News

By Voyage AI · November 12, 2024

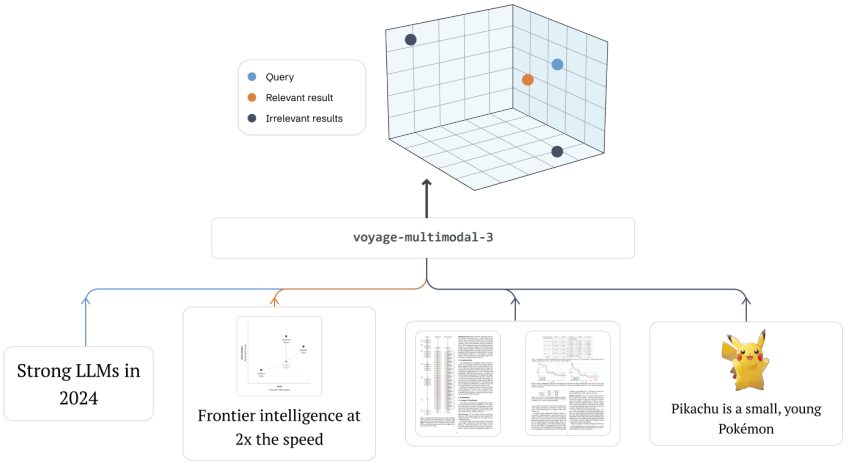
voyage-multimodal-3: all-in-one embedding model for interleaved text, images, and screenshots

> TL;DR — We are excited to announce `voyage-multimodal-3`, a new state-of-the-art for multimodal embeddings and a big step forward towards seamless RAG and semantic search for documents rich with both visuals and text. Unlike existing multimodal embedding models, `voyage-multimodal-3` is capable of vectorizing interleaved texts + images and capturing key visual features from screenshots of PDFs, slides, tables, figures, and more, thereby eliminating the need for complex document parsing. `voyage-multimodal-3` improves retrieval accuracy by an average of 19.63% over the next best-performing multimodal embedding model when evaluated across 3 multimodal retrieval tasks (20 total datasets).

Two months ago, we released the `voyage-3` and `voyage-3-lite` series of multilingual text embedding models, providing best-in-class performance across a variety of datasets. Today, we're excited to introduce `voyage-multimodal-3`, our first multimodal embedding model and a big step toward RAG and semantic search for knowledge bases rich with both visuals and text.

`voyage-multimodal-3` supports text and content-rich images such as screenshots of texts, figures, tables, PDFs, slide decks, and more. The resultant vectors capture critical textual and visual features such as font size, text location, whitespace, etc. This eliminates the need for heuristic-based document parsing, which often struggles with accuracy when layouts are complex or interspersed with figures and photos. Unlike existing multimodal embedding models that handle either a single text or image input, `voyage-multimodal-3` allows for interleaved texts and images for maximum flexibility. Our sample notebook demonstrates all of these features.

`voyage-multimodal-3` has an architecture that is similar to that of modern vision-language transformers. This makes it a significant departure from existing multimodal embedding models, including, but not limited to, OpenAI CLIP large (`clip-vit-large-patch14-336`) and Cohere multimodal v3 (`embed-multimodal-v3.0`).

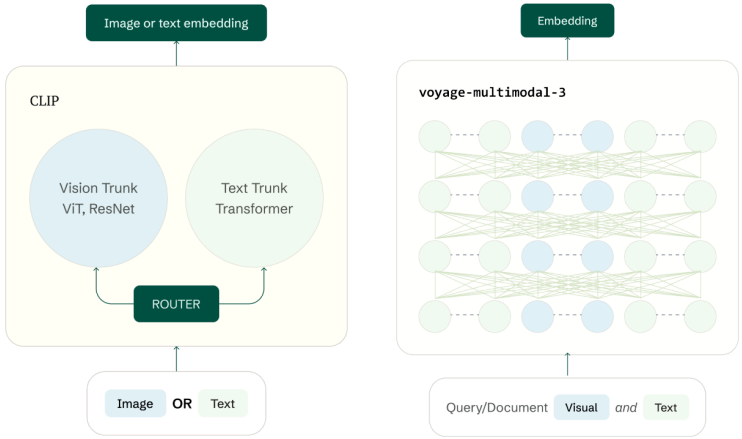


In a set of evaluations across 20 multimodal retrieval datasets and 34 text retrieval datasets, we found that `voyage-multimodal-3`:

1. Outperforms OpenAI CLIP large and Cohere multimodal v3 by an average of 41.44% (a 2.1x improvement) and 43.37% (a 2.2x improvement) on table/figure retrieval, 26.54% and 25.84% on document screenshot retrieval, and 6.55% and 5.86% on text-to-photo retrieval, respectively.
2. Outperforms OpenAI v3 large and Cohere multimodal/English[1] v3 by 5.13% and 13.70% on text-only datasets, respectively.

Support for Interleaved Text & Images

All existing commonly used multimodal embedding models (such as Amazon Titan Multimodal G1, Google Vertex AI multimodal, and Cohere multimodal v3) are based on OpenAI's CLIP, which processes different modalities of data through independent networks. In other words, images *must* be vectorized through the vision tower, while text *must* be vectorized through the text tower, preventing these models from being able to processing interleaved data.

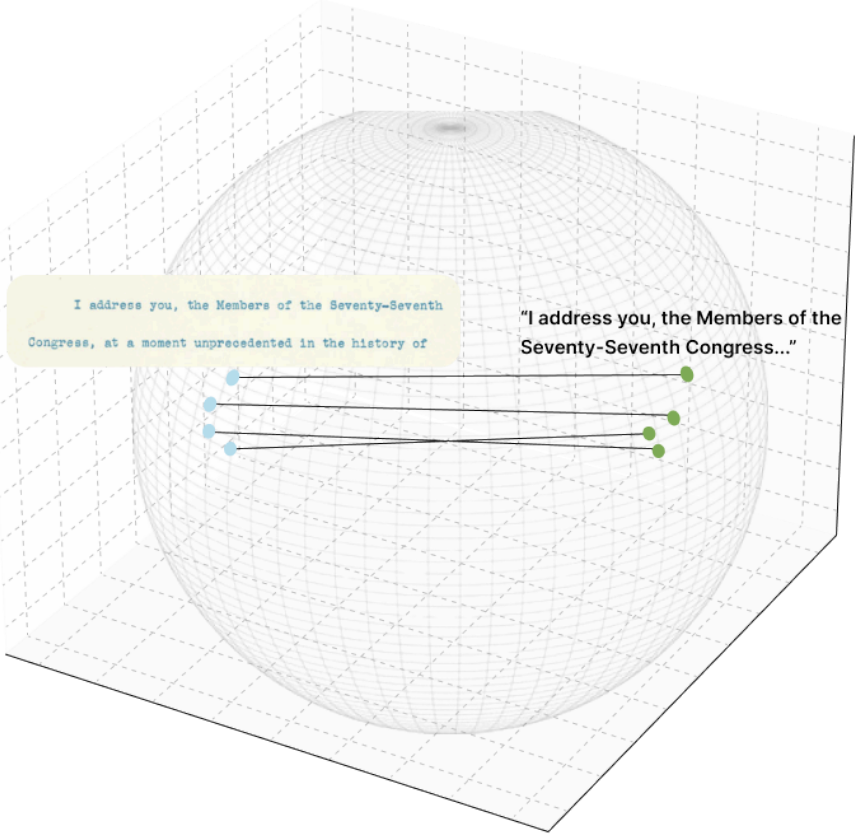


In contrast, `voyage-multimodal-3` vectorizes both modalities of data directly within the same transformer encoder, ensuring that both text and visual features are treated as part of a unified representation rather than distinct components. This mimics the model architecture of the latest vision-language models, only for vectorization rather than generation. As a result, interleaved texts and images, document screenshots, PDFs with complex layouts, annotated images, etc can

be vectorized in a way that preserves the contextual relationship between visual and textual information.

Mixed Modality Search with Screenshots

All CLIP-like models perform poorly on mixed-modality search due to a phenomenon known as the modality gap. As illustrated in the figure below, the closest vector to the snippet "I address you, members of the Seventy-Seventh Congress..." is not its screenshot, but other texts. This leads to search results that are skewed towards items of the same modality; in other words, text vectors will be closer to irrelevant texts than relevant images in the embedding space.



To illustrate this issue quantitatively, we conducted an experiment involving mixed-modality data. We created two sets of PyTorch documentation with identical content: one set as plain text (strings) and and the other set as screenshots. By combining a subset of text-based documentation with screenshots of remaining subset, we created a series of mixed-modality datasets. Each dataset represented a different proportion of text and screenshots, ranging from 0% to 100% screenshots. We then evaluated the retrieval accuracy of various multimodal models on these datasets, reporting the normalized discounted

cumulative gain (NDCG@10) for each model across different screenshot ratios.

As shown above, CLIP-based models experience a decline in retrieval quality as the proportion of screenshots increases up to 90%, highlighting a retrieval bias influenced by modality. Moreover, these models perform poorly when all text is converted to images.

In contrast, `voyage-multimodal-3` is not only the most performant for all ratios, but also has little-to-no performance drop across the board, indicating that the vectors truly capture the semantic content

contained in the screenshots. This robustness is due to the model's unique approach of processing all input modalities through the same backbone.

With `voyage-multimodal-3`, there is no longer a need for screen parsing models, layout analysis, or any other complex text extraction pipelines; you can easily vectorize a knowledge base containing both pure-text documents as well unstructured data (such as PDFs/slides/webpages/etc) — screenshots are all you need.

Evaluation Details

Datasets. We evaluate `voyage-multimodal-3` across 20 multimodal datasets spanning three different tasks: table/figure retrieval, document screenshot retrieval (the ViDoRe benchmark), and text-to-photo retrieval. We also evaluate `voyage-multimodal-3` on a standard text retrieval task spanning 34 datasets in 6 domains (law, finance, conversation, code, web, and tech).

For all datasets, the query is text, while the document could be a figure, photo, text, document screenshot, or a combination of these. For each task, we use prior top-performing models as the baseline. Alongside task names, we provide each task's corresponding description and datasets used in the table below:

Task	Description	Datasets
Table/figure retrieval	Table/figure retrieval measures the strength of a model's ability to match an image containing a table or figure (charts, graphs, etc) with descriptions, captions, or other textual queries which reference the figure.	charxiv, mmtab-test, ChartQA, Chartve, FintabnetQA, PlotQA,
Document screenshot retrieval	In this category, models are used to match queries with scans or screenshots of	Energy, Healthcare Industry, Artificial Intelligence, Government Report,

Task	Description	Datasets
	documents containing both text and charts. We use all datasets from the ViDoRe benchmark for this task.	InfoVQA, DocVQA, ArxivQA, TabFQuad, TAT-DQA, Shift Project
Text-to-photo retrieval	This is the typical text-to-image matching used by CLIP and other CLIP-like models, where queries are associated with the most semantically relevant photos.	meme-cap, mm-imdb, winoground, docci
Standard text retrieval	Standard text retrieval retrieves	LeCaRDv2, LegalQuAD legal_summari

Task	Description	Datasets
	relevant documents by matching query strings with document strings.	zation, AILA_casedocs, AILA_statutes, rag-benchmark-finance-apple-10K-2022, financebench, TAT-QA, finance-alpaca-csv fiqa-personal-finance-dataset, finance-financialmodelingprep-stock-news-sentiments-rss-feed, ConvFinQA, finqa, hc3_finance, dialogsum, QAConv, HQA-

Task	Description	Datasets
		data, LeetCodeCpp-new, LeetCodeJava-new, LeetCodePython-new, humaneval, mbpp, ds1000-referenceonly, ds1000, apps_5doc, Huffpostsports, Huffpostscience, Doordash, Healthforcalifornia, Cohere, 5GEdge, OneSignal, Langchain, PyTorch1024

Note that the standard text retrieval task encompasses all datasets used to evaluate `voyage-3`

and `voyage-3-lite` except long context and multilingual datasets. See our previous blog post for more information.

Models. For the three multimodal tasks, we evaluate `voyage-multimodal-3` alongside four alternative multimodal embedding models: OpenAI CLIP large (`clip-vit-large-patch14-336`), Amazon Titan Multimodal Embeddings G1 (`amazon.titan-embed-image-v1`), Cohere multimodal v3 (`embed-multimodal-v3.0`), and SigLIP So400M (`siglip-so400m-patch14-384`). We also evaluate ColQwen2 v0.1 (`colqwen-v0.1`), a late interaction model that outputs many embeddings per document.

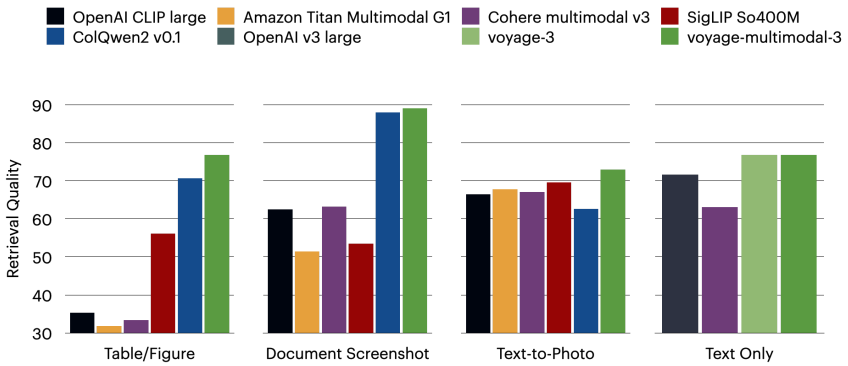
For the standard text retrieval task, we evaluate `voyage-multimodal-3` alongside OpenAI v3 large (`text-embeddings-3-large`), Cohere multimodal/English[1] v3, and `voyage-3`.

Metrics. Given a query, we retrieve the top 10 results by cosine similarity and report the NDCG@10.

Results

Multimodal retrieval. As shown in the figure below, `voyage-multimodal-3` outperforms OpenAI CLIP large, Amazon Titan Multimodal G1, Cohere multimodal v3, SigLIP So400M, and ColQwen2 v0.1 by:

- 41.44%, 45.00%, 43.37%, 20.66%, and 6.14% on table/figure retrieval, respectively
- 26.54%, 37.68%, 25.84%, 35.62%, and 0.98% on document screenshot retrieval, respectively
- 6.55%, 5.16%, 5.86%, 3.42%, and 10.34% on text-to-photo retrieval, respectively



Standard text retrieval. As shown in the figure below, `voyage-multimodal-3` outperforms OpenAI v3 large and Cohere multimodal/English[1] v3 by 5.13% and 13.70%, respectively. The performance of `voyage-multimodal-3` is 0.05% better than that of `voyage-3`, making the two comparable in terms of retrieval accuracy for pure text documents.

All evaluation results are available in this spreadsheet.

Try voyage-multimodal-3 now!

`voyage-multimodal-3` is available today! The first 200 million tokens are free. To get started, check out our sample notebook, or 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.

---

[1] Cohere multimodal v3 uses Cohere English v3 (`embed-english-v3.0`) for the text tower, which makes the both models' vectors identical on pure text. To minimize confusion, we use "Cohere multimodal v3" as the only label in the charts.

Tags:

Leave a Reply