ID: 22K-4080
Sec: Bai-7B

1) MLOps is the practice of managing the entire machine learning lifecycle with automation, reliability, and repeatability. It solves the problem of messy, inconsistent workflows by standardizing how data is prepared, models are trained, deployed, versioned, and monitored in production. The goal is to keep models stable, traceable, and easy to update as conditions change.

2) LLMOps focuses on building, deploying, and maintaining large language models, which behave differently from traditional ML systems. The difference is that LLMs rely more on prompts, context, embeddings, retrieval systems, and safety constraints rather than fixed numerical features and training pipelines. LLMOps must manage dynamic behavior, hallucination risks, and rapid updates to prompts or model configurations.

3) Three challenges unique to LLMOps include controlling hallucinations, evaluating outputs that have no single correct answer, and managing prompt versions or retrieval databases that heavily influence outputs. These issues matter far more in LLMs compared to standard ML

models.

4) Continuous monitoring means checking model performance, drift, latency, and reliability after deployment so issues get caught early. For LLmOps, an example would be tracking hallucination rates or monitoring how often the chatbot gives unsafe or off-policy responses.

5) Scaling an LLm requires distributing large parameter weights across multiple GPUs or servers and handling heavy memory and inference loads. In contrast, scaling a simple classifier usually just means hosting more lightweight instances behind a load balancer because the compute footprint is small.

Scenario answers:

1- They can reuse data pipelines, CI/CD workflows, model versioning, API deployment processes, logging infrastructure, and monitoring tools for uptime and latency. The general operational backbone of their mLOps system remains useful.

2- They need new capabilities like prompt management, retrieval augmentation, LLm-specific evaluation methods, safety and bias filters, output moderation, and dynamic prompt or system message versioning. They may also need vector databases and tools to evaluate the quality of generated text.

3- The monitoring loop should track biased or incorrect responses using automated detectors and human review samples. Any flagged outputs feed into a feedback system that updates prompts, improves retrieval content, or adjusts safety rules. Logs from real user interactions are periodically reviewed, evaluated, and used to trigger prompt revisions or model updates while keeping all changes versioned.