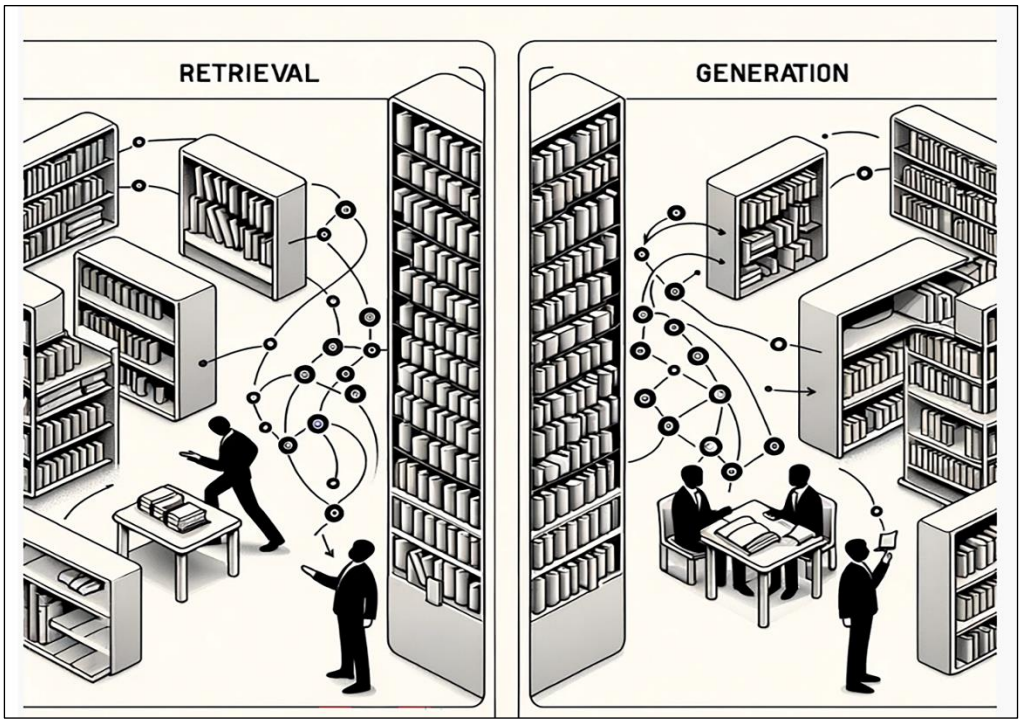
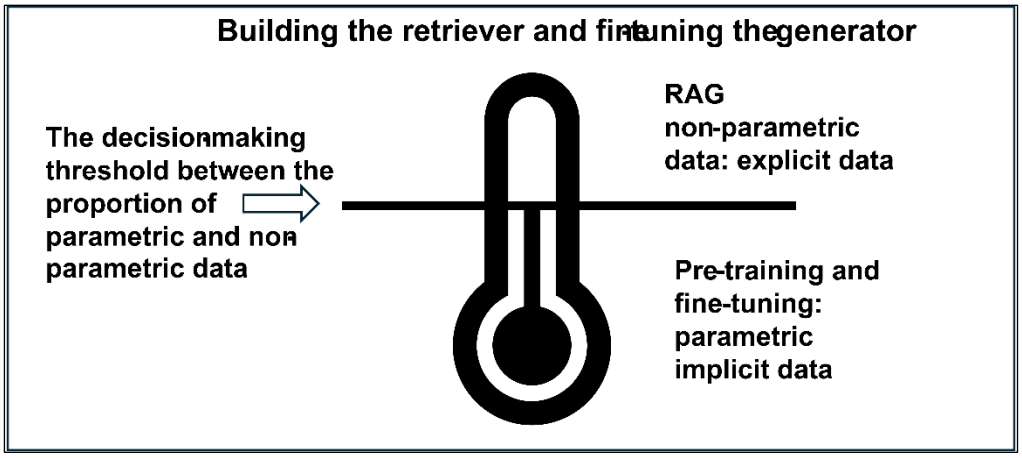
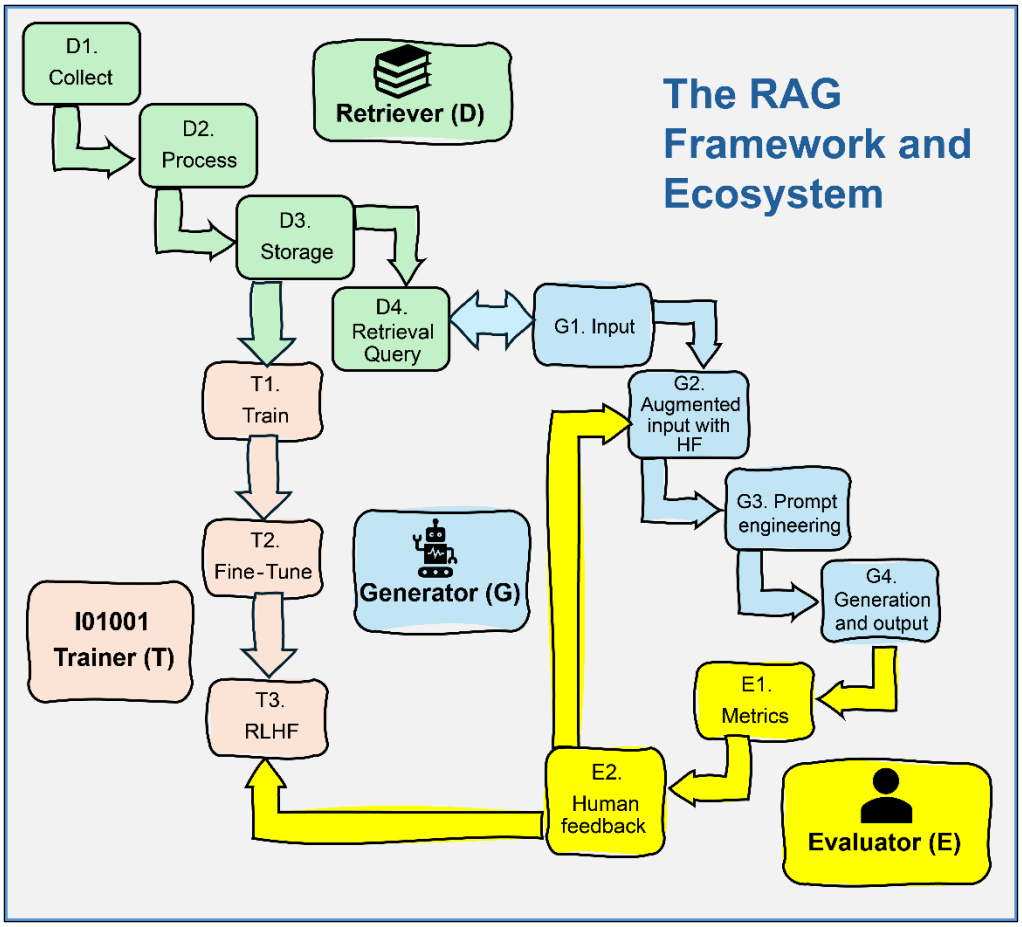
Chapter 1: Why Retrieval Augmented Generation?

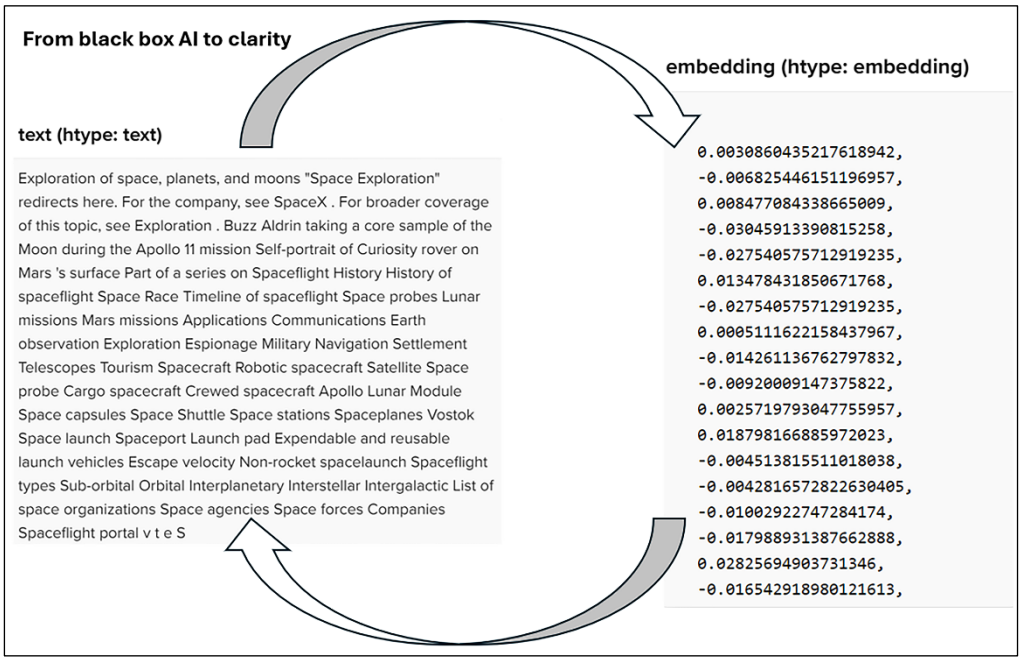


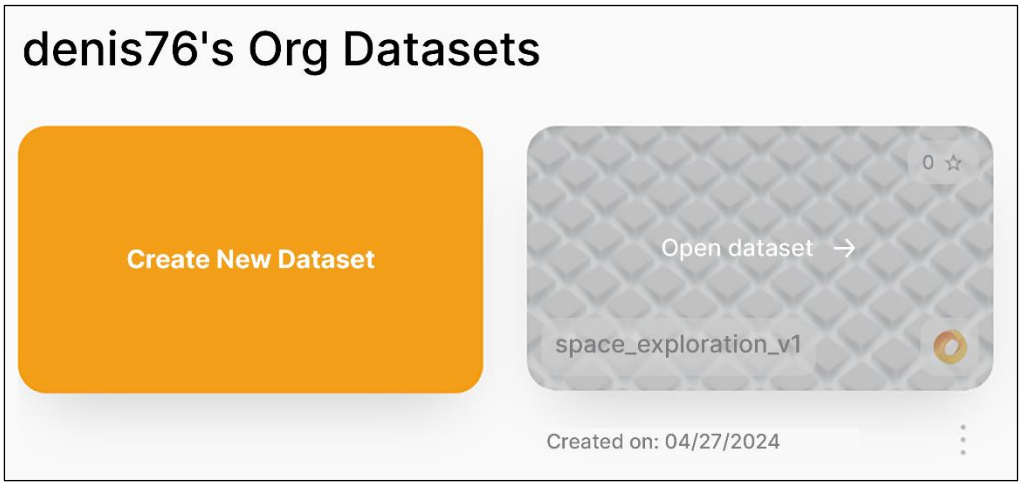


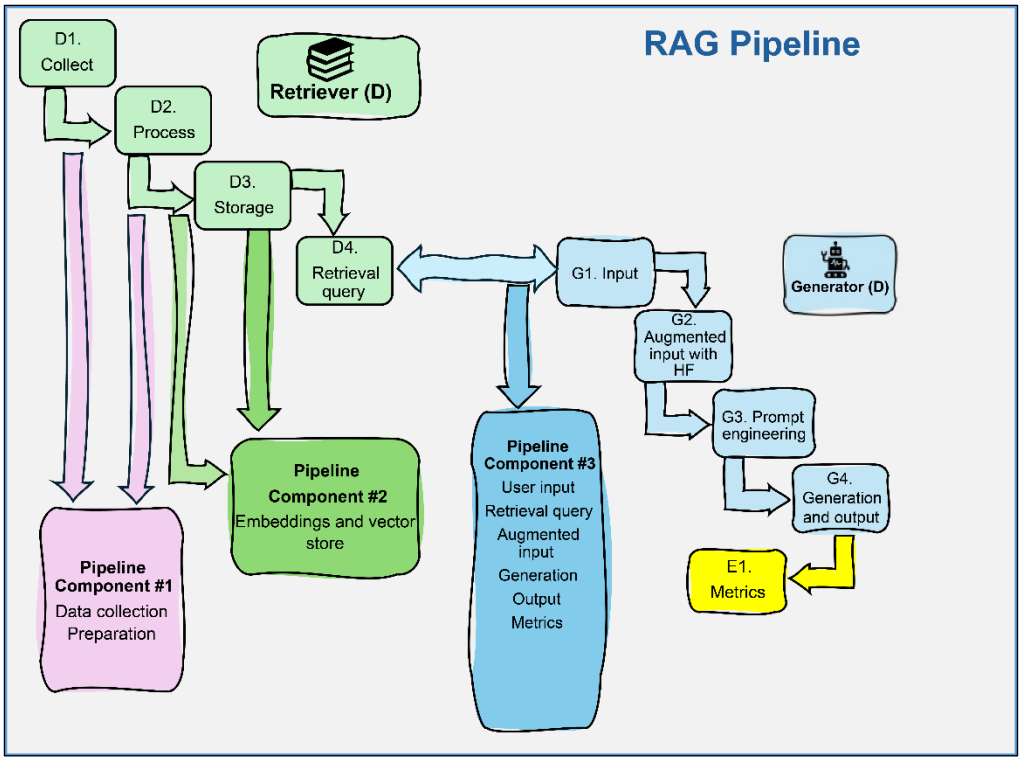


accurate	adapt	additional
0.000000	0.000000	0.000000
0.214779	0.000000	0.000000
0.000000	0.000000	0.000000
0.000000	0.000000	0.000000
0.000000	0.000000	0.236328

Chapter 2: RAG Embedding Vector Stores with Deep Lake and OpenAI

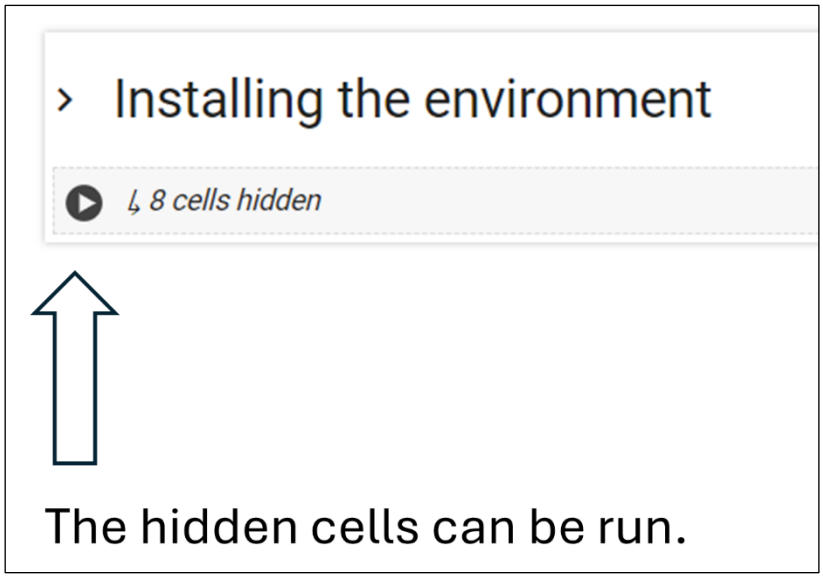




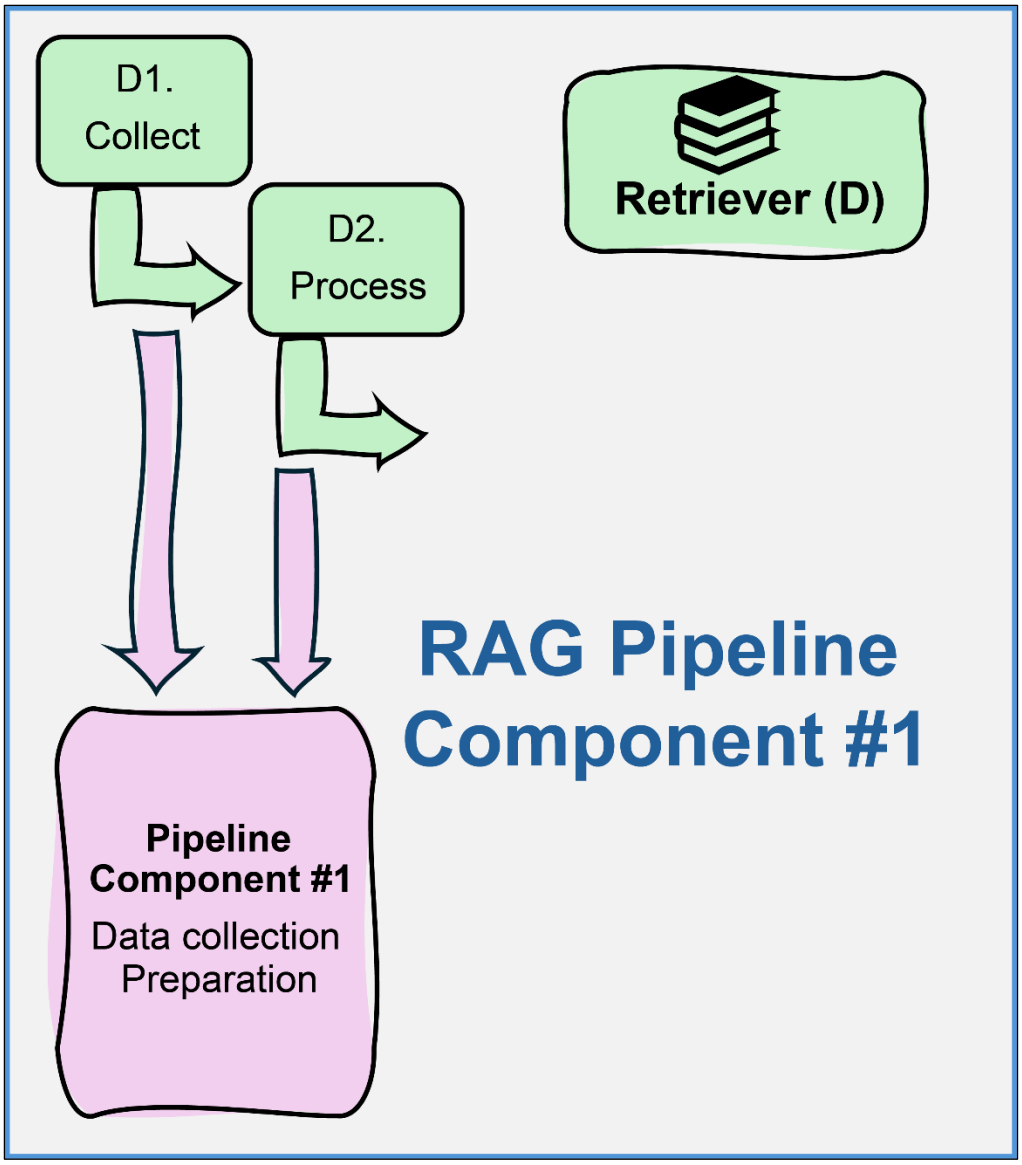


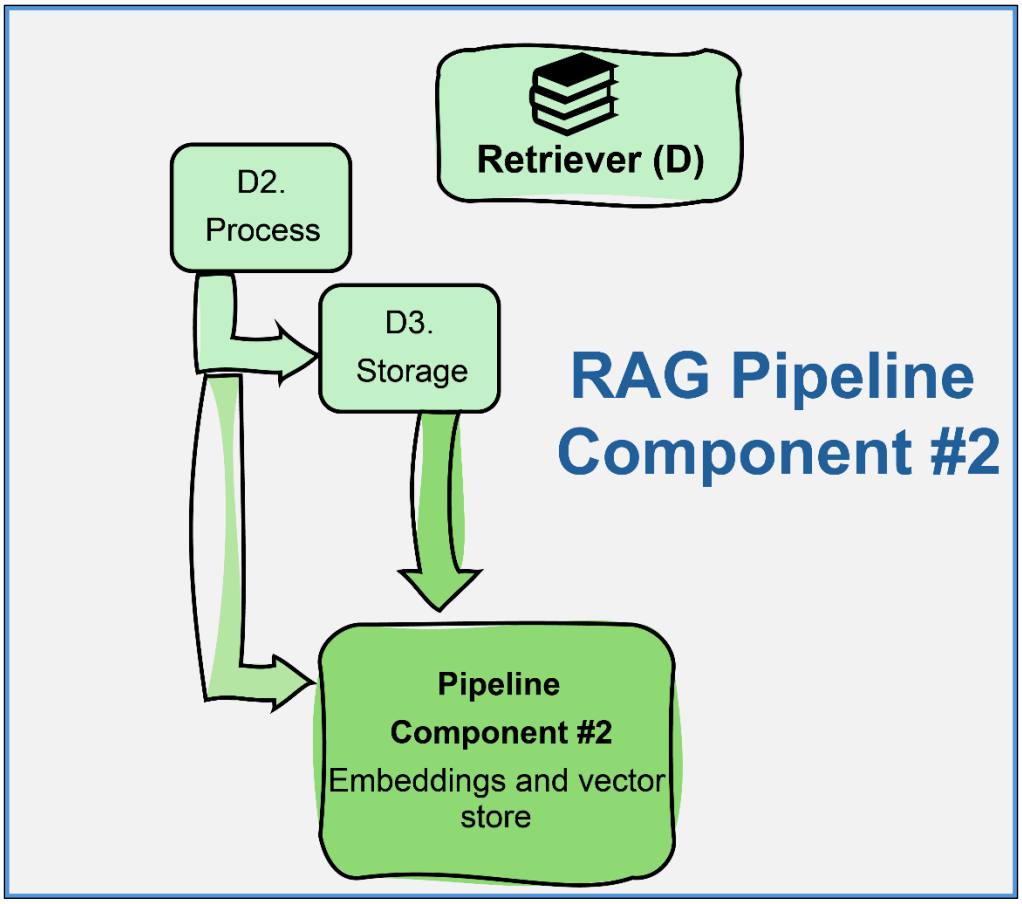
1. Installing the environment

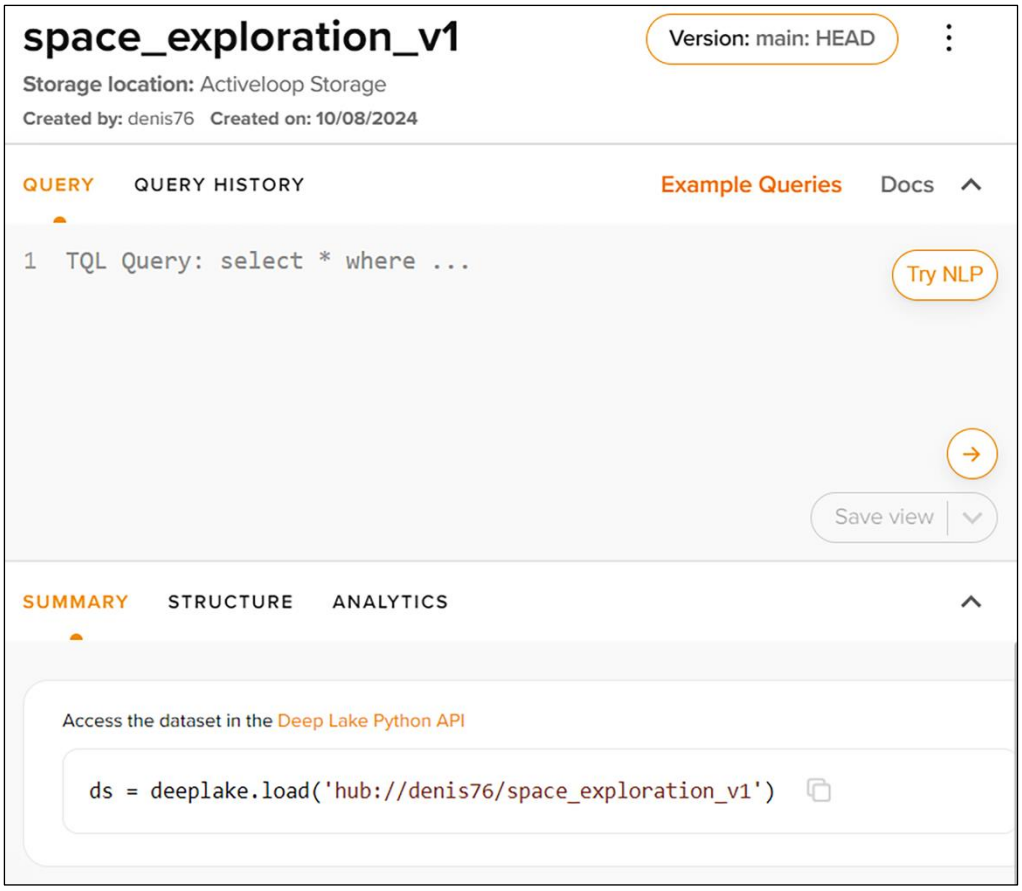
Click on the arrow to hide the installation cells.

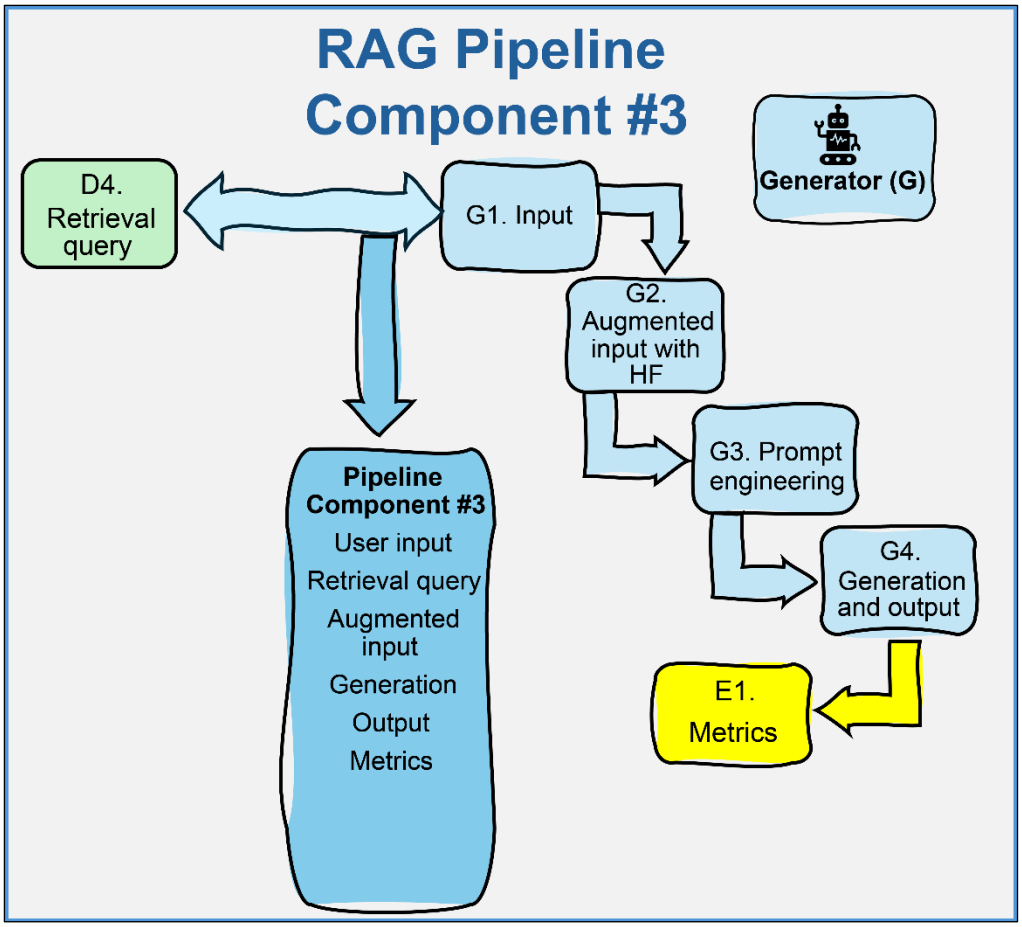


The hidden cells can be run.

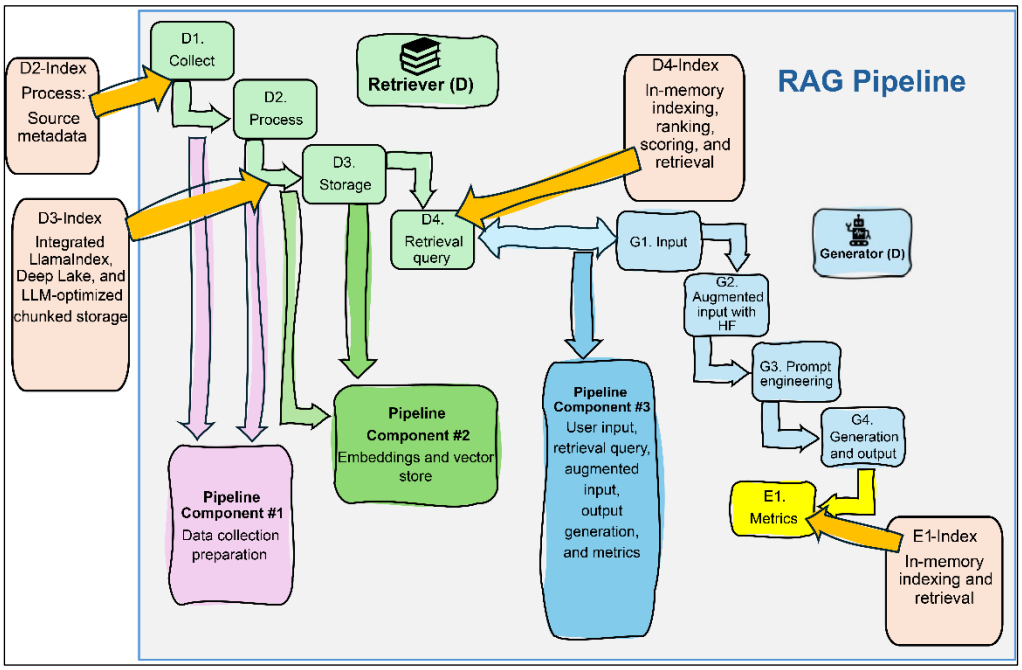

D1.
Collect
D2.
Process
Retriever (D)
RAG Pipeline
Component #1
Pipeline
Component #1
Data collection
Preparation

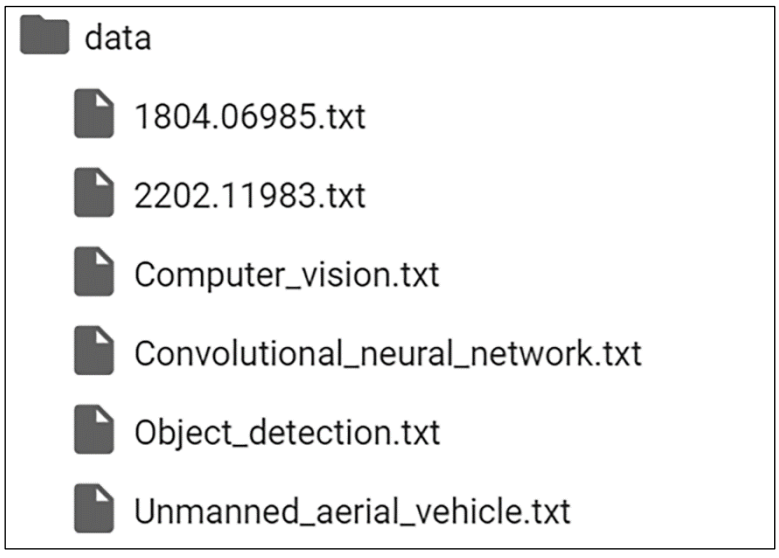

Retriever (D)
D2.
Process
D3.
Storage
RAG Pipeline
Component #2
Pipeline
Component #2
Embeddings and vector
store


space_exploration_v1
Version: main: HEAD
Storage location: Activeloop Storage
Created by: denis76 Created on: 10/08/2024
QUERY
QUERY HISTORY
Example Queries
Docs
1 TQL Query: select * where ...
Try NLP
Save view
SUMMARY
STRUCTURE
ANALYTICS
Access the dataset in the Deep Lake Python API
ds = deeplake.load('hub://denis76/space_exploration_v1')


RAG Pipeline Component #3
Generator (G)
D4. Retrieval query
G1. Input
G2. Augmented input with HF
G3. Prompt engineering
G4. Generation and output
E1. Metrics
Pipeline Component #3
User input
Retrieval query
Augmented input
Generation
Output
Metrics

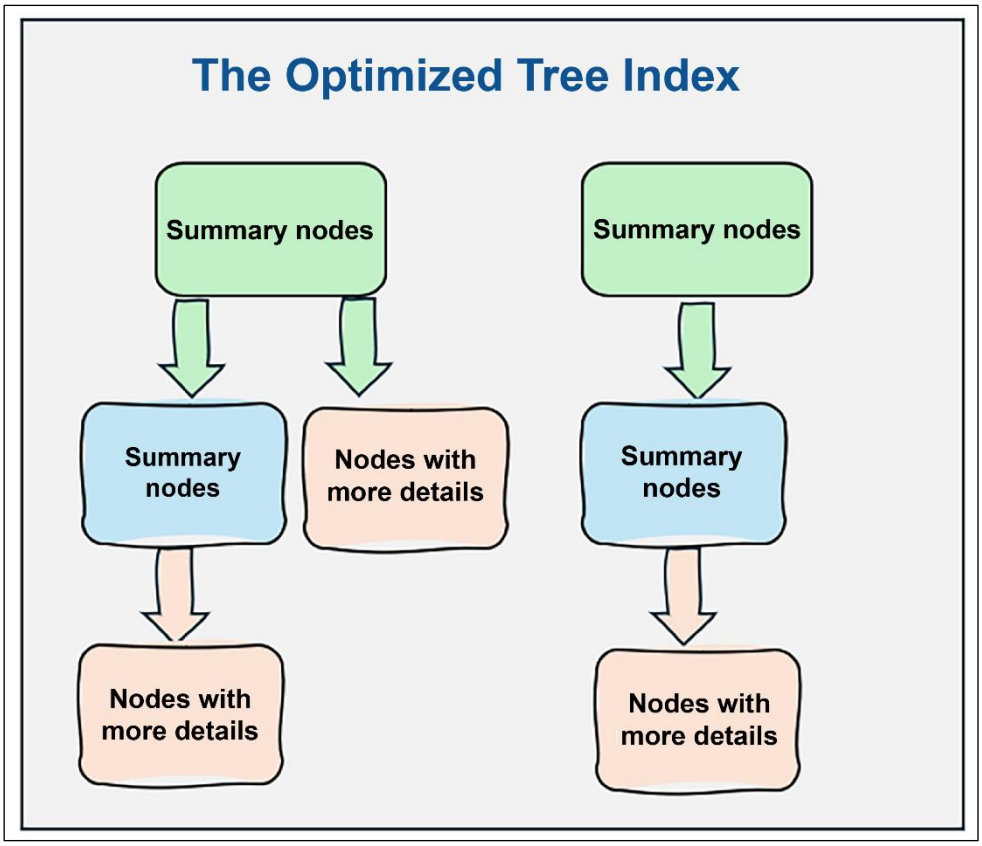
Chapter 3: Building Index-Based RAG with LlamaIndex, Deep Lake, and OpenAI





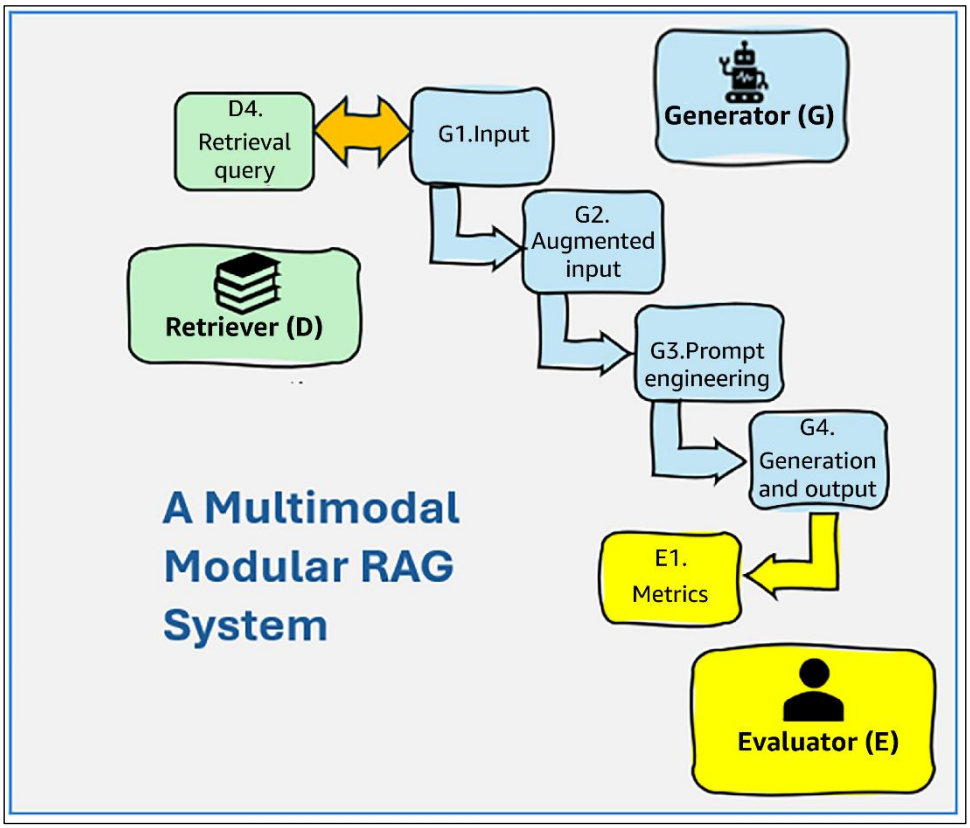
tensor	htype	shape	dtype	compression
text	text	(81, 1)	str	None
metadata	json	(81, 1)	str	None
embedding	embedding	(81, 1536)	float32	None
id	text	(81, 1)	str	None

Node ID	Score	Text
4befdb13-305d-42db-a616-5d9932c17ac8	0.833274	['These activities can be carried out with
bdefbbba-1b1a-4aea-812c-a2f6acb4dfb9	0.828414	['Degree of autonomy edit Drones could
381255c4-3bdb-455a-be73-d99cc6b46537	0.826526	['UAVs with generally nonlethal payloads



	Keyword	Document ID
0	black	48696b4b-978e-46b1-a5b3-f451d5265f42
1	asymptotically flat	48696b4b-978e-46b1-a5b3-f451d5265f42
2	high energy physics	48696b4b-978e-46b1-a5b3-f451d5265f42
3	entropy	48696b4b-978e-46b1-a5b3-f451d5265f42
4	extreme binary black hole geometry	48696b4b-978e-46b1-a5b3-f451d5265f42
...	...	...

Chapter 4: Multimodal Modular RAG for Drone Technology



	embedding	id	metadata	text
0	[-0.0009671939187683165, 0.010151553899049759,...	[a89cdb8c-3a85-42ff-9d5f-98f93f414df6]	[{'file_path': '/content/data/1804.06985.txt',...	[High Energy Physics Theory arXiv1804.06985 h...
1	[-0.015508202835917473, 0.017525529488921165, ...	[e9748376-1181-46a7-a376-d03b3baa82a5]	[{'file_path': '/content/data/2202.11983.txt',...	[Computer Science Computer Vision and Pattern...



	image	boxes	labels
0	b'\xff\xd8\xff\xe0\x00\x10JFIF\x00\x01\x01\x00...	[[1221.0, 84.0, 16.0, 33.0], [1235.0, 71.0, 18...	[pedestrian, pedestrian, tricycle, pedestrian,...
1	b'\xff\xd8\xff\xe0\x00\x10JFIF\x00\x01\x01\x00...	[[351.0, 936.0, 305.0, 114.0], [0.0, 818.0, 22...	[car, car, car, car, car, car, car, car, car, ...
2	b'\xff\xd8\xff\xe0\x00\x10JFIF\x00\x01\x01\x00...	[[699.0, 716.0, 26.0, 54.0], [600.0, 604.0, 22...	[truck, car, van, car, car, car, car, car, car...
3	b'\xff\xd8\xff\xe0\x00\x10JFIF\x00\x01\x01\x00...	[[417.0, 77.0, 57.0, 54.0], [387.0, 109.0, 31....	[ignored regions, ignored regions, car, car, c...
4	b'\xff\xd8\xff\xe0\x00\x10JFIF\x00\x01\x01\x00...	[[794.0, 617.0, 97.0, 115.0], [803.0, 539.0, 7...	[car, car, car, car, car, car, car, car, car, ...
...	...	...	...
6466	b'\xff\xd8\xff\xe0\x00\x10JFIF\x00\x01\x01\x00...	[[683.0, 710.0, 7.0, 6.0], [681.0, 717.0, 5.0,...	[car, car, car, car, awning-tricycle, car, car...
6467	b'\xff\xd8\xff\xe0\x00\x10JFIF\x00\x01\x01\x00...	[[1761.0, 676.0, 207.0, 155.0], [1539.0, 648.0...	[car, car, car, car, car, van, car, car, car, ...
6468	b'\xff\xd8\xff\xe0\x00\x10JFIF\x00\x01\x01\x00...	[[0.0, 501.0, 105.0, 93.0], [412.0, 594.0, 172...	[car, van, pedestrian, pedestrian, pedestrian,...
6469	b'\xff\xd8\xff\xe0\x00\x10JFIF\x00\x01\x01\x00...	[[200.0, 604.0, 112.0, 64.0], [311.0, 560.0, 1...	[van, truck, van, car, car, van, others, car, ...
6470	b'\xff\xd8\xff\xe0\x00\x10JFIF\x00\x01\x01\x00...	[[60.0, 144.0, 16.0, 21.0], [51.0, 120.0, 15.0...	[people, pedestrian, others, people, people, m...

6471 rows × 3 columns

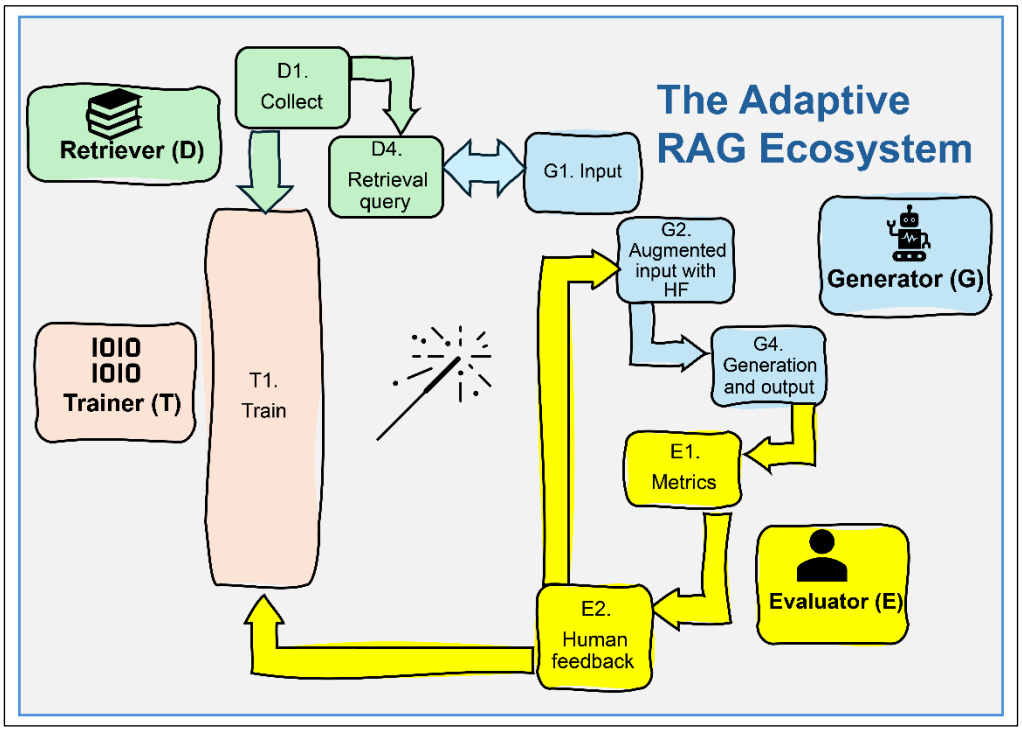


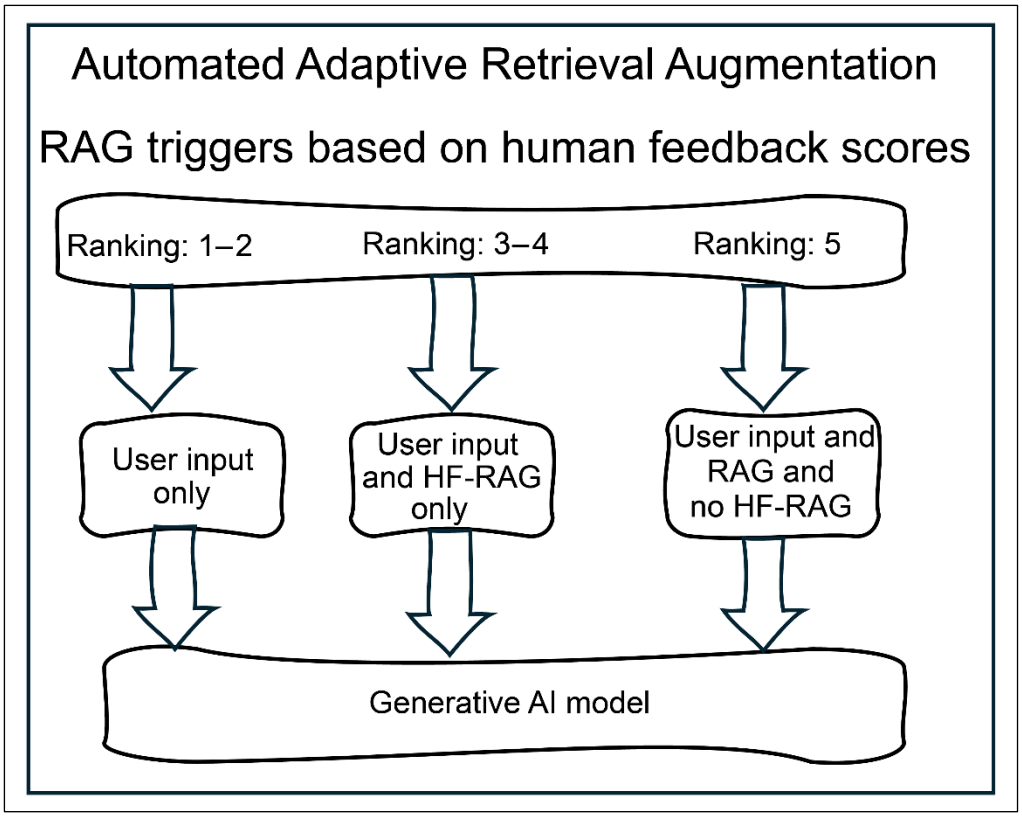


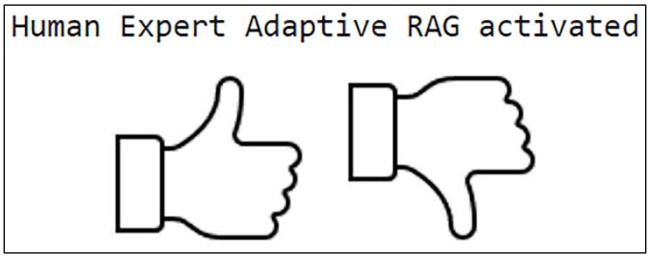


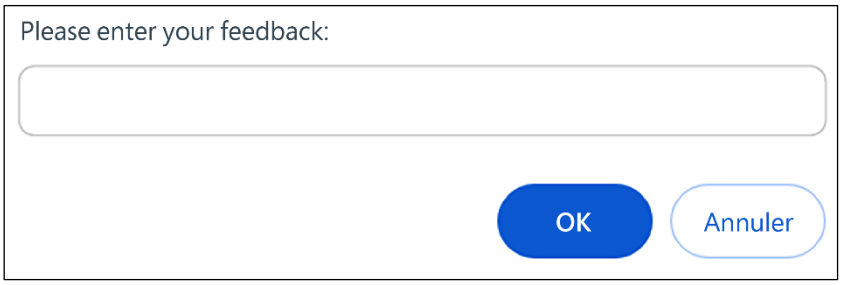


Chapter 5: Boosting RAG Performance with Expert Human Feedback

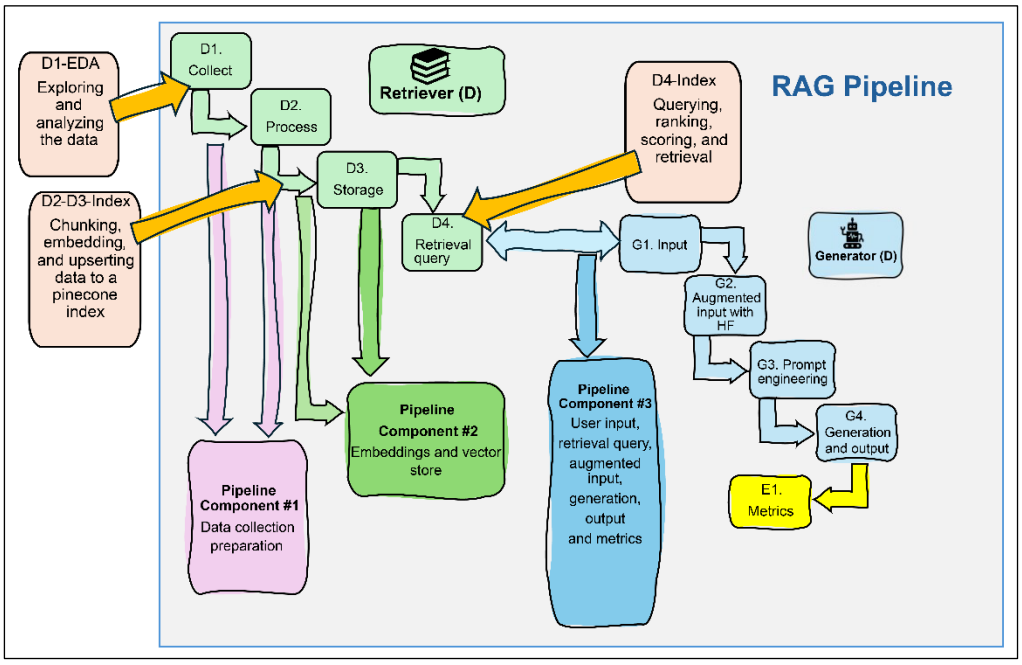



Automated Adaptive Retrieval Augmentation
RAG triggers based on human feedback scores
Ranking: 1–2
Ranking: 3–4
Ranking: 5
User input only
User input and HF-RAG only
User input and RAG and no HF-RAG
Generative AI model


Human Expert Adaptive RAG activated


Please enter your feedback:
OK
Annuler

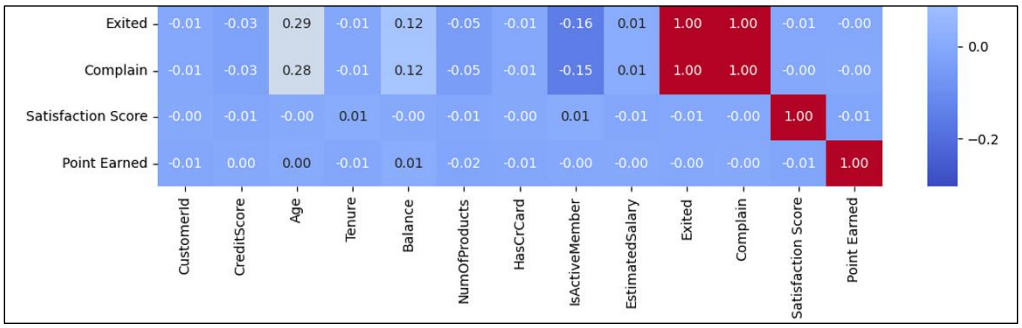
Chapter 6: Scaling RAG Bank Customer Data with Pinecone



	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
0	15634602	619	42	2	0.00	1	1	1	101348.88
1	15647311	608	41	1	83807.86	1	0	1	112542.58
2	15619304	502	42	8	159660.80	3	1	0	113931.57
3	15701354	699	39	1	0.00	2	0	0	93826.63
4	15737888	850	43	2	125510.82	1	1	1	79084.10
...	...	...	...	...	...	...	...	...	...
9995	15606229	771	39	5	0.00	2	1	0	96270.64
9996	15569892	516	35	10	57369.61	1	1	1	101699.77
9997	15584532	709	36	7	0.00	1	0	1	42085.58
9998	15682355	772	42	3	75075.31	2	1	0	92888.52
9999	15628319	792	28	4	130142.79	1	1	0	38190.78

0000 rows × 14 columns

Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Complain
42	2	0.00	1	1	1	101348.88	1	1
41	1	83807.86	1	0	1	112542.58	0	1
42	8	159660.80	3	1	0	113931.57	1	1
39	1	0.00	2	0	0	93826.63	0	0
43	2	125510.82	1	1	1	79084.10	0	0
...	...	...	...	...	...	...	...	...
39	5	0.00	2	1	0	96270.64	0	0
35	10	57369.61	1	1	1	101699.77	0	0
36	7	0.00	1	0	1	42085.58	1	1
42	3	75075.31	2	1	0	92888.52	1	1
28	4	130142.79	1	1	0	38190.78	0	0



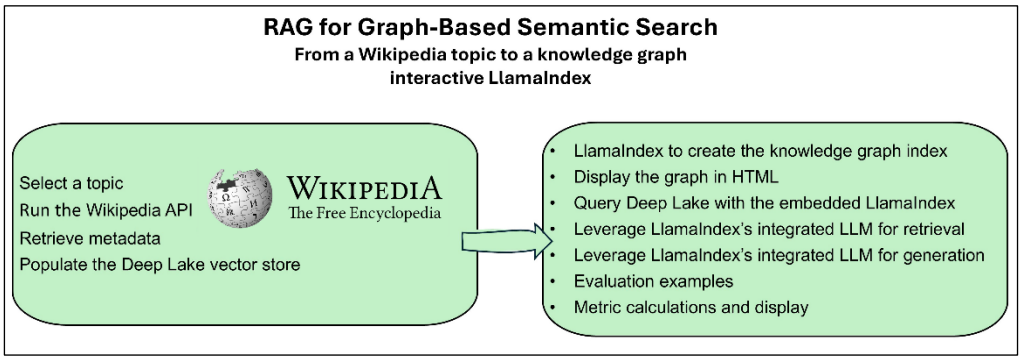
Exited	Complain	Satisfaction Score	Card Type	Point Earned	class
1	1	2	DIAMOND	464	0
0	1	3	DIAMOND	456	0
1	1	3	DIAMOND	377	0
0	0	5	GOLD	350	1
0	0	5	GOLD	425	1
...	...	...	...	...	...
0	0	1	DIAMOND	300	1
0	0	5	PLATINUM	771	1
1	1	3	SILVER	564	0
1	1	2	GOLD	339	0
0	0	3	DIAMOND	911	1

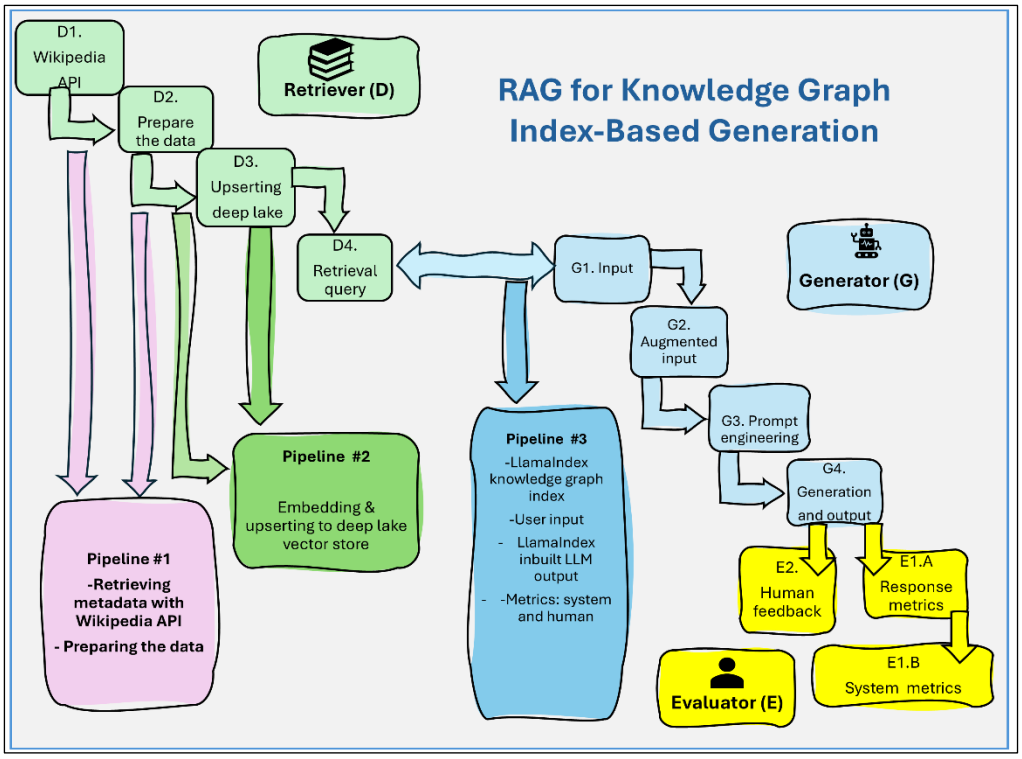
bank-index-50000

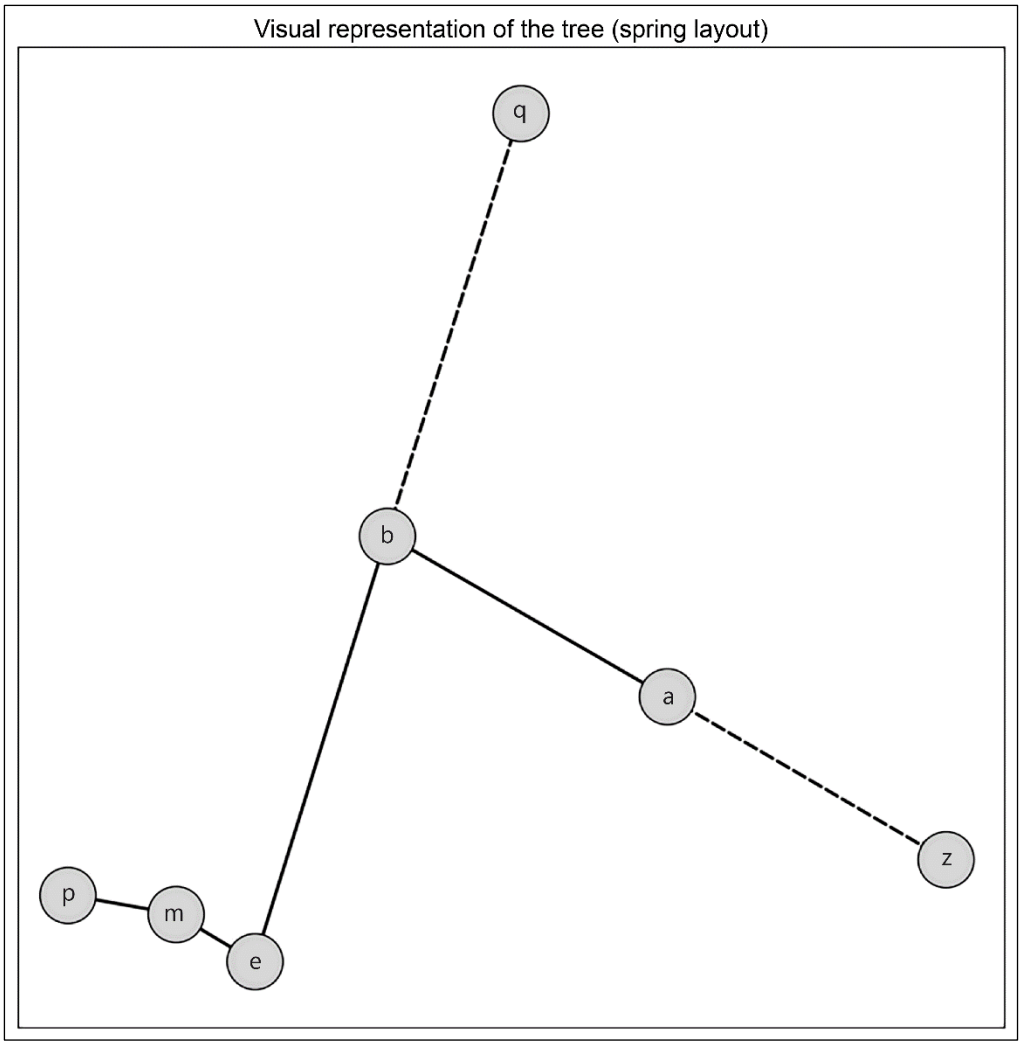
Host: https://bank-index-50000-ae302c8.svc.aped-4627-b74a.pinecone.io

Cloud: AWS • **Region:** us-east-1 • **Type:** Serverless • **Dimension:** 1536

Chapter 7: Building Scalable Knowledge-Graph-Based RAG with Wikipedia API and LlamaIndex






Visual representation of the tree (spring layout)
q
b
a
z
p
m
e

Pipeline 1: Collecting and preparing the documents

`Wikipedia_API.ipynb`

- **Selects a Wikipedia topic**
- **Generates a summary of the root Wikipedia page**
- **Retrieves the links from the root page**

`Wikipedia_API.ipynb`

- **Generates summaries, citations, and URLs for each page.**
- **Saves the information in a citations file and a URL file**

`Knowledge_Graph_Deep_Lake_LlamaIndex_OpenAI_RAG.ipynb`

- **Fetches the documents of the URLs**
- **Cleans the documents**
- **Loads and aggregates them for the next step: upserting**

Pipeline 3: Knowledge Graph Index-Based RAG

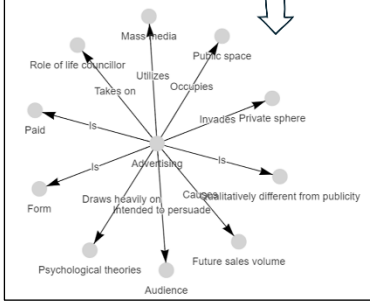
`Knowledge_Graph_Deep_Lake_LlamaIndex_OpenAI_RAG.ipynb`

Generating the knowledge base index with LlamaIndex and displaying the graph

Interacting with the graph index

- **User input**
- **LLM RAG parameters**

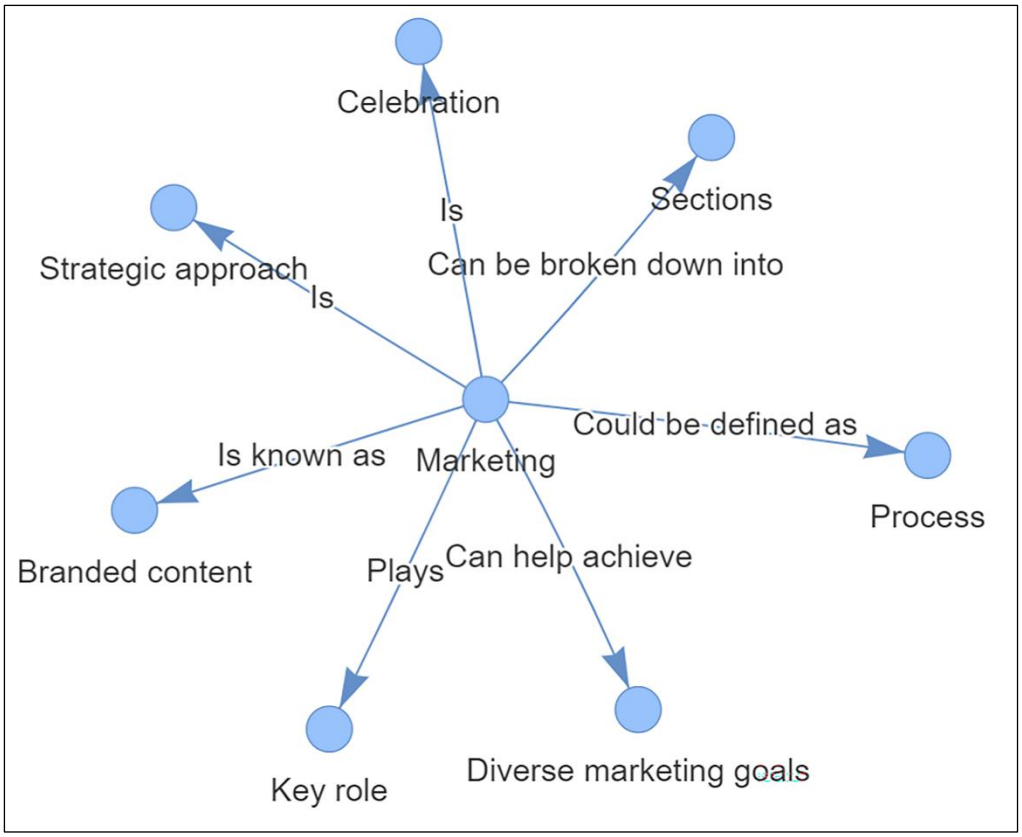
Installing the similarity score packages and defining the functions



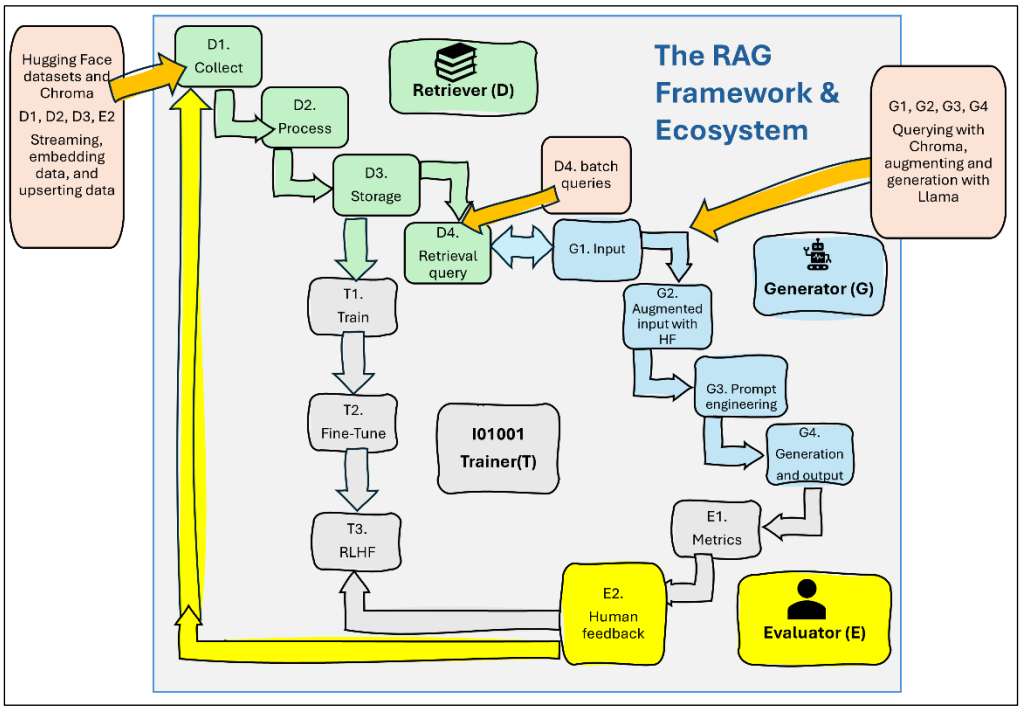
Evaluations

- **Reranking the output vectors**
- **Running examples and applying metrics and human feedback**
- **Global metric calculations and display**

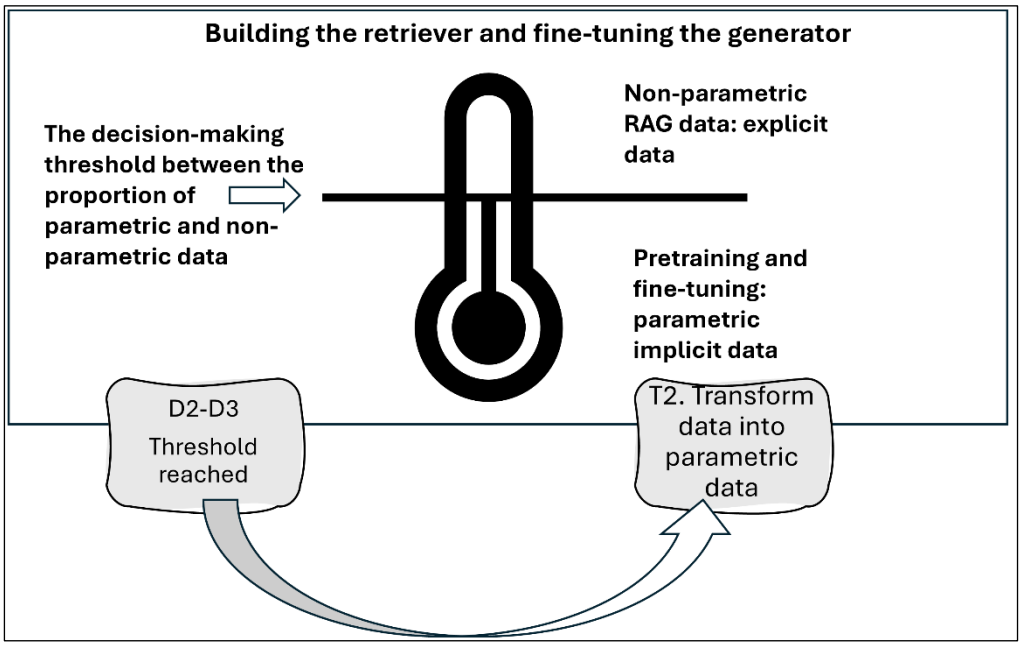
```
Mean: 0.68 Median: 0.71 Standard Deviation:
0.15 Variance: 0.02 Minimum: 0.45 Maximum:
0.90 Range: 0.46 25th Percentile (Q1): 0.56
75th Percentile (Q3): 0.80 Interquartile
Range (IQR): 0.24 F1 Score: 0.77 Precision:
0.83 Recall: 0.71 Accuracy: 0.70
```

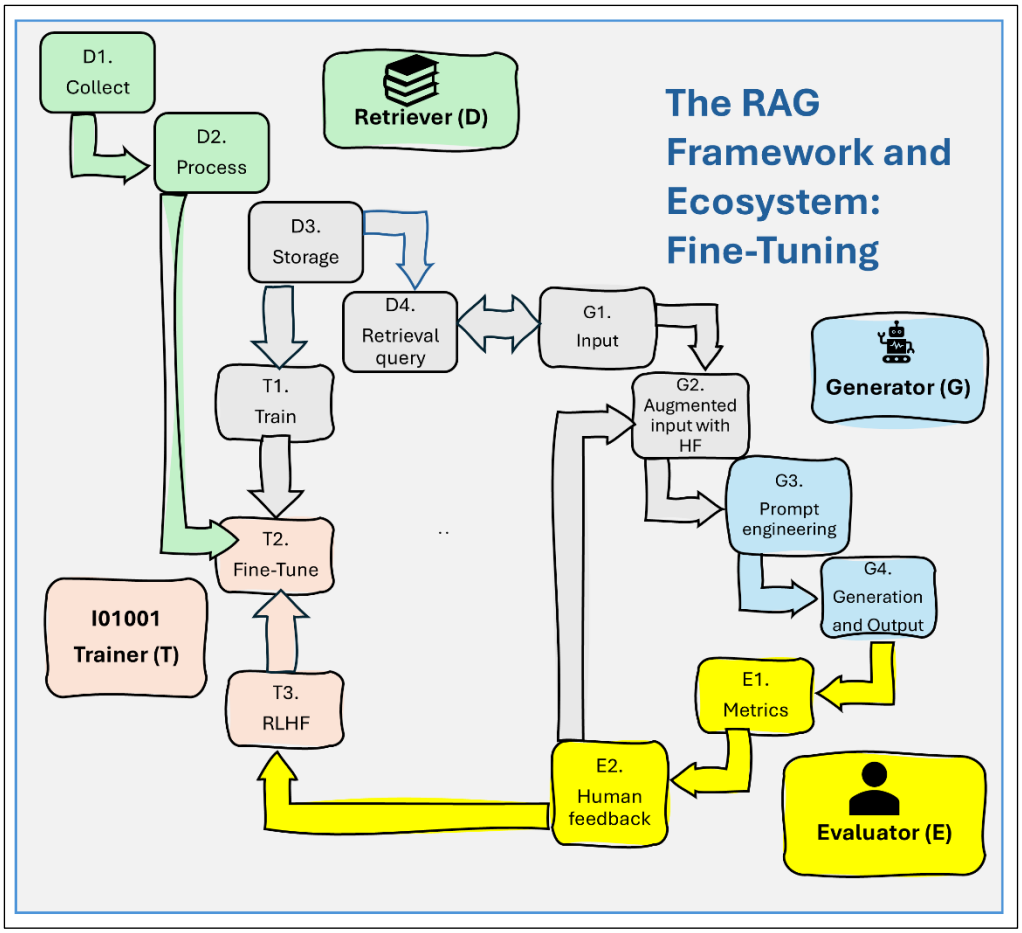

Celebration
Is
Sections
Can be broken down into
Strategic approach
Is
Could be defined as
Is known as
Marketing
Process
Branded content
Plays
Can help achieve
Key role
Diverse marketing goals

Chapter 8: Dynamic RAG with Chroma and Hugging Face Llama



Chapter 9: Empowering AI Models: Fine-Tuning RAG Data and Human Feedback



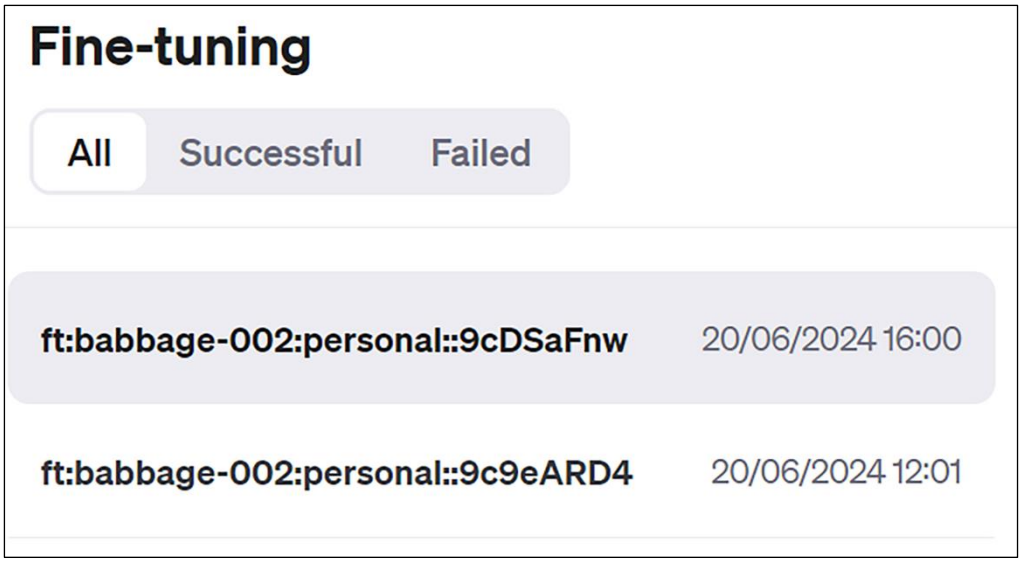


	question	correct_answer	support
0	What type of organism is commonly used in prep...	mesophilic organisms	Mesophiles grow best in moderate temperature, ...
1	What phenomenon makes global winds blow northe...	coriolis effect	Without Coriolis Effect the global winds would...
2	Changes from a less-ordered state to a more-or...	exothermic	Summary Changes of state are examples of phase...
3	What is the least dangerous radioactive decay?	alpha decay	All radioactive decay is dangerous to living t...
4	Kilauea in hawaii is the world's most continuo...	smoke and ash	Example 3.5 Calculating Projectile Motion: Hot...

index	messages
0	{'role': 'system', 'content': 'Given a science question, provide the correct answer with a detailed explanation.'},{'role': 'user', 'content': 'What type of organism is commonly used in preparation of foods such as cheese and yogurt?'}, {'role': 'assistant', 'content': 'mesophilic organisms Explanation: Mesophiles grow best in moderate temperature, typically between 25°C and 40°C (77°F and 104°F). Mesophiles are often found living in or on the bodies of humans or other animals. The optimal growth temperature of many pathogenic mesophiles is 37°C (98°F), the normal human body temperature. Mesophilic organisms have important uses in food preparation, including cheese, yogurt, beer and wine.'}
1	{'role': 'system', 'content': 'Given a science question, provide the correct answer with a detailed explanation.'},{'role': 'user', 'content': 'What phenomenon makes global winds blow northeast to southwest or the reverse in the northern hemisphere and northwest to southeast or the reverse in the southern hemisphere?'},{'role': 'assistant', 'content': 'coriolis effect Explanation: Without Coriolis Effect the global winds would blow north to south or south to north. But Coriolis makes them blow northeast to southwest or the reverse in the Northern Hemisphere. The winds blow northwest to southeast or the reverse in the southern hemisphere.'}

	Job ID	Created At	Status	Model	Training File	Error Message	Fine-Tuned Model
0	ftjob-O1OEE7eEyFNJsO2Eu5otzWA8	2024-09-02 15:00:50	running	gpt-4o-mini-2024-07-18	file-EUPGmm1yAd3axrQ0pyoeAKuE	None	None
1	ftjob-gQGiuvPMvSop0tzGaDn1NMql	2024-09-02 14:26:35	succeeded	gpt-4o-mini-2024-07-18	file-1OyEhi0D2b1kcL54JbQ3P1Pa	None	ft:gpt-4o-mini-2024-07-18:personal::A32qtJOo
2	ftjob-oVB0RAcwn3NEi4u0qOMeMZUF	2024-09-02 14:11:28	succeeded	gpt-4o-mini-2024-07-18	file-0dxQmL84uLME7ehnGljQAxit	None	ft:gpt-4o-mini-2024-07-18:personal::A32VfYIz

coriolis effect	Without Coriolis Effect the global winds would blow north to south or south to north. But Coriolis makes them blow northeast to southwest or the reverse in the Northern Hemisphere. The winds blow northwest to southeast or the reverse in the southern hemisphere.

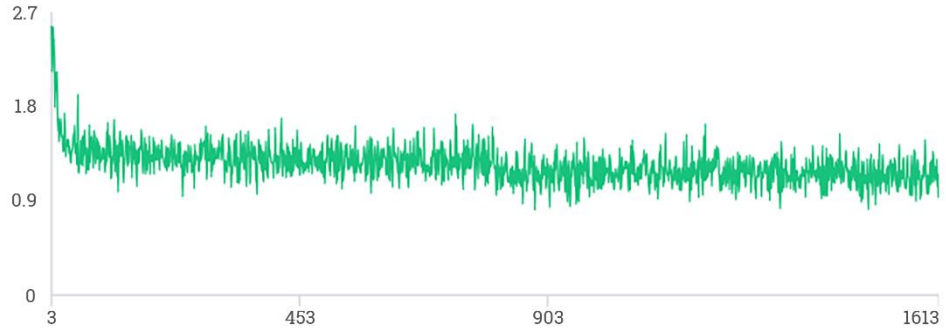


MODEL

ft:babbage-002:personal::9cDSaFnw

Status	Succeeded
Job ID	ftjob-DrZsgyCQsCSvl5wJkWY5omJf
Base model	babbage-002
Output model	ft:babbage-002:personal::9cDSaFnw
Created at	20 juin 2024, 16:00
Trained tokens	4861130
Epochs	5
Batch size	4
LR multiplier	0.1
Seed	1369873366

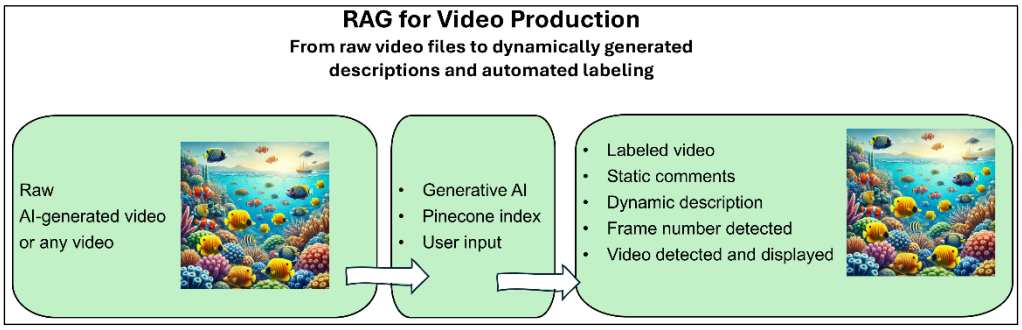
Training loss 1.1570

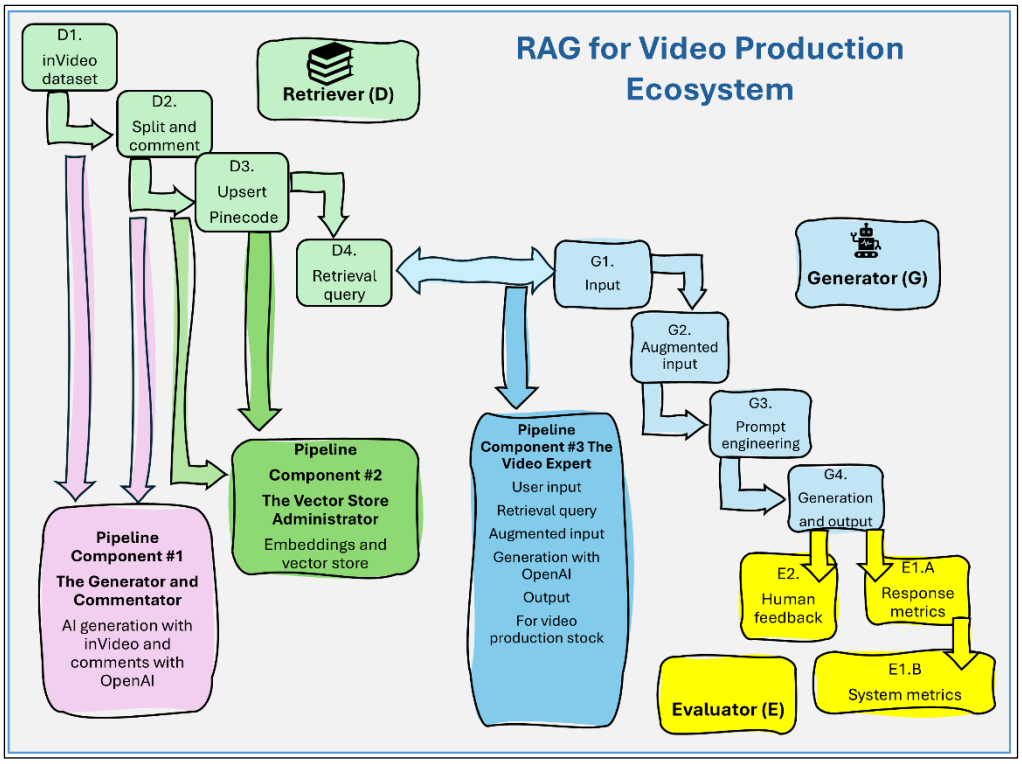


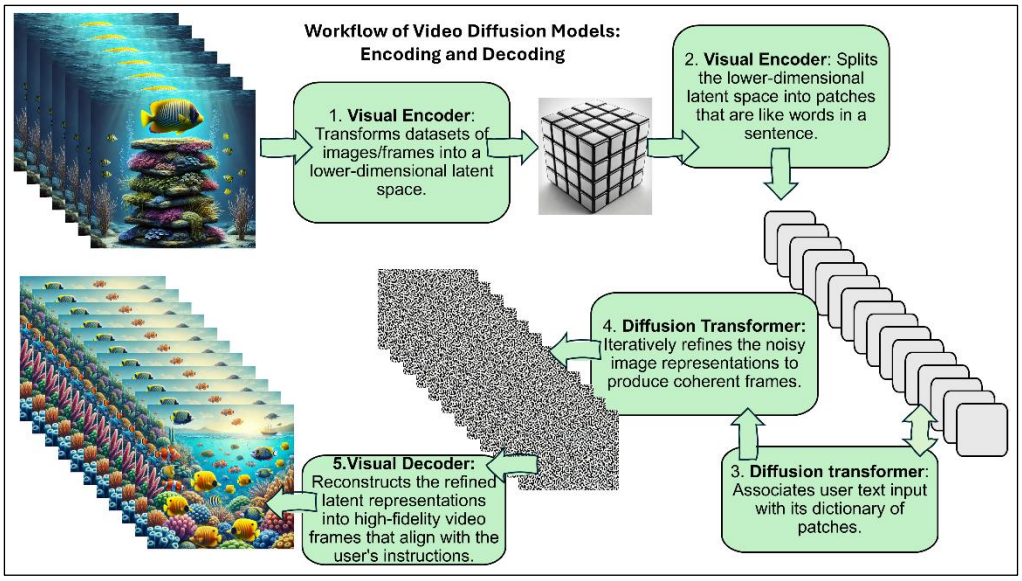
Messages | Metrics

Time	Step	Training loss
18:22:58	1613	1.1570
18:22:55	1612	0.9320
18:22:39	1611	1.0838
18:22:37	1610	1.0339
18:22:37	1609	1.2993
18:22:37	1608	1.1136
18:22:35	1607	1.1126
18:22:35	1606	1.2236
18:22:34	1605	1.0821

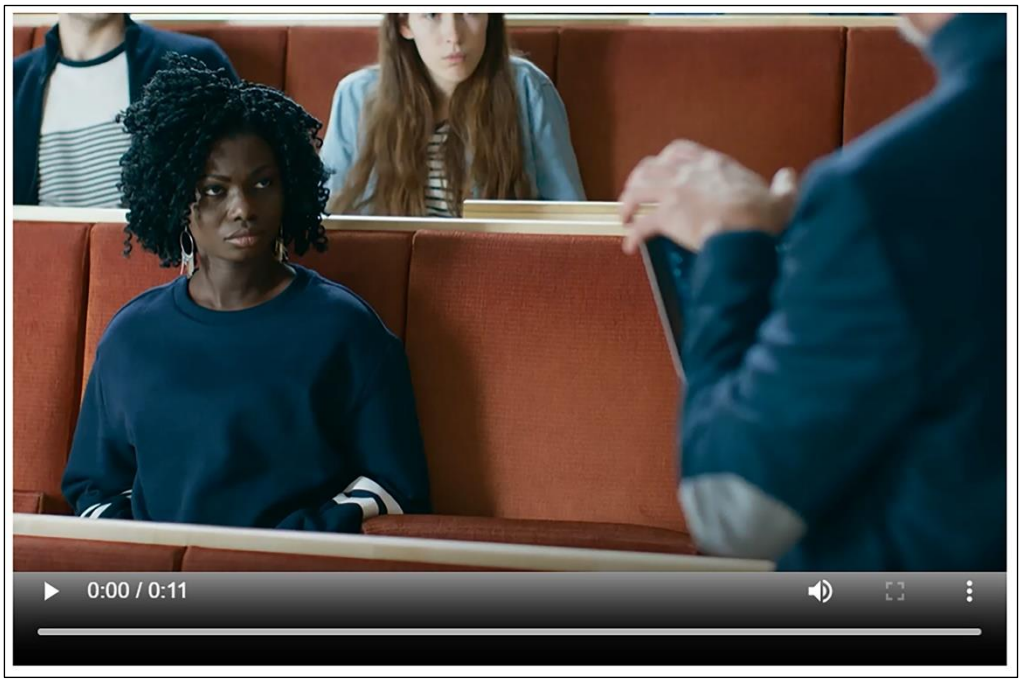
Chapter 10: RAG for Video Stock Production with Pinecone and OpenAI





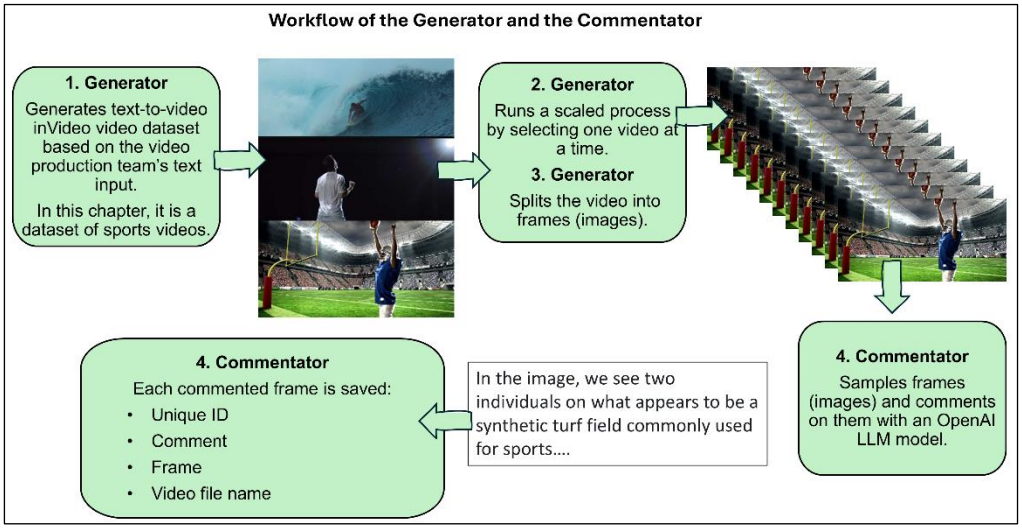

Workflow of Video Diffusion Models:
Encoding and Decoding
1. Visual Encoder: Transforms datasets of images/frames into a lower-dimensional latent space.
2. Visual Encoder: Splits the lower-dimensional latent space into patches that are like words in a sentence.
3. Diffusion transformer: Associates user text input with its dictionary of patches.
4. Diffusion Transformer: Iteratively refines the noisy image representations to produce coherent frames.
5.Visual Decoder: Reconstructs the refined latent representations into high-fidelity video frames that align with the user's instructions.

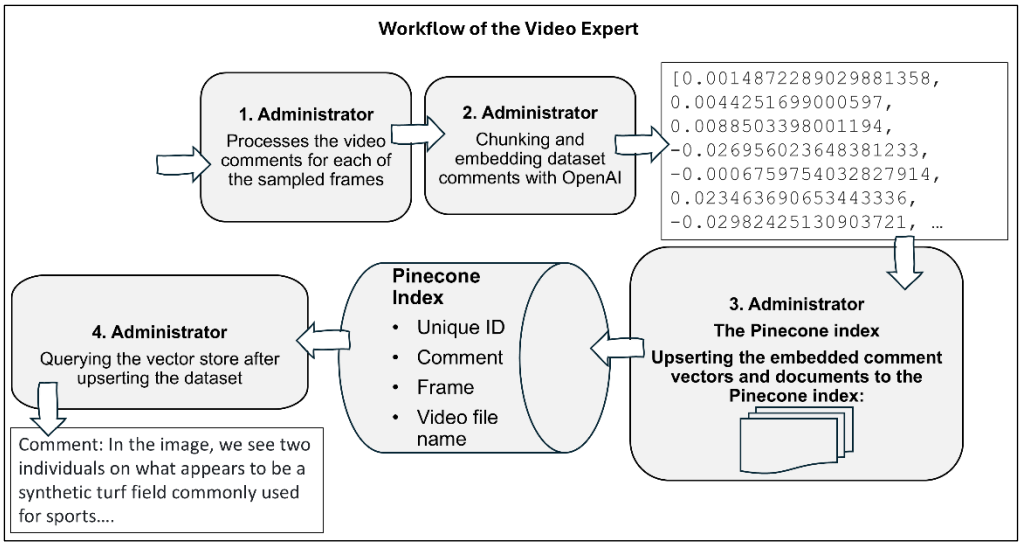


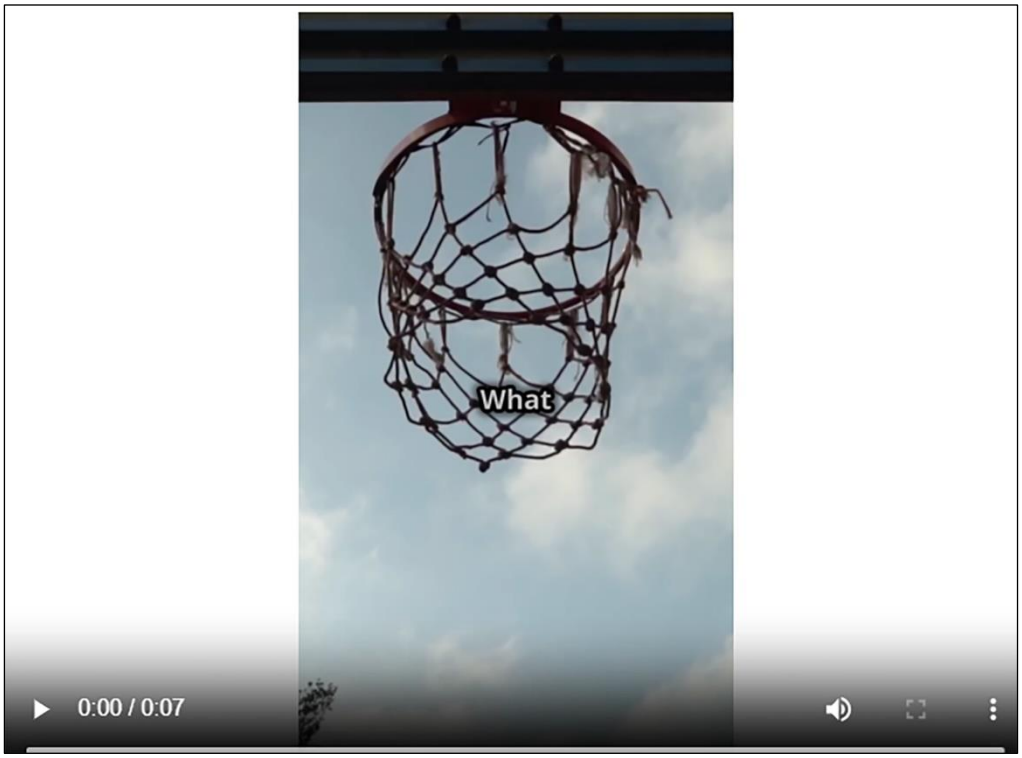

0:00 / 0:11


person

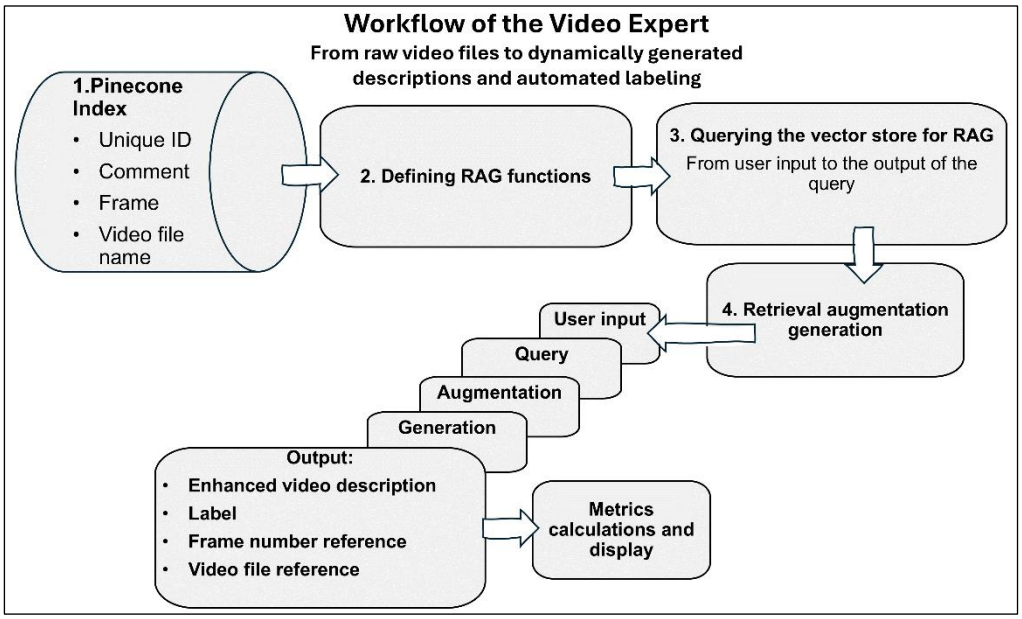

An


Workflow of the Generator and the Commentator
1. Generator
Generates text-to-video inVideo video dataset based on the video production team's text input.
In this chapter, it is a dataset of sports videos.
2. Generator
Runs a scaled process by selecting one video at a time.
3. Generator
Splits the video into frames (images).
4. Commentator
Samples frames (images) and comments on them with an OpenAI LLM model.
In the image, we see two individuals on what appears to be a synthetic turf field commonly used for sports....
4. Commentator
Each commented frame is saved:
• Unique ID
• Comment
• Frame
• Video file name


Workflow of the Video Expert
1. Administrator
Processes the video comments for each of the sampled frames
2. Administrator
Chunking and embedding dataset comments with OpenAI
[0.0014872289029881358,
0.0044251699000597,
0.0088503398001194,
-0.026956023648381233,
-0.0006759754032827914,
0.023463690653443336,
-0.02982425130903721, …
3. Administrator
The Pinecone index
Upserting the embedded comment vectors and documents to the Pinecone index:
Pinecone Index
Unique ID
Comment
Frame
Video file name
4. Administrator
Querying the vector store after upserting the dataset
Comment: In the image, we see two individuals on what appears to be a synthetic turf field commonly used for sports….


What
0:00 / 0:07




Workflow of the Video Expert
From raw video files to dynamically generated descriptions and automated labeling
1.Pinecone Index
Unique ID
Comment
Frame
Video file name
2. Defining RAG functions
3. Querying the vector store for RAG
From user input to the output of the query
4. Retrieval augmentation generation
User input
Query
Augmentation
Generation
Output:
Enhanced video description
Label
Frame number reference
Video file reference
Metrics calculations and display