Cost Optimization

Monitor and optimize infrastructure costs without sacrificing performance

Right-sizing

Match instance types to actual resource usage

Auto-scaling

Scale down during low-traffic periods

Spot Instances

Use for batch training jobs (up to 90% savings)

Model Optimization

Quantization, pruning to reduce compute needs

🛠️ Tools & Platforms

- AWS Cost Explorer
- GCP Cost Management
- Kubecost
- FinOps practices