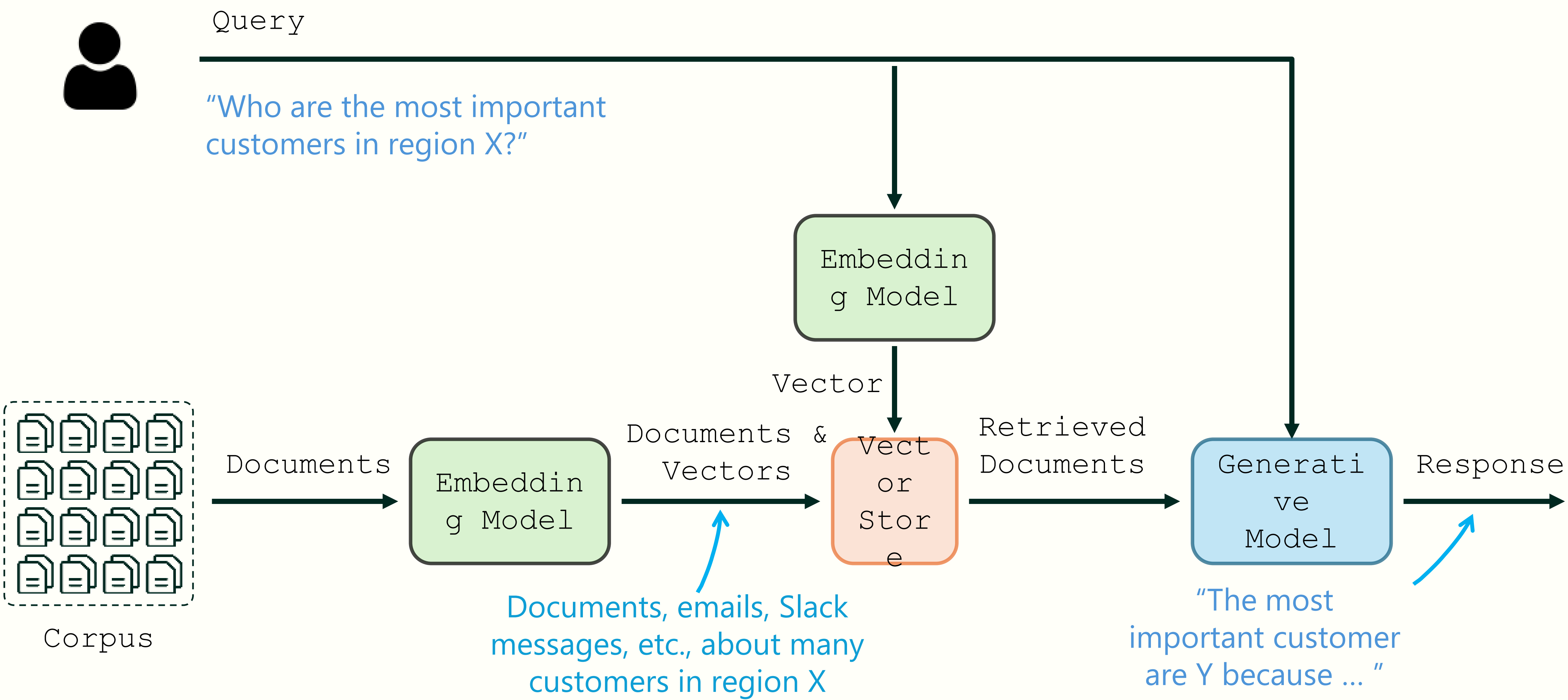
Voyage AI

Embedding models

(RAG) is the Predominant Approach for Enterprise Gen AI



Query

"Who are the most important customers in region X?"

Embedding Model

Vector

Documents

Embedding Model

Documents & Vectors

Vector Store

Retrieved Documents

Generative Model

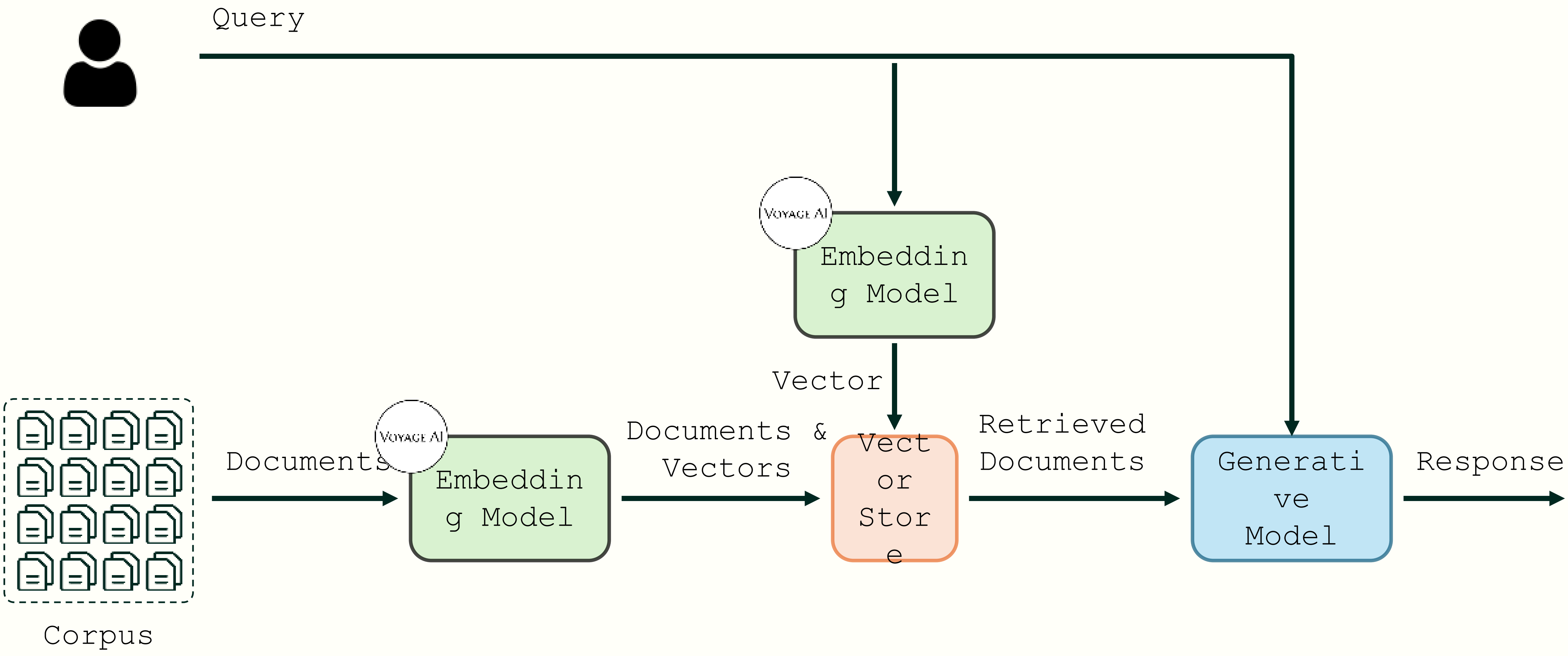
Response

Corpus

Documents, emails, Slack messages, etc., about many customers in region X

"The most important customer are Y because … "

(RAG) is the Predominant Approach for Enterprise Gen AI



Query

VOYAGE AI

Embeddin g Model

Vector

Corpus

Documents

VOYAGE AI

Embeddin g Model

Documents & Vectors

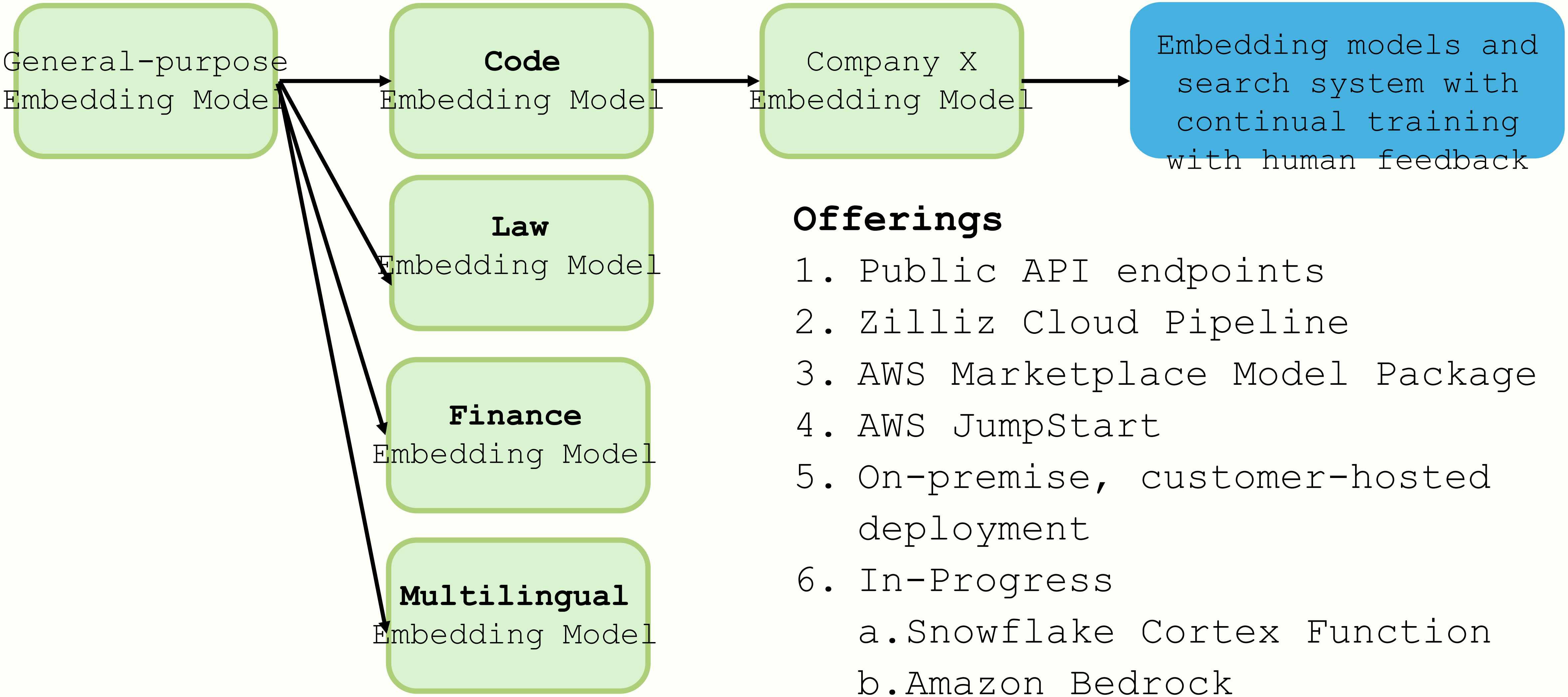
Vect or Stor e

Retrieved Documents

Generati ve Model

Response

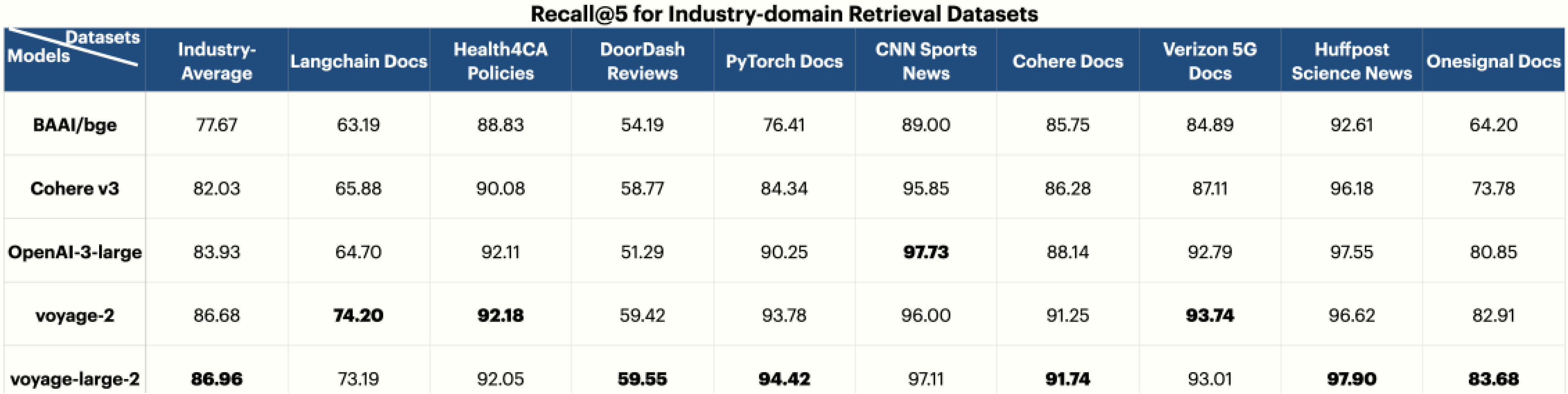
Voyage AI Products and Offerings


General-purpose Embedding Model
Code Embedding Model
Company X Embedding Model
Embedding models and search system with continual training with human feedback
Law Embedding Model
Finance Embedding Model
Multilingual Embedding Model

Offerings

1. Public API endpoints
2. Zilliz Cloud Pipeline
3. AWS Marketplace Model Package
4. AWS JumpStart
5. On-premise, customer-hosted deployment
6. In-Progress
   a. Snowflake Cortex Function
   b. Amazon Bedrock

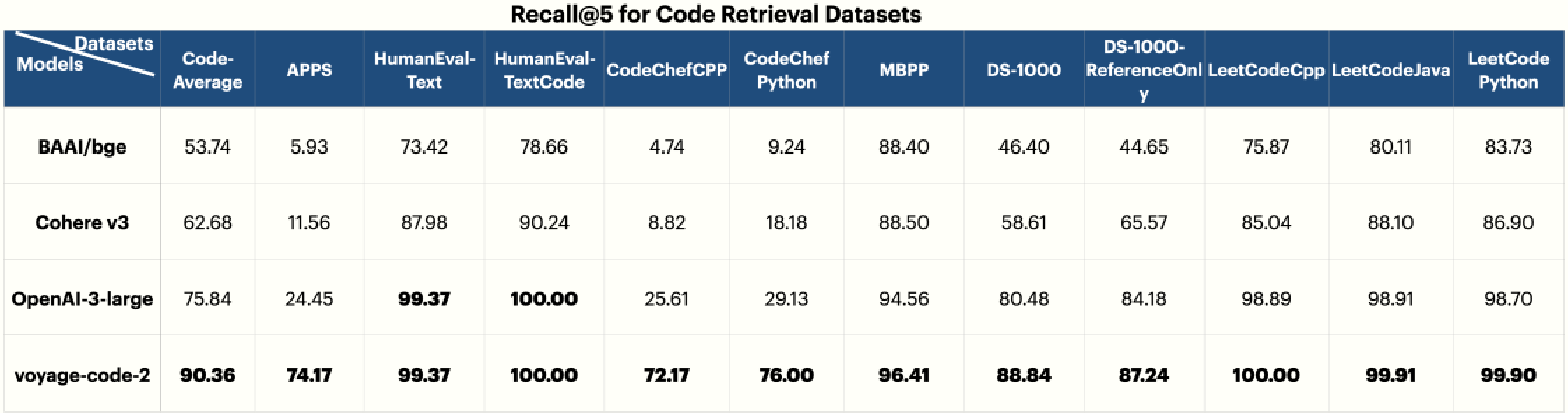
Voyage Models are State-of-the-Art in Retrieval Quality



Recall@5 for Industry-domain Retrieval Datasets

Datasets Models	Industry-Average	Langchain Docs	Health4CA Policies	DoorDash Reviews	PyTorch Docs	CNN Sports News	Cohere Docs	Verizon 5G Docs	Huffpost Science News	Onesignal Docs
BAAI/bge	77.67	63.19	88.83	54.19	76.41	89.00	85.75	84.89	92.61	64.20
Cohere v3	82.03	65.88	90.08	58.77	84.34	95.85	86.28	87.11	96.18	73.78
OpenAI-3-large	83.93	64.70	92.11	51.29	90.25	**97.73**	88.14	92.79	97.55	80.85
voyage-2	86.68	**74.20**	**92.18**	59.42	93.78	96.00	91.25	**93.74**	96.62	82.91
voyage-large-2	**86.96**	73.19	92.05	**59.55**	**94.42**	97.11	**91.74**	93.01	**97.90**	**83.68**

The Code Embedding Model Excels on Code Retrieval Tasks



Recall@5 for Code Retrieval Datasets

Datasets Models	Code-Average	APPS	HumanEval-Text	HumanEval-TextCode	CodeChefCPP	CodeChef Python	MBPP	DS-1000	DS-1000-ReferenceOnly	LeetCodeCpp	LeetCodeJava	LeetCode Python
BAAI/bge	53.74	5.93	73.42	78.66	4.74	9.24	88.40	46.40	44.65	75.87	80.11	83.73
Cohere v3	62.68	11.56	87.98	90.24	8.82	18.18	88.50	58.61	65.57	85.04	88.10	86.90
OpenAI-3-large	75.84	24.45	**99.37**	**100.00**	25.61	29.13	94.56	80.48	84.18	98.89	98.91	98.70
voyage-code-2	**90.36**	**74.17**	**99.37**	**100.00**	**72.17**	**76.00**	**96.41**	**88.84**	**87.24**	**100.00**	**99.91**	**99.90**

Early Pilot Partners

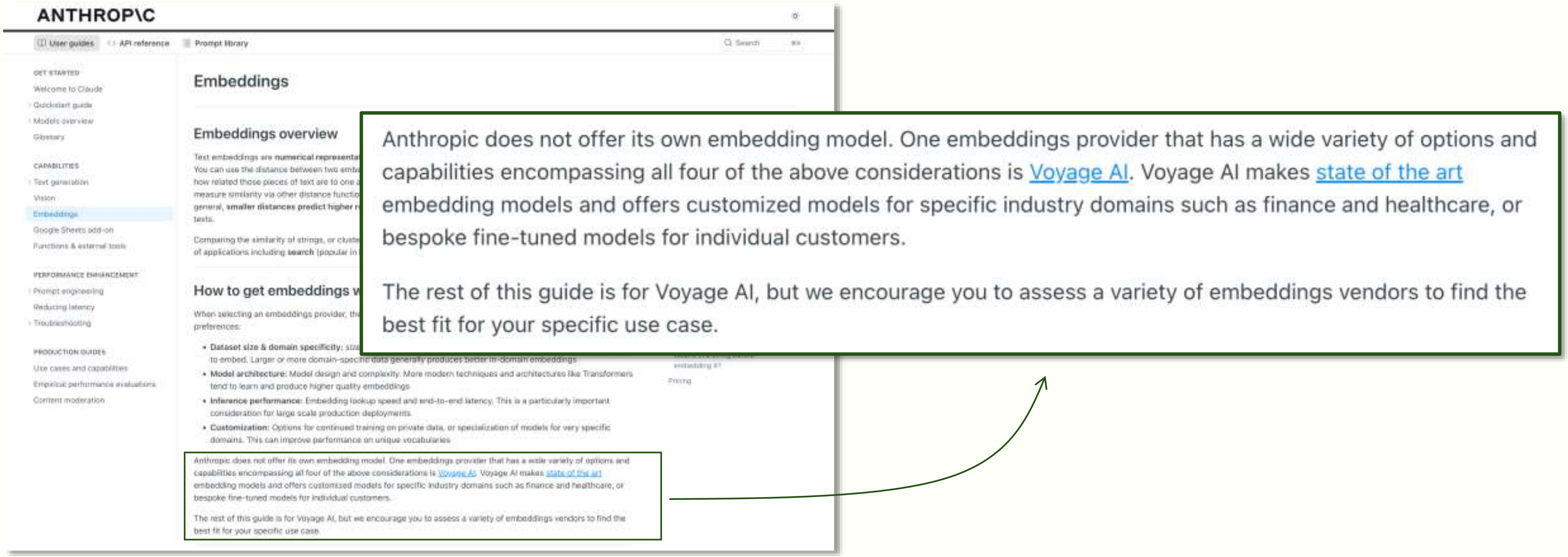
Harvey. Cursor replit

kapa.ai zoom Notion

LangChain LlamaIndex

ANTHROPIC weaviate zilliz snowflake

Trusted by Partners like ANTHROP\C



Embeddings

Embeddings overview

Text embeddings are **numerical representat**
You can use the distance between two emb
how related those pieces of text are to one a
measure similarity via other distance functio
general, **smaller distances predict higher r**
texts.

Comparing the similarity of strings, or cluste
of applications including **search** (popular in

How to get embeddings w

When selecting an embeddings provider, th
preferences:

- **Dataset size & domain specificity:** size
  to embed. Larger or more domain-specific data generally produces better in-domain embeddings
- **Model architecture:** Model design and complexity. More modern techniques and architectures like Transformers tend to learn and produce higher quality embeddings
- **Inference performance:** Embedding lookup speed and end-to-end latency. This is a particularly important consideration for large scale production deployments
- **Customization:** Options for continued training on private data, or specialization of models for very specific domains. This can improve performance on unique vocabularies

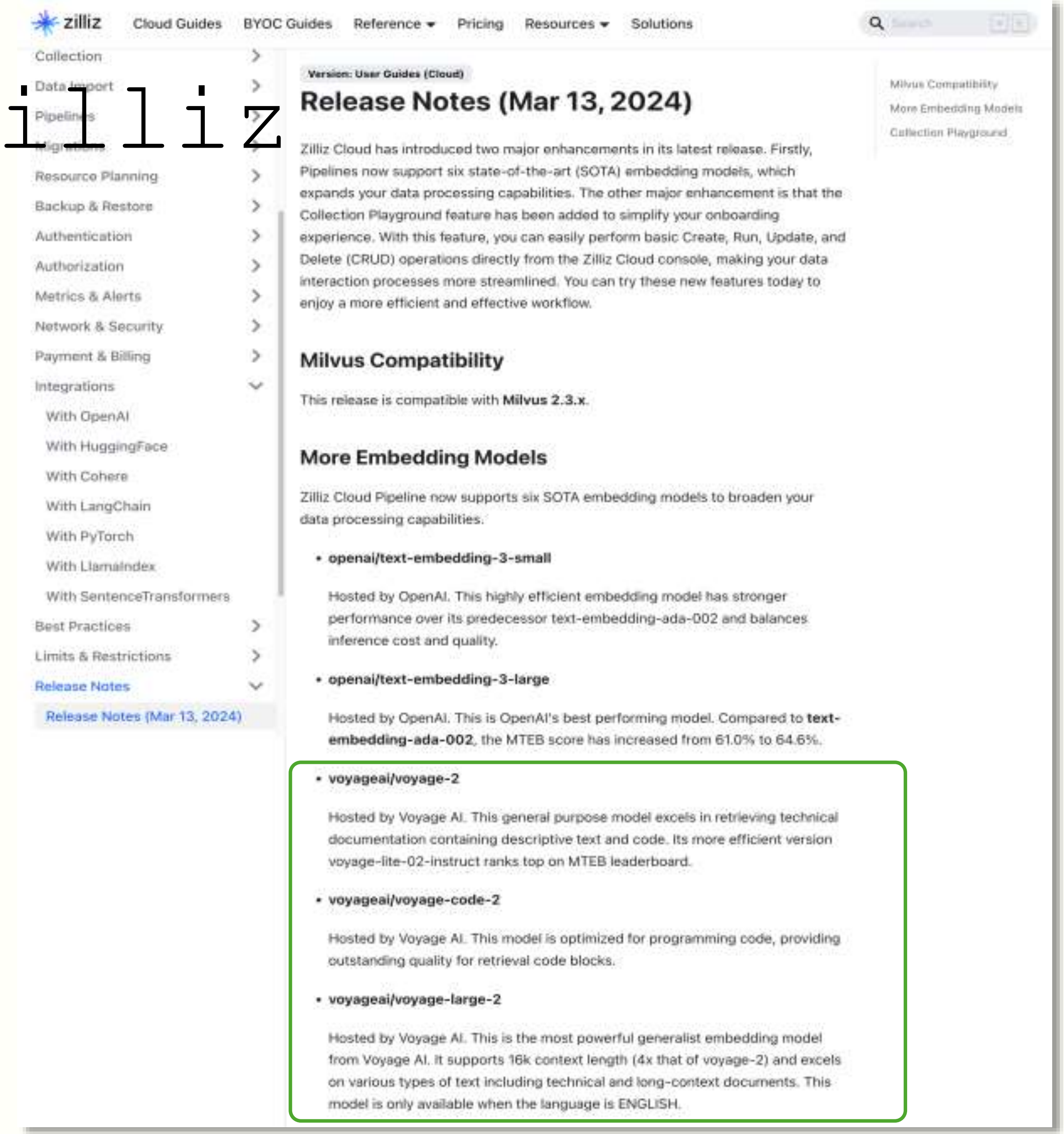
Anthropic does not offer its own embedding model. One embeddings provider that has a wide variety of options and capabilities encompassing all four of the above considerations is Voyage AI. Voyage AI makes state of the art embedding models and offers customized models for specific industry domains such as finance and healthcare, or bespoke fine-tuned models for individual customers.

The rest of this guide is for Voyage AI, but we encourage you to assess a variety of embeddings vendors to find the best fit for your specific use case.

Anthropic does not offer its own embedding model. One embeddings provider that has a wide variety of options and capabilities encompassing all four of the above considerations is Voyage AI. Voyage AI makes state of the art embedding models and offers customized models for specific industry domains such as finance and healthcare, or bespoke fine-tuned models for individual customers.

The rest of this guide is for Voyage AI, but we encourage you to assess a variety of embeddings vendors to find the best fit for your specific use case.

Trusted by Zilliz



Version: User Guides (Cloud)

Release Notes (Mar 13, 2024)

Zilliz Cloud has introduced two major enhancements in its latest release. Firstly, Pipelines now support six state-of-the-art (SOTA) embedding models, which expands your data processing capabilities. The other major enhancement is that the Collection Playground feature has been added to simplify your onboarding experience. With this feature, you can easily perform basic Create, Run, Update, and Delete (CRUD) operations directly from the Zilliz Cloud console, making your data interaction processes more streamlined. You can try these new features today to enjoy a more efficient and effective workflow.

Milvus Compatibility

This release is compatible with **Milvus 2.3.x**.

More Embedding Models

Zilliz Cloud Pipeline now supports six SOTA embedding models to broaden your data processing capabilities.

- **openai/text-embedding-3-small**

  Hosted by OpenAI. This highly efficient embedding model has stronger performance over its predecessor text-embedding-ada-002 and balances inference cost and quality.

- **openai/text-embedding-3-large**

  Hosted by OpenAI. This is OpenAI's best performing model. Compared to **text-embedding-ada-002**, the MTEB score has increased from 61.0% to 64.6%.

- **voyageai/voyage-2**

  Hosted by Voyage AI. This general purpose model excels in retrieving technical documentation containing descriptive text and code. Its more efficient version voyage-lite-02-instruct ranks top on MTEB leaderboard.

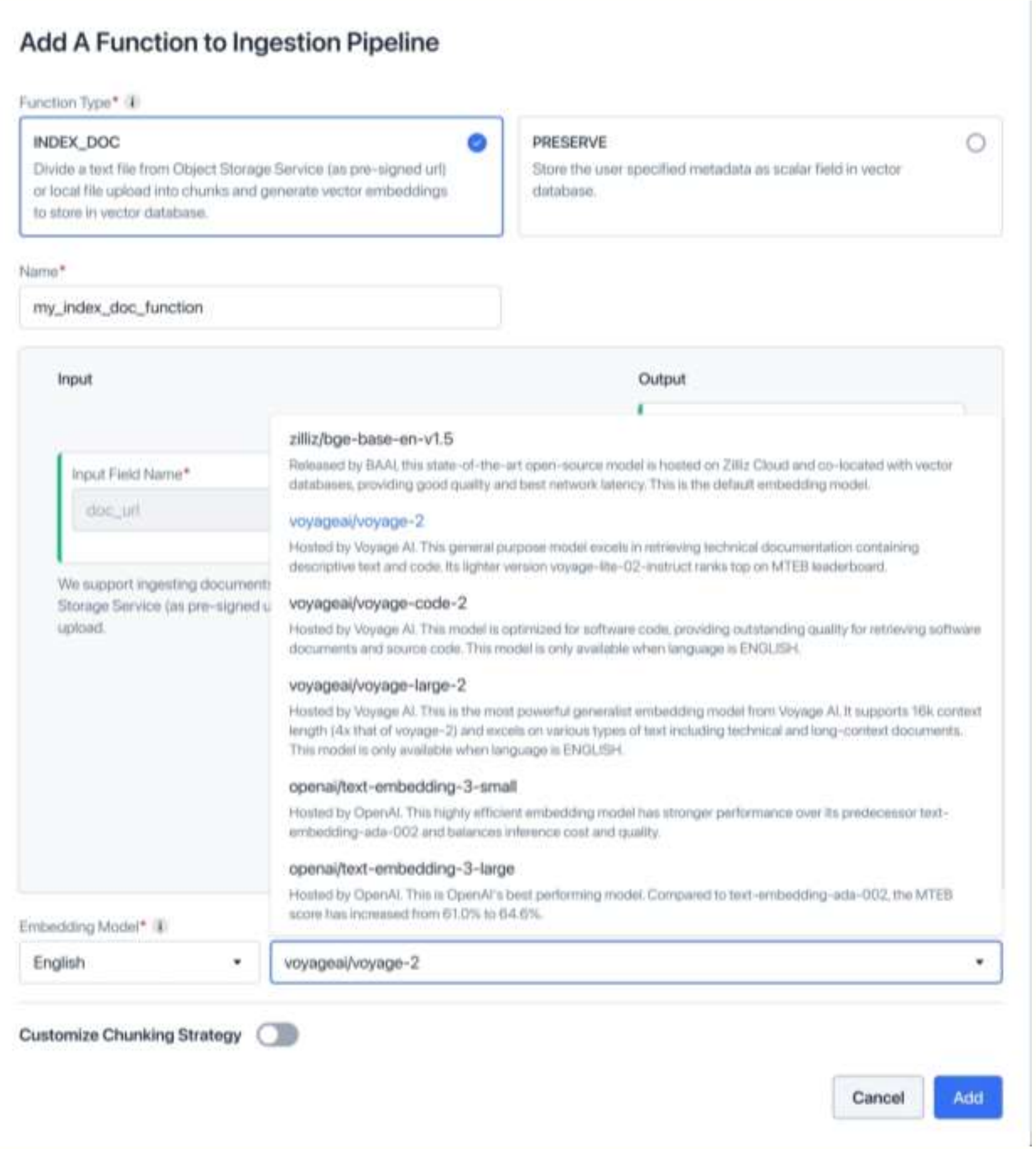
- **voyageai/voyage-code-2**

  Hosted by Voyage AI. This model is optimized for programming code, providing outstanding quality for retrieval code blocks.

- **voyageai/voyage-large-2**

  Hosted by Voyage AI. This is the most powerful generalist embedding model from Voyage AI. It supports 16k context length (4x that of voyage-2) and excels on various types of text including technical and long-context documents. This model is only available when the language is ENGLISH.

Use Voyage Embeddings on Zilliz Cloud Pipeline



Add A Function to Ingestion Pipeline

Function Type*

● **INDEX_DOC**
Divide a text file from Object Storage Service (as pre-signed url) or local file upload into chunks and generate vector embeddings to store in vector database.

○ **PRESERVE**
Store the user specified metadata as scalar field in vector database.

Name*: my_index_doc_function

Input

Input Field Name*: doc_url

We support ingesting document
Storage Service (as pre-signed u
upload.

Output

zilliz/bge-base-en-v1.5
Released by BAAI, this state-of-the-art open-source model is hosted on Zilliz Cloud and co-located with vector databases, providing good quality and best network latency. This is the default embedding model.

voyageai/voyage-2
Hosted by Voyage AI. This general purpose model excels in retrieving technical documentation containing descriptive text and code. Its lighter version voyage-lite-02-instruct ranks top on MTEB leaderboard.

voyageai/voyage-code-2
Hosted by Voyage AI. This model is optimized for software code, providing outstanding quality for retrieving software documents and source code. This model is only available when language is ENGLISH.

voyageai/voyage-large-2
Hosted by Voyage AI. This is the most powerful generalist embedding model from Voyage AI. It supports 16k context length (4x that of voyage-2) and excels on various types of text including technical and long-context documents. This model is only available when language is ENGLISH.

openai/text-embedding-3-small
Hosted by OpenAI. This highly efficient embedding model has stronger performance over its predecessor text-embedding-ada-002 and balances inference cost and quality.

openai/text-embedding-3-large
Hosted by OpenAI. This is OpenAI's best performing model. Compared to text-embedding-ada-002, the MTEB score has increased from 61.0% to 64.6%.

Embedding Model*: English ▾ voyageai/voyage-2 ▾

Customize Chunking Strategy