SambaNova SYSTEMS

USE CASE

SambaNova Enterprise Knowledge Retrieval

World record performance and accuracy for secure generative AI

Challenge

Like many large organizations, the U.S government faces challenges when it comes to enterprise knowledge management and retrieval. Some of the current challenges include information overload, siloed knowledge, and efficient search methods.

Objective

Many of the challenges associated with knowledge retrieval can be solved through the use of artificial intelligence. AI systems can deliver tailored information to government employees based on their roles which ensures they have access to the most relevant knowledge for their job. AI-powered search algorithms can understand context and intent, providing more accurate and relevant results. This can significantly improve the efficiency of information retrieval.

Distinguishing Factors

- Model/Data Ownership
- Full Stack Solution
- On-site/Cloud Services
- API based application intelligence
- Web Parsing
- Embedding and Retrieval

Justification

SambaNova offers an "Enterprise Knowledge Retriever" as part of their AI starter kit. This allows for the implementation of semantic search workflow, meaning they can search based on context instead of specific keywords. This improves AI enabled systems by grounding results from relevant and current sources. SambaNova Suite also has role-based access control which will allow users to easily retrieve knowledge that is specific to their role and responsibility.

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