# LLMOps

---

**LLMOps includes:**

1. **Model Selection & Evaluation**
    - Choosing base models (e.g., GPT-4, LLaMA, Mistral, Claude)
    - Benchmarking for task relevance
2. **Data Management**
    - Gathering, cleaning, labeling, and versioning of training & inference datasets
3. **LLM Fine-Tuning / Instruction Tuning**
    - Full fine-tuning (adjusting all weights — resource intensive)
    - Parameter-efficient fine-tuning (e.g., LoRA, PEFT, QLoRA)
4. **Model Deployment & Inference Optimization**
    - Serving large models efficiently using quantization, batching, etc.
5. **Monitoring & Observability**
    - Tracking model behavior (latency, hallucination rates, toxicity, etc.)
6. **Retraining & Model Versioning**
    - Updating models based on new data or feedback
7. **Governance & Security**
    - Handling privacy, bias mitigation, compliance (e.g., GDPR)
8. **Integration with Applications**
    - Using LangChain, LlamaIndex, RAG pipelines, etc.