An end-to-end framework for production-ready LLM systems by building your LLM twin

Data Collection Pipeline

aws

Digital Data

- Medium → Medium ETL
- Substack → Substack ETL
- LinkedIn → LinkedIn ETL
- GitHub → GitHub ETL

NoSQL DB (mongoDB)

1. CDC

2. Queue

Feature Pipeline

aws

Streaming Ingestion Pipeline

- Articles
- Posts
- Code

3.

bytewax

Superlinked

4.

Metadata & Text

Training Pipeline

Qwak

Experiment Tracker (comet)

LLM Fine-tuning

Data to Prompt Layer

Metadata & Text

Qdrant Retrieval Clients

5.

6. LLM Canditate

7. Model Registry (comet)

Evaluate LLM Candidate

Accepted LLM

Vector DB Retrieval Clients

- Articles
- Posts
- Code

Vector DB (qdrant)

Use "Vector DB Retrieval Clients"

Inference Pipeline

Load & Quantize LLM 8.

Twin LLM

Query to Prompt Layer

Metadata & Text

Qdrant Retrieval Clients

REST API

9. "Write a post about..."

Generated Post

10.

Prompt Monitoring (comet)

prompts & answers

Qwak

Data Collection Pipeline

Digital Data Sources

- Medium
- Substack
- LinkedIn
- GitHub

Medium & Substack ETL (Articles)

Login User → Crawl Article → Parse HTML → Clean & Normalize → Load to NoSQL DB → Next?

Yes → Crawl Article

LinkedIn ETL (Posts)

Login User → Clone Repository → Parse Repository File Tree → Clean & Normalize → Load to NoSQL DB → Next?

Yes → Clone Repository

GitHub ETL (Code)

Login User → Crawl Feed → Parse HTML → Clean & Normalize → Load to NoSQL DB

NoSQL DB

mongoDB

CDC

Added

Updated

Deleted

RabbitMQ Queue

Feature Pipeline

aws

Streaming Ingestion Pipeline

- Articles
- Posts
- Code

bytewax

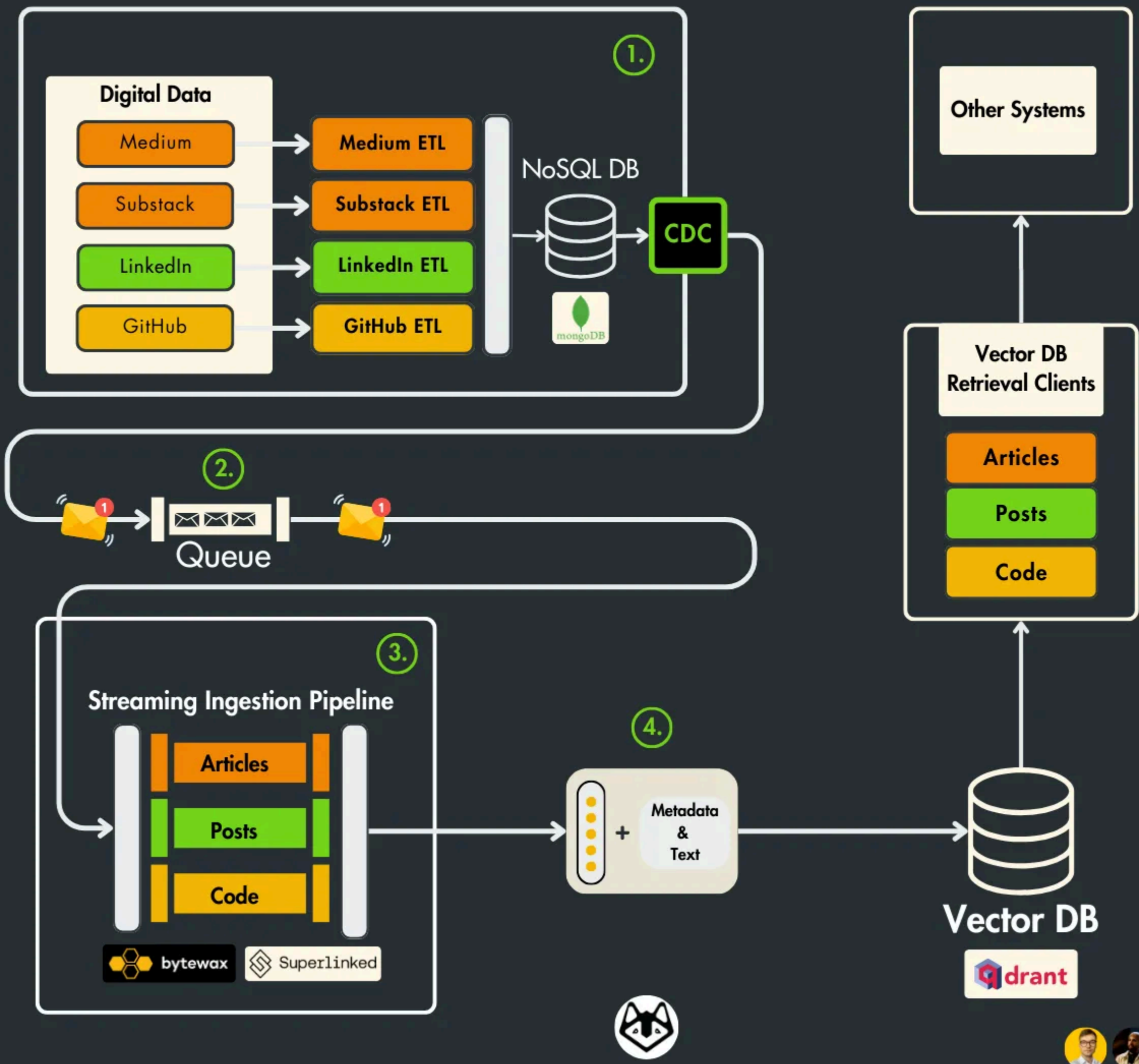
Superlinked

Metadata & Text

Vector DB

qdrant

An end-to-end framework for production-ready LLM systems by building your LLM twin



SOTA Python Streaming Pipelines for Fine-tuning LLMs and RAG — in Real-Time!

NoSQL DB

1.

CDC

mongoDB

Event types:
→ Article
→ Post
→ Code

Other Components

Vector DB Retrieval Clients

Articles

Posts

Code

6.

2.

RabbitMQ Queue

Metadata & Text

Vector DB

qdrant

Feature Store

Metadata & Text

3.

4.

5.

Raw Dispatcher

Cleaning Dispatcher

Qdrant Output

Chunking Dispatcher

Embedding Dispatcher

Qdrant Output

Article Model

map

Cleaned Article Model

map

Chunked Article Model

map

Embedded Article Model

Post Model

map

Cleaned Post Model

map

Chunked Post Model

map

Embedded Post Model

Code Model

map

Cleaned Code Model

map

Chunked Code Model

map

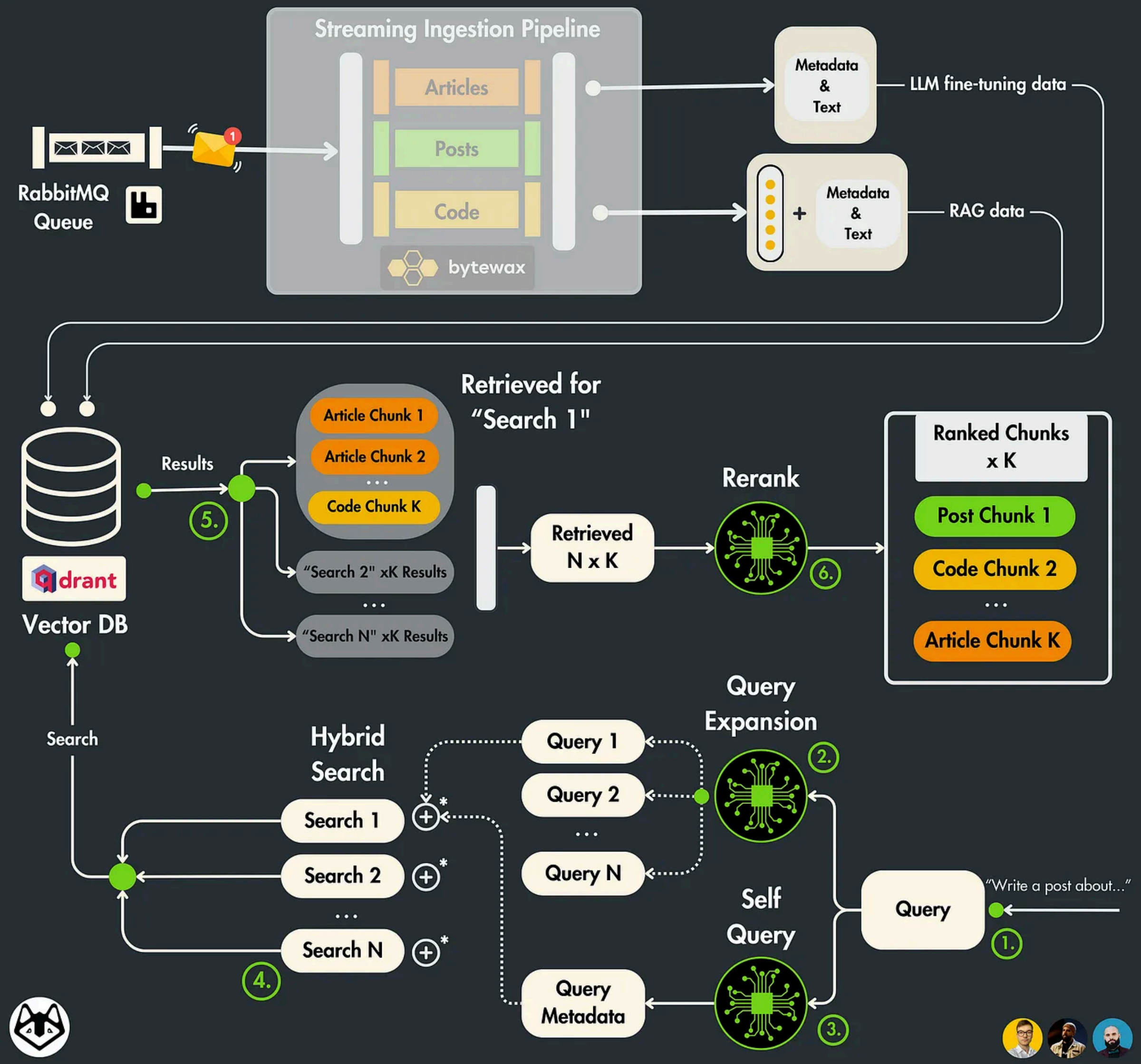
Embedded Code Model

bytewax

bytewax

bytewax

Advanced Retrieval-Augmented Generation (RAG): Retrieval Python Module Architecture



Custom LLM Finetuning Dataset Generation using Knowledge Distillation

Vector DB

id vector metadata
id vector metadata
id vector metadata
id vector metadata

qdrant

query
query

Dataset Generation

Articles
Posts
Code

Raw Data Formatting

template template template

Python

Samples
Samples
Samples

ingest

GPT3.5-Turbo

enhance

Instruction Dataset

Dataset Versioning

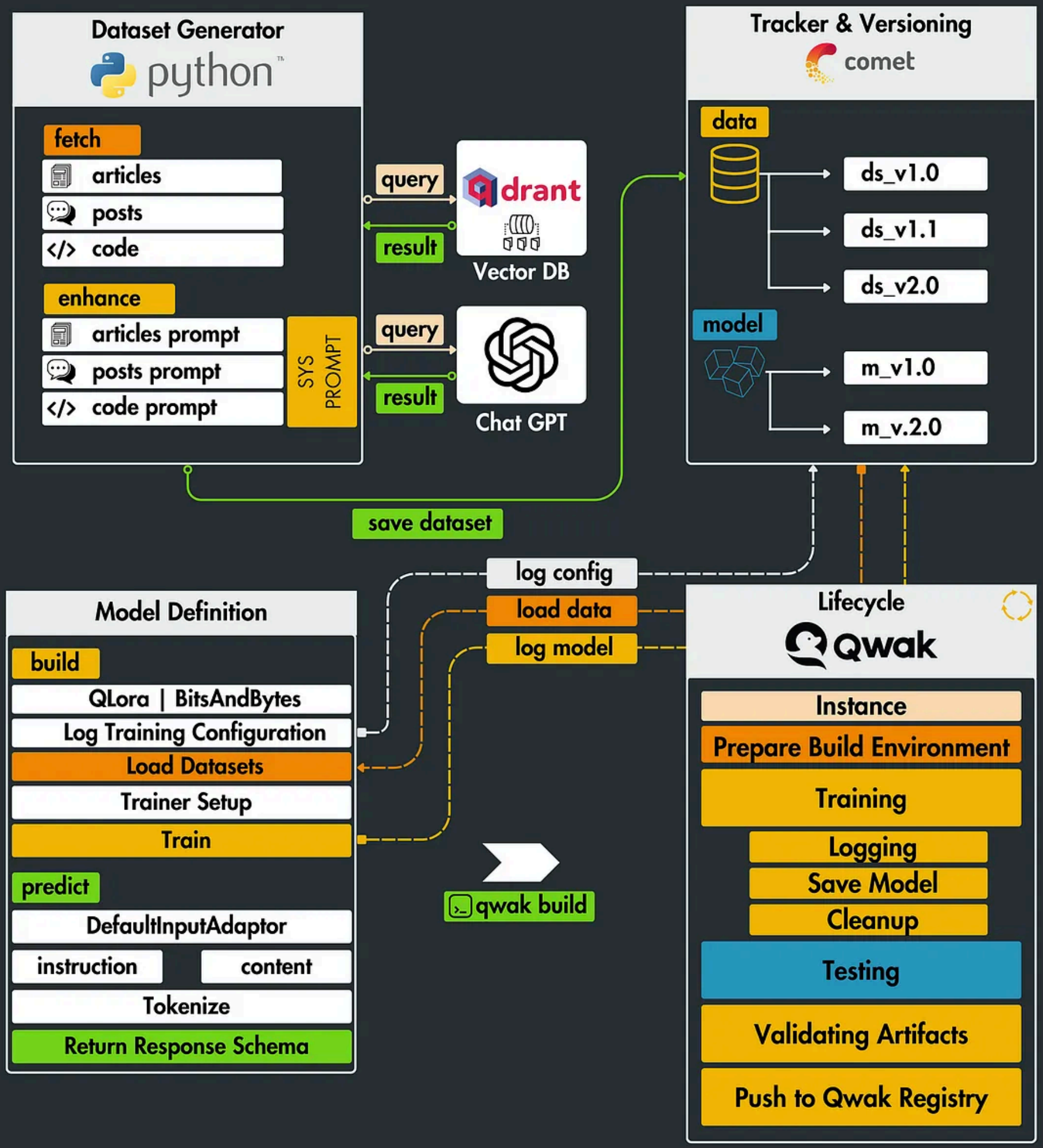
cleaned_v1.0
cleaned_v1.2
cleaned_v2.0

comet

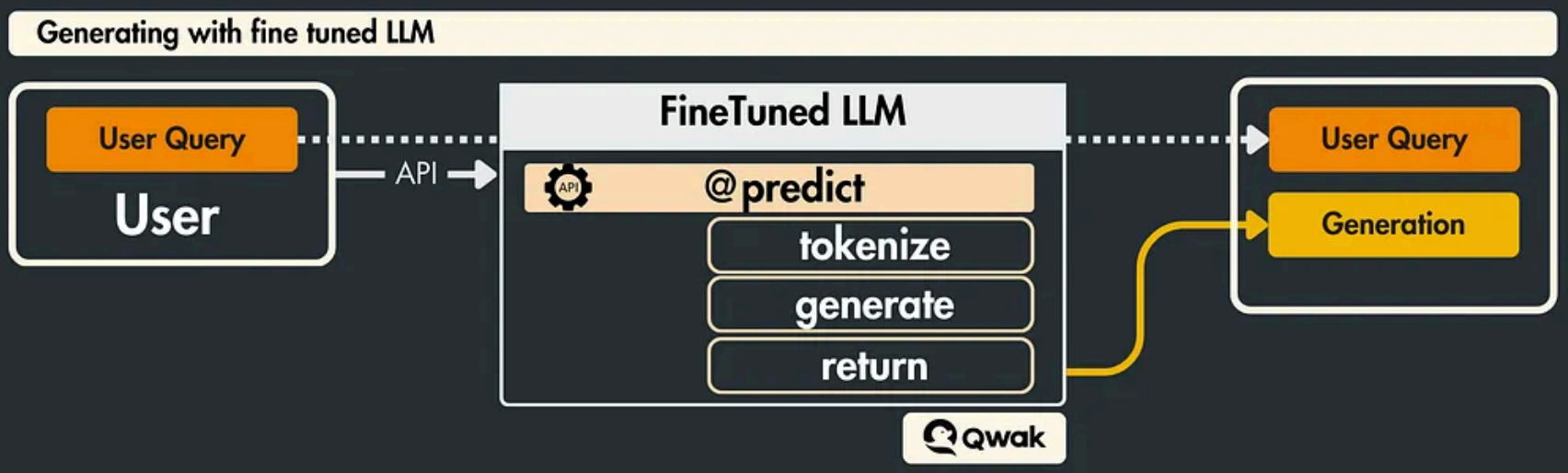
download Artifact

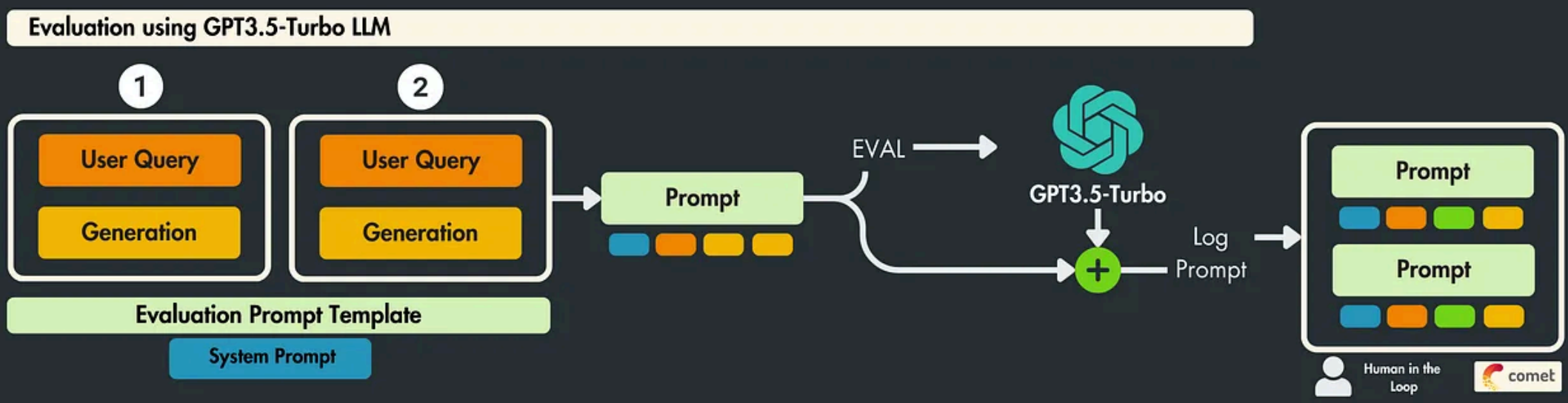
version Log Artifact

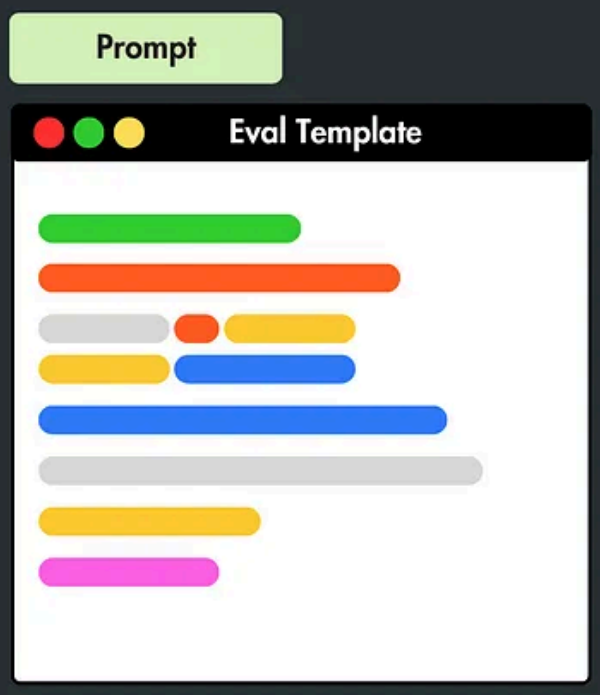
LLM Twim finetuning workflow on Qwak

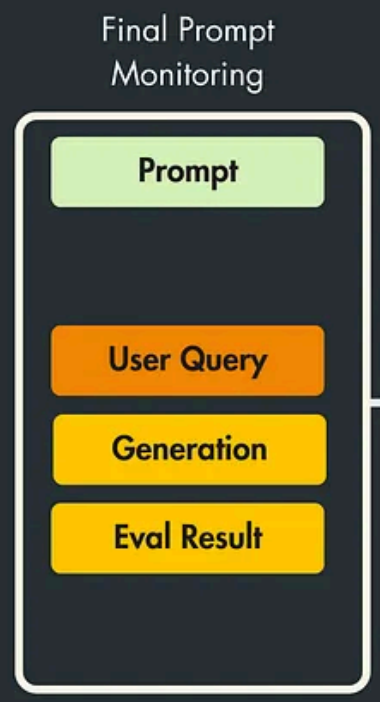


Evaluating the fine-tuned LLM Twin model

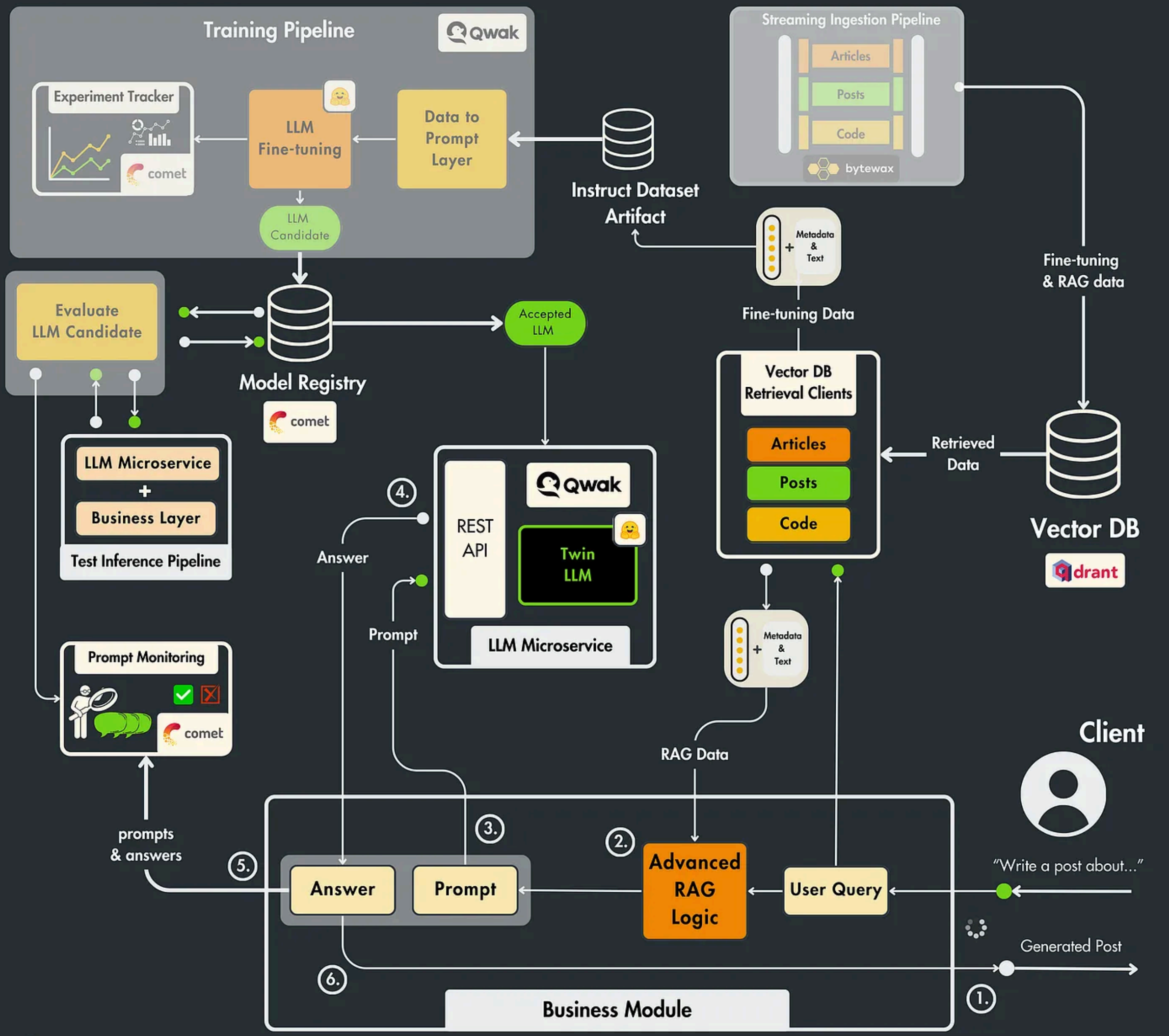




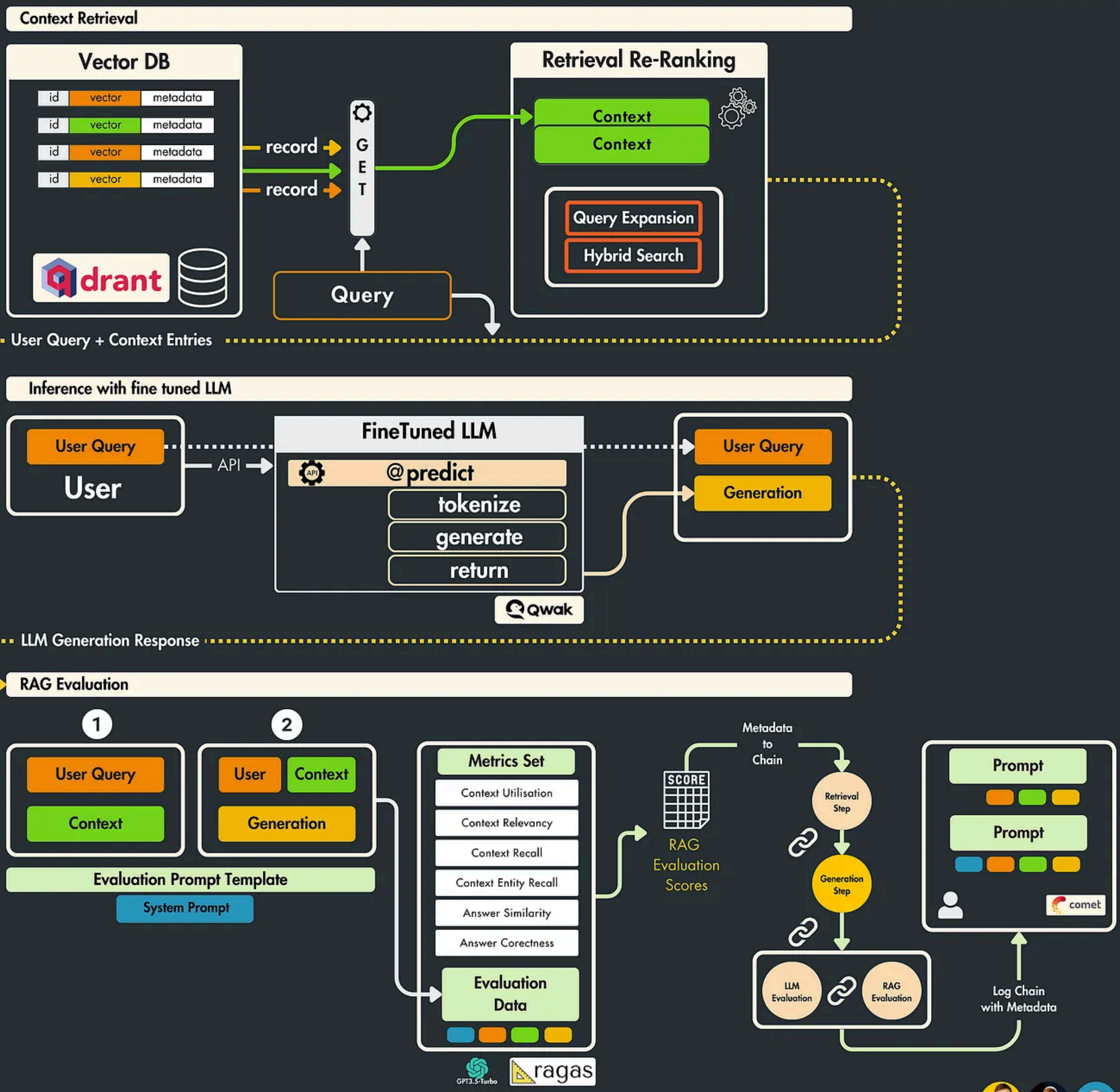




Architect scalable and cost-effective LLM & RAG inference pipelines



Evaluating the LLM-Twin RAG



Build a scalable RAG ingestion pipeline using 74.3% less code