Jensen Huang Says AGI Has Arrived With GPT-6 Astra. The Data Is More Complicated

TechRounder PDF Edition

Live article:
https://www.techrounder.com/ai/jensen-huang-says-agi-has-arrived-with-gpt-6-astra-the-data-is-more-complicated/

By Vipin PG | Published September 7, 2026 | Updated September 7, 2026 | Format: Analysis | 7 min read

In brief

Nvidia CEO Jensen Huang says artificial general intelligence has arrived.

Nvidia CEO Jensen Huang says artificial general intelligence has arrived.

He made the declaration on September 6 while congratulating OpenAI on GPT-6 Astra, the company's new flagship model. His post came three days after OpenAI President Greg Brockman told reporters, "Welcome to the AGI era."

Those statements have turned Astra's launch into a much larger argument about whether AI has crossed the line into artificial general intelligence.

The benchmark data gives Astra a strong case for being one of the most capable AI systems released so far. OpenAI reports scores of 98% on FrontierMath Tier 4, 99.9% on ARC-AGI-3 and 100% on ExploitBench. The model is also designed for long-running computer tasks, software development, scientific work, cybersecurity and professional document workflows.

The 99.9% ARC-AGI-3 score needs some explanation, though. Independent testing by ARC Prize produced a much lower result when Astra was placed inside its standard evaluation setup.

That difference tells us something useful about where advanced AI is heading. The model matters, and so does the system running around it.

What Jensen Huang actually said

Huang's post tied Astra's performance directly to Nvidia hardware.

He said the model had been trained on more than 100,000 Nvidia Grace Blackwell NVLink72 GPUs, described the progression from ChatGPT to OpenAI's o1 reasoning models and then Astra, and finished with a direct statement: "AGI has arrived."

He also said another 400,000 GPUs were coming online.

The infrastructure reference matters because Nvidia is one of the biggest commercial beneficiaries of the current AI buildout. Its latest quarterly results showed $96.2 billion in revenue, with $89 billion coming from its Data Center business.

For Nvidia, bigger AI models and longer-running AI agents mean continued demand for training and inference hardware. Huang's comments therefore carry two kinds of weight. He is one of the most influential figures in the AI industry, and Nvidia also supplies much of the computing infrastructure behind that industry.

OpenAI has been more careful with the AGI label

Brockman’s wording during Astra’s launch was strong, although he framed it as his personal view rather than an agreed scientific conclusion.

He said he believed there was a case that AI had reached this stage and suggested that people looking back several years from now might identify this period, and perhaps Astra itself, as the beginning of the AGI era.

OpenAI’s own definition of AGI focuses on economics. Its charter describes AGI as highly autonomous systems that outperform humans at most economically valuable work.

That definition creates a difficult measurement problem. A model can perform exceptionally well on coding, mathematics or reasoning tests without proving that it can outperform humans across most economically valuable jobs.

OpenAI CEO Sam Altman has also acknowledged the problem with the terminology. He has described AGI as poorly defined and has questioned how useful the term remains.

There is still no industry-wide test that produces a simple "AGI achieved" result.

Astra’s published benchmark scores

OpenAI’s launch material includes several unusually high results:

- 98% on FrontierMath Tier 4
- 99.9% on ARC-AGI-3
- 100% on ExploitBench

These tests measure different things.

FrontierMath focuses on difficult mathematical problems. ExploitBench tests cybersecurity skills against known software vulnerabilities. ARC-AGI-3 is designed to test whether an AI system can enter an unfamiliar interactive environment, work out what is happening and learn how to complete tasks without receiving normal instructions about the rules or objective.

ARC-AGI-3 is particularly relevant to the AGI discussion because it tries to measure adaptation rather than simple recall.

That is also where Astra’s results become more complicated.

The same Astra model scored 62.7% and 99.9%

ARC Prize tested GPT-6 Astra using two different setups.

With its Standard harness and maximum reasoning effort, Astra scored 62.7%. The reported evaluation cost was $26,098.

When Astra ran through OpenAI’s Provider Adapter at high reasoning effort, it reached 99.9% at a reported cost of $18,817.

A more direct comparison uses the same maximum reasoning level. Astra scored 62.7% with the Standard harness and 98.6% through the Provider Adapter.

The underlying Astra model did not change between those tests. The surrounding software did.

Why the Provider Adapter changes the result

The Standard ARC Prize setup gives the model a more limited way to preserve information while it moves through an environment. The model can carry forward notes that it decides to keep.

OpenAI's Provider Adapter can maintain opaque reasoning state between requests and compact long interactions. Astra can therefore continue working with more of its previous reasoning intact instead of repeatedly rebuilding its understanding of the task.

The effect is large.

The New Stack's analysis of ARC Prize data found that the Provider Adapter used fewer tokens and completed shared solved tasks much faster than the standard setup.

The result shows how difficult it is becoming to describe AI performance with a model name and a single percentage.

GPT-6 Astra inside one runtime scored 62.7%. GPT-6 Astra inside a more capable runtime reached almost 100%.

Both results describe Astra. They describe different systems built around Astra.

The 62.7% result is still a large advance

ARC Prize treats Astra as a major improvement even under its stricter Standard harness.

Earlier frontier systems had performed far below the level Astra reached. ARC-AGI-3 was created to make brute-force scaling and memorized answers less useful by placing models inside unfamiliar interactive environments.

Astra showed that it could explore those environments, infer rules and use compact internal representations to guide later actions.

ARC Prize also reported that Astra required fewer actions than the median human baseline on most of the tested levels.

Its researchers described the performance as a step-change in capability.

They stopped short of calling the result proof of AGI.

ARC-AGI-3 environments are still bounded and deterministic. Real workplaces, research problems, human relationships, business decisions and physical environments contain far more uncertainty.

Astra has crossed OpenAI's Critical cybersecurity threshold

One Astra result has a clearer definition than AGI.

OpenAI says GPT-6 Astra is its first model to reach the Critical level for cybersecurity capability under its Preparedness Framework.

At this level, a model can find previously unknown vulnerabilities and develop exploits against hardened systems with much less step-by-step human guidance.

During testing, Astra found previously unknown vulnerabilities, including two zero-day vulnerabilities that OpenAI said it was reporting to the affected maintainers.

OpenAI has placed additional restrictions and monitoring around these abilities. Some stronger cybersecurity functions are being distributed through controlled programs rather than being made freely available to every user.

The 100% ExploitBench result also comes with a warning from OpenAI's own system card. The company says historical vulnerability information may have appeared in Astra's training data, which could make the benchmark score look better than it would on completely unseen vulnerabilities.

OpenAI ran other cybersecurity evaluations using newer vulnerabilities and still concluded that Astra had crossed its Critical capability threshold.

AI benchmarks are becoming system benchmarks

The ARC-AGI-3 result exposes a problem that will become harder to ignore as AI agents improve.

A modern AI product can include the base model, persistent state, memory, context compression, browsing, software tools, computer access, retry logic and monitoring.

Change those components and the performance of the same model can change dramatically.

This already happens in coding agents. A strong coding model can produce different results depending on the tools, context handling and execution environment provided by the agent software around it.

Astra gives us a clean example because ARC Prize tested the same model under two different harnesses and published both results.

A headline saying "Astra scored 99.9% on ARC-AGI-3" is technically supported by the Provider Adapter run. Readers also need to know that the Standard harness produced 62.7%.

Those numbers answer different questions.

Has GPT-6 Astra reached AGI?

There is no agreed test that can settle that claim today.

Huang believes the threshold has been crossed. Brockman thinks this period may later be remembered as the beginning of the AGI era. Gary Marcus has rejected the declaration and argued that claims of AGI need clearer definitions and stronger evidence.

ARC Prize takes a more measurable position. Astra has made a large advance on its benchmark, and saturation of that benchmark does not prove AGI.

OpenAI's own economic definition sets an even broader bar. Evidence would need to show that highly autonomous AI systems can outperform humans across most economically valuable work. Current benchmark results cover narrower slices of that claim.

Astra still changes the practical conversation around advanced AI. It can operate computers, work through unfamiliar tasks, perform advanced scientific and mathematical reasoning, write and execute software, and carry out cybersecurity work at a level that OpenAI now considers Critical.

The ARC-AGI results add another lesson. More of an AI system's capability is moving into the runtime around the model. Memory, preserved reasoning state and tool access can turn the same model into a substantially more capable agent.

Huang's "AGI has arrived" statement will probably remain the headline attached to Astra's launch. The benchmark data gives us something more useful to track: how much work these systems can complete independently, how well those results hold up under neutral testing, and how much of the performance comes from the model versus the software around it.

We'll keep watching Astra's independent evaluations as researchers test it outside OpenAI's own runtime and publish more comparable results.