Java Me Smarter

Unleashing AI Power with Quarkus

KubeCon | CloudNativeCon

THE LINUX FOUNDATION OPEN SOURCE SUMMIT | AI_dev Open Source GenAI & ML Summit

China 2024

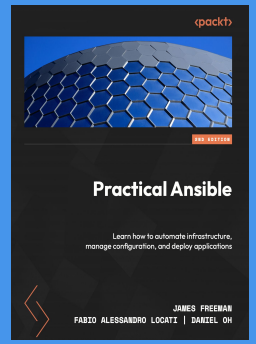
Daniel Oh

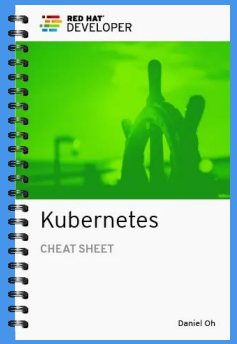
- Developer Advocate at Red Hat
  - Cloud Native Runtimes
  - AI, GitOps, and Serverless
- JAVA Champion
- Cloud Native Computing Foundation Ambassador
- App Development WG Co-Chair in CNCF TAG App Delivery
- Public Speaker & Published Author
- Living in Boston, MA USA

@danieloh30

@danieloh30

danieloh30

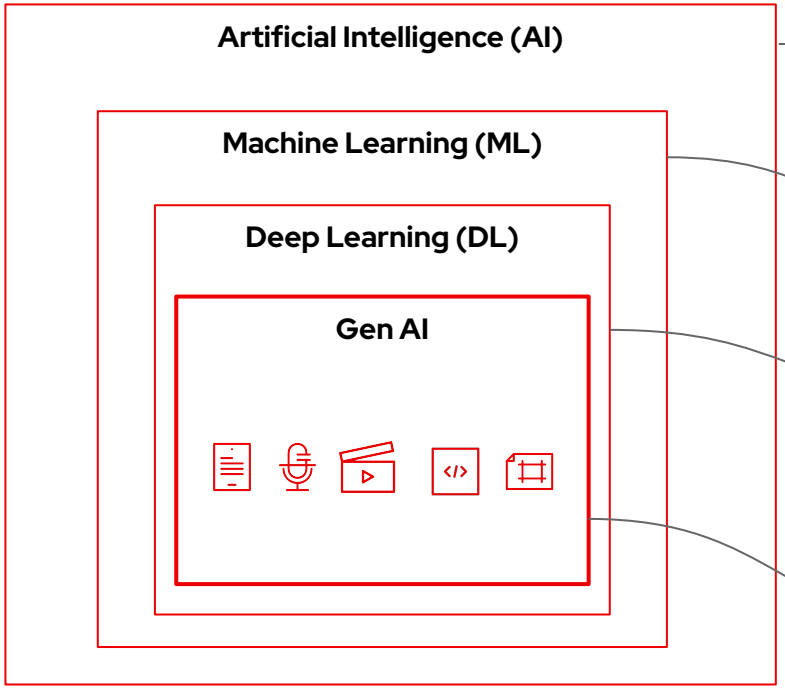




The Journey to Generative AI and LLMs

@danieloh30

What really is Generative AI?



Artificial Intelligence
A multidisciplinary field of Computer Science that aims to create systems capable of emulating and surpassing human-level intelligence

Machine Learning
A subfield for training computers to learn and make decisions from data, without explicit programming

Deep Learning
A subset from ML involved neural networks enabling complex pattern recognition and analysis from data

Gen AI Learning
Subfield focusing on generating new content such as text, images, code and more

@danieloh30

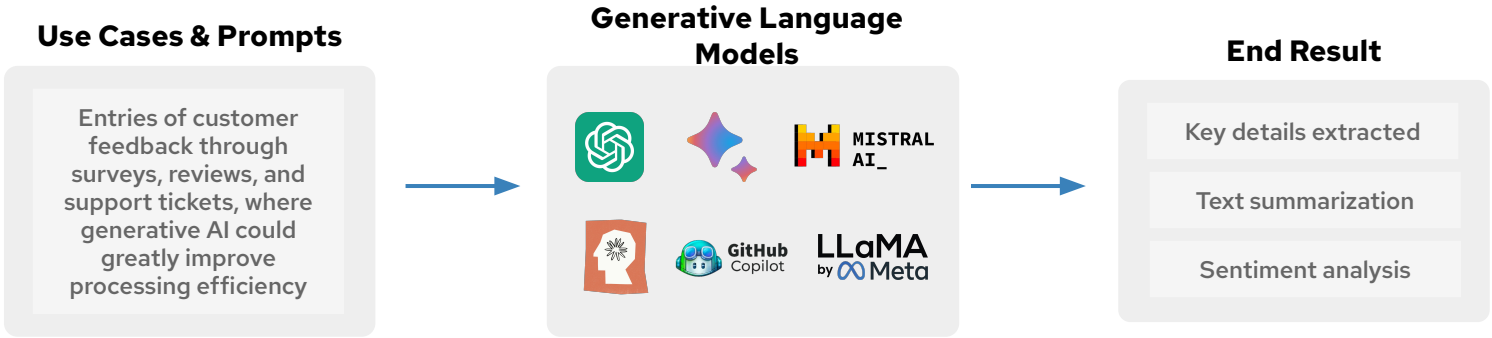
How does Generative AI really work?

Generative AI

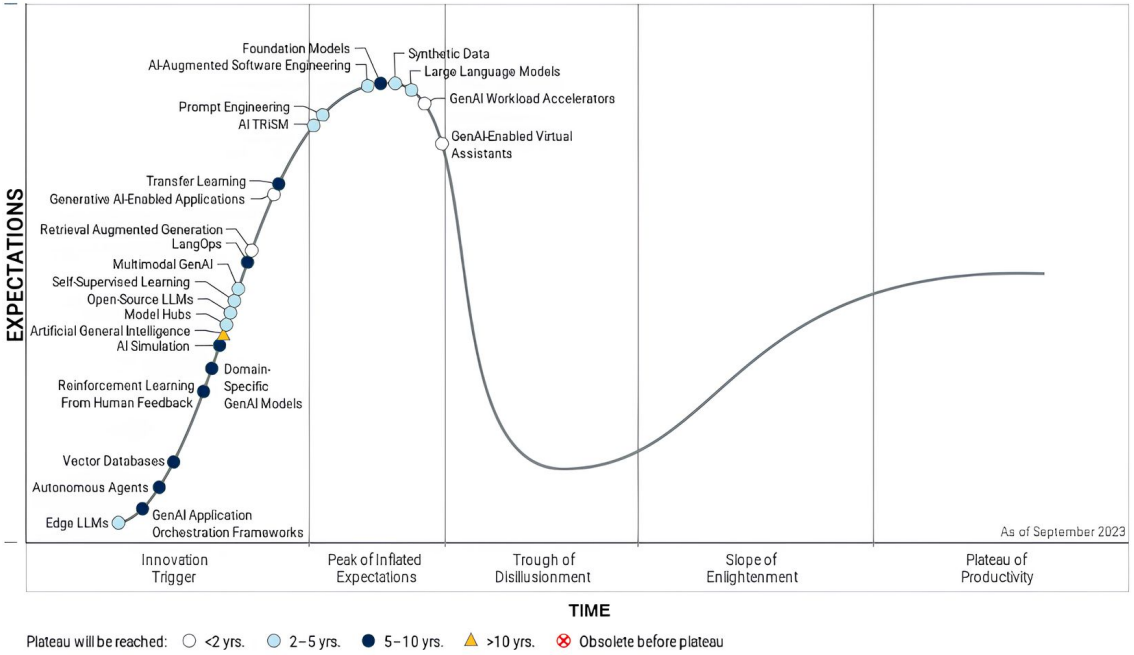
Generative AI utilizes AI models to create text, images, videos, code, or data in response to prompts

LLMs

Large Language Models are a specialized type of AI, trained on massive datasets to generate outputs in natural, conversational language



The AI adoption journey with the rise of Gen AI



Gartner

Gartner

> *More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026.*

https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026

The Gen AI Journey for an Application Developer

@danieloh30

The Journey of Adopting Generative AI

Ideation & Prototyping

- Find LLMs
- Try prompts
- Experiment with your data
- Benchmarking

As a developer, how do I **evaluate** models and pick the best one for my use case?

Building & Refining

- Connect to data source
- Exception handling
- Limited fine tuning
- Retrieval-Augmented Generation (RAG)
- Chaining
- Evaluate flows

As a developer, how do I **build** an application with LLMs?

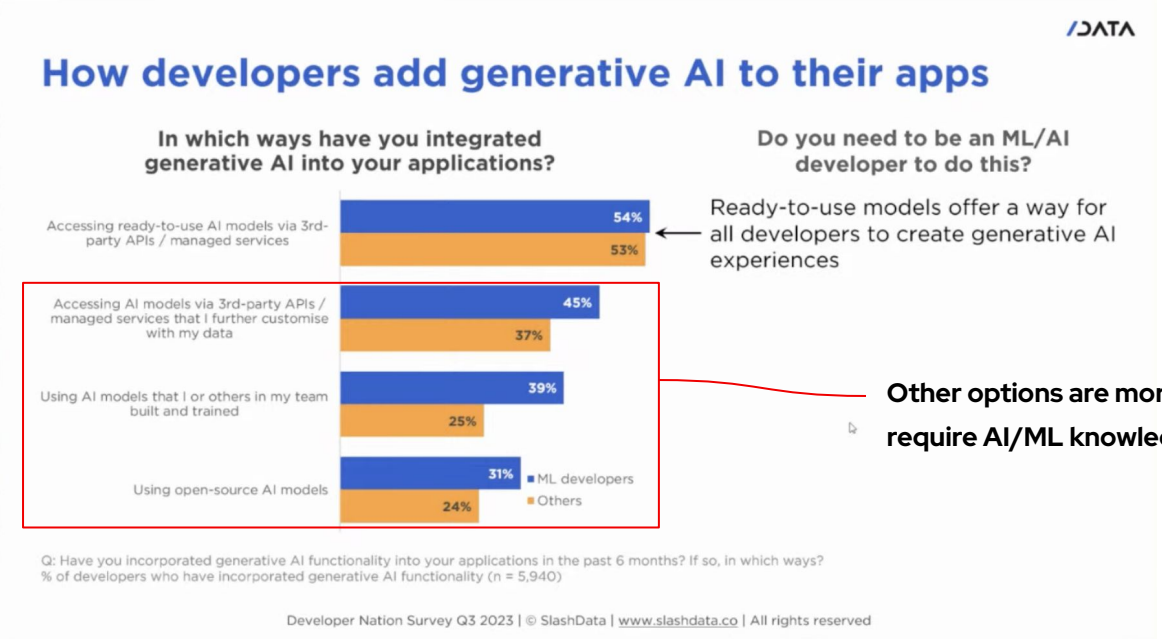
Operationalizing

- Model serving
- Endpoints
- Monitoring
- Integrate with apps

As a developer, how do I **deploy** my application with LLMs?

@danieloh30

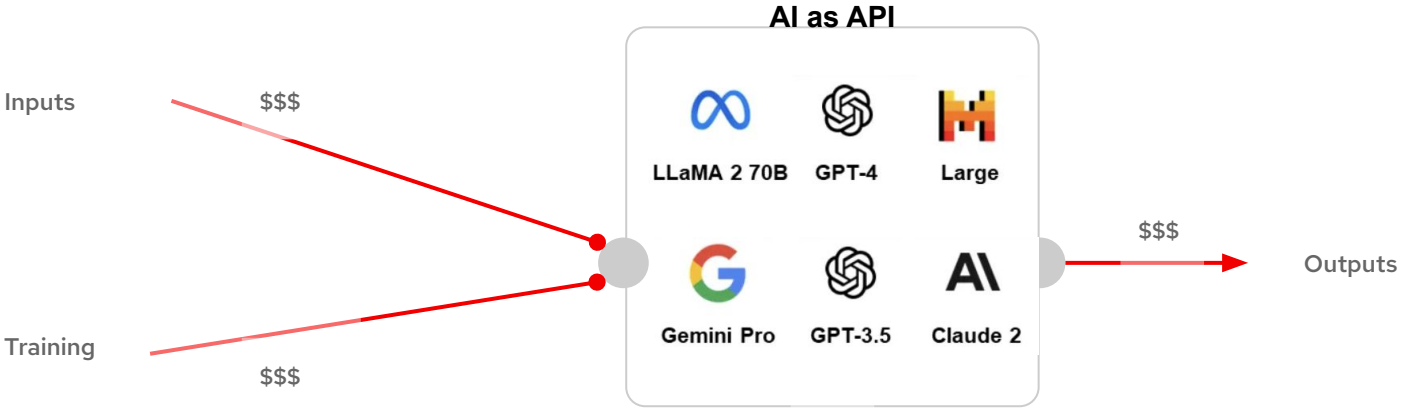
Developer's don't have to be AI/ML experts



AI/ML expertise is not a significant factor for integrating ready-to-use models

@danieloh30

Those APIs are costly... and challenging to test against



$$$


OpenAI pricing

GPT-4

Model	Prompt	Completion
8K context	$0.03 / 1K tokens	$0.06 / 1K tokens
32K context	$0.06 / 1K tokens	$0.12 / 1K tokens

Chat

Model	Input	Output
4K context	$0.0015 / 1K tokens	$0.002 / 1K tokens
16K context	$0.003 / 1K tokens	$0.004 / 1K tokens

InstructGPT

Model	Usage
Ada	$0.0004 / 1K tokens
Babbage	$0.0005 / 1K tokens
Curie	$0.0020 / 1K tokens
Davinci	$0.0200 / 1K tokens

Image models

Resolution	Price
1024×1024	$0.020 / image
512×512	$0.018 / image
256×256	$0.016 / image

Embedding models

Model	Usage
Ada v2	$0.0001 / 1K tokens
Ada v1	$0.0040 / 1K tokens
Babbage v1	$0.0050 / 1K tokens
Curie v1	$0.0200 / 1K tokens
Davinci v1	$0.2000 / 1K tokens

Fine-tuning models

Model	Training	Usage
Ada	$0.0004 / 1K tokens	$0.0016 / 1K tokens
Babbage	$0.0006 / 1K tokens	$0.0024 / 1K tokens
Curie	$0.0030 / 1K tokens	$0.0120 / 1K tokens
Davinci	$0.0300 / 1K tokens	$0.1200 / 1K tokens

Audio models

Model	Usage
Whisper	$0.006 / minute (rounded to the nearest second)

Nebuly

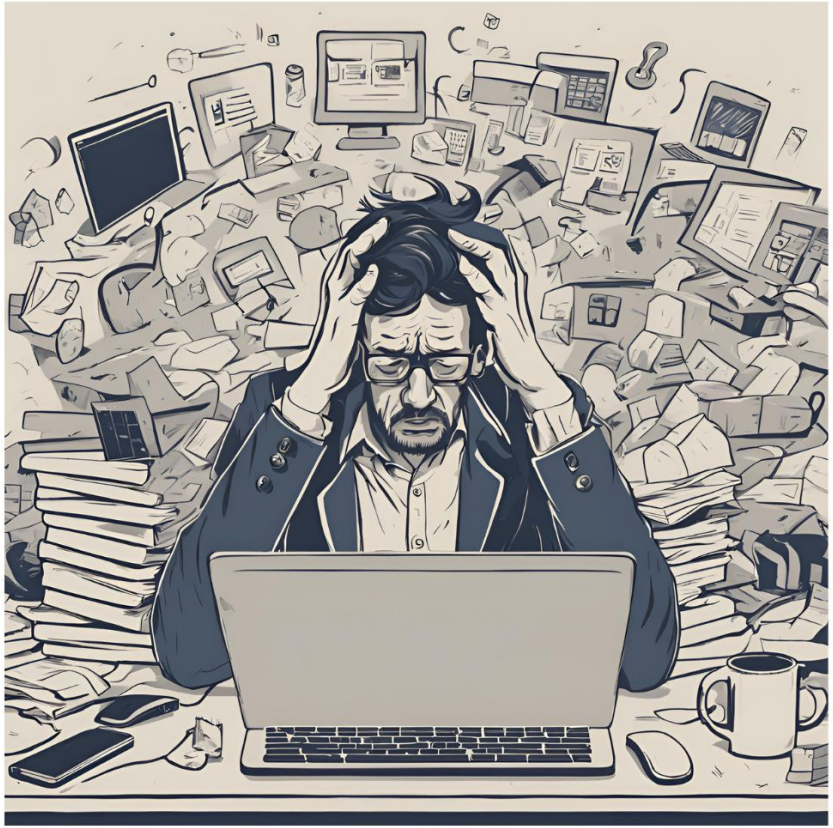
# of tokens used and costs randomly exploded over night

Cost for GPT failed requests:

- Issue from OpenAI side
- Timeout in Application

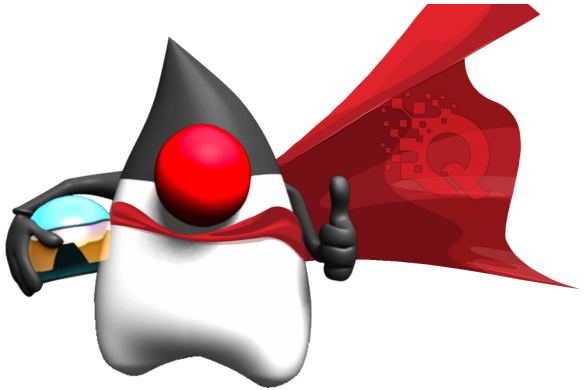
@danieloh30

Still Feeling Stuck in a Java Rut?



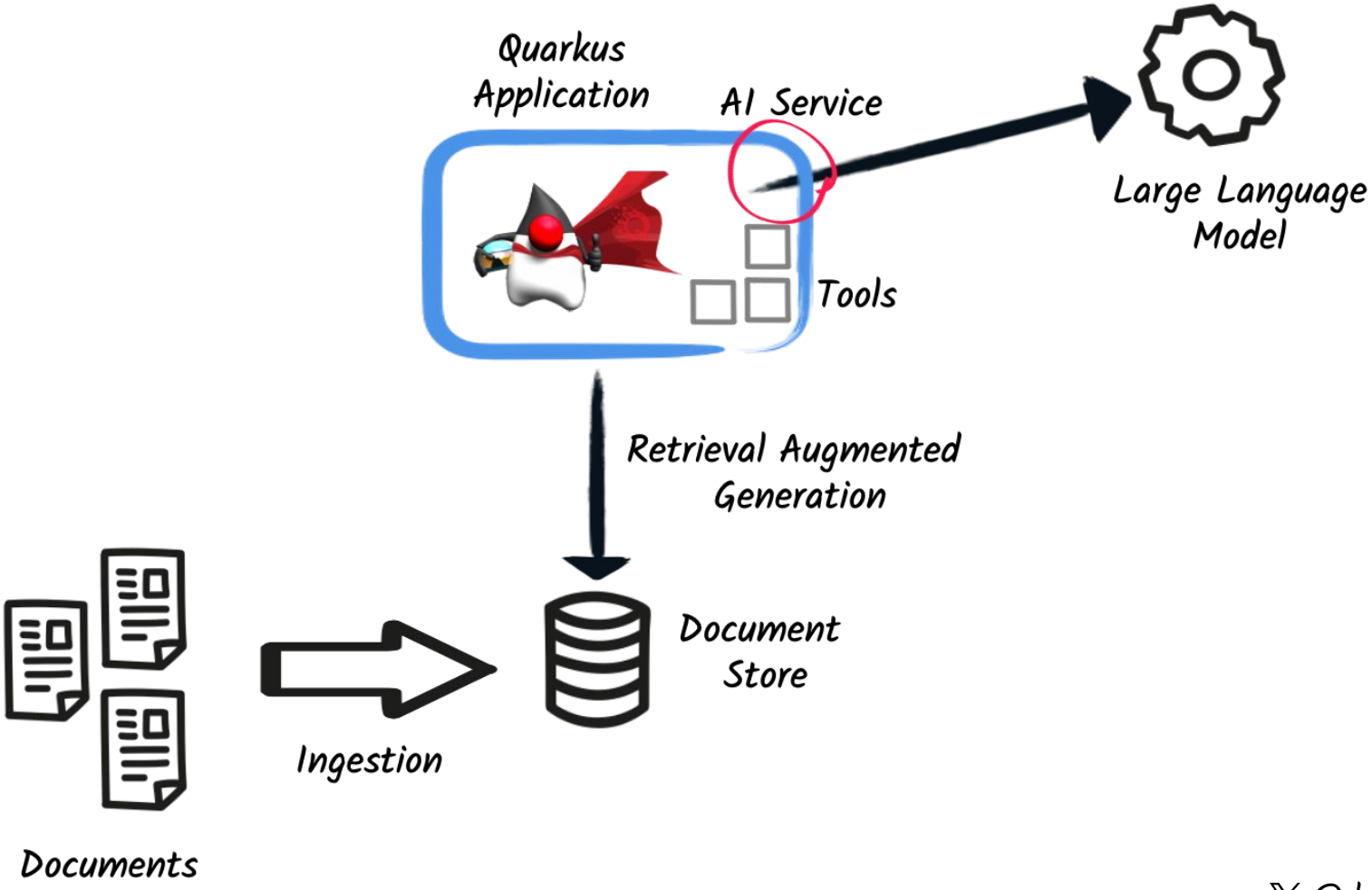
@danieloh30

Langchain4j



QUARKUS

@danieloh30


Quarkus
Application
AI Service
Large Language
Model
Tools
Retrieval Augmented
Generation
Document
Store
Ingestion
Documents

```
<dependency>
    <groupId>io.quarkiverse.langchain4j</groupId>
    <artifactId>quarkus-langchain4j-openai</artifactId>
    <version>0.17.0</version>
</dependency>
```






```
package io.quarkiverse.langchain4j.samples;

import dev.langchain4j.service.SystemMessage;
import dev.langchain4j.service.UserMessage;
import io.quarkiverse.langchain4j.RegisterAiService;

@RegisterAiService(
        tools = EmailService.class
)
public interface MyAiService {

    @SystemMessage("You are a professional poet")
    @UserMessage("""
                Write a poem about {topic}. The poem should be {lines}
lines long. Then send this poem by email.
            """)
    String writeAPoem(String topic, int lines);
}
```

DEMO

Agentic AI applications

Prompts

- Interacting with the model for asking questions
- Interpreting messages to get important information
- Populating Java classes from natural language
- Structuring output

Agent and Tools

- Mixing business code with model
- Delegating to external services
- Describing tools in a manner most beneficial to the agent

Embedding Documents (RAG)

- Adding specific knowledge to the model
- Asking questions about supplied documents
- Natural queries

Document Store

- Redis Store
- Chroma Store
- Infinispan Store
- Pinecone Store
- PgVector (PostgreSQL) Store
- In-Process Embeddings
- Loading CSV files
- Neo4j Store

Observability

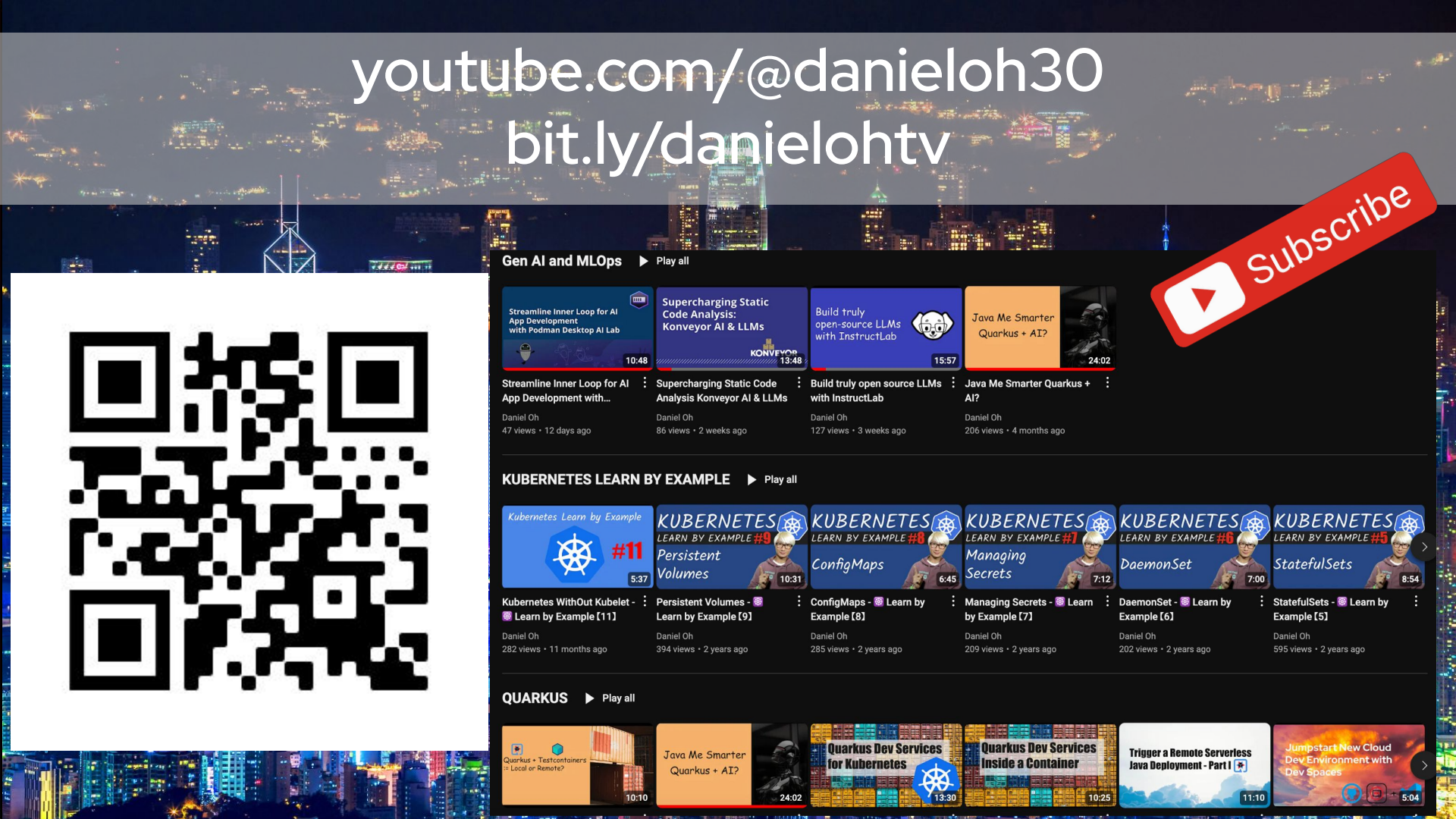
- Metrics
- Tracing
- Auditing

Red Hat

Advanced Topics

- Fault Tolerance
- WebSockets
- Enabling and Disabling Integrations

youtube.com/@danieloh30
bit.ly/danielohtv



THANK YOU!
QUESTIONS?

KubeCon | CloudNativeCon

THE LINUX FOUNDATION
OPEN SOURCE SUMMIT | AI_dev Open Source GenAI & ML Summit

China 2024