How vector search makes Vincent collections faster and smarter

Written by Sasha Perrin



When Vincent by Clio's Collections launched in 2024, the premise was to give legal teams a persistent, searchable index of their own documents. Upload your files, build your collection, and Vincent will search through them and find what you need.

As usage of Collections grew, so did the ambition of what firms wanted to do with them. Initially, firms would upload a handful of documents to their collections. Over time, they wanted to be able to upload hundreds or even thousands of documents. Firms began asking whether they could index entire practice group libraries, matter archives, or firm-wide knowledge bases. Unfortunately, the answer at the time was "not yet."

In late 2025, Clio's product and engineering teams did an assessment of Collections. The internal indexing system we had built in-house was not broken, but also wasn't built for where the feature was heading. So we began to search for something built specifically to scale.

How vector search finds what a keyword search misses

When a document is uploaded into Vincent by Clio, an embedding model converts each passage into a numerical "vector," or a point in semantic space that captures the meaning of that text. When a user queries a Vincent Collection, the system finds vectors that are geometrically close to the query vector, surfacing passages that are semantically similar to what was asked, even if they share no words.

This is what allows Vincent to understand legal language in the way it's actually used by lawyers. A contract clause describing "unforeseeable circumstances beyond the control of either party" is semantically similar to a force majeure provision. A keyword search would not find this contract clause because it doesn't understand that the underlying meanings are similar. A vector search does.

For legal work, a missed clause in due diligence or a missed precedent in research has significant downstream consequences. The retrieval layer underneath an AI platform is the part of the system that determines whether that platform's output can be trusted.

While Vincent's vector search was the right initial approach, the infrastructure behind it couldn't scale to meet what enterprise legal teams were demanding of it.

Scaling vector search with the right infrastructure

Vector search at millions of documents, with consistent sub-second retrieval, reliable indexing, and the ability to handle variable and unpredictable load is an infrastructure problem.

As Vincent's Collections grew, the limits of our internal indexing system became clear. Larger Collections slowed performance. Indexing new documents introduced a lag between upload and availability. And our existing architecture was not well-suited for the advanced usage our enterprise clients needed.

Turbopuffer is built for high-speed vector search at scale. It is not a general-purpose database with vector capabilities added on. It was designed from the ground up for high query volume, variable Collection sizes, and consistent performance regardless of how many documents are in the index. The "turbo" in the name represents how it was engineered for sub-second retrieval even at millions of vectors. The "puffer" in the name represents how compute scales up as the Collection size grows and contracts when demand drops.

With Turbopuffer, new documents are processed and ready for search almost immediately after upload. Collections that previously showed performance degradation at large volumes now handle significantly more documents without any change in retrieval speed or accuracy. And the architecture now supports the kind of sophisticated querying that was difficult to process with the infrastructure we had before.

Using vector search across your documents

The most significant consequence of this infrastructure change is what it enables for Vincent's agentic capabilities. Vincent by Clio can execute complex, multi-step legal tasks from a single instruction, including drafting, analysis, research, and strategy development, without requiring the user to orchestrate each step. Underpinning this is the growing network of legal-specific skills that Vincent draws on autonomously depending on what the overall task requires.

Collections are now queryable by those agentic skills. That means Vincent can reach into a firm's uploaded document library as part of an autonomous workflow, surfacing relevant documents while drafting, cross-referencing matter history during analysis, or pulling from a practice group's institutional knowledge base without being explicitly directed to do so.

For this to work at scale, the retrieval layer needs to be fast enough to not become a bottleneck, accurate enough to surface genuinely relevant content, and capable of handling the document volumes that large firms operate with. Turbopuffer has delivered precisely that.

How Vincent handles your data

Because Turbopuffer processes data on Vincent's behalf, it is classified as a subprocessor under GDPR and other applicable data protection frameworks. Customers were notified in advance of its introduction, in line with our data protection obligations.

Turbopuffer does not store original documents, only vector embeddings, which are mathematical representations of meaning used for search. Vincent by Clio retains the original files. Data does not leave the customer's existing regional environment. It runs in the same AWS region as the customer's Vincent instance (e.g., Ireland for UK and EU customers, U.S. East for U.S. customers, Sydney for Australian customers; etc). Customer data is never used to train AI models. These are zero-data retention standards applied consistently across the stack.

What this unlocks next

The move to Turbopuffer is the foundation, not the finish line. The new architecture supports capabilities that were difficult or impossible to build on what we had before, including significantly larger Collections, more sophisticated document analysis, and deeper integrations with firm knowledge bases that allow Vincent to operate with a richer, more complete picture.

The agentic layer is already live. The retrieval layer that makes it useful at scale is now in place. The next phase is building on both, expanding on what Vincent can find, what it can do with what it finds, and how seamlessly that happens within the workflows that legal teams already rely on.

The goal is to shape a version of Vincent by Clio that knows your firm's work as well as your best lawyers do. This infrastructure is the latest step in getting there.

Ready to see how Vincent by Clio's upgraded Collections would perform at your firm?

Request a Demo Today

About Clio

Clio is the global leader in legal AI, powering legal work across the profession, from law firms of every size to corporate legal teams, government agencies, and courts. Its Intelligent Legal Work Platform unites the business and practice of law, connecting the people, information, and processes behind legal work. Clio's AI brings the context together with legal sources drawn from Clio Library, a growing collection of more than one billion legal documents. Over 400,000 legal professionals in more than 130 countries rely on Clio to run their legal world. Learn more at www.clio.com.