

Carlos Mendoza

AI Model Deployment Engineer | LLM Production Specialist

Vancouver, Canada | (604) 555-6729 | carlos.mendoza@example.com

linkedin.com/in/carlosmendoza-ai | github.com/cmendoza-llm-prod

Professional Summary

Practical AI Model Deployment Engineer with 4+ years of experience bringing research models to production environments. Specialized in creating robust, scalable, and reliable services around large language models. Strong focus on latency optimization, cost management, and implementing MLOps best practices for AI systems.

Professional Experience

Senior ML Engineer, Deployment

ProductionAI Systems, Vancouver, Canada (Aug 2021 - Present)

- Designed and implemented production serving infrastructure for LLMs handling 1M+ daily requests.
- Reduced 95th percentile latency by 65% through optimized batching and caching strategies.
- Created automated monitoring system detecting 90% of model degradation issues before user impact.
- Developed CI/CD pipelines for continuous model updates with automated canary testing.

Machine Learning Engineer

AI Solutions Inc., Toronto, Canada (Mar 2019 - Jul 2021)

- Built API services and middleware for deploying NLP models to production.
- Implemented dynamic batching systems improving inference throughput by 40%.
- Created automated testing frameworks for evaluating model performance in production.
- Developed data logging and feedback loops for continuous model improvement.

Software Engineer

TechCloud Services, Seattle, WA (Jun 2017 - Feb 2019)

- Developed and maintained cloud-based microservices.
- Implemented performance monitoring and alerting systems.
- Created deployment automation for web services.

Technical Skills

- **Deployment Technologies:** Docker, Kubernetes, AWS Lambda, Azure Functions
 - **MLOps:** MLflow, Kubeflow, Weights & Biases, DVC
 - **Monitoring & Observability:** Prometheus, Grafana, ELK Stack, Datadog
 - **Programming Languages:** Python, Go, JavaScript
 - **ML Serving:** TorchServe, TensorRT, ONNX Runtime, vLLM, FastAPI
 - **Cloud Platforms:** AWS, GCP, Azure
 - **Testing & CI/CD:** GitHub Actions, Jenkins, PyTest, Locust
-

Education

Bachelor of Science, Computer Science

University of British Columbia, Vancouver, Canada (Graduated: May 2017)

- Specialization in Software Engineering - Capstone Project: “Automated Deployment Systems for ML Models”

Certifications

- AWS Certified DevOps Engineer - Professional (2023)
 - Google Cloud Professional Machine Learning Engineer (2022)
 - Certified Kubernetes Administrator (CKA) (2021)
 - Azure AI Engineer Associate (2020)
-

Project Portfolio

- **ScalableLLM Platform:** End-to-end system for deploying, serving, and monitoring LLMs
 - **Inference Optimization Toolkit:** Library of techniques for optimizing LLM inference
 - **ModelMonitor:** Comprehensive monitoring system for ML models in production
-

Conference Presentations

- Speaker at MLOps Summit 2023: “Practical Approaches to LLM Deployment”
 - Workshop Leader at DevOps Days Vancouver 2022: “CI/CD for Machine Learning Models”
-

Open Source Contributions

- Contributor to TorchServe, focusing on batching optimizations
 - Maintainer of “LLM-Deployment-Toolkit” - open-source templates for LLM serving
-

Languages: English (fluent), Spanish (native), Portuguese (conversational)