

LLM Deployment to Cloud - Comprehensive DevOps Guide

For: Beginners with DevOps background but no ML experience

Table of Contents

This guide provides everything you need to know about deploying Large Language Models (LLMs) to AWS, Azure, and GCP as a DevOps engineer.

1. **Fundamentals** - Start here to understand what LLMs are and why they're deployed

- Basic concepts about LLMs
- Why deploy LLMs to cloud
- Architecture overview
- Common terminology explained

2. **AWS Deployment** - Complete guide for AWS

- AWS services for LLM deployment
- Step-by-step deployment guides
- Real-world architectures

3. **Azure Deployment** - Complete guide for Azure

- Azure services for LLM deployment
- Deployment patterns
- Integration patterns

4. **GCP Deployment** - Complete guide for GCP

- GCP services for LLM deployment
- Deployment options
- Best practices

5. **Monitoring & Observability** - Track and debug your LLM deployments

- Metrics to monitor
- Logging strategies
- Alerting setup

6. **Cost Optimization** - Keep your bill manageable

- Cost drivers explained
- Optimization strategies
- Billing strategies

7. **Best Practices** - Learn from the industry

- Security practices
- Performance optimization
- Reliability patterns

8. **Use Cases** - Real-world examples

- Chatbot deployment
- Code generation services
- Document summarization
- And more...

Quick Start Path

If you're completely new:

1. Start with [01-Fundamentals/01-LLM-Basics.md](#)
2. Read [01-Fundamentals/02-Architecture-Overview.md](#)
3. Choose your platform (AWS/Azure/GCP)
4. Follow the deployment guide for your platform

If you have cloud experience:

1. Skip fundamentals, jump to your preferred platform section
2. Reference cost optimization and monitoring as needed

Key Takeaways

- **LLMs are large AI models** that can understand and generate human-like text
- **Cloud deployment** makes them scalable, cost-effective, and accessible
- **DevOps role** is crucial for managing infrastructure, monitoring, and costs
- **No ML knowledge needed** - your cloud infrastructure expertise is enough

How to Use This Guide

Each markdown file is self-contained but references other files when needed. You can:

- Read sequentially for comprehensive understanding
- Jump to specific topics based on your needs
- Use as a reference while working on deployments
- Share specific sections with your team

Last Updated: January 2026

Target Audience: DevOps Engineers new to LLM deployment

Prerequisite: Basic understanding of cloud platforms (AWS/Azure/GCP)