DATA+AI SUMMIT BY databricks

A Practitioner's perspective on LLMOps

Hien Luu
06/12/2024

Implementing GenAI at scale in the enterprise is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

Cassie Kozyrkov

About Me

Head of ML Platform @ DoorDash

Hien Luu

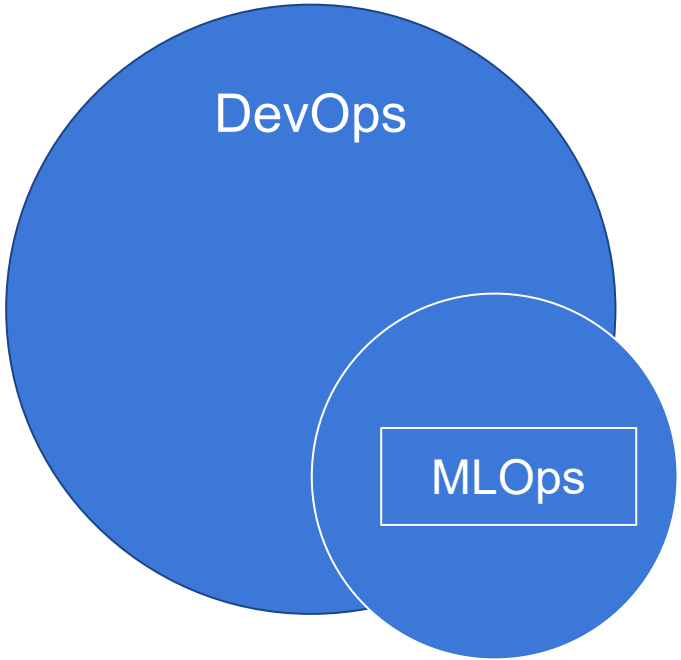
Prior work experience:

Uber

Agenda

- Unique Challenges of LLMs Application
- LLMs Application Archetypes
- LLM Use Cases
- Practitioner’s Perspective on LLMOps

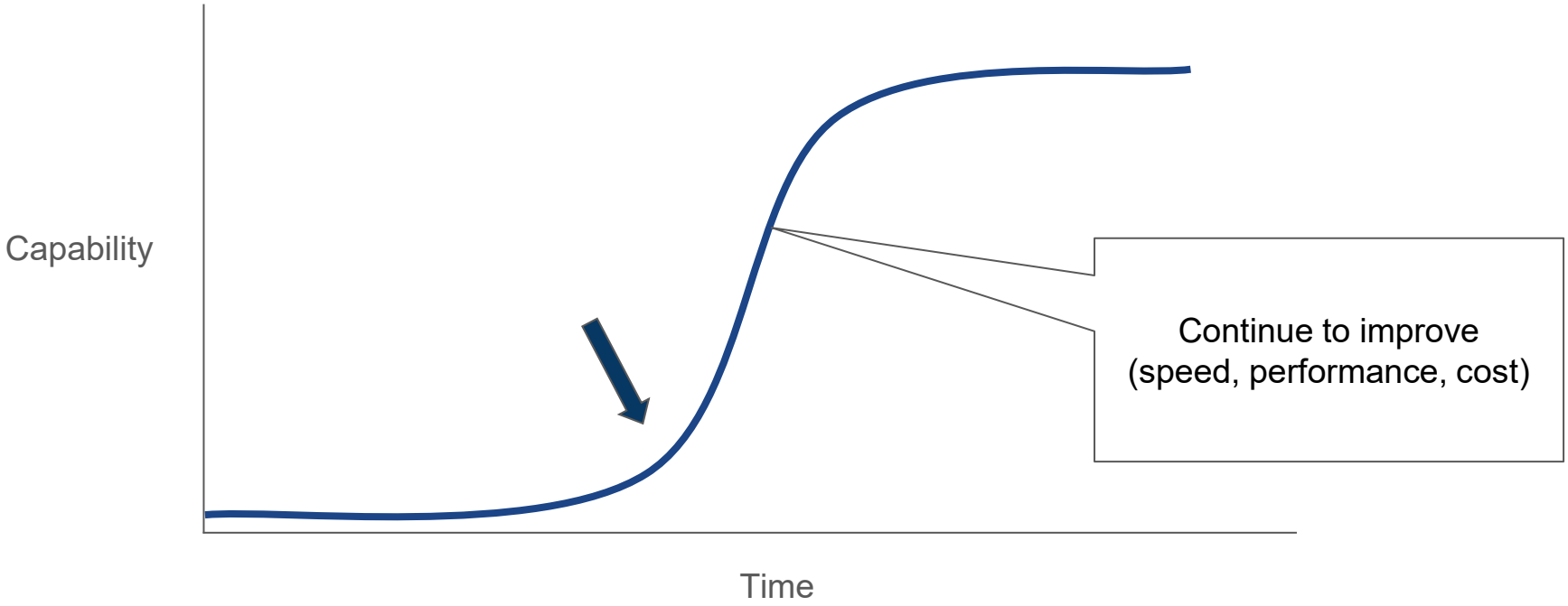
Is MLOps Still Relevant?



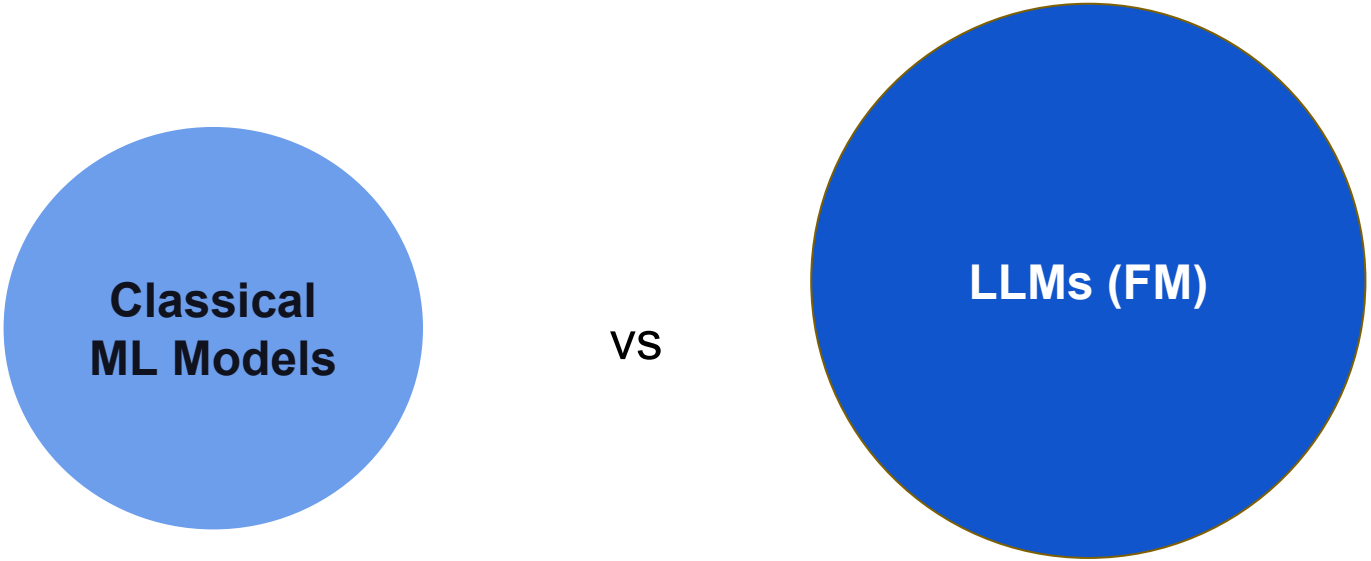
- Automation
- Collaboration
- Reproducibility
- CI/CD
- Observability

Tools and Best Practices

GenAI Advancement



LLM Application Unique Challenges



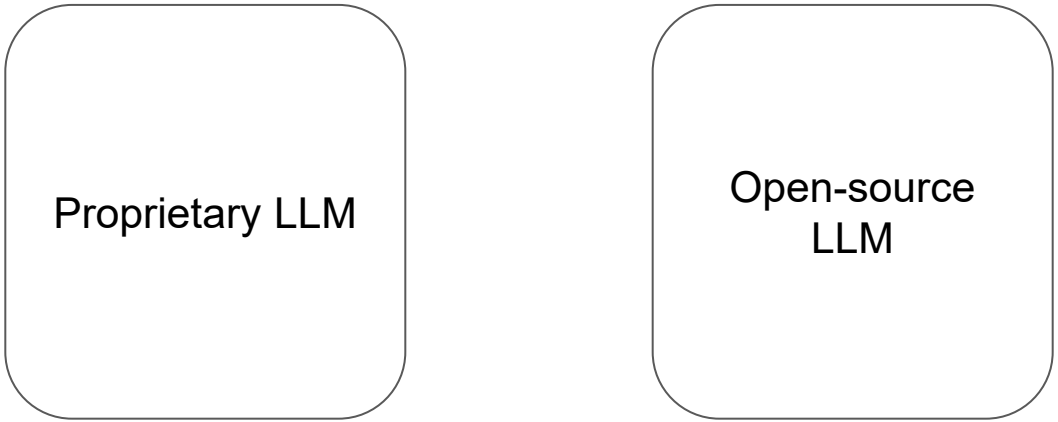
LLM Application Unique Challenges

- General purpose tasks
- Less end-2-end control
- Compute is the new oil
- GPUs are a must
- Hallucination

Prompt Engineering is the New Programming Language

LLM Application Unique Challenges

AI's Linux Moment



Owning your own LLMs is critical and achievable

LLM Application Unique Challenges

Model Quality & Performance

Computational Requirements

Memory & Size Constraints

Interpretability & Bias

Inference Related Challenges

[The New Era of Efficient LLM Deployment]

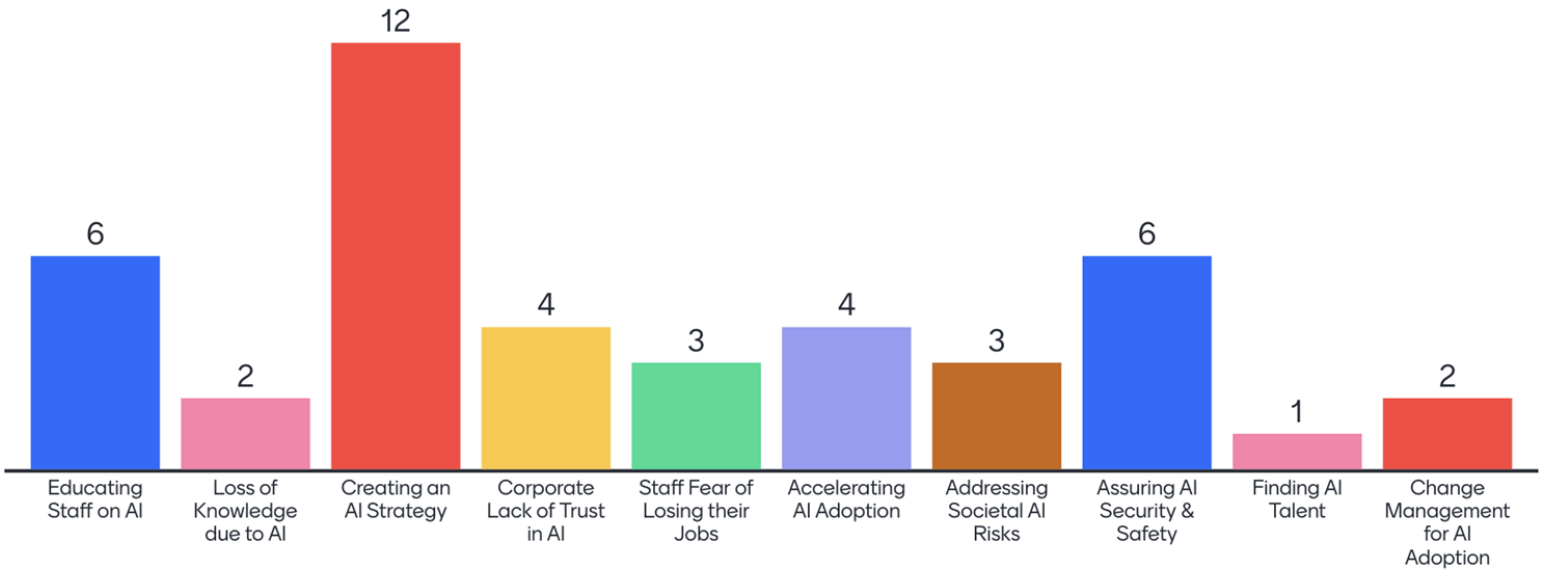
LLM Application Unique Challenges

The King Doing the Dishes?

LLM Application Unique Challenges

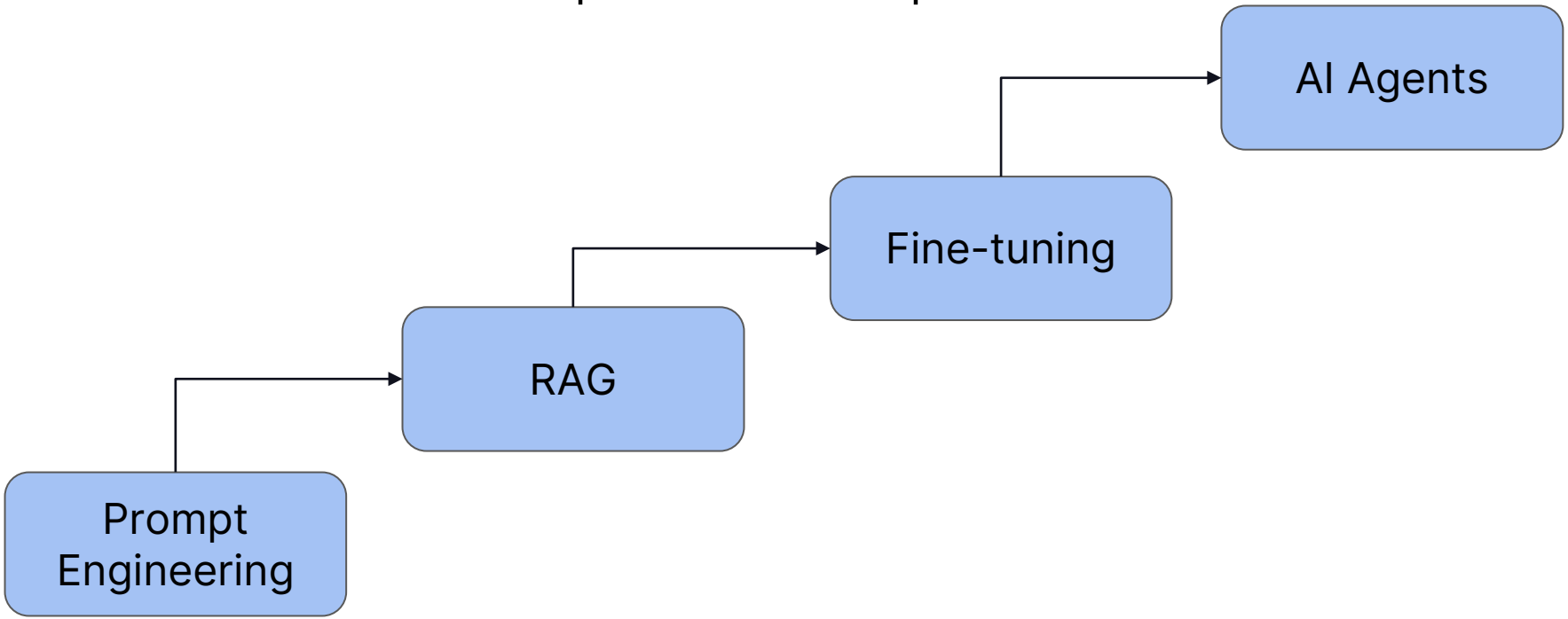
LLM Application Unique Challenges

GenAI Strategy (Where, Which, When)



LLM Application Archetypes

Operational Perspective



LLM Application Archetypes

Prompt Engineering

Operationalizing Prompts

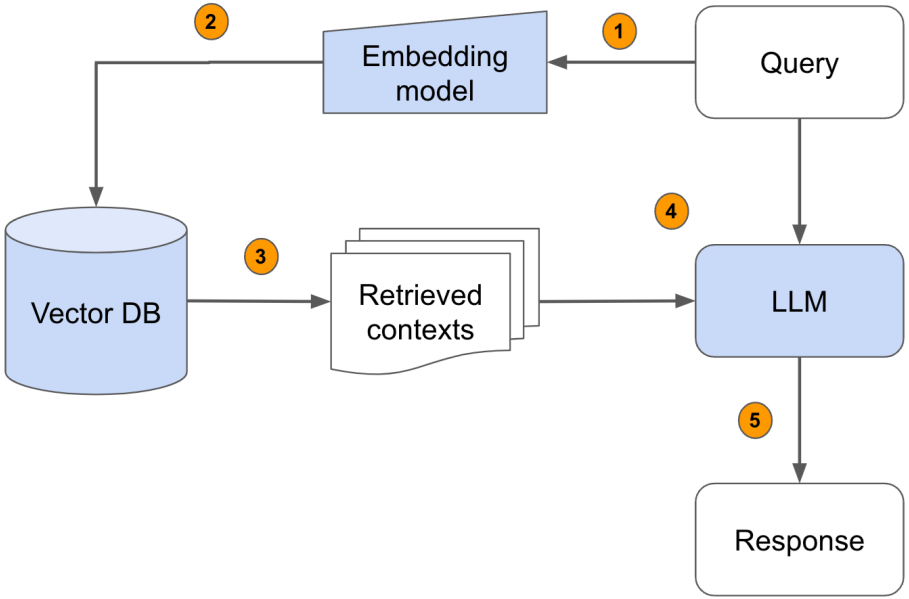
- Prompt Development/Testing
- Prompt Store
- Prompt Versioning
- Prompt + Response Logging
- Prompt + Response Monitoring
- Prompt A/B testing

Prompt & Prosper

LLM Application Archetypes

RAG

Operationalizing RAG Applications



AnyScale RAG related blog

LLM Application Archetypes

Operationalizing RAG Applications

RAG

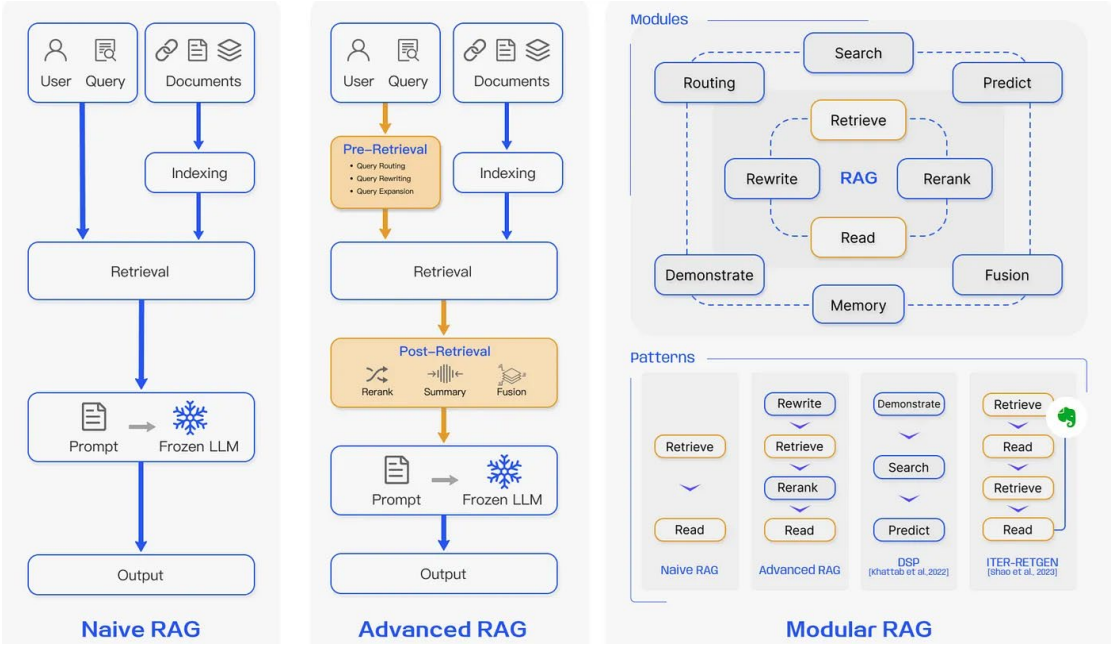
Contextual Data Pipelines

AnyScale RAG related blog

LLM Application Archetypes

RAG

Operationalizing RAG Applications



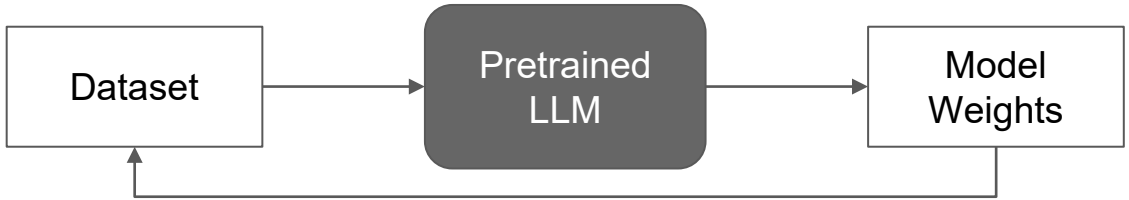
Naive, Advanced, Modular RAG

LLM Application Archetypes

Operationalizing Fine-tuning Pipelines

Fine-tuning LLM

- Automation, version control, reproducibility
- Distributed training infrastructure
  - DeepSpeed, PEFT, GPUs



LLM Application Archetypes

Operationalizing Fine-tuning Pipelines

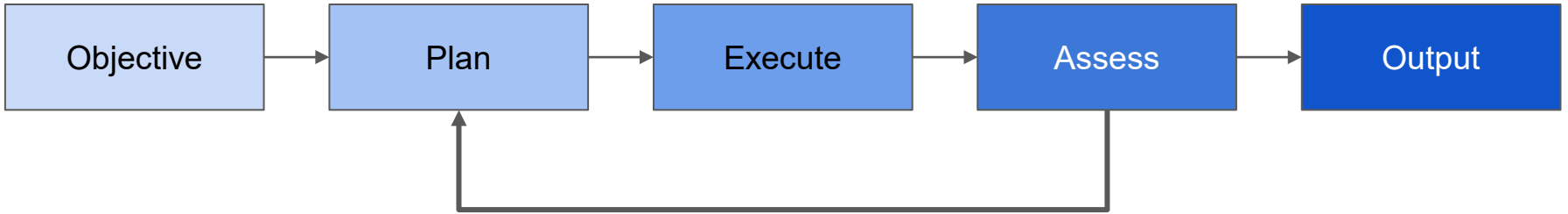
Fine-tuning LLM

- Similar to classical model training & serving
- Optimization techniques
  - vLLM, MLC, CudaGraph, MQA, Quantization, TensorRT
- Deeper understanding of GPUs, TTFT, TPOT

Self-hosted LLM

LLM Application Archetypes

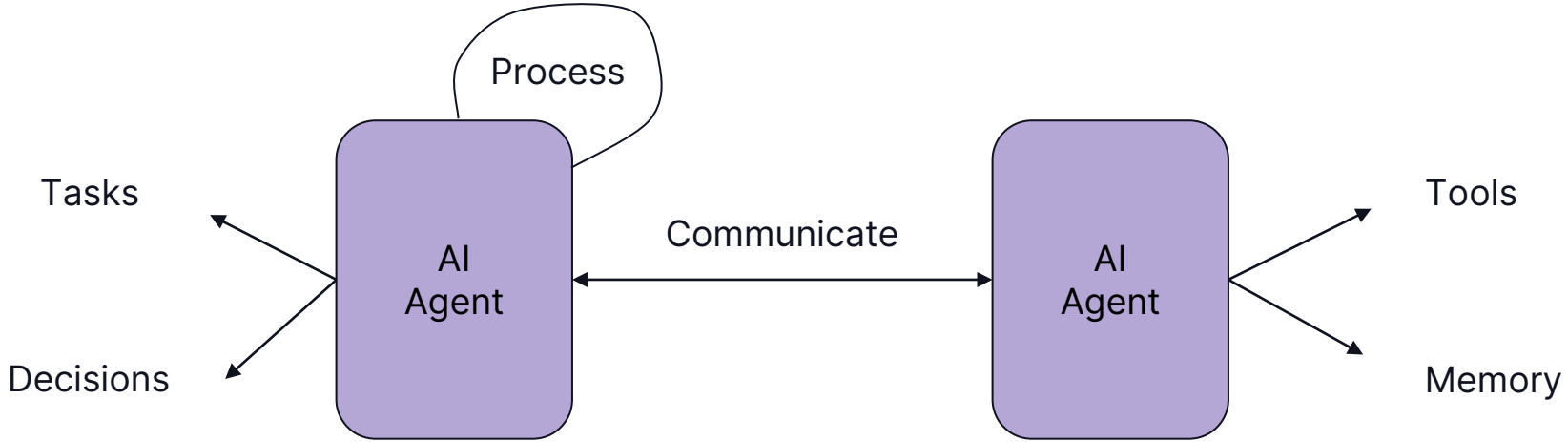
AI Agents



AI Agentic Workflow Will Drive Massive AI Progress

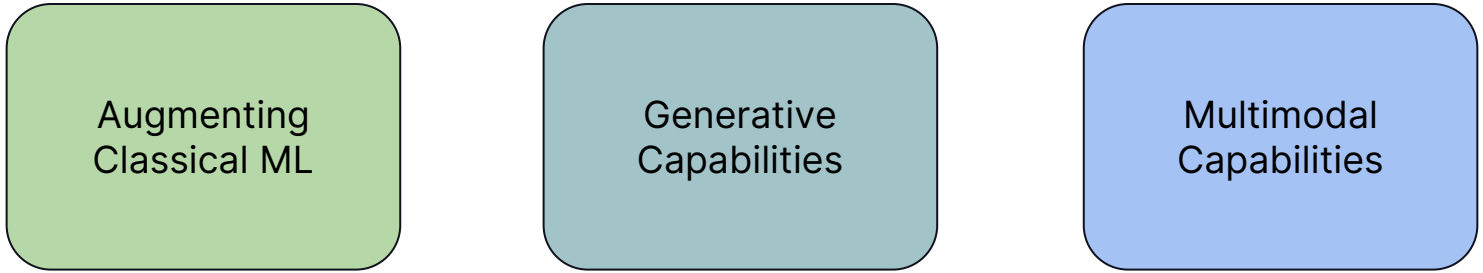
LLM Application Archetypes

Operationalize AI Agents - Components & Patterns



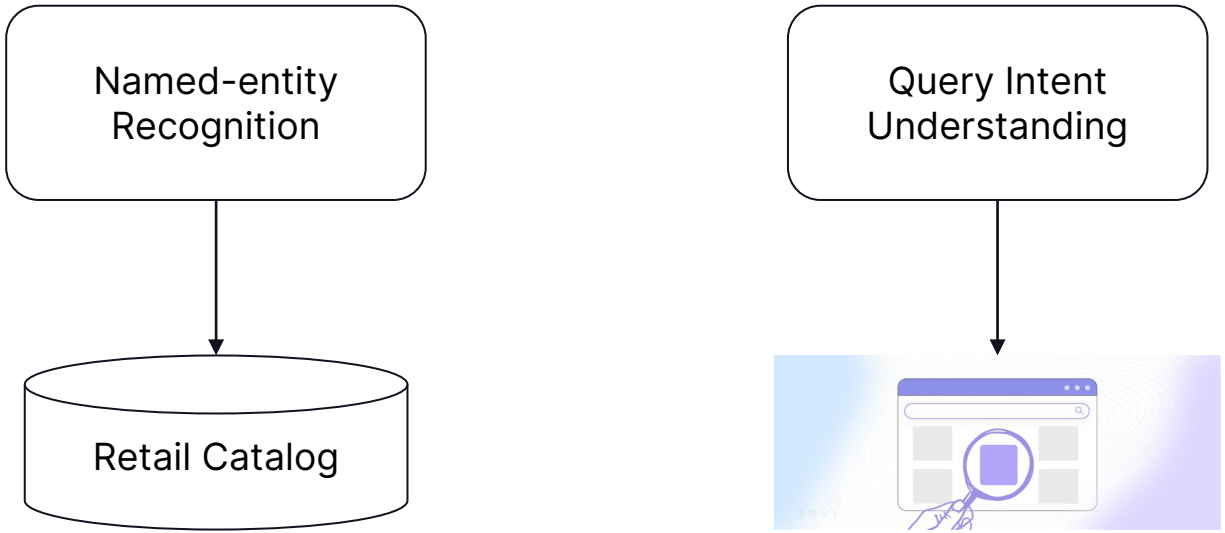
LLM Use Cases

Use Case Categories



LLM Use Cases

Augmenting Classical ML



Building DoorDash's Product Knowledge Graph w/ LLMs

LLM Use Cases

Generative Capabilities

Mom's Lasagna

A large savory slice of our 5-layered homemade lasagna baked in our bolognese sauce, topped with mozzarella.

$17.99+

Baked Spaghetti

A huge serving of spaghetti baked in your favorite homemade pasta sauce and topped with our Italian mozzarella.

$13.99+

Fettuccine Alfredo

A large serving of fettuccine smothered in a creamy Alfredo sauce topped with mozzarella and baked to perfection.

$17.99+

Craft Your Own Mac & Cheese

Start with a base of our aged white cheddar macaroni and cheese, and add one of your favorite toppings from our list of herbs, veggies, meats, and cheeses.

$17.99+

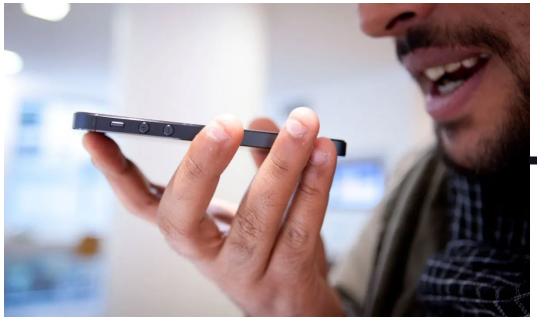
Manicotti

2 Large pasta tubes filled with a ricotta cheese filling, baked in your favorite homemade pasta sauce covered in all natural mozzarella and served with 2...

$14.99

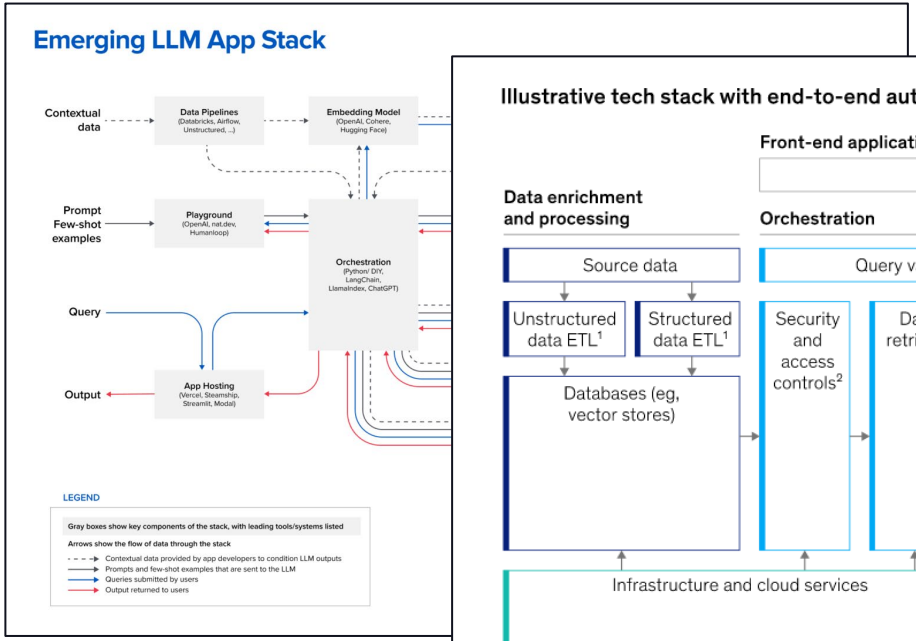
LLM Use Cases

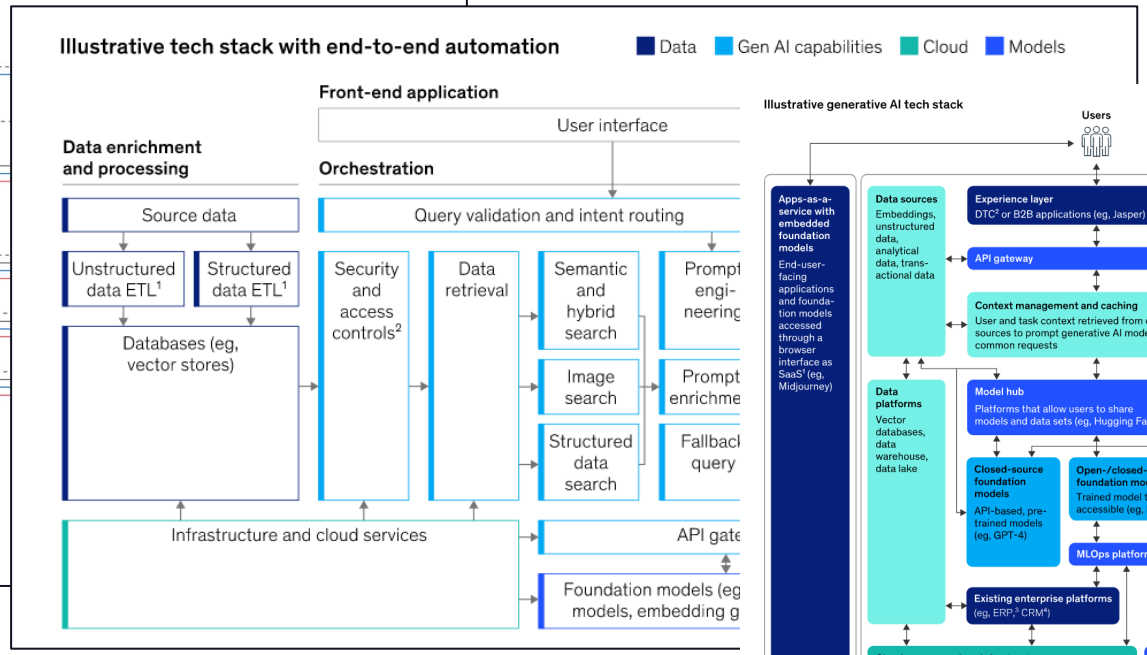
Multimodal Capabilities - Voice Ordering



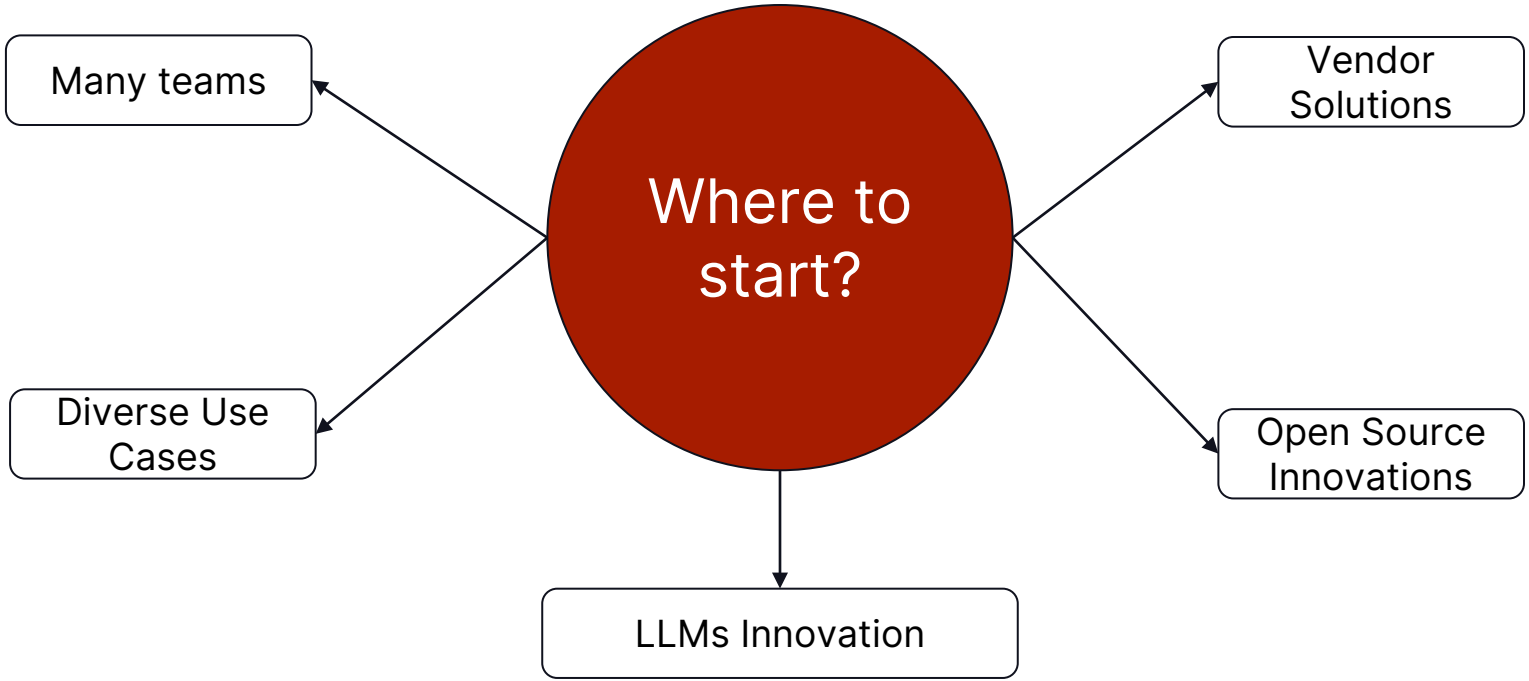
DoorDash Introduces AI and Agent-Powered Voice Ordering Solution

Practitioner's Perspective

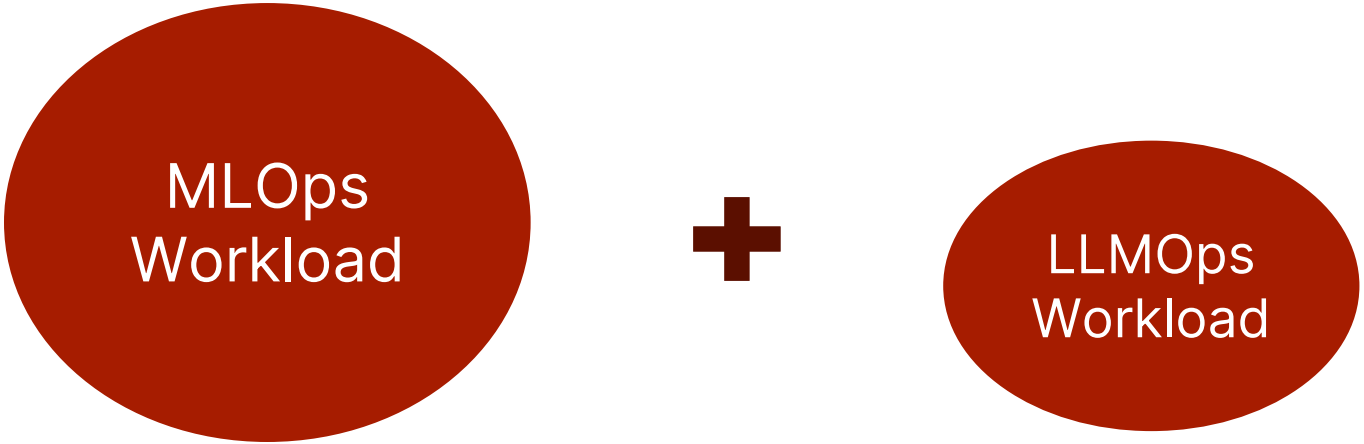




Practitioner's Perspective



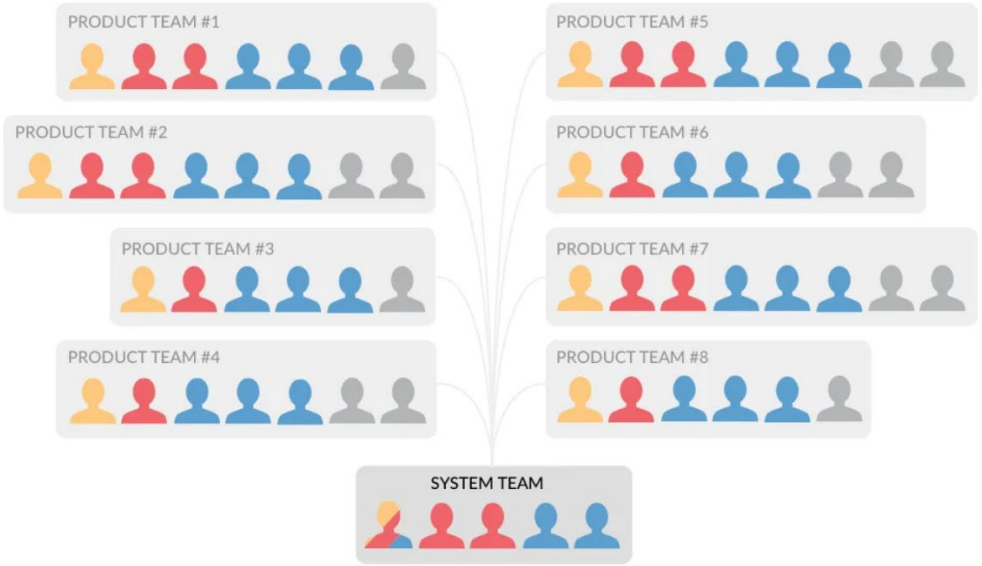
Practitioner's Perspective



Practitioner's Perspective

Understanding Use Cases & User Needs

List of
use case

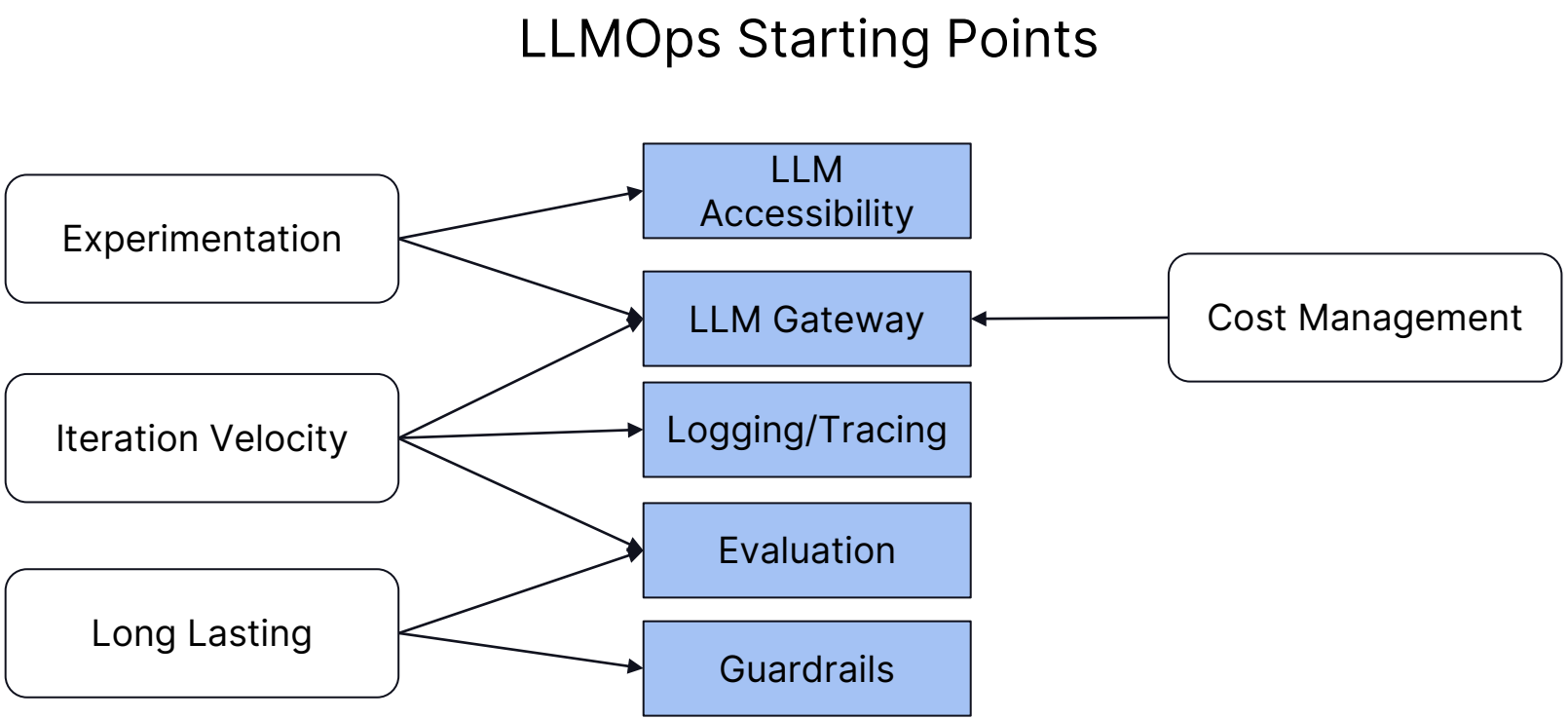


Practitioner's Perspective

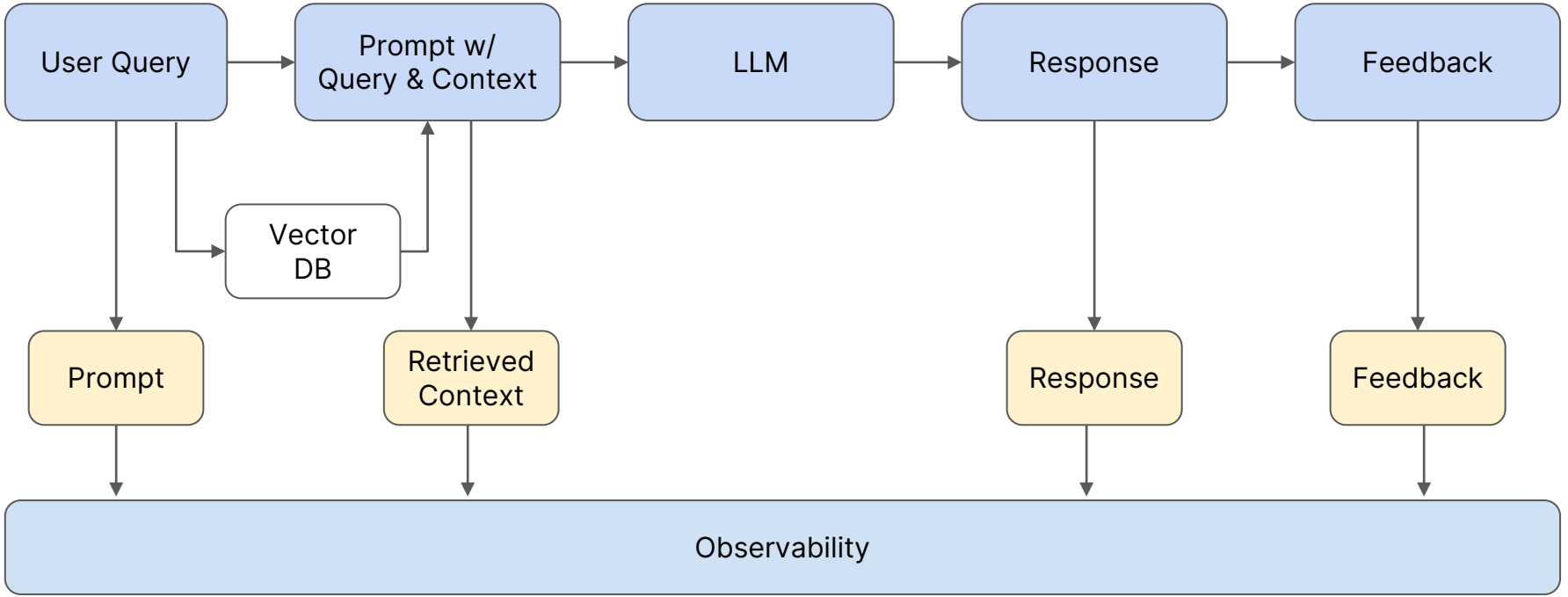
LLMOps Goals

- Experimentation
- Cost Management
- Iteration Velocity
- Long Lasting

Practitioner's Perspective



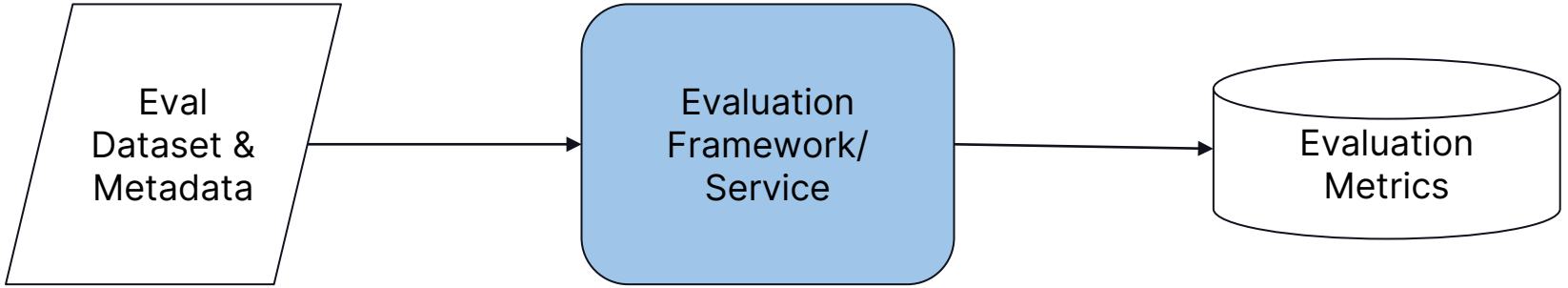
Practitioner's Perspective



Emerging Architectures for LLM Applications

Practitioner's Perspective

Evaluation - One of the Most Impactful Parts



Model Comparison

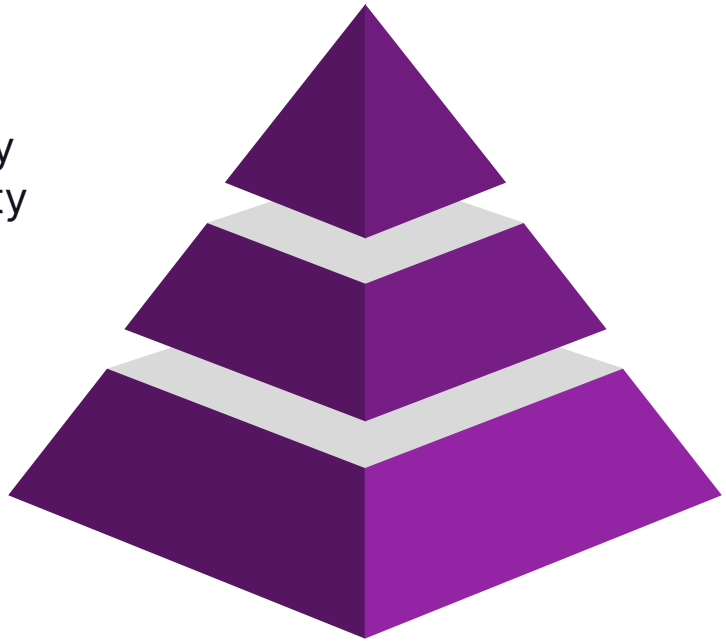
Bias Detection

Satisfaction & Trust

Practitioner's Perspective

Hierarchy of Needs from Demo to Production

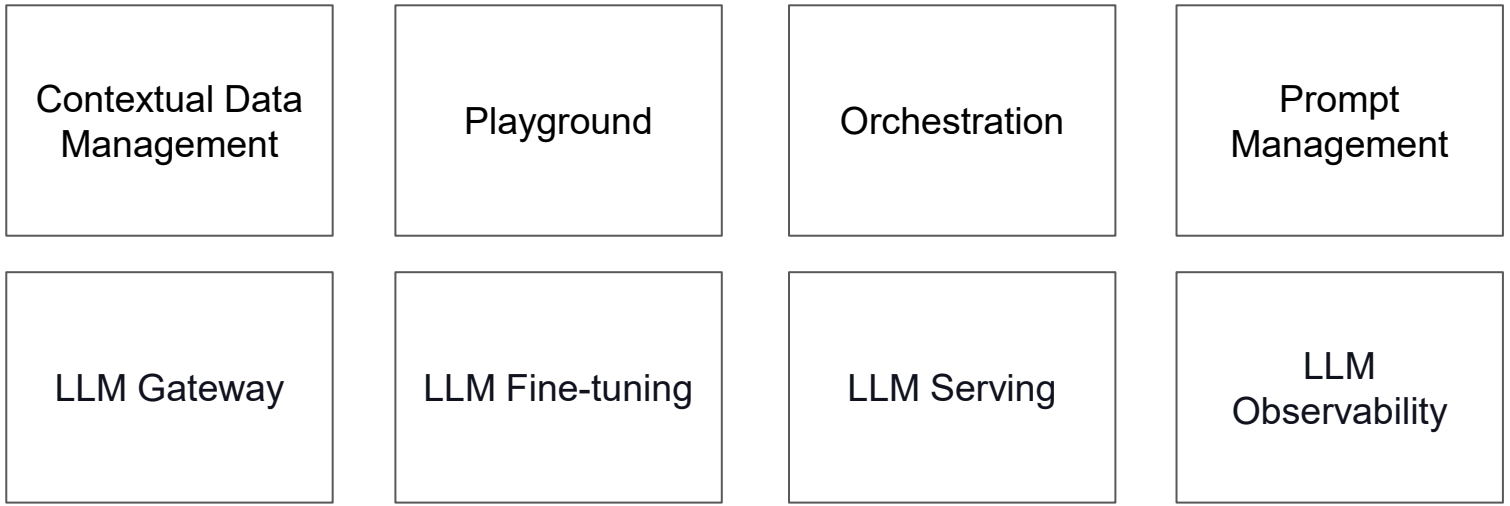
- Reliability
- Scalability
- Security
- Cost



- Harmlessness
- Factual Consistency
- Usefulness

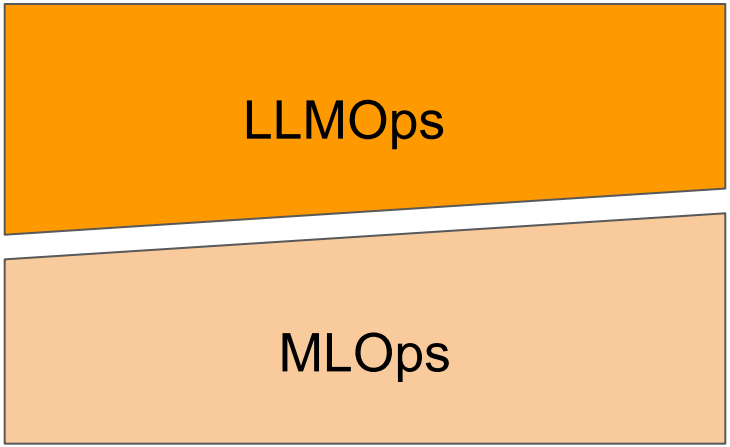
What We've Learned From A Year of Building with LLMs

Practitioner's Perspective



Use cases, archetype, domain, adoption stage

Practitioner's Perspective



Key Takeaways

- GenAI technology is at the inflection point
- LLM applications represent a different set of challenges
- Understand your goals and hierarchy of needs
- Focus less on infrastructure and more lasting value
- Craw, walk, run

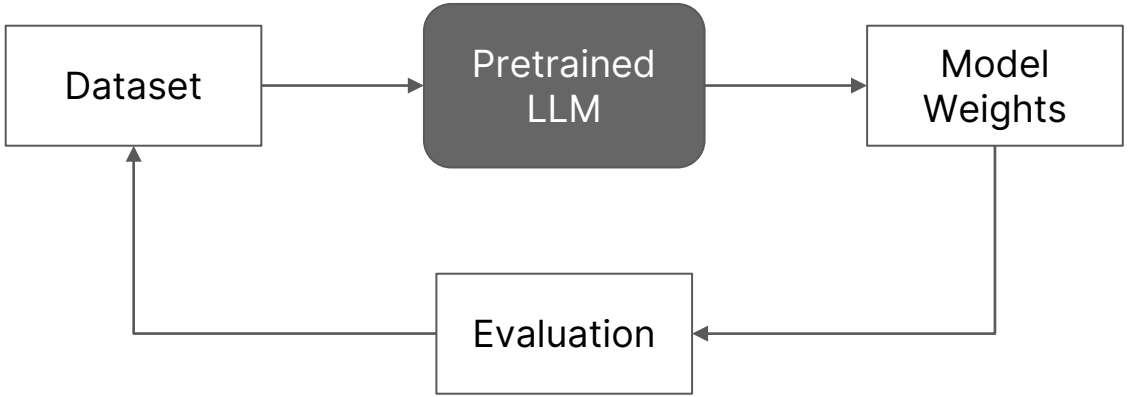
LLM Application Unique Challenges

Guardrails & Evaluation

LLM Application Archetypes

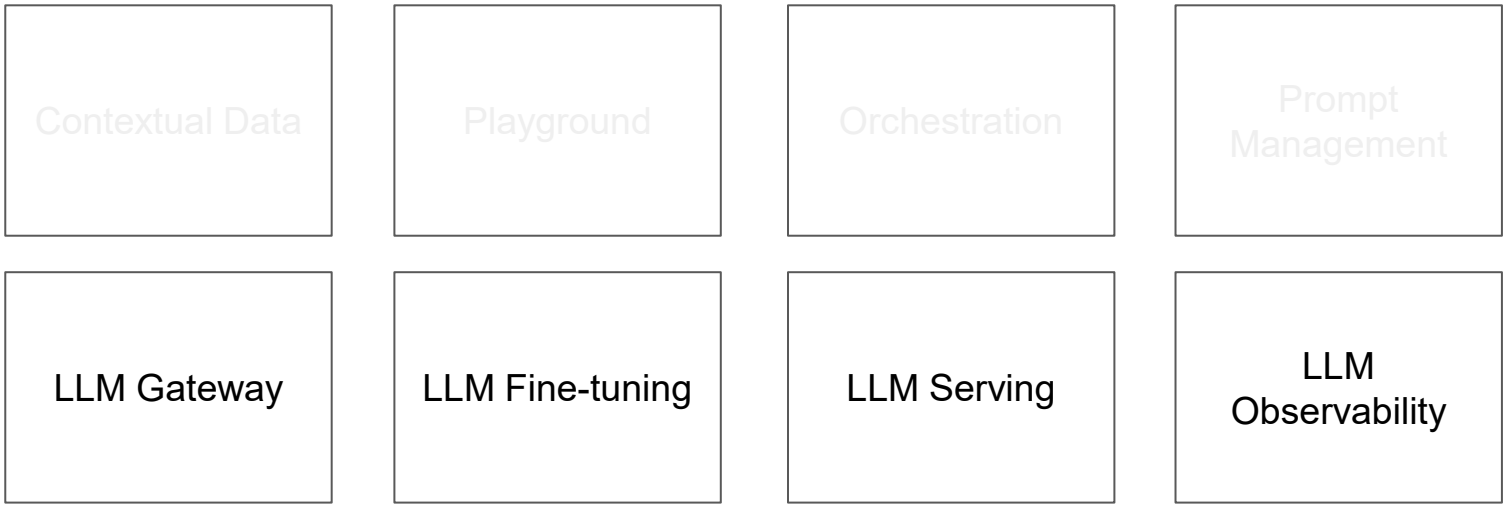
Operationalizing Fine-tuning Pipelines

Fine-tuning LLM



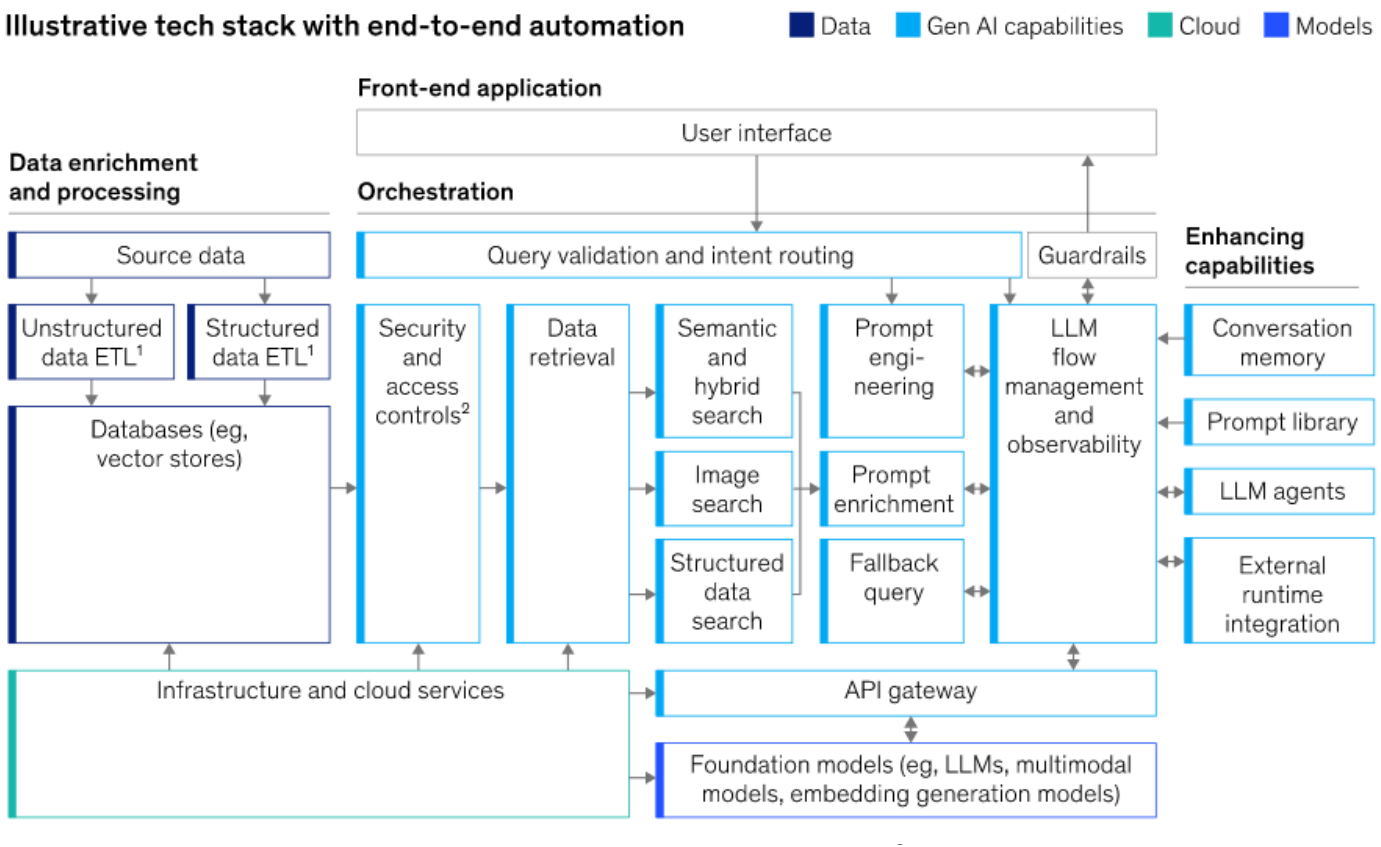
Customize LLMs

LLMOps Emerging Stack



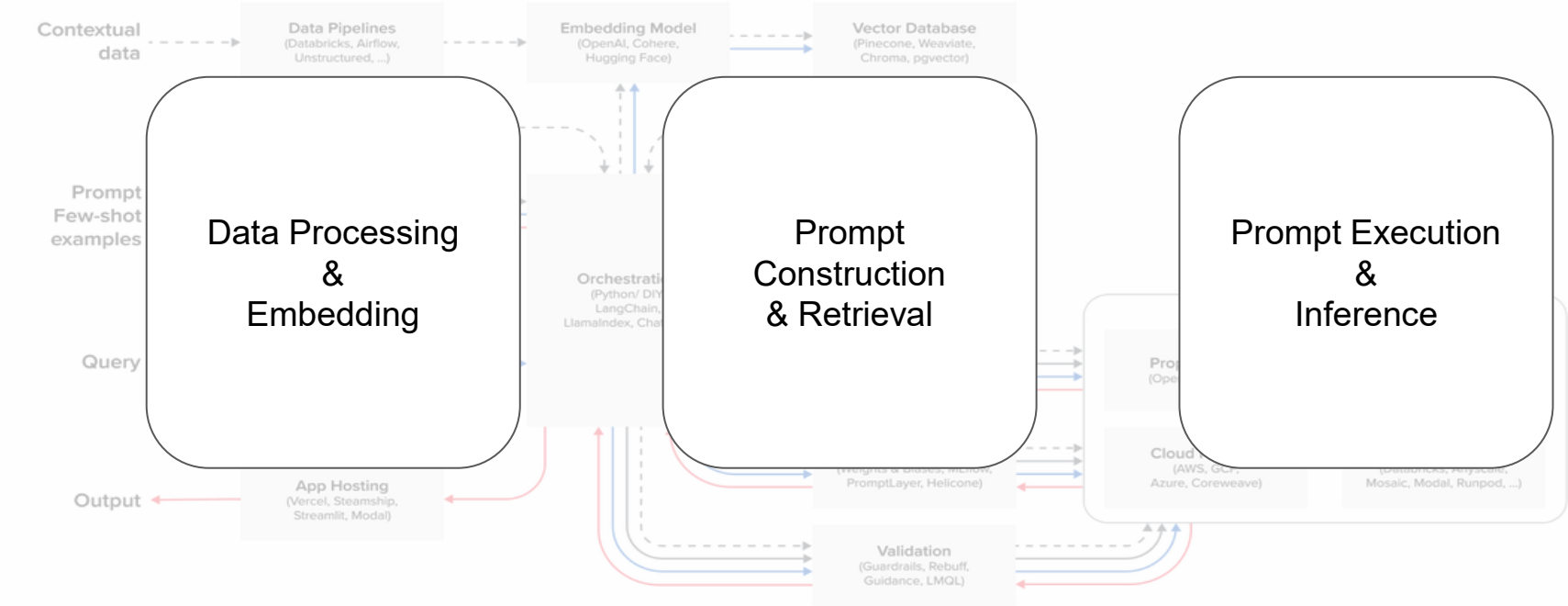
Use cases, archetype, domain, adoption stage

LLMOps Emerging Stack



LLMOps Emerging Stack

A16Z - Emerging Architecture for LLM Application



Resources

-
-
-

*Building **good** products on top of LLMs is incredibly difficult*

Mira Murati, CTO of OpenAI