S&P Global
Mobility

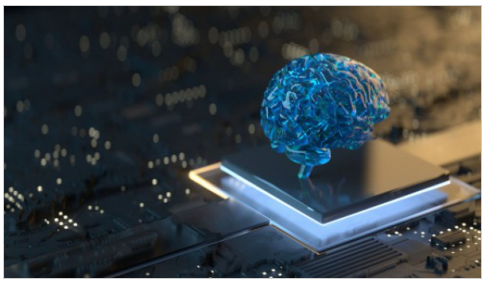
AutoTechInsight Insights
December 2025

Driving the future: NVIDIA's vision for AI in automotive

22-Dec-2025 10:30 GMT

Matthew Beecham
S&P Global
Supply Chain and Technology, Automotive

Q&A with NVIDIA



Source: Getty image/ Just_Super

NVIDIA is redefining the trajectory of automotive technology, with a vision for AI that extends far beyond the traditional concept of AI-defined vehicles. In a recent interview with S&P Global Mobility, NVIDIA shared insights on the evolution and integration of generative AI, agentic AI and physical AI within the mobility sector. This multilayered approach starts with GenAI, which creates realistic scenarios for autonomous vehicle (AV) training, and advances to agentic AI — capable of autonomous reasoning and decision-making — laying the groundwork for physical AI, where models execute real-world actions with safety and precision.

A landmark achievement highlighted in the discussion is NVIDIA's open-source Alpamayo-R1 (AR1) model, the industry's first vision-language-action (VLA) model with chain-of-thought reasoning. This breakthrough enables vehicles to interpret complex environments, anticipate novel situations and make safe decisions, even in scenarios not previously encountered. The model's open-source nature accelerates industrywide innovation, allowing partners to adapt and refine the technology for their unique needs.

NVIDIA's holistic ecosystem strategy, encompassing the Hyperion hardware platform, advanced data generation tools such as Cosmos, and neural reconstruction engines, empowers automotive partners to build, test and validate robust AV solutions. By tackling the challenge of long-tail scenarios and promoting safety through embedded monitoring and redundancy, NVIDIA is driving the future of mobility. Its commitment to open collaboration and cutting-edge AI sets a new benchmark for safety, adaptability and industry progress in autonomous driving.

To learn more, Owen Chen, Senior Principal Analyst at S&P Global Mobility, spoke to Ali Kani, vice president of Automotive at NVIDIA.

[Source: NVIDIA]

Key takeaways:

- ***NVIDIA is driving industry progress with open-source, state-of-the-art AI models***
  NVIDIA's release of the Alpamayo VLA model, featuring chain-of-thought reasoning, marks a significant milestone for automotive AI. By open-sourcing this technology, NVIDIA empowers the industry to fine-tune, adapt and enhance the model for safer autonomous vehicles, accelerating collective innovation and safety standards.
- ***The shift from software-defined to AI-defined vehicles is underway***
  NVIDIA highlights the industry transition from traditional software-defined vehicles (which rely on static, updatable code) to AI-defined vehicles that continuously learn and improve through reinforcement learning and advanced data workflows. This evolution demands new infrastructure and expertise, and NVIDIA is investing in comprehensive cloud-to-car solutions

to support this shift.

- ***Comprehensive ecosystem and platform strategy is core to NVIDIA's approach***
  NVIDIA's open-source strategy goes beyond just software. The company provides a full-stack solution, including hardware (Hyperion platform), datasets, synthetic data generation (Cosmos) and neural reconstruction tools, enabling partners to build, test and validate advanced autonomous systems. This ecosystem approach is seen as a larger opportunity than hardware or software sales alone.
- ***Enabling robust solutions for long-tail and low-frequency scenarios***
  NVIDIA addresses a major industry challenge: Handling rare or complex driving events that are not well-represented in real-world data. By combining reasoning models, synthetic scenario generation (Cosmos) and neural reconstruction, NVIDIA equips partners with the tools to train AI that can safely manage long-tail events, raising the bar for autonomous vehicle safety and reliability.

The following is an edited transcript of the conversation.

S&P Global Mobility: Could you tell us more about your open-source system? We would appreciate hearing about any recent progress or updates you can share with us.

Ali Kani: Safety in automotive applications is always top of mind because it's critical to design AI with safety and security as priorities. At NVIDIA, we implement all best practices in our DRIVE platform to ensure this. For example, our latest open-source model — Alpamayo — is noteworthy, because it incorporates reasoning capabilities. In automotive environments, you can't anticipate every possible scenario, so the model must be able to break down situations, evaluate potential actions and select the safest outcome. This reasoning ability allows the system to make correct decisions even when encountering scenarios it hasn't seen before.

Alpamayo is the world's first automotive vision-language-action model with chain-of-thought reasoning released to the industry, which is foundational for safe autonomous vehicles (AVs). We're proud to have open-sourced it, enabling others to adapt, fine-tune and improve the software with their own data.

Beyond the model itself, safety is addressed in design through features like safety monitors embedded within our models — similar to parity and ECC [Error-Correcting Code] protection in memory systems. Our AV stack includes tens of thousands of monitors to detect and respond to errors or corruption, ensuring the system ignores unreliable results.

System architecture also incorporates diversity for redundancy. Our stack features both an end-to-end model and a separate functionally safe stack, so two diverse algorithms work together with classical safety guardrails to ensure consistently safe decisions.

All these innovations reflect NVIDIA's significant investment in developing safe AI for automotive applications, and our ongoing commitment to lead in this area.

From your perspective, how has NVIDIA's AI-defined vehicle evolved over the past two years? What are the key differentiators between an AI-defined vehicle and a software-defined vehicle? Are these concepts in conflict, or do they work in parallel, like different cards in a deck?

The way we view this is rooted in the evolution from embedded software. Traditionally, the industry relied on embedded software — a base platform with code that could be updated over time, which

is the essence of a software-defined vehicle. However, AI introduces a significant shift: It's constantly learning and improving through data. With reinforcement learning, we monitor actions, reward positive outcomes and penalize negative ones, so the model continually gets better.

Building an AI-defined vehicle means creating a development loop designed for ongoing improvement, enhancing functionality, safety and security. Unlike conventional software, the AI model is self-learning, but it also requires a robust training workflow: From car to cloud, training and testing data, and quickly identifying and fixing issues.

That's why we invest heavily in training infrastructure, simulation computers, in-car computers and the seamless loop connecting them. This infrastructure is optimized to accelerate AI learning and minimize errors. While many companies can build software-defined computers, developing an AI-defined vehicle demands a completely different approach and expertise. That's what we're focused on — making this advanced infrastructure available across the ecosystem, from cloud to car.

In your opinion, does the concept of an AI-defined vehicle include agentic AI? How do you view the transition from generative AI to agentic AI within the automotive industry? I'm aware that companies like Google and Microsoft are already working on agentic AI with various industries. Does NVIDIA have any plans in this area as well?

Yes, I see this as a natural evolution. As you mentioned, it started with generative AI — models that generate new content, such as text or images. Agentic AI is the next step, where AI can autonomously think, answer questions, complete tasks and achieve goals.

Both generative and agentic AI are highly relevant for autonomous driving. For example, with our Cosmos transfer, we use generative AI to create and simulate scenarios for testing AV systems. That's an application of generative AI. The next step is agentic AI, where the model actively thinks and decides on actions, such as answering questions in the car. Beyond that, we move towards physical AI, which is essentially agentic AI applied to real-world actions, like making driving decisions in a vehicle.

All these elements are necessary to build a truly autonomous vehicle. Generative AI is essential for scenario generation, while agentic and physical AI models are critical for reasoning, planning and executing control actions. The evolution from generative to agentic to physical AI reflects the increasing sophistication required for self-driving technology.

How does NVIDIA view the role of open-source autonomous driving models? Is NVIDIA aiming to build an open-source platform for the entire industry? How do you see open source supporting and advancing the industry as a whole?

What I'd say is that NVIDIA's go-to-market approach has remained consistent over the years. Let me explain, and then I'll address your question directly.

If you recall our early days in gaming, we developed acceleration libraries like PhysX and CUDA, then partnered with game developers to integrate these libraries into their applications. This collaboration enabled developers to create outstanding games optimized for our platform, ultimately delivering better experiences to end customers and creating a positive feedback loop.

Today, our approach is similar, but the platform has evolved. In automotive, we've created the Hyperion platform — a comprehensive computer architecture and sensor suite that provides a foundation for development. This makes it easy for partners to build their own applications on top of our platform. Alongside Hyperion, we're offering libraries, models and reference code, and we're

open-sourcing many of these assets. This allows our partners to customize and improve the models for their specific needs — because, for example, a car from one brand should drive differently than one from another.

We're also releasing datasets, such as the physical AI dataset for autonomous driving that we released recently. By providing the platform, libraries, models and datasets, we're enabling the industry to innovate more efficiently and build safer, more advanced automotive applications.

Ultimately, our goal is to make it easier for partners to develop and enhance their own solutions, leading to safer products for the industry as a whole.

We would like to dive deeper into the Alpamayo-R1 model. I understand the model is referred to as VLA — vision language action. Why is it called VLA instead of VLM or a word model?

What I'd say is that while we might use different terminology, a "world model" is indeed important for a car — it allows for displaying the environment inside the vehicle, typically using models like BEV [battery-electric vehicle] transformers that detect both dynamic and static objects. Of course, we have that capability.

However, what we've announced is a vision-language-action (VLA) model with reasoning. If you want a model that truly understands what it's seeing and can then make decisions for automotive applications, VLA is the current state-of-the-art. For example, Tesla's Elon Musk has recently discussed adding reasoning to their end-to-end model, with plans to implement it soon. In [mainland] China, companies like XPeng and Li Auto are also moving towards vision language models with reasoning.

This is the direction the industry is heading, and that's why we released our VLA model — to provide the industry with a foundation to build upon, improve and accelerate progress. The world model is more closely tied to legacy sensor sets and architectures, and since it's already a common feature in the industry, we haven't released a separate version of it yet.

Using the [mainland Chinese] automotive industry as an example, they're leading software development in this area. I've noticed that current end-to-end models focus more on "what" rather than "why," largely due to limited reasoning capabilities. Another challenge is that these models struggle with low-frequency or long-tail scenarios, since they can only imitate data they've already seen. High-quality, diverse data is rare, with human driving data covering only about 0.1% of possible scenarios. Could you help us understand how reasoning in VLA models addresses these challenges? Specifically, how does reasoning VLA help solve the issues related to low-frequency and long-tail scenarios?

A reasoning model is effective for long-tail events because it can "think" through scenarios it hasn't seen before. Its architecture breaks down problems into steps, generates multiple possible solutions and selects the safest option, even for unfamiliar situations.

However, the model alone isn't enough. While real-world data is valuable, it's impossible to capture every scenario. That's why synthetic data is so important. Our Cosmos synthetic data generation engine lets you create specific scenarios you need more data for — like an uncontrolled left turn with multiple pedestrians. Cosmos Transfer, our foundational world model, generates a wide range of these long-tail scenarios, enabling your reasoning model to train on more diverse data and better handle rare events.

When you encounter a unique real-world scenario, our open-source neural reconstruction engine allows you to recreate it and add variations (such as different emergency vehicles with sirens). This helps further enrich your training data.

By combining Cosmos for synthetic scenarios, neural reconstruction for expanding real-world cases, and a reasoning model with safety guardrails, you have a robust approach to confidently address long-tail and low-frequency events in autonomous driving.

The feedback and solutions are really impressive.

Absolutely, no one else in the world is offering such a complete solution as NVIDIA. Others may provide software, but we go further. We offer datasets with scenarios, a synthetic data generation engine, a neural reconstruction engine to capture and augment issues, plus all the necessary libraries, AI models, tools, and datasets for safe development. No other company has invested as broadly in the physical AI market as NVIDIA.

From a PR perspective, Tesla claims they can achieve this, but it hasn't happened yet.

That's true, and their approach is different because they're building it solely for themselves.

You're enabling others to do it.

Exactly.

From an investment perspective, it takes significant resources and a long-term commitment — 5 to 10 years is not easy. NVIDIA's deep investment makes this possible. On the commercialization side, we know NVIDIA's foundation is in chips, both for data centers and vehicle hardware. Beyond hardware, there's software licensing and the full-stack solution revenue split. With AR1 being open source, it seems NVIDIA's monetization strategy is shifting from hardware and software sales to ecosystem value capture, which could boost industry adoption. Having strong hardware and software is important, but building a robust ecosystem — like Android, Llama or DeepSeek — is even more crucial for monetization. What's your view on this?

I agree — the platform and ecosystem are the bigger opportunity. Training these algorithms requires massive infrastructure, as seen with Tesla's public investment in 100,000 GPUs [graphics processing units]. If we provide a useful platform and algorithms that partners can fine-tune in the cloud, it represents a multihundred-billion-dollar opportunity. Supporting the ecosystem with AI libraries and tools is our largest strategic focus. However, it's important to note that we don't open source production-ready software for specific platforms; that requires a different business model and rigorous safety validation. Our open-source development models are meant for partners to adapt and make safe for their own products, while production software is provided directly to partners with necessary validation.

As the industry moves from Level 2+ to Level 4, NVIDIA has announced collaborations with Uber and several others. Is the NVIDIA-Uber partnership intended to support OEMs [original equipment manufacturers] in launching robo-taxi services? If so, which partners are currently engaged? For example, Mercedes-Benz's collaboration is focused on consumer vehicles, not robo-taxis. Which OEMs are working with NVIDIA on robo-taxi initiatives?

Great question. The Uber partnership involves several steps. First, we've defined the Hyperion

platform architecture for Level 4 and robo-taxi-ready vehicles. This platform ensures safety and redundancy, making it easy for Uber to standardize with OEMs and software partners. We've announced that NVIDIA, along with companies like Wayve, Waabi and WeRide, are developing software for this architecture. OEMs are also building robo-taxi-ready vehicles on Hyperion — Mercedes-Benz S-Class, Lucid, Nuro (using Lucid cars), and Stellantis are among those announced. We're engaging with more OEMs and AV software partners to grow the Hyperion ecosystem. Additionally, tier 1 suppliers like Magna, Continental (Aumovio) and Desay SV are building Hyperion-compatible components and vehicles. The ecosystem includes both software companies and OEMs, allowing flexibility in software development. Partners can develop their own, collaborate or work directly with NVIDIA.

CONTACTS

The Americas
+1 877 863 1306

Europe, Middle East & Africa
+44 20 7176 1234

Asia-Pacific
+852 2533 3565

www.spglobal.com/mobility