NEW HORIZON · AI DIGEST

The Long Read

13 August 2026

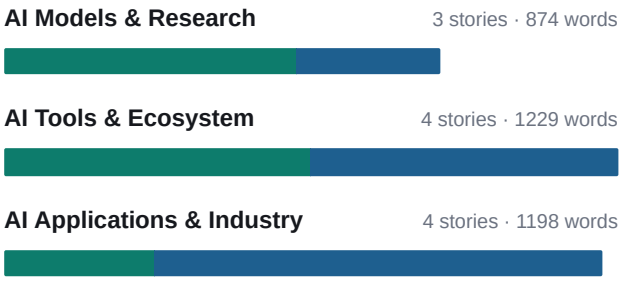
11	3	3301	5
STORIES	SECTIONS	WORDS	HIGH IMPACT

5 high impact · 6 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 · 3301 words · 3 sections · 4 charted



AI Models & Research

01 DeepSeek V4 Pro 0813: Flagship 1.6T MoE Model Exits Preview

02 Liquid AI Ships LFM2.5-VL-3B: Vision-Language Model That Runs on Your Phone

03 AI4AI at Test-Time: Can Strong Models Transfer Skills to Weak Ones Without Training?

AI Tools & Ecosystem

04 195TB of Credentials Leaked in LiteLLM Supply-Chain Attack Impacting 2,500 Organizations

05 Made by Google 2026: Pixel 11, Gemini Intelligence, and Sign Language AI

06 Hinton, Ng, and Fei-Fei Li Unite on Openness -- With Caveats

07 Amazon Will Train AI on Twitch Streamers' Content by Default Unless They Opt Out

AI Applications & Industry

08 Cognition AI in Talks to Raise at $40B Valuation as Devin Nears $1B ARR

09 Lovable Raises $400M at $13.3B Valuation -- Vibe-Coding Startup Doubles in Eight Months

10 Thrive Holdings Raises $2B at $12B to Bring OpenAI-Powered Tools to Accounting and IT

11 Blacksmith Hits $550M Valuation: AI Code-Testing Startup Grows 10x in Under a Year

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

3 stories · 2 high · 1 medium

01 MEDIUM IMPACT openrouter.ai

DeepSeek V4 Pro 0813: Flagship 1.6T MoE Model Exits Preview

> DeepSeek's 1.6T parameter Mixture-of-Experts model, V4 Pro 0813, has moved from preview to general availability via OpenRouter.

The DeepSeek V4 Pro 0813 model is now accessible through OpenRouter, marking its exit from the preview phase. This flagship architecture utilizes a 1.6 trillion parameter Mixture-of-Experts (MoE) design. Unlike multi-provider endpoints where traffic is load-balanced across several backends, this specific model is hosted by a single provider. OpenRouter forwards every request directly to this host without making internal routing decisions, simplifying the inference path but removing automatic failover to alternative providers for this specific slug.

Integration relies on OpenRouter's standard OpenAI-compatible API, requiring only a base URL swap and the specific model slug to deploy. While listed prices exist, the actual cost per token varies based on caching efficiency and provider-specific discounts, often resulting in lower effective rates than the headline figures. Performance metrics tracked for this endpoint include throughput measured in tokens per second, total round-trip latency, and time-to-first-token (TTFT). Uptime data reflects the success rate of requests over the preceding 30 days, though the single-provider configuration means retries on next-best providers do not apply if the primary host fails.

Usage signals indicate growing adoption in public applications, with token volume and request traffic trends available to gauge production suitability. The model's benchmark scores place it within a specific percentile rank among all models currently hosted on the platform, offering a comparative view of its capabilities against the broader ecosystem. For engineers evaluating deployment, the key distinction lies in the direct forwarding mechanism; there is no intermediary routing logic to optimize for cost or speed dynamically, placing the reliability burden entirely on the singular hosting provider while maintaining the convenience of a unified API interface.

KEY FACTS

MODEL NAME **DeepSeek V4 Pro 0813**	ARCHITECTURE **1.6T MoE**
HOSTING CONFIGURATION **Single provider, direct forwarding**	API COMPATIBILITY **OpenAI-compatible**

WHY IT MATTERS

Engineers can now integrate the 1.6T MoE architecture into production pipelines using standard OpenAI SDKs, but must account for the lack of automatic provider failover inherent in this single-host deployment.

Read the original at openrouter.ai →

02 HIGH IMPACT huggingface.co

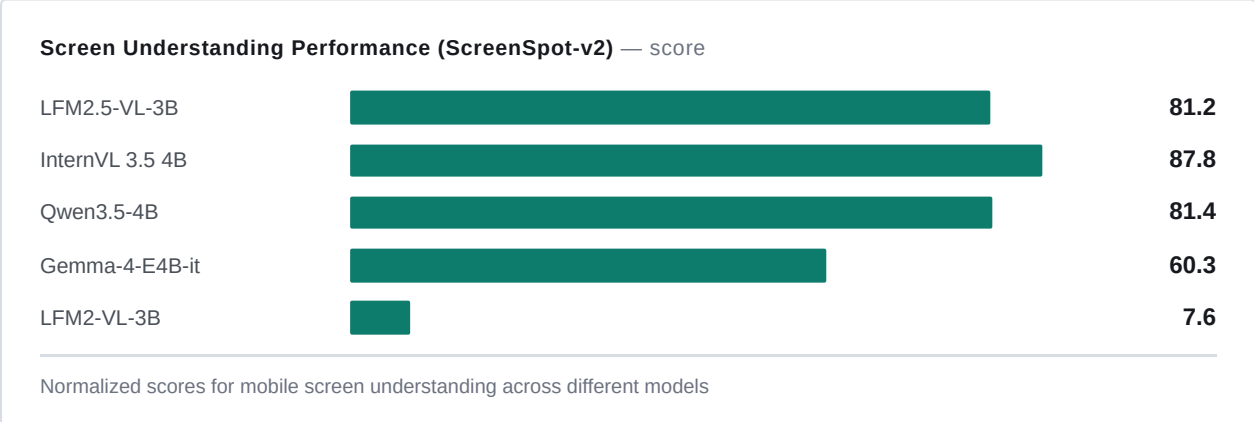
Liquid AI Ships LFM2.5-VL-3B: Vision-Language Model That Runs on Your Phone

> Liquid AI has released LFM2.5-VL-3B, a 3.1-billion-parameter vision-language model optimized for on-device inference that outperforms larger competitors in screen understanding and grounding tasks.

The new model pairs a SigLIP2 400M NaFlex vision encoder with the LFM2.5-2.6B text backbone, pre-trained on approximately 34 trillion tokens including four times more vision data than previous iterations. To support non-Latin scripts, Liquid AI extended the tokenizer in place, doubling the vocabulary to 128,000 without retraining from scratch. Post-training involves supervised fine-tuning with knowledge distillation and Antidoom training, followed by multi-reward reinforcement learning.

Benchmark results position LFM2.5-VL-3B as a leader in its size class for real-world image tasks. On ScreenSpot-v2 Mobile, the model scores 81.2, drastically surpassing the prior LFM2-VL-3B score of 7.6 and exceeding the 69.7 achieved by Qwen3.5-2B. It also leads in grounding with an 87.9 average on RefCOCO, compared to 57.1 for its predecessor. In text-only evaluations, instruction following on IFEval reaches 82.3, while tool use on ToolSandbox hits 59.5, placing it on par with Gemma-4-E2B-it and Qwen3.5-2B despite having fewer parameters.

Inference performance targets edge deployment across CPU and GPU environments. The model fits within 3 GB of memory and decodes at 228 tokens per second on an M5 Max, 116 tokens per second on a Ryzen AI Max+ 395, and maintains 20 tokens per second on a Galaxy S26 Ultra. For server-side workloads, it achieves roughly 11,000 tokens per second output throughput at high concurrency on an H100, approximately double the rate of 4B-class models tested. Day-one support includes llama.cpp, MLX, vLLM, SGLang, and ONNX.



KEY FACTS

PARAMETER COUNT **3.1B**	VISION ENCODER **SigLIP2 400M NaFlex**
PRE-TRAINING TOKENS **34T**	VOCABULARY SIZE **128K**
MEMORY FOOTPRINT **3 GB**	MOBILE THROUGHPUT **20 tokens/s**

> WHY IT MATTERS
>
> Practitioners can now deploy capable vision-language agents directly on mobile devices or high-volume CPU clusters without sacrificing grounding accuracy or tool-use reliability. The specific optimization for UI understanding enables new classes of automated testing and accessibility tools running entirely offline.

Read the original at huggingface.co →

---

03 HIGH IMPACT arXiv.org

AI4AI at Test-Time: Can Strong Models Transfer Skills to Weak Ones Without Training?

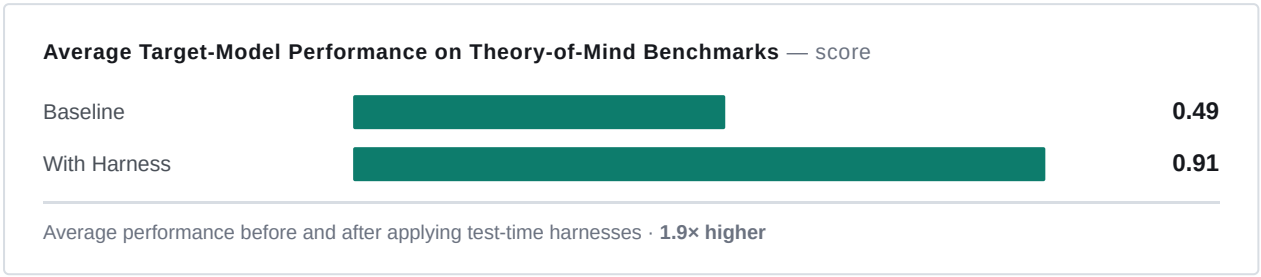
> Strong builder models can nearly double the performance of weaker target models at inference time by constructing deterministic code harnesses rather than updating parameters.

The paper "AI4AI at Test-Time: Strong-to-Weak Capability Transfer via Harnesses," submitted to arXiv on 12 Aug 2026, proposes a shift from training-time distillation to test-time scaffolding. Instead of using teacher forcing or on-policy distillation to update a small model's weights, the authors employ a stronger builder model to generate inference-time harnesses. These harnesses act as external cognitive structures that guide the weaker target model through tasks without any modification to its underlying parameters. The methodology involves the builder model utilizing 5% of the available data as a validation set to iteratively refine the harness over multiple rounds before final evaluation on the full test set.

Empirical results across four representative Theory-of-Mind benchmarks demonstrate a substantial performance increase. The average score for target models rose from 0.49 to 0.91 when equipped with these generated harnesses. The analysis attributes this gain not to expanded reasoning depth or broader sampling by the target model, but to the offloading of unstable reasoning steps into deterministic code. Key mechanisms include benchmark-specific routing logic and strict enforcement of answer formats. The study notes that the quality of the resulting harness improves monotonically with the reasoning effort expended by the builder model, while platform-specific effects remain modest relative to the builder's inherent capability.

This approach identifies weaker target models as the primary beneficiaries of strong-to-weak transfer via harnessing. The findings position inference-time harness design as a functional complement to conventional distillation techniques. By enabling the transfer of cognitive structure without retraining, this method offers a pathway to deploy smaller models with reliability profiles previously achievable only through parameter updates or larger model deployment. The work suggests that future system architectures may increasingly rely on dynamic, code-based scaffolding generated by capable models to stabilize outputs from efficient, smaller counterparts.

The distinction between this method and prior work lies in the locus of adaptation: previous distillation methods alter the student model's internal state, whereas this technique constructs an external execution environment. The reliance on deterministic code to handle reasoning instability marks a specific technical departure from prompting strategies that merely encourage the model to "think harder." The results hold across the tested benchmarks, indicating that the structural constraints provided by the harness are more effective than increasing the target model's computational breadth during inference.



KEY FACTS

SUBMISSION DATE
12 Aug 2026

PERFORMANCE BASELINE
0.49

PERFORMANCE WITH HARNESS
0.91

VALIDATION DATA USAGE
5%

BENCHMARK COUNT
4

BENCHMARK TYPE
Theory-of-Mind

WHY IT MATTERS

Engineers can now improve weak model reliability on specific tasks by generating code wrappers with strong models, bypassing the cost and latency of retraining or distillation pipelines. This enables immediate performance gains for deployed small models using only inference-time compute.

Read the original at arXiv.org →

SECTION 2 OF 3

AI Tools & Ecosystem

4 stories · 2 high · 2 medium

04 HIGH IMPACT Ars Technica

195TB of Credentials Leaked in LiteLLM Supply-Chain Attack Impacting 2,500 Organizations

> A 40-minute compromise of the LiteLLM Python package in March exposed 195TB of credentials from over 2,500 organizations, including Microsoft, Amazon, and Salesforce.

Security firms CloudSEK and Hudson Rock disclosed that a supply-chain attack on LiteLLM resulted in the exfiltration of 195TB of sensitive data. The breach occurred when victims downloaded compromised versions of the open-source AI tool from the official Python Package Index repository. During a 40-minute window in March, attackers extracted cloud keys, repository tokens, SSH keys, Kubernetes secrets, package publishing credentials, environment variables, and AI provider keys. The affected entities include major technology firms such as Microsoft, Amazon, Cisco, Samsung, and Salesforce.

The intrusion vector traces back to a prior supply-chain campaign that initially infected Trivy, a widely used vulnerability scanner. This initial compromise subsequently spread to other software components, including KICS and the Telnyx Python SDK, before reaching LiteLLM. Researchers attribute the operation to TeamPCP, a group described as a ramshackle but capable gang largely composed of teenagers. Independent security researcher Kevin Beaumont corroborated the validity of the leaked data, confirming it contains significant volumes of sensitive content from multiple victim organizations.

Beaumont characterized the incident as a massive supply-chain breach driven by poor AI security practices rather than an inherent threat from AI technology itself. He noted that the attackers exploited organizations obsessed with rushing AI deployment while maintaining poor DevOps security standards. The exposure encompasses a broad spectrum of access secrets, ranging from infrastructure-level SSH keys to specific AI provider credentials, potentially allowing attackers to gain unauthorized access to critical systems across more than 2,500 distinct organizations.

KEY FACTS

DATA VOLUME
195TB

ORGANIZATIONS IMPACTED
2500

EXPOSURE WINDOW
40 minutes

INCIDENT MONTH
March

ATTRIBUTED GROUP
TeamPCP

WHY IT MATTERS

Developers using LiteLLM or dependent on the compromised supply chain must immediately rotate all cloud keys, SSH keys, and AI provider credentials generated or accessed during the March window. The incident highlights the risk of transitive dependencies where vulnerabilities in tools like Trivy can cascade into AI orchestration layers.

Read the original at Ars Technica →

---

05 MEDIUM IMPACT TechCrunch

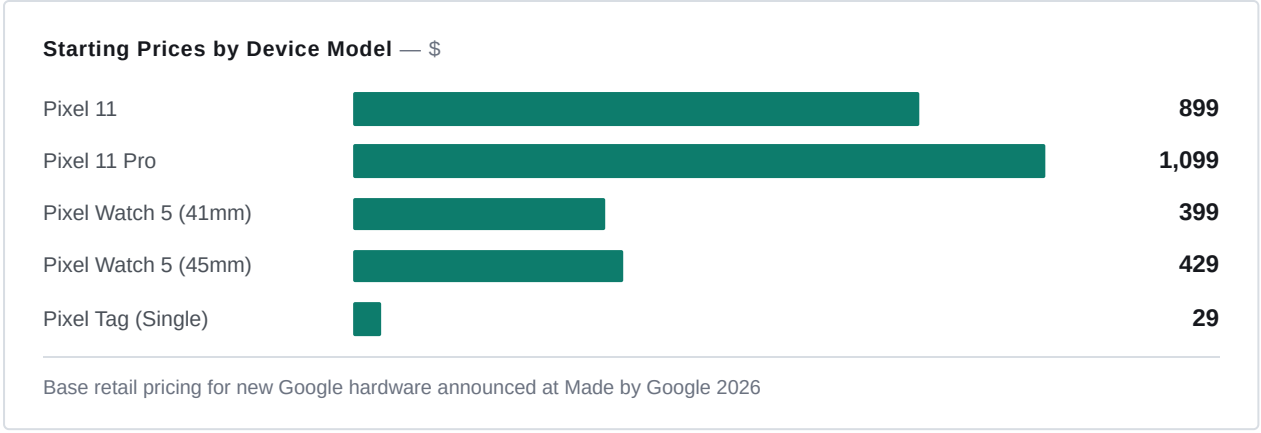
Made by Google 2026: Pixel 11, Gemini Intelligence, and Sign Language AI

> Google's Made by Google 2026 event delivered hardware upgrades centered on durability and storage while expanding Gemini's accessibility to include American Sign Language translation.

The Pixel 11 lineup introduces significant hardware shifts driven by supply constraints and design changes. The base Pixel 11 features a camera bar that is more than 40% thinner than the previous generation, utilizing an all-glass surface across the device width. Base storage doubles to 256GB, eliminating the 128GB option, which contributes to a starting price of $899, a $100 increase over the Pixel 10. The Pixel 11 Pro and Pro XL models prioritize physical resilience, boasting an anti-scratch display coating with more than twice the scratch resistance of the Pixel 10 Pro. These Pro models start at $1,099, up from $999. The Pixel 11 Pro Fold achieves a 10% weight reduction and is almost 1mm thinner, featuring a glass-fiber composite back and redesigned hinge that make it three times more durable than its predecessor. It includes a 48-megapixel main camera and 30x Super Zoom.

On the software front, Gemini integration expands beyond standard voice commands to address specific accessibility and interaction gaps. A major update extends Live Transcribe to support American Sign Language, using the Pixel Camera to translate signs directly into text. Google also introduced Rambler, a voice-input feature engineered to parse run-on sentences, filler words, and unstructured speech without requiring precise phrasing. Circle to Search functionality now integrates directly into the Pixel Camera interface, enabling object identification and translation without exiting the camera app. These updates shift Gemini from a reactive assistant to a more contextual layer within the device's core sensors.

Google officially entered the item-tracking market with Pixel Tag, a direct competitor to Apple's AirTag priced at $29 individually or $99 for a four-pack. The tags leverage Android's Find Hub network for location tracking and integrate with Pixel Watch and Pixel Buds for voice-controlled locating via Gemini. The Pixel Watch 5 adds longitudinal health metrics, providing monthly summaries of blood pressure and insulin resistance trends through the Google Health app. Pricing starts at $399 for the 41mm model and $429 for the 45mm version, with a Stephen Curry edition available for $579.



KEY FACTS

PIXEL 11 BASE STORAGE
256GB

PIXEL 11 STARTING PRICE
$899

CAMERA BAR THICKNESS REDUCTION
40%

PRO DISPLAY SCRATCH RESISTANCE
2x

FOLD DURABILITY INCREASE
3x

PIXEL TAG SINGLE UNIT PRICE
$29

WHY IT MATTERS

Developers building on Android must account for the new Rambler voice parsing capabilities and the expanded Camera API access for Circle to Search. The removal of the 128GB SKU on base devices alters storage assumptions for on-device model deployment.

Read the original at TechCrunch →

---

06 MEDIUM IMPACT TechCrunch

Hinton, Ng, and Fei-Fei Li Unite on Openness -- With Caveats

> Three AI pioneers converged at the Ai4 conference to argue against industry consolidation while diverging sharply on the specific risks of open-weight models.

At the Ai4 conference in Las Vegas last week, Geoffrey Hinton, Fei-Fei Li, and Andrew Ng addressed the tension between AI safety and the necessity of open distribution. While all three opposed a future where a handful of major labs control the pace of progress, their definitions of risk and solutions differed. Ng framed openness as a competitive imperative for the United States, warning that lobbying and fear-mongering could cede soft power to China if American firms fail to produce cost-efficient open models. He argued that if Chinese open-weight models gain adoption across Asia and Africa, they could shape global norms regarding democracy and human rights.

Hinton offered a stricter technical distinction between open source code and open-weight models. He noted that while inspecting code allows bug detection, releasing trained parameters enables bad actors to fine-tune expensive foundation models for cyber attacks at a fraction of the original training cost. Despite these reservations, Hinton conceded that the barrier to accessing big models has already disappeared, making open weights a permanent fixture. He maintained that worrying about superintelligent agents is valid and should not

be dismissed as fear-mongering, even as he acknowledged AI's potential to boost productivity in healthcare and education.

Li rejected the binary framing of the debate entirely, advocating for a nuanced approach akin to nuclear physics regulation. She pointed out that scientific papers in that field remain open while uranium is strictly controlled, suggesting AI ecosystems can operate with mixed levels of openness. Citing the Human Genome Project as a model, she described an infrastructure where public discovery coexists with lucrative private business models. The consensus among the three was that regulation is unavoidable to prevent tech CEOs from unilaterally deciding AI's trajectory, though the mechanism for balancing innovation, safety, and geopolitical competition remains undefined.

KEY FACTS

EVENT
Ai4 conference in Las Vegas

KEY SPEAKERS
Geoffrey Hinton, Fei-Fei Li, Andrew Ng

HINTON'S RISK FOCUS
Fine-tuning foundation models for cyber attacks

NG'S GEOPOLITICAL CONCERN
Chinese model adoption in Asia and Africa

LI'S ANALOGY
Nuclear physics regulation

> **WHY IT MATTERS**
>
> Practitioners should anticipate regulatory frameworks that differentiate between code transparency and weight release rather than applying blanket open or closed mandates. Strategic planning must now account for geopolitical pressures where model efficiency directly influences international soft power.

Read the original at TechCrunch →

---

07 HIGH IMPACT TechCrunch

Amazon Will Train AI on Twitch Streamers' Content by Default Unless They Opt Out

> Amazon will train generative AI models on Twitch creators' audio and video content by default, requiring manual opt-out to prevent data usage.

Twitch has updated its terms to include creator content in the training datasets for Amazon's generative AI models unless streamers explicitly disable the feature. The change applies to thousands of hours of recorded livestreams, capturing both voice and video data valuable for model development. Unlike standard industry practices where public data is scraped indiscriminately, Twitch implemented a specific toggle within channel settings under the security and privacy tab labeled "training for generative AI." However, the default state remains enabled, a decision Twitch Chief Product Officer Mike Minton justified during a live stream addressing nearly 3,000 users. When asked why the system was not opt-in, Minton stated, "If this was opt-in, nobody would opt in," acknowledging the community's prevalent opposition to generative AI training on personal content.

The announcement generated immediate confusion regarding the temporal scope of data collection. During the same broadcast, when a user questioned whether their historical videos had already been utilized for training,

Minton admitted he did not know what Amazon had previously used or excluded from model training efforts. This uncertainty compounds concerns among creators who monetize their accounts and cannot rely on private account settings as a workaround, a constraint noted by Twitch Head of Community Mary Kish. Kish compared the situation to Meta's approach, where public Facebook and Instagram content is used for AI training, noting that UK residents have a statutory opt-out while others must rely on privacy settings that are non-viable for professional broadcasters.

Technical practitioners deploying systems reliant on third-party data must note the friction in this acquisition pipeline. While Meta scrapes public content automatically, Twitch's implementation requires navigating a specific user interface path to exclude data, yet relies on user awareness to trigger that exclusion. The company framed the update as adding a setting to let users opt out rather than announcing a new data licensing agreement, a semantic distinction that obscured the scale of the shift until community backlash forced clarification. For engineers building models on social media data, this highlights a growing divergence between platform-level defaults and creator consent mechanisms, potentially introducing noise or legal risk if opt-out rates surge following this disclosure.

KEY FACTS

DEFAULT SETTING **Opt-out**	LIVE AUDIENCE SIZE **3000**
SETTING LOCATION **Channel Settings > Security and Privacy**	DATA TYPES **Audio and Video**

> **WHY IT MATTERS**
>
> Developers relying on social media datasets for training must verify inclusion status, as Twitch's default-opt-in policy may soon yield high opt-out rates that degrade dataset quality or introduce compliance gaps.

Read the original at TechCrunch →

SECTION 3 OF 3

AI Applications & Industry

4 stories 1 high 3 medium

08 HIGH IMPACT TechCrunch

Cognition AI in Talks to Raise at $40B Valuation as Devin Nears $1B ARR

> Cognition AI is negotiating a new funding round at a $40 billion valuation, driven by Devin's rapid ascent to a $1 billion annualized revenue run rate.

Cognition, the developer behind the AI coding agent Devin, has entered discussions with investors for a new capital raise that would value the company at $40 billion. This potential markup represents a significant increase from the $26 billion valuation established during its most recent funding round in May, where the startup secured $1 billion in capital. According to sources cited by Bloomberg, the basis for this steep revaluation is Devin's achievement of a $1 billion annualized revenue run rate. This financial milestone arrives only three months after CEO Scott Wu confirmed to TechCrunch that the company had reached a $492 million run rate in May, indicating a more than doubling of revenue velocity in a single quarter.

The acceleration in enterprise adoption underpins these financial figures. During the previous reporting period, Wu noted that enterprise usage of Devin had grown by 50% month-over-month for six consecutive months leading up to May. The platform's deployment strategy focuses on specific engineering workflows rather than generalist code generation. Wu explicitly stated that Devin is not marketed as a replacement for human developers; instead, it targets long-tail maintenance tasks that are often deprioritized by engineering teams. These use cases include modernizing legacy software stacks and migrating applications between different infrastructure platforms. This positioning has facilitated adoption among large-scale organizations, with the company listing Mercedes-Benz, NASA, and Goldman Sachs as current customers.

The speed at which Cognition has scaled its revenue suggests a shift in how enterprises are budgeting for AI agents. Moving from a $492 million run rate to $1 billion in approximately 90 days implies an aggressive expansion in seat count or contract value among existing clients. While the specific terms of the new round remain unconfirmed, the $40 billion target valuation would place Cognition among the most valuable private AI application companies globally. The distinction between the tool's function—handling grunt work versus replacing engineers—appears critical to its sales narrative, allowing it to integrate into established workflows at major institutions without triggering immediate workforce reduction concerns.

KEY FACTS

TARGET VALUATION
$40 billion

PREVIOUS VALUATION
$26 billion

CURRENT ARR
$1 billion

MAY ARR
$492 million

ENTERPRISE GROWTH RATE
50% month-over-month

KEY CUSTOMERS
Mercedes-Benz, NASA, Goldman Sachs

> WHY IT MATTERS
>
> The rapid doubling of Devin's revenue run rate in three months signals that enterprises are moving quickly from pilot programs to production-scale deployments for AI coding agents. Builders should anticipate increased competition for legacy modernization projects as these tools become standard infrastructure.

Read the original at TechCrunch →

---

09 MEDIUM IMPACT TechCrunch

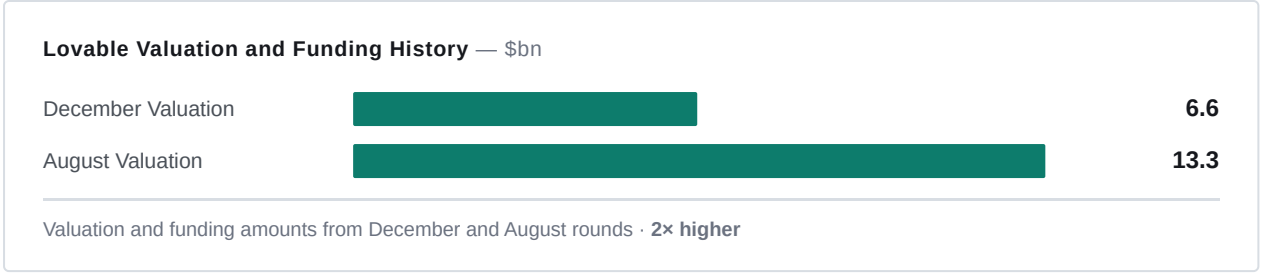
Lovable Raises $400M at $13.3B Valuation -- Vibe-Coding Startup Doubles in Eight Months

> Vibe-coding platform Lovable has secured a $400 million Series C round, doubling its valuation to $13.3 billion in eight months.

Lovable confirmed Wednesday that it raised $400 million in a Series C financing round led by Menlo Ventures and the Scaleup Europe Fund, with participation from more than a dozen other investors. This transaction values the European startup at $13.3 billion, a significant increase from its previous $6.6 billion valuation established in December during a $330 million round also led by Menlo Ventures, which included CapitalG as co-lead. The rapid valuation growth coincides with the company reporting $500 million in annualized run rate revenue as of June.

The funding supports expanding infrastructure demands driven by user growth; Lovable now hosts 60 million projects attracting 900 million monthly visitors. To handle this scale, the company has moved beyond relying solely on frontier model options to offering its own in-house trained AI model. Infrastructure commitments have scaled accordingly, highlighted by a multiyear deal signed with Google Cloud in June that represents a fivefold increase in usage. The investment roster for this round includes Regent, an investment firm that also owns TechCrunch.

Beyond its core platform, Lovable is deploying capital into the broader ecosystem through strategic backing of other European startups. One such investment is in Atech, a Danish startup developing vibe-coding software specifically designed for tech hardware creation. This expansion suggests a shift from purely providing a coding interface to influencing the underlying toolchain and hardware design layers within the European AI development landscape. The combination of proprietary modeling, increased cloud capacity, and ecosystem investment marks a transition from early-stage experimentation to established infrastructure provider.



KEY FACTS

SERIES C RAISE
$400 million

POST-MONEY VALUATION
$13.3 billion

ANNUALIZED RUN RATE REVENUE
$500 million

HOSTED PROJECTS
60 million

MONTHLY VISITORS
900 million

CLOUD USAGE INCREASE
5x

WHY IT MATTERS

The shift to in-house models and a fivefold increase in cloud usage signals that high-volume vibe-coding platforms are outgrowing reliance on third-party APIs, requiring practitioners to evaluate hybrid inference architectures for cost and latency control.

Read the original at TechCrunch →

---

10 MEDIUM IMPACT TechCrunch

Thrive Holdings Raises $2B at $12B to Bring OpenAI-Powered Tools to Accounting and IT

> Thrive Holdings secured $2 billion in new funding at a $12 billion valuation to expand its model of embedding OpenAI-derived workflows into traditional service sectors.

Thrive Holdings, a spinout of Thrive Capital, closed a $2 billion round led by SoftBank, D1 Capital Partners, and Altimeter Capital. The firm operates as a private equity vehicle specifically for AI integration, acquiring traditional businesses in accounting and information technology to retrofit them with proprietary AI stacks. OpenAI holds an ownership stake in Thrive Holdings following a December 2025 agreement that included deploying OpenAI employees directly into Thrive's portfolio companies to accelerate adoption. This hands-on implementation strategy mirrors recent moves by Anthropic, which partnered with private equity to launch Ode, signaling a broader industry shift toward embedded engineering teams rather than pure API distribution.

The capital injection follows operational metrics from Thrive's two existing platforms: Current (accounting) and Shield (IT). Current, comprising over 50 firms and 2,000 professionals, deployed self-improving tax agents called TaxAI. These agents processed more than 7,000 tax returns with 98% accuracy, reducing preparation times by over 30%. Shield, hosting approximately 20 companies, reported a 36x speedup in help desk resolution times and doubled its deployment of custom AI agents in the last month alone. The combined portfolio now exceeds 70 businesses.

A portion of the new funding will launch a third vertical focused on regulatory services for the built environment, targeting bottlenecks in data centers, manufacturing, and infrastructure projects. Founding

members Anuj Mehndiratta and Kareem Zaki stated the initiative aims to automate research, permit preparation, and compliance tracking while maintaining human sign-off for field work and safety standards. The strategy targets fragmented, operationally complex sectors where AI can compress regulatory timelines without replacing professional judgment.

This raise validates a business model where value creation stems from deep operational integration rather than software licensing alone. By taking equity stakes in legacy firms and inserting elite engineering resources directly into workflows, Thrive bypasses the typical friction of enterprise sales cycles. The involvement of OpenAI personnel suggests that future model iterations may be co-developed or fine-tuned specifically for these high-volume, domain-specific tasks, creating a feedback loop between real-world usage data and model capability.

KEY FACTS

FUNDING AMOUNT **$2 billion**	VALUATION **$12 billion**
PORTFOLIO SIZE **70+ businesses**	TAX RETURNS PROCESSED **7,000+**
TAX PREP ACCURACY **98%**	HELP DESK SPEEDUP **36x**

> **WHY IT MATTERS**
>
> Practitioners should note the pivot toward equity-based integration models where AI vendors embed teams directly into client operations, potentially offering higher performance gains than standard API deployments but requiring deeper structural alignment.

Read the original at TechCrunch →

---

11 MEDIUM IMPACT TechCrunch

Blacksmith Hits $550M Valuation: AI Code-Testing Startup Grows 10x in Under a Year

> AI code-testing startup Blacksmith has secured a $550 million valuation following a $45 million Series B round as it scales to serve over 5,000 customers.

Blacksmith closed a $45 million Series B financing round led by Peak XV Partners, assigning the company a post-money valuation of $550 million. This represents a tenfold increase from its $60 million valuation during a $10 million Series A less than a year ago. Existing backers GV and Y Combinator participated, bringing total capital raised to $58.5 million since the company's founding in 2024. The funding targets the expanding bottleneck of software validation caused by the rapid adoption of AI coding tools like Cursor, OpenAI's Codex, and Anthropic's Claude Code.

The startup's growth metrics reflect a sharp uptake in demand for automated testing infrastructure. Customer count surged from more than 700 to over 5,000 in under twelve months, including enterprise clients such as Mercury, Supabase, Clerk, Ashby, and Expensify. While CEO Aditya Jayaprakash declined to disclose a specific updated annualized revenue run rate (ARR), he confirmed the figure has grown from an initial $10

million to "tens of millions of dollars" with a team of approximately 30 employees. Notably, several large customers now exceed $1 million in annual spend on the platform.

Technically, Blacksmith has evolved from a cloud provider for continuous integration (CI) workloads into a broader validation suite. Its latest offering, Codesmith, functions as an AI coding agent capable of automatically fixing failed code checks before they reach production. This positions the company against a crowded field including GitHub Actions, native validation features in Codex and Claude Code, and cloud-based testing services from Amazon Web Services, Microsoft Azure, and Google Cloud. Blacksmith differentiates through testing speed and cost efficiency, with plans to expand further into tools that assist developers in writing, validating, and merging software.

KEY FACTS

SERIES B VALUATION
$550 million

SERIES B RAISE
$45 million

CUSTOMER COUNT GROWTH
700 to 5,000+

TOTAL FUNDING
$58.5 million

INITIAL ARR
$10 million

CURRENT HEADCOUNT
30

> **WHY IT MATTERS**
>
> As AI generation accelerates code output, the constraint shifts from writing to validating; Blacksmith's rapid scaling suggests teams are prioritizing automated repair agents over manual review pipelines.

Read the original at TechCrunch →

Sources

01 DeepSeek V4 Pro 0813: Flagship 1.6T MoE Model Exits Preview
https://openrouter.ai/deepseek/deepseek-v4-pro-0813

02 Liquid AI Ships LFM2.5-VL-3B: Vision-Language Model That Runs on Your Phone
https://huggingface.co/blog/LiquidAI/lfm2-5-vl-3b

03 AI4AI at Test-Time: Can Strong Models Transfer Skills to Weak Ones Without Training?
https://arxiv.org/abs/2608.12307

04 195TB of Credentials Leaked in LiteLLM Supply-Chain Attack Impacting 2,500 Organizations
https://arstechnica.com/security/2026/08/terabytes-of-credentials-leaked-in-massive-supply-chain-attack/

05 Made by Google 2026: Pixel 11, Gemini Intelligence, and Sign Language AI
https://techcrunch.com/2026/08/12/google-unveils-pixel-11-lineup-new-airtag-rival-and-gemini-features-at-made-by-google-2026/

06 Hinton, Ng, and Fei-Fei Li Unite on Openness -- With Caveats
https://techcrunch.com/2026/08/12/as-ai-safety-concerns-mount-three-pioneers-make-the-case-for-staying-open/

07 Amazon Will Train AI on Twitch Streamers' Content by Default Unless They Opt Out
https://techcrunch.com/2026/08/12/amazon-will-train-on-twitch-streamers-content-by-default-unless-they-opt-out/

08 Cognition AI in Talks to Raise at $40B Valuation as Devin Nears $1B ARR
https://techcrunch.com/2026/08/12/ai-coding-startup-cognition-reportedly-already-in-talks-to-raise-at-40b-valuation/

09 Lovable Raises $400M at $13.3B Valuation -- Vibe-Coding Startup Doubles in Eight Months
https://techcrunch.com/2026/08/12/lovable-confirms-new-13-3b-valuation-raises-another-400m/

10 Thrive Holdings Raises $2B at $12B to Bring OpenAI-Powered Tools to Accounting and IT
https://techcrunch.com/2026/08/12/openai-backed-thrive-holdings-raises-2b-to-bring-ai-to-the-enterprise/

11 Blacksmith Hits $550M Valuation: AI Code-Testing Startup Grows 10x in Under a Year
https://techcrunch.com/2026/08/12/blacksmiths-valuation-jumps-10x-to-550m-as-ai-coding-fuels-software-validation/

About this document. Every story in the 13 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 2026-08-13T05:34:55Z · model qwen3.5:397b-cloud · New Horizon, new-horizon.tech · To unsubscribe or manage your subscription, see the footer of the daily email.