News

By Voyage AI · January 7, 2025

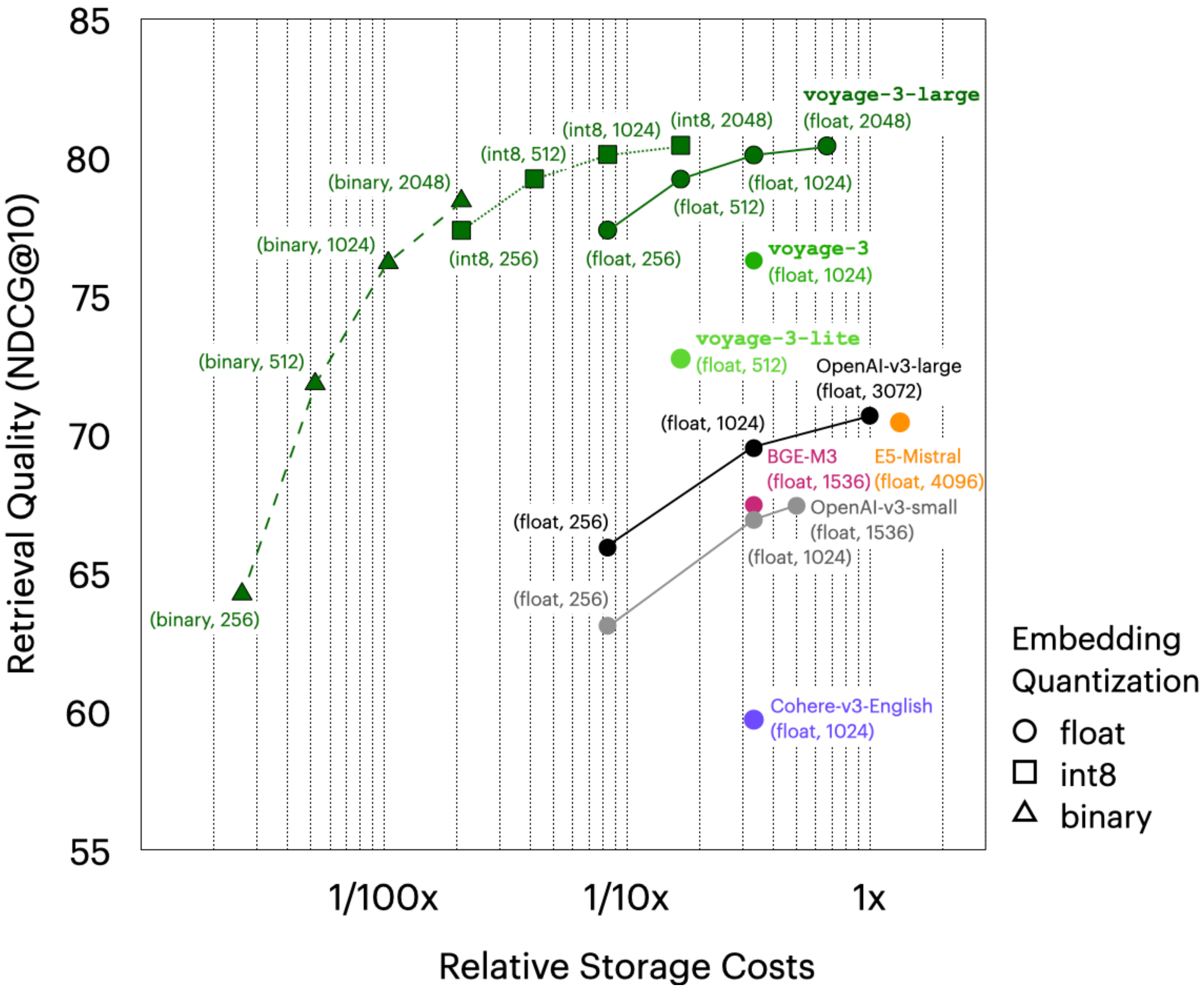
voyage-3-large: the new state-of-the-art general-purpose embedding model

TL;DR – Introducing **voyage-3-large**, a new state-of-the-art general-purpose and multilingual embedding model that ranks first across eight evaluated domains spanning 100 datasets, including law, finance, and code. It outperforms OpenAI-v3-large and Cohere-v3-English by an average of 9.74% and 20.71%, respectively. Enabled by Matryoshka learning and quantization-aware training, **voyage-3-large** supports smaller dimensions and int8 and binary quantization that dramatically reduce vectorDB costs with minimal impact on retrieval quality.

We are excited to announce **voyage-3-large**, the latest addition to our Voyage 3 series, offering the best general-purpose and multilingual retrieval quality, which:

- outperforms OpenAI-v3-large and Cohere-v3-English by an average of 9.74% and 20.71%, respectively, across 100 datasets, spanning eight diverse domains, including law, finance, and code
- supports 2048, 1024, 512, and 256 dimensional embeddings enabled by Matryoshka learning
- offers multiple embedding quantization options, including 32-bit floating point, signed and unsigned 8-bit integer and binary precision—while minimizing quality loss.
- supports a 32K-token context length, compared to OpenAI (8K) and Cohere (512)

The following figure illustrates the tradeoff between retrieval quality and storage cost (which is proportion to the number of bits per vector). We see that `voyage-3-large` with int8 precision and 1024 dimensions is only 0.31% below `voyage-3-large` with float precision and 2048 dimensions, despite using 8x less storage. Further, it is still 9.44% better than OpenAI-v3-large with float precision and 3072 dimensions, despite using 12x less storage. Even with 512-dimensional binary embeddings, `voyage-3-large` outperforms OpenAI-v3-large (3072-dimensional float embeddings) by 1.16%, requiring just 200x less the storage costs—e.g., $20K in monthly storage costs will drop to $100!



The flexibility in precisions and dimensionality was enabled by Matryoshka and quantization-aware training. Please check out our previous blog post for more details.

Of note, while `voyage-3-large` establishes a new accuracy-cost frontier, `voyage-3` and `voyage-3-lite` still offer cutting-edge retrieval quality—outperforming OpenAI-v3-large by an average of 5.60% and 2.06%, respectively—at lower inference costs and latency.

The evaluation results used to generate this plot are available in this spreadsheet.

Evaluation Details

Datasets. We evaluate on 100 datasets, spanning eight domains, technical documentation, code, law, finance, web reviews, multilingual, long documents, and conversations. Each dataset consists of a corpus (e.g., technical documentation, court opinions) and queries (e.g., questions, summaries). The following table list the datasets in the eight categories except multilingual, which includes 62 datasets covering 26 languages. A list of all evaluation datasets is available in this spreadsheet.

Category	Descriptions	Datasets
TECH	Technical documentation	Cohere, 5G, OneSignal, LangChain, PyTorch
CODE	Code snippets, docstrings	LeetCodeCpp-rtl, LeetCodeJava-rtl, LeetCodePython-rtl, HumanEval-rtl, MBPP-rtl, DS1000-referenceonly-rtl, DS1000-rtl, APPS-rtl
LAW	Cases, court opinions, statutes, patents	LeCaRDv2, LegalQuAD, LegalSummarization, AILA casedocs, AILA statutes
FINANCE	SEC filings, finance QA	RAG benchmark (Apple-10K-2022), FinanceBench, TAT-QA-rtl, Finance Alpaca, FiQA-Personal-Finance-rtl, Stock News

Category	Descriptions	Datasets
		Sentiment, ConvFinQA-rtl, FinQA-rtl, HC3 Finance
WEB	Reviews, forum posts, policy pages	Huffpostsports, Huffpostscience, Doordash, Health4CA
LONG-CONTEXT	Long documents on assorted topics: government reports, academic papers, and dialogues	NarrativeQA, QMSum, SummScreenFD, WikimQA
CONVERSATION	Meeting transcripts, dialogues	Dialog Sum, QA Conv, HQA

Models. We evaluate `voyage-3-large` alongside several alternatives, including: OpenAI-v3 small (`text-embedding-3-small`) and large (`text-embedding-3-large`), E5-Mistral (`intfloat/e5-mistral-7b-instruct`), BGE-M3 (`BAAI/bge-m3`), Cohere-v3-English (`embed-english-v3.0`), `voyage-3`, and `voyage-3-lite`. For domain-specific evaluation, we also include `voyage-law-2`, `voyage-finance-2`, `voyage-code-2`, and `voyage-code-3`.

Metrics. Given a query, we retrieve the top 10 documents based on cosine similarities and report the normalized discounted cumulative gain (NDCG@10), a standard metric for retrieval quality and a variant of the recall.

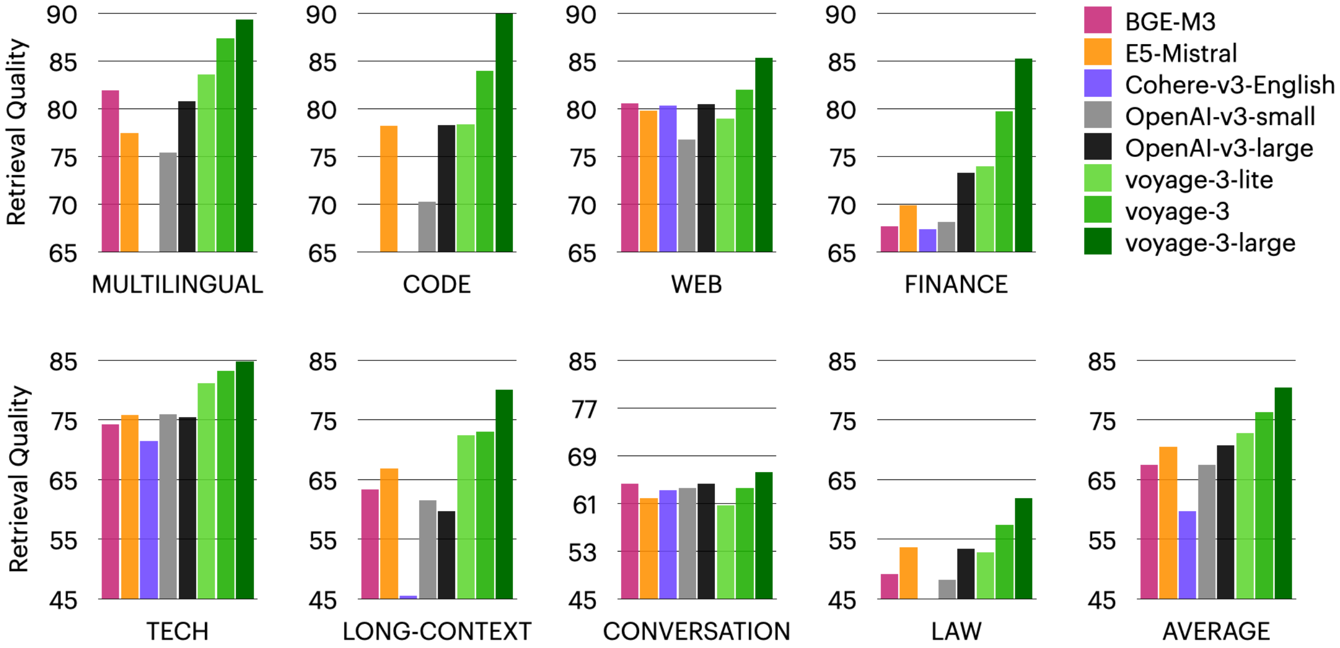
Results

All the evaluation results are available in this spreadsheet.

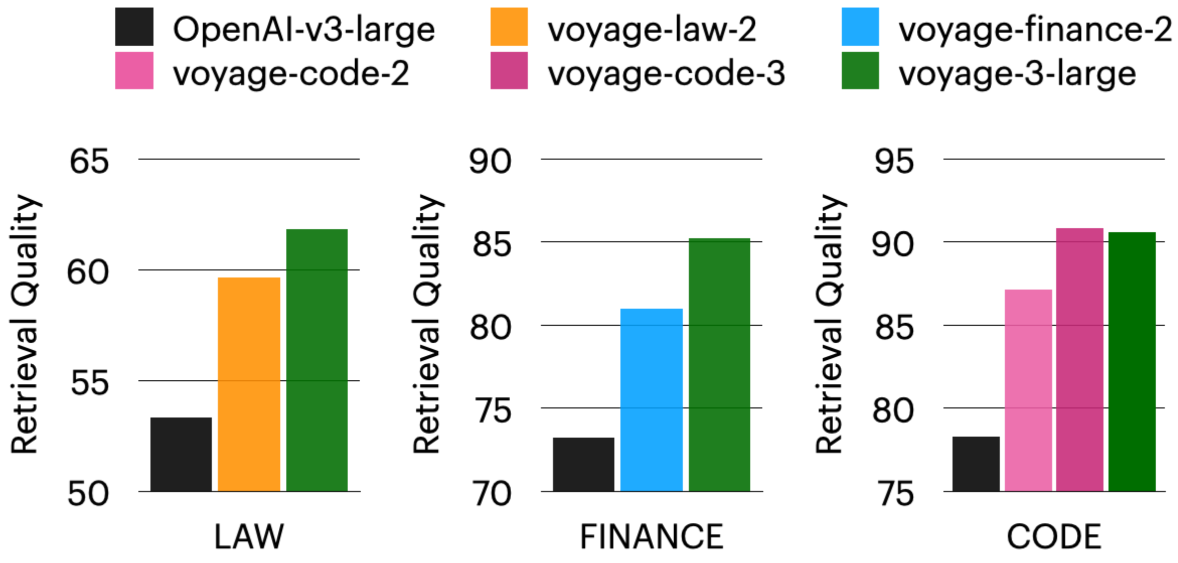
On average, `voyage-3-large` outperforms OpenAI-v3-large by:

- 10.58% and 11.47% at 1024 and 256 dimensions, respectively
- 8.56% at 1/24 the storage cost (int8 512 vs. 3072 dimensions)
- 1.16% at 1/200 the storage cost (binary 512 vs. float 3072 dimensions)

Domain-specific quality. The bar charts below illustrate the average retrieval quality of `voyage-3-large` with full precision and 2048 dimensions, both overall and for each domain. Overall, `voyage-3-large` is the top-performing model, surpassing `voyage-3`, `voyage-3-lite`, and OpenAI-v3-large by an average of 4.14%, 7.68%, and 9.74%, respectively.



Notably, it also outperforms Voyage's domain-specific models from the previous series 2 generations, `voyage-law-2`, `voyage-finance-2`, and `voyage-code-2`, on the corresponding domains. (However, `voyage-code-3` remains the best for code retrieval.)



Binary rescoring. Finally, users sometimes retrieve a set of documents with binary embeddings (e.g., 100 in our evaluation) and then rescore the retrieved documents with full-precision embeddings. For `voyage-3-large`, binary rescoring yields up to 5.84% improvement in retrieval quality when applied on top of standard binary retrieval.

Try voyage-3-large!

`voyage-3-large` is available today! The first 200 million tokens are free. To get started, head over to our docs to learn more. If you're also interested in fine-tuned embedding models, we'd love to hear from you—please email us at contact@voyageai.com. Follow us on X (Twitter) and LinkedIn, and join our Discord for more updates.

Tags:

Leave a Reply

Write a comment...

Comment

---

www.voyageai.com