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.

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)