captain™ v2 API launch + product updates

We've been busy this holiday season, shipping optimizations throughout the stack. I'm thrilled to announce the stable release of our v2 API, among other product updates:

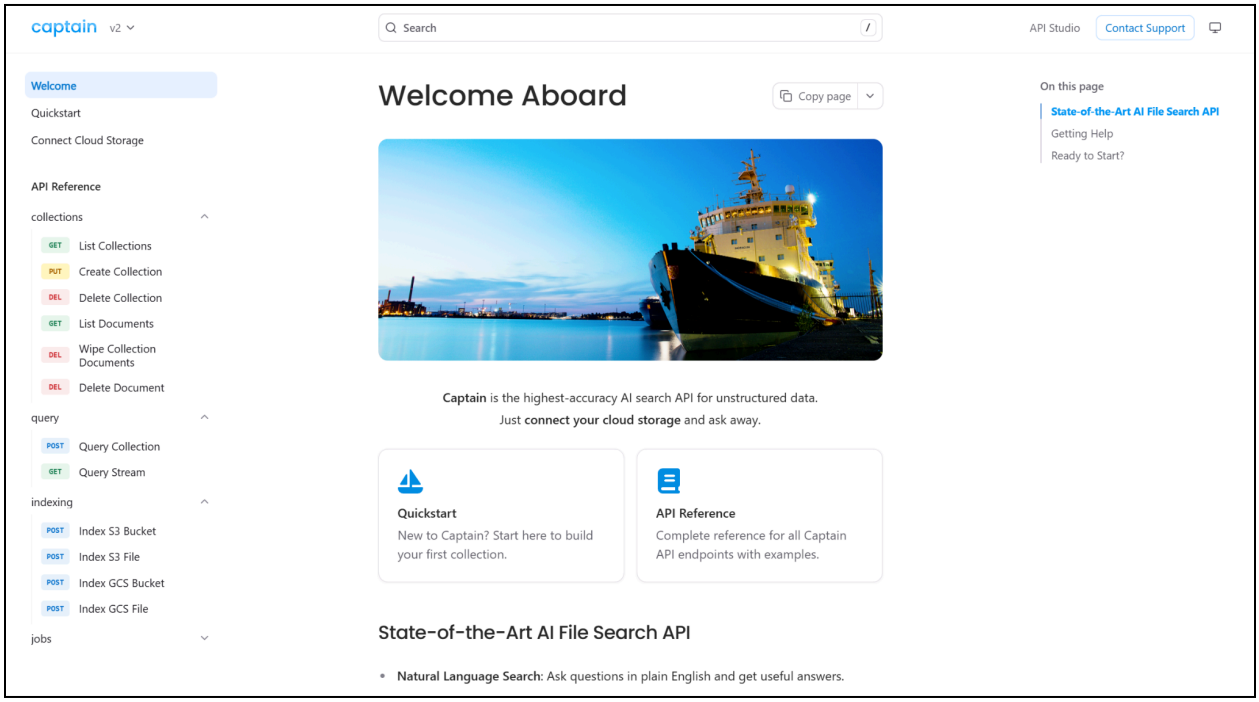
TL;DR

- Increased search precision for legal, healthcare, and financial datasets
- Search latency slashed **3x**
- v2 REST API now stable
- Brand new interactive documentation site - docs.runcaptain.com

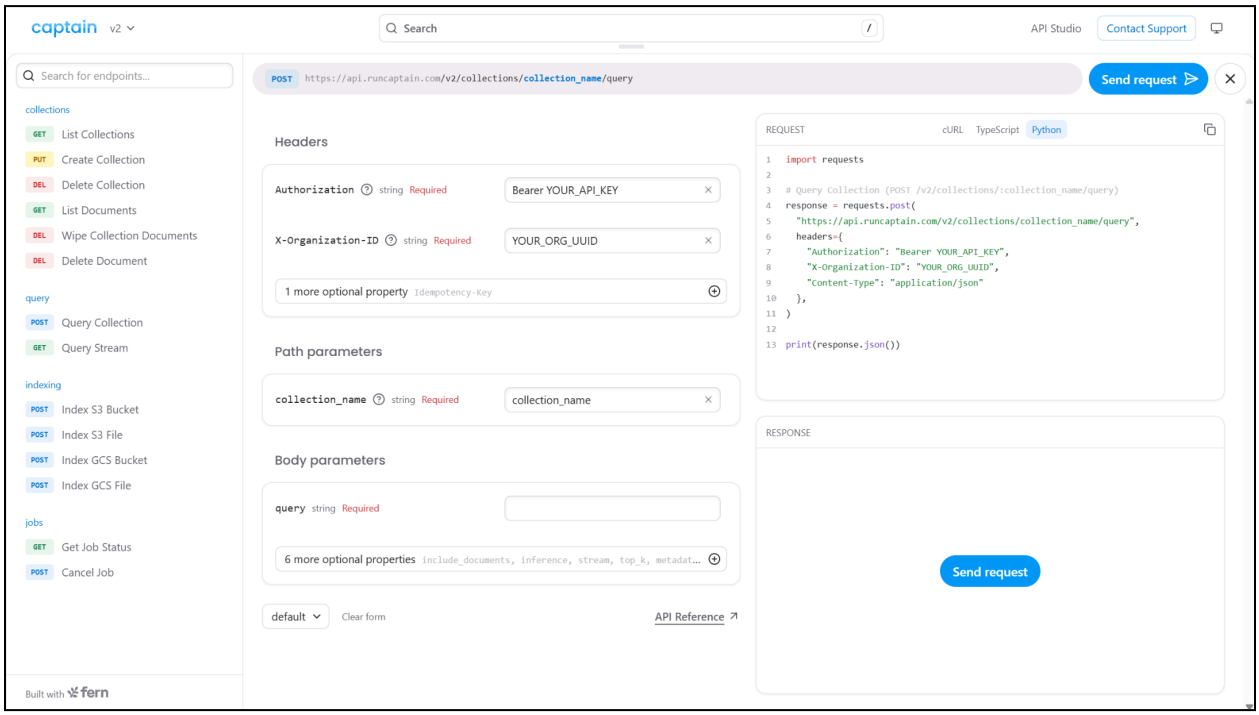
- Python & TypeScript SDKs out later this month
- Hacker News RAG Search by Captain is underway

New Interactive Docs

We've invested in a beautiful new docs site with an interactive API explorer. You can now supply your API key and test out requests from within the docs!

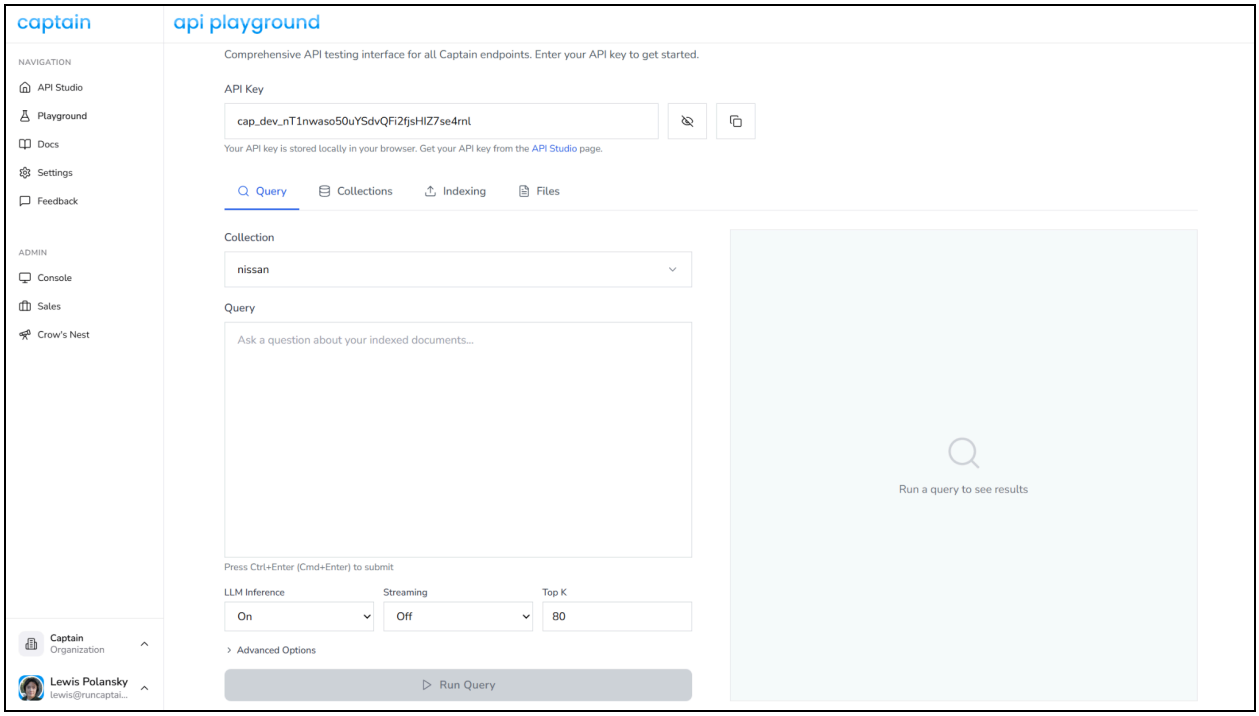


The new Captain Docs (docs.runcaptain.com/welcome)



Interactive API Explorer

In addition to the interactive docs, for on-the-fly API development, the Captain Playground has been redesigned for non-technical users to experience the power of high-accuracy RAG within a non-programmatic interface.



The new Captain Playground

Precision Optimizations

Our retrieval precision has been further optimized for healthcare, legal, and financial RAG use cases. We have employed a number of techniques to increase accuracy, namely long-context document embedding and automatic domain detection.

Long-context document embedding entails the embedding of the entire context of a file and supplementing it with the narrow chunk content. This provides both broad and narrow semantic value during similarity lookups.

Automatic domain detection allows Captain to index domain-specific terms within a BM25 index (to supplement dense vector search). This development solves the long-standing issue within RAG (and search more broadly) of domain-specific terms getting 'lost' without proper vector correlations.

By combining these two techniques and weighing them, search results provide a better understanding of complex document narratives and niche industry terms like drug names or non-English legal terms.

Latency Optimizations

We have seamlessly migrated away from vector storage, resulting in an over 3x drop in latency. Our experimentation has also shown better recall when object storage search is used as opposed to leveraging vector databases.

This past December we successfully transferred all non-textual knowledge data into object storage. There was no disruption to service uptime.

We remain committed to leveraging state-of-the-art embedding models, but going forward, our embeddings will be stored as objects.

What's next?

Captain has partnered with the Hacker News team to develop an RAG forum search across the popular aggregator's over 19+ years of posts and comments. More on that in a few weeks.

Our Python and TypeScript SDKs are underway; these will be out later this month.

Happy Shipping!

Lewis Polansky
CEO & Co-Founder

Jan 3, 2026