Cerence Pioneers Automotive-Specific LLM in Collaboration with NVIDIA, Powering the Future of In-Car Experiences

December 19, 2023

Cerence has created an LLM-based platform leveraging its extensive automotive dataset and tech stack to deliver enhanced experiences for end users

BURLINGTON, Mass., Dec. 19, 2023 (GLOBE NEWSWIRE) -- Cerence Inc. (NASDAQ: CRNC), AI for a world in motion, today introduced a pioneering, automotive-specific large language model, CaLLM™, powered by NVIDIA technology. CaLLM™ (Cerence Automotive Large Language Model) serves as the foundation for Cerence's next-generation in-car computing platform, running on the NVIDIA DRIVE platform. Cerence is working with NVIDIA to solve several key automaker challenges: accelerating time to market by enabling new user experiences to be deployed via a cloud integration with both existing embedded systems and new generative AI-powered domains, giving OEMs an advantage as drivers demand increasingly intelligent features and capabilities inside the car.

As automakers and mobility OEMs look to leverage generative AI and LLMs to improve the user experience, CaLLM™ meets several critical needs. It has a unique level of automotive-specific intelligence, leveraging Cerence's extensive automotive expertise and fine-tuned, growing automotive dataset encompassing billions of tokens to deliver integrated in-car user experiences beyond those of general knowledge LLMs. CaLLM™ is distinctly able to support automotive functions, features, and requirements and is deeply customizable for automakers through training, fine-tuning, and bespoke applications. Its features include user personalization as well as built-in information retrieval for Cerence's generative AI-powered applications like Cerence Car Knowledge.

Cerence is building CaLLM™ with the NVIDIA AI foundry service, which includes NVIDIA AI Foundation Models, NVIDIA AI Enterprise software, and NVIDIA accelerated computing. Cerence will first train CaLLM™ on its extensive dataset leveraging NVIDIA DGX Cloud and NVIDIA DGX systems and then develop the capabilities required for in-car user experiences. Second, to help achieve an ultrafast user experience, Cerence will deploy CaLLM™ in an NVIDIA-accelerated infrastructure with NVIDIA AI Enterprise, an end-to-end, cloud-native, secure software platform that accelerates training, fine-tuning, and inferencing of LLMs with the NVIDIA NeMo framework and the NVIDIA TensorRT-LLM optimization library.

"We are delivering a breakthrough in-car computing platform that leverages our vertical tech stack to allow users to complete tasks across applications through a unified conversational interface," said Iqbal Arshad, Chief Technology Officer, Cerence. "When paired with NVIDIA's industry-leading AI frameworks and platforms, our exclusive dataset, deep relationships with automakers, and extensive market penetration put us in the ideal position to lead with our pioneering automotive-specific LLM, CaLLM™, and our own in-car computing platform. As we look to the future, we see this proprietary architecture also benefiting industry leaders outside of automotive."

"Generative AI and automotive LLMs are expected to enable new, intuitive, and personalized user experiences, capabilities, and features both inside and outside the car," said Ali Kani, Vice President of Automotive, NVIDIA. "By tapping NVIDIA's core expertise in cloud and edge technologies, Cerence can more quickly train, scale, and deploy these models within their in-car user-experience platform to enable safer, smarter, and more enjoyable rides."

CaLLM™ serves a critical role in building the cars of the future, serving as the foundation for Cerence's new in-car computing platform. This platform is the future of in-car interaction, an automotive- and mobility-specific assistant that provides an integrated in-cabin experience. While previous solutions required multi-step interactions to take place as distinct, separate steps – resulting in too many manual interactions and a high cognitive load for the driver – Cerence's new platform combines all aspects of a user's interaction into a seamless and intuitive conversational interface. Cerence's new platform will enable smooth transitions for OEMs by reimagining their existing applications within the conversational interface while offering new generative AI-powered services.

CaLLM™ and Cerence's new platform are already in proof of concept with leading global automakers and will be demonstrated at CES 2024 in Cerence's booth, 6627, in the West Hall. To book an appointment, reach out to ces.booking@cerence.com. To learn more about Cerence, visit www.cerence.com, and follow the company on LinkedIn and Twitter.

About Cerence Inc.
Cerence (NASDAQ: CRNC) is the global industry leader in creating unique, moving experiences for the mobility world. As an innovation partner to the world's leading automakers and mobility OEMs, it is helping advance the future of connected mobility through intuitive, AI-powered interaction between humans and their vehicles, connecting consumers' digital lives to their daily journeys no matter where they are. Cerence's track record is built on more than 20 years of knowledge and 475 million cars shipped with Cerence technology. Whether it's connected cars, autonomous driving, e-vehicles, or two-wheelers, Cerence is mapping the road ahead. For more information, visit www.cerence.com.

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

Source: Cerence Operating Company