Introduction

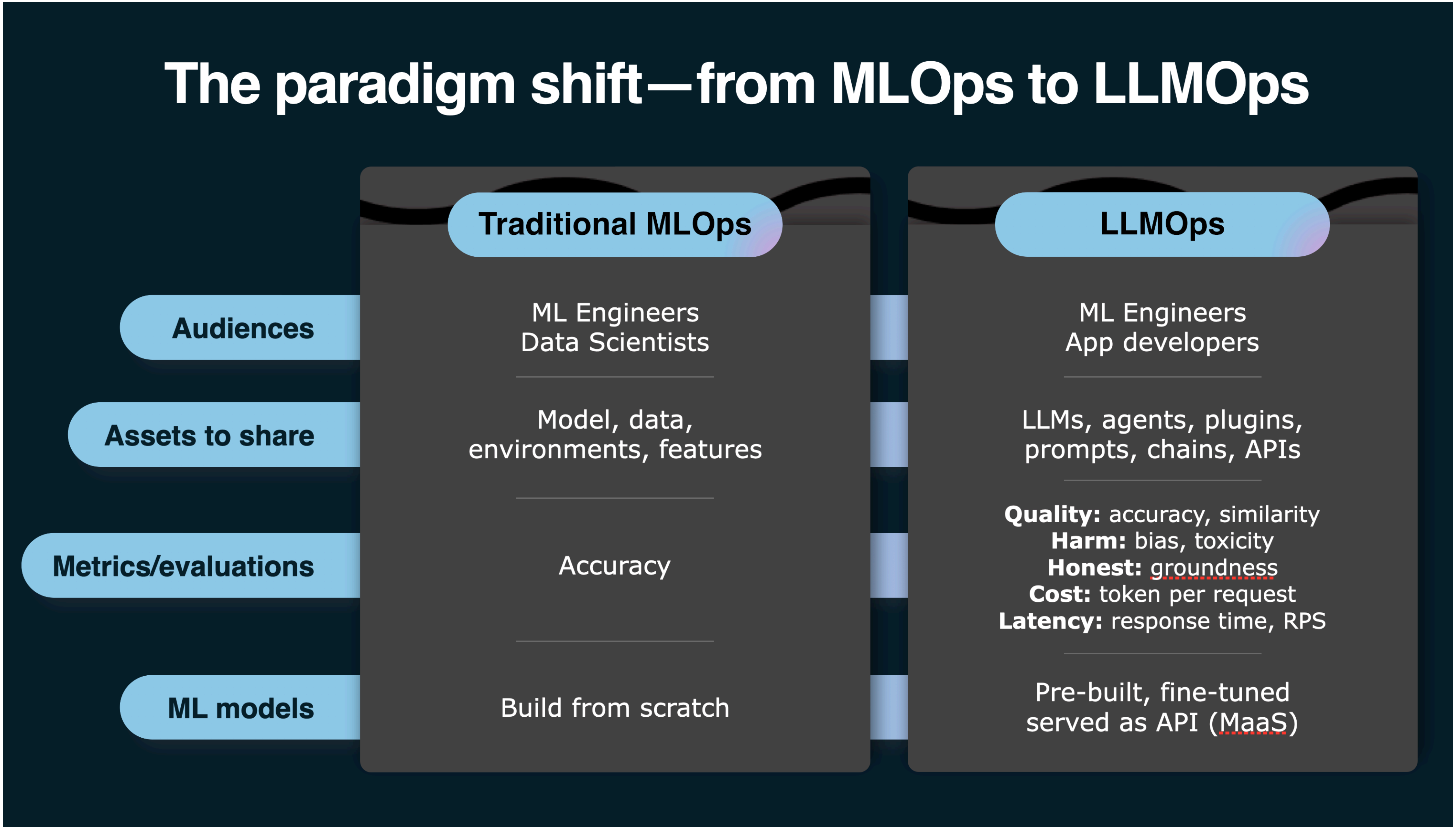
In this chapter, you will:

- Understand the Paradigm Shift from MLOps to LLMOps
- The LLM Lifecycle
- Lifecycle Tooling
- Lifecycle Metrification and Evaluation

Understand the Paradigm Shift from MLOps to LLMOps

LLMs are a new tool in the Artificial Intelligence arsenal, they are incredibly powerful in analysis and generation tasks for applications, however this power has some consequences in how we streamline AI and Classic Machine Learning tasks.

With this, we need a new Paradigm to adapt this tool in a dynamic, with the correct incentives. We can categorize older AI apps as "ML Apps" and newer AI Apps as "GenAI Apps" or just "AI Apps", reflecting the mainstream technology and techniques used at the time. This shifts our narrative in multiple ways, look at the following comparison.



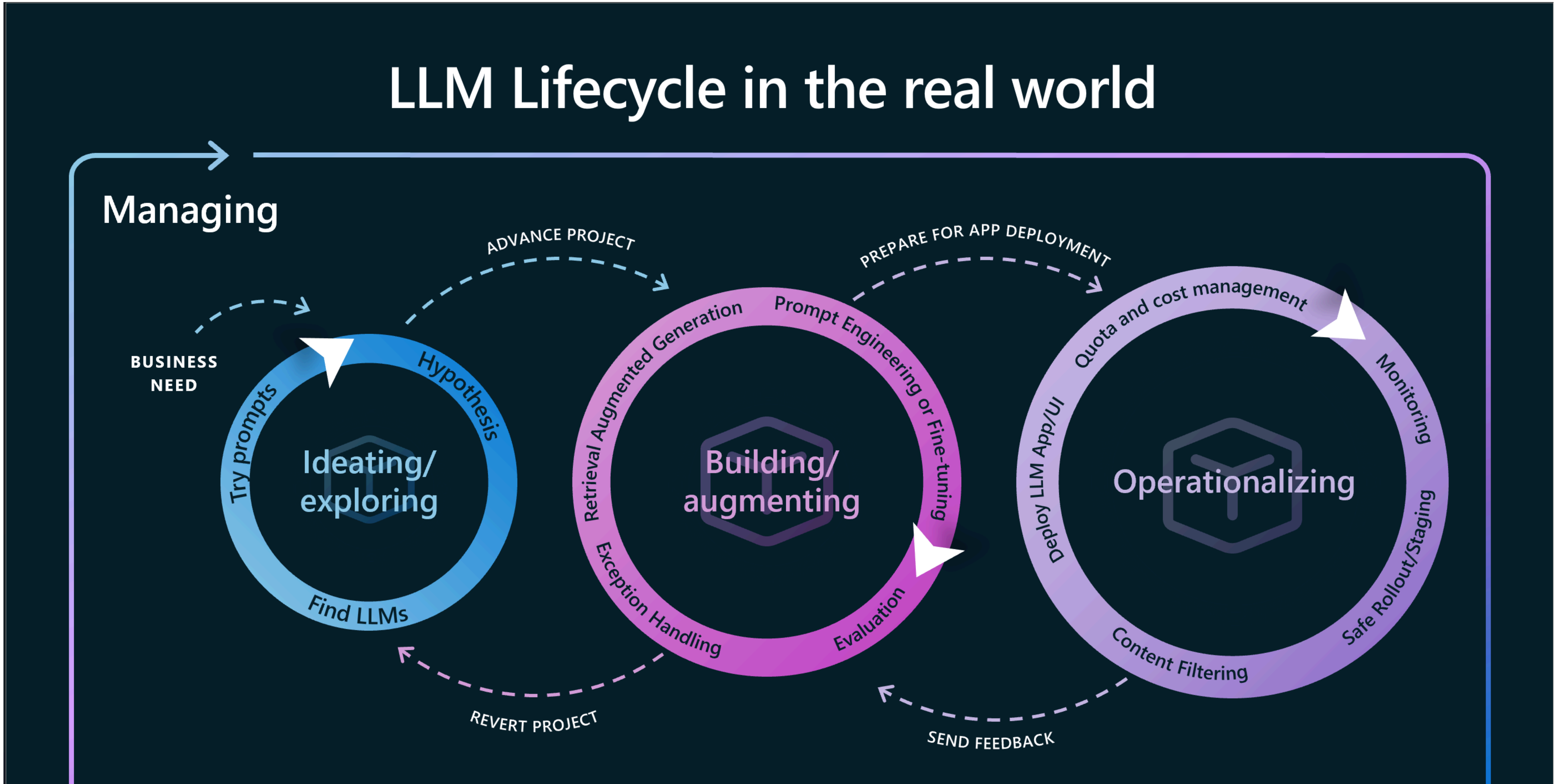
The paradigm shift—from MLOps to LLMOps

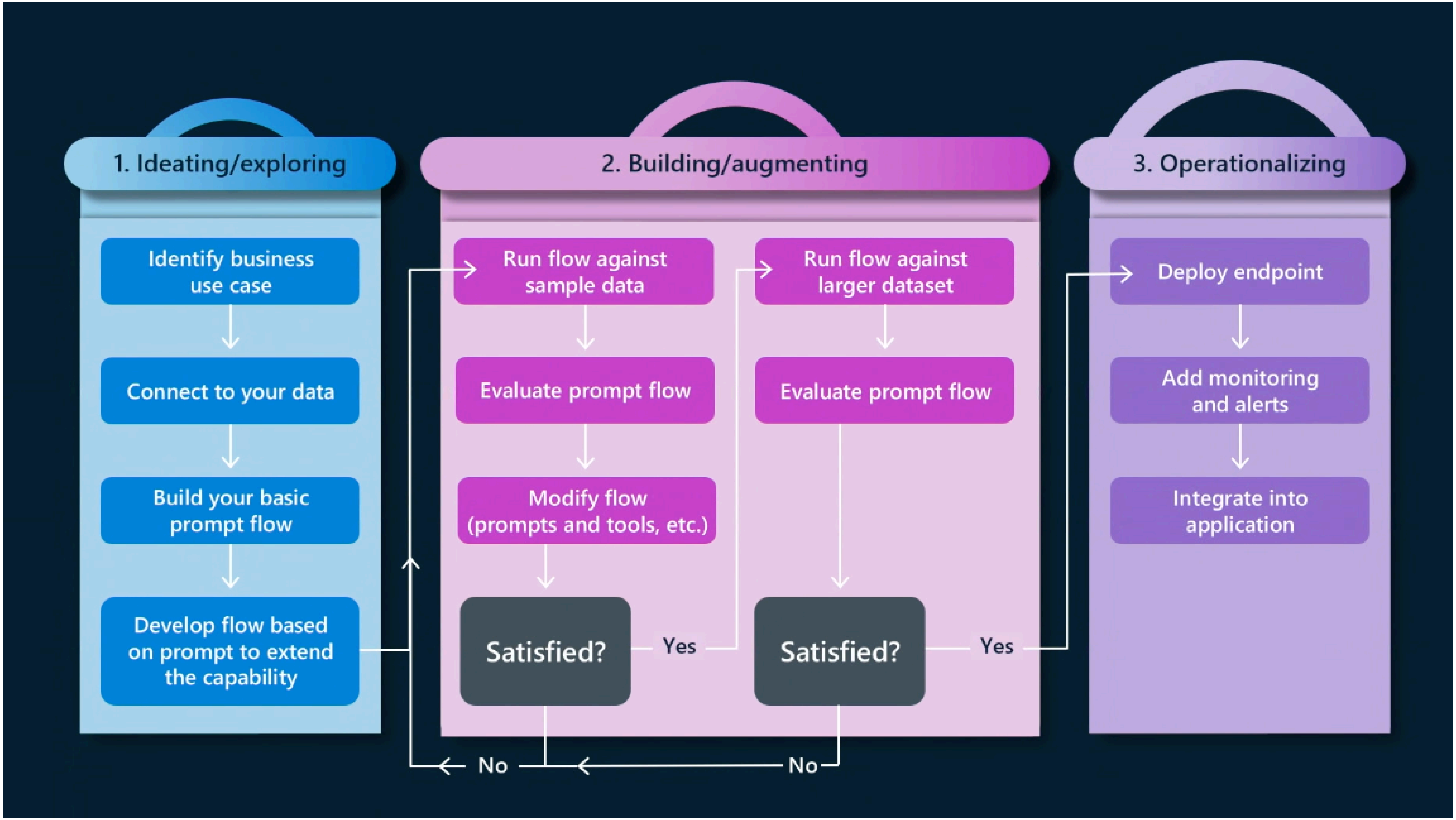
	Traditional MLOps	LLMOps
Audiences	ML Engineers Data Scientists	ML Engineers App developers
Assets to share	Model, data, environments, features	LLMs, agents, plugins, prompts, chains, APIs
Metrics/evaluations	Accuracy	**Quality:** accuracy, similarity **Harm:** bias, toxicity **Honest:** groundness **Cost:** token per request **Latency:** response time, RPS
ML models	Build from scratch	Pre-built, fine-tuned served as API (MaaS)

Notice that in LLMOps, we are more focused on the App Developers, using integrations as a key point, using "Models-as-a-Service" and thinking in the following points for metrics.

- Quality: Response quality
- Harm: Responsible AI
- Honesty: Response groundedness (Makes sense? It is correct?)
- Cost: Solution Budget
- Latency: Avg. time for token response

The LLM Lifecycle





1. Ideating/exploring

- Identify business use case
- Connect to your data
- Build your basic prompt flow
- Develop flow based on prompt to extend the capability

2. Building/augmenting

- Run flow against sample data
- Evaluate prompt flow
- Modify flow (prompts and tools, etc.)
- Satisfied?
- Run flow against larger dataset
- Evaluate prompt flow
- Satisfied?

3. Operationalizing

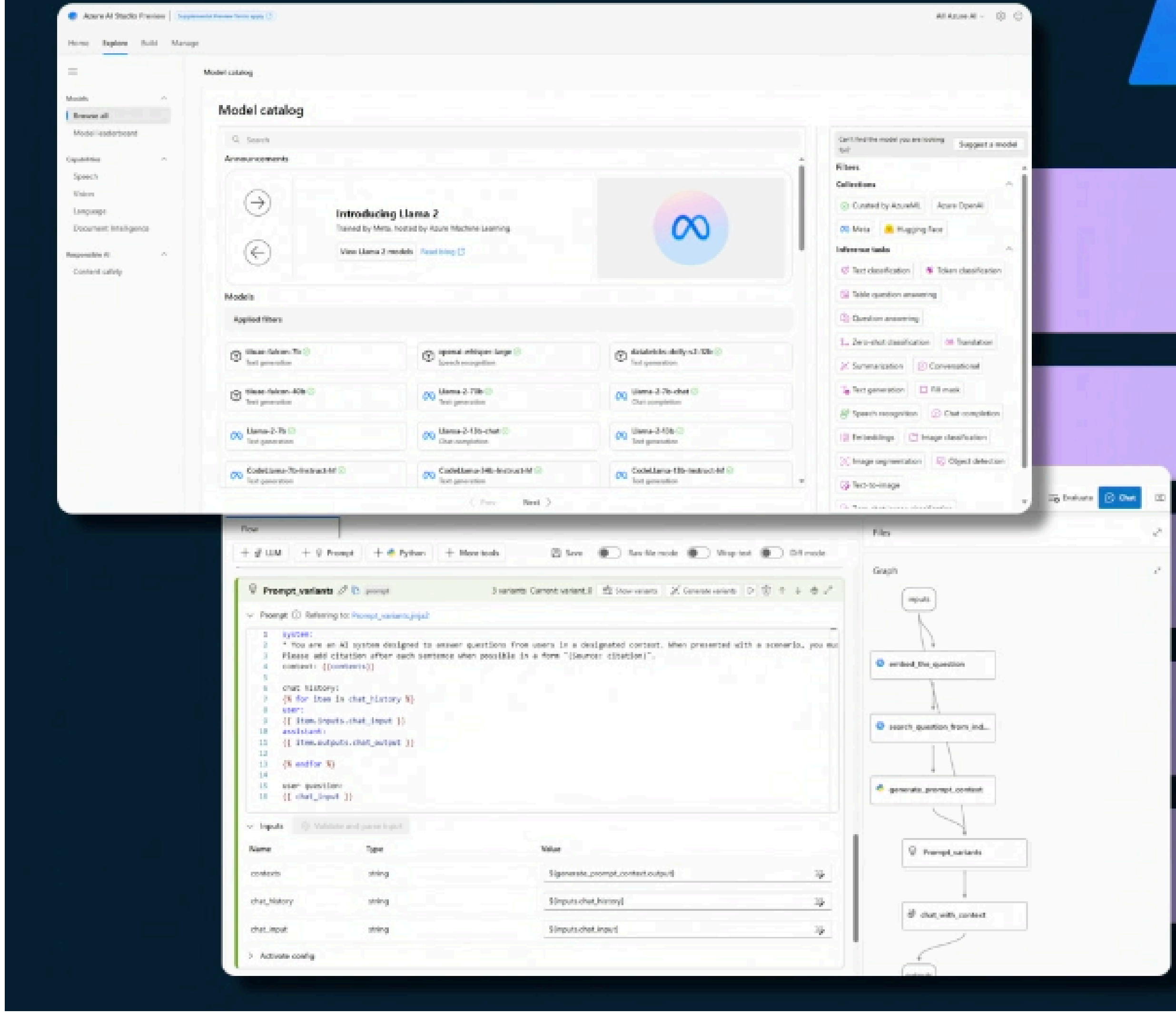
- Deploy endpoint
- Add monitoring and alerts
- Integrate into application

Lifecycle Tooling

For Tooling, Microsoft provides the Azure AI Platform and PromptFlow facilitate and make your cycle easy to implement and ready to go.

The Azure AI Platform, allows you to use AI Studio. AI Studio is a web portal allows you to Explore models, samples and tools. Managing your resources, UI development flows and SDK/CLI options for Code-First development.

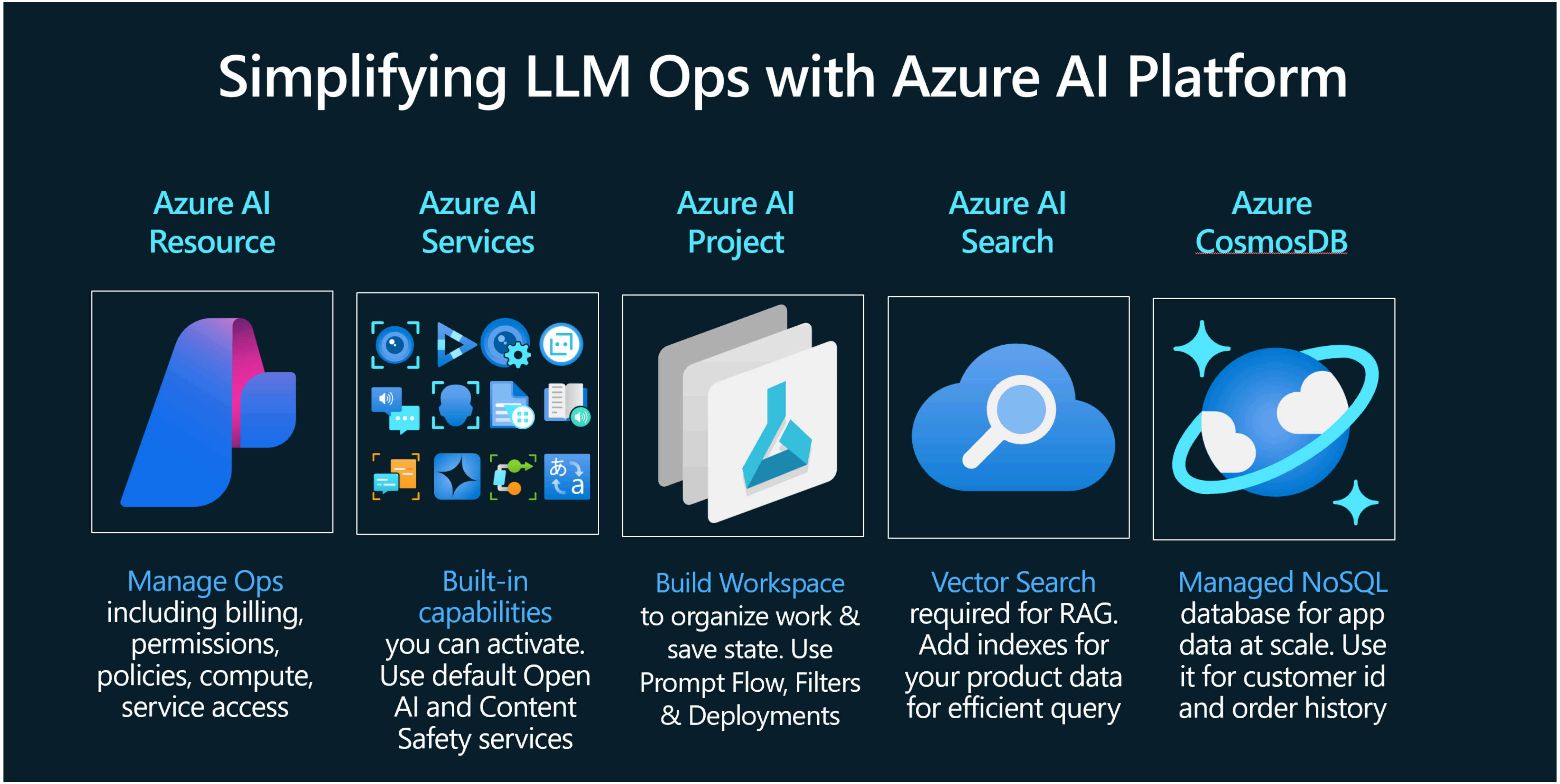
Azure AI is a platform for Generative AI



```
system:
* You are an AI system designed to answer questions from users in a designated context. When presented with a scenario, you mu
Please add citation after each sentence when possible in a form "(Source: citation)".
context: {{contexts}}

chat history:
{% for item in chat_history %}
user:
{{ item.inputs.chat_input }}
assistant:
{{ item.outputs.chat_output }}

{% endfor %}

user question:
{{ chat_input }}
```

Name	Type	Value
contexts	string	${generate_prompt_context.output}
chat_history	string	${inputs.chat_history}
chat_input	string	${inputs.chat_input}

- Access to thousands of LLMs from OpenAI, Meta, Hugging Face
- Data grounding with RAG
- Prompt engineering/evaluation
- Built-in safety and responsible AI
- Continuous monitoring for LLMs



Simplifying LLM Ops with Azure AI Platform

Azure AI Resource

Manage Ops including billing, permissions, policies, compute, service access

Azure AI Services

Built-in capabilities you can activate. Use default Open AI and Content Safety services

Azure AI Project

Build Workspace to organize work & save state. Use Prompt Flow, Filters & Deployments

Azure AI Search

Vector Search required for RAG. Add indexes for your product data for efficient query

Azure CosmosDB

Managed NoSQL database for app data at scale. Use it for customer id and order history

Construct, from Proof-of-Concept(POC) until large scale applications with PromptFlow:

Streamlining LLM App Dev with Prompt Flow

VS Code Extension

Design & Build your LLM App as a DAG with inputs, nodes, outputs

Visual DAG Functions

Extend & Run your prompt flow with visual & function tools

Evaluation Metrics

Test & Tune your prompt flow for quality and responsible AI

Cloud Deployment

Push & Deploy your prompt flow to Azure for app integrations

Azure AI Studio UI

Integrate & Iterate with cloud hosted runtime & LLM app endpoint