Cerence Introduces Pioneering Embedded Small Language Model, Purpose-Built for Automotive

Novembre 13, 2024

CaLLM™ Edge is developed and optimized in collaboration with Microsoft, leveraging its Phi-3 family of small language models

BURLINGTON, Mass., Nov. 13, 2024 (GLOBE NEWSWIRE) -- Cerence Inc. (NASDAQ: CRNC), AI for a world in motion, today introduced CaLLM™ Edge, its pioneering, automotive-grade, embedded small language model (SLM). This new SLM is available within the existing Cerence solutions portfolio and will power its next-generation AI assistant platform, enabling an intelligent, seamless user experience regardless of connectivity. Developed and optimized in partnership with Microsoft, CaLLM Edge is available directly to Cerence's automaker customers as well as in the Microsoft Azure AI model catalog.

CaLLM (Cerence Automotive Large Language Model) Edge is fine-tuned on Microsoft's Phi-3 family of small language models, using Cerence's extensive automotive dataset to deliver highly specialized AI that can handle a variety of automotive use cases. With 3.8 billion parameters, 4k context size, and 4-bit quantization, this model fits comfortably when embedded in the automotive headunit. Its core capabilities include implicit and explicit car control commands (for example, temperature, windows and doors, seat position) and point-of-interest search and navigation, as well as conversational interaction like, "What's the most popular movie ever shot in Hollywood?" followed by "Can you tell me more about the plot?"

Compatible with major automotive platforms, CaLLM Edge is available in both embedded-only deployments, meaning it can function independently without any connectivity, as well as hybrid or cloud-first deployments in which the SLM serves as one method of answering queries and as backup when connectivity is lost. For users, this means always-on access to key generative AI-style features and information, even when not connected to the cloud, as well as improved data privacy, with data staying on board in the car rather than being sent to the cloud. For automakers, CaLLM Edge delivers not only improved assistant performance, but also cost efficiency – by leveraging a fully embedded SLM model, OEMs can keep costs under control while still delivering a generative AI-based experience for their drivers.

"CaLLM Edge fundamentally transforms the way users can interact with their systems, enabling them to access the rich, responsive experiences they've come to expect from cloud-based systems no matter where they are," said Nils Schanz, EVP, Product & Technology, Cerence. "As we continue to leverage our embedded and cloud expertise and advancements in generative AI and LLMs throughout our product architecture, we not only further expand the capabilities of our current solutions portfolio but also advance our next-gen platform. We're proud to partner with Microsoft to unite our deep vertical expertise and strength in embedded solutions with their industry-leading language model capabilities."

"Adapted AI models are becoming increasingly necessary in order to deliver intelligent, capable agent experiences across industries. We're pleased to work with Cerence to offer CaLLM Edge in the Azure AI model catalog, giving automotive organizations the opportunity to build AI solutions in Azure AI Studio and Microsoft Copilot Studio," said Satish Thomas, Corporate Vice President, Business & Industry Solutions, Microsoft. "Cerence's unique automotive dataset brings incredible value, enabling a high level of accuracy and relevance throughout the user interaction – all in an embedded architecture that means the experience is available at all times."

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

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 500 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.

Contact Information

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