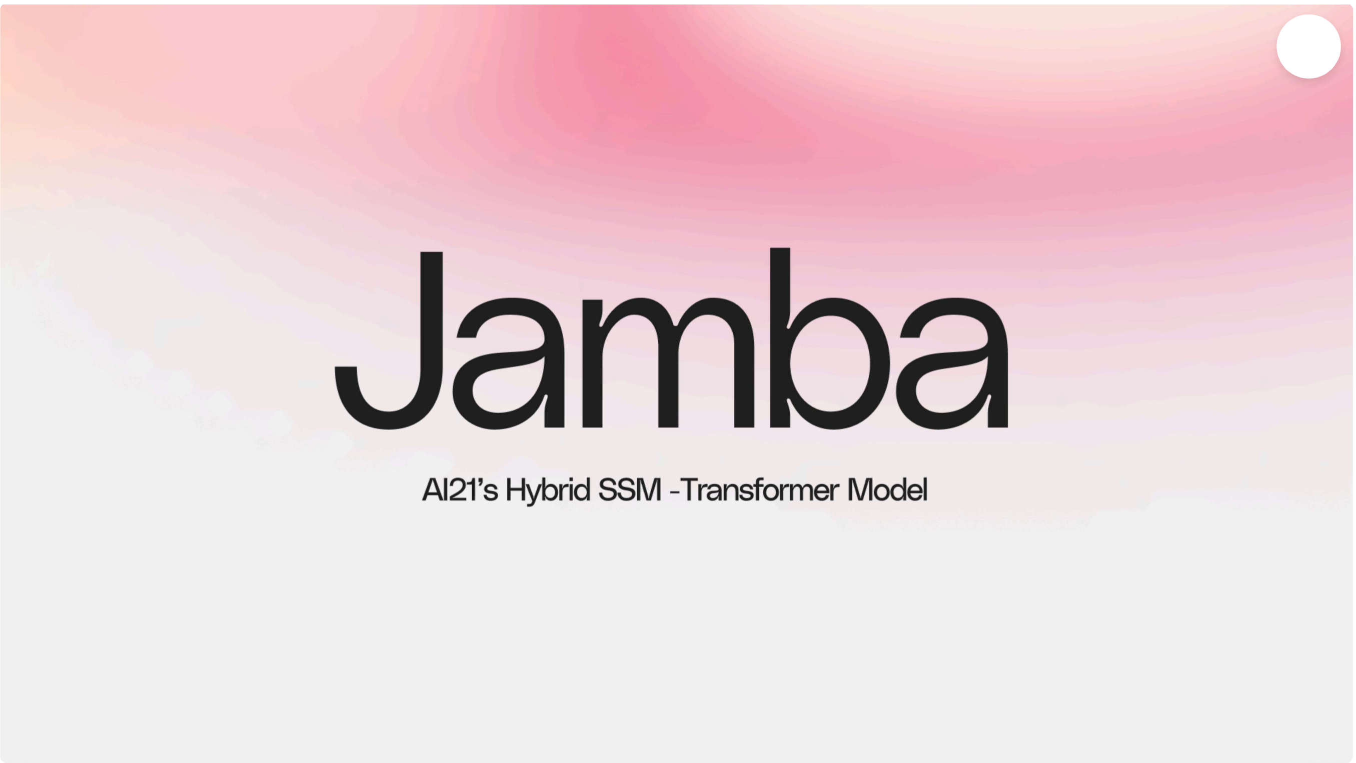


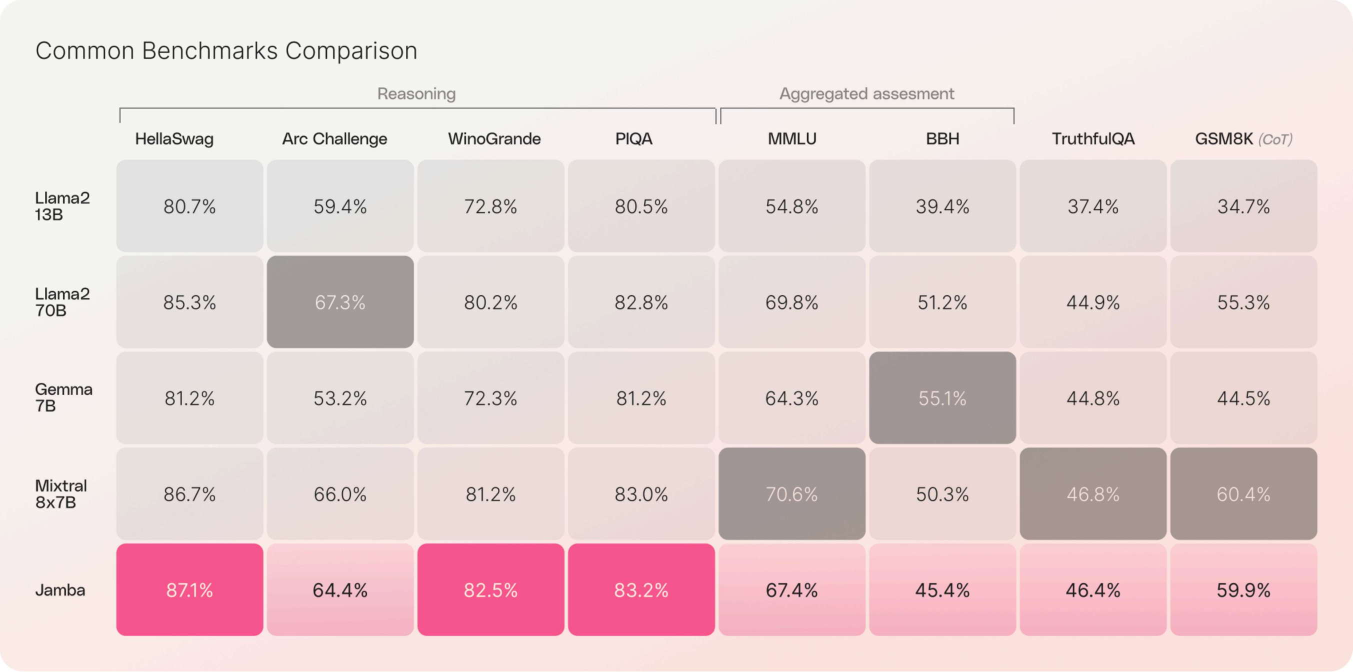
Announcements

Introducing Jamba: AI21's Groundbreaking SSM-Transformer Model

Debuting the first production-grade Mamba-based model delivering best-in-class quality and performance.

March 28, 2024

We are thrilled to announce Jamba, the world's first production-grade Mamba based model. By enhancing Mamba Structured State Space model (SSM) technology with elements of the traditional Transformer architecture, Jamba compensates for the inherent limitations of a pure SSM model. Offering a 256K context window, it is already demonstrating remarkable gains in throughput and efficiency—just the beginning of what can be possible with this innovative hybrid architecture. Notably, Jamba outperforms or matches other state-of-the-art models in its size class on a wide range of benchmarks.



Common Benchmarks Comparison

	Reasoning				Aggregated assesment			
	HellaSwag	Arc Challenge	WinoGrande	PIQA	MMLU	BBH	TruthfulQA	GSM8K *(CoT)*
Llama2 13B	80.7%	59.4%	72.8%	80.5%	54.8%	39.4%	37.4%	34.7%
Llama2 70B	85.3%	67.3%	80.2%	82.8%	69.8%	51.2%	44.9%	55.3%
Gemma 7B	81.2%	53.2%	72.3%	81.2%	64.3%	55.1%	44.8%	44.5%
Mixtral 8x7B	86.7%	66.0%	81.2%	83.0%	70.6%	50.3%	46.8%	60.4%
Jamba	87.1%	64.4%	82.5%	83.2%	67.4%	45.4%	46.4%	59.9%

In releasing Jamba with open weights, licensed under Apache 2.0, we invite further discoveries and optimizations that build off this exciting advancement in model architecture. We can't wait to see what you'll build.

Jamba will also be accessible from the NVIDIA API catalog as NVIDIA NIM inference microservice, which enterprise applications developers can deploy with the NVIDIA AI Enterprise software platform.

Key Features

- First production-grade Mamba based model built on a novel SSM-Transformer hybrid architecture
- 3X throughput on long contexts compared to Mixtral 8x7B
- Democratizes access to a massive 256K context window
- The only model in its size class that fits up to 140K context on a single GPU
- Released with open weights under Apache 2.0
- Available on Hugging Face and coming soon to the NVIDIA API catalog

Jamba Offers the Best of Both Worlds

Jamba's release marks two significant milestones in LLM innovation: successfully incorporating Mamba alongside the Transformer architecture and advancing the hybrid SSM-Transformer model to production-grade scale and quality.

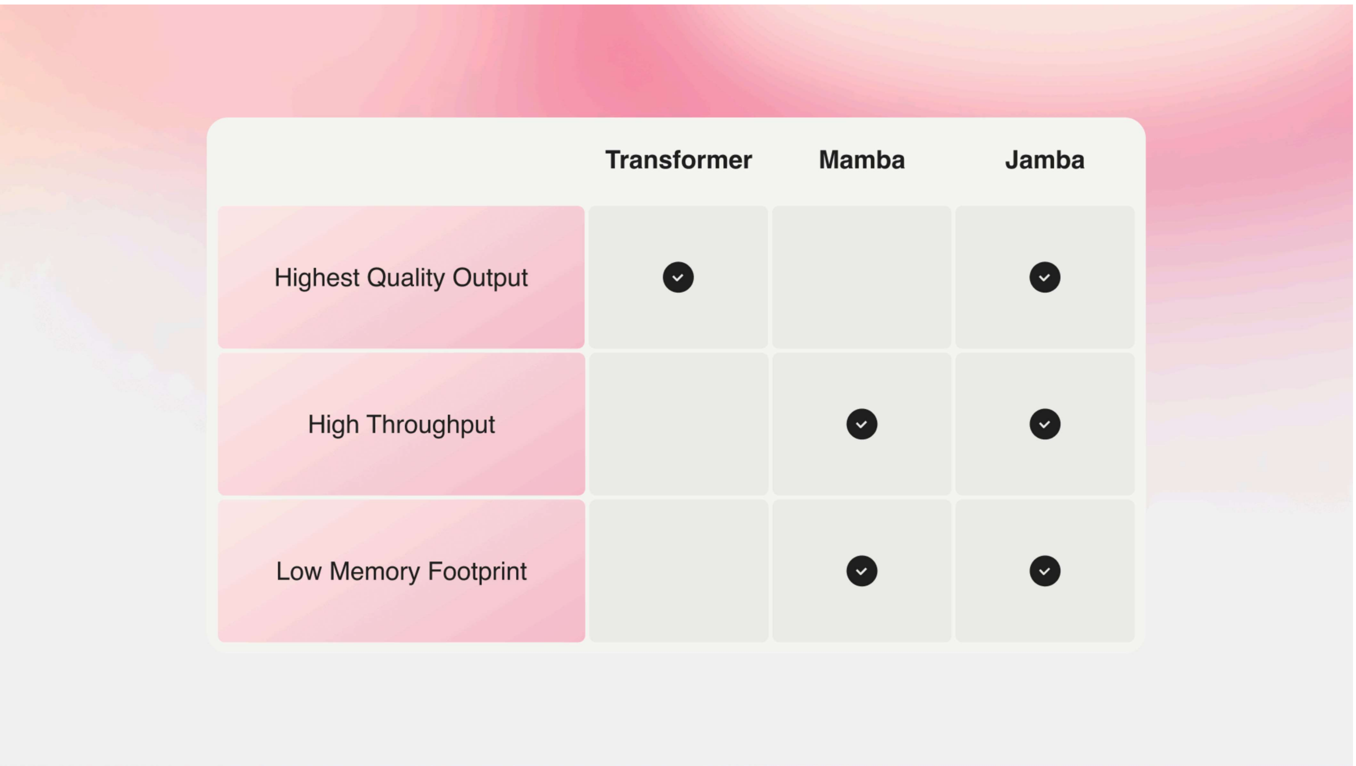
Until now, LLMs have been primarily built on the conventional Transformer architecture. While undoubtedly powerful, this architecture presents two major drawbacks:

- **Large memory footprint:** Transformer's memory footprint scales with context length. This makes it challenging to run long context windows or numerous parallel batches without extensive hardware resources, limiting widespread opportunities to experiment and deploy.

- **Slow inference as context grows:** Transformer's attention mechanism scales quadratically with sequence length and slows down throughput, as each token depends on the entire sequence that came before it—placing long context use cases outside the scope of efficient production.

Proposed by researchers at Carnegie Mellon and Princeton Universities, Mamba addresses exactly those shortcomings, opening a new frontier of possibility for language model development. However, without attention over the entire context, this architecture struggles to match the same output quality of the best existing models, especially on recall-related tasks.

To capture the best that both Mamba and Transformer architectures have to offer, we developed the corresponding Joint Attention and Mamba (Jamba) architecture. Composed of Transformer, Mamba, and mixture-of-experts (MoE) layers, Jamba optimizes for memory, throughput, and performance—all at once.



	Transformer	Mamba	Jamba
Highest Quality Output	✓		✓
High Throughput		✓	✓
Low Memory Footprint		✓	✓

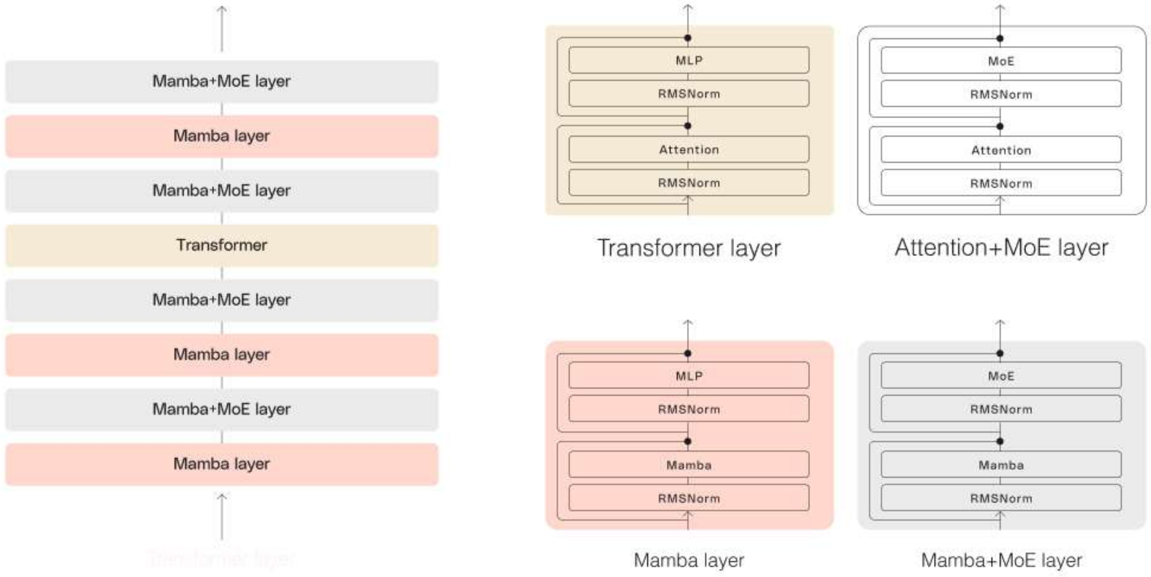
Jamba's MoE layers allow it to draw on just 12B of its available 52B parameters at inference, and its hybrid structure renders those 12B active parameters more efficient than a Transformer-only model of equivalent size.

While some have experimented with scaling Mamba, none have scaled it beyond 3B parameters. Jamba is the first hybrid architecture of its kind to reach a production-grade scale.

Building for Scale with Jamba Architecture

Several core architectural innovations were required to successfully scale Jamba's hybrid structure.

As depicted in the diagram below, AI21's Jamba architecture features a blocks-and-layers approach that allows Jamba to successfully integrate the two architectures. Each Jamba block contains either an attention or a Mamba layer, followed by a multi-layer perceptron (MLP), producing an overall ratio of one Transformer layer out of every eight total layers.



(a) Jamba block

(b) Different types of layers defined by Jamba architecture

Diagram showing (a) a single Jamba block, (b) Different types of layers. Jamba implementation includes 4 Jamba blocks, each containing 8 layers, a 1/7 ratio of attention/Mamba layers, and MoE applied every 2 layers

The second feature is the utilization of MoE to increase the total number of model parameters while streamlining the number of active parameters used at inference—resulting in higher model capacity without a matching increase in compute requirements. To maximize the model's quality and throughput on a single 80GB GPU, we optimized the number of MoE layers and experts used, leaving enough memory available for common inference workloads.

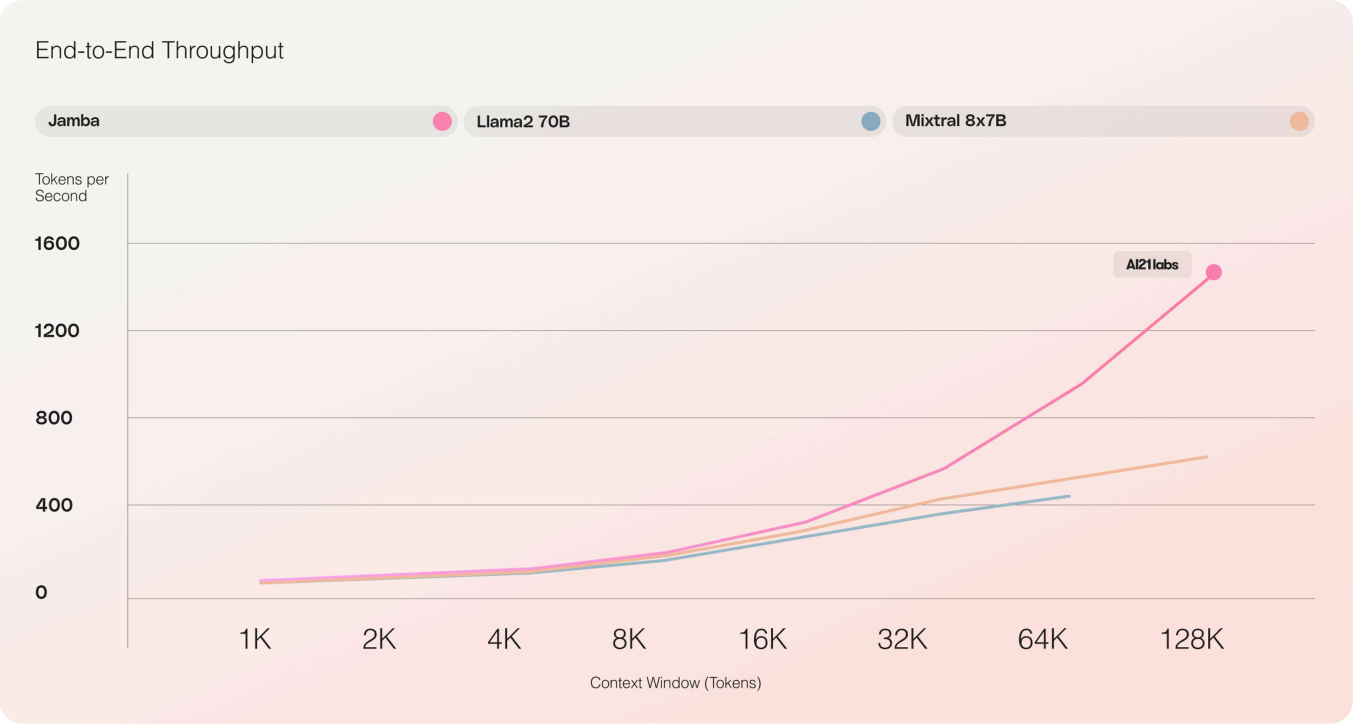
For further details on Jamba's novel architecture, read the full whitepaper.

Unprecedented throughput and efficiency

Based on our initial evaluations, Jamba is excelling across key measurements, such as throughput and efficiency. While its preliminary performance has already hit impressive milestones, we’re excited to see how these benchmarks will only continue to improve as the community pushes this new technology further through experimentation and optimization.

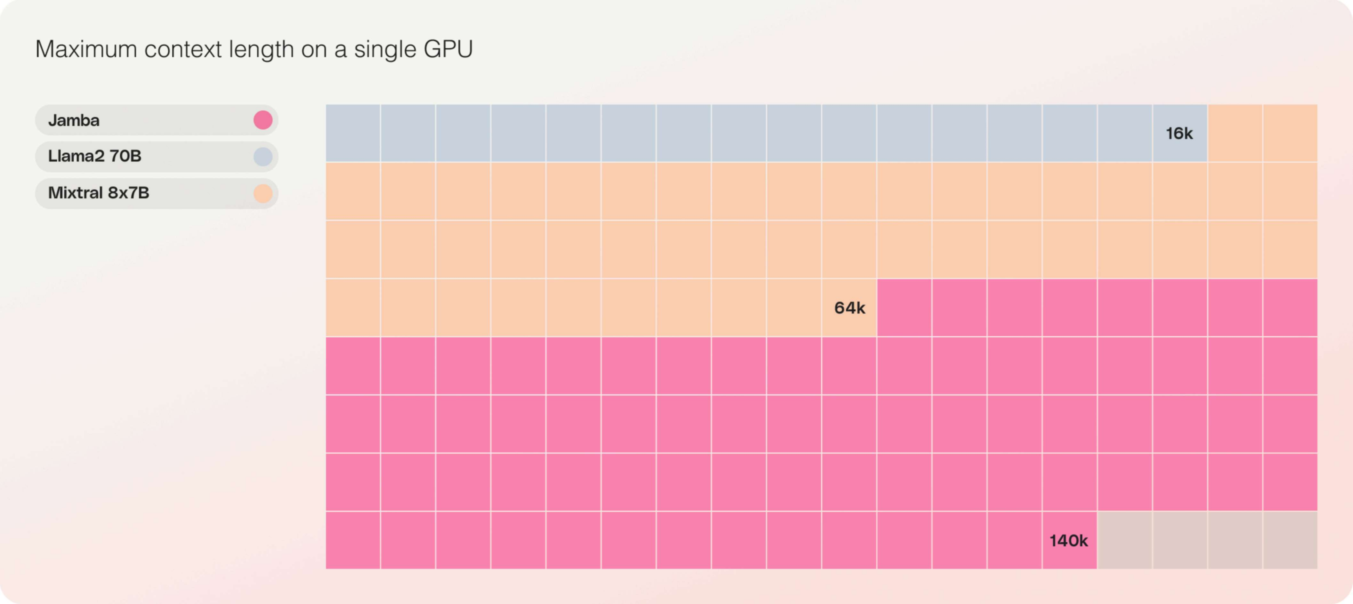
Efficiency

Delivers 3x throughput on long contexts, making it a more efficient model than Transformer-based models of comparable size like Mixtral 8×7B.



Cost

Jamba can fit 140K context on a single GPU, enabling more accessible opportunities for deployment and experimentation than currently available with other open source models of a similar size.



We expect these already encouraging gains to be further enhanced with future optimizations, such as better MoE parallelism, faster Mamba implementations, and more.

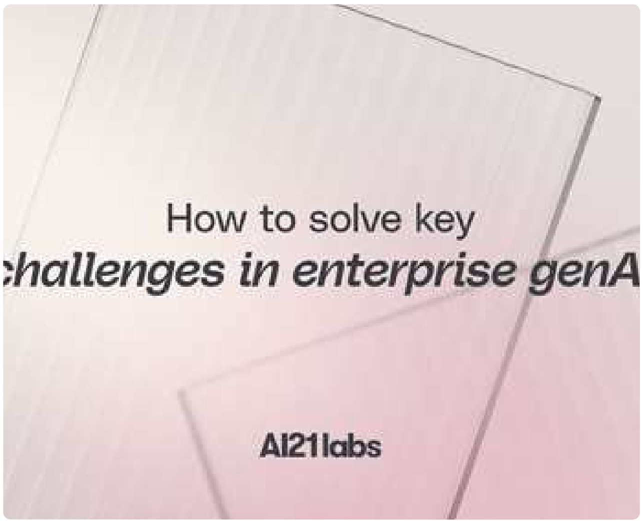
Start building with Jamba

You can start working with Jamba on Hugging Face. As a base model, Jamba is intended for use as a foundation layer for fine tuning, training, and developing custom solutions and guardrails should be added for responsible and safe use. An instruct version will soon be available in beta via the AI21 Platform. To share what you're working on, give feedback, or ask questions, join the conversation on Discord.

AI21 builds reliable, practical, and scalable AI solutions for the enterprise. To learn how genAI solves key business challenges, schedule time with an expert.

Email address Subscribe

Related Articles



Blog posts

Enterprise Generative AI: Key Challenges and How to Solve Them

Learn about some of the key challenges facing organizations when deploying enterprise LLM, and how mature LLM implementations address these challenges.

Blog posts

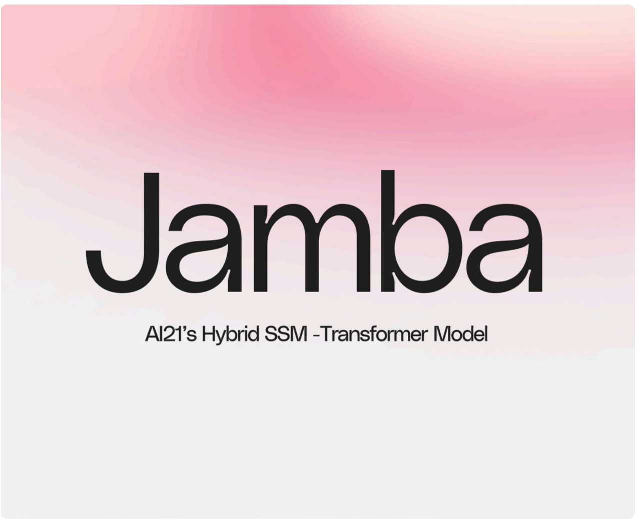
Enterprise GenAI: Definition, Challenges, and Solutions

Learn about the concerns that organizations are facing today as AI adoption increases, and how enterprise AI is evolving to meet them.

Blog posts

Top Ten GenAI Enterprise Use Cases

How do companies use Generative AI as a strategic tool for tackling real business challenges? Start with the problems today's GenAI applications can solve. Scoping high-impact use cases helps overcome common roadblocks facing enterprise teams. It's also a practical start for a smooth GenAI integration.



Announcements

Introducing Jamba: AI21's Groundbreaking SSM-Transformer Model

Debuting the first production-grade Mamba-based model delivering best-in-class quality and performance.

AI21 labs