

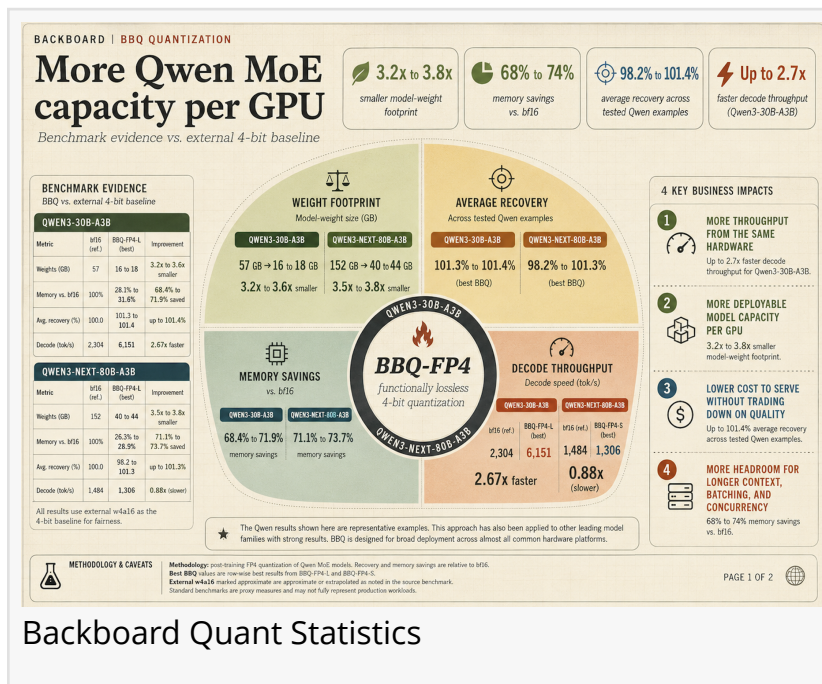
Backboard.io Announces AI Model Compression of Up to 70 Percent, Developed in Ottawa

Backboard.io unveils a model-compression breakthrough, a frontier-class coding suite, a multi-model chat app, and the world's #1-ranked AI memory.

OTTAWA, CANADA, July 1, 2026 /EINPresswire.com/ -- Hundreds of billions in global capital is going toward new AI hardware, yet supply remains constrained. GPU lead times run into quarters, data-centre construction is backlogged against power and cooling limits, and grid interconnection queues in North America now stretch years. The result is a gap between the compute organizations want and the compute they can obtain. All of this spend is to support frontier lab LLM's growing in size and hardware requirements.

“

The mandate is there. Enterprises want an option they can trust. We are here, and you do not need to wait. Major enterprises around the world are already working with us.”
Robert Imbeault, CEO



Backboard Quant Statistics

[Backboard.io](https://backboard.io) has taken a different approach: it is building on the premise that long-term value in AI accrues in memory, context, and user data rather than in the model itself. On that basis the company has built an AI memory system, a unified API to run it, a harness that runs open-source models on coding benchmarks, and model compression intended to extend advanced AI to on-premise and edge deployments.

Today the company announced four developments that address three issues common to organizations using AI:

cost, data shared with outside providers, and the use of unapproved tools. All were built in Nepean, Ontario, in Prime Minister Mark Carney's riding, by a team composed entirely of

graduates of Canadian universities, colleges, and CEGEPs.

BackboardQuant: more usage per GPU

BackboardQuant (BBQ) compresses AI models by up to 70 percent. In internal testing, compressed models retained accuracy comparable to their full-precision versions while running up to 2.7x faster, allowing one GPU to handle the workload of two or three. For the current compute shortage, the effect is measurable: instead of waiting on new hardware and the power and construction timelines behind it, a cloud provider or enterprise can serve up to three times the usage from the same installed fleet, and at a lower cost per unit served. Developed in Backboard's research lab, BBQ ships built into the company's enterprise deployments.

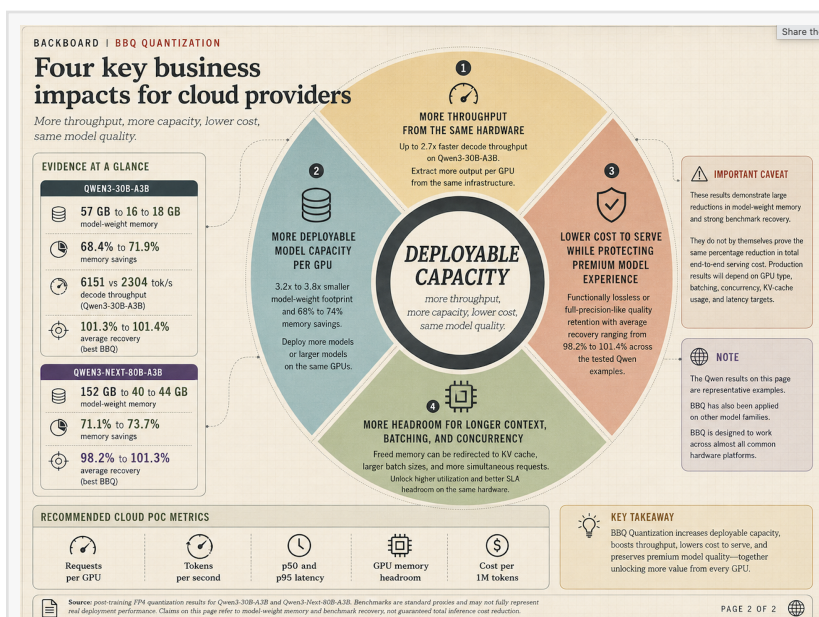
[Backboard Studio](#): coding performance at lower cost

Backboard Studio performs comparably to coding tools from major AI labs on public benchmarks, at up to 90 percent lower cost. On Terminal-Bench 2.1, an independent benchmark for agentic coding, Backboard Studio scores 79.8 percent running Claude Opus 4.8, compared with published results of 78.2 percent for GPT-5.5 and 78.9 percent for Opus 4.8 on the same public harness. Running the open-source GLM 5.2, it scores above 72 percent without a proprietary model. A built-in token optimizer reduces frontier model usage by up to 30 percent on like-to-like comparisons. It runs on cloud or self-hosted, so source code does not leave the customer's environment. Available now.

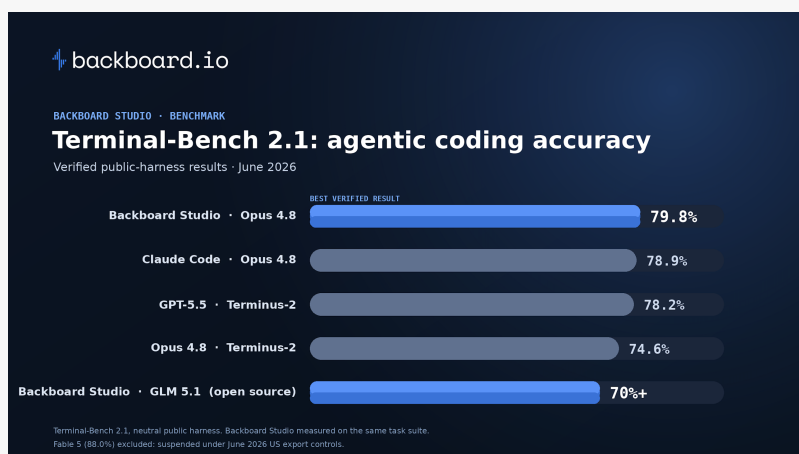
[Nash](#): multiple AI models in one app

Nash gives users access to thousands of AI models across text and images in a single chat app while keeping user memory separate from the models. For enterprises, it addresses shadow AI by providing one sanctioned application in place of the unapproved tools employees adopt on their own. Available now for consumer and enterprise use at hellonash.ai.

AI memory benchmarks



Backboard Quant Implications



Backboard Leading Coding Benchmarks

Backboard ranks first on LoCoMo and LongMemEval, two independent benchmarks for AI memory. The results are public and reproducible: LoCoMo, LongMemEval. Memory stays inside the customer's environment, under the customer's control.

Sovereign deployment

Backboard's full stack, including the API, the application layer, and the models, can run inside a customer's own cloud. Data remains in place and the system operates under the customer's governance. Combined with model compression, this lowers the hardware required to run advanced AI in place, extending it from large data centres to on-premise and edge deployments. For governments, hospitals, banks, and critical infrastructure, this supports running AI without moving data to outside providers.

About Backboard.io

Backboard.io is a Canadian AI company that helps organizations run AI more affordably, more securely, and under their own control. Its products include BackboardQuant, which compresses AI models by up to 70 percent (enterprise deployments); Backboard Studio, a coding suite with a desktop IDE and the R-CLI; Nash (hellonash.ai), a multi-model, multi-modal chat app; and an AI memory system ranked first on the LoCoMo and LongMemEval benchmarks.

Maya Ellis

Backboard IO

[email us here](#)

Visit us on social media:

[LinkedIn](#)

[Instagram](#)

[YouTube](#)

[X](#)

This press release can be viewed online at: <https://www.einpresswire.com/article/923113831>

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.

© 1995-2026 Newsmatics Inc. All Right Reserved.