MongoDB Sets a New Standard for Retrieval Accuracy with Voyage 4 Models for Production-Ready AI Applications

January 15, 2026

Tavily and TinyFish among customers using MongoDB to build and scale AI-powered features and workloads.

SAN FRANCISCO, Jan. 15, 2026 /PRNewswire/ -- MongoDB, Inc. (NASDAQ: MDB) today announced an industry-first expansion of its AI capabilities at MongoDB.local San Francisco, bringing together its core database with Voyage AI's world-class embedding and reranking models to deliver a unified data intelligence layer for production AI. By integrating these models directly into MongoDB's platform infrastructure, developers can now build and operate sophisticated applications at scale with reduced risk of hallucinations, without the need to move or duplicate data.

To support developers moving AI applications into production, MongoDB introduced a set of new AI capabilities designed to simplify how intelligent applications are built and operated. The company unveiled five embedding models from Voyage AI, MongoDB's embedding and retrieval model suite, Automated Embedding for MongoDB Community Vector Search, embedding and reranking AI model APIs in Atlas, and an AI-powered data operations assistant for MongoDB Compass and Atlas Data Explorer. These capabilities strengthen MongoDB's position as the leading AI-ready data platform, trusted by more than 60,000 customers running mission-critical workloads. Voyage AI models are available through MongoDB Atlas via API, integrated with MongoDB Community through managed Automated Embedding, and remain fully available as a standalone platform independent of MongoDB.

"The biggest challenge customers face with AI isn't experimentation, it's operating reliably at scale," said Fred Roma, Senior Vice President of Product and Engineering at MongoDB. "Developers want fewer moving parts and clearer paths from prototype to production. With today's launches, MongoDB is raising the bar, helping teams reduce complexity and focus on building AI applications that perform in real-world, mission-critical environments."

Transforming data into AI intelligence
As projects move into production, many organizations are discovering that their existing data stacks were never designed to support context-aware, retrieval-intensive workloads at scale. Developers are left managing fragmented combinations of operational databases, vector stores, and model APIs, which introduces complexity, latency, and operational risk at the exact moment speed and reliability matter most. This fragmentation has become a primary barrier to AI innovations, translating into real customer-facing impact.

MongoDB addresses this by unifying the core capabilities needed to build and run AI applications in production in a single data platform. Instead of stitching together an operational database, a vector store, and multiple pipelines, teams can keep operational data and retrieval capabilities together, reducing latency and synchronization overhead. The result is a simpler architecture, faster iteration, and AI applications that are built to run reliably in production, not just in demos. New capabilities include:

- **State-of-the-art accuracy with models from Voyage AI:** The general availability of the new Voyage 4 series continues giving developers high performing embedding models—which outperform Gemini and Cohere on the public RTEB leaderboard—for more accurate retrieval at lower cost. The Voyage 4 series includes the general-purpose voyage-4 embedding model, which strikes a balance between retrieval accuracy, cost, and latency, the flagship voyage-4-large model for the highest retrieval accuracy, voyage-4-lite for optimized latency and cost, and an open-weights voyage-4-nano for local development and testing, or on-device applications.
- **Facilitated context extraction from video, images, and text:** The general availability of the new voyage-multimodal-3.5 model expands support for interleaved text and images to now include video. Voyage AI's voyage-multimodal-3 was the first production-grade embedding model to handle interleaved text and images, voyage-multimodal-3.5 advances this unified processing approach, more effectively vectorizing multimodal data together to best capture key semantic meaning from tables, graphics, figures, slides, PDFs, and more. This helps developers eliminate the significant effort required for complex document parsing, which can reduce retrieval accuracy and lead to less trustworthy applications.
- **Automated Embedding for MongoDB Vector Search:** Automatically generate and store high-fidelity embeddings using Voyage AI whenever data is inserted, updated, or queried. By handling embedding generation natively within the database, MongoDB removes the need for separate embedding pipelines or external model services. Embeddings stay fresh as data changes, helping retrieval to remain accurate and AI applications to maintain reliable context. The result is a simpler architecture with fewer moving parts, making it easier for teams to build and run AI-enabled applications in production.

Automated Embedding is available in public preview with support in our drivers (e.g. Javascript, Python, Java, etc) and AI Frameworks like LangChain and LangGraph (Python). Available today for MongoDB Community, and coming soon on MongoDB Atlas.

"We were looking for extremely accurate embedding models, and Voyage AI provided accuracy at scale," says Sudheesh Nair, Cofounder and CEO of TinyFish. "The Python APIs that Voyage comes out of the box with are also extremely lightweight and very fast."

"Today, companies need to move extremely fast, and at very lean startups, you need to only focus on what you are building," said Rotem Weiss, CEO of Tavily. "MongoDB allows us to focus on what matters most, our customers and our business."

For the first time, developers can build and run AI applications with operational data, semantic understanding, and retrieval in a single system. MongoDB's Atlas Embedding and Reranking API exposes Voyage AI models natively within Atlas, allowing teams to ship AI features with enterprise-grade security, performance, and reliability infrastructure. An intelligent assistant for MongoDB Compass and Atlas Data Explorer is now generally available, delivering natural-language, AI-powered assistance for everyday data operations, such as query optimization. MongoDB also introduced a new AI skills certification to help teams scale data strategies, accelerate time to market, and reduce costs–the first in a broader set of AI skill offerings planned this year.

To learn more about these new capabilities and to get started, please find the wrap blog here.

About MongoDB
Headquartered in New York, MongoDB's mission is to empower innovators to create, transform, and disrupt industries with software. MongoDB's unified data platform was built to power the next generation of applications, and MongoDB is the most widely available, globally distributed database on the market. With integrated capabilities for operational data, search, real-time analytics, and AI-powered data retrieval, MongoDB helps organizations everywhere move faster, innovate more efficiently, and simplify complex architectures. Millions of developers and more than 60,000 customers across industries—including over 75% of the Fortune 100—rely on MongoDB for their most important applications. To learn more, visit mongodb.com.

Press Contact:
press@mongodb.com

View original content to download multimedia:https://www.prnewswire.com/news-releases/mongodb-sets-a-new-standard-for-retrieval-accuracy-with-voyage-4-models-for-production-ready-ai-applications-302662558.html

SOURCE MongoDB, Inc.