SambaNova® SYSTEMS

PRODUCT SHEET

Samba-1

Samba-1: 1 Trillion Parameters, One Model, One Platform - All Your Apps

Features:

- 1.3 Trillion parameter Composition of Experts.
- Made up of best in class base models including Llama, Mistral, Falcon, DeepSeek, Llava
- 92 state of the art experts in a wide range of tasks, domains and modalities.
- Sophisticated routing techniques enable users to automatically route to the best expert and developers to create their own routing strategies.

Benefits:

- Secure. Your own dedicated Samba-1 instance that can be fine tuned on your own private data.
- Affordable at enterprise scale. 10X cost and power reduction for production scale inference versus similar solutions.
- Accurate for business tasks. Complex business tasks require systems that understand detailed and diverse content and queries.
- Broadly applicable. Able to handle the range and diversity of business tasks.
- Manageable. A full stack solution, and over a trillion parameters deployed as one endpoint.
- Open standards. Own and control your model, providing governance and avoid vendor lock in.

What is Samba-1

Samba-1 is a pioneering 1 trillion parameter model designed to revolutionize enterprise AI. A high-performing Composition of Experts that combines the comprehensive power of trillion-parameter models with the precision and efficiency of specialized models. This unique model offers a single API endpoint, enabling the orchestration of domain-specific experts across various fields such as finance, legal, and engineering, along with task-specific experts for operations like summarization, extraction, and content editing across multiple modalities including text, code, and images. Samba-1 facilitates a seamless integration of these diverse models, providing configurable access and permissions that reflect an organization's structure, thus ensuring data segregation and security.

Features of Samba-1

Leverage diverse models from one endpoint: Utilize the single API endpoint to create knowledge-rich applications with a wide range of domain and task experts across different configurations and modalities, including text, code, and images. Easily integrate these experiments into your applications for real-world use.

Customize and fine-tune models for specific tasks: Tailor Samba-1's Composition of Experts model with your own data for fine-tuning, ensuring the model aligns perfectly with your specific business requirements and enhances performance on targeted tasks.

Develop applications with advanced reasoning capabilities: Harness the power of Samba-1's automated routing to build applications that can make complex decisions, perform tasks, and provide solutions tailored to customer needs by combining expert models with API calls and data queries.

Configurable Access and Permissions: Samba-1 enables flexible access controls, allowing precise management of model permissions to align with organizational roles, enhancing security and ensuring compliance.

Total AI Stack Control: With options for on-premise and air-gapped deployments, Samba-1 ensures complete data privacy and security, enabling the use of sensitive data for model training without external exposure.

Optimize application performance and efficiency: Benefit from the efficiency and lower operational costs of running Samba-1 in SambaStudio on the SN40L system. Enjoy 10x reduction in inference cost and power consumption over alternative solutions.

Llama 2			Mistral
Swallow-7b-instruct-v0.1	SambaLingo-70b-Hungarian-Chat	NexusRaven-V2-13B	sqlcoder-7b
Swallow-13b-instruct-v0.1	SambaLingo-70b-Arabic-Chat	llama-2-13b-chat-hf	BioMistral-7B
Swallow-70b-instruct-v0.1	SambaLingo-70b-Thai-Chat	Xwin-Math-13B-V1.0	BioMistral-7B-DARE
Swallow-7b-NVE-instruct-hf	SambaLingo-Bulgarian-Chat	llama-2-13b-hf	BioMistral-7B-TIES
Swallow-70b-NVE-instruct-hf	SambaLingo-Japanese-Chat	Xwin-Math-70B-V1.0	BioMistral-7B-SLERP
ELYZA-japanese-llama-2-7b	SambaLingo-Thai-Chat	deepseek-llm-7b-chat	Saul-Instruct-v1
Nous-Hermes-llama-2-7b	SambaLingo-Arabic-Chat	deepseek-llm-67b-chat	WestLake-7B-v2-laser-truthy-dpo
Nous-Hermes-Llama2-13b	SambaLingo-Hungarian-Chat	tulu-2-7b	EmertonMonarch-7B
CodeLlama-70b-Python-hf	SambaLingo-Russian-Chat	tulu-2-dpo-7b	typhoon-7b
CodeLlama-70b-Instruct-hf	SambaLingo-Slovenian-Chat	GOAT-70B-Storytelling	Genstruct-7B
Magicoder-S-DS-6.7B	SambaLingo-Turkish-Chat	llama-2-70b-chat-hf	Mistral-7B-OpenOrca
Magicoder-S-DS-6.7B-16k	SambaLingo-Serbian-Chat	llama-2-70b-hf	OpenHermes-2p5-Mistral-7B
sqlcoder-7b-2	medicine-chat	tulu-2-dpo-13b	Nous-Hermes-2-Mistral-7B-DPO
sqlcoder-70b-alpha	Xwin-Math-7B-V1.0	autoj-13b	Starling-LM-7B-beta
llama-2-7b-chat-hf	TableLlama	tulu-2-13b	Mistral-7B-Instruct-v0.2
law-chat	lumos_web_agent_ground_iterative	tulu-2-dpo-70b	zephyr-7b-beta
finance-chat	lumos_web_agent_plan_iterative	tulu-2-70b	Mistral-T5-7B-v1
Explore-LM-7B-Rewriting	SambaCoder-nsql-llama-2-70b	UniNER-7B-all	Lil-c3po
llama-2-7b-hf		nsql-llama-2-7b	v1olet_merged_dpo_7B
LlamaGuard-7b			Rabbit-7B-DPO-Chat
			DonutLM-v1
			Snorkel-Mistral-PairRM-DPO

Llama 3	Falcon	DeepSeek	BLOOMChat	Gemma
Meta-Llama-3-8B	Falcon-40b-instruct	deepseek-coder-1.3b-instruct	BLOOMChatv2-8k	codegemma-7b
Meta-Llama-3-8B-Instruct		deepseek-coder-6.7b-instruct	BLOOMChatv2-2k	codegemma-7b-it
Meta-Llama-Guard-2-8B		deepseek-coder-33b-instruct		gemma-7b
Meta-Llama-3-70B		Deepseek-llm-7b-chat		gemma-7b-it
Meta-Llama-3-70B-Instruct		Deepseek-llm-67b-chat		
Meta-Llama-3-70B-Instruct				

Inference Performance

Samba-1 can be deployed on a single SN40L node, while other systems would need many nodes to run a model of this size.

*Consult your SambaNova point of contact for assistance on sizing.

SambaNova Suite & Samba-1

To learn more about how SambaNova Systems can accelerate and transform your organization with generative AI, **schedule a meeting.**

Learn more at SambaNova.AI

linkedin.com/company/sambanova

@SambaNovaAI

info@sambanova.ai

Customers turn to SambaNova to quickly deploy state-of-the-art AI capabilities to meet the demands of the AI-enabled world. Our purpose-built enterprise-scale AI platform is the technology backbone for the next generation of AI computing. We enable customers to unlock the valuable business insights trapped in their data. Our flagship offering, SambaNova Suite, overcomes the limitations of legacy technology to power the large complex foundation models that enable customers to discover new services and revenue streams, and boost operational efficiency. Headquartered in Palo Alto, California, SambaNova Systems was founded in 2017 by industry luminaries, and hardware and software design experts from Sun/Oracle and Stanford University. Investors include SoftBank Vision Fund 2, funds and accounts managed by BlackRock, Intel Capital, GV, Walden International, Temasek, GIC, Redline Capital, Atlantic Bridge Ventures, Celesta, and several others.