🎓 AI Engineering & LLM Systems: Complete Study Guide (2026)

Section 1: Foundations of Artificial Intelligence

Artificial Intelligence (AI) simulates human intelligence through probabilistic systems.

- **1.1 Data Science vs. ML vs. AI:** Understanding the hierarchy of intelligent systems.
- **1.2 Neural Networks (NN):** Input, Hidden, and Output layers; Weights and Biases.
- **1.3 Activation Functions:** Non-linearity via ReLU, Sigmoid, Tanh, and Softmax.
- **1.4 The Learning Process:** Forward propagation, Loss functions, and Backpropagation.

Section 2: Mathematical Essentials for AI

The "engine" under the hood of every model.

- **2.1 Linear Algebra:** Tensors, Matrix multiplication, Eigenvalues, and SVD (foundational for Embeddings).
- **2.2 Calculus:** Derivatives, Gradients, and Chain Rule (used in Gradient Descent).
- **2.3 Probability & Statistics:** Bayes' Theorem, Distributions, and Hypothesis Testing for reasoning under uncertainty.

Section 3: Large Language Models (LLMs) & Transformers

Modern NLP is built on the Transformer architecture (2017).

- **3.1 Transformer Architecture:** Self-Attention, Multi-Head Attention, and Positional Encodings.
- **3.2 Tokenization:** Byte-Pair Encoding (BPE) and SentencePiece.
- **3.3 Inference Strategies:** Greedy search, Beam search, Nucleus (Top-p) sampling, and Temperature.
- **3.4 Memory Management:** KV Caching and Context Window optimization.

Section 4: Retrieval-Augmented Generation (RAG)

Connecting LLMs to live, private data sources.

- **4.1 The RAG Pipeline:** Load, Split (Chunking), Embed, Store, and Retrieve.
- **4.2 Vector Databases:** Similarity search using HNSW and IVF indexing in databases like ChromaDB or Pinecone.
- **4.3 Similarity Metrics:** Cosine Similarity vs. Euclidean Distance.
- **4.4 Advanced RAG:** Reranking, Hybrid search, and Query transformation.

Section 5: AI Agents & Tool Use

Autonomous systems that plan, act, and use external tools.

- **5.1 Reasoning Patterns:** Chain-of-Thought (CoT) and the **ReAct** (Reason + Act) loop.
- **5.2 Agent Architectures:** Tool-calling, Memory management, and Multi-step planning.
- **5.3 Human-in-the-Loop:** Designing escalation paths for high-stakes tasks.

Section 6: Advanced Model Optimization

Fine-tuning models for specific domain expertise.

- **6.1 Fine-Tuning Methods:** Supervised Fine-Tuning (SFT) vs. Instruction Tuning.
- **6.2 Parameter-Efficient Fine-Tuning (PEFT):** LoRA (Low-Rank Adaptation) and QLoRA.
- **6.3 Alignment:** Reinforcement Learning from Human Feedback (RLHF) and DPO (Direct Policy Optimization).

Section 7: AI Evaluation & MLOps

Moving from "vibe coding" to production-grade engineering.

- **7.1 Evaluation Frameworks:** Benchmarking with MMLU and RAGAS (Faithfulness & Relevancy).
- **7.2 Deployment & Scaling:** FastAPI, Docker, and Model Quantization for faster inference.
- **7.3 Monitoring:** Drift detection, Latency tracking, and Cost management.

Section 8: Ethics, Safety & Governance

Responsible AI development.

- **8.1 Hallucinations & Bias:** Detecting and mitigating systemic errors in training data.
- **8.2 Safety Guardrails:** Input/Output filtering and adversarial testing (Red Teaming).
- **8.3 Regulation:** Auditability and compliance in the 2026 AI landscape.
-