Introduction to MLOps and LLMOps

Disclaimer: *This text was generated for demonstration purposes and is used solely to showcase the functionality of the MLOps-LLMOps project.*

MLOps (Machine Learning Operations)

MLOps is a set of practices that combines machine learning, DevOps, and data engineering to automate and streamline the deployment, monitoring, and maintenance of machine learning models in production. It focuses on the entire lifecycle of models, including data preprocessing, training, validation, deployment, monitoring, and retraining.

Key principles of MLOps include reproducibility, scalability, automation, and collaboration between data scientists and operations teams. Tools commonly used in MLOps pipelines include Docker, Kubernetes, CI/CD platforms, MLflow, and monitoring tools like Prometheus and Grafana.

LLMOps (Large Language Model Operations)

LLMOps is an emerging subset of MLOps specifically for large language models (LLMs), such as GPT, LLaMA, or Ollama models. It involves practices for deploying, fine-tuning, monitoring, and optimizing LLMs in production.

Key challenges in LLMOps include prompt management, latency optimization, cost control, and responsible AI practices. Observability is crucial — metrics such as prompt counts, token usage, and inference latency are often tracked. LLMOps also emphasizes retrieval-augmented generation (RAG), where external knowledge sources are used to enhance model responses.

RAG Pipelines

A RAG system combines vector databases, embeddings, and LLM inference to provide context-aware responses. The pipeline typically:

- Ingests text documents (like PDFs) and generates embeddings.
- Stores embeddings in a vector database such as ChromaDB.
- Queries the database with user prompts to retrieve relevant chunks.
- Combines the retrieved chunks with the prompt and sends it to the LLM.

RAG improves the accuracy and relevance of LLM responses while keeping the system memory-efficient and scalable.

Benefits of MLOps and LLMOps

- Ensures reproducibility and consistency across deployments.
- Improves collaboration between data scientists, engineers, and operations.
- Enables continuous monitoring and retraining of models.
- Supports responsible AI practices and mitigates risks in production environments.