Cerence AI Expands Collaboration with NVIDIA to Advance its CaLLM Family of Language Models

January 3, 2025

Company will leverage the NVIDIA AI Enterprise software platform, including the NVIDIA TensorRT-LLM open-source library and the NVIDIA NeMo framework, to optimize performance

BURLINGTON, Mass., Jan. 03, 2025 (GLOBE NEWSWIRE) -- Cerence Inc. (NASDAQ: CRNC) (“Cerence AI”), a global industry leader in voice AI, today announced an expanded collaboration with NVIDIA to advance the capabilities of its CaLLM™ family of language models, including its cloud-based Cerence Automotive Large Language Model (CaLLM) and its CaLLM Edge embedded small language model. Through this collaboration, CaLLM is powered by NVIDIA AI Enterprise, an end-to-end, cloud-native software platform, and some aspects of CaLLM Edge are powered by NVIDIA DRIVE AGX Orin.

Integrating agentic frameworks with in-car conversations in both cloud and embedded forms requires a comprehensive, cross-disciplinary effort combining hardware, software, and UX domain expertise. Working alongside NVIDIA hardware and software engineers, Cerence AI enhanced its ability to meet production timelines and productize generative AI innovation for automotive. Specifically, Cerence AI has accelerated the development and deployment of CaLLM by leveraging the NVIDIA AI Enterprise software platform, including NVIDIA TensorRT-LLM and NVIDIA NeMo, an end-to-end framework to build, customize, and deploy generative AI applications into production. As a result, Cerence AI has optimized and customized its CaLLM family of models to:

- Deliver faster in-vehicle assistant performance on NVIDIA accelerated computing and SoCs
- Develop an automotive-optimized implementation of NVIDIA NeMo Guardrails, helping ensure Cerence-powered systems can navigate the nuances of in-car interaction
- Implement and optimize an agentic architecture on CaLLM Edge via NVIDIA DRIVE AGX Orin, helping advance the next generation of in-vehicle user experiences

Overall, this expanded collaboration with NVIDIA equips Cerence AI with scalable, reliable tools and resources to develop next-generation user experiences in partnership with its automaker customers. This, in turn, facilitates enriched driver experiences intended to deliver advanced performance, reduced latency, enhanced privacy and security, and robust protection against malicious or unwanted interactions.

“By optimizing the performance of our CaLLM family of language models, we are delivering cost savings and improved performance to our automaker customers, who are running quickly to deploy generative AI-powered solutions to their drivers,” said Nils Schanz, Executive Vice President, Product & Technology, Cerence AI. “As we advance our next-gen platform, with CaLLM as its foundation, these advanced capabilities will deliver faster, more reliable interaction to drivers, enhancing their safety, enjoyment and productivity on the road.”

“Large language models are offering vast, new user experiences, but complexities in size and deployment can make it difficult for developers to get AI-powered solutions into the hands of end users,” said Rishi Dhall, Vice President of Automotive, NVIDIA. “Through this expanded collaboration, Cerence AI is deploying advanced NVIDIA AI and accelerated computing technologies to optimize its LLM development and deployment.”

To learn more about Cerence AI, visit www.cerence.ai, and follow the company on LinkedIn.

About Cerence Inc.
Cerence Inc. (NASDAQ: CRNC) is a global industry leader in creating intuitive, seamless, AI-powered experiences across automotive and transportation. Leveraging decades of innovation and expertise in voice, generative AI, and large language models, Cerence powers integrated experiences that create safer, more connected, and more enjoyable journeys for drivers and passengers alike. With more than 500 million cars shipped with Cerence technology, the company partners with leading automakers, transportation OEMs, and technology companies to advance the next generation of user experiences. Cerence is headquartered in Burlington, Massachusetts, with operations globally and a worldwide team dedicated to pushing the boundaries of AI innovation. For more information, visit www.cerence.ai.

Kate Hickman | Tel: 339-215-4583 | Email: kate.hickman@cerence.com