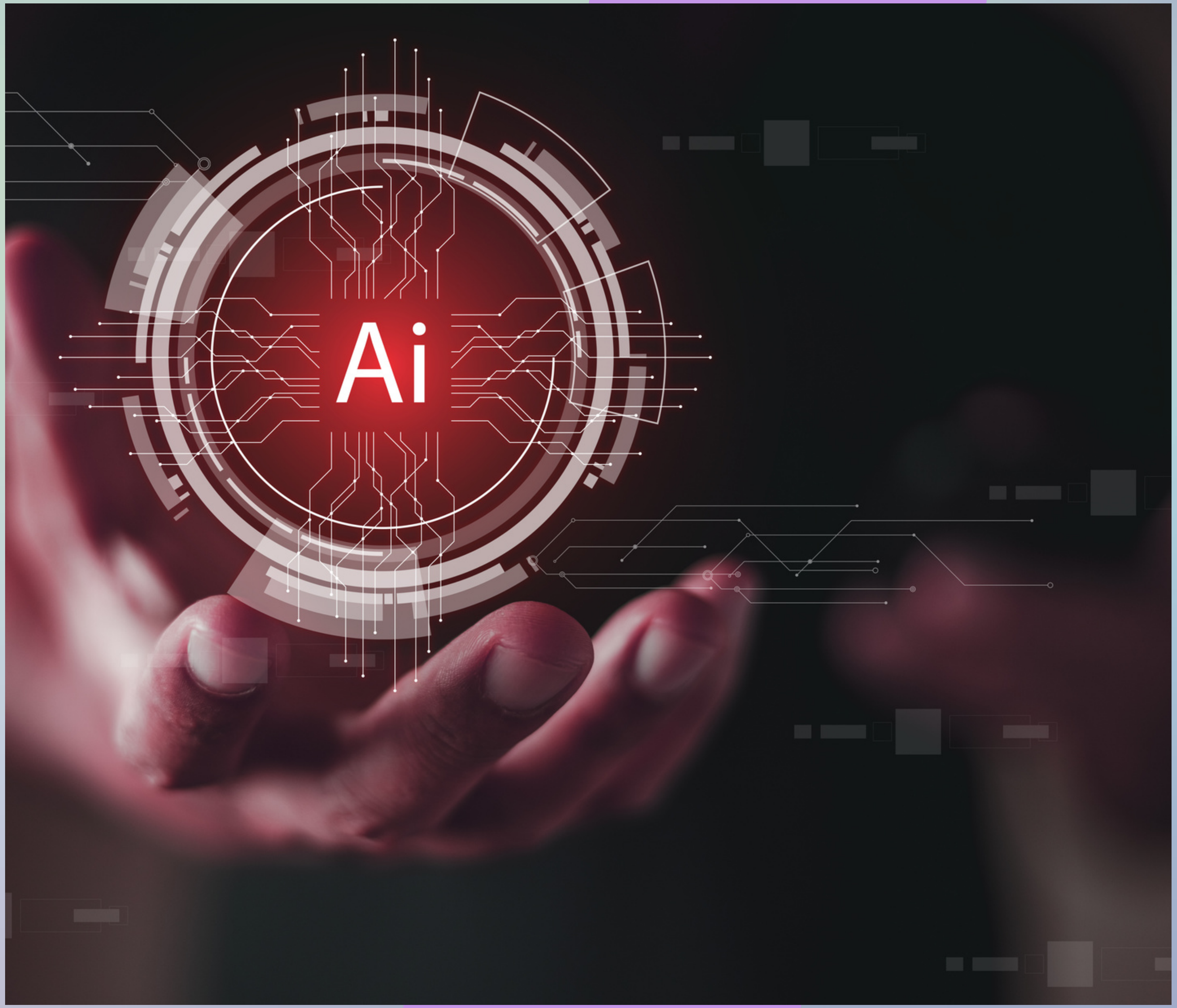
LLMOPs

LLMs in Production

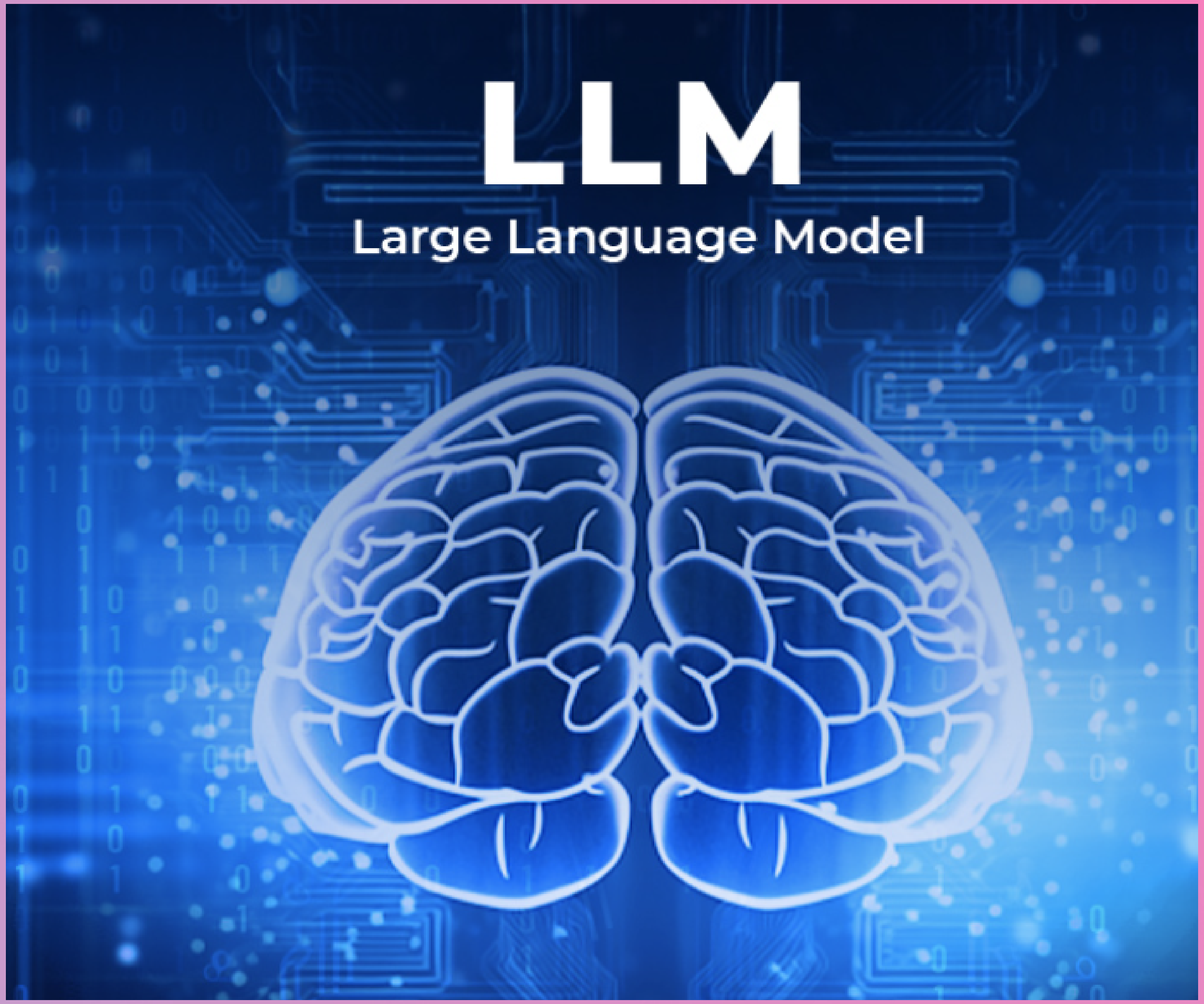
SHREYAS PUTTARAJU



Introduction

Large Language Model Ops (LLMOps) refers to the practices, techniques, and tools used for managing large language models in production environments.

The latest advancements in LLMs, highlighted by releases such as OpenAI's GPT, Google's Bard, and Databricks' Dolly, are driving considerable growth in enterprises building and deploying LLMs. As a result, there is a need to establish best practices for operationalizing these models.



LLMOPs World

Providers	Tools	Frameworks	Infrastructure
OpenAI	Pinecone	Hugging Face	Azure
Meta	mlflow™	LangChain	databricks
Google	Snorkel	LlamaIndex	Vertex.ai

The Importance of LLMops Cannot Be Overstated for Success!

LLM Challenges

Essential Factors for Training Your LLMs:

- Facilitates tailor-made solutions yet demands more effort in training.
- Necessitates a substantial volume of premium data.
- Consumes a considerable amount of processing power.
- Calls for sophisticated tools for practical application.
- Leads to cost-efficiency and enhanced response times.
- Enhances confidentiality, security, and processing speed.

What to consider for your LLM projects?

A strategical way to select right tools and services

Selecting the right LLM will depend on:

- Cost, speed, and quality
- Model size
- Training data
- Application

Other key considerations:

- Optimizing for convenience allows using LLM providers (e.g., ChatGPT or Claude)
- Customizability may require using open-source base models
- Privacy may require training internal custom models and using your own data

QUALITY

GPT-4, Gemini, Llama-2

SPEED

GPT-3.5, Claude-Instant

OPEN SOURCE

Llama-2, Google Flan-T5

Choosing the Right Tools for LLM Operations

Pinecone: Select Pinecone for vector database services, which are essential for similarity search in large-scale LLM applications.

MLflow: Ideal for experiment tracking, model versioning, and deployment, MLflow streamlines the machine learning lifecycle, particularly for teams.

Snorkel: Use Snorkel if you require a tool to rapidly create, model, and manage training data, enabling weak supervision for faster LLM training.

Frameworks to Enhance LLM Development

🤗 Hugging Face

Hugging Face: A go-to choice for pre-trained models and datasets, Hugging Face accelerates development with its extensive transformer-based model repository.

🦜🔗 LangChain

LangChain: Leverage LangChain for building applications that chain LLM functionalities together, simplifying the process of creating complex AI-powered solutions.

LlamaIndex

LlamaIndex: Consider LlamaIndex when you need a specialized search interface for LLMs, as it provides enhanced capabilities for indexing and retrieving information.

Infrastructure to Power Your LLMs

Azure

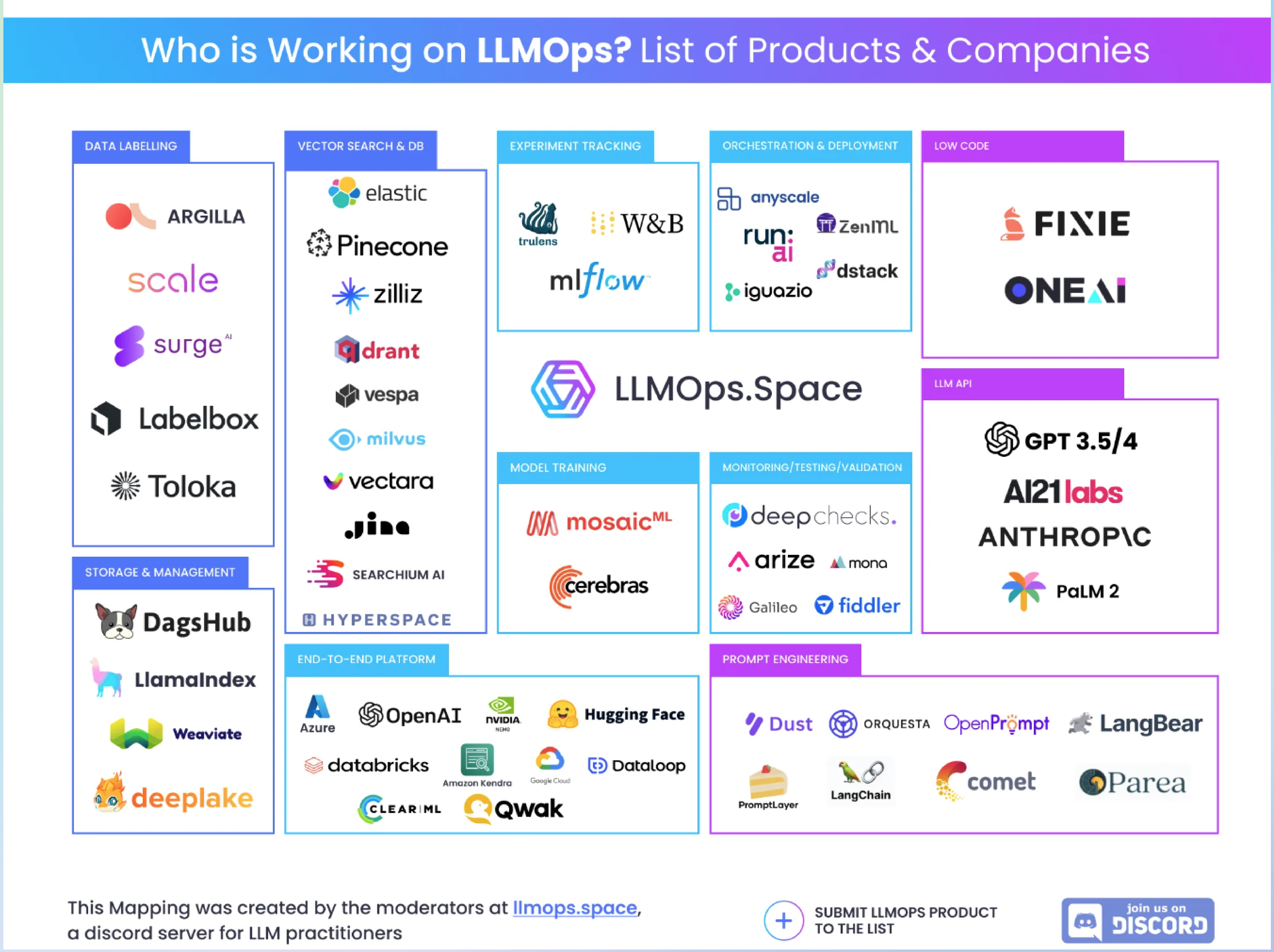
Azure: Microsoft's Azure offers robust cloud computing services with AI-specific infrastructure that's scalable for training and deploying LLMs.

databricks

Databricks: For a unified analytics platform that facilitates collaboration between data scientists and engineers, Databricks is a top contender, especially for Spark-based environments.

Vertex.ai

Vertex.ai: Google Cloud's Vertex.ai provides an end-to-end managed machine learning platform, perfect for deploying and maintaining LLMs with minimal effort.



Who is Working on **LLMOps?** List of Products & Companies

DATA LABELLING
- ARGILLA
- scale
- surge AI
- Labelbox
- Toloka

VECTOR SEARCH & DB
- elastic
- Pinecone
- zilliz
- qdrant
- vespa
- milvus
- vectara
- jina
- SEARCHIUM AI
- HYPERSPACE

EXPERIMENT TRACKING
- trulens
- W&B
- mlflow

ORCHESTRATION & DEPLOYMENT
- anyscale
- run:ai
- ZenML
- dstack
- iguazio

LOW CODE
- FIXIE
- ONE AI

LLMOps.Space

MODEL TRAINING
- mosaicML
- cerebras

MONITORING/TESTING/VALIDATION
- deepchecks.
- arize
- mona
- Galileo
- fiddler

LLM API
- GPT 3.5/4
- AI21 labs
- ANTHROP\C
- PaLM 2

STORAGE & MANAGEMENT
- DagsHub
- LlamaIndex
- Weaviate
- deeplake

END-TO-END PLATFORM
- Azure
- OpenAI
- NVIDIA NeMo
- Hugging Face
- databricks
- Amazon Kendra
- Google Cloud
- Dataloop
- CLEAR ML
- Qwak

PROMPT ENGINEERING
- Dust
- ORQUESTA
- OpenPrompt
- LangBear
- PromptLayer
- LangChain
- comet
- Parea

This Mapping was created by the moderators at llmops.space, a discord server for LLM practitioners

+ SUBMIT LLMOPS PRODUCT TO THE LIST

join us on DISCORD

LLM Ecosystem: A Comprehensive Overview

- **Data Labeling:** Essential for training accurate models. Tools like ARGILLA and Labelbox allow teams to annotate datasets efficiently.
- **Vector Search & DB:** Services like Pinecone and Jina offer scalable solutions for similarity search in large datasets, a cornerstone for responsive LLM applications.
- **Experiment Tracking:** Tools such as MLflow and Weights & Biases (W&B) provide robust platforms for tracking ML experiments, crucial for iterative improvement.
- **Orchestration & Deployment:** Technologies like Anyscale and ZenML streamline the deployment of machine learning models, enabling scalable and manageable operations.
- **Low Code Platforms:** FIXIE and ONE AI offer user-friendly interfaces for deploying AI without extensive coding, democratizing access to LLM technologies.

LLM Ecosystem: A Comprehensive Overview

- **Storage & Management:** Solutions like Weaviate and DeepLake manage and store large volumes of data, ensuring quick retrieval and efficient handling.
- **Model Training:** Cerebras and MosaicML provide advanced systems and frameworks for efficient and scalable model training.
- **Monitoring/Testing/Validation:** Tools like Deepchecks and Arize AI are crucial for maintaining model quality and performance.
- **LLM API:** GPT-3.5/4 and others offer powerful APIs for easy integration of LLM capabilities into various applications.
- **End-to-End Platforms:** Azure and Databricks provide comprehensive platforms that support the entire machine learning workflow from development to deployment.
- **Prompt Engineering:** LangChain and PromptLayer specialize in crafting prompts that maximize the performance of LLMs in specific tasks.

Summary: Navigating the World of LLMOps

- LLMOps encapsulates the ecosystem required to build, deploy, and maintain Large Language Models effectively.
- It encompasses a variety of tools and platforms, each serving distinct but interconnected functions, from data labeling to model deployment.
- Key to LLMOps is the integration of these tools to streamline the entire lifecycle of an LLM, ensuring efficiency and scalability.
- The goal of LLMOps is to provide a structured approach to managing the complexities of LLMs while optimizing for performance and cost.
- As the field grows, LLMOps continues to evolve, focusing on automation, best practices, and the development of more sophisticated operational frameworks.

Thank you