AI21 labs Google Cloud

AI21's Jamba 1.6 Large

Building success on Google Cloud

Biggest GenAI Concerns for Enterprises

Accuracy & Quality Answers (Low Hallucination)

GenAI is only useful if it provides high quality, accurate output with a low risk of misinformation.

Data Protection & Privacy

GenAI applications must satisfy stringent IT security and data policies given regulatory, compliance and privacy concerns.

Cost Effectiveness & Low TCO

The deployment of LLMs can be intricate, necessitating significant technical expertise and robust infrastructure.

Enterprise Grade Support

Given how new GenAI is, the ability to reach a human being for help when critical issues hit is a must.

AI21labs Google Cloud

Built for the enterprise: Powerful and privately deployable long-context models

Accuracy & Quality Answers (Low Hallucination)

- Quality answers with fast speed and low latency for long context input enterprise use cases (> 100K tokens).
- Ranks among best in low hallucination rate.

Data Protection & Privacy

- Can be deployed via private deployment in a cloud VPC or on-premise to meet stringent regulatory, compliance and privacy requirements.

Cost Effectiveness & Low TCO

- Groundbreaking Mamba+Transformer architecture allows for compelling token prices and efficient VM usage, leading to overall lower cost and TCO.

Enterprise Grade Support

- Multiple enterprise-level support packages.
- AI21 is valued by customers for its white glove support, all the way from pre-sales to post-implementation.

AI21labs Google Cloud

3 Reasons Why Financial Services Companies Choose AI21labs

1. **Novel model architecture** delivers exceptional value, especially on long contexts.

2. **More secure deployment options** Including Google Cloud's Vertex AI and private VPC deployment.

3. **Strategic partnership and high-touch support** from pre-sales to post-implementation

AI21's Jamba 1.6 Large offers financial services companies the best all-around LLM that is commercially supported and can be deployed privately via VPC, efficiently and reliably handling your longest context use cases.

And the perfect choice for your longest context use cases.

AI21labs Google Cloud

Powerful & Efficient Long Context Models: Built for the Financial Space

Lightning speed

- Speed-optimized hybrid architecture.
- Low latency.
- Quick interaction for rapid turnaround.
- Accelerate traditionally labor-intensive tasks e.g. capital markets research rapidly.

Long context, short wait

- Efficiently process long contexts without compromising on quality and accuracy
- Handles increasing task complexity without compromising performance.
- Minimizes hallucinations.

Built-in security with Google Cloud

- Advanced data encryption.
- Rigorous access controls.
- Able to be deployed privately for organizations with stringent data security requirements e.g. those who handle financial reporting and compliance.

Bank-breaking not required

- Hybrid architecture provides scalable options for enterprise, from chatbots to complex financial analysis.
- Efficient processing to keep costs in check.
- High performance without the high price tag.

AI21labs Google Cloud

True Data Clarity with Jamba 1.6 Large

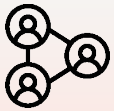
Unmatched accuracy and performance to handle all your financial services needs.

Processing Capabilities

- High-quality data analysis, technical documentation and long-form content creation.
- Easily process lengthy documents like loan applications, legal agreements and financial reports, and integrate/analyze data from diverse sources.

Speed and Efficiency

- Real time insights to accelerate decision making on investment and risk assessment.
- Optimize customer onboarding, loan processing and claims management to yield faster turnaround times and improved efficiency.



Organization-Wide Benefits

- Improved customer support, with personalized financial advice, more responsive chatbots and faster service resolutions.
- Automations reduce operational costs and lead to cost savings in various areas.

High Memory, Low Latency

- Access and “remember” a customer’s entire interaction history to provide personalized service and more relevant recommendations.
- Analyze streaming data, like market trends and financial news, to provide real-time insights and alerts.

AI21labs Google Cloud

The Google Cloud Difference: Better Together

Performance

Whether it's handling real-time interactions or processing large volumes of data, the Jamba Model Family on Google Cloud offers the flexibility that is required for diverse enterprise applications.

Scalability

Scale applications with minimal infrastructure concerns by taking advantage of the Jamba Model Family's on Vertex AI's fully managed infrastructure. This ensures scalability with ease.

Affordability

Google Cloud customers seeking advanced AI capabilities without the high cost will find the Jamba Model Family's cost-effective design makes it a true leader in the very fast-moving AI landscape.

Private Deployments

The Jamba Model Family's options for private deployment in a cloud VPC or on-premise, plus Google Cloud's comprehensive security, provide peace of mind for businesses handling confidential information.

AI21labs Google Cloud

Build with the Jamba 1.6 Model on Google Cloud's Vertex AI

Experiment, customize and deploy on Vertex AI to keep pace with innovation

Build & Evaluate

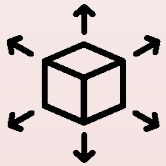
- With simple API calls or the Gen AI Evaluation Service offered through Vertex AI.
- All in the intuitive Vertex AI environment.

Optimize

- Scale Jamba Model Family applications using Google Cloud's AI-optimized infrastructure.
- Manage costs with pay-as-you-go pricing and auto-scaling to meet enterprise demands.

Deploy

- Deploy confidently with robust security, data privacy, and compliance.
- Protect your data and AI applications with Google Cloud's comprehensive security features.



Craft

- Create and orchestrate agents powered by the Jamba Model Family using Vertex AI's comprehensive set of tools.

AI21labs Google Cloud

No task is too complex when you have powerful, accurate, and more secure AI tools at your fingertips.

Jamba 1.6 Large
on Google Cloud is more secure and efficient.

AI21labs Google Cloud

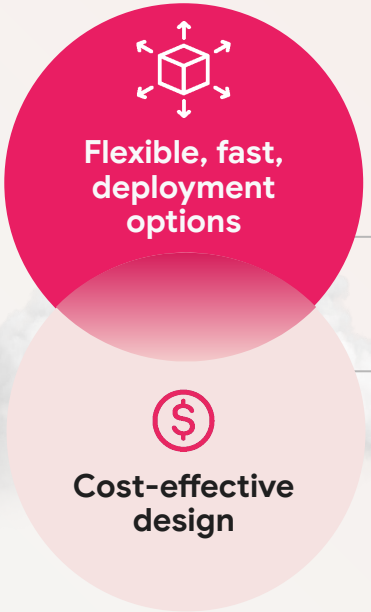
Performance Without Compromise

With Jamba 1.6 Large's standout long-context capabilities, you can manage large-scale, complex tasks with ease and accuracy

- 256K context window to handle large amounts of data
- Ultra-efficient Mamba-Transformer MoE architecture.
- Developer-friendly features (function calling, JSON mode output, document objects, citation mode) come standard.
- Maintains outstanding quality at lightning speeds, even as context grows.

More Secure Deployment that Suits Your Enterprise

Meet even the most stringent data security standards, without breaking the bank



- Deploy confidently with Google Cloud's robust security, data privacy and compliance.
- Scale your applications in production with Google Cloud's fully managed infrastructure.
- Benefit from flexible pay-as-you go pricing while meeting enterprise-level demands.
- Hands-on management for enterprises with bespoke needs.

Thank You

AI21 labs Google Cloud