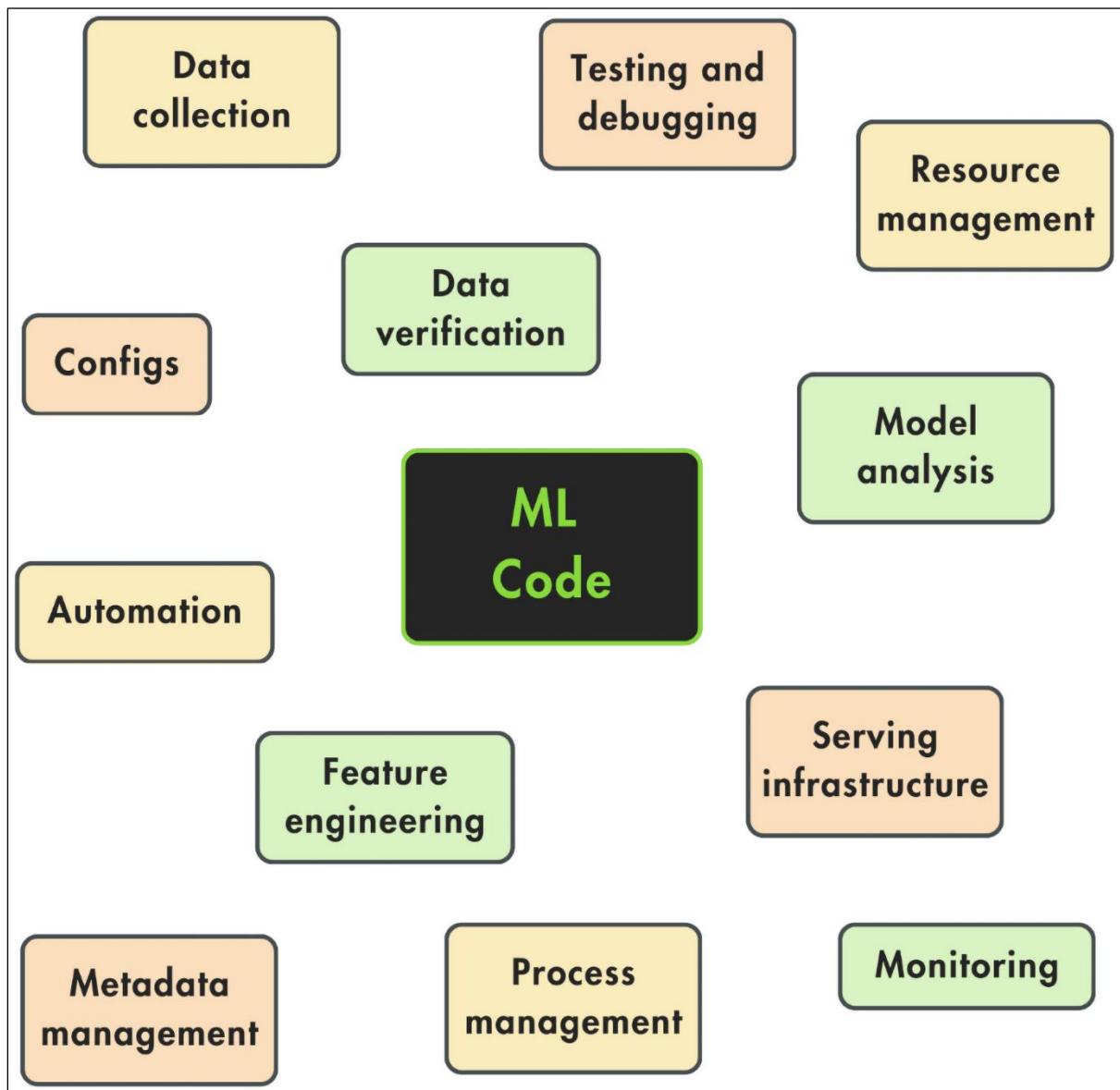
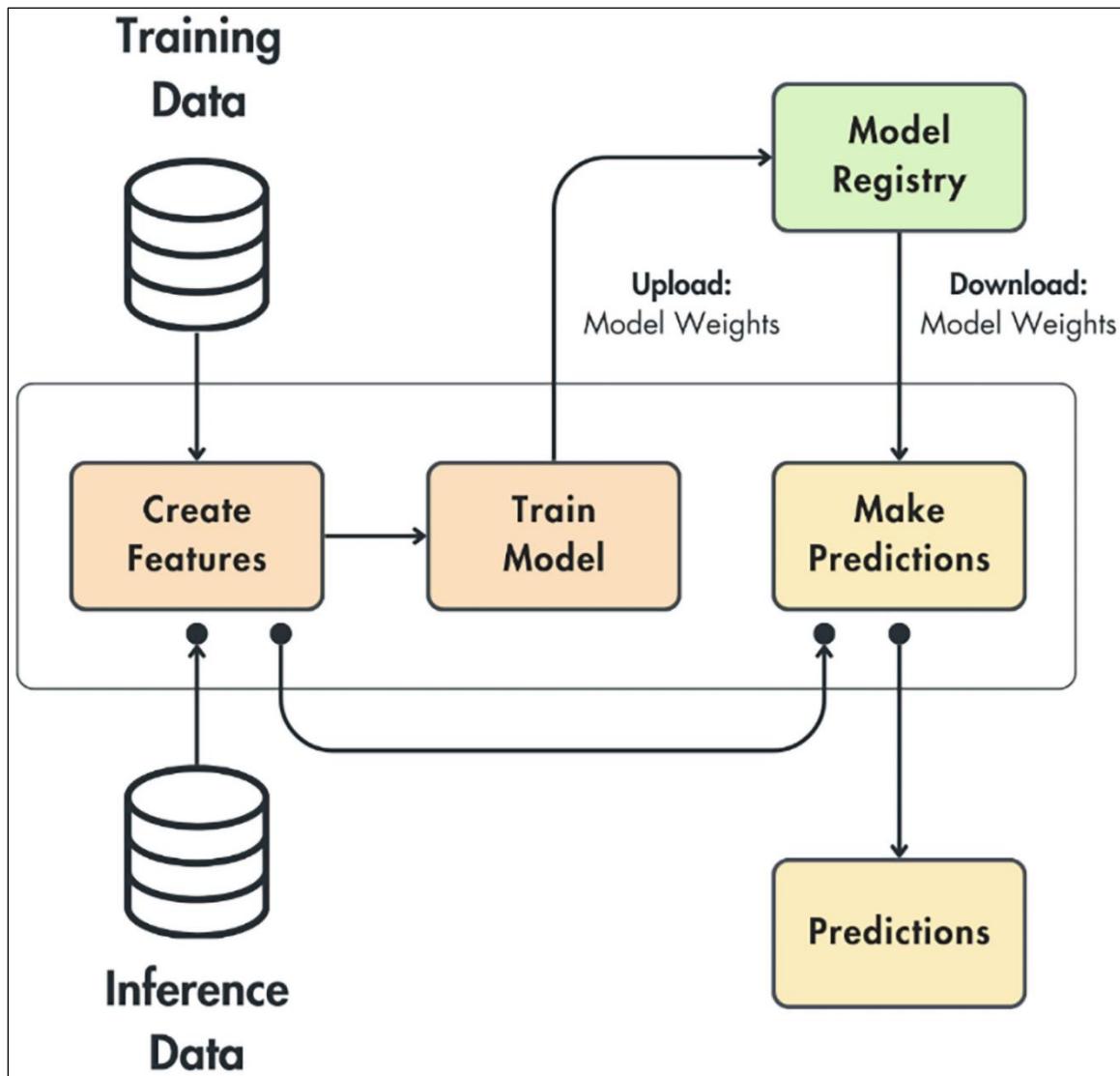
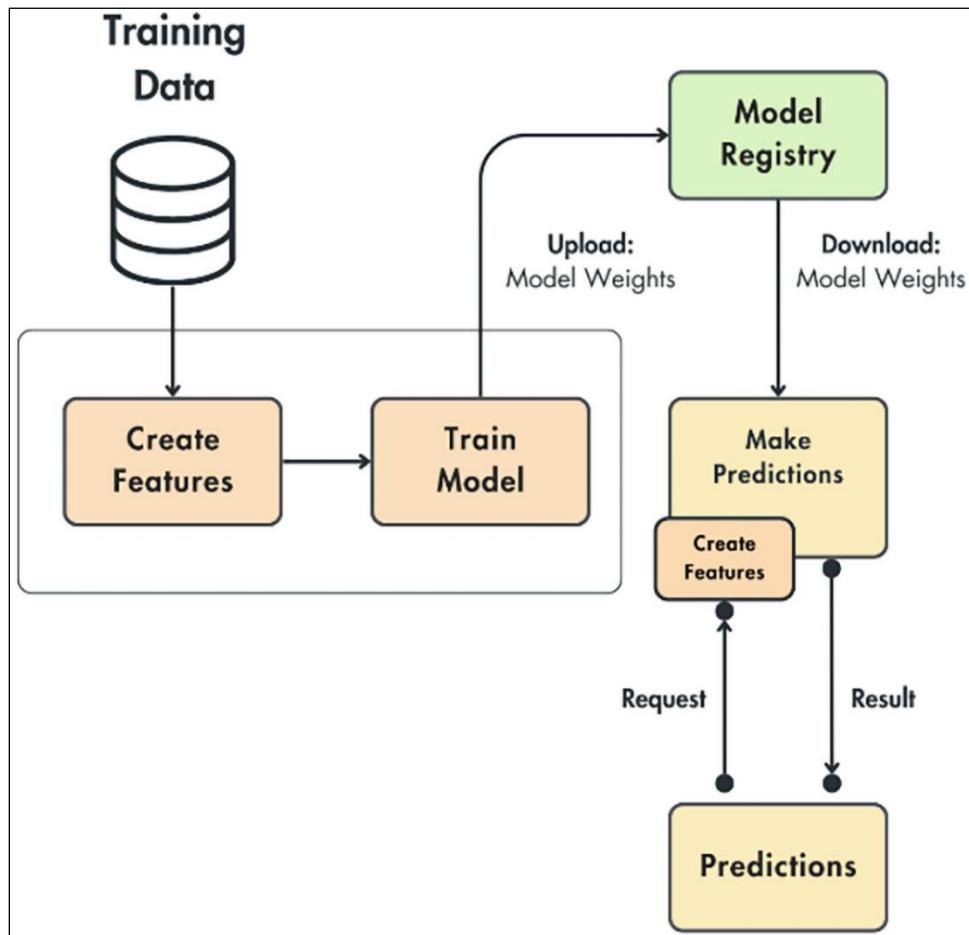
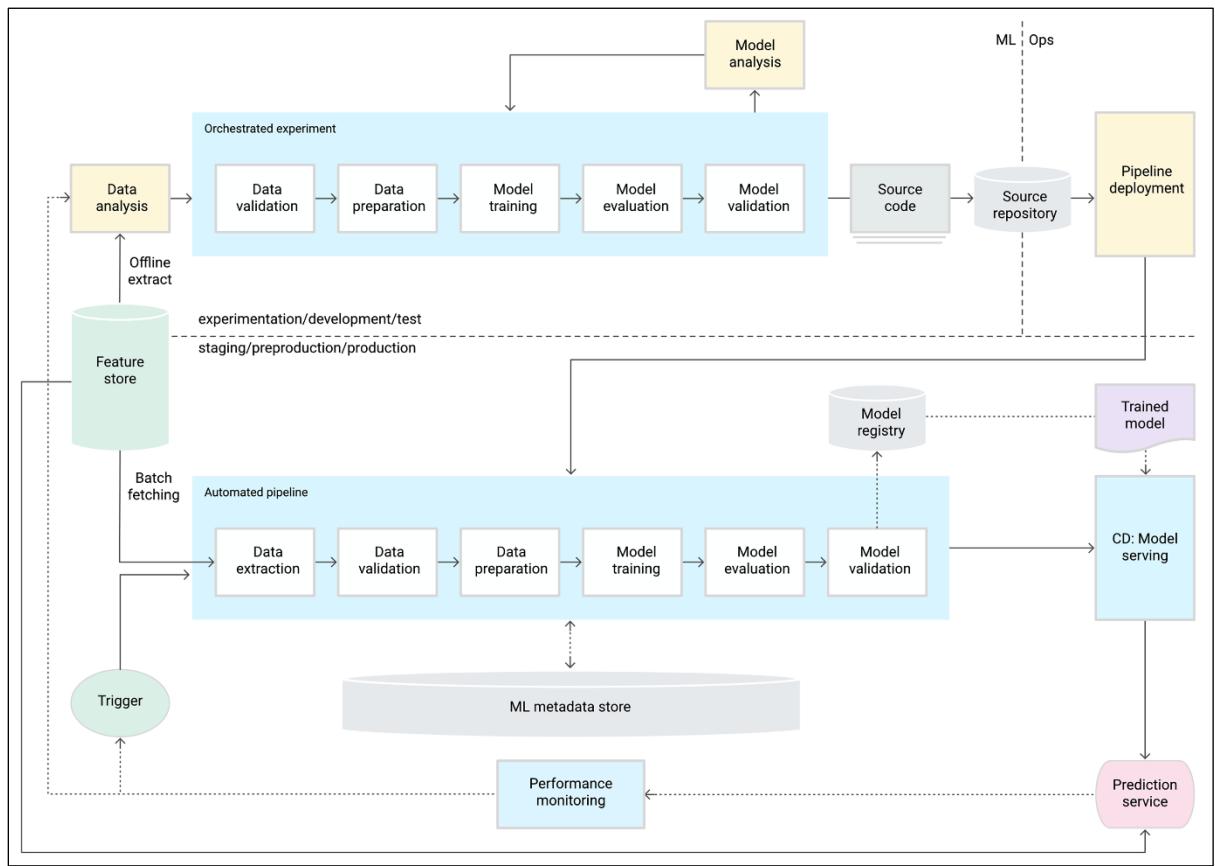


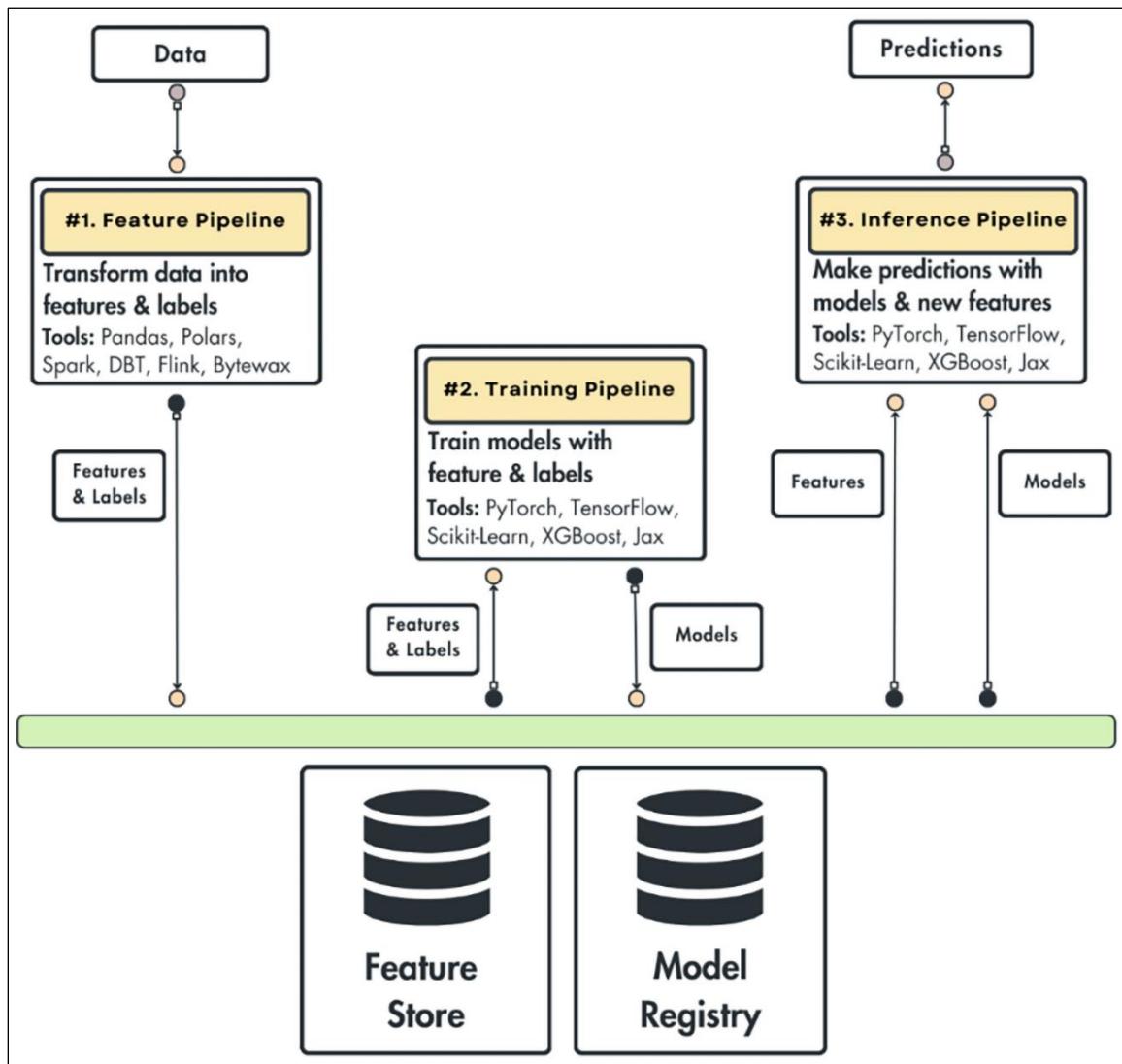
Chapter 1: Understanding the LLM Twin Concept and Architecture

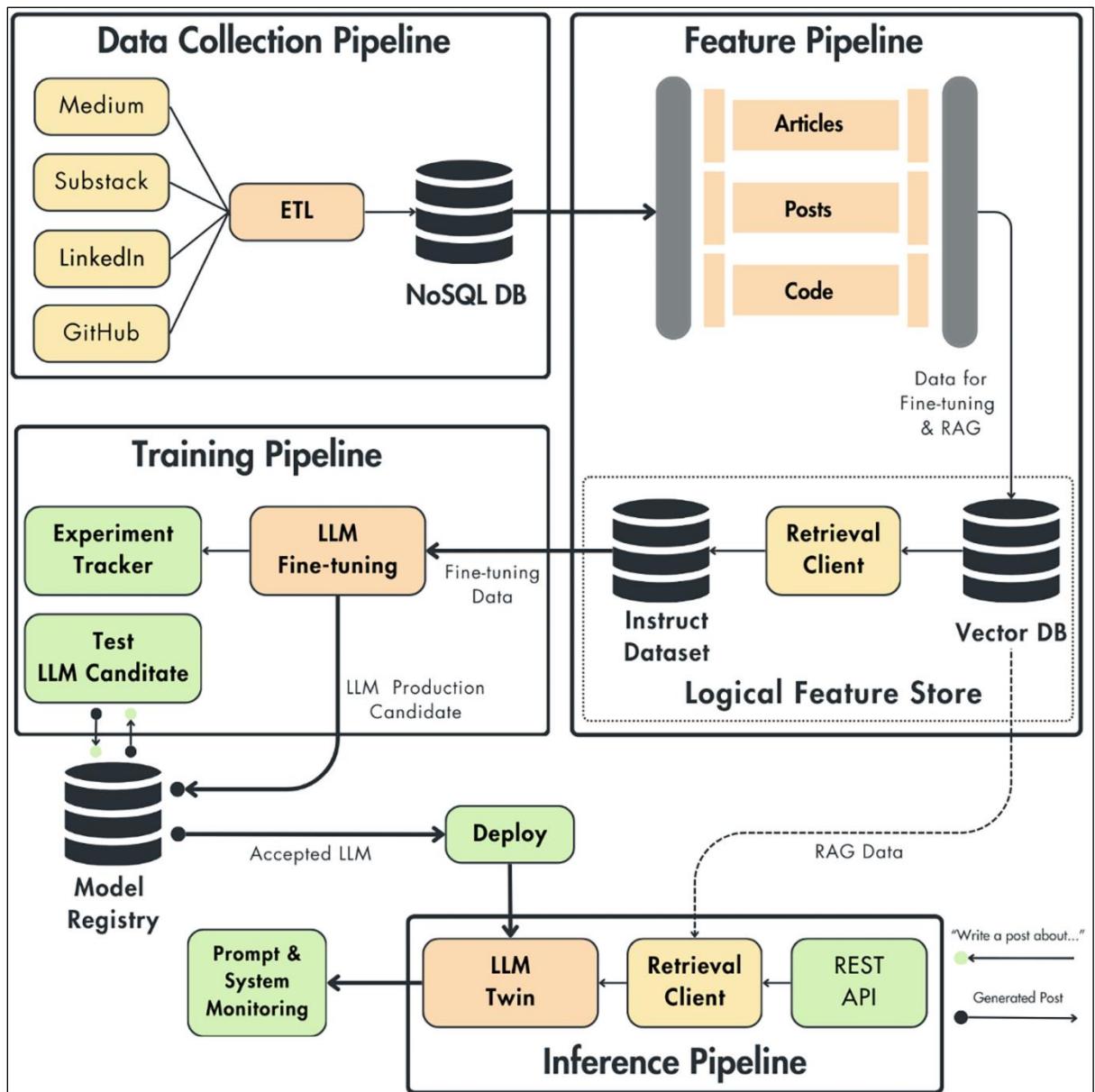












Chapter 2: Tooling and Installation

The screenshot shows the GitHub repository page for `mlbonne/TwinLlama-3.1-8B`. The page includes the repository name, a 'like' button (3 likes), and tags for `Text Generation`, `Transformers`, `Safetensors`, and `mlabon`. It also displays the license as `apache-2.0`. Below the repository information, there are links for `Model card`, `Files and versions` (selected), and `Community` (3 members).

The screenshot shows the Pipelines section of the repository. On the left, a sidebar lists `default` (Production Setup 0/2), `Overview`, `Pipelines` (selected and highlighted with a red border), `Models`, `Artifacts`, and `Stacks`. The main area displays a list of pipelines: `digital_data_etl`, `export_artifact_to_json`, `generate_datasets`, `generate_instruct_datasets`, and `feature_engineering`. Each pipeline entry includes a green circular icon with a checkmark.

Run	Stack	Repository	Created at	Author
digital_data_etl_run_2024_09_26_15_38_51 1187442f	default		26/09/2024, 15:38:51	D default
digital_data_etl_run_2024_09_24_12_29_57 d1632846	default		24/09/2024, 12:29:58	D default
digital_data_etl_run_2024_09_24_12_29_31 a94b23d4	default		24/09/2024, 12:29:31	D default
digital_data_etl_run_2024_09_24_12_28_40 40530bbb	default		24/09/2024, 12:28:41	D default
digital_data_etl_run_2024_08_26_09_14_06 906dff30	default		26/08/2024, 11:14:06	D default

[digital_data_etl_run_2024_09_26_15_38_51](#)

```

graph TD
    A[get_or_create_user] --> B[user]
    B --> C[crawl_links]
    C --> D[crawled_links]
  
```

Run Insights

Overview Configuration

Details

- ID:** 1187442f-6a11-43c2-b455-ffa6254477cc
- Status:** completed
- Pipeline:** digital_data_etl
- Author:** D default
- Start Time:** 26/09/2024, 15:38:51
- End Time:** 26/09/2024, 15:38:51

Stack

Artifact Store default

get_or_create_user completed

Overview Code Logs Configuration

Logs

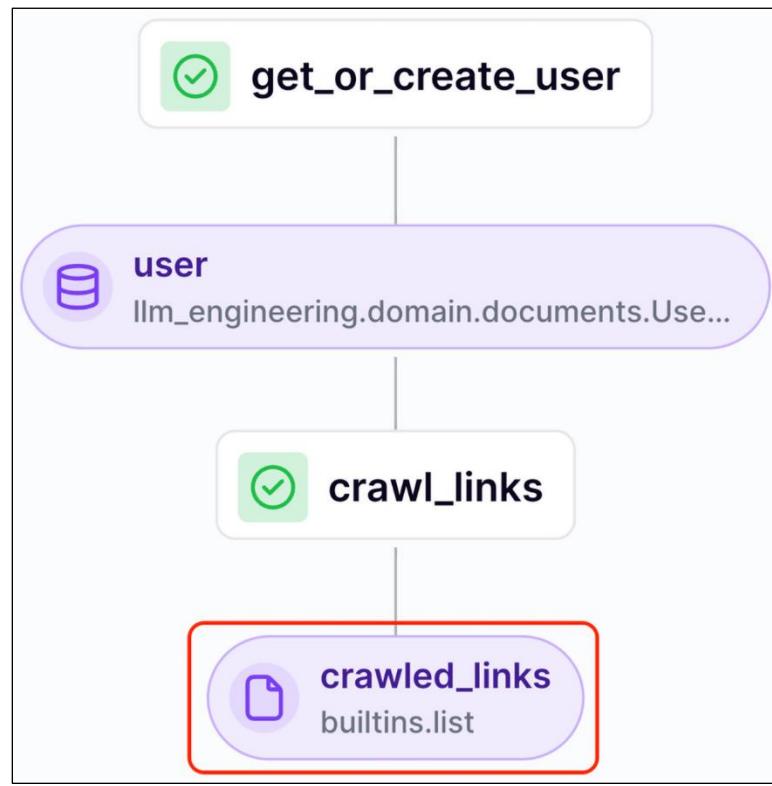
```
Step get_or_create_user has started.  
2024-09-26 15:38:51.597 | INFO | steps.etl.get_or_create_user:get_or_create_user:11  
- Getting or creating user: Maxime Labonne  
Step get_or_create_user has finished in 0.073s.  
Step get_or_create_user completed successfully.
```

LLM-Engineers-Handbook Public

Edit Pins Watch 2 ▾

main 7 Branches 0 Tags Go to file Add file Code ▾

Author	Commit Message	Date
iusztinpaul	docs: Improve README	d3f23f4 · 17 hours ago 116 Commits
	.github/workflows fix: Loading Settings from ZenML secrets	2 months ago
	.vscode feat: Add DE pipeline logic	4 months ago
	code_snippets feat: Add custom ODM example	last week
	configs docs: Extend README	yesterday
	dummy_dataset added finetuning script v1	2 months ago
	images docs: Update README with .env details	3 days ago
	llm_engineering docs: Extend README	yesterday
	pipelines feat: Add dataset generation logic with preference support	2 weeks ago
	steps feat: Add dataset generation logic with preference support	2 weeks ago
	tools feat: Add dataset generation logic with preference support	2 weeks ago



cb8f8ac7-30bd-48fa-b5a2-e2958096c15a

 **crawled_links** 8

 Overview  Metadata  Visualization

> Uncategorized

▼ mlabonne.github.io

successful	2
total	2

▼ maximelabonne.substack.com

successful	24
total	24

8bba35c4-8ff9-4d8f-a039-08046efc9fdc

instruct_datasets 10

 Overview

 Metadata

 Visualization

▼ Uncategorized

data_categories articles

storage_size 493.23 KB

test_split_size 0.1

> schema

▼ train_num_samples_per_category

articles 738

▼ test_num_samples_per_category

articles 82

LLM-Engineering / configs /



iusztinpaul feat: Add dataset generation

Name



..

 digital_data_etl_alex_vesa.yaml

 digital_data_etl_maxime_labonne.yaml

 digital_data_etl_paul_iusztin.yaml

 end_to_end_data.yaml

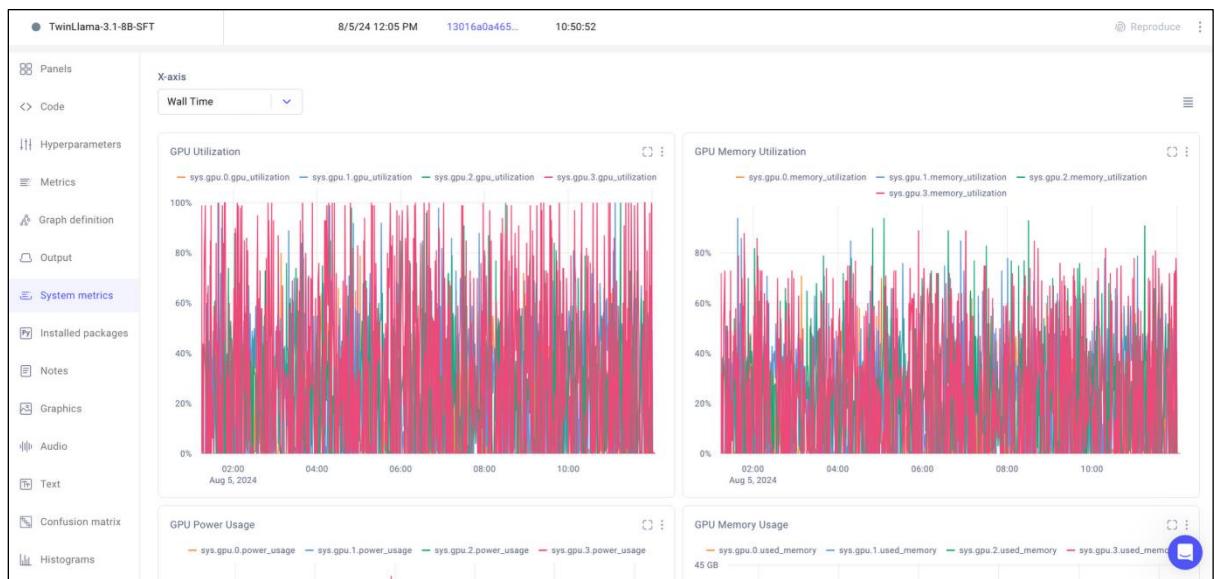
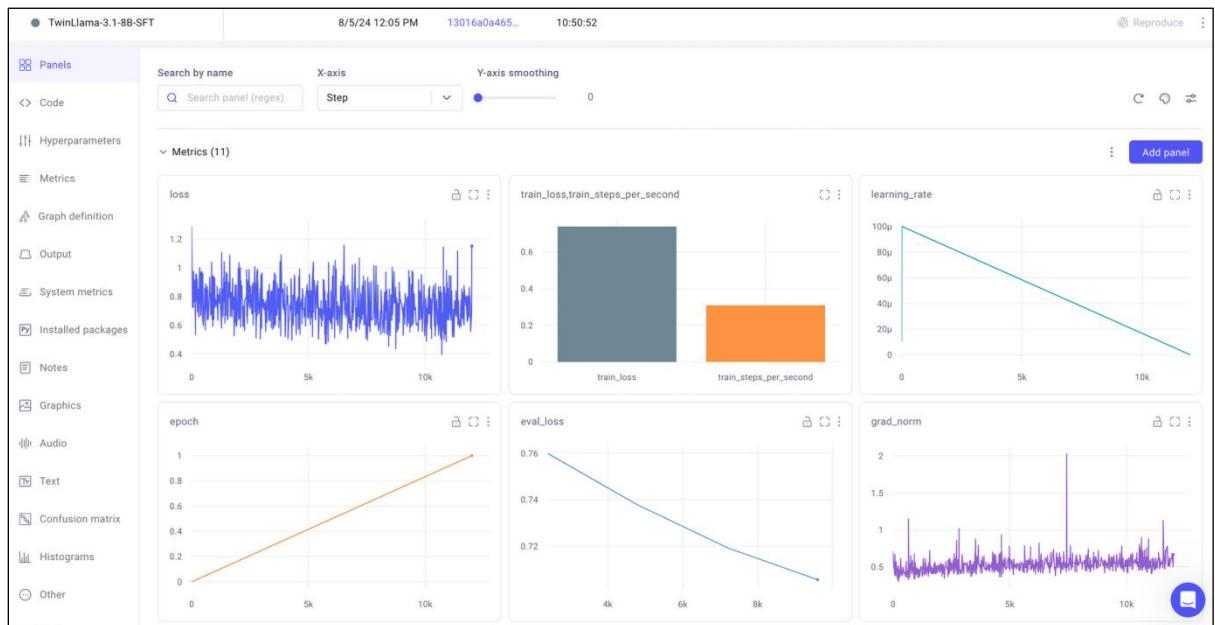
 export_artifact_to_json.yaml

 feature_engineering.yaml

 generate_instruct_datasets.yaml

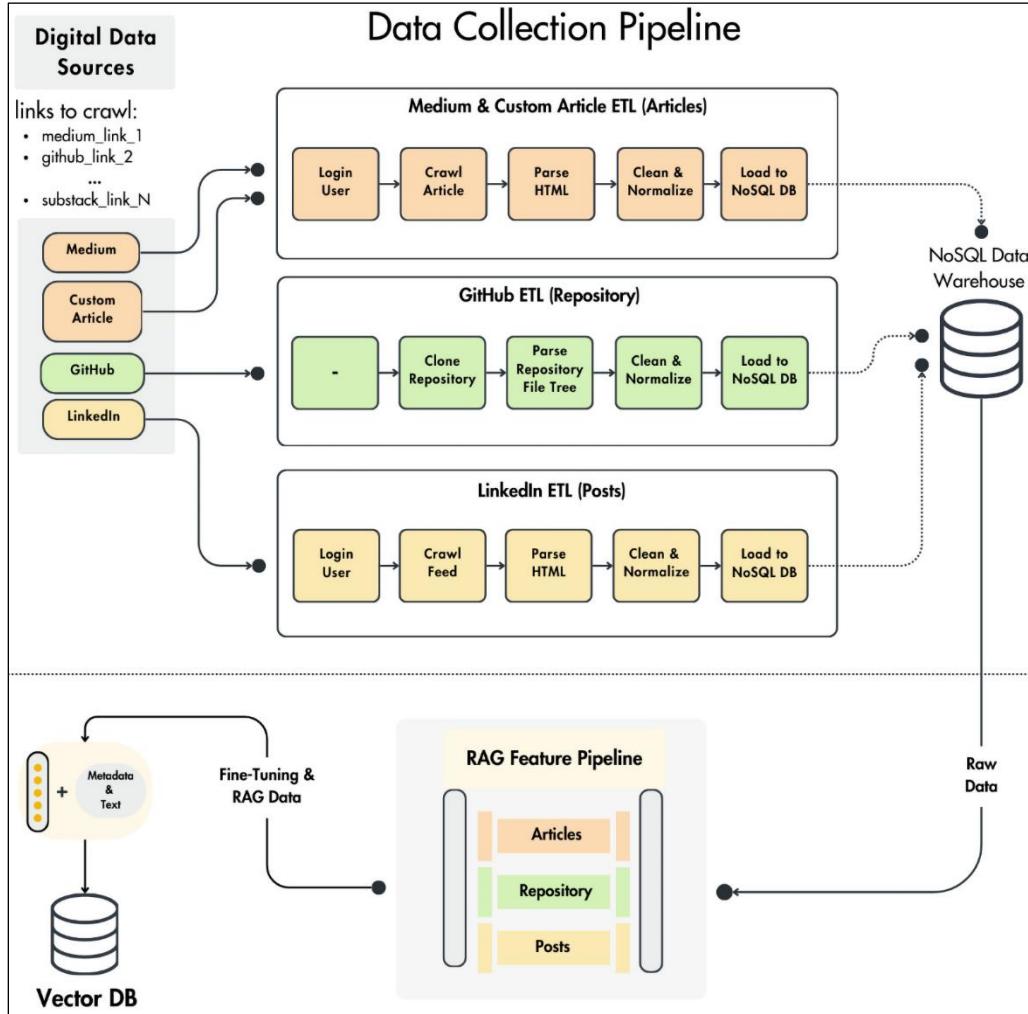
 generate_preference_datasets.yaml

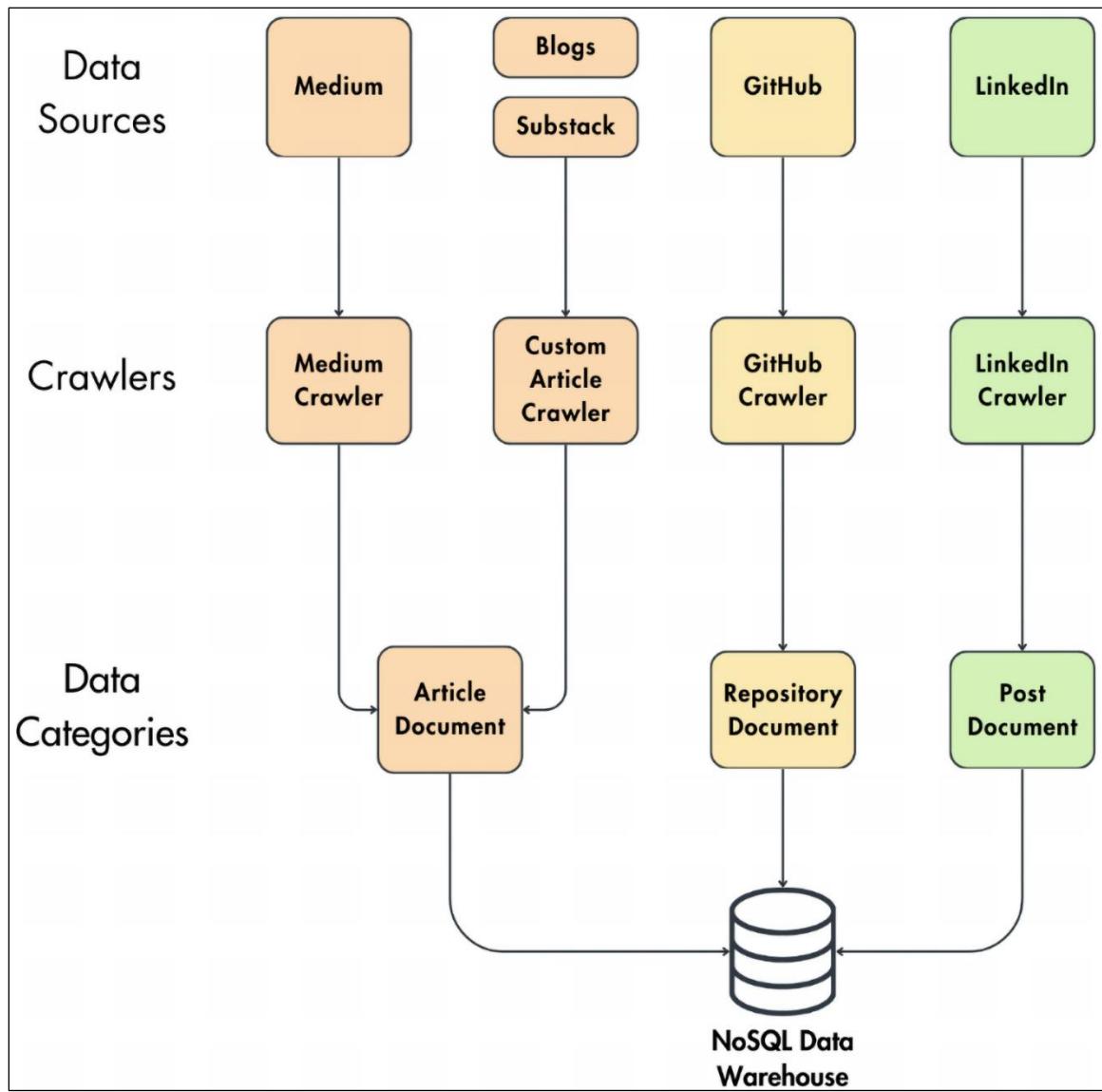
 training.yaml

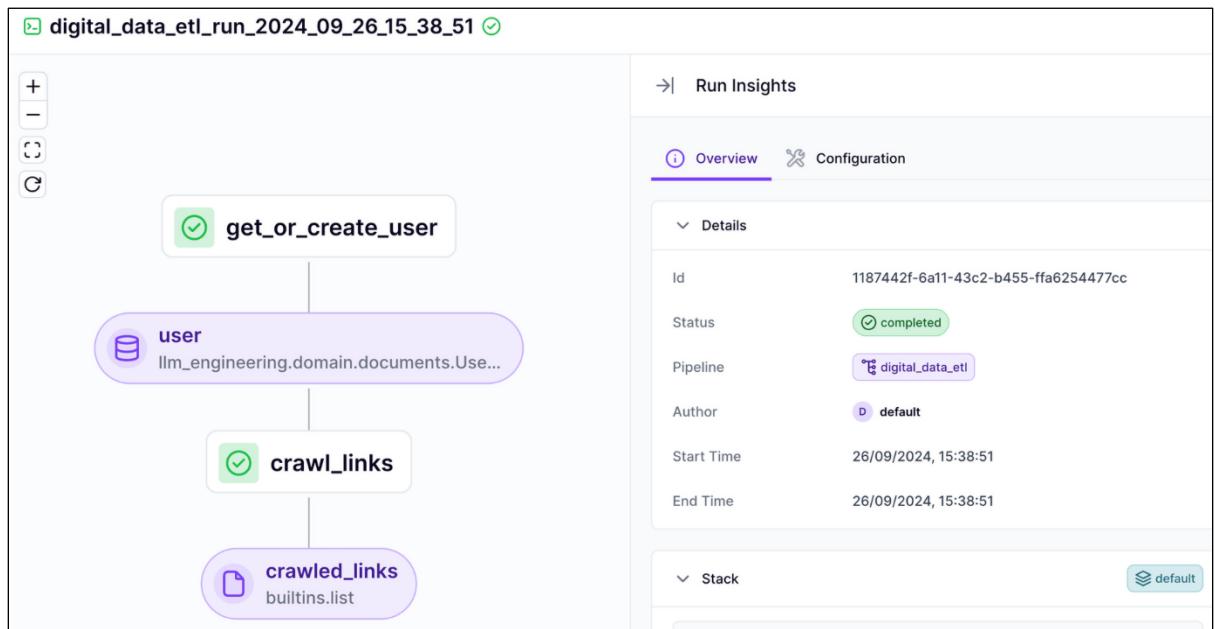


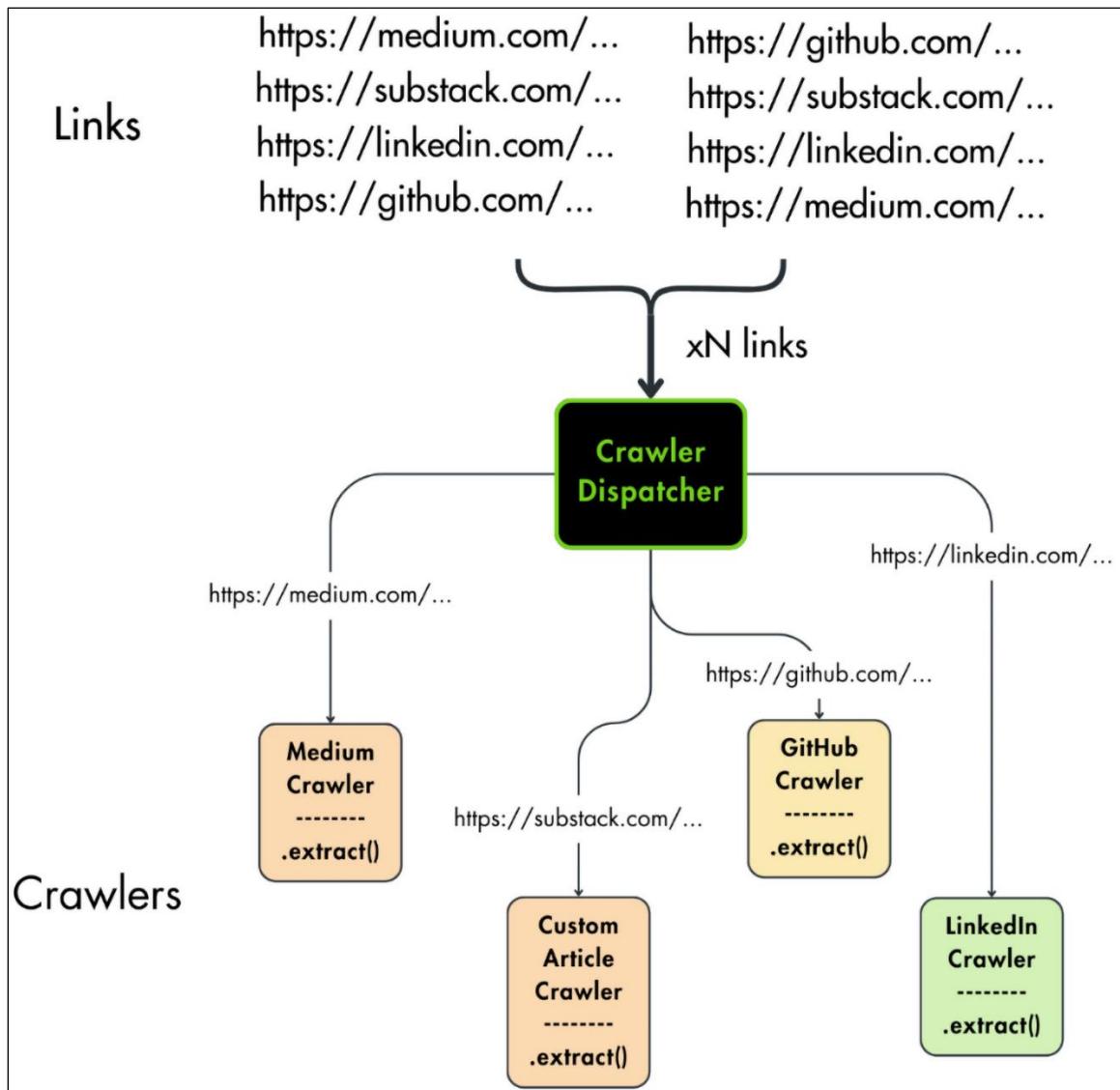
Permissions	Groups	Tags (2)	Security credentials	Last Accessed
Permissions policies (2)				
Permissions are defined by policies attached to the user directly or through groups.				
<div style="display: flex; justify-content: space-between;">Filter by Type<input type="text" value="Search"/> All types< 1 > @</div>				
<input type="checkbox"/> Policy name	Type	Attached via		
<input type="checkbox"/> AdministratorAccess	AWS managed - job function	Directly		
<input type="checkbox"/> IAMUserChangePassword	AWS managed	Directly		

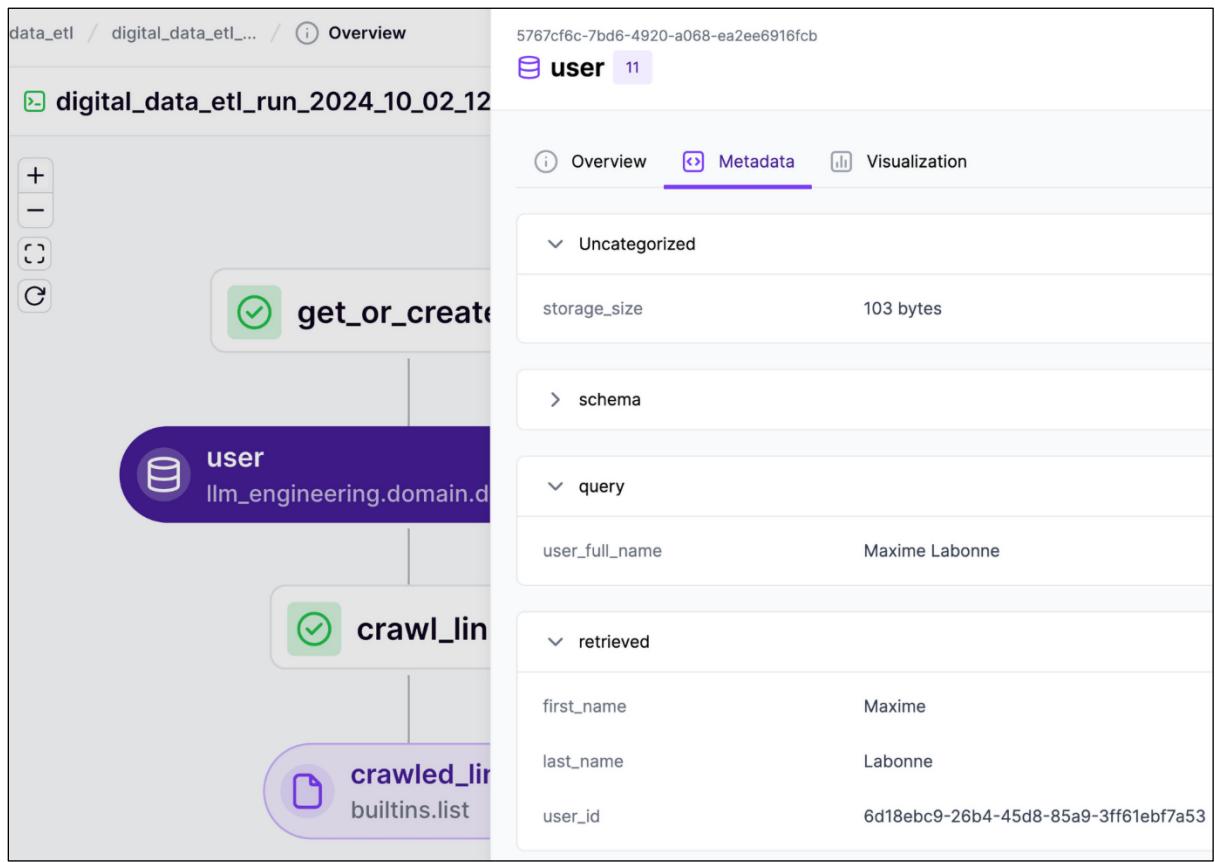
Chapter 3: Data Engineering

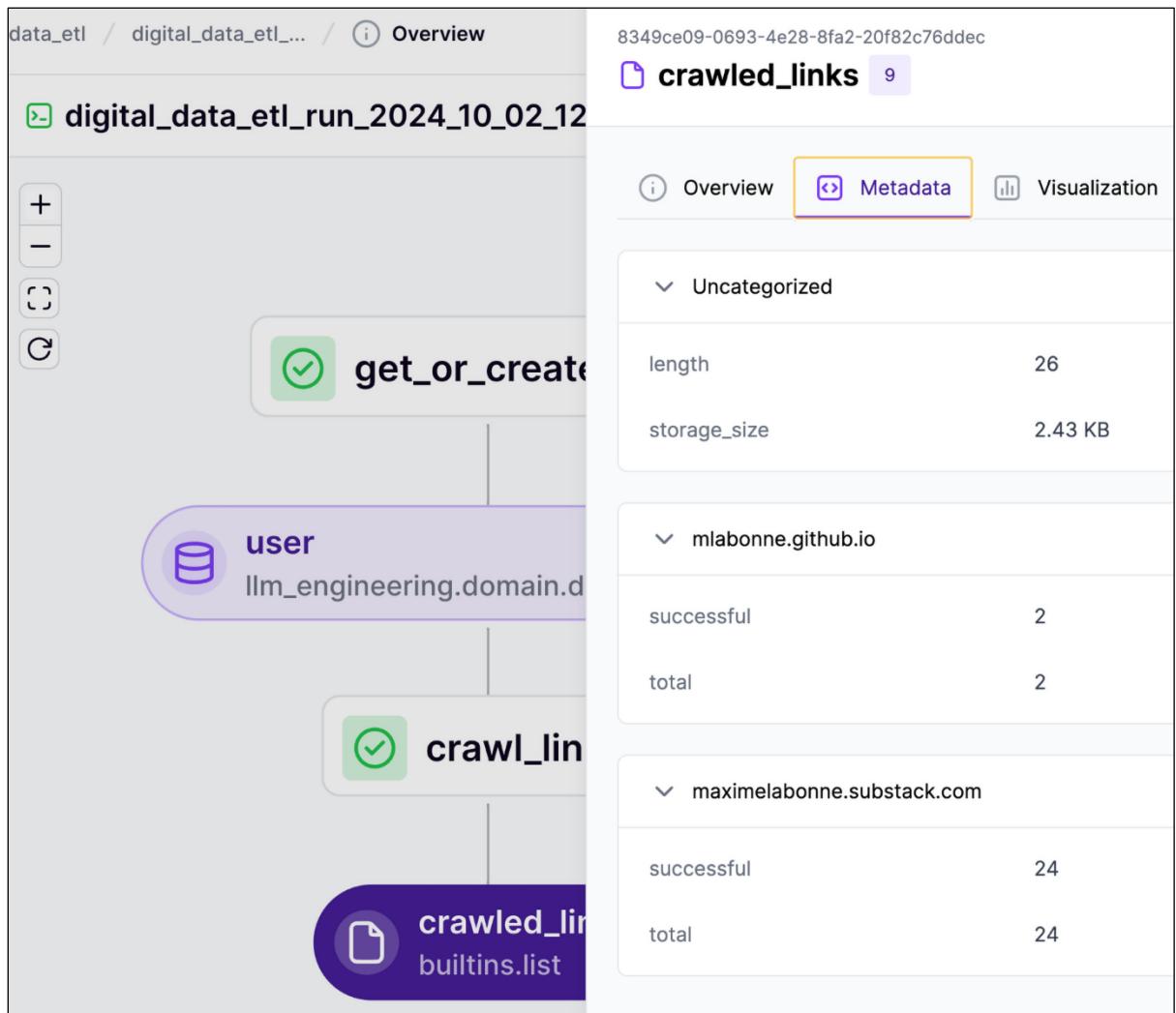












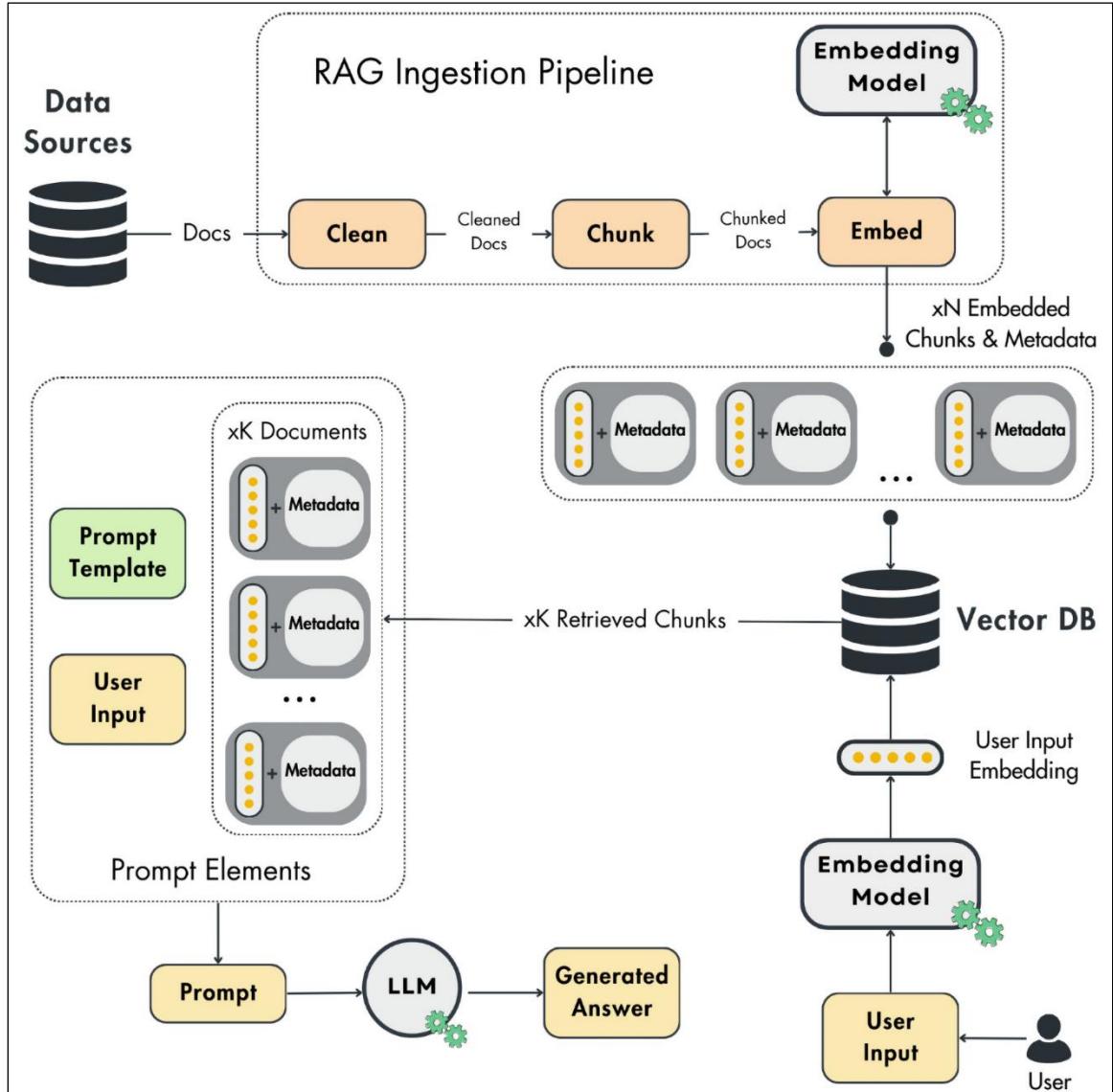
[LLM-Engineers-Handbook](#) / [configs](#) / [digital_data_etl_paul_iusztin.yaml](#) ↗

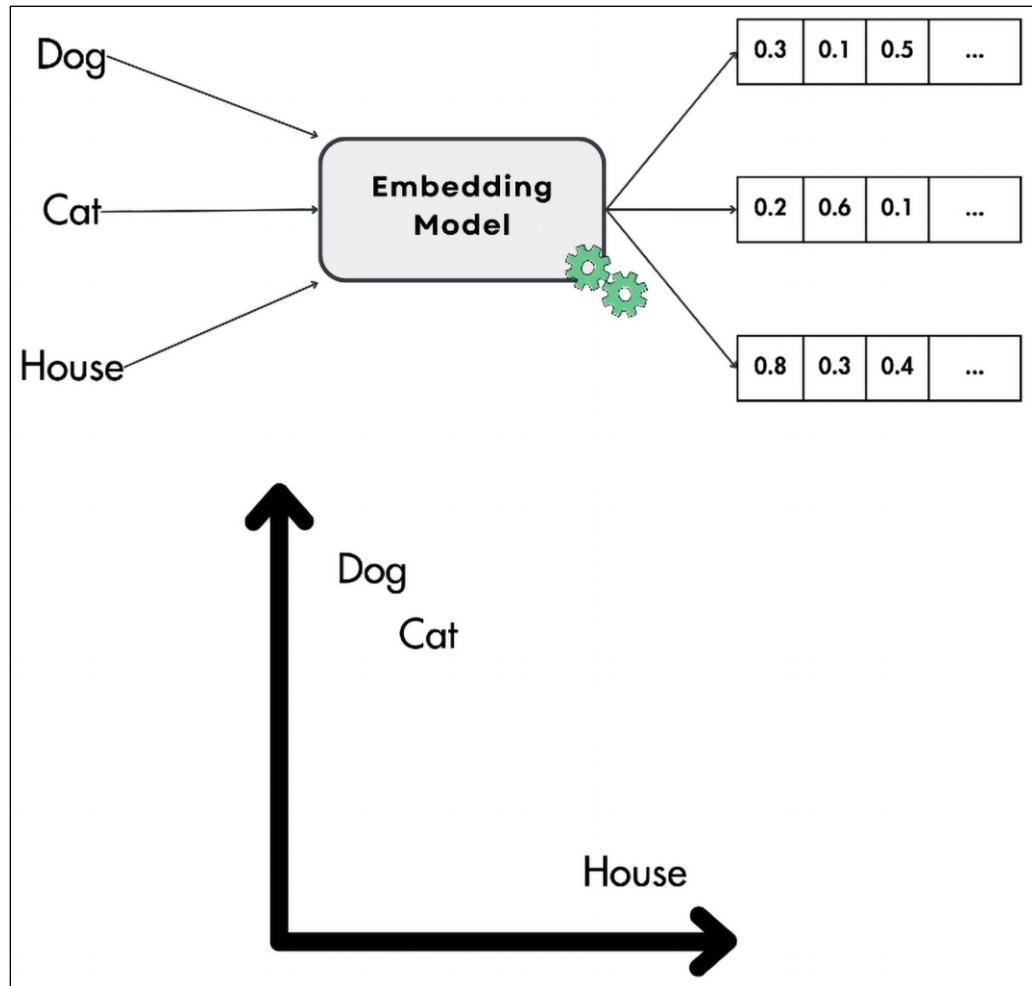
 **iusztinpaul** docs: Comment out Medium urls

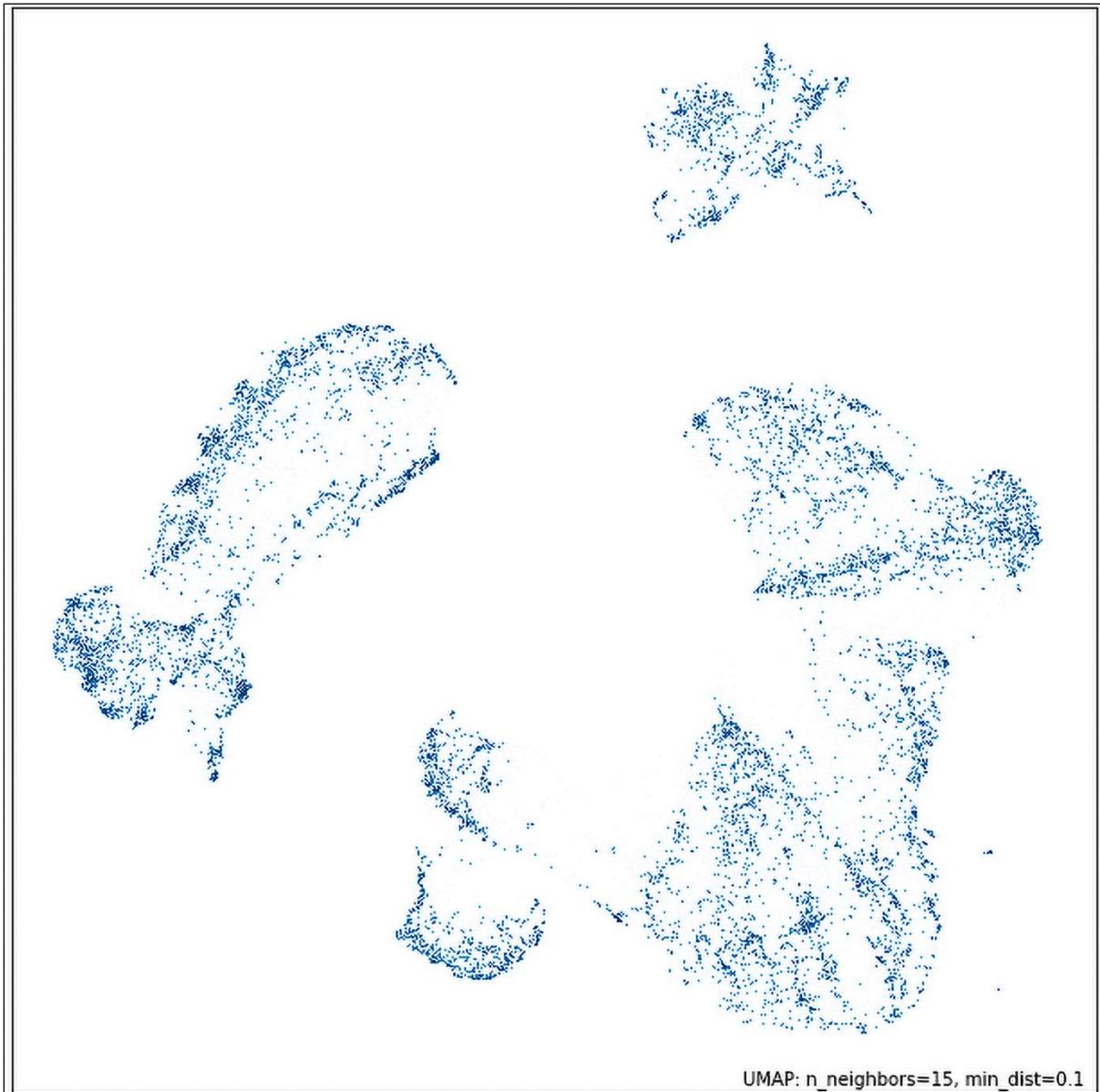
[Code](#) [Blame](#) 62 lines (61 loc) · 4.62 KB

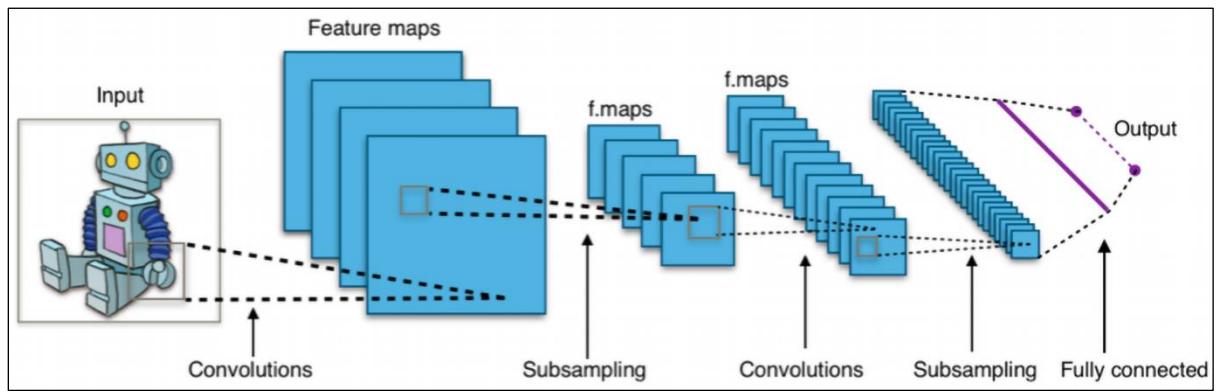
```
1   settings:
2     docker:
3       parent_image: 992382797823.dkr.ecr.eu-central-1.amazonaws.com/zenml-rlwlcs:latest
4       skip_build: True
5     orchestrator.sagemaker:
6       synchronous: false
7
8   parameters:
9     user_full_name: Paul Iusztin # [First Name(s)] [Last Name]
10    links:
11      # Medium (only articles that are not under the paid wall work)
12      # - https://medium.com/decodingml/an-end-to-end-framework-for-production-ready-llm-systems-by-building-your-llm-twin-2cc6bb01141f
13      # - https://medium.com/decodingml/a-real-time-retrieval-system-for-rag-on-social-media-data-9cc01d50a2a0
14      # - https://medium.com/decodingml/sota-python-streaming-pipelines-for-fine-tuning-llms-and-rag-in-real-time-82eb07795b87
15      # - https://medium.com/decodingml/the-4-advanced-rag-algorithms-you-must-know-to-implement-5d0c7f1199d2
16      # - https://medium.com/decodingml/architect-scalable-and-cost-effective-llm-rag-inference-pipelines-73b94ef82a99
17      # Substack
18      - https://decodingml.substack.com/p/real-time-feature-pipelines-with?r=1ttoeh
19      - https://decodingml.substack.com/p/building-ml-systems-the-right-way?r=1ttoeh
20      - https://decodingml.substack.com/p/reduce-your-pytorchs-code-latency?r=1ttoeh
21      - https://decodingml.substack.com/p/llm-agents-demystified?r=1ttoeh
```

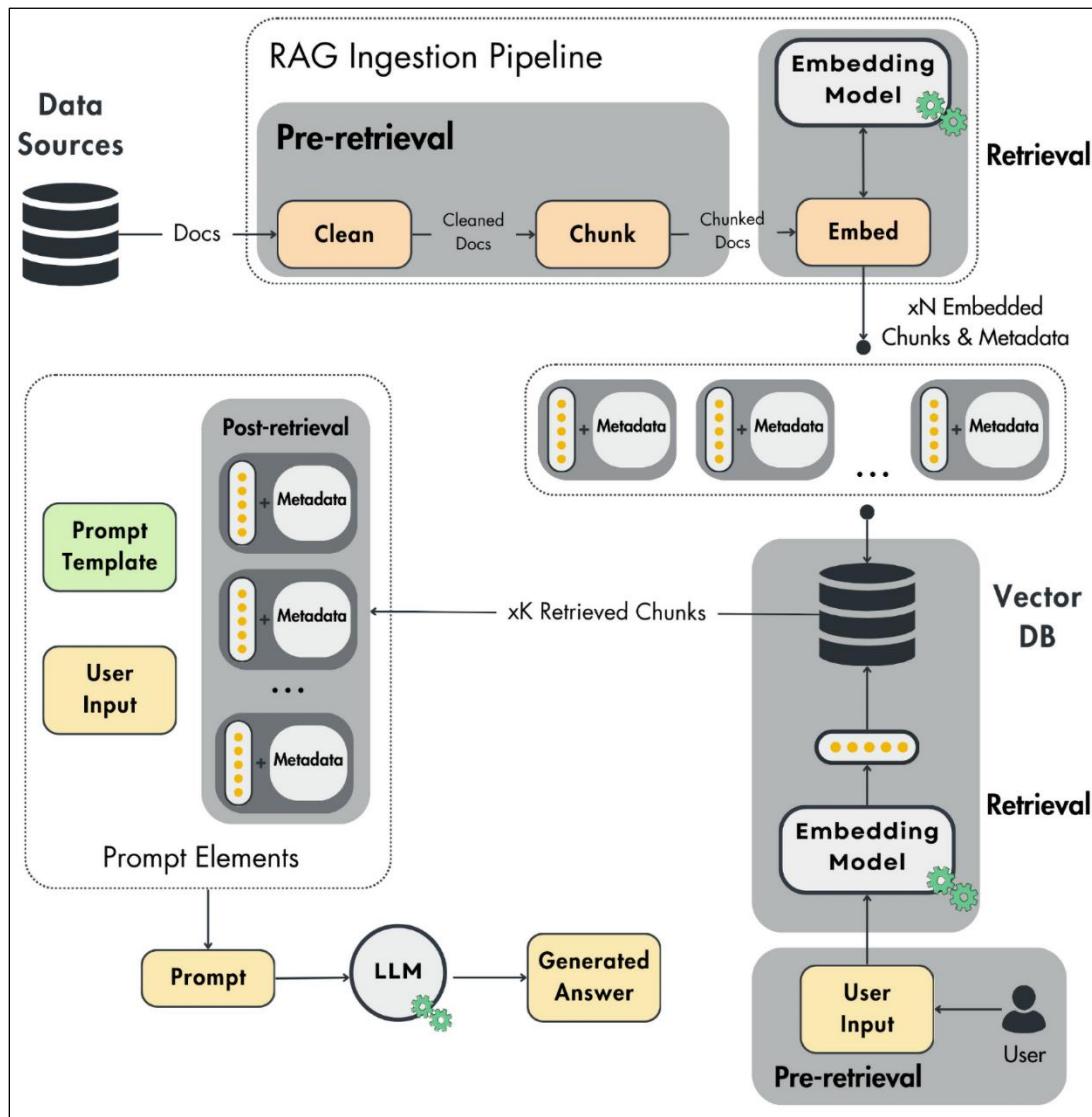
Chapter 4: RAG Feature Pipeline

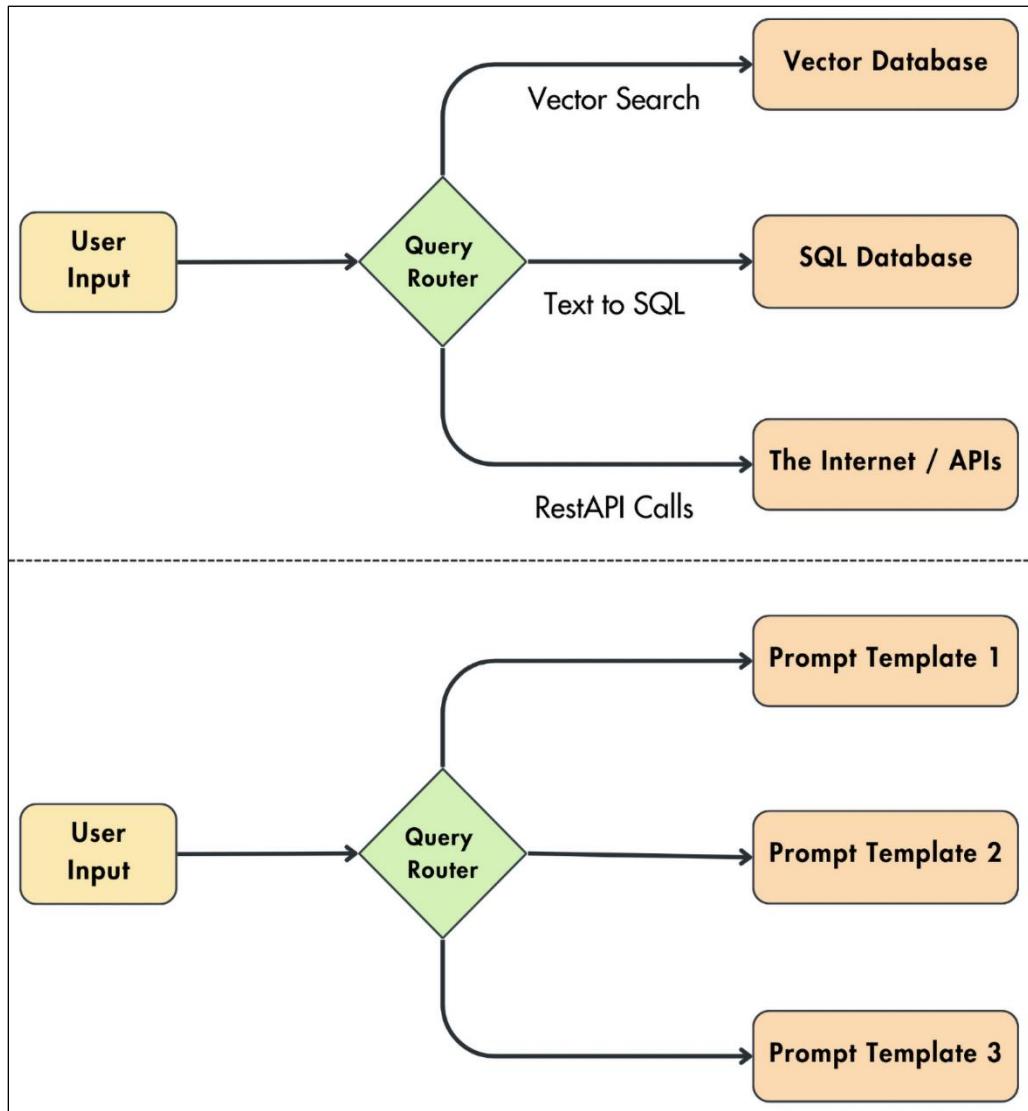


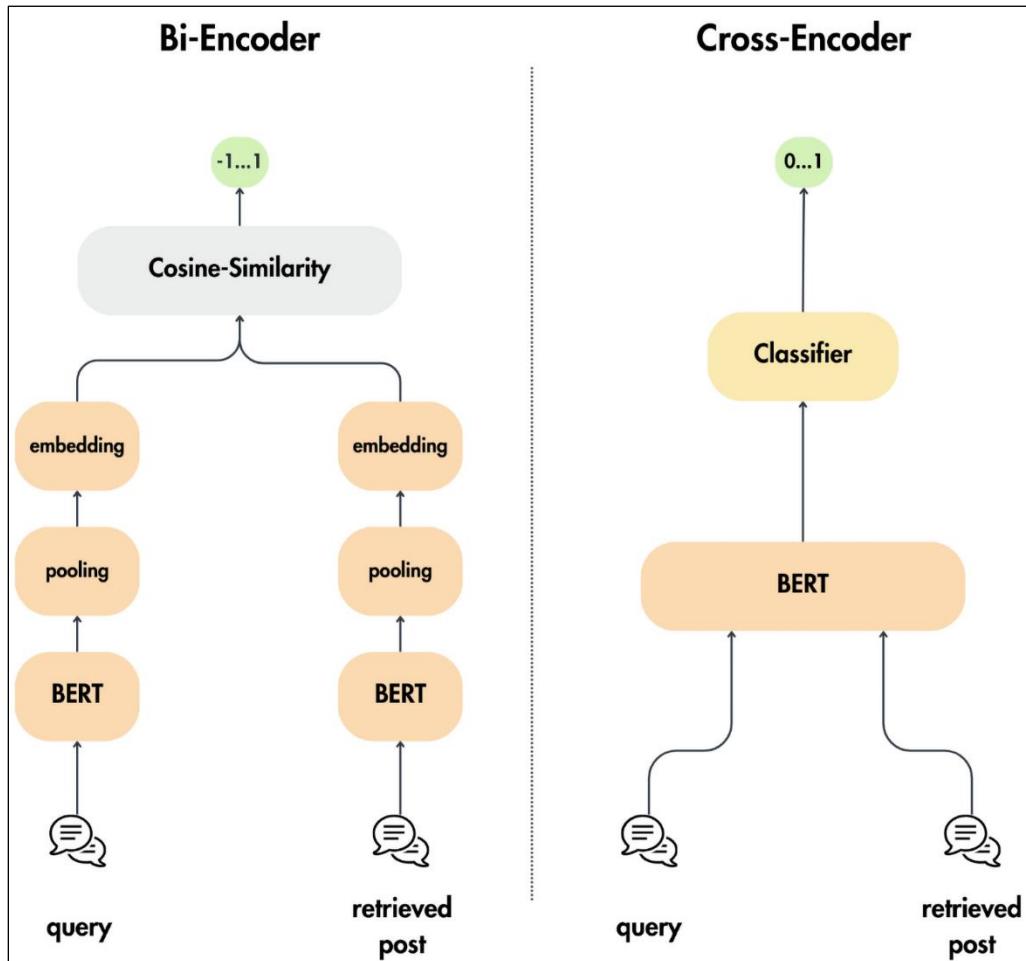


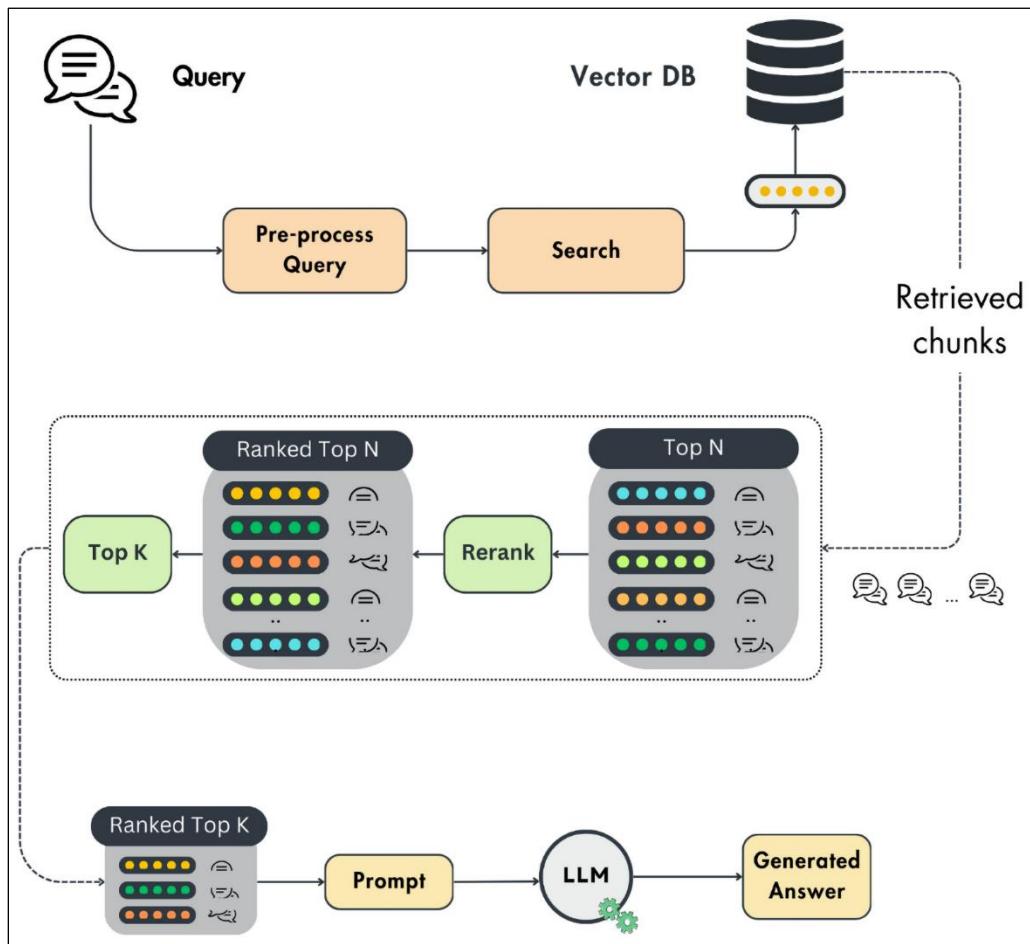


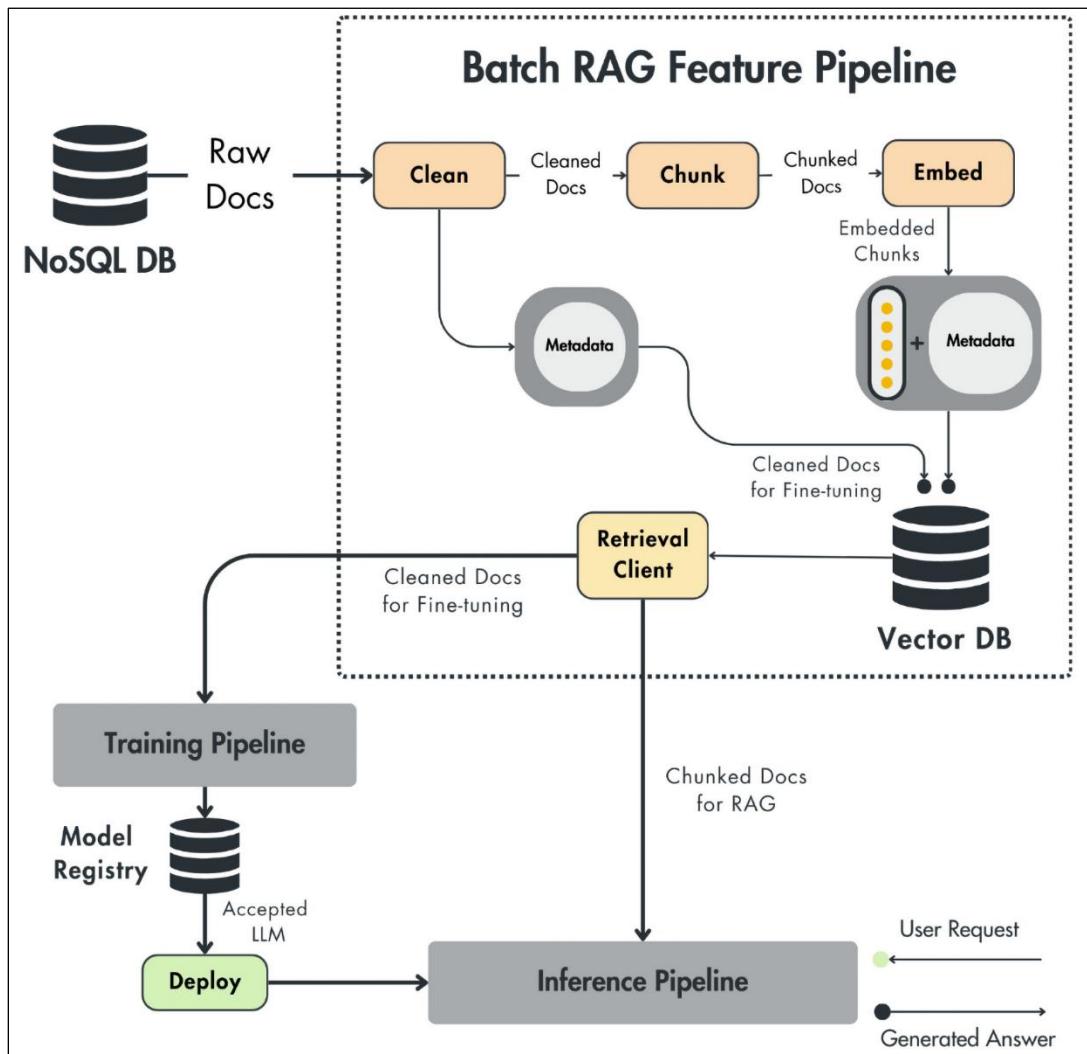


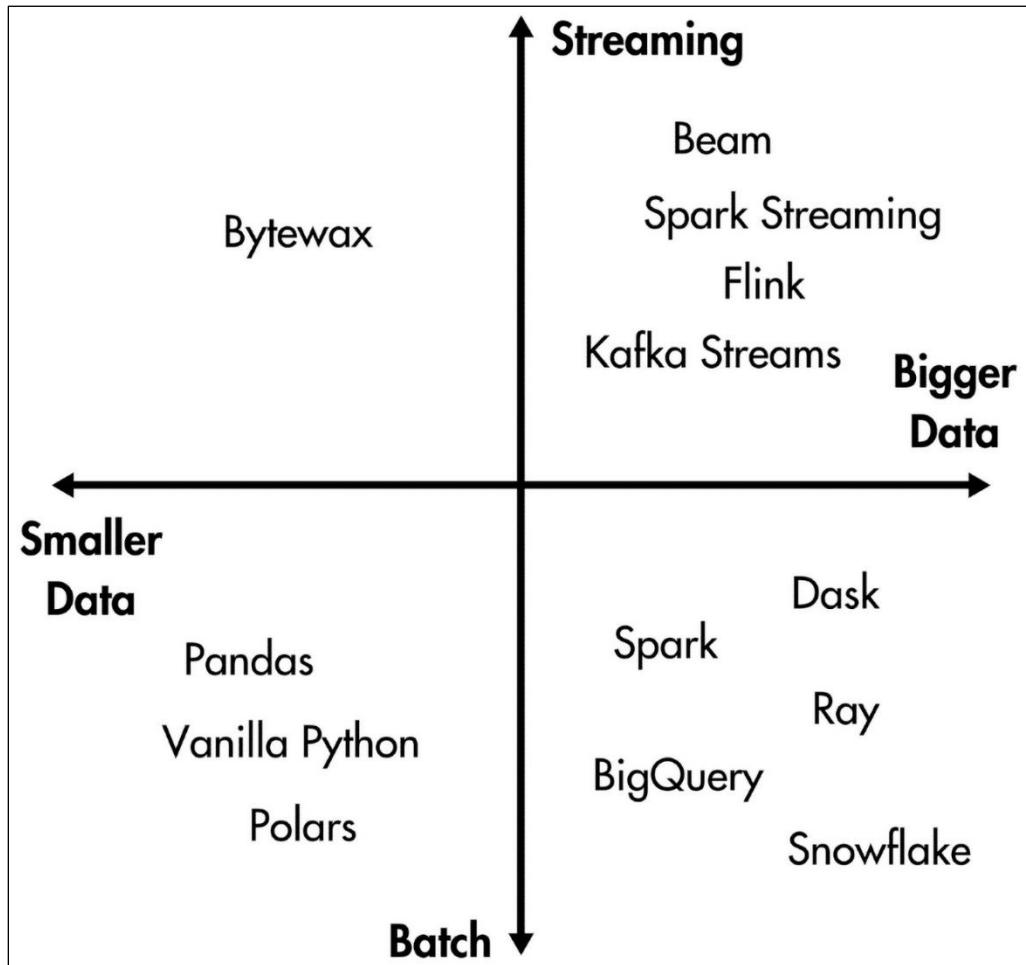




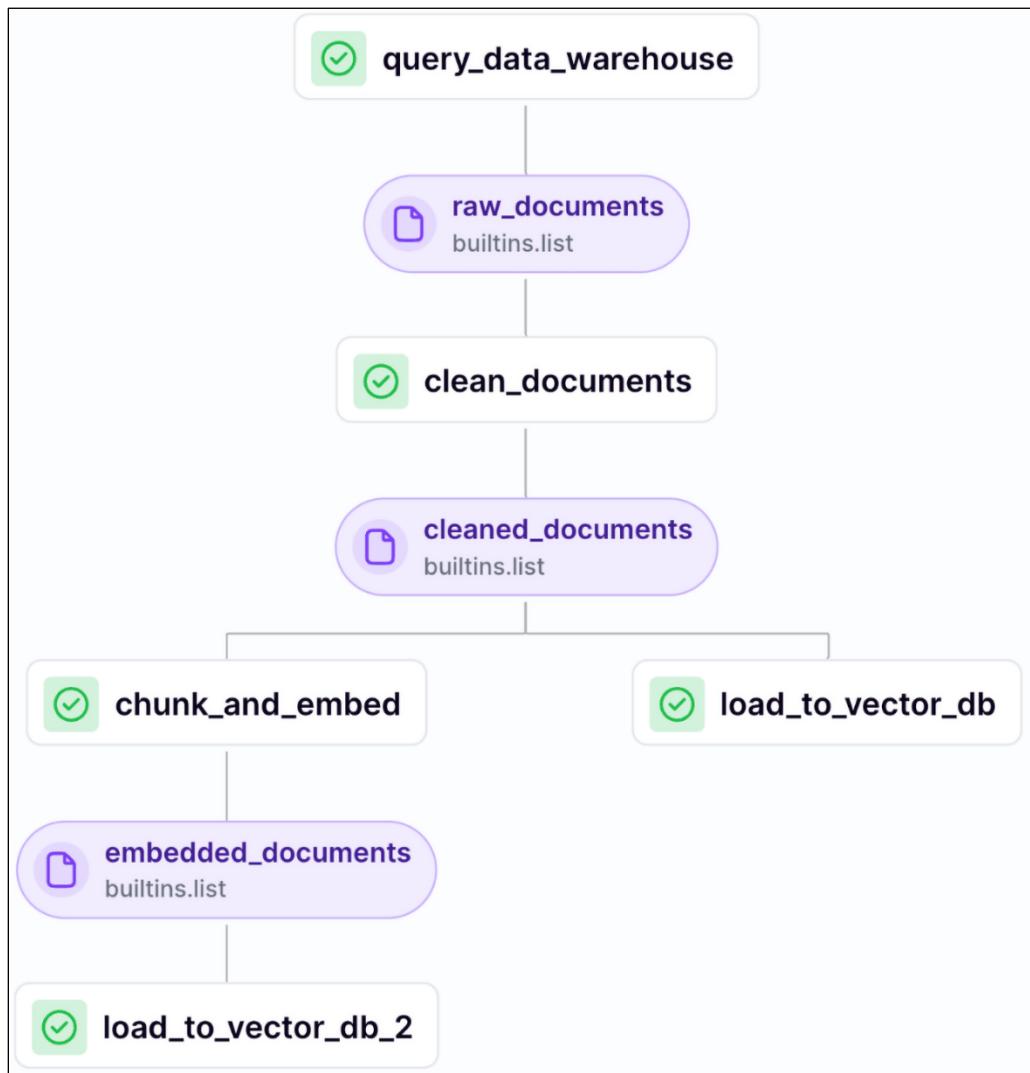








feature_engineering					
Run	Version	Stack	Repository	Created at	Author
feature_engineering_run_2024_06_29_09_58_41 3e31e84e	33	default		29/06/2024, 09:58:42	D default
feature_engineering_run_2024_06_29_09_50_29 ad7abecce	33	default		29/06/2024, 09:50:31	D default
feature_engineering_run_2024_06_29_09_41_26 caccfd39	33	default		29/06/2024, 09:41:26	D default



f80ace2c-af55-42d7-964b-012c81f17511

raw_documents 63

Overview Metadata Visualization

> Uncategorized

▼ articles

	num_documents	
0	76	Paul Iusztin
1		Maxime Labonne
2		Alex Vesa

56260613-5a4d-45bb-b17a-27479d5151e2

📄 **embedded_documents** 28

ⓘ Overview

🔗 Metadata

📊 Visualization

▼ Uncategorized

length	2373
num_chunks	2373
num_documents	76
num_embedded_chunks	2373
storage_size	22.34 MB

▼ chunking

▼ articles

chunk_overlap	50
chunk_size	500
num_chunks	2373

> authors

56260613-5a4d-45bb-b17a-27479d5151e2

📄 embedded_documents 28

ⓘ Overview 📊 Metadata 📈 Visualization

➤ Uncategorized

➤ chunking

⌄ embedding

⌄ articles

embedding_model_id sentence-transformers/all-MiniLM-L6-v2

embedding_size 384

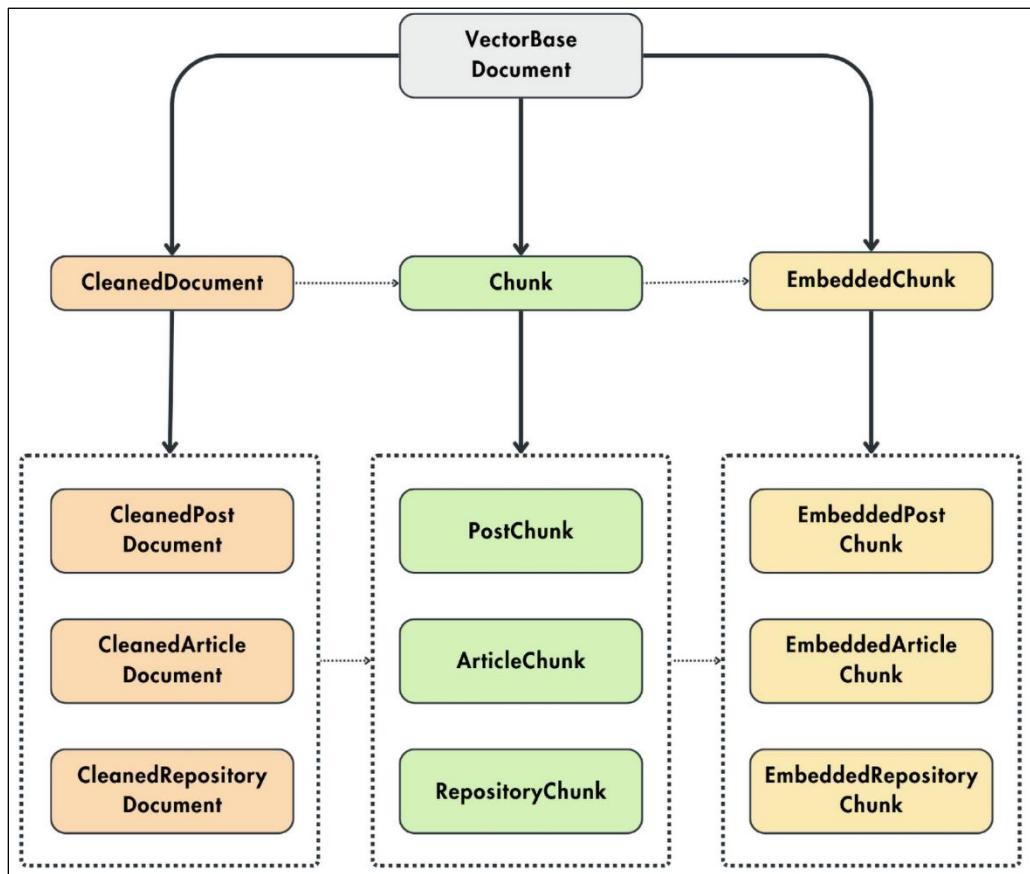
max_input_length 256

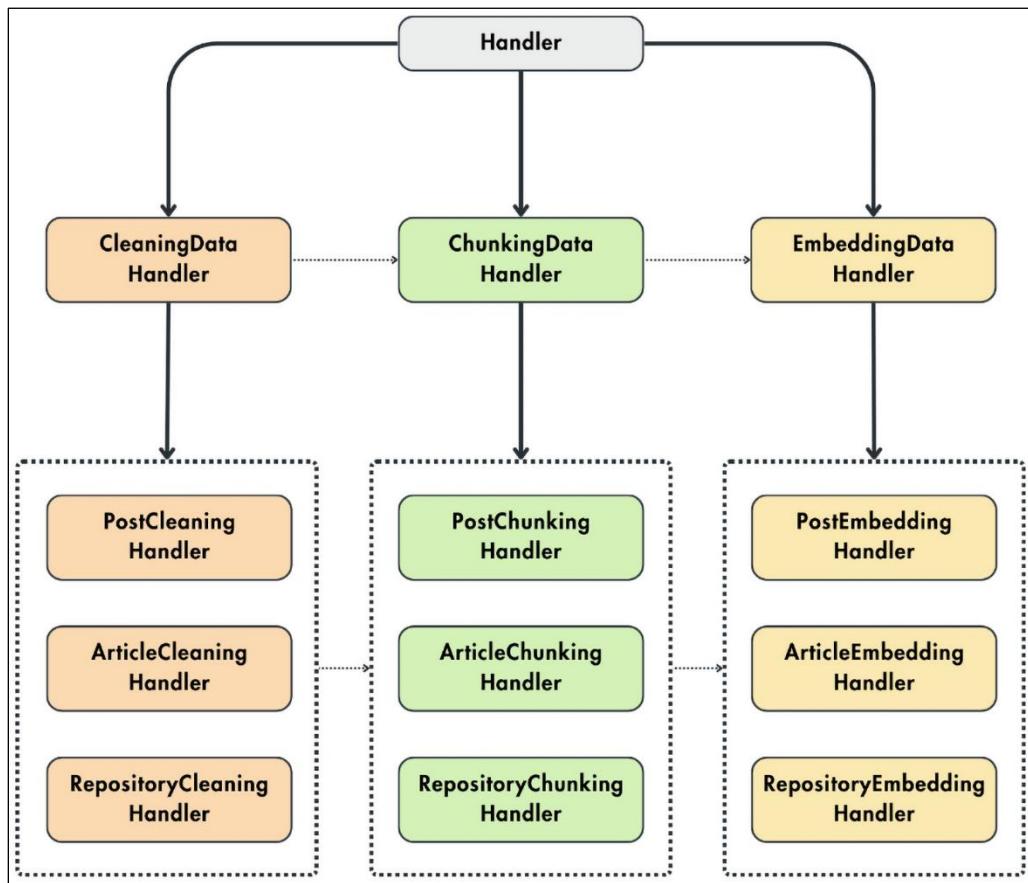
⌄ authors

0 Paul Iusztin

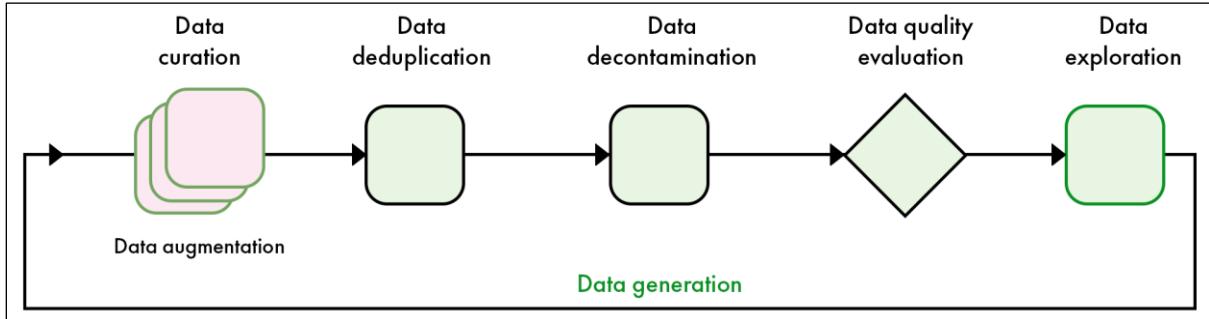
1 Maxime Labonne

2 Alex Vesa

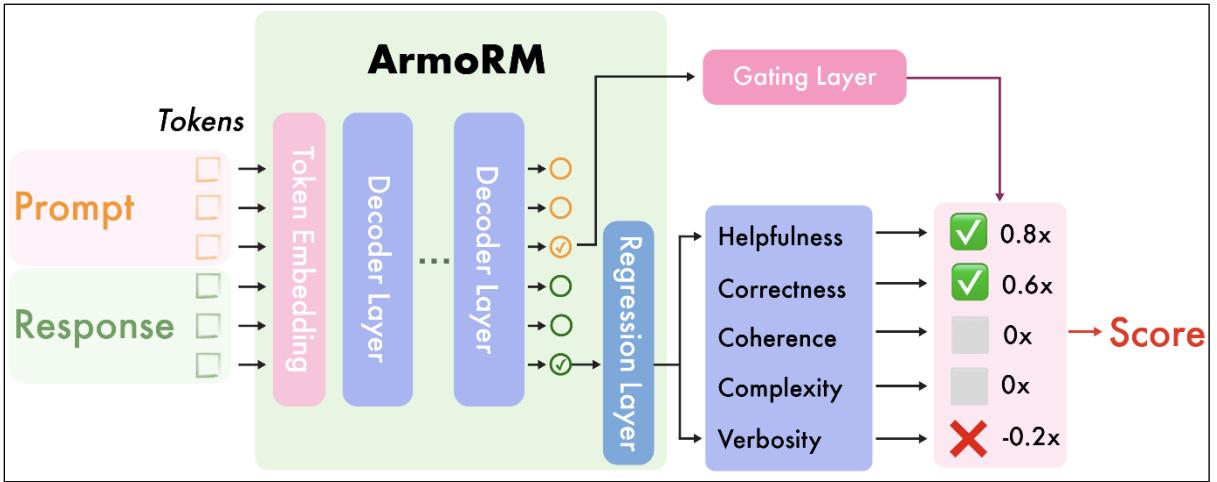




Chapter 5: Supervised Fine-Tuning



Screenshot of the Hugging Face Datasets interface, showing a list of available datasets. The interface includes filters for 'Main', 'Tasks', 'Libraries', 'Languages', and 'Licenses'. On the left, there are filters for 'Modality' (3D, Audio, Geospatial, Image, Tabular, Text, Time-series), 'Size (rows)' (with a slider from <1K to >1T), and 'Format' (json, csv, parquet, imagefolder, soundfolder, webdataset, text). The main list displays 178,406 datasets, each with a preview, last updated date, size, and likes count. Some examples shown include fka/awesome-chatgpt-prompts, Open-Orca/OpenOrca, gsdf/EasyNegative, togethercomputer/RedPajama-Data-1T, tiiuae/falcon-refinedweb, bigcode/the-stack, databricks/databricks-dolly-15k, HuggingFaceFW/fineweb, OpenAssistant/oasst1, Anthropic/hh-rlhf, Nerfgun3/bad_prompt, allenai/dolma, anon8231489123/ShareGPT_Vicuna_unfiltered, and QingyiSi/Alpaca-CoT.



Pending

Filters Sort

1 of 5557

Prompt

Which is more problematic: That good things regularly happen to bad people, or that bad things regularly happen to good people? Respond as if you were an intellectual with a confident and principled take one way or another.

Response

As an intellectual, it is important to acknowledge that both situations can be problematic and have significant ethical implications.

If good things regularly happen to bad people, it can create a sense of injustice and unfairness. It may lead to a lack of trust in the justice system and other institutions, and can cause feelings of resentment and anger in those who perceive themselves as being good and deserving of positive outcomes. Additionally, if bad behavior is consistently rewarded, it can reinforce negative actions and discourage positive actions, creating a toxic and unhealthy culture.

On the other hand, if bad things regularly happen to good people, it can also create a sense of injustice and unfairness. It may lead to feelings of hopelessness and disillusionment in those who are victimized, and can cause a breakdown of social cohesion and trust. Additionally, if good behavior is consistently punished, it can discourage positive actions and lead to a culture of cynicism and despair.

Ultimately, as an intellectual, I believe that both situations are problematic and require careful consideration and action. We must work to create a just and equitable society that rewards positive behavior and punishes negative behavior, regardless of who is involved. This requires a commitment to fairness, empathy, and a willingness to take action to address injustice when we see it.

Is the response relevant for the given prompt? *

Yes No

Does the response include any of the following?

1 Hate Speech 2 Inappropriate content 3 Not English
 4 Personal information 5 Sexual content
 6 Untruthful info 7 Violent content

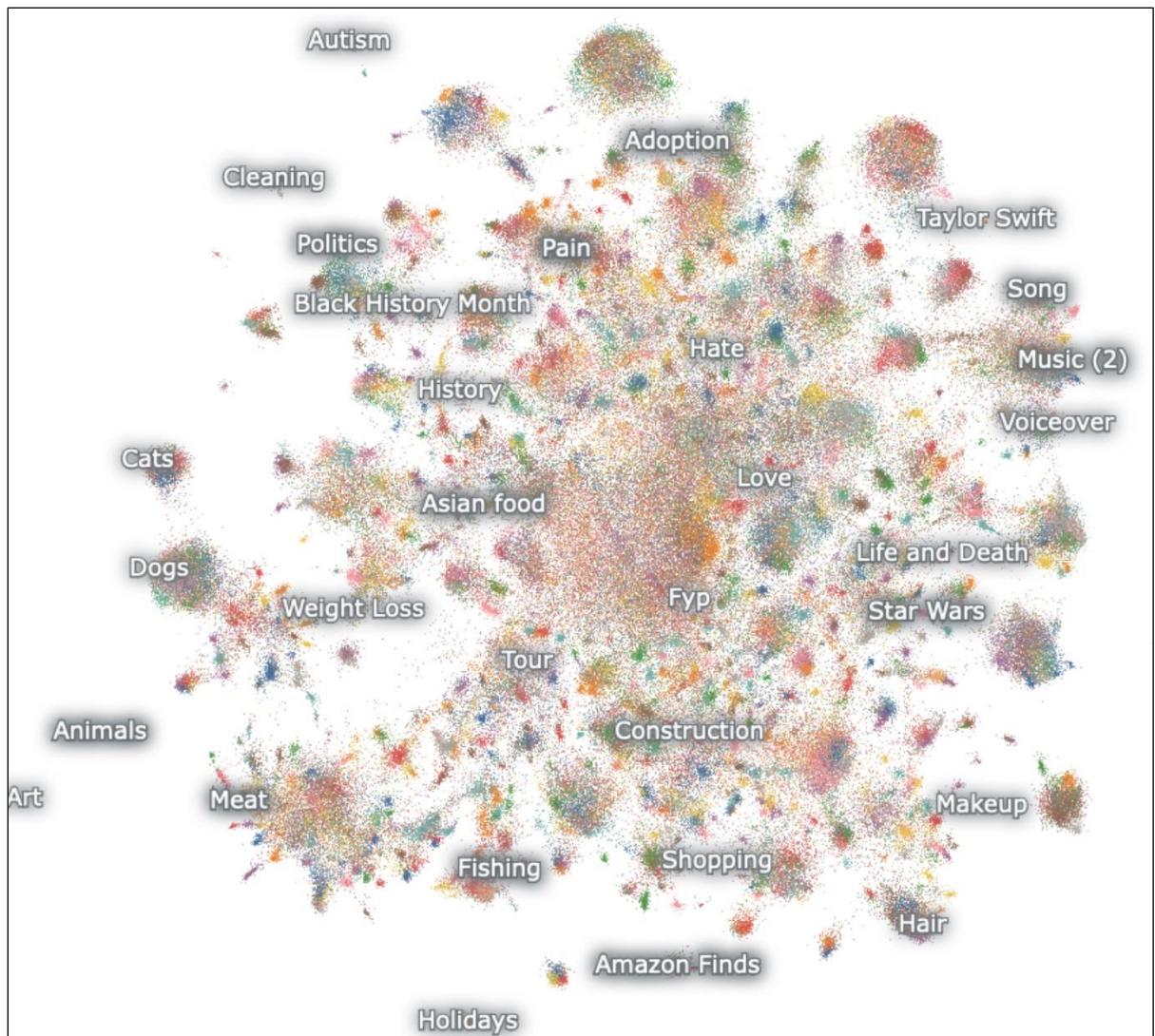
Rate the quality of the response: *

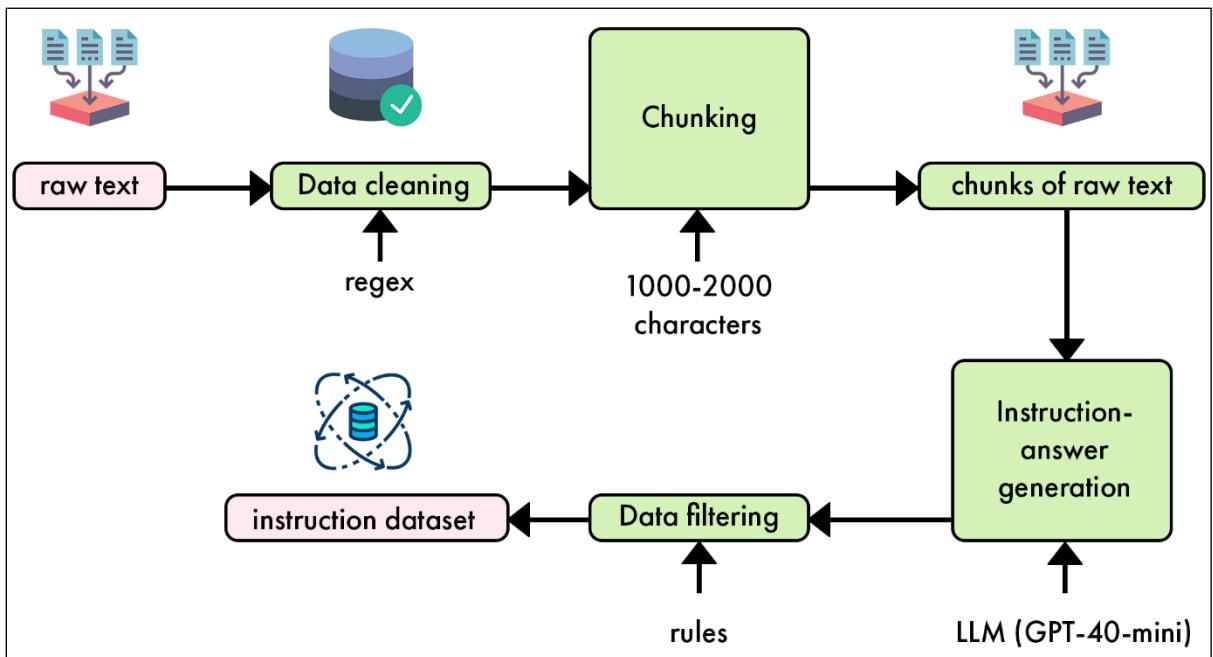
1 2 3 4 5 6 7 8 9 10

Provide a correction to the response:

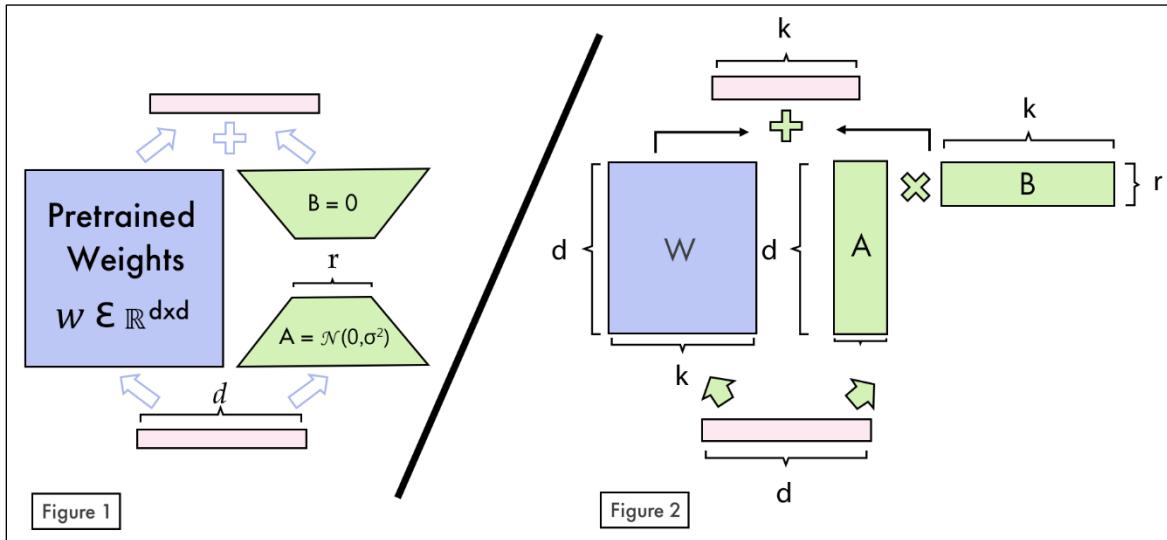
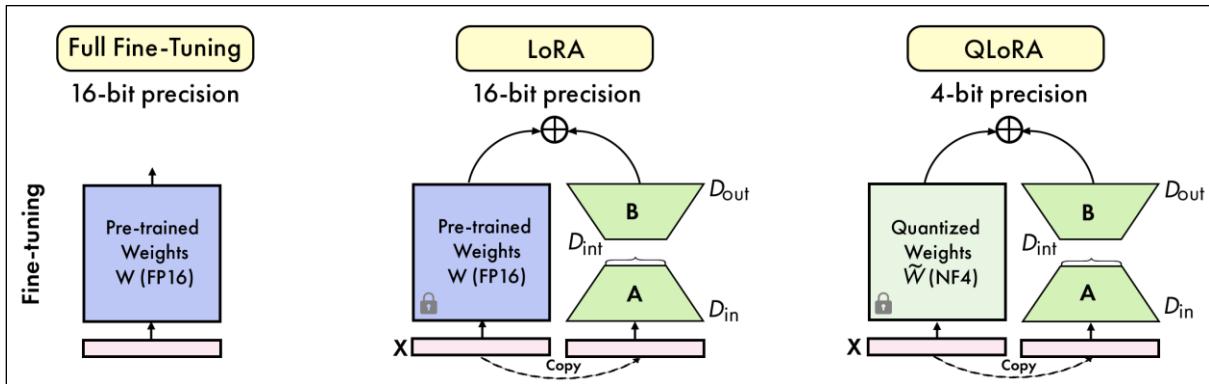
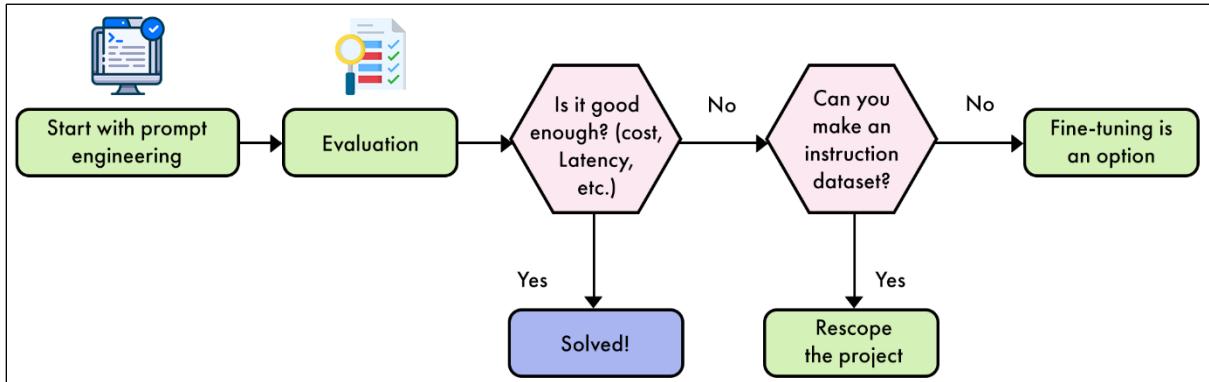
Discard Save as draft Submit

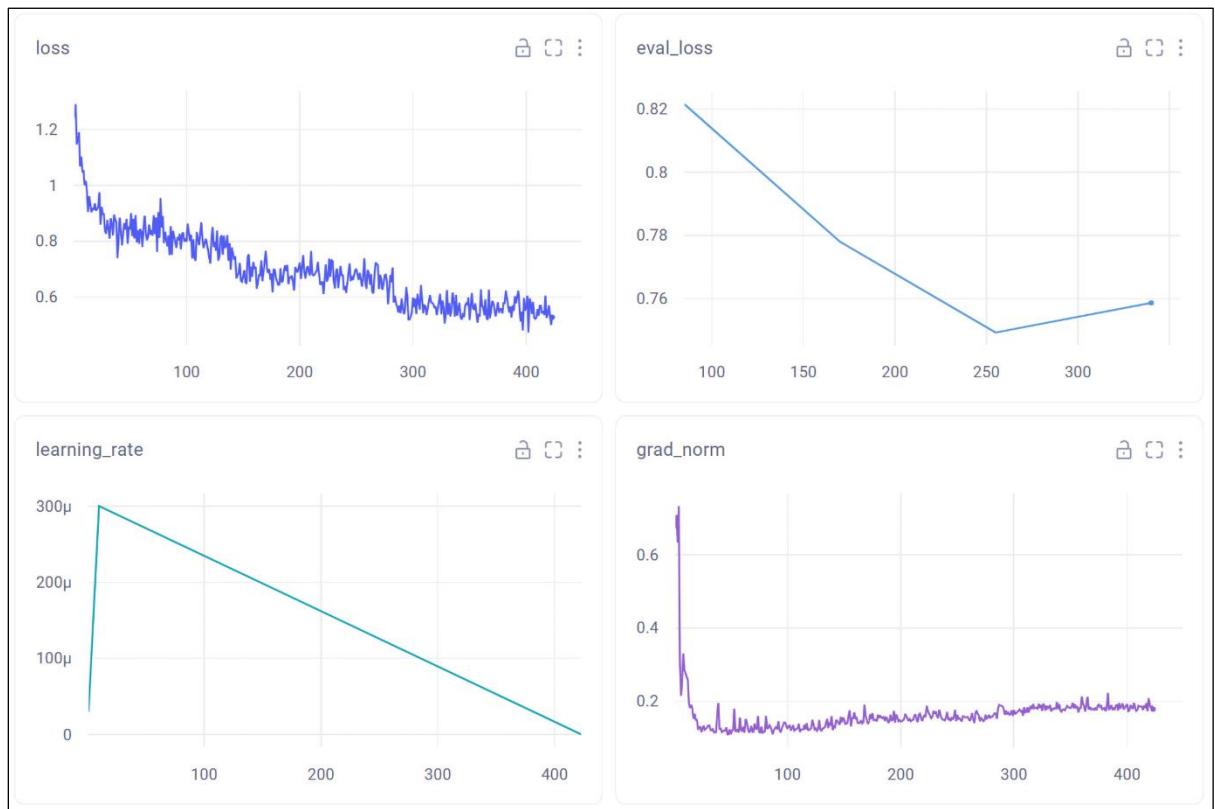
GUIDELINES PROGRESS 28 of 5590 0.50%



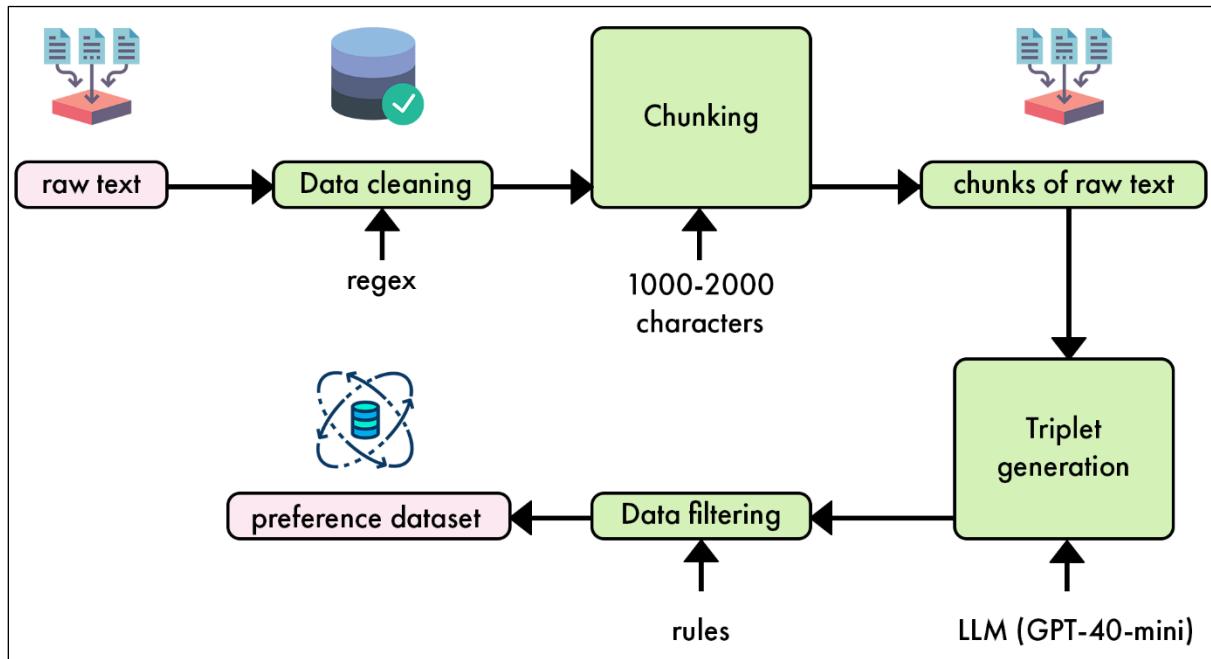
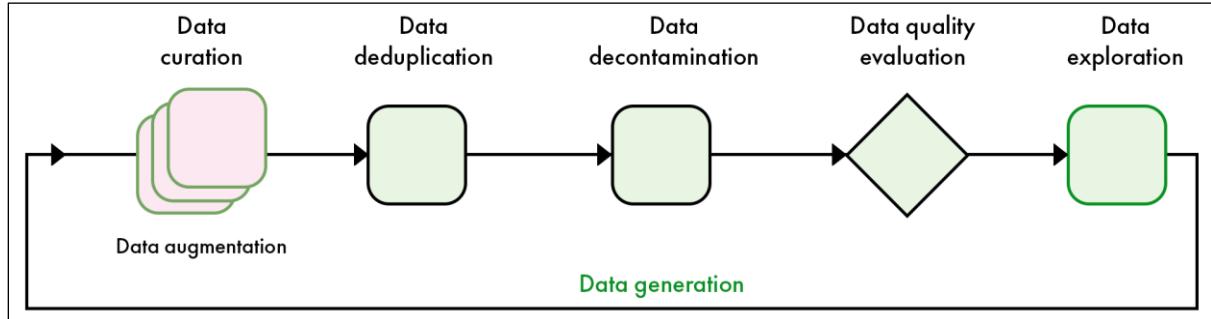


instruction string · lengths	output string · lengths
57-67 28.2%	561-628 21.7%
Explain how list comprehensions improve iteration speed in Python.	List comprehensions serve as a highly efficient mechanism for iterating over lists in Python, enabling operations to be expressed in a single line. This method eliminates the need to explicitly call the .append method during iteration, which contributes to its speed. Python processes list comprehensions similarly to traditional for loops at the bytecode level, yet the overhead of function calls is reduced, resulting in a performance boost. As evidenced by benchmarks, list comprehensions can achieve speeds that are 200 times faster than conventional looping methods.
Define what an LLM Twin is and its purpose.	An LLM Twin is an AI-driven character designed to replicate your own writing style, personality, and voice. By leveraging large language...
Describe how audio narrations can enhance the reading experience.	Audio narrations can significantly enhance the reading experience by adding an auditory dimension to the text. This feature allows readers to...
Describe the role of a vector database in embedding storage.	A vector database serves as a crucial repository for storing embeddings along with their associated metadata, such as the embedded text, source...
Explain the concept of quantization in the context of LLMs.	Quantization in the context of LLMs refers to the process of reducing the precision of the numbers used to represent the model's weights. Thi...

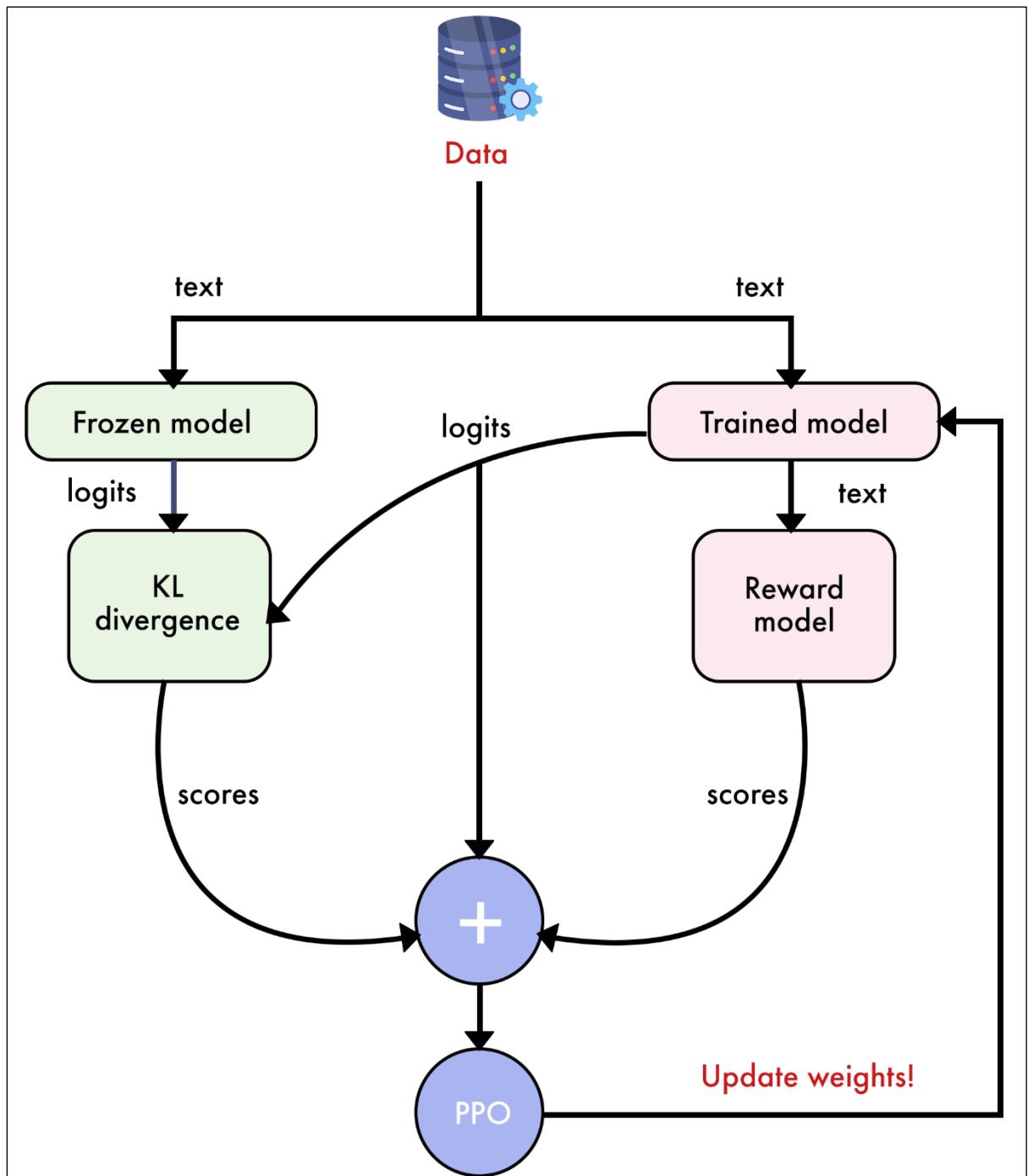


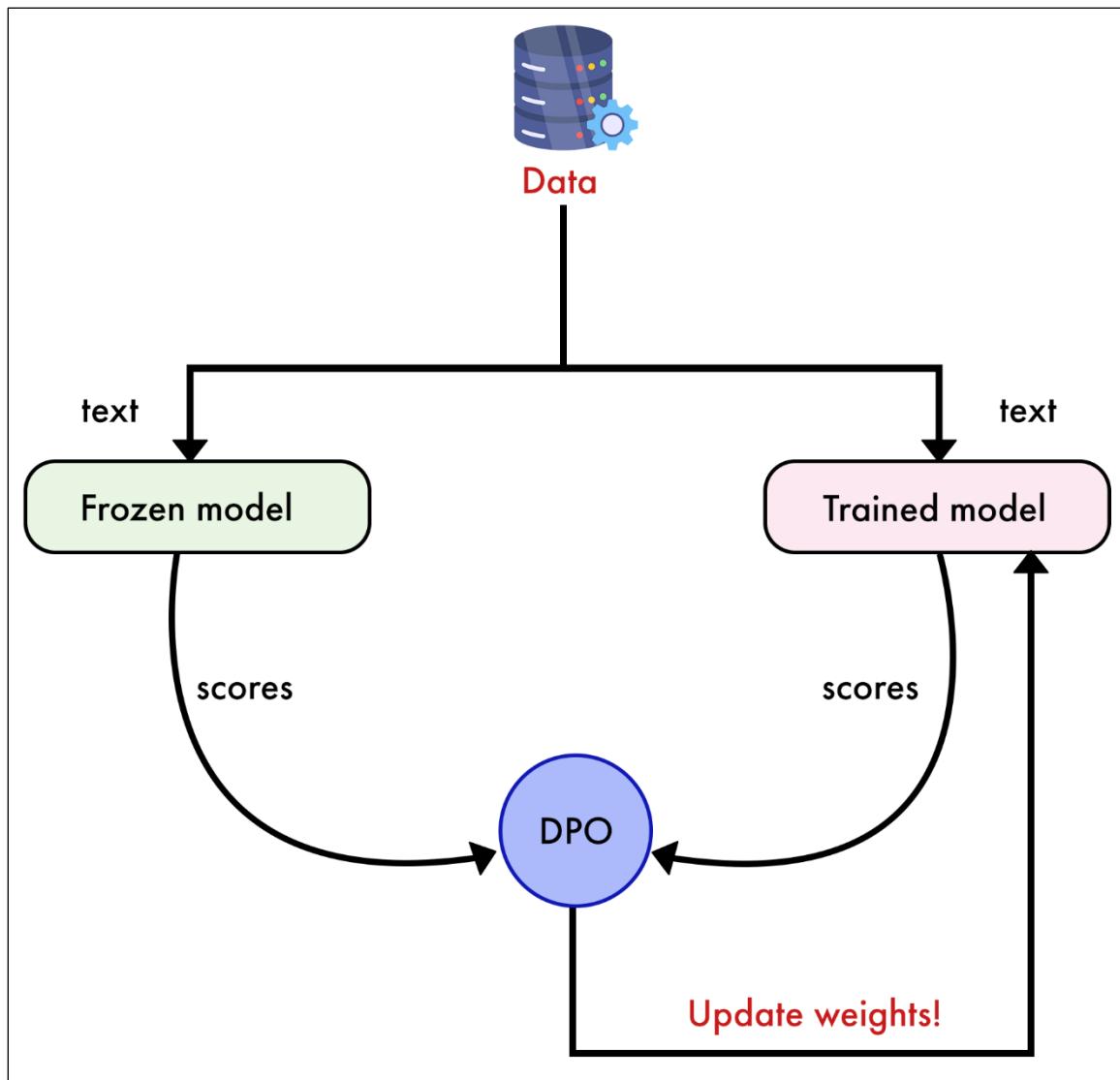


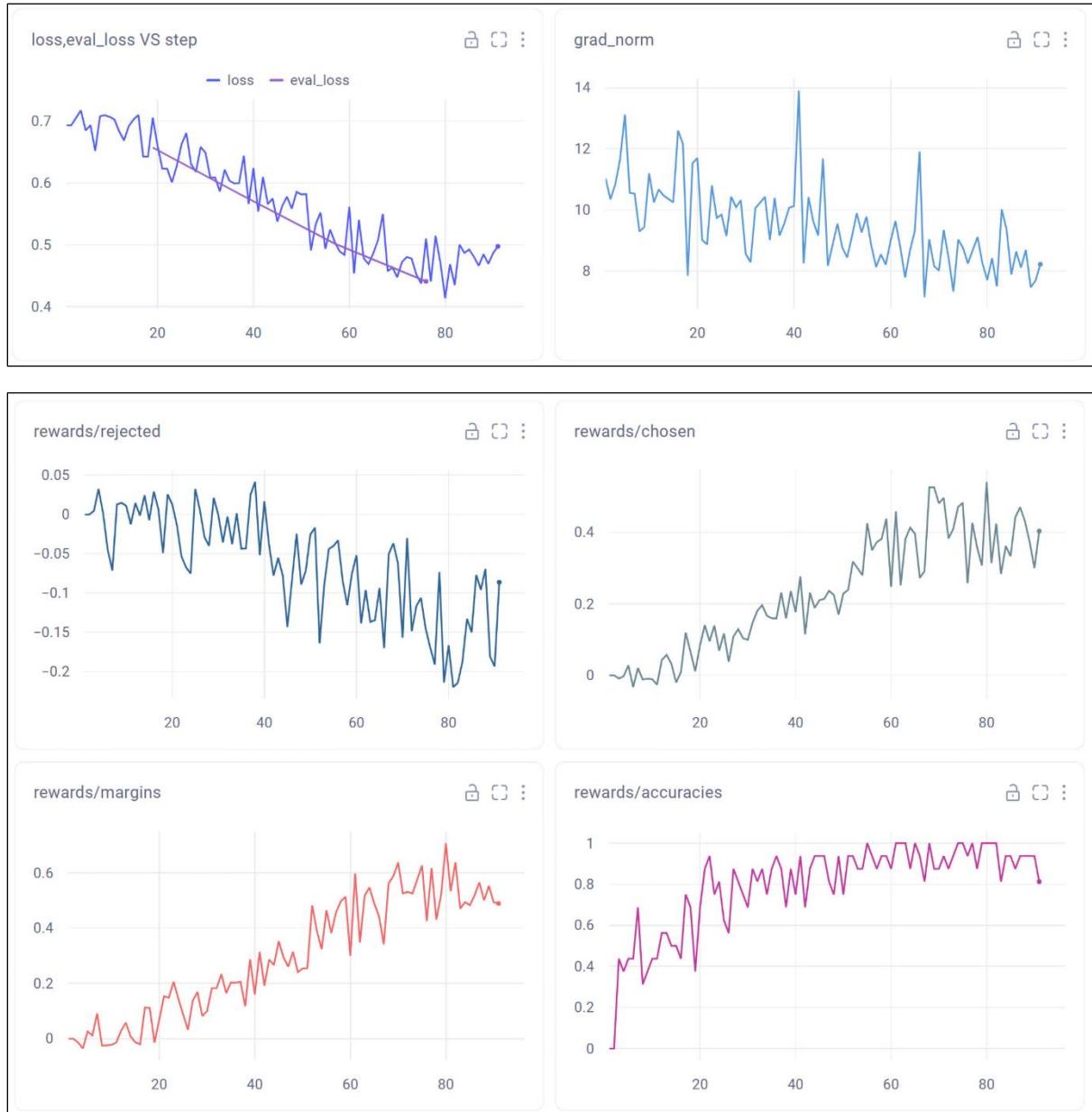
Chapter 6: Fine-Tuning with Preference Alignment



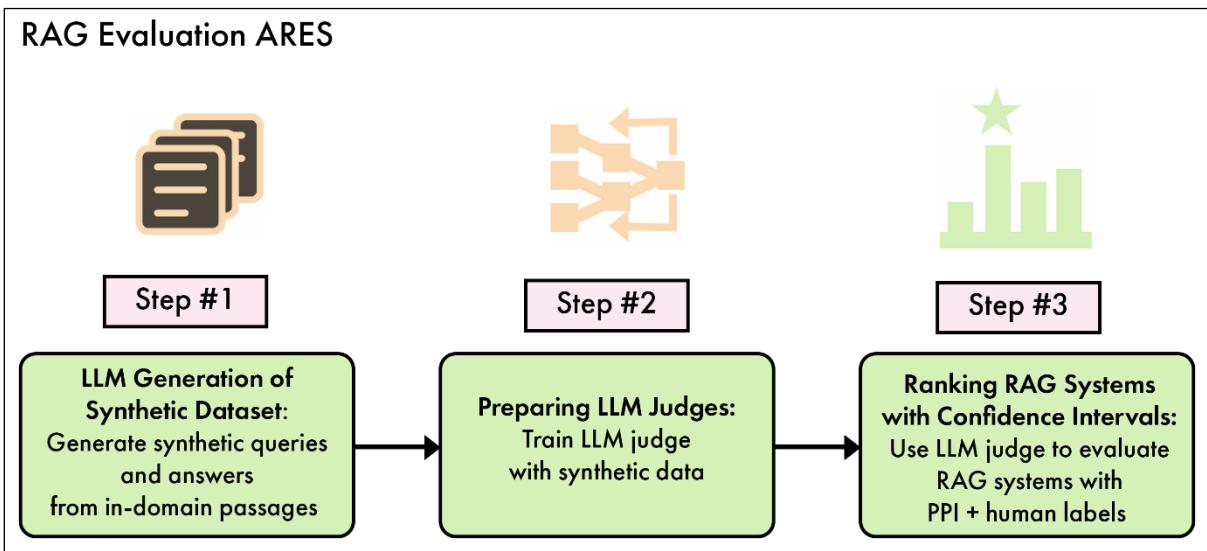
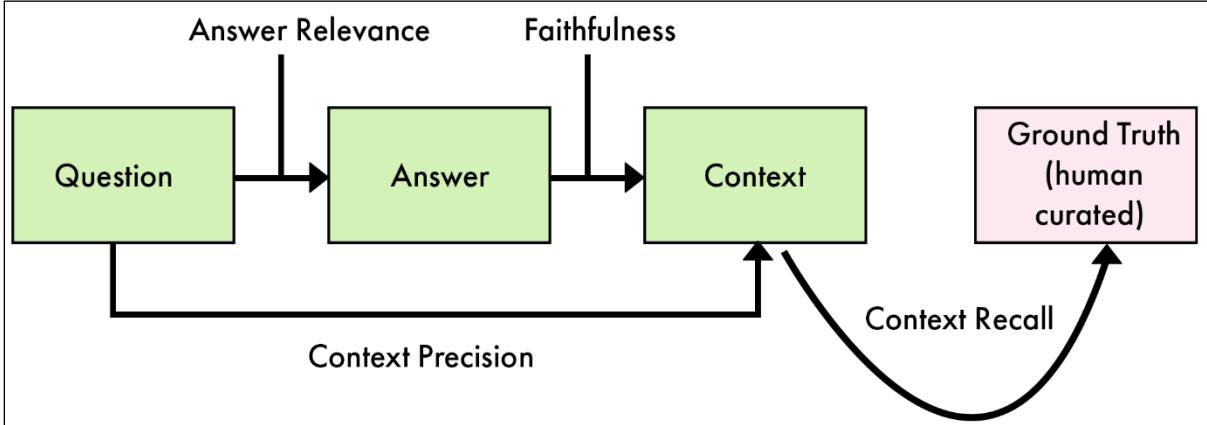
prompt string · lengths	rejected string · lengths	chosen string · lengths
 68-79 20.5%	 101-136 32.4%	 100-155 63.4%
What is the purpose of instruction datasets in guiding language models?	Instruction datasets are designed to efficiently guide a language model toward a specific task, such as news classification.	Instruction datasets offer an efficient way to guide a language model toward a specific task like news classification.
What components will you learn to build in the Hands on LLMs course?	In the Hands on LLMs course, you will learn to build a real-time streaming pipeline, a fine-...	There are 3 components you will learn to build during the course n a real time streaming...
What is the purpose of the function fetch_all_cleaned_content?	The function fetch_all_cleaned_content is used to efficiently retrieve a list of cleaned...	To easily fetch data from a Qdrant database, you can utilize the Python function...
What does the scroll method do in the context of data retrieval?	The scroll method is used to start scrolling through the database, allowing for the...	Initialize the Scroll Start by scrolling through the database using the scroll method,...



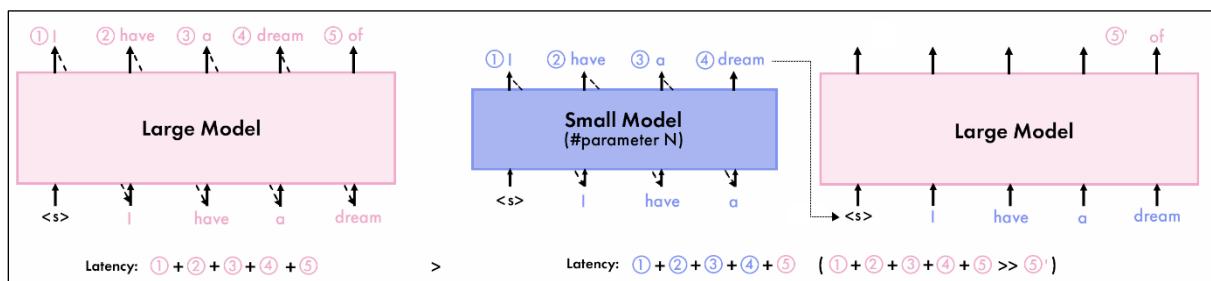
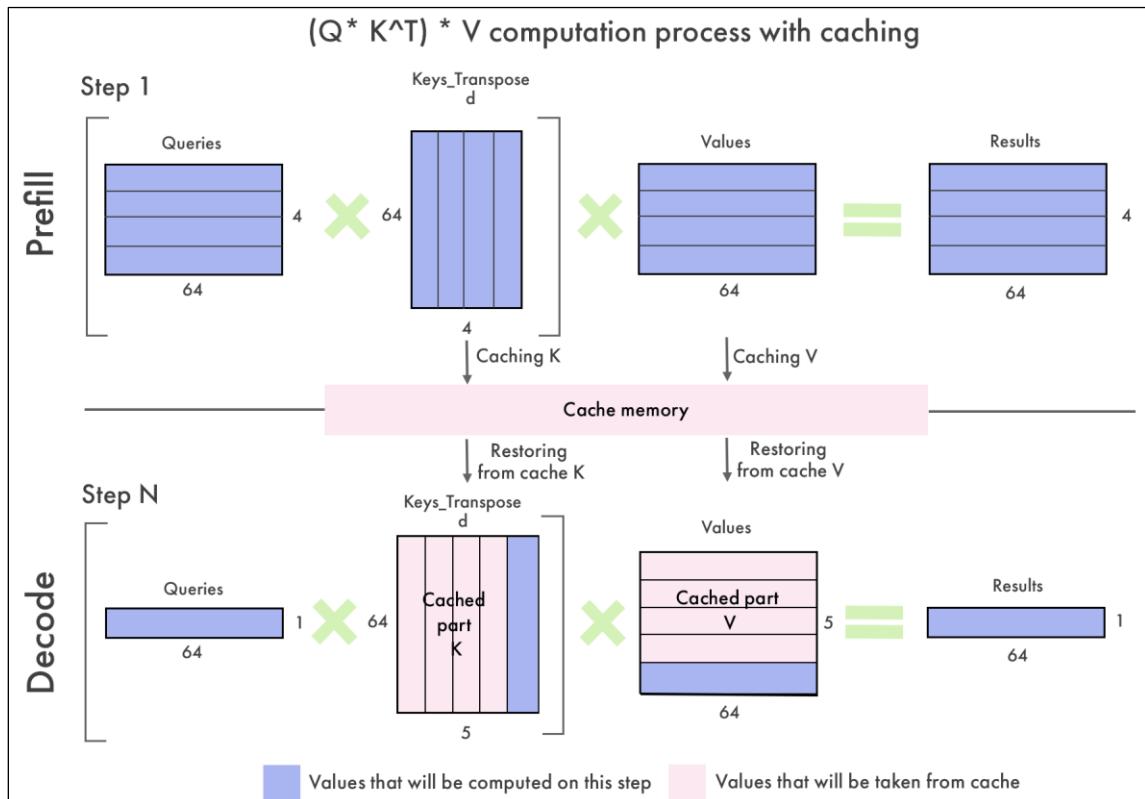
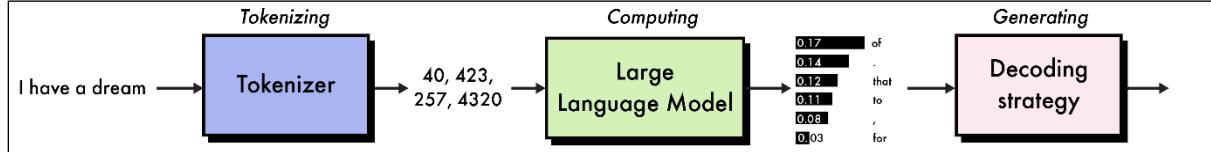


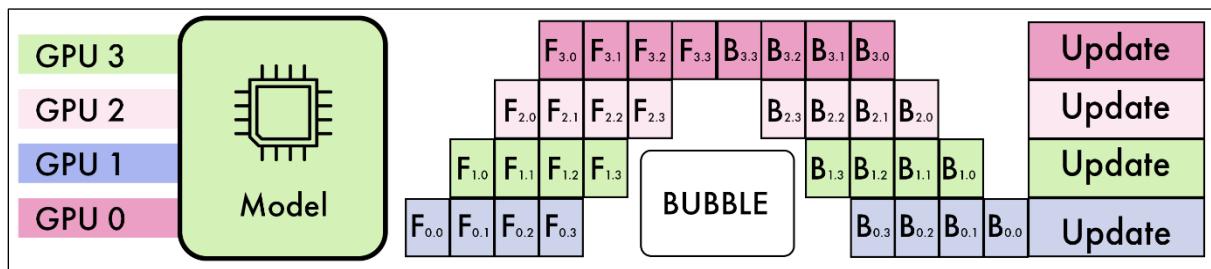
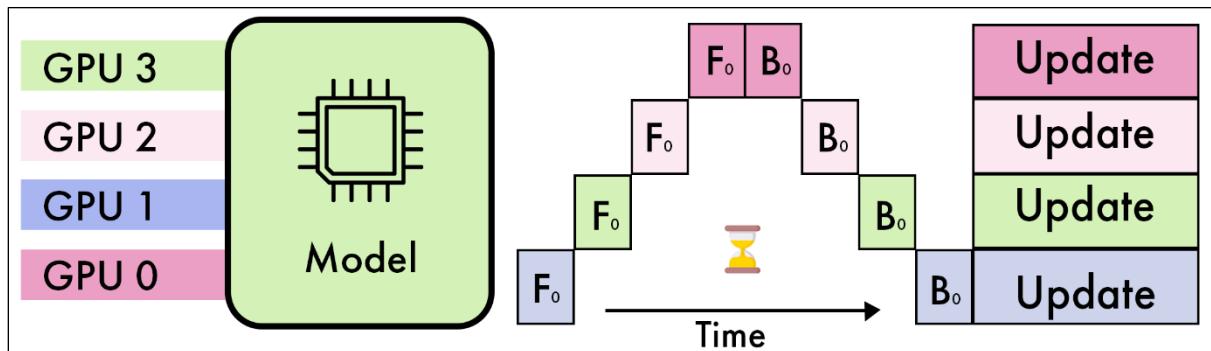
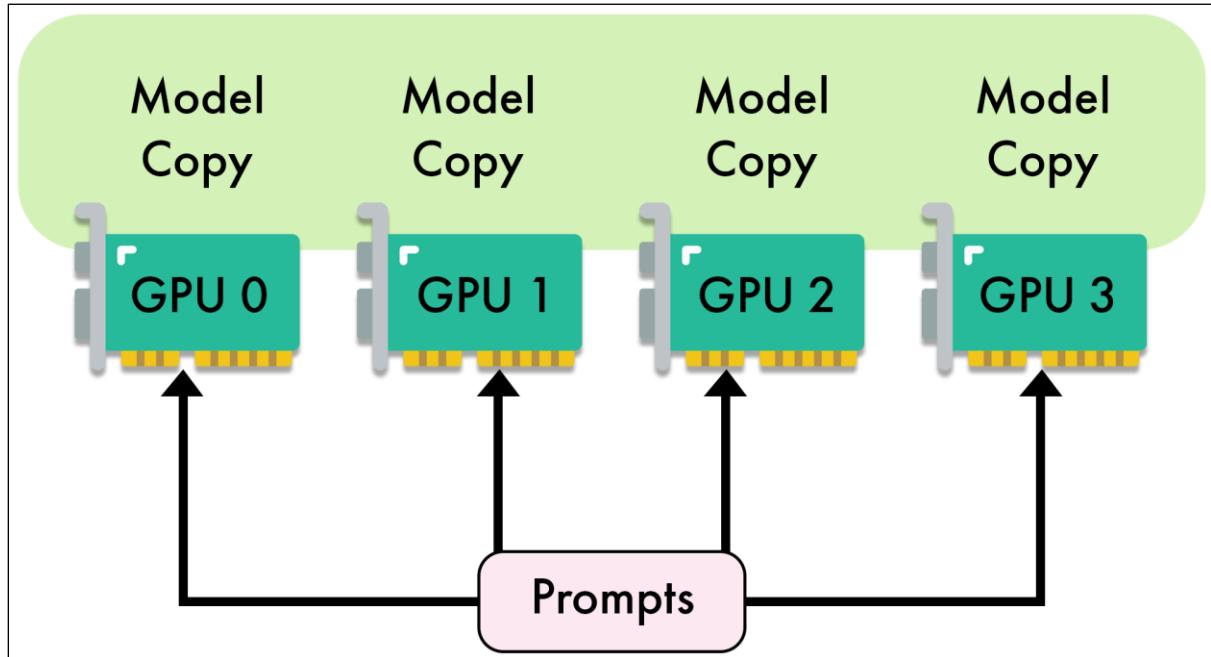


Chapter 7: Evaluating LLMs



Chapter 8: Inference Optimization





Without parallelism

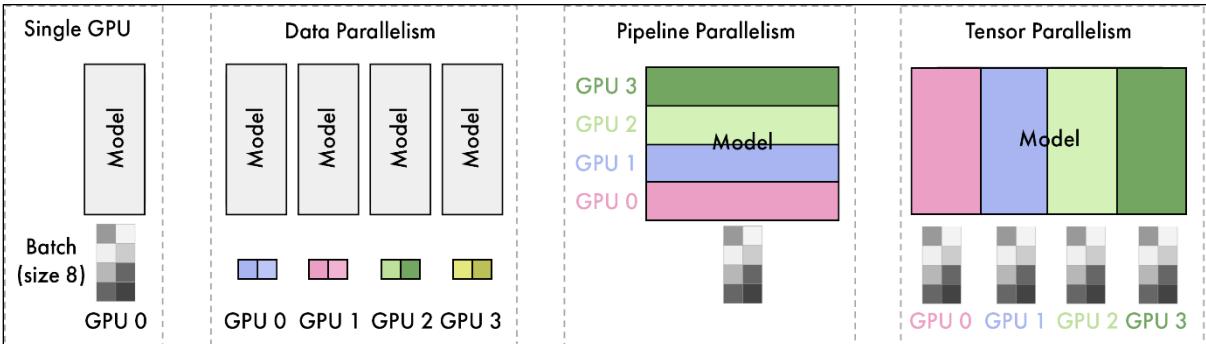
$$\begin{matrix} X \\ \begin{array}{|c|c|}\hline & & \\ \hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} \end{matrix} \cdot \begin{matrix} W \\ \begin{array}{|c|c|}\hline & & \\ \hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} \end{matrix} = \begin{matrix} Y \\ \begin{array}{|c|c|}\hline & & \\ \hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} \end{matrix}$$

With tensor parallelism

$$\begin{matrix} X \\ \begin{array}{|c|c|}\hline & & \\ \hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} \end{matrix} \cdot \begin{matrix} W_0 \\ \begin{array}{|c|}\hline & \\ \hline \end{array} \end{matrix} = \begin{matrix} Y_0 \\ \begin{array}{|c|}\hline & \\ \hline \end{array} \end{matrix}$$

$$\begin{matrix} X \\ \begin{array}{|c|c|}\hline & & \\ \hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} \end{matrix} \cdot \begin{matrix} W_1 \\ \begin{array}{|c|}\hline & \\ \hline \end{array} \end{matrix} = \begin{matrix} Y_1 \\ \begin{array}{|c|}\hline & \\ \hline \end{array} \end{matrix}$$

$+ \rightarrow Y$

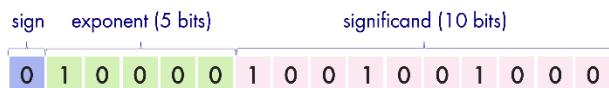


32-bit float (FP32)



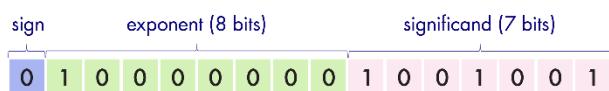
$$(-1)^0 \times 2^{128-127} \times 1.5707964 = 3.1415927$$

16-bit float (FP16)



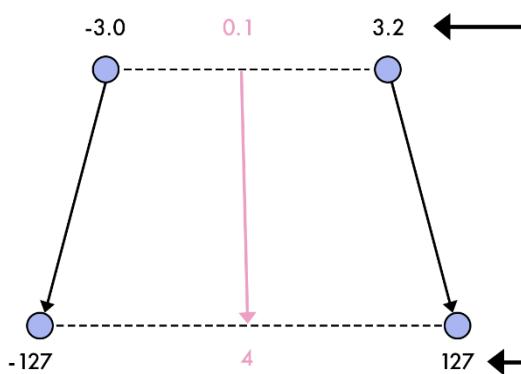
$$(-1)^0 \times 2^{128-127} \times 1.571 = 3.141$$

bfloat16 (BF16)

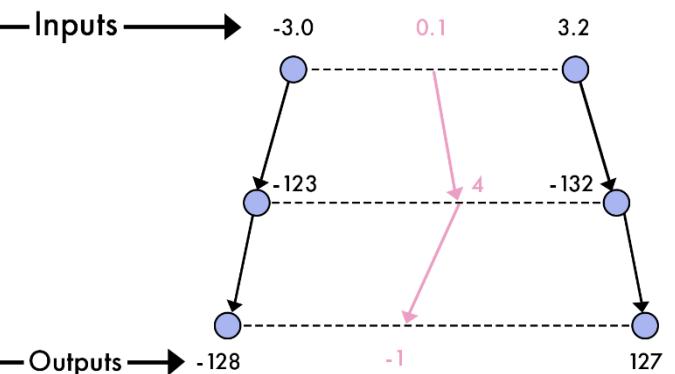


$$(-1)^0 \times 2^{128-127} \times 1.5703125 = 3.140625$$

Absmax quantization



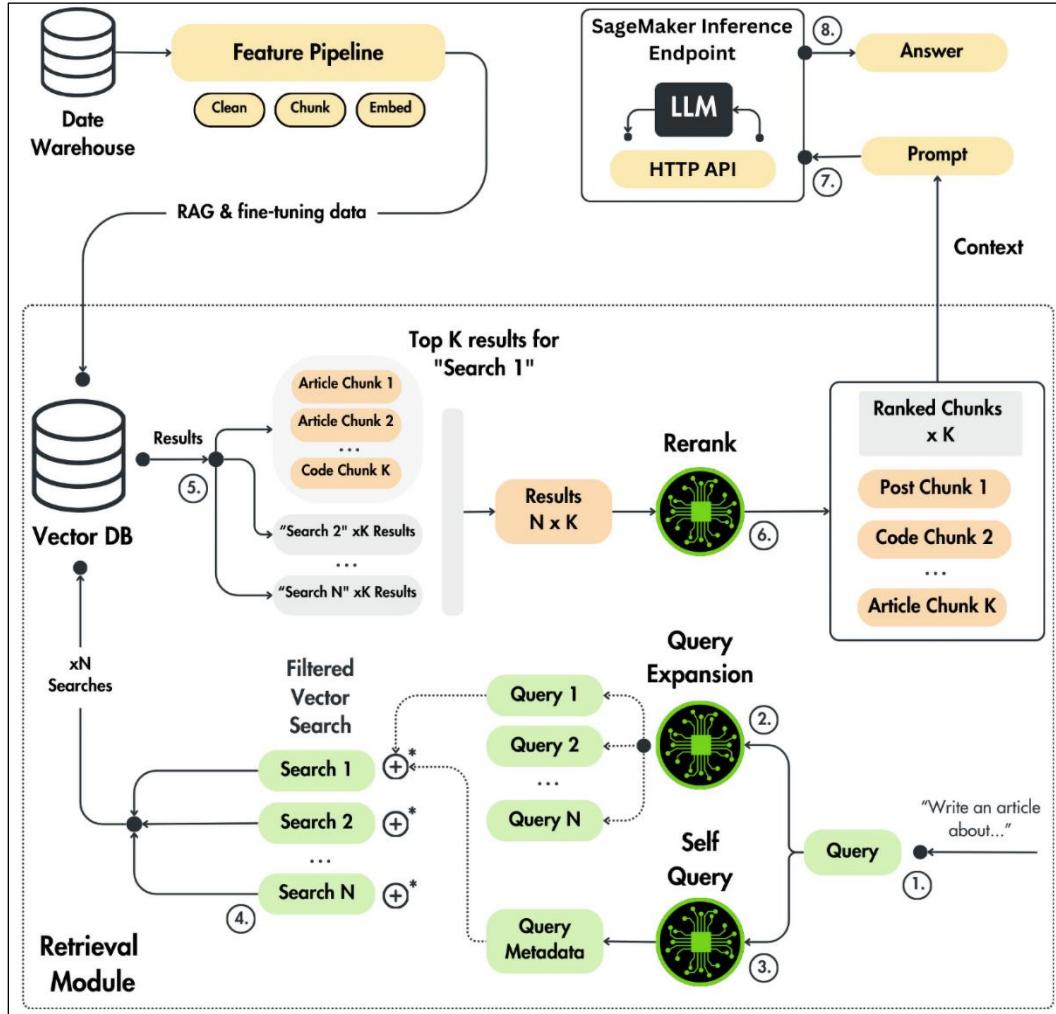
Zero-point quantization

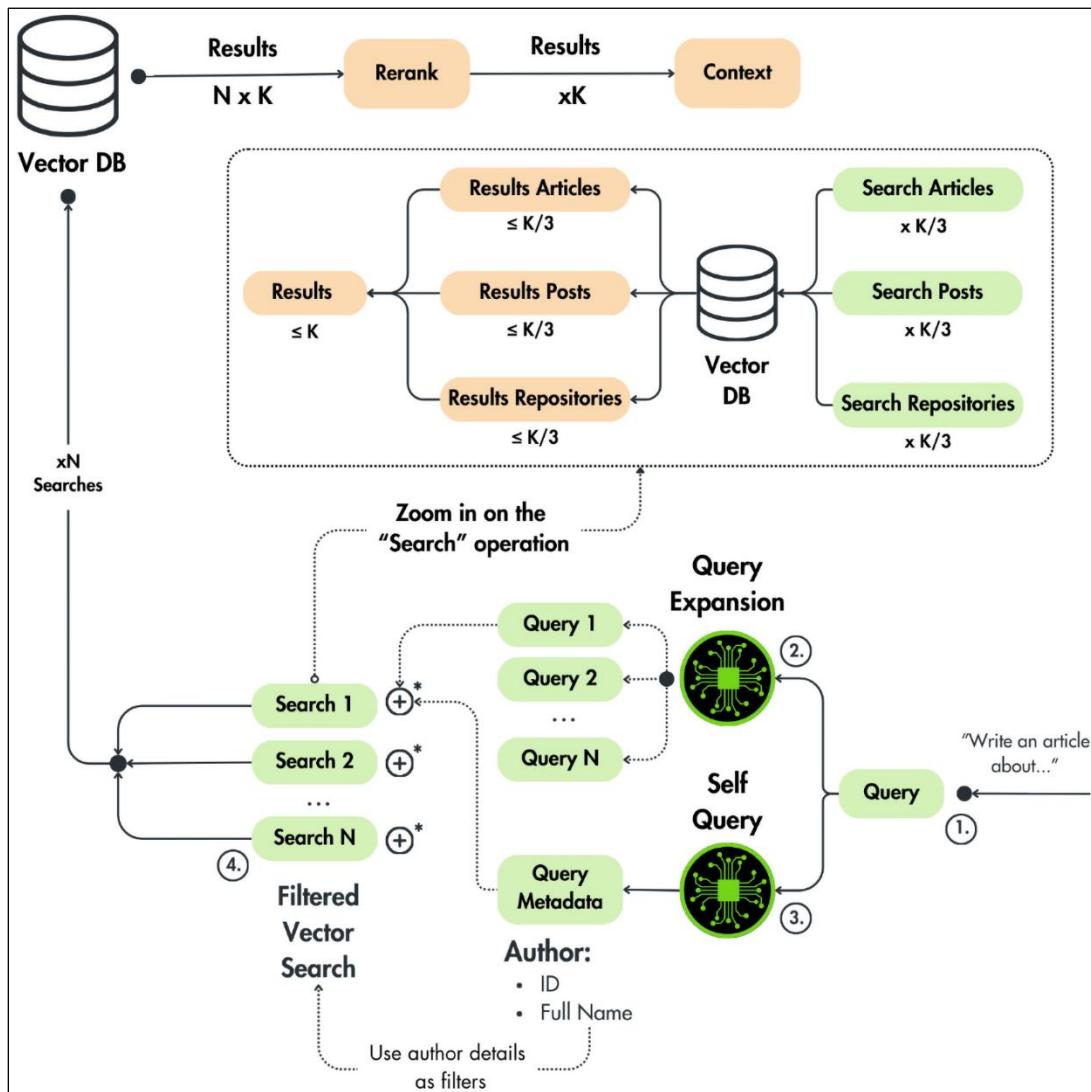


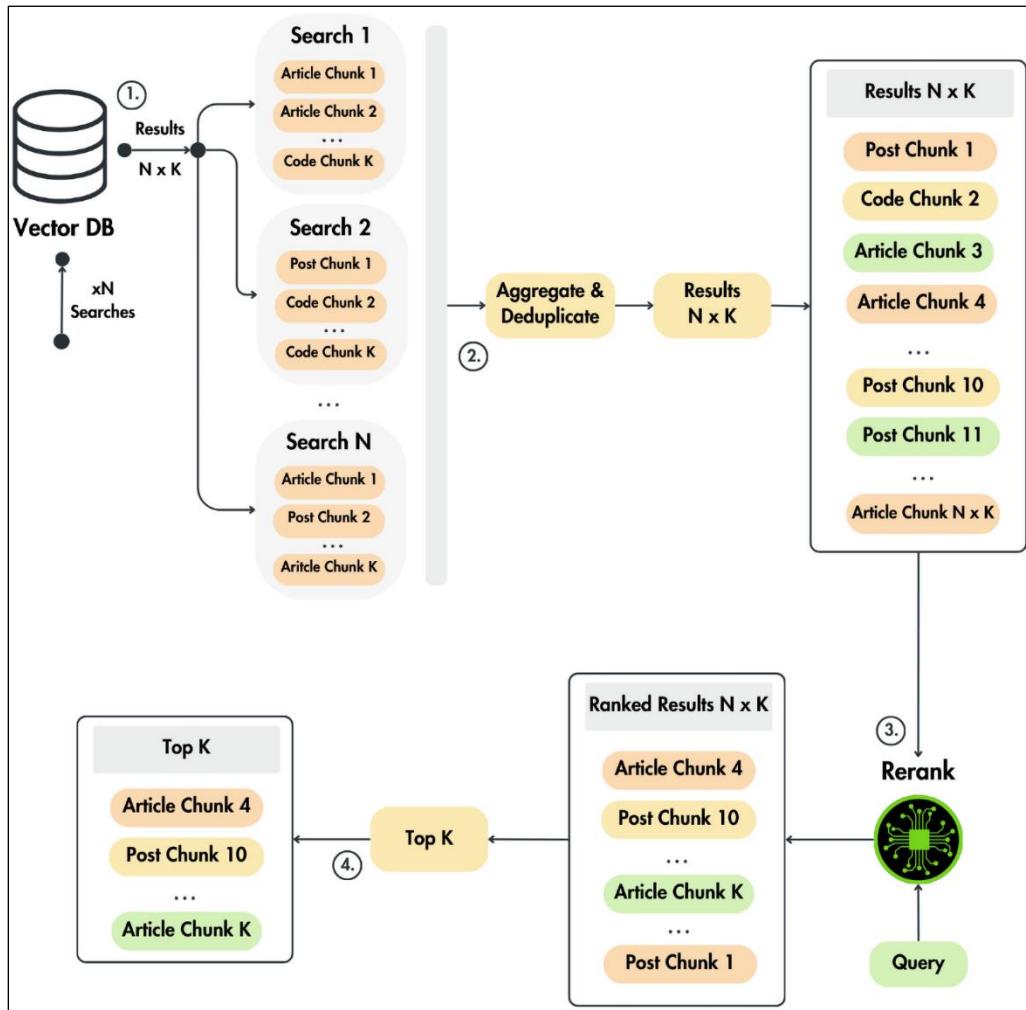
$$X_{F16} = \begin{bmatrix} 2 & 45 & -1 & 17 & -1 \\ 0 & 12 & 3 & -63 & 2 \\ 1 & 37 & 1 & -83 & 0 \end{bmatrix} . W_{F16} = \begin{bmatrix} -1 & 0 \\ 2 & 0 \\ 0 & -2 \\ 1 & -2 \\ -1 & 2 \end{bmatrix}$$

. **Corresponding rows**

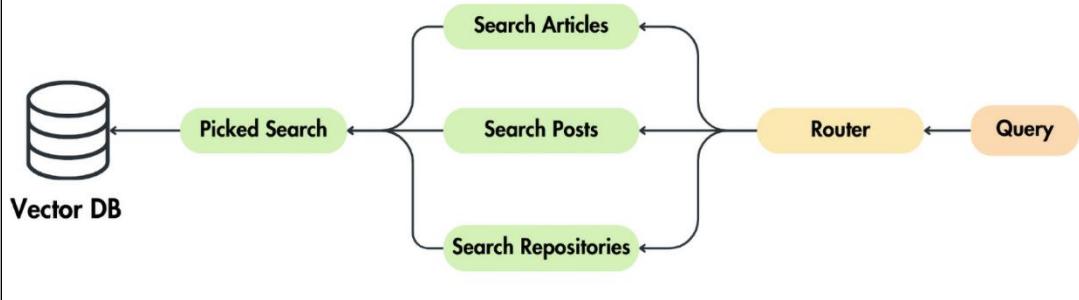
Chapter 9: RAG Inference Pipeline







Routing



Conversation History

- User Prompt
- Generated Answer
- ...
- User Prompt
- Generated Answer

Option 1:
Keep only the latest K

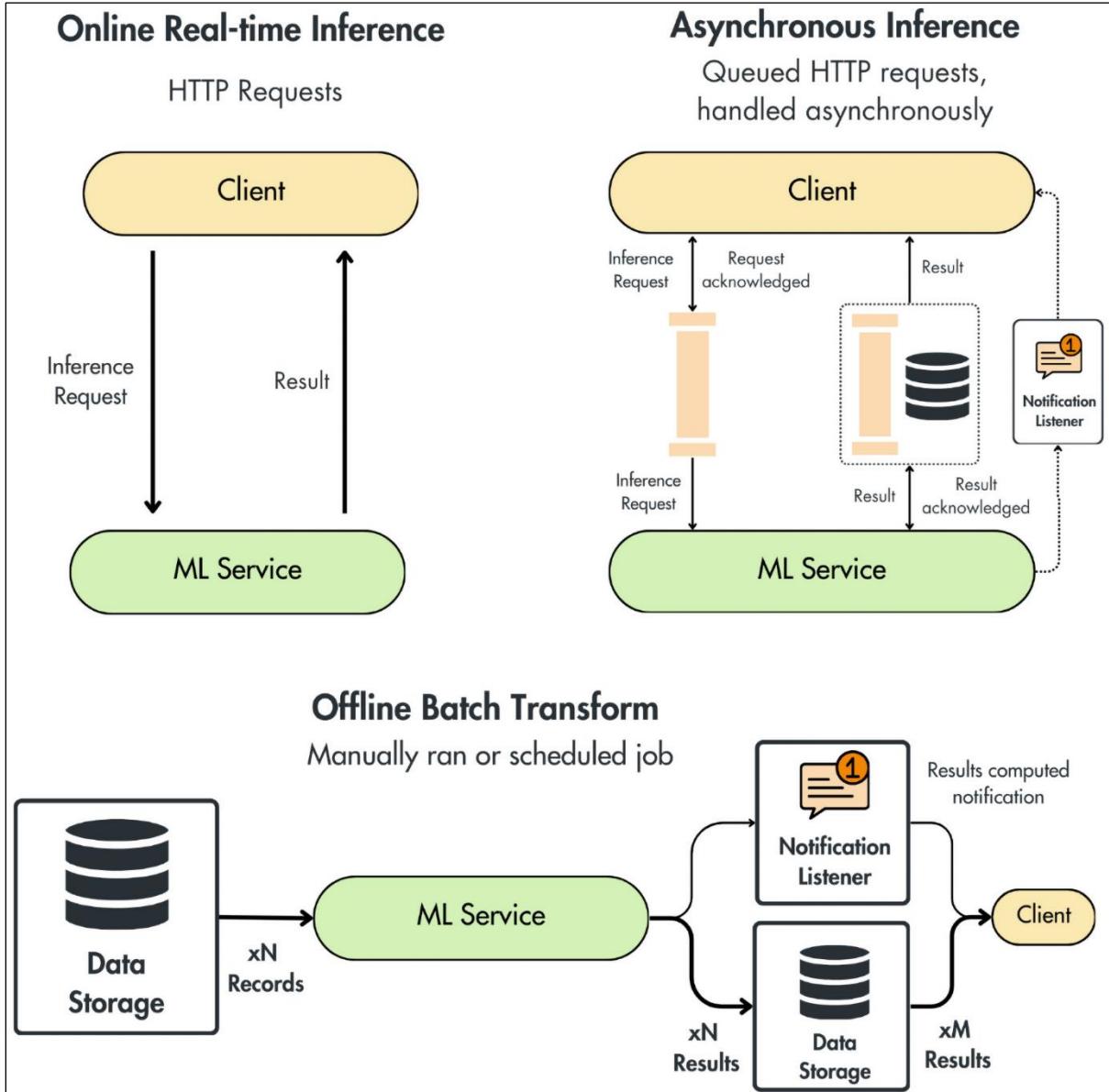


Memory

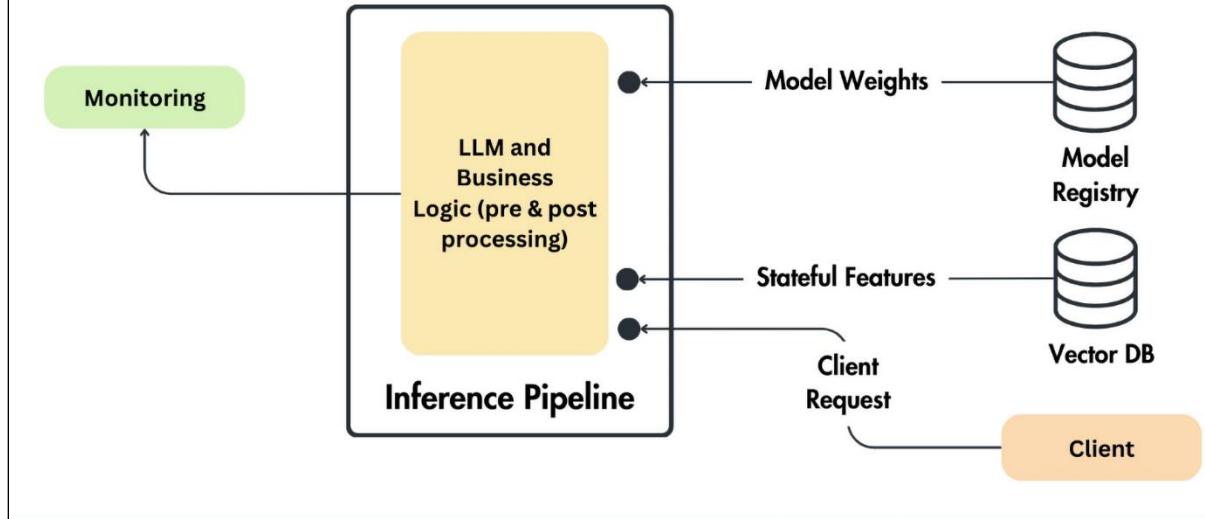
Option 2:
Summary + latest K

- Conversation Summary
- User Prompt
- Generated Answer
- User Prompt
- Generated Answer
- User Prompt
- Generated Answer

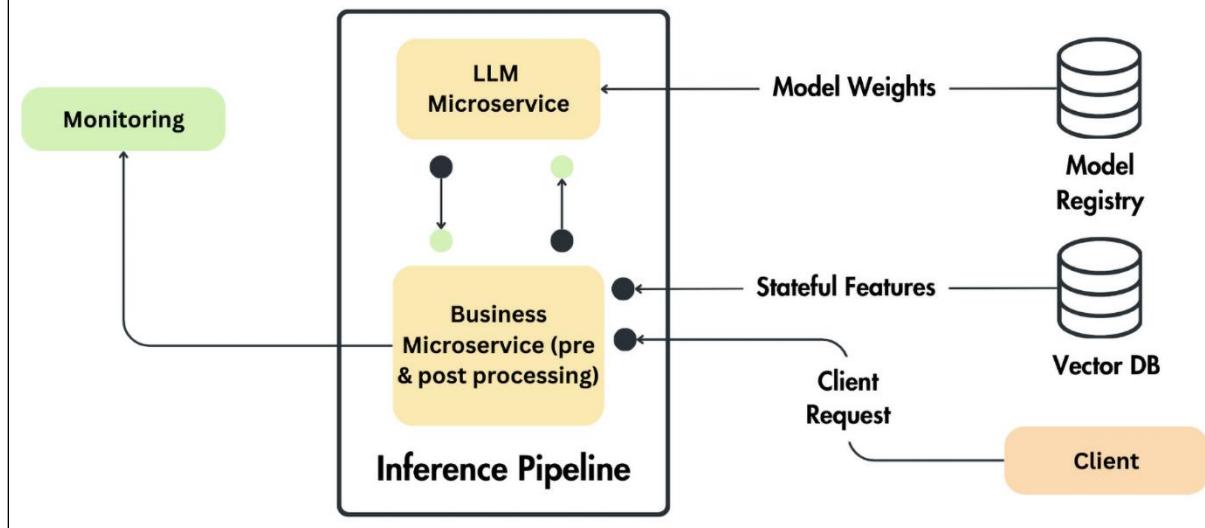
Chapter 10: Inference Pipeline Deployment

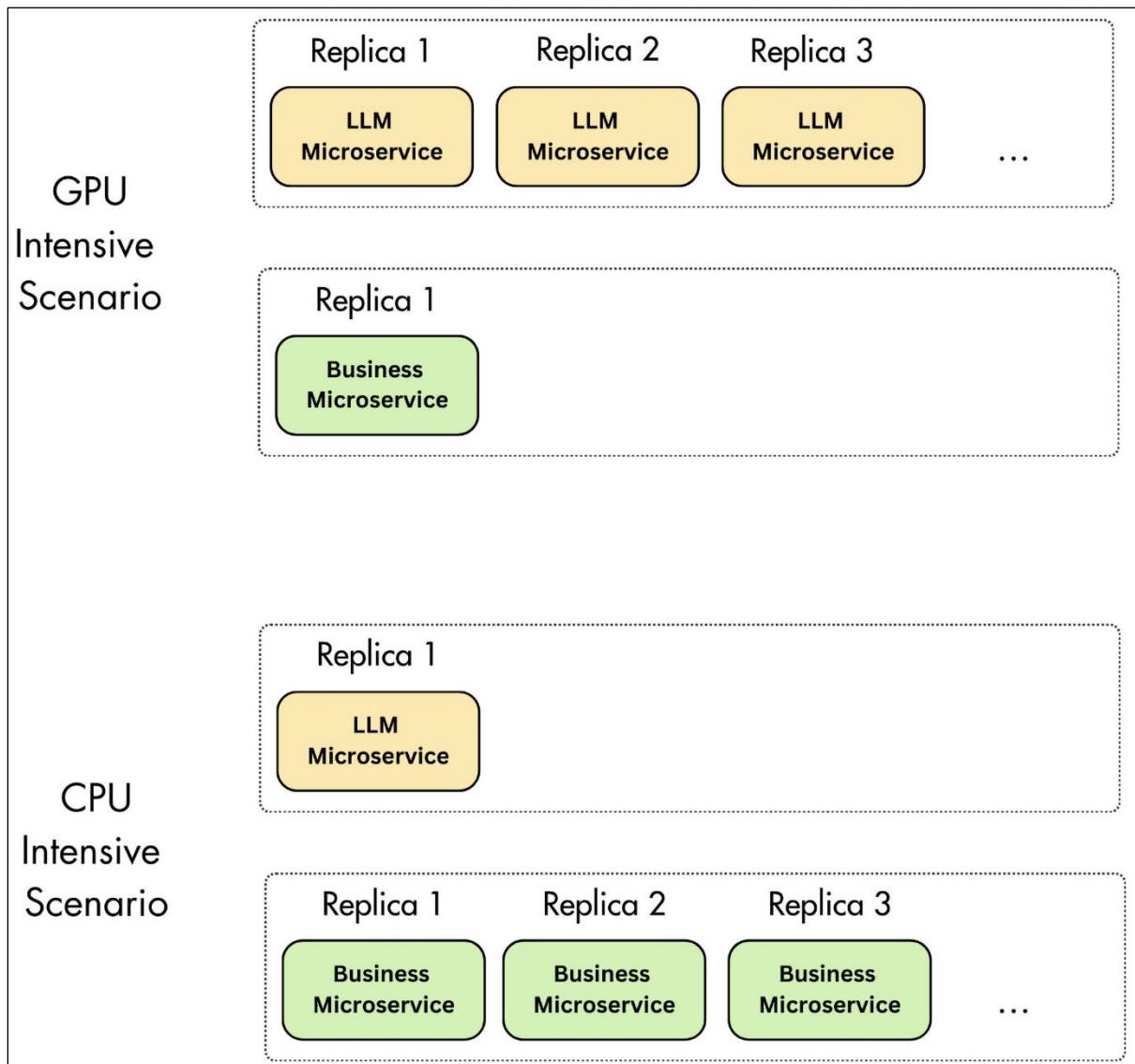


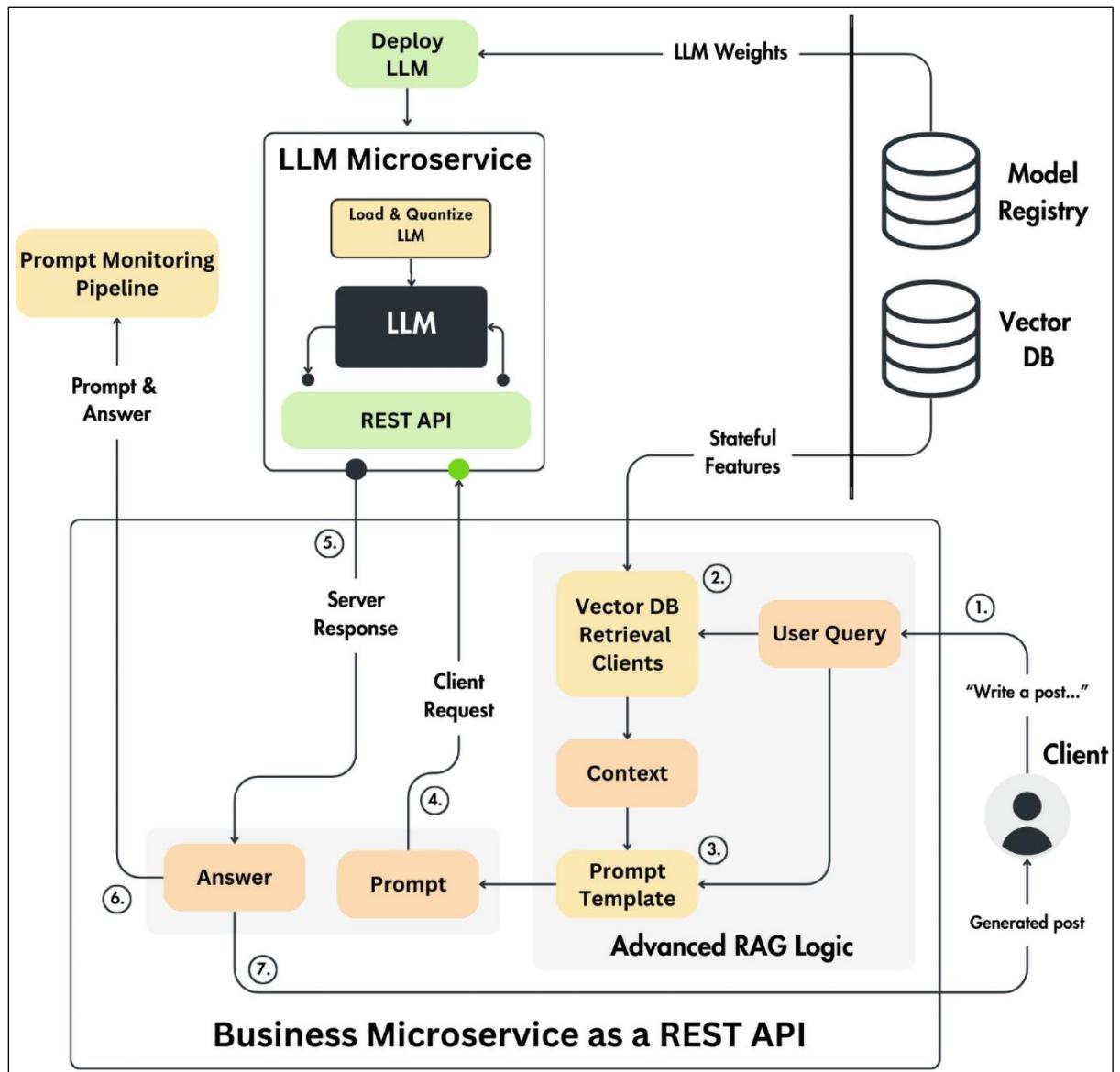
Monolith

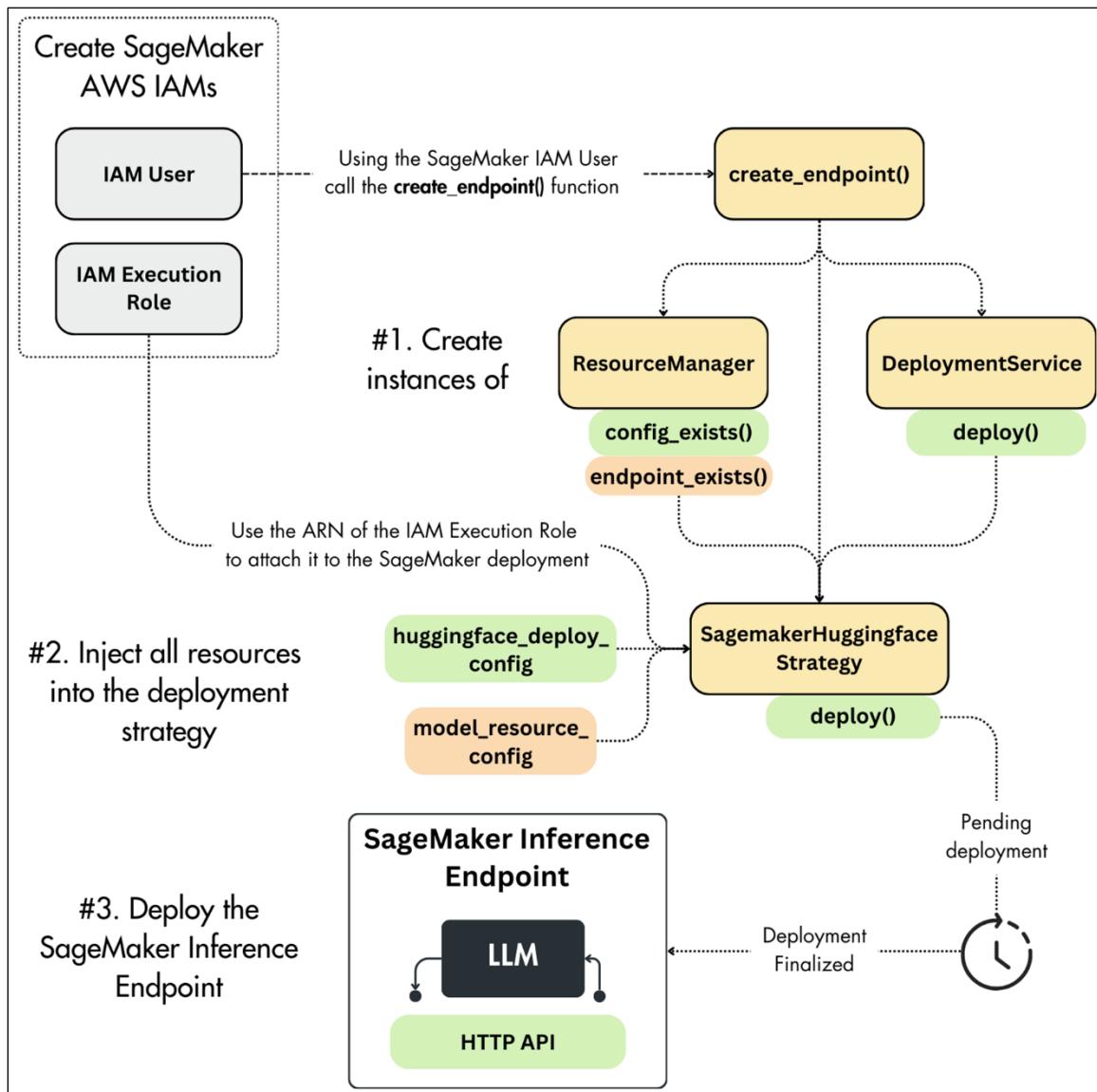


Microservices









Amazon SageMaker > Dashboard

Dashboard

Recent activity

Recent activity within the

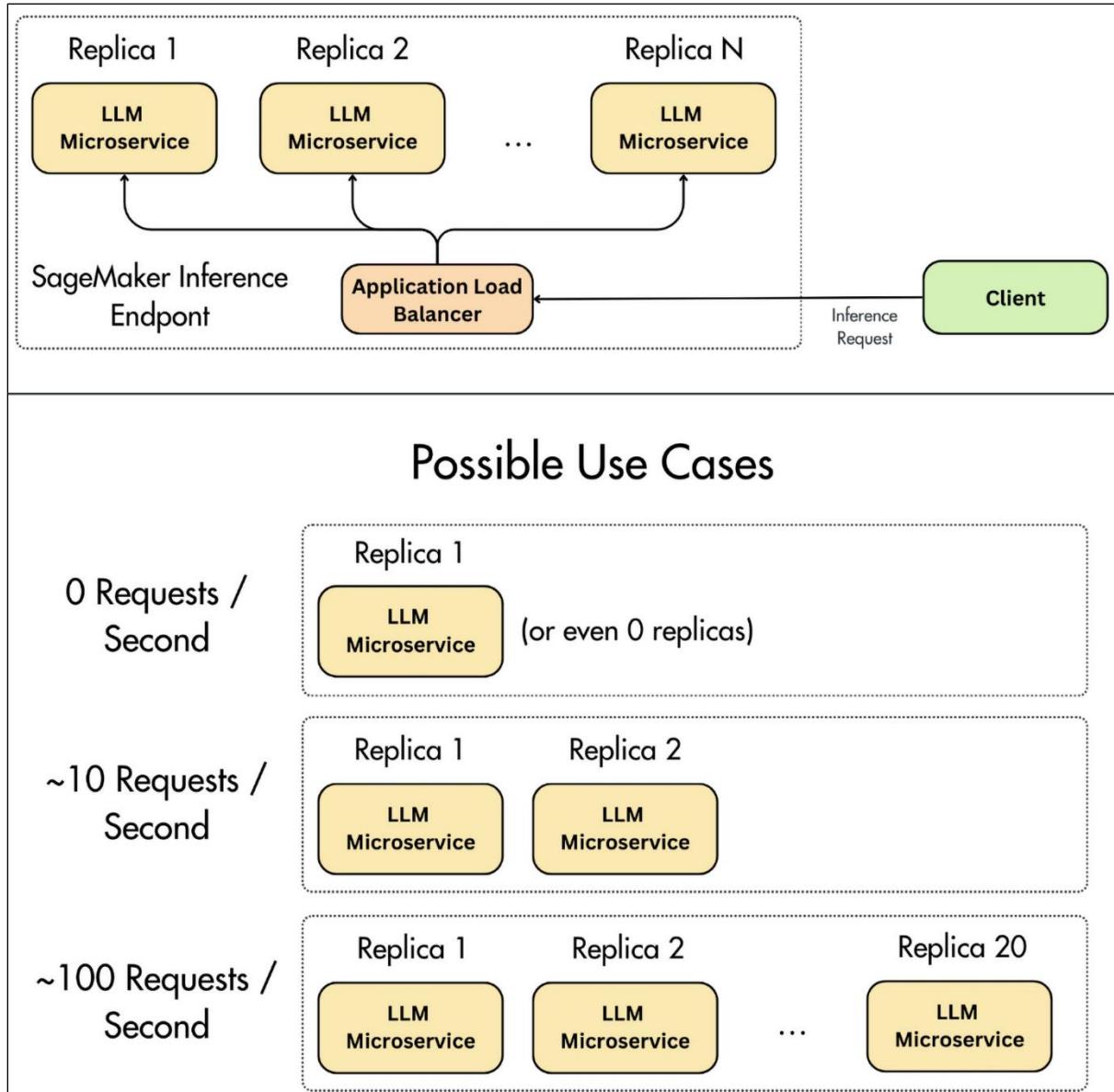
Ground Truth	Notebook	Training	Inference	Processing	Canvas
Labeling jobs	Notebook instances	Training jobs	Models	Processing jobs	Endpoints
No recent activity.	No recent activity.	No recent activity.	<input checked="" type="radio"/> 1 Created	<input checked="" type="radio"/> 28 Completed	No recent activity.
Hyperparameter tuning jobs			Endpoints		
No recent activity.			<input checked="" type="radio"/> 1 Created		
Batch transform jobs					
No recent activity.					

[Open SageMaker Domain](#)

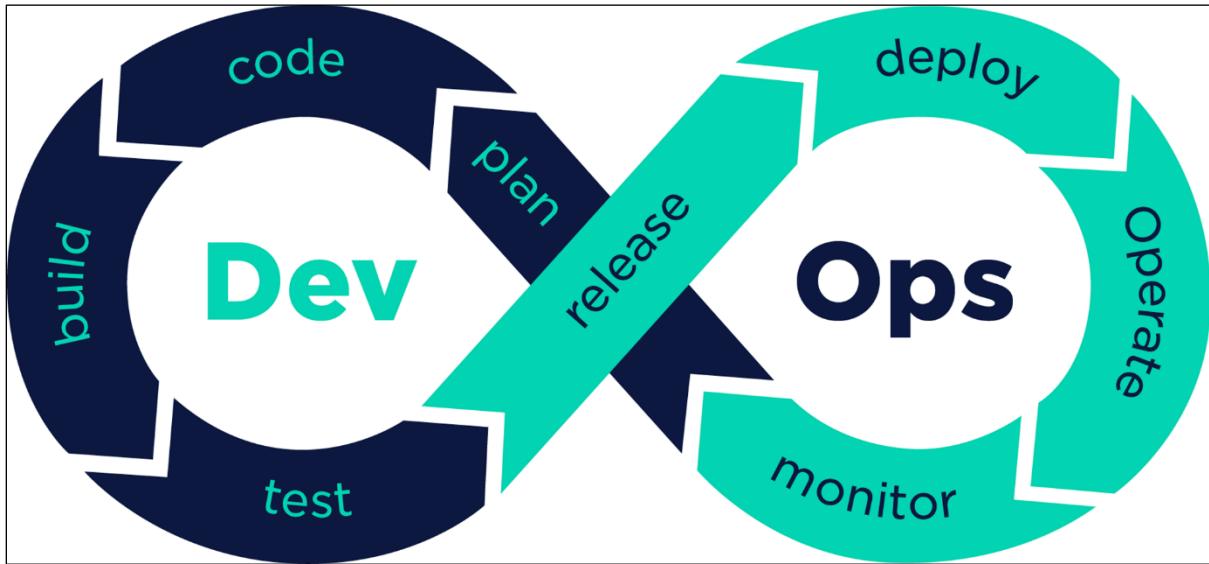
Amazon SageMaker > Endpoints

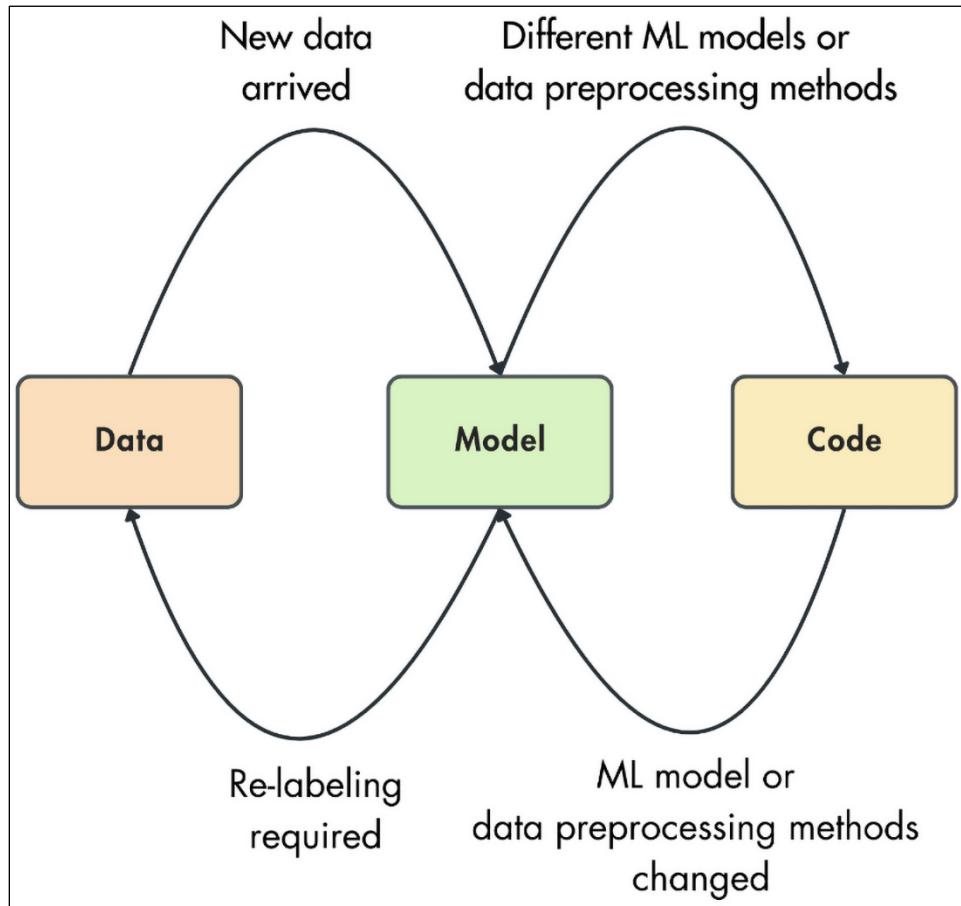
Endpoints

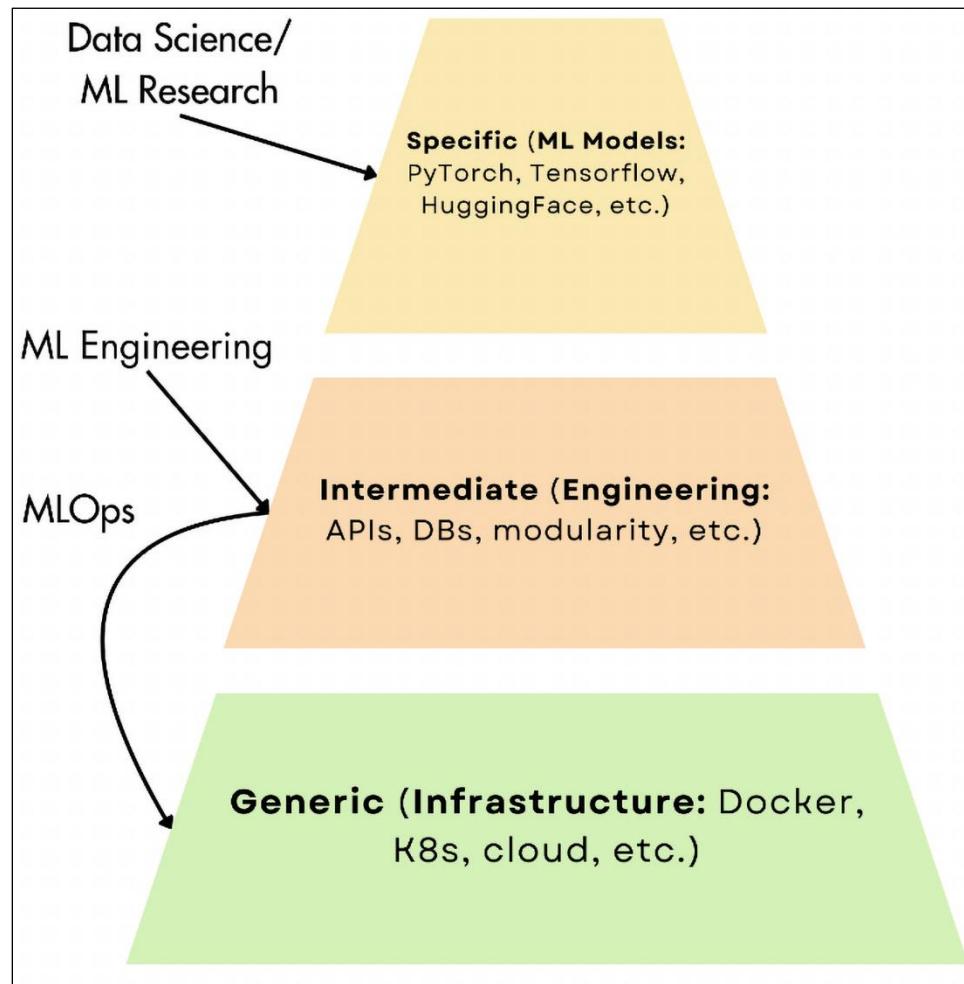
Name	ARN	Creation time	Status	Last updated
twin	arn:aws:sagemaker:eu-central-1:992382797823:endpoint/twin	8/17/2024, 8:38:34 PM	<input checked="" type="radio"/> Creating	8/17/2024, 8:38:34 PM

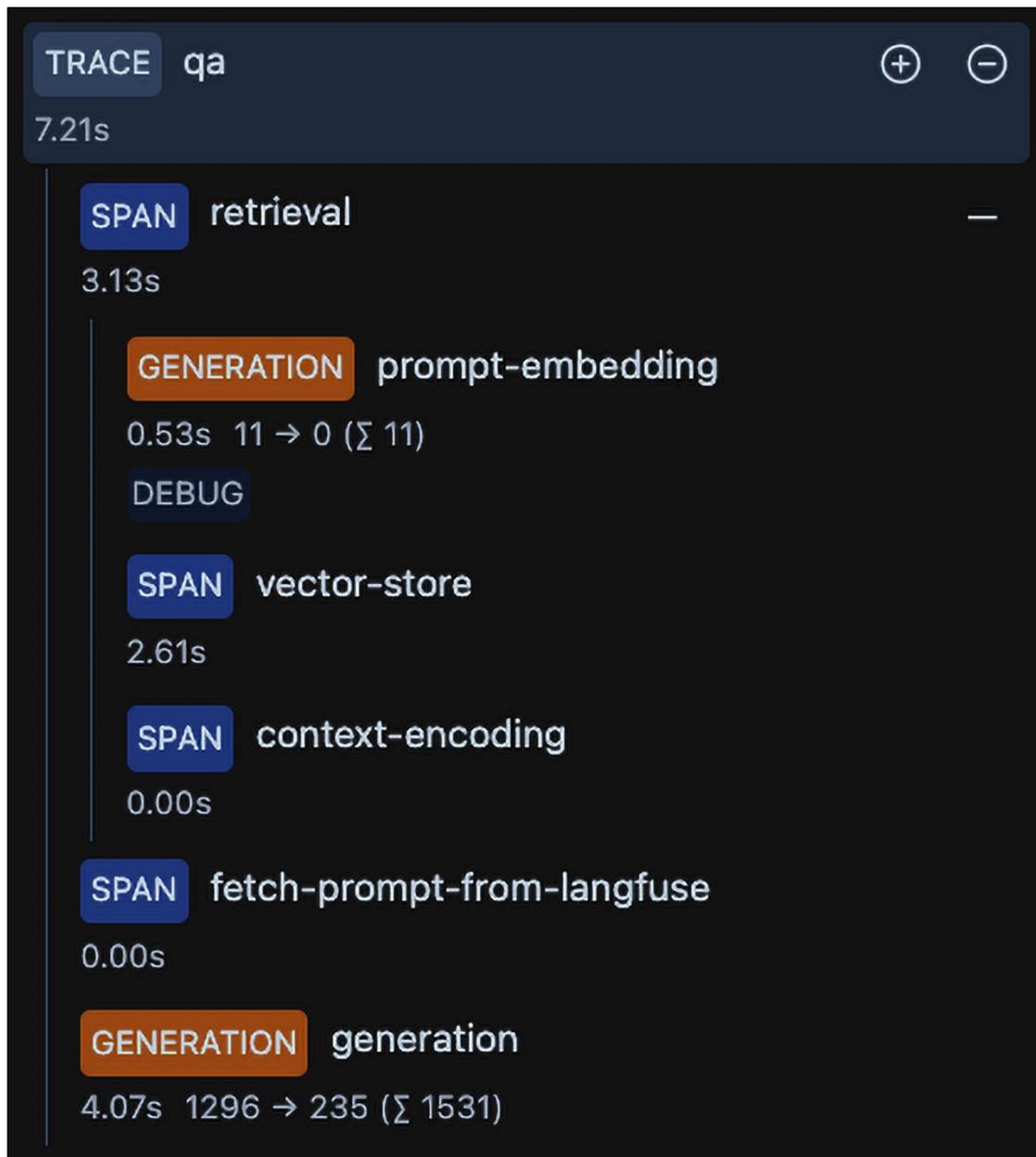


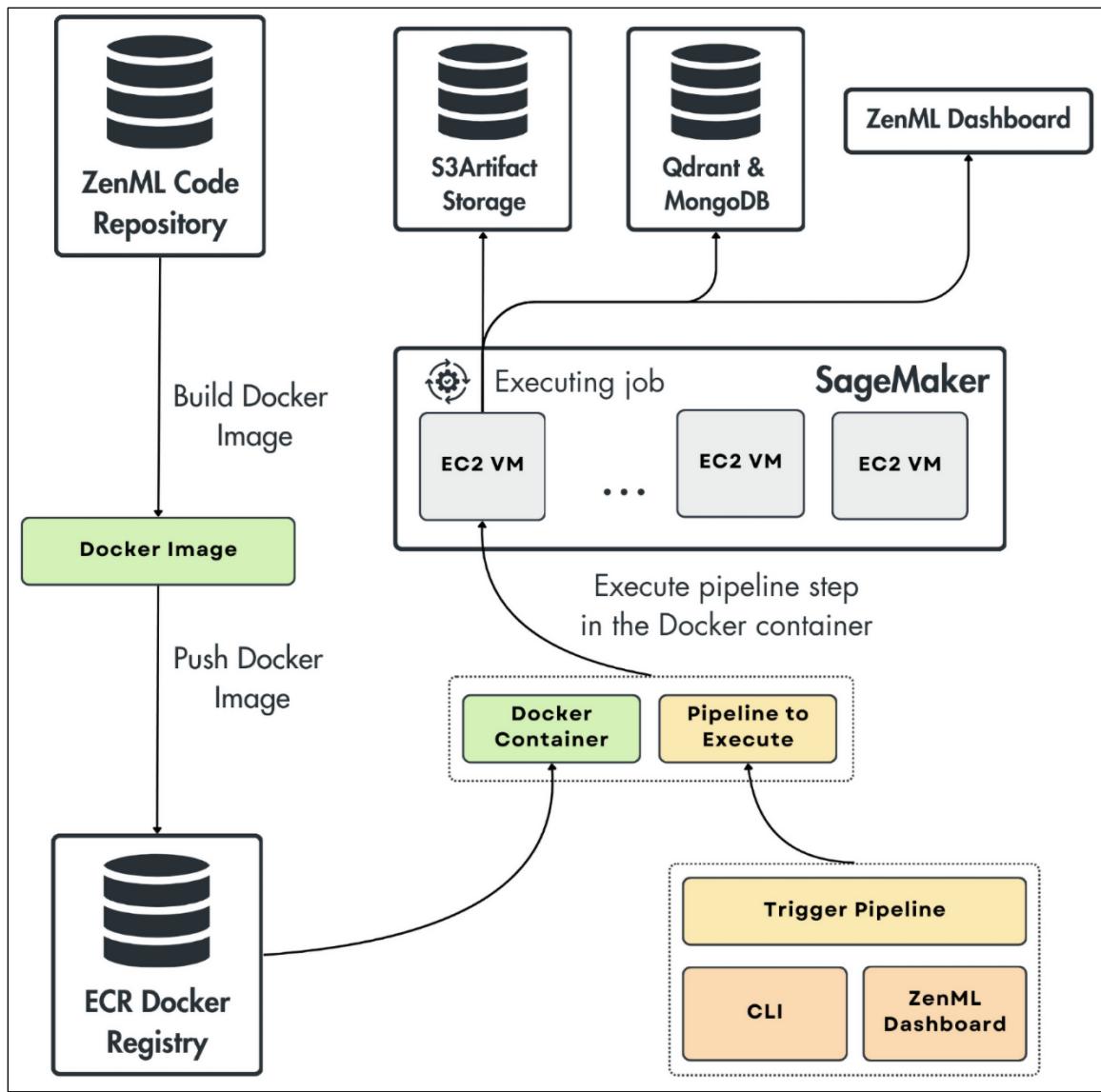
Chapter 11: MLOps and LLMOps

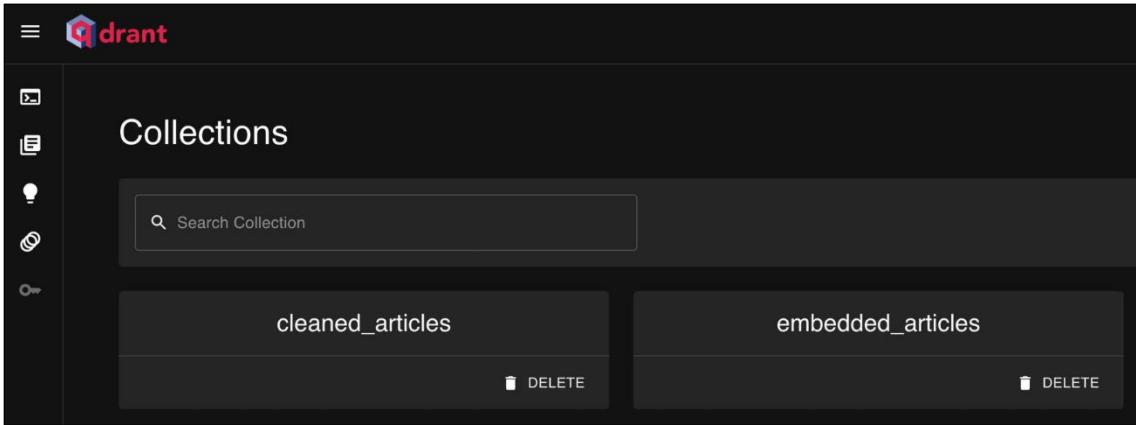




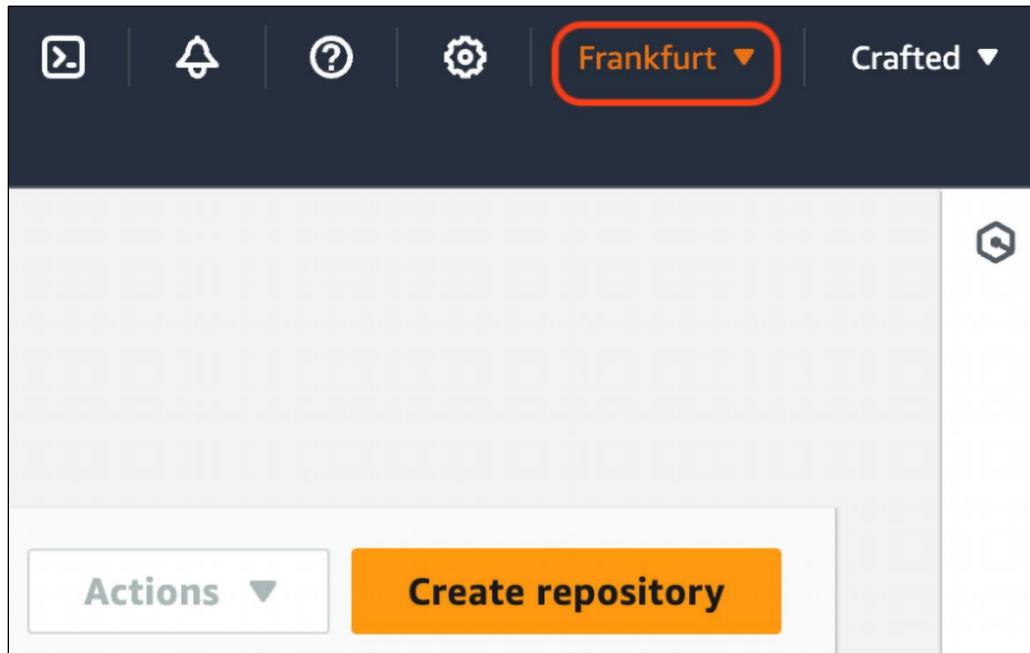








Private repositories						
Repositories (2)		Actions		Create repository		
Repository name	URI	Created at	Tag immutability	Scan frequency	Encryption type	
crawler	992382797823.dkr.ecr.eu-central-1.amazonaws.com/crawler	March 21, 2024, 15:58:08 (UTC+02)	Disabled	Manual	AES-256	
zenml-rwlcs	992382797823.dkr.ecr.eu-central-1.amazonaws.com/zenml-rwlcs	August 10, 2024, 14:52:34 (UTC+03)	Disabled	Manual	AES-256	

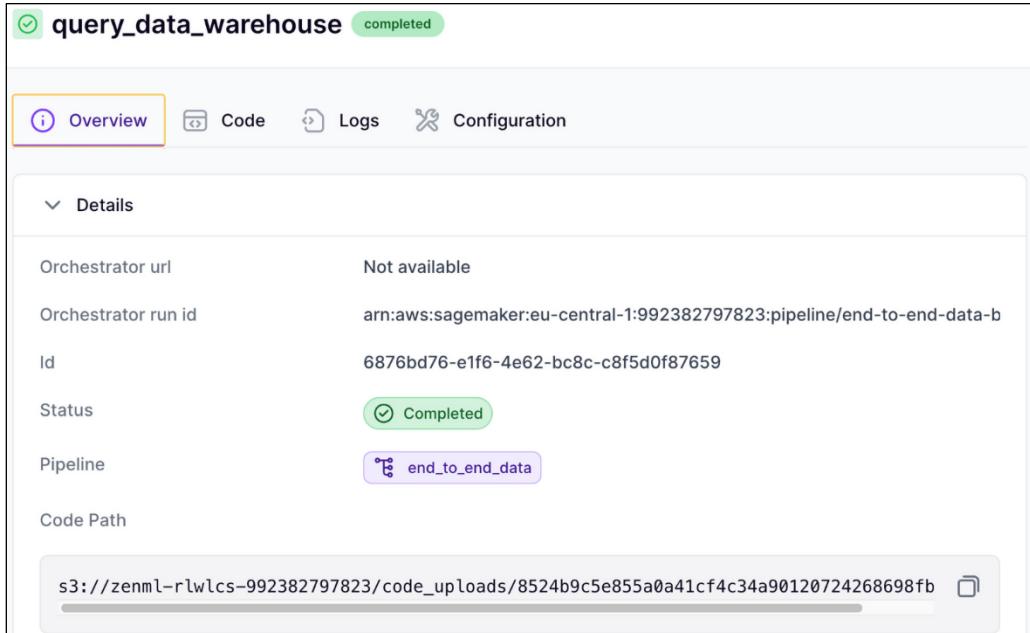
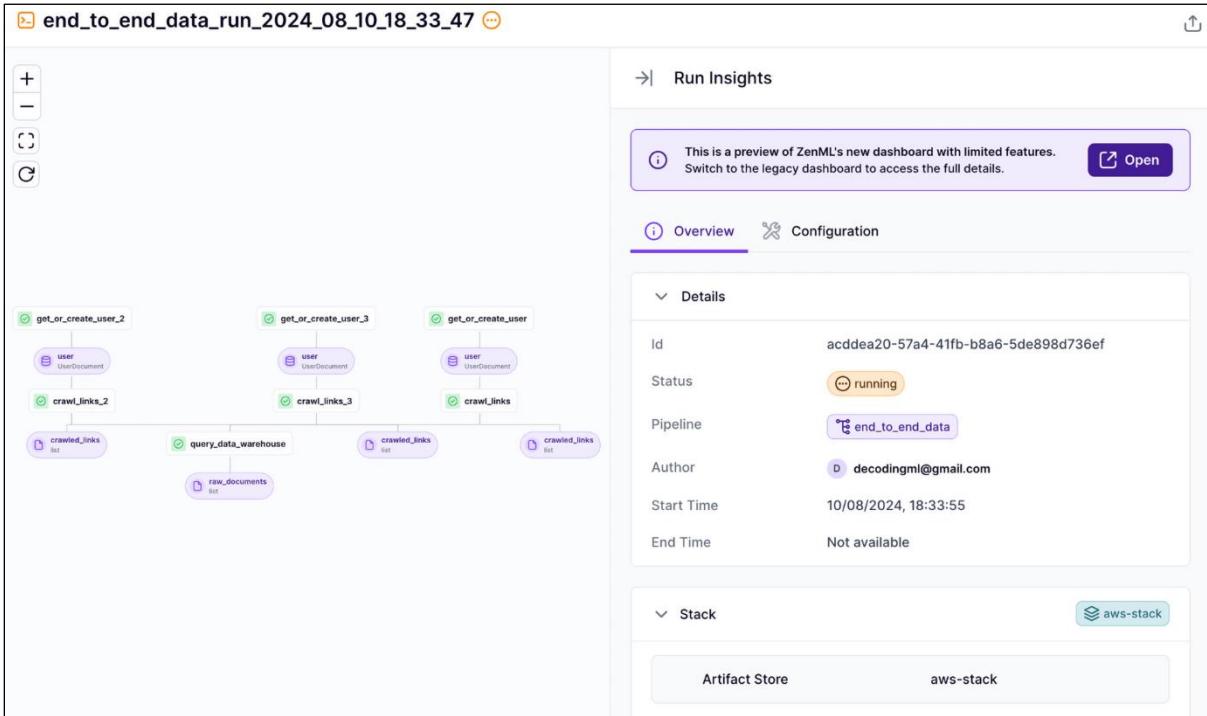


Amazon ECR > Private registry > Repositories > zenml-rlwlcs

zenml-rlwlcs

Images (1)

<input type="checkbox"/>	Image tag	Artifact type	Pushed at	Size (MB)	Image URI
<input type="checkbox"/>	latest	Image	August 10, 2024, 18:24:58 (UTC+03)	3494.10	<input type="button" value="Copy URI"/>



Amazon SageMaker

Getting started

Applications and IDEs

- Studio
- Canvas
- RStudio
- TensorBoard
- Profiler
- Notebooks

Admin configurations

- Domains
- Role manager
- Images
- Lifecycle configurations

SageMaker dashboard

Amazon SageMaker > Dashboard

AWS Marketplace

Find, buy, and deploy ready to use model packages, algorithms, and data products in AWS Marketplace

Dashboard

Recent activity

	Recent activity
Ground Truth	No recent activity.
Labeling jobs	No recent activity.
Notebook	No recent activity.
Notebook instances	No recent activity.
Training	No recent activity.
Training jobs	No recent activity.
Inference	No recent activity.
Models	No recent activity.
Processing	No recent activity.
Endpoints	No recent activity.
Hyperparameter tuning jobs	No recent activity.
Batch transform jobs	No recent activity.
Processing jobs	No recent activity.

Recent activity

	Recent activity
2 Failed	8 Created
1 Completed	9 Failed
5 Created	4 Running
73 Completed	73 Completed
86 Created	86 Created

Service Quotas > AWS services > Amazon SageMaker

Amazon SageMaker

Amazon SageMaker is a fully-managed service that enables data scientists and developers to quickly and easily build, train, and deploy machine learning models at scale.

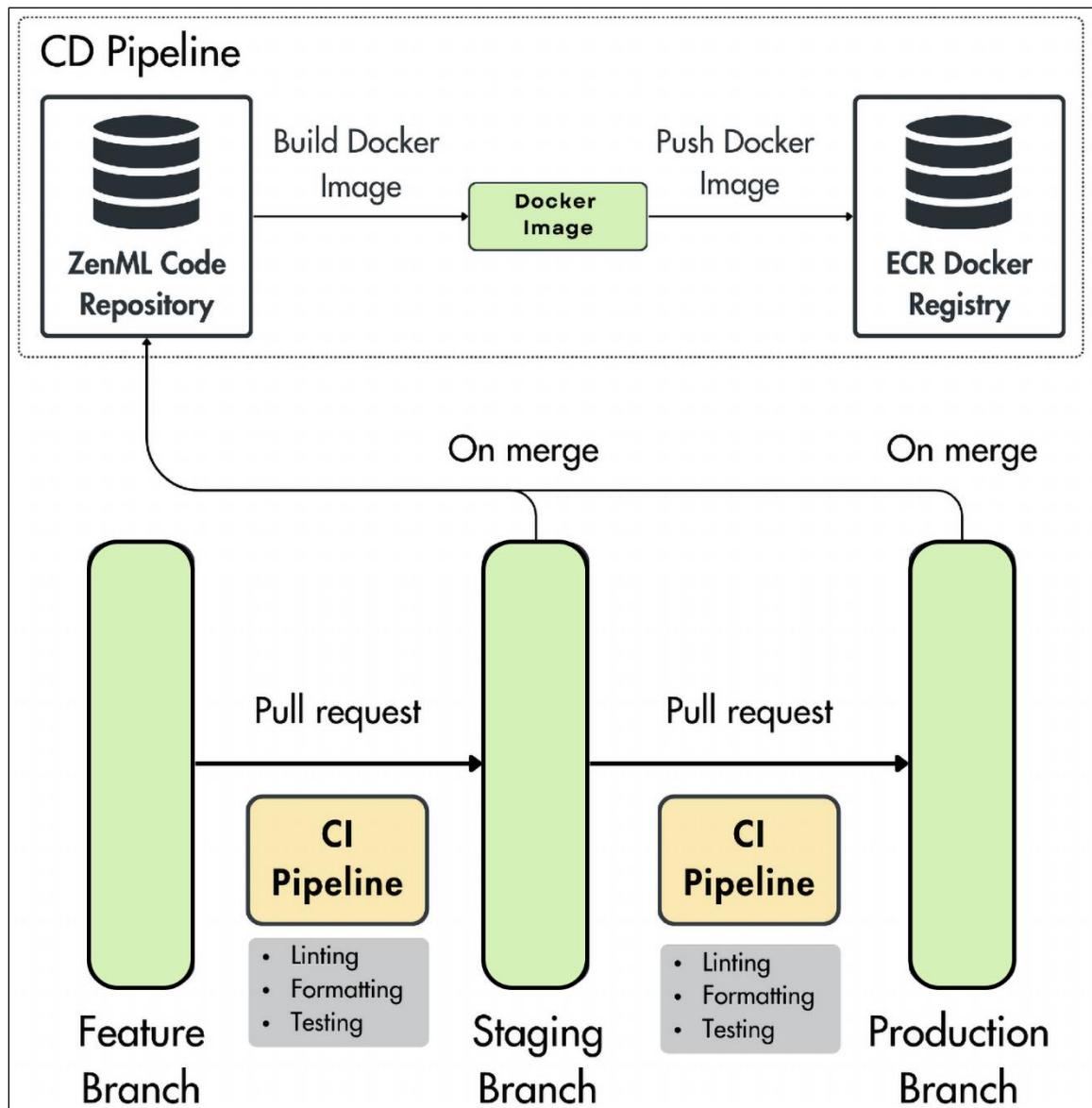
Service quotas info

View your applied quota values, default quota values, and request quota increases for quotas. [Learn more](#)

Request increase at account level

Q. ml.t3.medium X 7 matches

Quota name	Applied account-level quota value	AWS default quota value	Utilization	Adjustability
ml.t3.medium for cluster usage	10	0	0	Account level
ml.t3.medium for notebook instance usage	30	6	0	Account level
ml.t3.medium for processing job usage	10	4	1	Account level
RSessionGateway Apps running on ml.t3.medium instance	30	2	0	Account level
Studio CodeEditor Apps running on ml.t3.medium instances	6	6	0	Account level
Studio JupyterLab Apps running on ml.t3.medium instances	6	6	0	Account level
Studio KernelGateway Apps running on ml.t3.medium instance	30	2	0	Account level



PacktPublishing / LLM-Engineering

Code Issues Pull requests ① Actions Wiki Security ① Insights Settings

← CI

feat: Add Docker image and CD pipeline #49

Summary

Jobs

QA Test

Run details

Usage Workflow file

Triggered via pull request 3 days ago Status Success Total duration 2m 15s

iustzinpaul synchronize #4 zenml-deploy

ci.yaml on: pull_request

QA 18s

Test 2m 7s

PacktPublishing / LLM-Engineering

Code Issues Pull requests Actions Wiki Security ① Insights Settings

← CD

Merge pull request #4 from PacktPublishing/zenml-deploy #9

Summary

Jobs

Build & Push Docker Image

Run details

Usage Workflow file

Re-run triggered 21 minutes ago Status Success Total duration 7m 10s Artifacts 1

iustzinpaul → 4415577 main

cd.yaml on: push

Build & Push Docker Im... 7m 0s

zenml-vrsopg 992382797823.dkr.ecr.eu-central-1.amazonaws.com/zenml-vrsopg August 13, 2024, 09:42:20 (UTC+03)

General

Access

Collaborators and teams

Moderation options

Code and automation

- Branches
- Tags
- Rules
- Actions
- Webhooks
- Pages
- Custom properties

Security

Code security and analysis

Deploy keys

Secrets and variables

- Actions
- Codespaces
- Dependabot

Actions secrets and variables

Secrets and variables allow you to manage reusable configuration data. Secrets are **encrypted** and are used for sensitive data. [Learn more about encrypted secrets](#). Variables are shown as plain text and are used for **non-sensitive** data. [Learn more about variables](#).

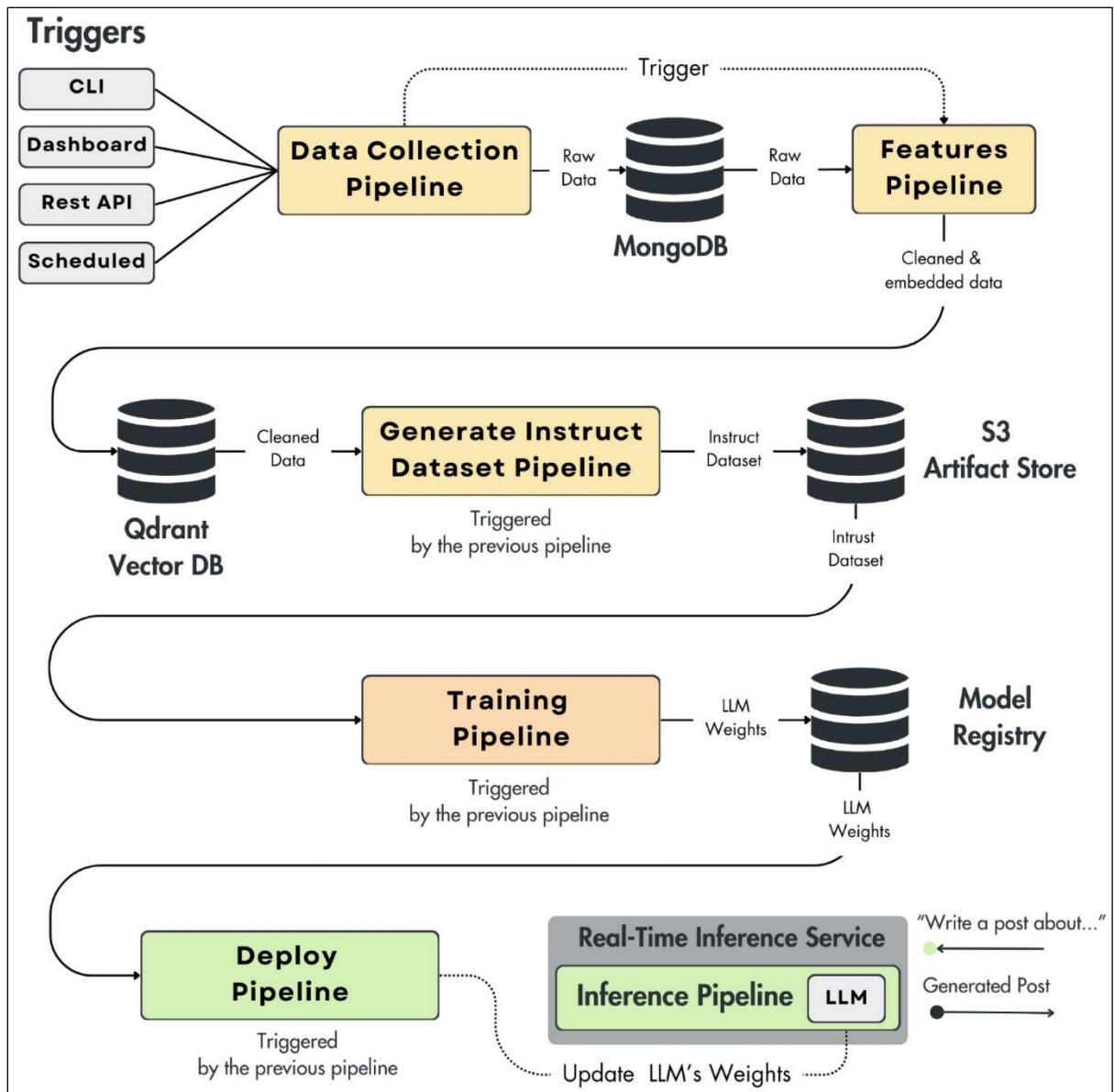
Anyone with collaborator access to this repository can use these secrets and variables for actions. They are not passed to workflows that are triggered by a pull request from a fork.

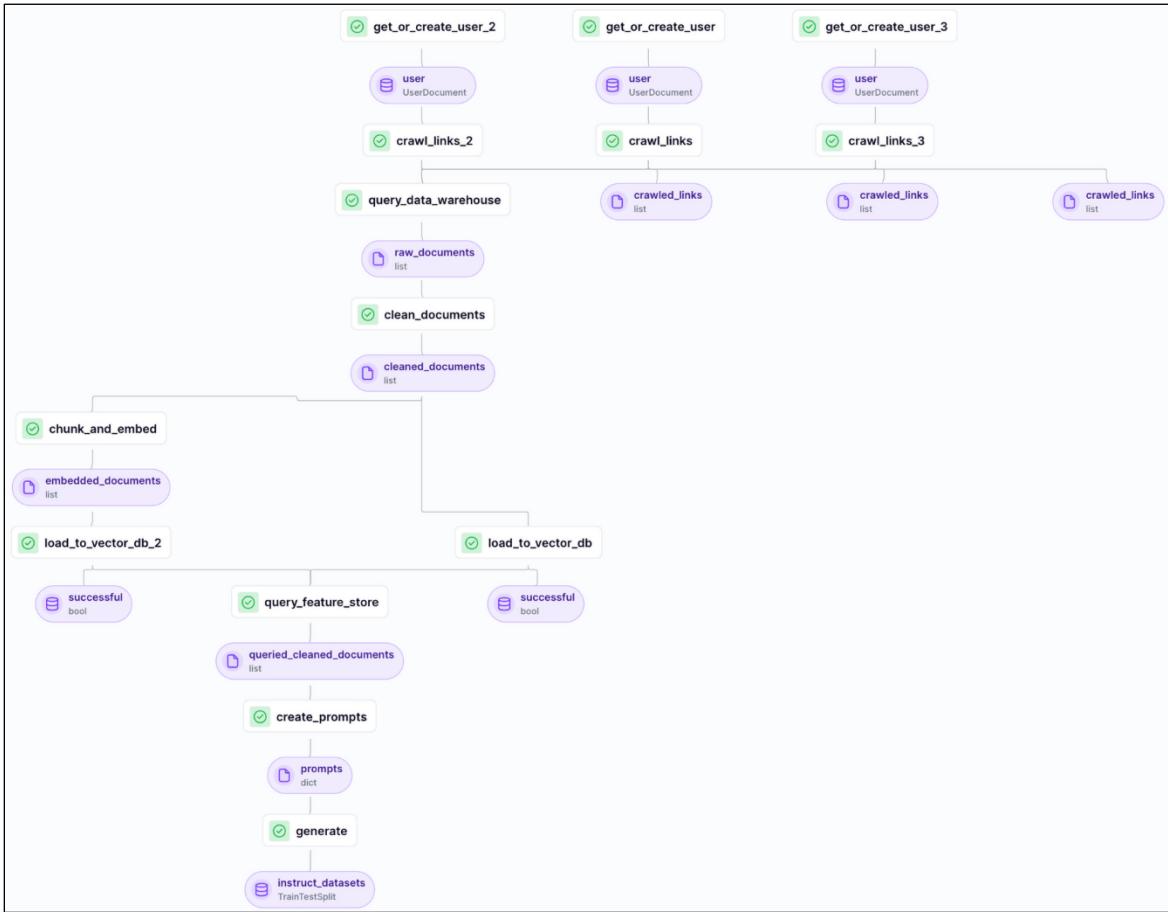
Secrets Variables

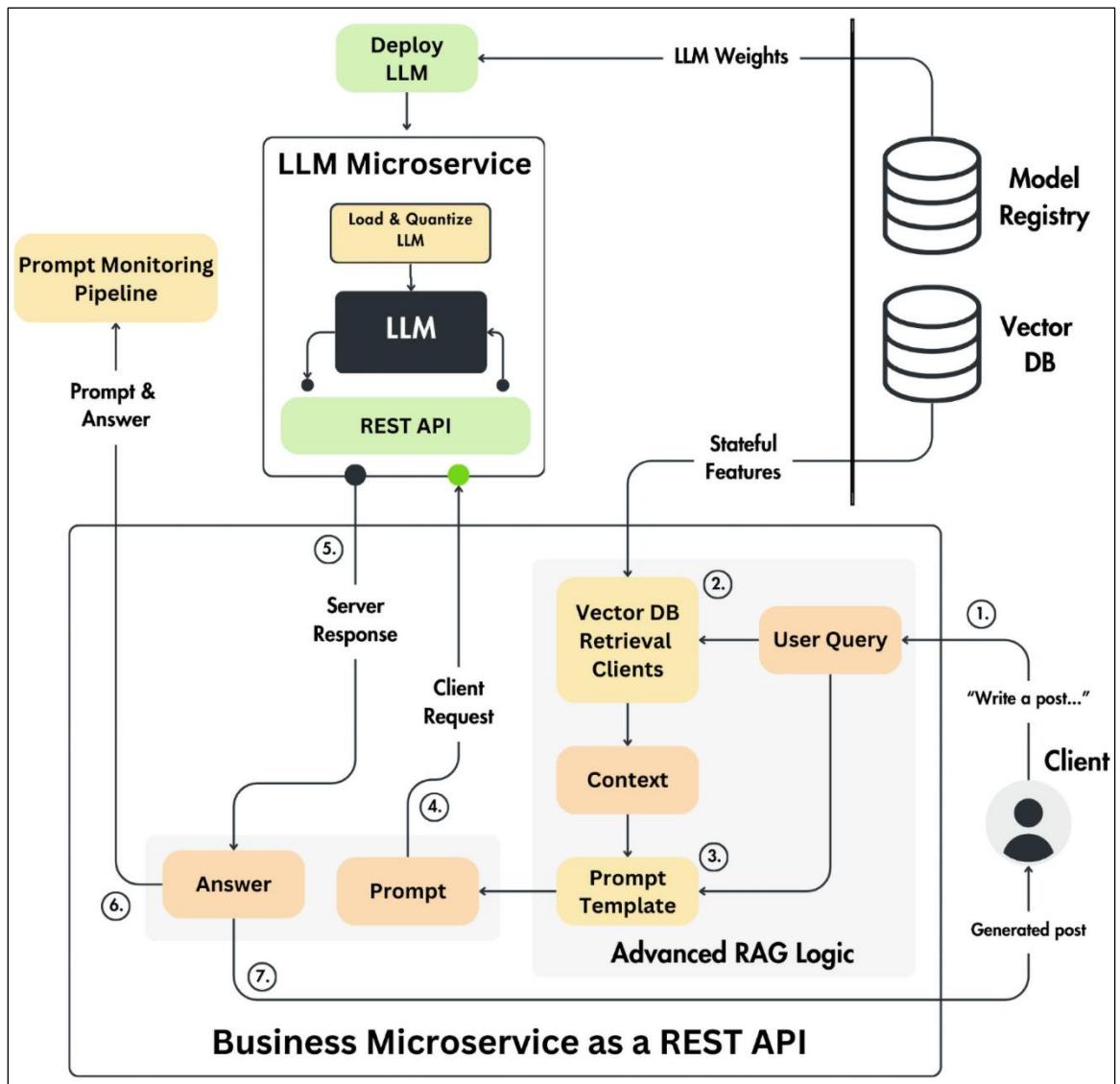
Repository secrets

New repository secret

Name	Last updated	Actions
AWS_ACCESS_KEY_ID	5 days ago	
AWS_ECR_NAME	13 minutes ago	
AWS_REGION	5 days ago	
AWS_SECRET_ACCESS_KEY	5 days ago	







Appendix: MLOps Principles

