Cerebras Powers Ultrafast Mode for OpenAI's GPT-5.6 Sol

August 13, 2026

Cerebras powers OpenAI's most capable model at up to 14x the speed, giving frontier intelligence with no compromise on latency

SUNNYVALE, Calif., Aug. 13, 2026 (GLOBE NEWSWIRE) -- Cerebras (NASDAQ: CBRS) today announced that it is powering Ultrafast mode, a new service tier in the OpenAI API for GPT-5.6 Sol. Available initially in limited preview to OpenAI customers, Ultrafast runs GPT-5.6 Sol at up to 750 output tokens per second and up to 14× faster than Standard processing.

GPT-5.6 Sol Ultrafast powered by Cerebras runs with the same intelligence as GPT-5.6 Sol Standard, enabling frontier intelligence at Cerebras' blistering fast speed.

Every major computing shift has been unlocked by a leap in speed, not just capability: the PC era needed the jump from kilohertz to gigahertz, and the internet needed the move from dial-up to broadband before it could reach everyone. AI is no different. Now that frontier models are demonstrably capable, the next constraint on adoption is how fast they run.

"GPT-5.6 Sol on Ultrafast is proof that speed and intelligence are no longer mutually exclusive," said Andrew Feldman, CEO and co-founder, Cerebras. "Together with OpenAI, we're putting frontier intelligence in the hands of users at unprecedented speed and changing what's possible with AI."

"By combining GPT-5.6 Sol with Cerebras' inference technology, we're exploring what becomes possible when customers can get the intelligence of our most capable models with significantly lower latency. We're starting with a small group of customers to learn where that speed creates meaningful value, and we'll use those learnings to inform how we expand the service over time," said Sachin Katti, VP Compute Strategy & GPT-Infra at OpenAI.

A New Speed-Intelligence Frontier

Until now, organizations needed to choose between the capabilities of larger models and the faster response times of smaller models. Ultrafast powered by Cerebras offers access to the full intelligence of GPT-5.6 Sol at speeds suited to work that cannot wait. Based on output speeds for Anthropic models reported by Artificial Analysis, Ultrafast is 5x faster than Claude Opus 4.8 in Fast mode, and 11x faster than Claude Fable 5. In doing so, it opens up a new region of speed-intelligence: frontier intelligence combined with unprecedented speed.

Benchmarks focused on economically valuable work, from programming to drafting legal documents, show how faster token generation translates directly into higher productivity for users.

On Humanity's Last Exam, a 2,500-question benchmark spanning graduate-level chemistry, economics and literature, GPT-5.6 Sol Ultrafast answered the full question set in just over 11 hours. This compares to more than three days of continuous compute for Claude Fable 5, with GPT-5.6 Sol Ultrafast reaching comparable accuracy nearly 7x faster. On GDP-Val, a benchmark of economically valuable knowledge-work tasks such as legal briefs, financial models, and engineering reports, Ultrafast delivered a 5.6x end-to-end speedup with no loss in quality.

Ultrafast's speed comes from Cerebras' Wafer-Scale Engine architecture, which keeps model weights on-chip — 44 GB of SRAM on each wafer-sized chip — rather than shuttling them between on-chip memory and off-chip storage as GPU-based inference must. This eliminates the memory-bandwidth bottleneck that constrains frontier-model inference speed on conventional hardware.

To get notified when capacity expands to more customers, please visit the Cerebras website.

About Cerebras Systems
Cerebras Systems (NASDAQ: CBRS) builds the world's fastest AI infrastructure. The Cerebras team of pioneering computer architects, computer scientists, AI researchers, and engineers of all types came together to make AI blisteringly fast through innovation and invention. We believe that when AI is fast, it will change the world. Leading global corporations, research institutes, and governments choose Cerebras to run their AI workloads. Cerebras solutions are available on premises and in the cloud. Visit cerebras.ai for more.

Cerebras Disclosure Information
Cerebras uses its investor relations page (investors.cerebras.ai), its X account (@cerebras), and its LinkedIn page

(linkedin.com/company/cerebras-systems/) to disclose material nonpublic information and for complying with its disclosure obligations under Regulation FD. Accordingly, investors should monitor these channels, in addition to following Cerebras' press releases, Securities and Exchange Commission (SEC) filings, public conference calls and public webcasts.

Forward-Looking Statements

This press release contains "forward-looking statements" within the meaning of applicable securities laws. All statements other than statements of historical fact could be deemed to be forward-looking. The words "may," "will," "shall," "should," "expects," "plans," "anticipates," "could," "intends," "target," "projects," "contemplates," "believes," "estimates," "predicts," "potential," "objective," or "continue," or the negative of these words or other similar terms or expressions that concern our expectations, strategy, plans, or intentions are intended to identify forward-looking statements, although not all forward-looking statements contain these identifying words. These forward-looking statements are subject to a number of risks and uncertainties, many of which involve factors or circumstances that are beyond Cerebras' control. These risks and uncertainties include, but are not limited to: Cerebras' ability to sustain and manage its growth, access borrowings and other sources of capital on acceptable terms, and deploy available capital to support growth; its history of net losses and ability to achieve and maintain profitability; its limited operating history at its current scale and ability to accurately forecast revenue and appropriately budget and manage expenses; its dependence on a limited number of significant customers, including OpenAI, Group 42 Holding Ltd, Mohamed bin Zayed University of Artificial Intelligence, and AWS, and the potential impact of any reduction in demand from, material adverse development in its relationships with, or failure to meet its obligations to, such customers; the timing, execution and expected benefits of its strategic customer, partner and financing arrangements; its historical reliance on sales of hardware systems and the early-stage, rapidly evolving market for its cloud-based offerings and AI infrastructure; its ability to secure sufficient data center capacity and capital to support its cloud-based offerings; its ability to launch new offerings and add new product capabilities; and its ability to compete effectively in the rapidly evolving and competitive market for AI computing solutions.

Cerebras' actual results could differ materially from those stated or implied in forward-looking statements due to a number of factors. Accordingly, undue reliance should not be placed on such statements. These forward-looking statements are made as of the date they were first issued and are based on information available to Cerebras together with Cerebras' expectations, estimates, forecasts, projections, beliefs, and assumptions as of such date. These forward-looking statements should not be relied upon as representing Cerebras' views as of any date subsequent to the date of this press release. Past performance is not necessarily indicative of future results. Cerebras undertakes no intention or obligation to update or revise any forward-looking statements, whether as a result of new information, future events, or otherwise, except as required by law.

Further information on potential risks that could affect actual results is included in Cerebras' most recent filings with the SEC, including in Cerebras' most recent Quarterly Report on Form 10-Q, copies of which may be obtained by visiting Cerebras' Investor Relations website at investors.cerebras.ai or the SEC's website at www.sec.gov.

Contacts

Kriselle Laran
Media Relations
pr@cerebras.ai

Sean Dorsey
Investor Relations
investors@cerebras.ai