# Production AI Systems and MLOps

## Production AI Systems

Production AI System Architecture: - Model serving and inference optimization - Load balancing and auto-scaling - Monitoring and observability - A/B testing and gradual rollouts MLOps for LLMs: - Version control for models and prompts - Continuous integration and deployment - Data drift detection and model retraining - Performance monitoring and alerting Security and Compliance: - Data privacy and protection - Model security and adversarial attacks - Compliance with regulations (GDPR, AI Act) - Ethical AI considerations and bias mitigation Cost Optimization: - Model compression and quantization - Efficient inference techniques - Resource allocation and scheduling - Cost monitoring and budget management

## Deployment and Scaling

Deployment Strategies: - Cloud deployment: AWS, Azure, GCP - Edge deployment: Mobile and IoT devices - Hybrid architectures: Cloud-edge integration - Serverless and containerized deployments Scaling Considerations: - Horizontal vs. vertical scaling - Model parallelism and distributed inference - Caching strategies for improved performance - Database optimization for vector search Reliability and Fault Tolerance: - Redundancy and failover mechanisms - Circuit breakers and rate limiting - Health checks and automated recovery - Disaster recovery and backup strategies