MLOps Engineer

AI / ML | SMB | Enterprise

[Company Name] is hiring an MLOps Engineer to build and maintain the infrastructure that takes machine learning models from research to reliable production systems. You will design training pipelines, model serving infrastructure, and monitoring systems that enable our ML team to iterate quickly and deploy models with confidence. This role is critical to bridging the gap between data science experimentation and production-grade AI systems.

Key Responsibilities

- Design, build, and maintain ML training and deployment pipelines that are reproducible and scalable
- Implement model serving infrastructure for real-time and batch inference workloads
- Build monitoring and alerting systems for model performance, data drift, and system health
- Manage ML experiment tracking, model versioning, and artifact storage
- Automate the end-to-end ML lifecycle from data ingestion through model deployment
- Collaborate with data scientists and ML engineers to productionize their models efficiently
- Optimize compute costs and resource utilization for GPU/CPU training and inference workloads

Required Skills & Experience

- 3+ years of experience in MLOps, DevOps, or infrastructure engineering with ML focus
- Strong Python skills and experience with ML frameworks (PyTorch, TensorFlow, or scikit-learn)
- Hands-on experience with ML pipeline orchestration tools (Kubeflow, Airflow, Vertex AI Pipelines, or SageMaker Pipelines)
- Experience with containerization (Docker) and orchestration (Kubernetes) for ML workloads
- Familiarity with cloud ML services on AWS (SageMaker), GCP (Vertex AI), or Azure (Azure ML)
- Experience with model serving frameworks (TorchServe, TensorFlow Serving, Triton, or BentoML)
- Understanding of CI/CD principles applied to ML systems
- Experience with experiment tracking and model registry tools (MLflow, Weights & Biases, or Neptune)

Nice-to-Have

- Experience with GPU cluster management and distributed training
- Familiarity with feature stores (Feast, Tecton, or Hopsworks)
- Knowledge of data versioning tools (DVC, LakeFS)
- Experience with LLM serving and optimization (vLLM, TensorRT-LLM)
- Infrastructure-as-code experience (Terraform, Pulumi)

Tech Stack

Python | Kubernetes | Docker | Kubeflow | MLflow | AWS SageMaker | Terraform | Airflow | Prometheus | Grafana

What We Offer

- Competitive salary and equity package
- Flexible remote or hybrid work arrangement
- Health, dental, and vision insurance
- Annual learning and development budget
- Generous PTO policy

Interview Process

1. Recruiter phone screen (30 min)
2. Technical phone screen covering infrastructure and ML pipeline fundamentals (45 min)
3. System design exercise: design an end-to-end ML deployment pipeline (60 min)
4. Hands-on coding round: infrastructure or pipeline automation task (60 min)
5. Culture fit and team interview with hiring manager and ML team members (45 min)