

Vespa.ai and Voyage AI by MongoDB Partner to Lower AI Search Costs and Boost Performance

New integration reduces AI search costs, improves retrieval quality, and increases performance by removing external API calls

TRONDHEIM, NORWAY, August 20, 2026 /EINPresswire.com/ -- [Vespa.ai](https://vespa.ai) today announced a

“

As AI becomes central to enterprise applications, enterprises shouldn't have to choose between cost and quality. Our partnership with Voyage AI delivers both, with greater reliability and scale.”

Jon Bratseth

partnership with Voyage AI by MongoDB to help enterprises run AI-powered search more efficiently, reduce costs and improve performance and reliability. The collaboration introduces a new architecture for processing AI search queries that removes one of the biggest ongoing expenses: repeatedly generating embeddings for every user query.

Voyage AI by MongoDB delivers state-of-the-art embedding and reranking models designed for high-performance retrieval, enabling organizations to build trustworthy AI applications with industry-leading accuracy.

Reducing a Major Cost in AI Search

In most AI search systems, every search query must be converted into a vector (an “embedding”) before results can be returned. At scale, this can mean millions or billions of API calls each month, driving high cost, latency and adding external dependencies.

The Vespa-Voyage AI integration takes a different approach. Companies can process queries directly within the query pipeline, avoiding these repeated external calls altogether.

A Simpler, More Efficient Architecture

In this architecture, documents are embedded once, while queries are processed within Vespa at the point of execution, using a model optimized for speed and efficiency. This removes the need to send queries to external services, resulting in lower operating costs, faster response times, and more reliable performance without dependency on external APIs.

[Learn more about this architecture.](#)

Built for Enterprise-Scale AI Applications

The joint solution is designed for organizations running high-traffic, real-time applications such as search, intelligent recommendations, and AI assistants. It enables teams to scale to large datasets and high query volumes, balance cost and accuracy across use cases, and operate within a single, unified system. Vespa also applies a two-step ranking process, retrieving results quickly before refining them for accuracy, improving relevance without slowing performance.

A Joint Vision for the Future of AI Search

"As AI becomes central to enterprise applications, the demands on performance and scale are increasing," said Jon Bratseth, CEO at Vespa.ai. "Enterprises shouldn't have to choose between cost and quality in AI search. By partnering with Voyage AI, we are enabling a new architecture that delivers both, while also improving reliability and scalability."

"Voyage AI's mission is to provide the best embedding models in the world," said Frank Liu.

"Together with Vespa, customers have more ways to deploy those models in order to power high-performing AI applications, with economic and operational efficiency."

Availability

The integrated solution is available today, with support for Voyage AI embeddings in Vespa Cloud and self-managed deployments.

[To learn more, visit the PyVespa guide.](#)

About Vespa.ai

Vespa.ai develops the Vespa AI Search Platform, enabling organizations to build AI search, recommendations, answer engines, AI agents, and other customer-facing AI applications within a unified distributed architecture.

Timothy Young

Vespa

[email us here](#)

Visit us on social media:

[LinkedIn](#)

This press release can be viewed online at: <https://www.einpresswire.com/article/935330382>

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.

© 1995-2026 Newsmatics Inc. All Right Reserved.