THE WALL STREET JOURNAL.

TUESDAY, MARCH 31, 2026 © 2026 Dow Jones & Company, Inc. All Rights Reserved.

EXCLUSIVE

Caltech Researchers Claim Radical Compression of High-Fidelity AI Models

PrismML says its 1-bit large language model achieves radical compression without sacrificing performance, lowering energy consumption

By Steven Rosenbush

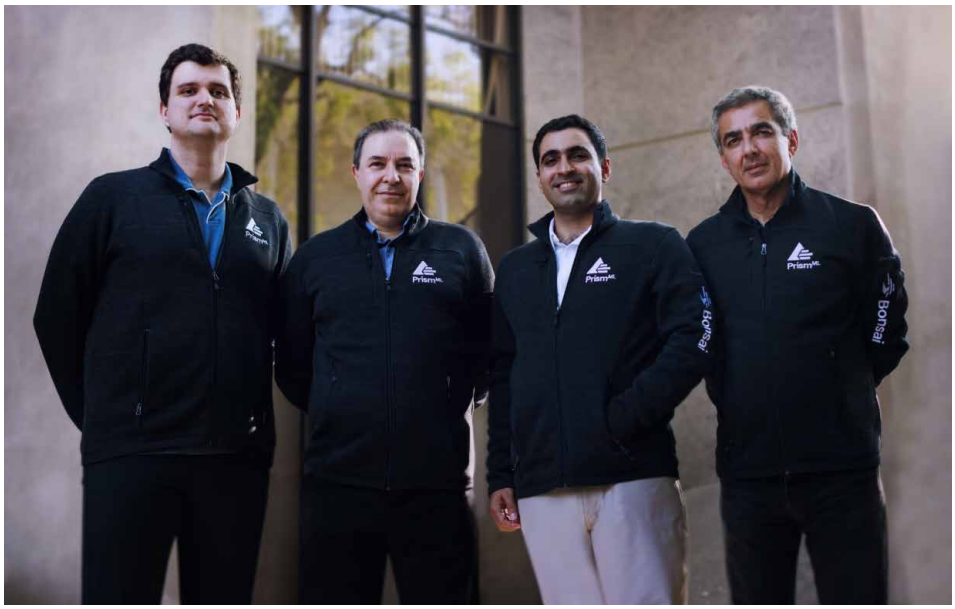
A team of researchers led by California Institute of Technology computer scientist and mathematician Babak Hassibi says it has created a large language model that radically compresses its size without compromising performance.

The company, PrismML, came out of stealth Tuesday and open-sourced its 1-bit technology model, enabling others to use it.

PrismML has developed an extreme form of compression that allows AI to run locally on phones, laptops and other devices, and enables data center build-outs that can do more with fewer resources and avoid ballooning energy costs, according to Hassibi.

"We spent years developing the mathematical theory required to compress a neural network without losing its reasoning capabilities," said Hassibi, chief executive of the venture. "We are creating a new paradigm for AI: one that adapts to diverse hardware environments and delivers maximum intelligence per unit of compute and energy," he said.

The other Caltech-affiliated co-founders include Sahin Lale, Omead Pooladzandi, and Reza Sadri, who is also vice president, strategy.



The PrismML team, from left: Sahin Lale, co-founder, Babak Hassibi, co-founder and CEO, Omead Pooladzandi, co-founder, and Reza Sadri, co-founder and vice president, strategy.

ENOCH KIM

The intellectual property is owned by Caltech, and PrismML is the sole exclusive licensee, Hassibi said.

The company raised $16.25 million in a SAFE and seed round with investors Khosla Ventures, Cerberus Capital and Caltech. A SAFE, or Simple Agreement for Future Equity, occurs when an investor gives a startup money in exchange for the right to receive equity later.

AI's future won't be defined by who can build the largest data centers, but by who can deliver the most intelligence per unit of energy and cost, according to investor Vinod Khosla. "So this is not a minor iteration. This is a major technical breakthrough," Khosla said. "It's a mathematical breakthrough, not just another tiny model."

© 2026 Dow Jones & Co. Inc.
Licensed Use: Post to corporate site
Licensed to: 43PR
Expiration Date: 10/6/2026

The Publisher's Sale Of This Reprint Does Not Constitute Or Imply Any Endorsement Or Sponsorship Of Any Product, Service, Company Or Organization.
Custom Reprints 800.803.9100 www.djreprints.com DO NOT EDIT OR ALTER REPRINT/REPRODUCTIONS NOT PERMITTED 2665454

DOW JONES

PrismML answers a need, Khosla said, for fast and small but high-performing models that serve a range of applications from voice conversation to some aspects of finance.

The models PrismML developed are designed to operate on consumer devices such as smartphones and laptops, as well as industrial-edge devices. The idea is to enable applications in robotics, wearables, and personal computing that were previously impractical, the company said.

"You can fit a much better model on a phone. That's a big deal. Of course on your phone or a mobile device, energy consumption is a very, very big deal," Khosla said.

The same efficiency gains that enable local deployment also allow data centers to operate more effectively, PrismML said.

While the broader tech industry fiercely debates whether the future of AI lies in transformers, diffusion models, or newer concepts, PrismML's mathematical framework can be applied to any of them, according to Hassibi.

How It Works

One way to describe an AI model is in terms of bits, which refers to the amount of code needed to render a numerical value in ones and zeros, the language of computing. Most AI models are written with a 16-bit level of precision, although some approaches employ 4-bits or fewer. PrismML has achieved a mathematical breakthrough that achieves a 1-bit model without compromising the reasoning, coding, and general knowledge capabilities of full-precision models, according to Hassibi. The mathematics are proprietary, but Hassibi said the effect was much like compressing a digital photograph without losing visual fidelity.

When it comes to running AI models, delays, known as latency, and energy consumption are tied to moving data in and out of memory. By reducing the units of data, or model weights, to a single bit represented by +1 or -1, PrismML's flagship 1-bit Bonsai 8B model can boost processing speeds by as much as eight times compared with a 16-bit model, Hassibi said. It can also achieve reductions in energy consumption of up to 75% to 80% on current hardware platforms, Hassibi said. If future hardware is designed specifically for one-bit models, it will eliminate the need for complex mathematical multiplications altogether, he said. The hardware would only need to perform simple addition and subtraction, which would drive energy consumption and latency down even further, according to Hassibi.

Amir Salek, senior managing director at Cerberus Capital Management, said he was convinced PrismML achieved a major mathematical breakthrough with the potential to improve the economics of AI.

By employing a 1-bit architecture, a two-terabyte model instantly becomes 150 gigabytes, according to PrismML. "Your bandwidth requirement significantly drops, your memory storage size significantly drops and the energy that you consume to move the data... is reduced in a big way," Salek said. He was previously founder and Head of Silicon for the Google Technical Infrastructure and Google Cloud businesses. Before that, he was founder and head of Nvidia's System-on-a-Chip Design organization.

Developers, researchers, and other users can download PrismML's open source 1-bit Bonsai 8B model free.

Bonsai 8B is an 8-billion parameter large language model, trained using Google v4 TPUs.

The model achieves high-fidelity reasoning and language understanding comparable to 16-bit models, but with a memory footprint of 1 gigabyte vs 16 gigabytes, according to PrismML. High-fidelity reasoning is the capability to successfully perform complex reasoning.

The company said it is also releasing 1-bit Bonsai 4 billion parameter and 1.7 billion parameter models, with 0.5 gigabytes and 0.24 gigabytes memory footprints, and even higher intelligence density.