Advanced LLM Concepts - Cheat Sheet

1. LoRA (Low-Rank Adaptation)

A technique to fine-tune large models efficiently by adding small trainable matrices to frozen weights. Saves compute and memory.

2. QLoRA (Quantized LoRA)

Combines LoRA with 4-bit quantization to enable fine-tuning large models on limited hardware (e.g., consumer GPUs).

3. Adapter Layers

Task-specific layers plugged into a frozen model. Enables switching tasks without retraining the full model.

4. Mixture of Experts (MoE)

Only a subset of model parameters (experts) are activated per input, reducing computation while scaling model size.

5. RLHF (Reinforcement Learning from Human Feedback)

Technique used in models like ChatGPT to align outputs with human preferences using reinforcement learning and a reward model.

6. Alignment & Safety

Methods to make LLMs ethical, non-biased, and safe. Includes red-teaming, moderation layers, and Constitutional AI.

7. LangChain & Agents

LangChain is a framework for building LLM-powered apps. Agents use tools, memory, and multi-step reasoning to solve tasks.

8. RAG (Retrieval-Augmented Generation)

Combines LLMs with a vector database to retrieve documents and generate grounded, real-time, factual answers.

Advanced LLM Concepts - Cheat Sheet

9. Memory in Agents

Mechanism to track prior context in conversations or plans. Types include short-term (in-context) and long-term (stored).