P08

OpenAI models have demonstrated the ability to improve Power, Performance, and Area (PPA) in production chip designs.

4

Deployment

P04/P09

NVIDIA's Vera Rubin platform achieved leading performance across MLPerf Training v6.1 and MLPerf Endpoints benchmarks.

5

Governance

P05/P10

Microsoft AI has released seven new models focused on price-performance, including MAI Cyber 2 Flash and MAI Code 2.

EVIDENCE REQUIRED AT EACH HANDOFF

Models & data to Compute & devices	Transfer test conditions, acceptance criteria, failure data, owner and decision date. Recheck integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.
Compute & devices to Infrastructure	Transfer test conditions, acceptance criteria, failure data, owner and decision date. Recheck integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.
Infrastructure to Deployment	Transfer test conditions, acceptance criteria, failure data, owner and decision date. Recheck integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.
Deployment to Governance	Transfer test conditions, acceptance criteria, failure data, owner and decision date. Recheck integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

COMMON CONTROL POINT

No theme moves to a commercial commitment until a named owner has reproduced the relevant evidence and documented integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

Priority map: maturity and next decision locus



P01	Fujitsu Japan AI: Clinical trial matching and 2027 roadmap System / adoption | cue: Clinical
P02	NVIDIA, IBM, and Fermilab: Quantum-Classical Codesign Platform System / adoption | cue: Platform
P03	OpenAI Chip Design: PPA Gains for Jalapeño and AMD Hardware Tech / process | cue: no system/business cue
P04	NVIDIA Vera Rubin: MLPerf Training v6.1 and Endpoints Results System / adoption | cue: platform
P05	Microsoft AI: Seven new models and Maia 200 specs for 2026 Tech / process | cue: no system/business cue
P06	OpenAI and AMD: AI-driven 14% PPA gains in GPU designs System / adoption | cue: integrated
P07	NVIDIA Nemotron for Science: Specialized Agents and LitmusBench v0.2 System / adoption | cue: adopters
P08	ModBERT: DNA LLM for Genetic Variants and Methylation Analysis Business / rules | cue: regulation
P09	CGMap: Interpretable Graph Learning for Cancer Gene Screening System / adoption | cue: integrating
P10	GPT-6.2 Astra and Sol: OpenAI's Professional Model Specs Tech / process | cue: no system/business cue

REPRODUCIBLE POSITIONING RULE

Horizontal position uses five discrete stages: Research, Prototype, Joint evaluation, Product adoption and Scale. The vertical bands classify the next decision locus: Tech / process, System / adoption, or Business / rules (including supply systems). Article-title cues take precedence; the overview is used only when the title is silent. These are categories, not impact ranks. Bubble diameter reflects the editorial score.

FACT is source-bound. BUSINESS READ, ACTION and UNKNOWN are explicit editorial layers.

P01 Fujitsu Japan AI: Clinical trial matching and 2027 roadmap

A02 | Models & data | Joint evaluation

2026-11-20

FACT

Fujitsu Japan, the National Cancer Center Hospital East, and Paradigm demonstrated an AI-driven system that reduced the patient review pool for clinical trials by 89%. By structuring unstructured medical records using large language models, the study identified 57 priority candidates from a respiratory medicine population. This technology aims to accelerate trial enrollment and address Japan's " drug loss" issue.

WHY IT MATTERS

The decision relevance is deployable systems, permissions, infrastructure and exit readiness. The signal should be tested at the application and operating-system level, not accepted from a headline alone.

BUSINESS READ

For operating companies, compare the reported change with current qualification, sourcing and product-roadmap assumptions. For suppliers, translate model or hardware capability into workload-level latency, energy, reliability and lifecycle evidence.

ACTION

Within 30 days, benchmark representative workflows and define permission, audit and shutdown requirements; record the evidence and owner against P01.

UNKNOWN

Still unproven or incompletely stated: integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

EDITORIAL SCORE 21/25 | MATURITY 3/5 | DECISION LOCUS: SYSTEM / ADOPTION

P02 NVIDIA, IBM, and Fermilab: Quantum-Classical Codesign Platform

A09 | Compute & devices | Research

2026-11-16

FACT

NVIDIA, IBM, and Fermi National Accelerator Laboratory have launched a modular quantum-classical codesign platform to advance error-correction and scientific applications. The initiative includes a new decoding algorithm, TCBP-JOBS, which accelerates quantum error-correction research by up to 13,000x. These developments aim to scale quantum computers for practical use in high-energy and nuclear physics.

WHY IT MATTERS

The decision relevance is deployable systems, permissions, infrastructure and exit readiness. The signal should be tested at the application and operating-system level, not accepted from a headline alone.

BUSINESS READ

For operating companies, compare the reported change with current qualification, sourcing and product-roadmap assumptions. For suppliers, translate model or hardware capability into workload-level latency, energy, reliability and lifecycle evidence.

ACTION

Within 30 days, benchmark representative workflows and define permission, audit and shutdown requirements; record the evidence and owner against P02.

UNKNOWN

Still unproven or incompletely stated: integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

EDITORIAL SCORE 20/25 | MATURITY 1/5 | DECISION LOCUS: SYSTEM / ADOPTION

FACT is source-bound. BUSINESS READ, ACTION and UNKNOWN are explicit editorial layers.

P03 OpenAI Chip Design: PPA Gains for Jalapeño and AMD Hardware

A10 | Infrastructure | Product adoption

2026-11-18

FACT

OpenAI models have demonstrated the ability to improve Power, Performance, and Area (PPA) in production chip designs, including a 13% area reduction in Jalapeño's vector unit. Collaborations with AMD achieved up to 14% PPA improvement on flagship GPU IP, compressing roughly a year of progress into less than five months. These results indicate that frontier models can reason at the level of expert microarchitects to optimize SystemVerilog RTL.

WHY IT MATTERS

The decision relevance is deployable systems, permissions, infrastructure and exit readiness. The signal should be tested at the application and operating-system level, not accepted from a headline alone.

BUSINESS READ

For operating companies, compare the reported change with current qualification, sourcing and product-roadmap assumptions. For suppliers, translate model or hardware capability into workload-level latency, energy, reliability and lifecycle evidence.

ACTION

Within 30 days, benchmark representative workflows and define permission, audit and shutdown requirements; record the evidence and owner against P03.

UNKNOWN

Still unproven or incompletely stated: integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

EDITORIAL SCORE 20/25 | MATURITY 4/5 | DECISION LOCUS: TECH / PROCESS

P04 NVIDIA Vera Rubin: MLPerf Training v6.1 and Endpoints Results

A07 | Deployment | Pilot

2026-11-16

FACT

NVIDIA's Vera Rubin platform achieved leading performance across MLPerf Training v6.1 and MLPerf Endpoints benchmarks, including a 10x generational gain in throughput per watt for agentic AI. These results demonstrate significant scaling efficiency and power optimization for training and agentic inference workloads.

WHY IT MATTERS

The decision relevance is deployable systems, permissions, infrastructure and exit readiness. The signal should be tested at the application and operating-system level, not accepted from a headline alone.

BUSINESS READ

For operating companies, compare the reported change with current qualification, sourcing and product-roadmap assumptions. For suppliers, translate model or hardware capability into workload-level latency, energy, reliability and lifecycle evidence.

ACTION

Within 30 days, benchmark representative workflows and define permission, audit and shutdown requirements; record the evidence and owner against P04.

UNKNOWN

Still unproven or incompletely stated: integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

EDITORIAL SCORE 19/25 | MATURITY 2/5 | DECISION LOCUS: SYSTEM / ADOPTION

FACT is source-bound. BUSINESS READ, ACTION and UNKNOWN are explicit editorial layers.

P05 Microsoft AI: Seven new models and Maia 200 specs for 2026

A06 | Governance | Research

2026-11-17

FACT

Microsoft AI has released seven new models focused on price-performance, including MAI Cyber 2 Flash and MAI Code 2. These models significantly reduce costs compared to GPT-6 and Claude Opus 5.5 while maintaining high quality for enterprise coding, security, and image editing.

WHY IT MATTERS

The decision relevance is deployable systems, permissions, infrastructure and exit readiness. The signal should be tested at the application and operating-system level, not accepted from a headline alone.

BUSINESS READ

For operating companies, compare the reported change with current qualification, sourcing and product-roadmap assumptions. For suppliers, translate model or hardware capability into workload-level latency, energy, reliability and lifecycle evidence.

ACTION

Within 30 days, benchmark representative workflows and define permission, audit and shutdown requirements; record the evidence and owner against P05.

UNKNOWN

Still unproven or incompletely stated: integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

EDITORIAL SCORE 19/25 | MATURITY 1/5 | DECISION LOCUS: TECH / PROCESS

P06 OpenAI and AMD: AI-driven 14% PPA gains in GPU designs

A11 | Models & data | Research

2026-11-18

FACT

OpenAI's models achieved up to 14% gains in power, performance, and area (PPA) on AMD's flagship GPU and single-digit gains on next-generation CPU blocks. These AI-authored RTL edits have already been integrated into designs sent for manufacturing, signaling a shift toward AI-automated chip design.

WHY IT MATTERS

The decision relevance is deployable systems, permissions, infrastructure and exit readiness. The signal should be tested at the application and operating-system level, not accepted from a headline alone.

BUSINESS READ

For operating companies, compare the reported change with current qualification, sourcing and product-roadmap assumptions. For suppliers, translate model or hardware capability into workload-level latency, energy, reliability and lifecycle evidence.

ACTION

Within 30 days, benchmark representative workflows and define permission, audit and shutdown requirements; record the evidence and owner against P06.

UNKNOWN

Still unproven or incompletely stated: integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

EDITORIAL SCORE 19/25 | MATURITY 1/5 | DECISION LOCUS: SYSTEM / ADOPTION

FACT is source-bound. BUSINESS READ, ACTION and UNKNOWN are explicit editorial layers.

P07 NVIDIA Nemotron for Science: Specialized Agents and LitmusBench v0.2

A08 | Compute & devices | Research

2026-11-16

FACT

NVIDIA has released Nemotron for Science, combining the Nemotron 3.5 Super VL model with the LitmusBench v0.2 chemistry dataset to enable domain-specific scientific agents. Early adopters including Argonne National Laboratory and various life sciences firms are using the framework to accelerate drug discovery and materials synthesis.

WHY IT MATTERS

The decision relevance is deployable systems, permissions, infrastructure and exit readiness. The signal should be tested at the application and operating-system level, not accepted from a headline alone.

BUSINESS READ

For operating companies, compare the reported change with current qualification, sourcing and product-roadmap assumptions. For suppliers, translate model or hardware capability into workload-level latency, energy, reliability and lifecycle evidence.

ACTION

Within 30 days, benchmark representative workflows and define permission, audit and shutdown requirements; record the evidence and owner against P07.

UNKNOWN

Still unproven or incompletely stated: integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

EDITORIAL SCORE 19/25 | MATURITY 1/5 | DECISION LOCUS: SYSTEM / ADOPTION

P08 ModBERT: DNA LLM for Genetic Variants and Methylation Analysis

A03 | Infrastructure | Research

Prior release 2026-11-13

FACT

ModBERT is a DNA large language model that integrates genetic variants and DNA methylation using a five-letter probabilistic genome representation. It outperforms models trained without methylation in classifying functional genomic assay peaks and predicts chromatin accessibility at held-out loci with an AUROC of 0.88. This framework enables person-specific predictions of gene regulation by moving beyond shared reference genomes.

WHY IT MATTERS

The decision relevance is deployable systems, permissions, infrastructure and exit readiness. The signal should be tested at the application and operating-system level, not accepted from a headline alone.

BUSINESS READ

For operating companies, compare the reported change with current qualification, sourcing and product-roadmap assumptions. For suppliers, translate model or hardware capability into workload-level latency, energy, reliability and lifecycle evidence.

ACTION

Within 30 days, benchmark representative workflows and define permission, audit and shutdown requirements; record the evidence and owner against P08.

UNKNOWN

Still unproven or incompletely stated: integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

EDITORIAL SCORE 19/25 | MATURITY 1/5 | DECISION LOCUS: BUSINESS / RULES

FACT is source-bound. BUSINESS READ, ACTION and UNKNOWN are explicit editorial layers.

P09 CGMap: Interpretable Graph Learning for Cancer Gene Screening

A01 | Deployment | Research

2026-11-16

FACT

CGMap is an interpretable graph-learning framework designed to prioritize potential cancer genes (CGs) by integrating sparse oncogenic signals from biomolecular networks. The system analyzed 12,022 genes across a pan-cancer setting, identifying 70 reliable candidates, including those previously unannotated in established databases.

WHY IT MATTERS

The decision relevance is deployable systems, permissions, infrastructure and exit readiness. The signal should be tested at the application and operating-system level, not accepted from a headline alone.

BUSINESS READ

For operating companies, compare the reported change with current qualification, sourcing and product-roadmap assumptions. For suppliers, translate model or hardware capability into workload-level latency, energy, reliability and lifecycle evidence.

ACTION

Within 30 days, benchmark representative workflows and define permission, audit and shutdown requirements; record the evidence and owner against P09.

UNKNOWN

Still unproven or incompletely stated: integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

EDITORIAL SCORE 19/25 | MATURITY 1/5 | DECISION LOCUS: SYSTEM / ADOPTION

P10 GPT-6.2 Astra and Sol: OpenAI's Professional Model Specs

A05 | Governance | Pilot

2026-11-17

FACT

OpenAI has released GPT-6.2 Astra for complex professional projects and GPT-6.2 Sol for versatile everyday work. Astra achieves higher benchmarks in scientific and legal tasks at significantly lower costs than previous models, while Sol provides a high-efficiency alternative for agentic workflows.

WHY IT MATTERS

The decision relevance is deployable systems, permissions, infrastructure and exit readiness. The signal should be tested at the application and operating-system level, not accepted from a headline alone.

BUSINESS READ

For operating companies, compare the reported change with current qualification, sourcing and product-roadmap assumptions. For suppliers, translate model or hardware capability into workload-level latency, energy, reliability and lifecycle evidence.

ACTION

Within 30 days, benchmark representative workflows and define permission, audit and shutdown requirements; record the evidence and owner against P10.

UNKNOWN

Still unproven or incompletely stated: integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.

EDITORIAL SCORE 19/25 | MATURITY 2/5 | DECISION LOCUS: TECH / PROCESS

Playbook and 0-5 year roadmap

STAKEHOLDER	NOW | 0-30 DAYS	NEXT | 1-12 MONTHS	LATER | 1-5 YEARS
Operating companies	benchmark representative workflows and define permission, audit and shutdown requirements	Run production-like qualification with explicit acceptance and stop criteria.	Build a portfolio and architecture that can absorb technology and supplier changes.
Materials, equipment and technology suppliers	translate model or hardware capability into workload-level latency, energy, reliability and lifecycle evidence	Create shared reliability, process-window and failure-analysis data with customers.	Standardize interfaces, traceability, lifecycle support and recovery services.
Trading, investment and legal teams	Map integration timing, operating cost, neutrality, permissions and measurable workflow outcomes.	Contract for capacity, quality, change notice, liability, alternatives and exit.	Diversify regional, technical and customer concentration risk.

ROADMAP

NOW | 0-30 DAYS
Evidence ledger
Owner: Strategy + quality
Deliverable: claim, source, condition, owner

NOW | 0-30 DAYS
Risk screen
Owner: Procurement + legal
Deliverable: unknowns, alternatives, stop

NEXT | 1-12 MONTHS
Joint qualification
Owner: R&D + supplier
Deliverable: process window and failure data

NEXT | 1-12 MONTHS
Commercial gate
Owner: Business + finance
Deliverable: unit economics and capacity

LATER | 1-5 YEARS
Platform strategy
Owner: Executive team
Deliverable: portfolio and interface roadmap

LATER | 1-5 YEARS
Service layer
Owner: Operations + channel
Deliverable: monitoring, maintenance and recovery

GO / NO-GO GATES

GATE	OWNER	REQUIRED EVIDENCE	NO-GO CONDITION
G1 Evidence	**Quality**	Source, test conditions and reproducibility	No traceable evidence
G2 Integration	**R&D**	Interface, process window and failure analysis	Cannot meet system conditions
G3 Commercial	**Business**	Capacity, total cost, contract and alternatives	Economics or supply cannot be supported
G4 Lifecycle	**Operations**	Monitoring, maintenance, change notice and exit	No accountable operating plan

ANALYZED ARTICLES | ITEMS 01-11

Every title and URL is displayed in full. URLs wrap without shortening and remain clickable. A verified-reference label means the translated HTML did not retain its original source URL.

ID	FULL ARTICLE TITLE	FULL SOURCE / VERIFIED REFERENCE URL - CLICKABLE
A01	CGMap: Interpretable Graph Learning for Cancer Gene Screening	https://www.nature.com/articles/s42256-026-01327-4
A02	Fujitsu Japan AI: Clinical trial matching and 2027 roadmap	https://jp.ibtimes.com/fujitsu-japan-ai-study-cuts-clinical-trial-patient-review-pool-89-105276
A03	ModBERT: DNA LLM for Genetic Variants and Methylation Analysis	https://doi.org/10.64898/2026.11.12.766258
A04	Gemini 4 Neon: Specs, Benchmarks, and Pricing for Google AI	https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-4-neon/
A05	GPT-6.2 Astra and Sol: OpenAI's Professional Model Specs	https://openai.com/index/introducing-gpt-6-2-astra-sol/
A06	Microsoft AI: Seven new models and Maia 200 specs for 2026	https://microsoft.ai/news/microsoft-ignite-2026/
A07	NVIDIA Vera Rubin: MLPerf Training v6.1 and Endpoints Results	https://blogs.nvidia.com/blog/mlperf-training-endpoints-vera-rubin/
A08	NVIDIA Nemotron for Science: Specialized Agents and LitmusBench v0.2	https://nvidianews.nvidia.com/news/nvidia-introduces-nemotron-for-science-to-accelerate-discoveries-with-agents
A09	NVIDIA, IBM, and Fermilab: Quantum-Classical Codesign Platform	https://nvidianews.nvidia.com/news/nvidia-announces-collaboration-with-ibm-and-fermi-national-accelerator-laboratory-to-launch-quantum-classical-codesign-platform-plus-error-correction-breakthrough-to-advance-practical-quantum-computing
A10	OpenAI Chip Design: PPA Gains for Jalapeño and AMD Hardware	https://openai.com/index/advancing-chip-design-with-our-models/
A11	OpenAI and AMD: AI-driven 14% PPA gains in GPU designs	https://aiweekly.co/alerts/openai-says-models-found-14-ppa-gains-in-amd-gpu-designs

AI_MachineLearning — Selected Articles

Date: 2026-11-21

Articles: 11

Table of Contents

#01 CGMap: Interpretable Graph Learning for Cancer Gene Screening

#02 Fujitsu Japan AI: Clinical trial matching and 2027 roadmap

#03 ModBERT: DNA LLM for Genetic Variants and Methylation Analysis

#04 Gemini 4 Neon: Specs, Benchmarks, and Pricing for Google AI

#05 GPT-6.2 Astra and Sol: OpenAI's Professional Model Specs

#06 Microsoft AI: Seven new models and Maia 200 specs for 2026

#07 NVIDIA Vera Rubin: MLPerf Training v6.1 and Endpoints Results

#08 NVIDIA Nemotron for Science: Specialized Agents and LitmusBench v0.2

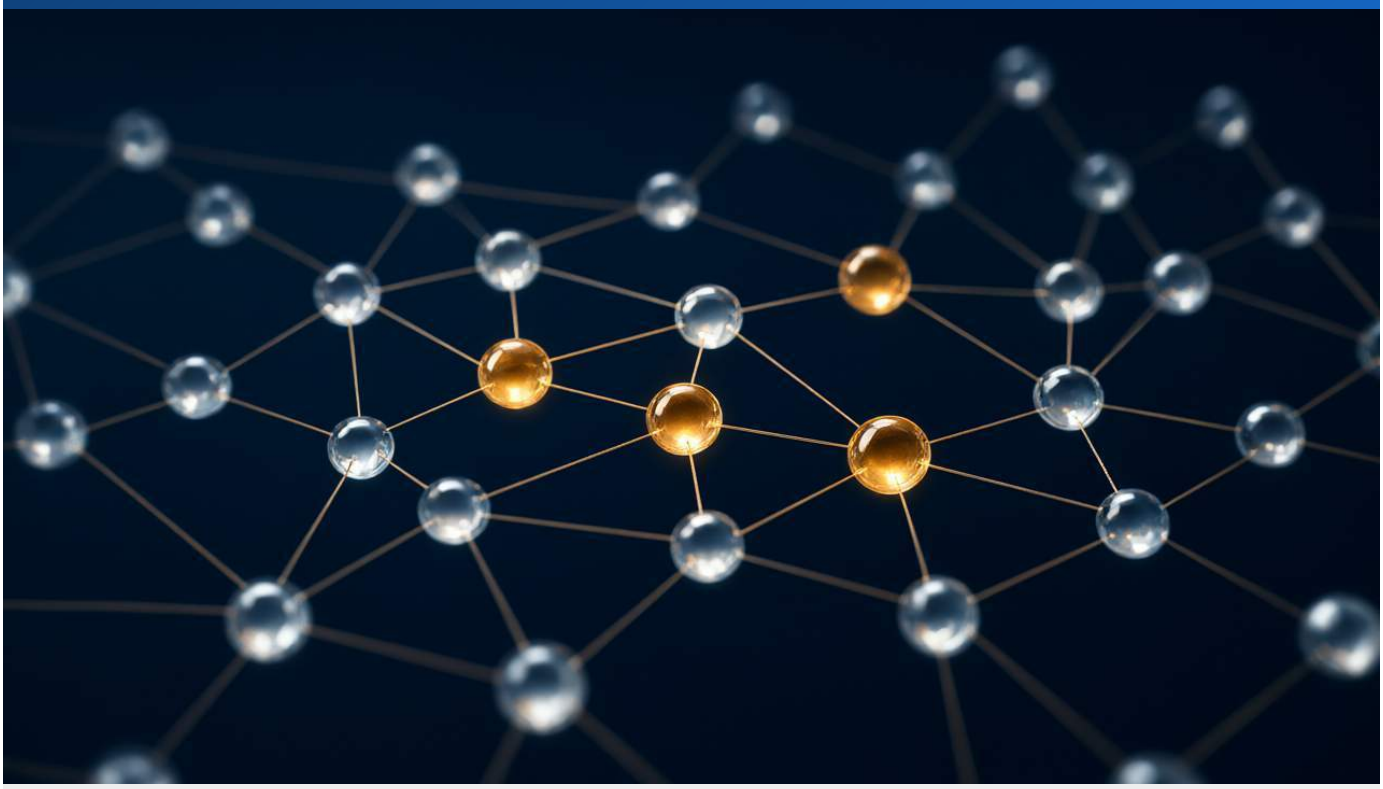
#09 NVIDIA, IBM, and Fermilab: Quantum-Classical Codesign Platform

#10 OpenAI Chip Design: PPA Gains for Jalapeño and AMD Hardware

#11 OpenAI and AMD: AI-driven 14% PPA gains in GPU designs

#01 CGMap: Interpretable Graph Learning for Cancer Gene Screening

Published November 16, 2026 Nature Machine Intelligence



OVERVIEW

CGMap is an interpretable graph-learning framework designed to prioritize potential cancer genes (CGs) by integrating sparse oncogenic signals from biomolecular networks. The system analyzed 12,022 genes across a pan-cancer setting, identifying 70 reliable candidates, including those previously unannotated in established databases.

IN DEPTH

Key Findings

CGMap consistently outperformed existing methods in prioritizing cancer genes while maintaining computational efficiency. In a pan-cancer analysis of 12,022 genes, the framework identified 70 reliable candidates. Parallel loss-of-function assays across cancer and non-malignant cell lines confirmed that these prioritized candidates are associated with the suppression of tumor-related phenotypes, specifically reducing cell proliferation, migration, and invasion, and shifting epithelial-to-mesenchymal transition (EMT)-related features toward a non-mesenchymal state.

Technical Details

The framework utilizes a width-oriented parallel architecture to localize spatially dispersed CG signals. Key technical components include:

- **Distance-Resolved Channels:** These channels localize signals across multiple topological scales to prevent representation homogenization.
- **Global Diffusion Module:** This module quantifies gene cooperativity within pathway contexts. It generates traceable virtual interactions that map predictions to potential regulatory hypotheses, providing intrinsic interpretability for the model's decisions.

Background & Context

Precision oncology requires the systematic screening of cancer genes, but integrating sparse and spatially dispersed oncogenic signals to find testable candidates remains a significant challenge. CGMap addresses this by using biomolecular networks to trace mechanistic cues. The study utilized several data sources for validation and benchmarking, including GEPIA2, TIMER 2.0, CancerSEA, and the CCG cohort from NCG7.2. The framework's development also incorporated the Optuna v.4 framework for automated hyperparameter search.

Significance & Outlook

By providing an interpretable method to prioritize candidates that were previously unannotated in established databases, CGMap offers a scalable approach to identifying novel cancer gene targets. The ability to map predictions to regulatory hypotheses allows researchers to move from computational prediction to biological verification more efficiently. The source code and precomputed data have been made available via GitHub and Zenodo.

Source: https://www.nature.com/articles/s42256-026-01327-4

Collected: November 21, 2026 | Automated Research System (AI)

#02 Fujitsu Japan AI: Clinical trial matching and 2027 roadmap

Published November 20, 2026 International Business Times Japan



OVERVIEW

Fujitsu Japan, the National Cancer Center Hospital East, and Paradigm demonstrated an AI-driven system that reduced the patient review pool for clinical trials by 89%. By structuring unstructured medical records using large language models, the study identified 57 priority candidates from a respiratory medicine population. This technology aims to accelerate trial enrollment and address Japan's "drug loss" issue.

IN DEPTH

Key Findings

A joint study conducted between June and October 2026 demonstrated that integrating large language models (LLMs) with the Paradigm Platform can significantly narrow the search for clinical-trial participants. In a simulated trial for patients with EGFR gene changes, the system identified 57 candidates for priority review. Upon physician review of the AI-cited records, 51 of these candidates were judged to be promising. The automated overnight screening process reduced the number of patient records requiring manual review by medical staff by approximately 89%.

Technical Details

The research utilized records from 22,334 respiratory medicine patients. Fujitsu Japan processed electronic medical records for 3,027 lung cancer patients within this group, performing anonymization and converting unstructured information into structured data. The LLM extracted critical eligibility data from genomic test results, treatment histories, and clinical notes. This structured data was then linked to the Paradigm Platform, which evaluated 540 patients with EGFR gene changes against five specific trial participation criteria.

Background and Context

Trial recruitment is traditionally labor-intensive, relying on physician experience and manual record examination. Modern eligibility requirements often involve complex biomarkers and genetic changes, with data scattered across various clinical documents. Furthermore, Japan faces a "drug loss" problem where medicines available overseas are not approved domestically. This is exacerbated by the prevalence of free-text information in Japanese electronic medical records and inconsistent data formats across different hospitals and countries.

Significance and Outlook

Fujitsu Japan targets the launch of a commercial service supporting candidate searches and enrollment during fiscal 2027 (ending March 2028). The service will connect with the Paradigm Platform to manage the process from identification to enrollment. Future development goals include expanding coverage to additional clinical departments and diseases, improving selection accuracy, and promoting adoption across Japanese medical institutions. While the AI identifies candidates, final eligibility and enrollment remain subject to physician determination and patient consent.

Source: https://jp.ibtimes.com/fujitsu-japan-ai-study-cuts-clinical-trial-patient-review-pool-89-105276

Collected: November 21, 2026 | Automated Research System (AI)

#03 ModBERT: DNA LLM for Genetic Variants and Methylation Analysis

Published November 13, 2026 bioRxiv



OVERVIEW

ModBERT is a DNA large language model that integrates genetic variants and DNA methylation using a five-letter probabilistic genome representation. It outperforms models trained without methylation in classifying functional genomic assay peaks and predicts chromatin accessibility at held-out loci with an AUROC of 0.88. This framework enables person-specific predictions of gene regulation by moving beyond shared reference genomes.

IN DEPTH

Key Findings

ModBERT demonstrates superior performance in classifying peaks from functional genomic assays compared to ab initio trained models and versions of ModBERT pretrained without methylation. The model effectively utilizes methylation data to classify 11 different histone marks and 69 transcription factor (TF) binding ChIP-seq data. In evaluation, ModBERT achieved an AUROC of 0.88 when predicting chromatin accessibility at held-out loci in held-out individuals. Furthermore, its predicted variant effects correlate with likely causal caQTLs (Spearman $\rho = 0.52$), while avoiding variants associated only through linkage.

Technical Details

The model utilizes a five-letter probabilistic genome representation where each position is a distribution over A, C, G, T, and 5mC (represented as M). Pretraining employed a soft-target masked language modeling objective using data from 242 individuals in the 1000 Genomes Long-Read Sequencing Consortium to encode CpG methylation levels and population allele frequencies. ModBERT also introduces probabilistic fine-tuning, allowing the representation to encode the specific genotype and methylation of a single individual for person-specific predictions from a single assay.

Background & Context

Most existing DNA large language models, including Nucleotide Transformer and DNABERT-2, are pretrained almost exclusively on reference genomes. Consequently, they fail to represent DNA methylation and genetic variants, which are the two primary sources of variation tied to regulatory differences between individuals. While long-read sequencing from Nanopore and PacBio can measure both factors from the same native DNA molecules, ModBERT is the first model designed to utilize this joint readout.

Significance & Outlook

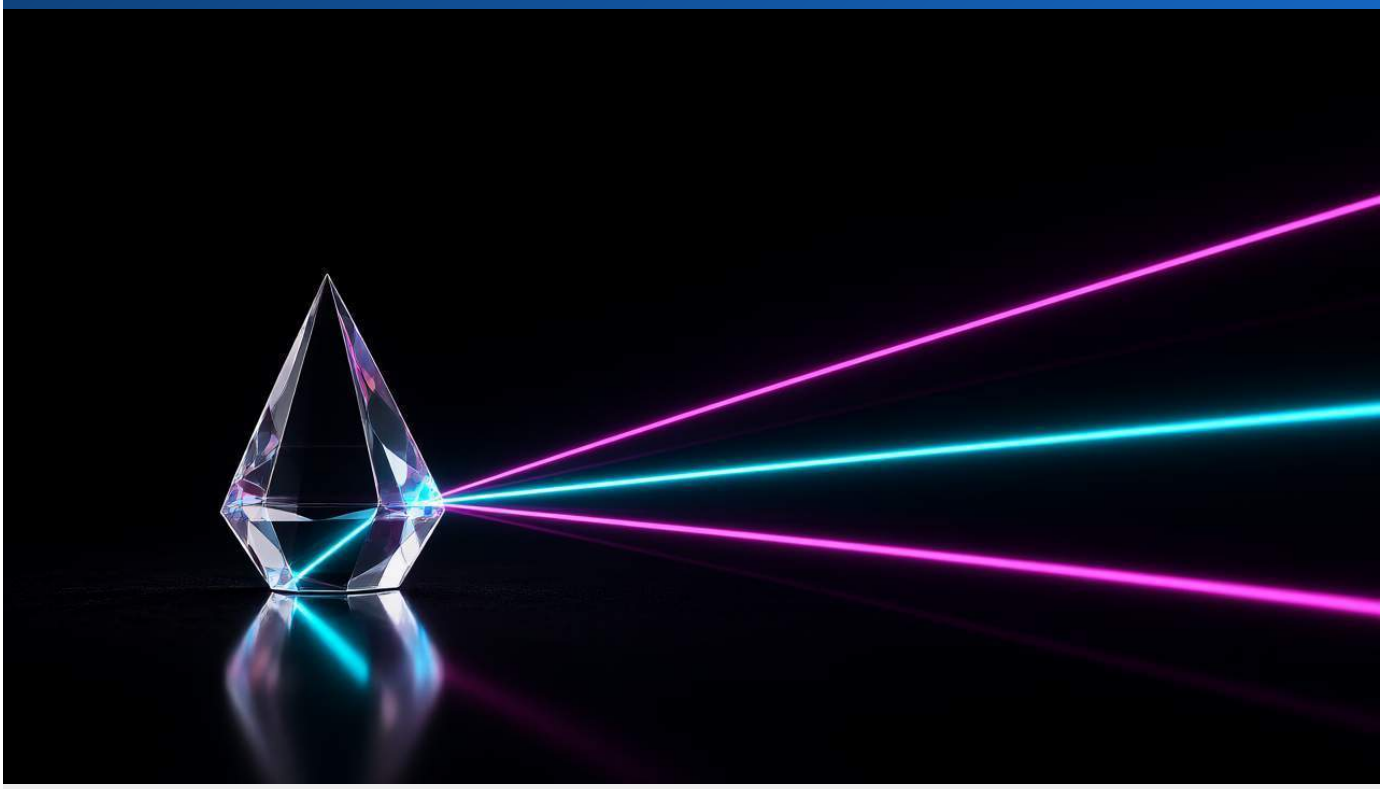
ModBERT establishes a framework for personal-genome language modeling by reading an individual's specific genetic and epigenetic sequence rather than a shared reference. This approach allows for the in silico dissection of methylation and variant effects on gene regulation. The authors note that this provides a principled method for prioritizing candidates for experimental validation as population-scale long-read cohorts, such as the Human Pangenome Reference Consortium and 1KGP, continue to grow.

Source: https://doi.org/10.64898/2026.11.12.766258

Collected: November 21, 2026 | Automated Research System (AI)

#04 Gemini 4 Neon: Specs, Benchmarks, and Pricing for Google AI

Published November 18, 2026　Google（The Keyword）　USA



OVERVIEW

Google has introduced Gemini 4 Neon, a model designed for fast end-to-end task completion in software engineering, science, and knowledge work. It achieves #1 rankings in several specialized benchmarks, including the Artificial Analysis Coding Agent Index and Legal Research Bench, while offering a 1 million-token context window.

IN DEPTH

Key Findings

Gemini 4 Neon ranks #3 overall on the Artificial Analysis Intelligence Index with a score of 61 at its xhigh thinking level and #2 on the Vals Index. It achieves #1 rankings on the Legal Research Bench, Tax Agent Bench, and Zapier's AutomationBench (56.3%). In scientific reasoning, it leads in Terminal-Bench Science (70.5%), RiemannBench (84.0%), and HLE-Diamond (71.3%).

Technical Details

The model features a 1 million-token context window and is priced at $3 per million input tokens ($0.15 per million cached input tokens) and $15 per million output tokens. Neon utilizes four scalable thinking levels to balance intelligence and token expenditure. At its default high thinking level, it costs 32% less per task than Carbon and 50% to 80% less than top competing frontier models on the Vals index and Agent Arena.

For coding and agentic workflows, Neon ranks #1 on the Artificial Analysis Coding Agent Index (58.6). In Google Antigravity, it achieves a state-of-the-art 39.0% on SWE-sweep. It is engineered to act as an Orchestrator directing parallel Hydrogen subagents, handling multi-file debugging and context compaction in long sessions.

Background & Context

Neon is designed to synthesize inputs across spreadsheets, PDFs, charts, audio, and video. It recorded its lowest fabrication rate to date in internal Google Antigravity testing and ranks #2 on steerability via the Agent Arena Index. Demonstrated capabilities in Google Antigravity include building a 3D lab bench via /swarm, creating a WebGL2 ski simulator with 10,200 lines of custom shaders, and converting a single image into a 100,000-cell Google Sheet using 500,000 formulas.

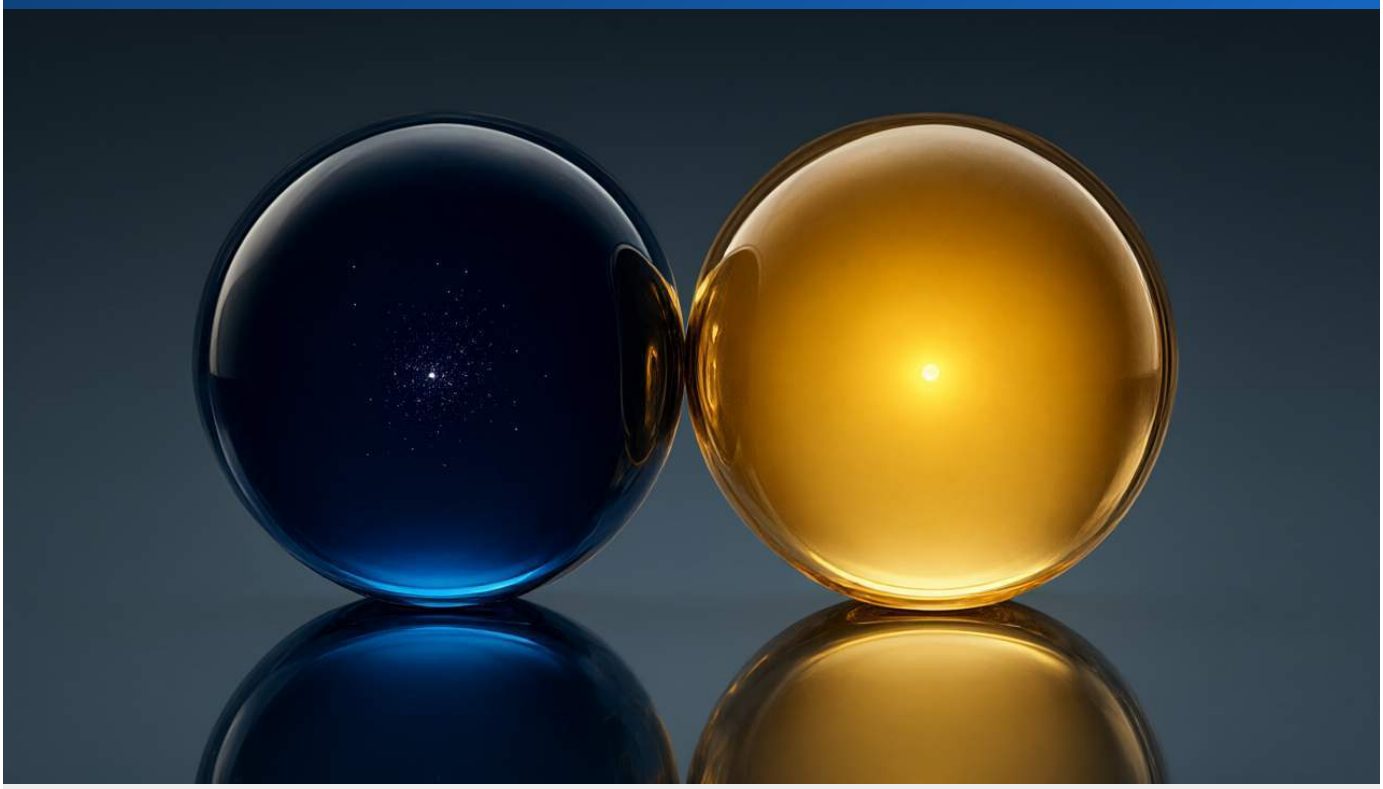
Significance & Outlook

Gemini 4 Neon is now available via the Gemini API in Google AI Studio, Gemini for business, and for Pro and Ultra subscribers in the Gemini app and Google Antigravity. By pairing frontier-level intelligence with reduced cost per task, the model targets production-ready workflows in enterprise and scientific sectors.

Source: https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-4-neon/

#05 GPT-6.2 Astra and Sol: OpenAI's Professional Model Specs

Published November 17, 2026 OpenAI USA



OVERVIEW

OpenAI has released GPT-6.2 Astra for complex professional projects and GPT-6.2 Sol for versatile everyday work. Astra achieves higher benchmarks in scientific and legal tasks at significantly lower costs than previous models, while Sol provides a high-efficiency alternative for agentic workflows.

IN DEPTH

Key Findings

GPT-6.2 Astra sets a new performance standard for professional work, delivering higher quality outputs with fewer tokens. On the Agents' Last Exam, Astra scored 5.2 points higher than GPT-6 Astra at less than one-third of the cost, and 5.5 points ahead of Claude Fable 5.2 at 5% of the cost. In the API, Astra's input and output token prices are 20% lower than GPT-6 Astra, with cached input priced at $0.40 per million tokens, input at $8, and output at $40.

Technical Details

Astra demonstrates specialized capabilities across several technical domains:

- **Science:** Astra scored 9 percentage points higher than GPT-6 Astra on Terminal-Bench-Science 0.1 and set a new high on the Vals Science Index.
- **Law and Finance:** On the Harvey LAB-AA v1.1 benchmark, Astra more than doubled the pass rate of Claude Fable 5.2 at one-sixth the cost. It also answered 62.5% of questions correctly on the AA-AnalystAgent test using California Medicaid expenditure reports.
- **Engineering:** The model supports precise computer use in CAD and simulation tools, capable of creating 3D models and Python solvers to simulate chassis dynamics in Blender.

GPT-6.2 Sol serves as the efficiency-focused model. In Cognition's FrontierCode 2.0, Sol came within 0.5 points of Astra at approximately one-fifth of the cost. On the LifeVal evaluation for everyday tasks, Sol outscored GPT-6 Astra at about one-seventh of the cost.

Background & Context

The models are designed to follow industry-specific templates and conventions in law, sales, finance, and accounting more closely than previous iterations. Internal testing via the GDPval v2 benchmark shows Astra is up 6.2 percentage points from GPT-6 Astra while using 22% fewer output tokens; Sol improved 5.3 points over GPT-6.1 Sol.

Significance & Outlook

The release focuses on reducing the cost of frontier performance at scale for businesses and developers. By improving visual reasoning for web design and providing higher precision in mechanical assembly and CAD, these models extend AI utility into physical engineering and professional document production.

Source: https://openai.com/index/introducing-gpt-6-2-astra-sol/

Collected: November 21, 2026 | Automated Research System (AI)

#06 Microsoft AI: Seven new models and Maia 200 specs for 2026

Published November 17, 2026 Microsoft AI USA



OVERVIEW

Microsoft AI has released seven new models focused on price-performance, including MAI Cyber 2 Flash and MAI Code 2. These models significantly reduce costs compared to GPT-6 and Claude Opus 5.5 while maintaining high quality for enterprise coding, security, and image editing.

IN DEPTH

Key Findings

Microsoft AI has introduced a family of seven new models designed to lower the cost of running agents 24/7. A primary highlight is MAI Cyber 2 Flash, which outperforms GPT-6 Astra on the CyberGym-E2E benchmark at less than 2% of the cost per task. Additionally, the Decision 1.1 model is 50x faster and 79x cheaper than GPT-6.1 Sol.

Technical Details

- **Coding:** MAI Code 2 achieves 98% of the quality of GPT-6.1 Sol at 40% of the price, and 96% of Claude Opus 5.5's performance at 5% of the price. MAI Code 2 Flash performs comparably to GPT-6 Luna and Claude Haiku 5.5 at 65% and 20% of the price, respectively.
- **Image and Audio:** MAI Image 2.7 ranks #1 on Artificial Analysis for image editing, beating GPT Image 2.6 Sunburst (max) at one-fifth of the cost. MAI Realtime 1, a speech-to-speech model, starts at $0.01 per audio minute.
- **Transcription and Decision:** Dragon Copilot Transcribe, based on MAI Transcribe 2, is used in 10 million monthly patient encounters. Decision 1.1 provides model routing to optimize the use of larger models.
- **Hardware:** The Maia 200 AI accelerator is now in production, delivering 1.7x more tokens per watt.

Background & Context

The release addresses the rising cost of token spend for enterprises. MAI Cyber 2 Flash is specifically designed for 24/7 monitoring and patching of production systems, with output token pricing 86% lower than MAI Cyber 1.1 Flash. MAI Code 2 is optimized for everyday enterprise tasks seen in GitHub Copilot, though it is not yet tuned for fundamental science research or long-horizon agentic work.

Significance & Outlook

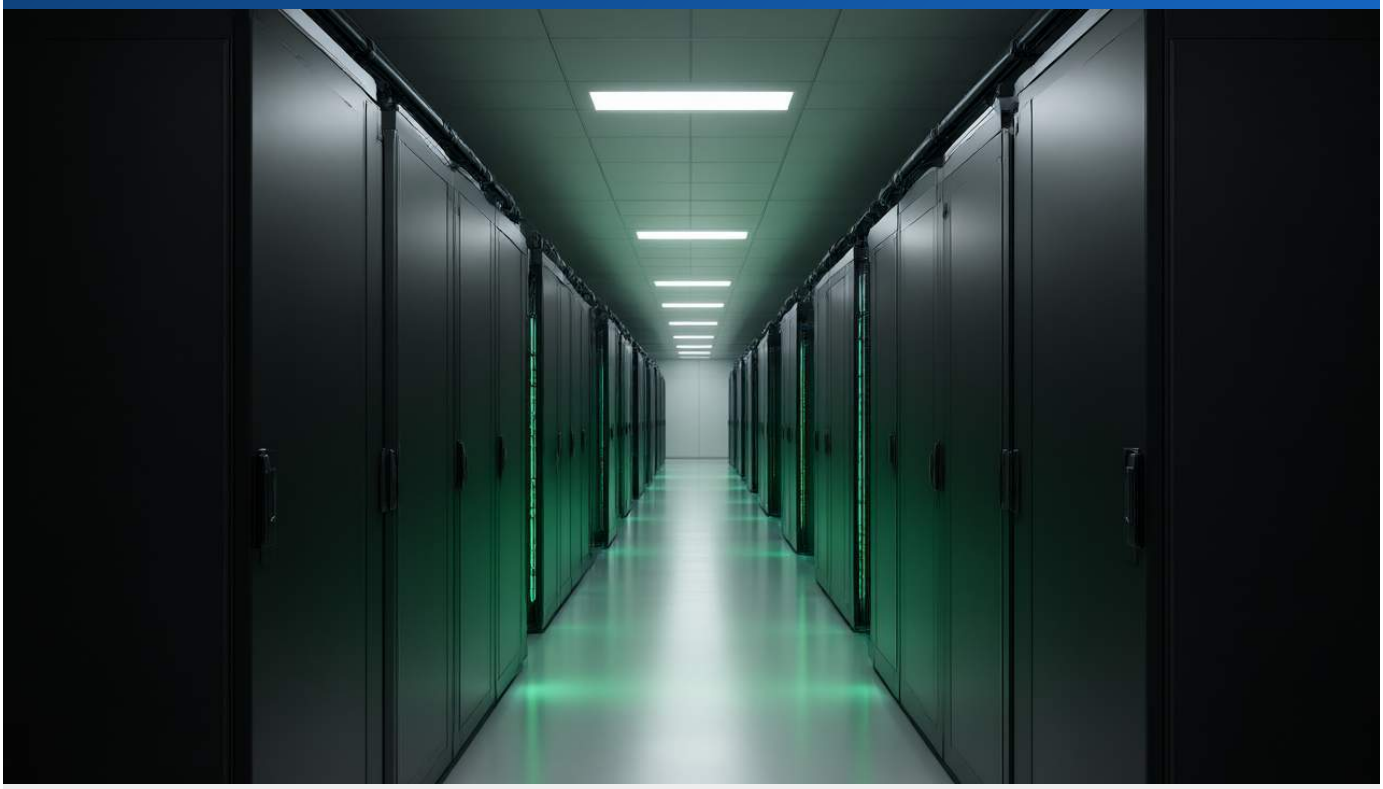
These developments enable high-volume deployment of frontier intelligence at a fraction of previous costs. For example, AT&T developed a custom MAI model for 5G network operation analytics that outperformed GPT-6.1 Sol at 1/50th of the cost. This shift allows defenders to maintain security in an era of cheap offensive tools.

Source: https://microsoft.ai/news/microsoft-ignite-2026/

Collected: November 21, 2026 | Automated Research System (AI)

#07 NVIDIA Vera Rubin: MLPerf Training v6.1 and Endpoints Results

Published November 16, 2026 NVIDIA Blog USA



OVERVIEW

NVIDIA's Vera Rubin platform achieved leading performance across MLPerf Training v6.1 and MLPerf Endpoints benchmarks, including a 10x generational gain in throughput per watt for agentic AI. These results demonstrate significant scaling efficiency and power optimization for training and agentic inference workloads.

IN DEPTH

Key Findings

NVIDIA swept every MLPerf Training v6.1 benchmark, becoming the only platform to provide results across all categories, including large language models, image generation, recommender systems, and reinforcement learning post-training. In its debut, the Vera Rubin NVL72 delivered up to 2x more performance per GPU than the NVIDIA GB300 NVL72 system. Additionally, Vera Rubin led all agentic benchmarks in MLPerf Endpoints; specifically, for the Kimi K3 model, Vera Rubin NVL72 systems achieved more than 10x the throughput per watt of the GB300 NVL72 at the same response speed.

Technical Details

NVIDIA demonstrated high scaling efficiency on the DeepSeek-V3 benchmark, where the Vera Rubin NVL72 scaled from 256 to 1,024 GPUs with approximately 97% efficiency, reducing training time from 16 minutes to about four. The company also executed benchmarks at a larger scale using 10,240 GPUs with GB300 systems, a scale matched by Microsoft Azure using 10,240 NVIDIA GB200 GPUs. This scaling is supported by NVIDIA NVLink for intra-system connectivity, and NVIDIA Quantum-X800 InfiniBand and Spectrum-X Ethernet for data center-wide scale-out networking.

Background and Context

MLPerf Training evaluates the speed of training AI models to a target accuracy, while MLPerf Endpoints tests systems on real agentic coding and enterprise workflows. The benchmarks reflect the increasing demand for AI agents that perform complex reasoning and tool-calling. Beyond MLPerf, Vera Rubin showed leading performance on independent benchmarks, including SemiAnalysis AgentX and Artificial Analysis AgentPerf.

Significance and Outlook

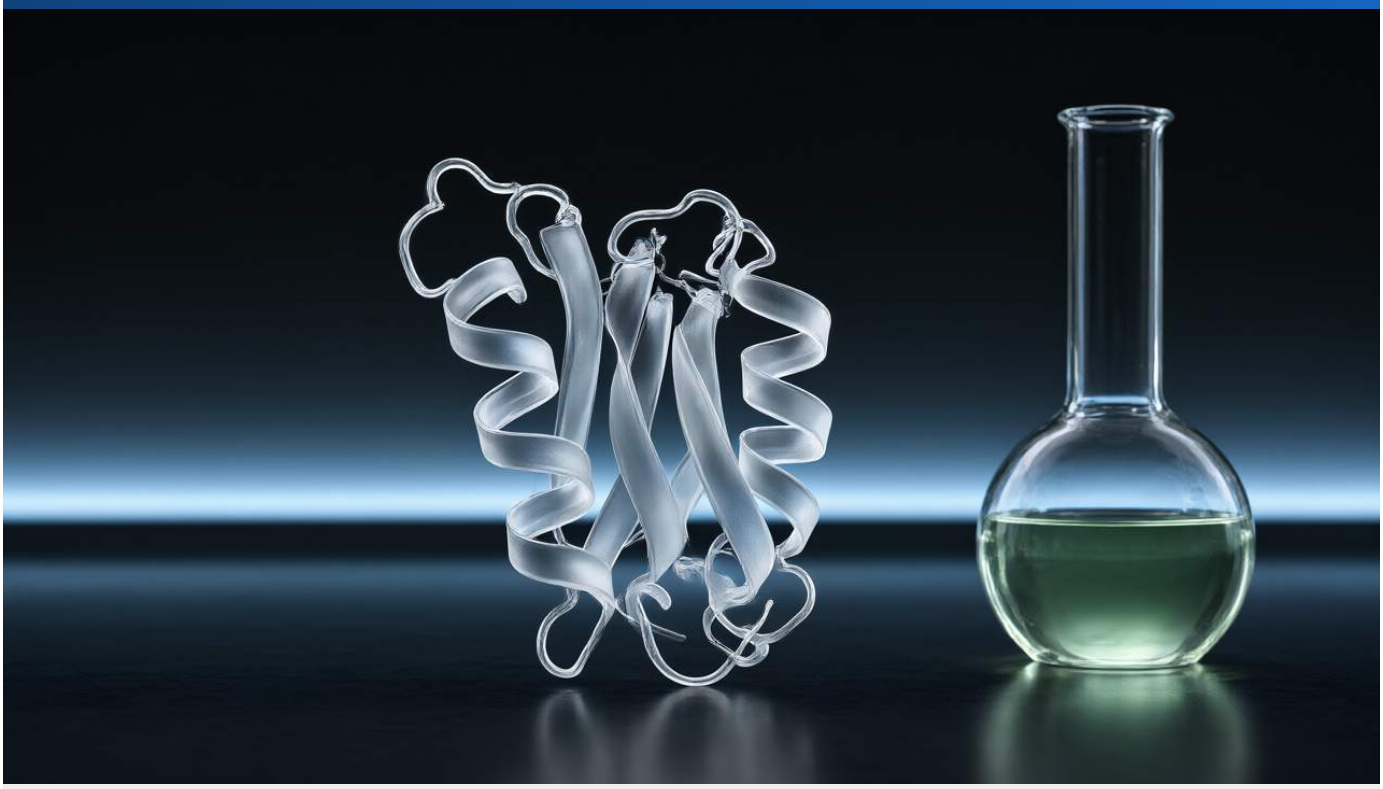
The results indicate that Vera Rubin improves the balance between serving a high volume of users and maintaining fast response times. NVIDIA notes that performance is expected to increase further through the continuous optimization of the CUDA-powered software stack. The ecosystem's breadth was highlighted by the participation of 20 organizations in MLPerf Training, including Azure, Google Cloud, Oracle, and Dell Technologies.

Source: https://blogs.nvidia.com/blog/mlperf-training-endpoints-vera-rubin/

Collected: November 21, 2026 | Automated Research System (AI)

#08 NVIDIA Nemotron for Science: Specialized Agents and LitmusBench v0.2

Published November 16, 2026 NVIDIA Newsroom USA



OVERVIEW

NVIDIA has released Nemotron for Science, combining the Nemotron 3.5 Super VL model with the LitmusBench v0.2 chemistry dataset to enable domain-specific scientific agents. Early adopters including Argonne National Laboratory and various life sciences firms are using the framework to accelerate drug discovery and materials synthesis.

IN DEPTH

Key Findings

NVIDIA Nemotron for Science provides researchers with the Nemotron 3.5 Super VL open model, post-training recipes, and the LitmusBench v0.2 dataset to build specialized AI agents. These agents are designed to reason through scientific problems and utilize CUDA-X accelerated tools for computation. Early implementations show significant performance gains: Lila Sciences achieved 61% accuracy in making quantum dots—surpassing GPT-6 Astra's 54%—at 10-20x lower inference cost.

Technical Details

The framework centers on the Nemotron 3.5 Super VL model, which can be post-trained with domain-specific data to improve performance on specialized tasks. A critical component is LitmusBench v0.2, an open dataset containing nearly 50,000 questions focused on chemistry reasoning, molecular properties, and experimental drug discovery. To support deployment, the model is optimized for local execution on 128GB NVIDIA DGX Spark systems, while DGX Spark and data center infrastructure allow scaling to supercomputing centers.

Background & Context

Several institutions and companies have already integrated the technology into their workflows:

- **Terray Therapeutics:** Post-trained the model to extract reactants and products from scientific diagrams, recovering approximately 76% of annotated reactions on the RxnScribe test.
- **Therna:** Collaborated with NVIDIA to design agentic mRNA therapeutics, nearly doubling accuracy in predicting protein-expression ranges compared to frontier models.
- **Argonne National Laboratory:** Integrating Nemotron into the StructBioReasoner multi-agent system for protein-based therapeutic candidates.
- **KTH Royal Institute of Technology:** Developed an agent that independently repaired 81% of broken GROMACS molecular dynamics setups in a single overnight run.

- **RIKEN:** Made Nemotron 3.5 Super VL available via the RiVault inference gateway on the RIKYU AI supercomputing platform.

Significance & Outlook

By providing open models and scientific data, NVIDIA aims to transform how researchers tackle complex challenges in cancer treatment, national security, and critical minerals. The ability to translate desired material properties into synthesis recipes and automate laboratory protocols reduces the gap between theoretical ideas and physical discoveries.

Source: https://nvidianews.nvidia.com/news/nvidia-introduces-nemotron-for-science-to-accelerate-discoveries-with-agents

Collected: November 21, 2026 | Automated Research System (AI)

#09 NVIDIA, IBM, and Fermilab: Quantum-Classical Codesign Platform

Published November 16, 2026 NVIDIA Newsroom USA



OVERVIEW

NVIDIA, IBM, and Fermi National Accelerator Laboratory have launched a modular quantum-classical codesign platform to advance error-correction and scientific applications. The initiative includes a new decoding algorithm, TCBP-JOBS, which accelerates quantum error-correction research by up to 13,000x. These developments aim to scale quantum computers for practical use in high-energy and nuclear physics.

IN DEPTH

Key Findings

NVIDIA has partnered with IBM and the U.S. Department of Energy's Fermi National Accelerator Laboratory (Fermilab) to develop a modular quantum computing codesign platform. A primary technical breakthrough is the release of TCBP-JOBS, a decoding algorithm within the CUDA-Q platform that delivers up to a 13,000x speedup for quantum error-correction research. This reduces the computation time for researching new quantum designs from 12 weeks to 45 minutes.

Technical Details

The codesign platform integrates NVIDIA DGX B200 systems with IBM quantum computers, including those powered by the IBM Quantum Nighthawk processor. Key technical integrations include:

- **Error Correction:** The TCBP-JOBS algorithm, available via the NVIDIA CUDA-Q QEC 0.9.0 library, accelerates the computationally intensive parts of decoding to identify and correct errors across physical qubits.
- **Workflow Integration:** NVIDIA is expanding Slurm to simplify the scheduling of quantum processors within supercomputing workflows, utilizing IBM's Quantum Resource Management Interface.
- **Modular Architecture:** The platform allows researchers to evaluate how classical systems scale alongside multiple linked quantum processors.

Background & Context

Quantum processors are currently limited by inherent errors that restrict computation complexity and duration. To address this, the collaboration leverages Fermilab's expertise in high-energy and nuclear physics (HENP) to define the requirements for future fault-tolerant quantum computers. Additionally, IBM and RIKEN are utilizing NVIDIA accelerated computing to advance quantum-classical chemistry research via RIKEN's JHPC-quantum GPU supercomputer.

Significance & Outlook

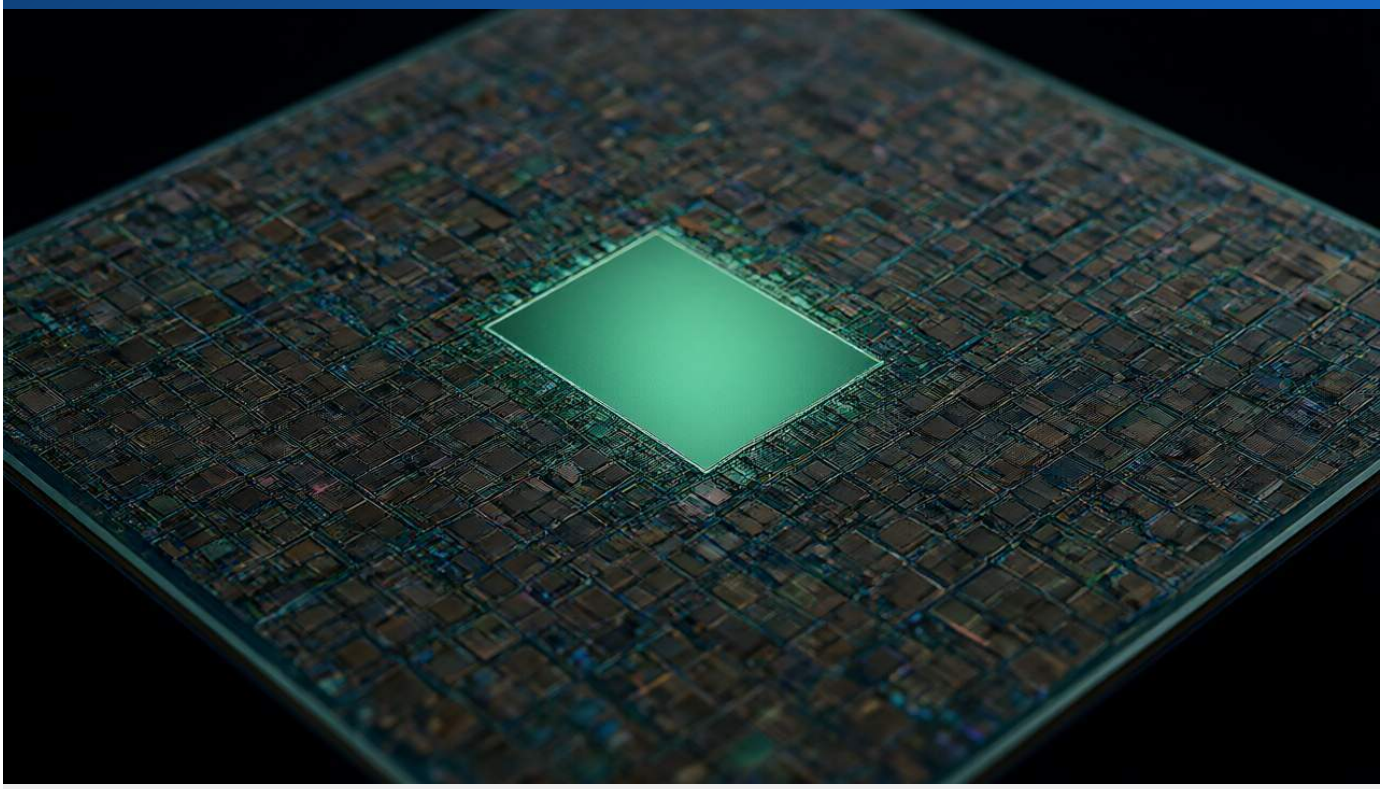
By combining accelerated computing with quantum hardware, the partners aim to move beyond the limitations of quantum hardware alone. The ability to rapidly evaluate error-correction codes and decoding algorithms shortens the path to practical fault-tolerant quantum computing. This infrastructure is designed to solve fundamental scientific problems in HENP and other fields that remain unreachable via classical computing alone.

Source: https://nvidianews.nvidia.com/news/nvidia-announces-collaboration-with-ibm-and-fermi-national-accelerator-laboratory-to-launch-quantum-classical-codesign-platform-plus-error-correction-breakthrough-to-advance-practical-quantum-computing

Collected: November 21, 2026 | Automated Research System (AI)

#10 OpenAI Chip Design: PPA Gains for Jalapeño and AMD Hardware

Published November 18, 2026 OpenAI USA



OVERVIEW

OpenAI models have demonstrated the ability to improve Power, Performance, and Area (PPA) in production chip designs, including a 13% area reduction in Jalapeño's vector unit. Collaborations with AMD achieved up to 14% PPA improvement on flagship GPU IP, compressing roughly a year of progress into less than five months. These results indicate that frontier models can reason at the level of expert microarchitects to optimize SystemVerilog RTL.

IN DEPTH

Key Findings

OpenAI models have delivered significant PPA improvements across both internal and third-party hardware. In the development of OpenAI's first custom inference chip, Jalapeño, model-proposed changes reduced standard-cell area by approximately 13% in the vector unit and 10% in the matrix unit. These optimizations were silicon proven during bring-up. Furthermore, a partnership with AMD resulted in up to 14% PPA improvement relative to heavily optimized baselines on flagship GPU IP, as well as single-digit percentage gains across key blocks of AMD's upcoming next-generation flagship CPU.

Technical Details

The models achieved these gains by rewriting SystemVerilog RTL and exploring alternative logic implementations. Specific improvements in the Jalapeño design include:

- **Vector Unit:** Optimizations in floating-point adders, multipliers, conversion functions, reduction logic, and the special-function unit.
- **Matrix Unit:** A multiplier redesign that reduced cell area by about 15% and latency from two clock cycles to one, and an adder redesign that reduced cell area by about 15% while maintaining two-cycle latency.

To verify these changes, OpenAI integrated frontier intelligence with Synopsys electronic design automation (EDA) tools to ensure that rewrites preserved block behavior and to evaluate trade-offs in power, performance, and area.

Background & Context

To measure capability progression, OpenAI tested models on two open-source 32-bit RISC-V cores: PicoRV32 and Ibex (used in OpenTitan). Models were tasked with increasing throughput (instructions per second), measured as the geometric mean of instructions per clock cycle multiplied by the maximum clock frequency estimated with Yosys. GPT-6.2 Astra achieved the highest average score on both cores, followed by GPT-6.1 Sol and GPT-5.6 Sol.

Significance & Outlook

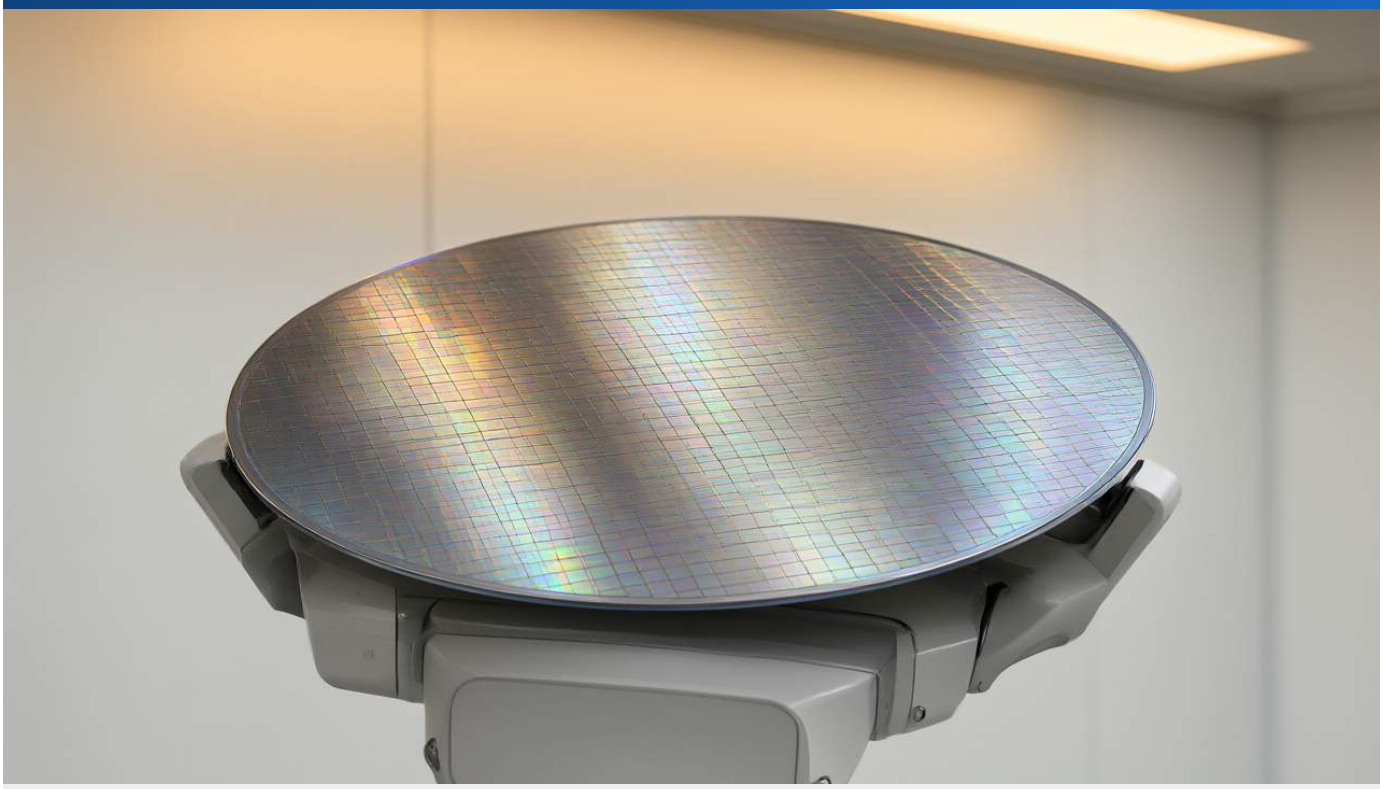
Evaluation data shows that models perform better when designing a core from written requirements than when optimizing existing designs. This suggests that starting from specifications allows models to omit non-essential features, such as performance counters, to prioritize throughput. These results demonstrate that AI models can now perform expert-level microarchitectural reasoning in production environments.

Source: https://openai.com/index/advancing-chip-design-with-our-models/

Collected: November 21, 2026 | Automated Research System (AI)

#11 OpenAI and AMD: AI-driven 14% PPA gains in GPU designs

Published November 18, 2026 AI Weekly USA



OVERVIEW

OpenAI's models achieved up to 14% gains in power, performance, and area (PPA) on AMD's flagship GPU and single-digit gains on next-generation CPU blocks. These AI-authored RTL edits have already been integrated into designs sent for manufacturing, signaling a shift toward AI-automated chip design.

IN DEPTH

Key Findings

OpenAI has demonstrated that its models can identify PPA (power, performance, and area) improvements of up to 14% in selected circuits for AMD's flagship GPU. These gains were measured against baselines that AMD had already heavily optimized. OpenAI characterized this result as achieving roughly a year of progress in less than five months.

Technical Details

The AI-driven optimizations were applied across several hardware projects:

- **AMD GPU:** Model-written edits were folded into designs sent for manufacturing, with others added to production files for future versions.
- **AMD CPU:** OpenAI's models proposed edits that resulted in verified single-digit percentage PPA gains across key blocks of a next-generation flagship CPU.
- **OpenAI Jalapeño:** On its own custom inference chip, OpenAI's models reduced standard-cell area by approximately 13% in the vector unit and 10% in the matrix unit. AI assistance enabled the transition from initial design to tapeout in nine months.

The technical workflow involved models rewriting circuits in SystemVerilog. Synopsys design tools were then used to verify that the rewrites maintained the original behavior and to measure the resulting tradeoffs.

Background & Context

As part of a multi-year agreement, OpenAI is partnering with Synopsys to develop "GPT-Synopsys." This model is designed to drive Synopsys' electronic design automation (EDA) tools in a manner similar to expert chip designers. The agreement encompasses joint development, revenue sharing, and go-to-market strategies, with the intention of turning this internal workflow into a commercial product.

Significance & Outlook

While the results are significant, OpenAI noted a caveat: all five tested models performed better when designing from written requirements than when improving existing code, as models could omit features present in original designs. Despite this, the integration of OpenAI-authored RTL into AMD designs sent to a fab marks a milestone in AI-assisted silicon engineering.

Source: https://aiweekly.co/alerts/openai-says-models-found-14-ppa-gains-in-amd-gpu-designs