SambaNova SYSTEMS

USE CASE

SambaNova Internal Chat with Custom Data

World record performance and accuracy for secure generative AI

Challenge

The need for internal chatbots with custom data has greatly increased within agencies of the U.S. government. Many agencies see the need for an internal chatbot to help improve efficiency, enhance user experience, and handle complex information within government operations. The need for custom data integration is crucial to ensure that these chatbots can provide accurate and secure services tailored to specific agency requirements.

Objective

SambaNova offers an AI starter kit with SambaNova Suite, this will allow you to build chat capabilities optimized and customized with your data. Within this AI starter kit your organization will have access to the most recent and performant models that can be fine tuned to create a chatbot that fits your organization needs.

Distinguishing Factors

- Model/Data Ownership
- Full Stack Solution
- On-site/Cloud Services
- Data Management
- Fine Tuned Expert Models
- Validation Testing Suite

Justification

SambaNova's platform comes with pre-trained models that can be customized and fine-tuned with private data to meet your needs. We designed SambaNova Suite in a way that simplifies the deployment and management of generative AI models. This reduces the complexity and cost associated with running these models on legacy systems, making it easier for integrating advanced AI capabilities into operations.

Solution

SambaNova Suite, combined with the SambaNova Composition of Experts (CoE) model, delivers the highest combination of performance and accuracy for generative AI. A complete hardware/software platform, SambaNova Suite can be deployed as a fully configured rack-level solution either on-premises, including air-gapped environments, or as a cloud-based solution. The CoE model provides state-of-the-art accuracy across a wide range of use cases.

Key Features

Performance, Accuracy, and Efficiency
Only SambaNova delivers performance of over 1000 tokens/s at full precision, on as few as 16 sockets. This is an unrivaled combination of performance and accuracy with a small footprint, dramatically reducing power consumption.

Total AI Stack Control
With options for on-premises and air-gapped deployments, SambaNova ensures complete data privacy and security, enabling the use of sensitive data for model training without external exposure.

Model Ownership
Once a model is fine-tuned, it becomes the property of the customer in perpetuity, eliminating model ownership concerns.

Configurable Access and Permissions
The CoE model enables flexible access controls, allowing precise management of model permissions to align with organizational roles, enhancing security and ensuring compliance.

SambaNova Suite delivers the most accurate, integrated full stack generative AI platform, optimized for enterprise and government organizations. Delivered through SambaNova's integrated AI platform, which can be deployed on-premises or through the cloud, SambaNova Suite seamlessly integrates into existing business processes to deliver transformative capabilities. With the capacity to be further refined using customer data, these models can deliver unparalleled accuracy while providing enterprise grade security and data governance.

Learn more