WPS306

Responsible Machine Learning In The Era Of Generative AI

Sam Palani
Sr Leader, Machine Learning

Kwadwo Ankrah-Kusi
AI/ML Solution Architect

Henry Jia
Sr AI/ML Solution Architect

What is responsible AI?

Fairness

How a system impacts different subpopulations of users (such as by sex or ethnicity)

Explainability

Mechanisms to understand and evaluate the outputs of an AI system

Robustness

Mechanisms to ensure an AI system operates reliably

Privacy and Security

Data used in accordance with privacy considerations and protected from theft and exposure

Governance

Processes to define, implement, and enforce responsible AI practices within an organization

Transparency

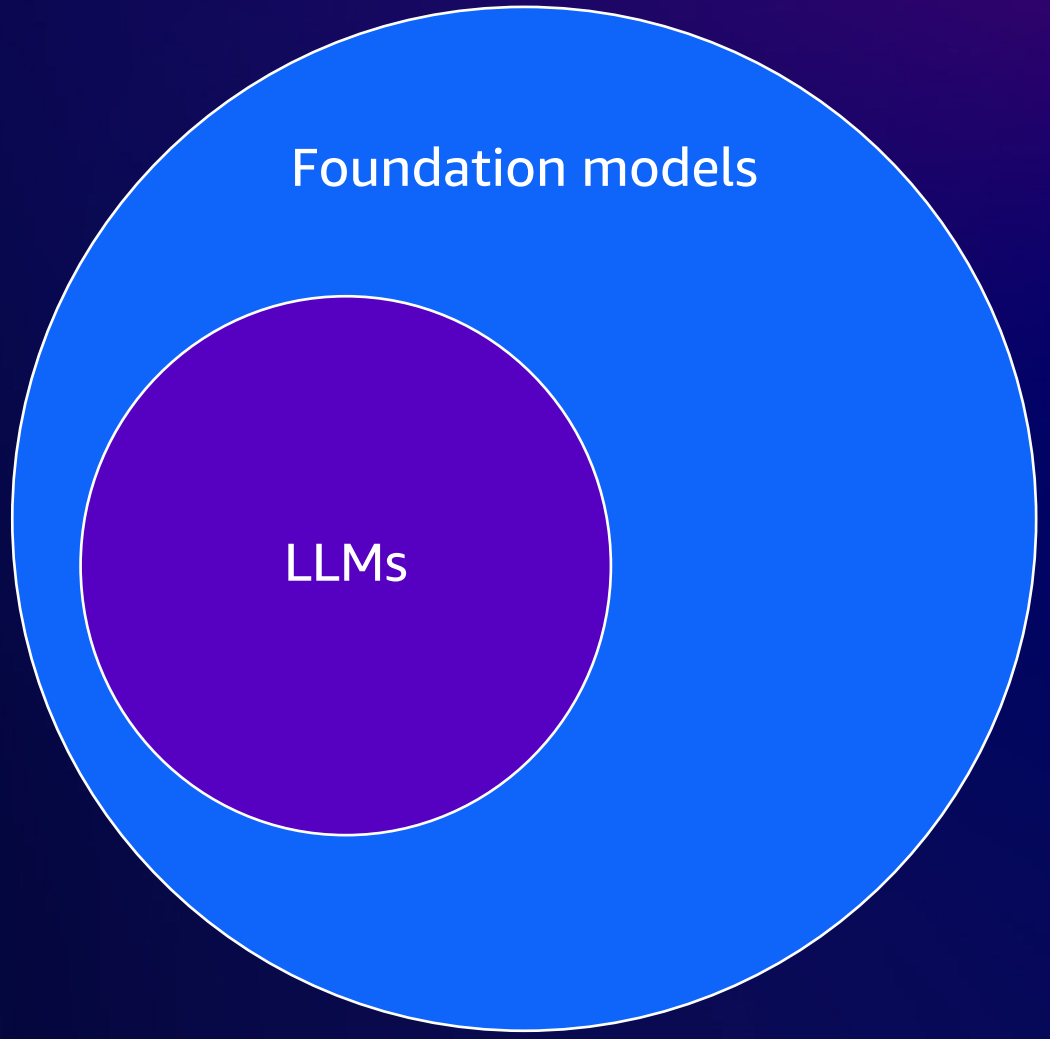
Communicating information about an AI system so stakeholders can make informed choices about their use of the system

Generative AI

Generative AI is a type of AI that can create new content and ideas, including conversations, stories, images, videos, and music. Like all AI, generative AI is powered by ML models—very large models that are pre-trained on vast amounts of data and commonly referred to as Foundation Models (FMs).

What are large language models (LLMs)

- Trained on text
- Excel at natural language prompts
- Have accelerated use cases such as question answering, code generation/explanation, summarization, etc.



What do LLMs do?

- Like all ML models, LLMs make predictions
- The next “token” in a sequence
- They produce a reasonable continuation of text based on the training data

Building an honest and responsible AI system requires

Bias mitigation	4.5%
Transparency	3.6%
Ethics	3.2%
Accountability	3.1%
Privacy	2.7%

Generative AI - Responsible AI Challenges

Legacy ML vs Foundation Model Challenges:

- Fairness
  - Narrow vs Open ended content
  - Context
- Privacy
  - Data Leak

Foundation Model Specific Challenges:

- Hallucinations
- Toxicity
- Intellectual Property
- Plagiarism
- Disruption Of Work

Addressing Responsible AI Challenges

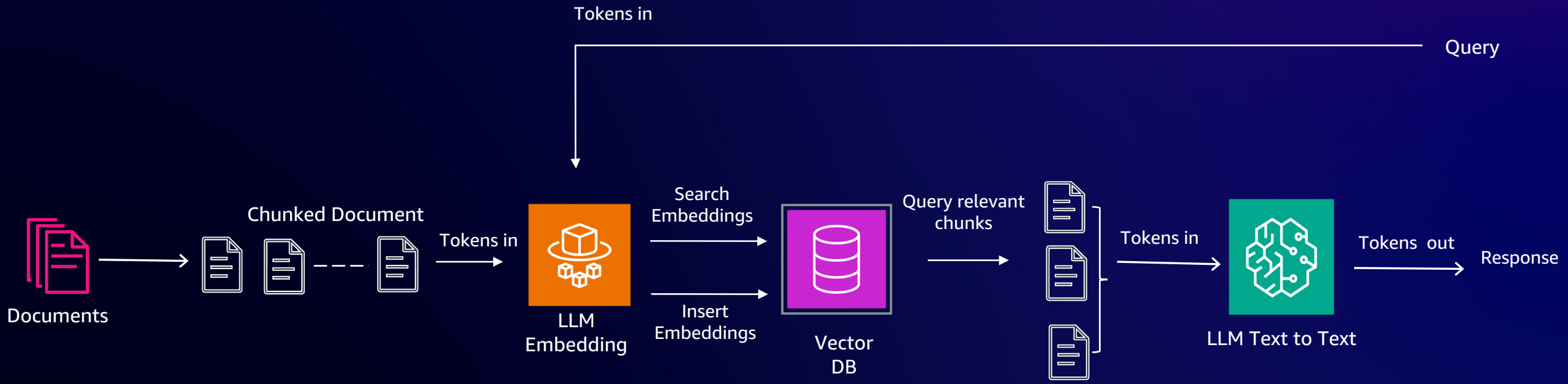
Data

- Toxicity
- Fairness
- Differential Privacy
- Model Disgorgement
- Fine Tuning

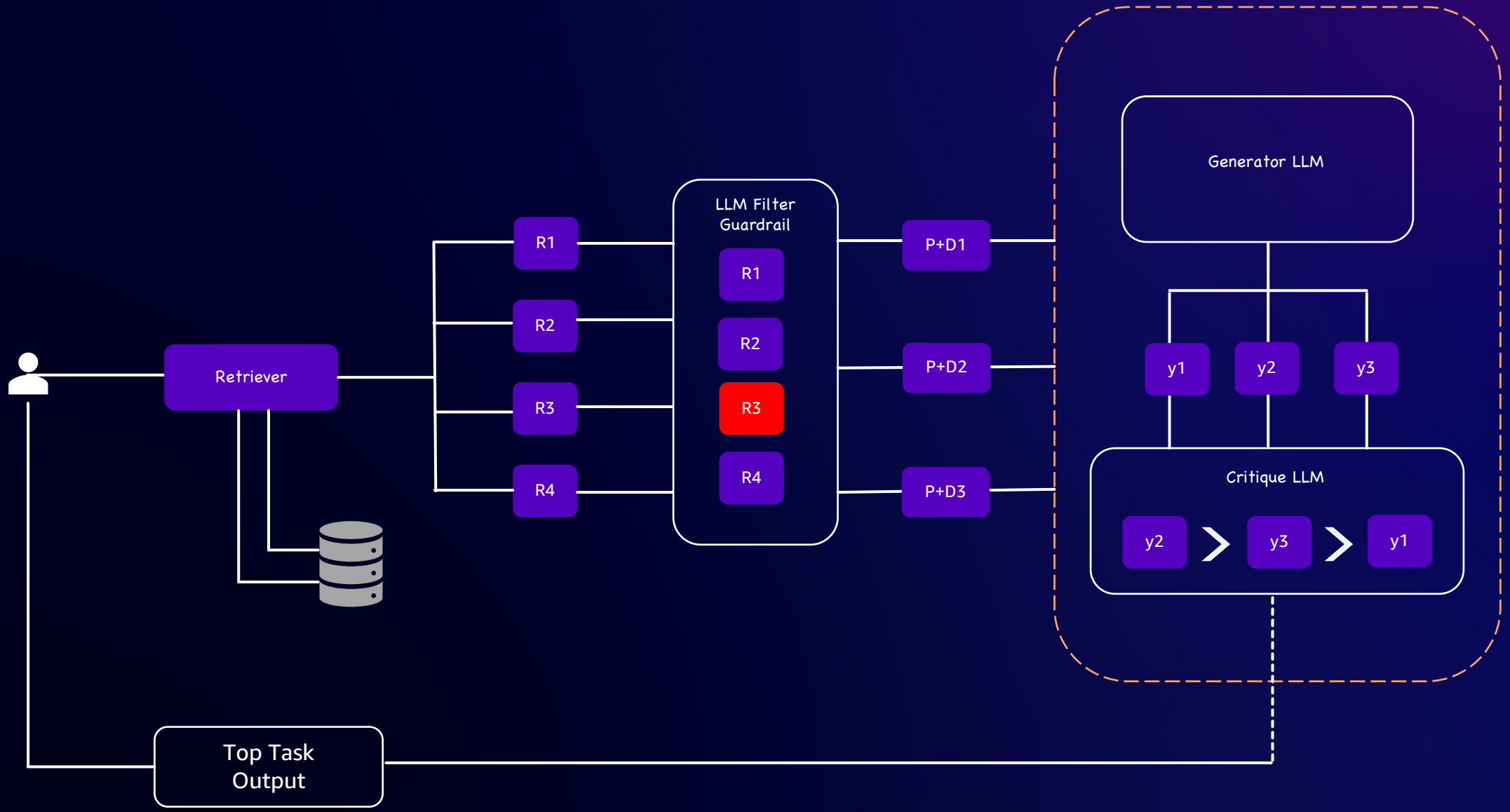
Output

- Guardrails
- Evaluations
- Watermarking
- Disclaimers
- Attributions
- Augmentations

Architecture Patterns – Grounding LLM



Self-Reflective RAG



References & Exploring Further

Getting Started with Amazon Bedrock

Llama2 on SageMaker Jumpstart

RAG on AWS

Responsible AI Hub

The Ethical Algorithm (Book)

Thank you!

Please complete the session survey in the mobile app

Sam Palani
sampal@amazon.com

Kwadwo Ankrah-Kusi
kankrah@amazon.com

Henry Jia
henryjia@amazon.com