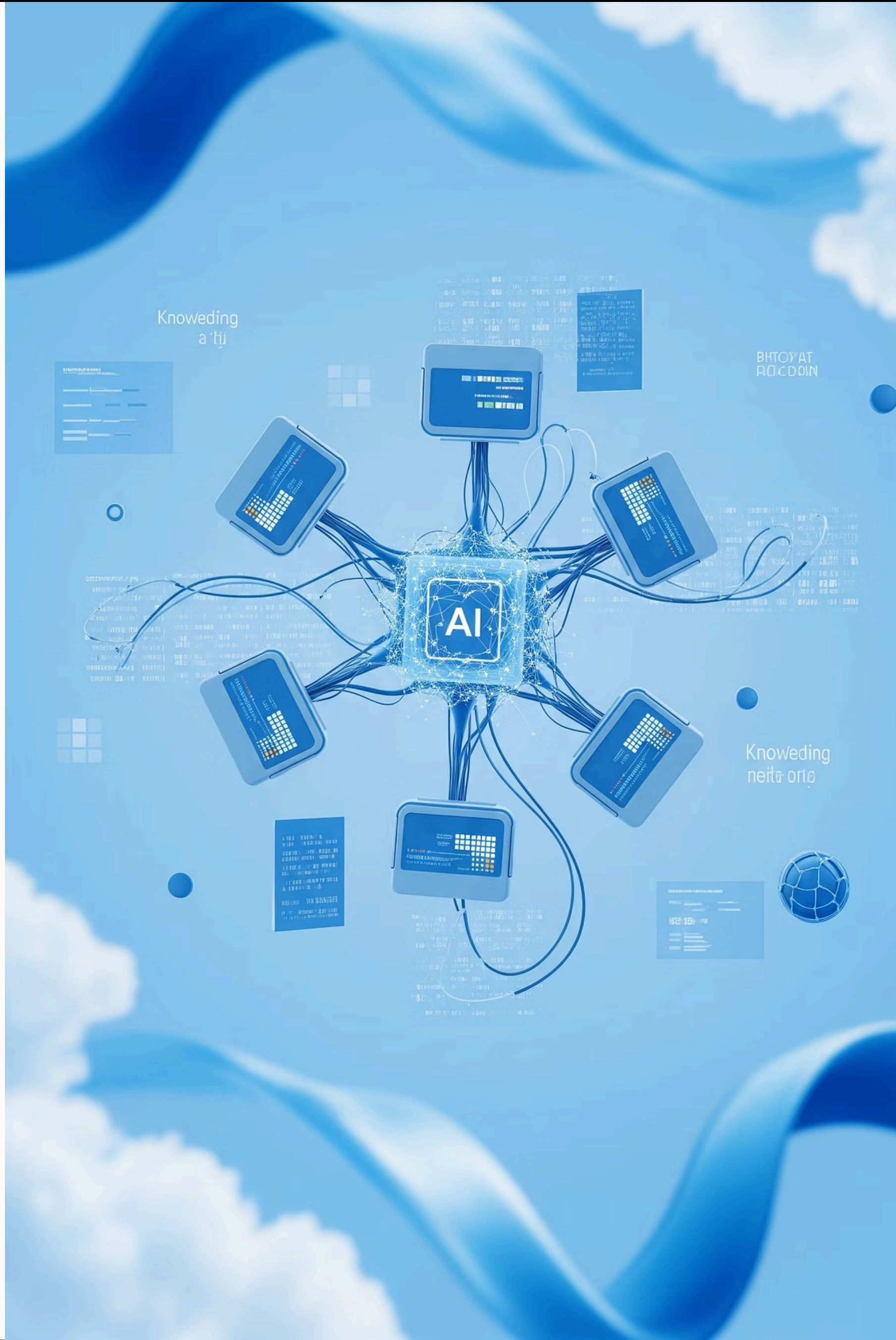
Day 9 – Fine-Tuning, Adapters & RAG in Agents

School of AI – 52-Week Agentic AI Master Program

Welcome to Day 9 of our program, where we'll explore how to enhance base LLMs for specialized agent functionality through three powerful approaches.



Why Extend Base LLMs?

Base language models provide impressive general capabilities, but agentic systems often require specialized knowledge and behaviors to perform effectively in specific domains.

Limitations of Base Models

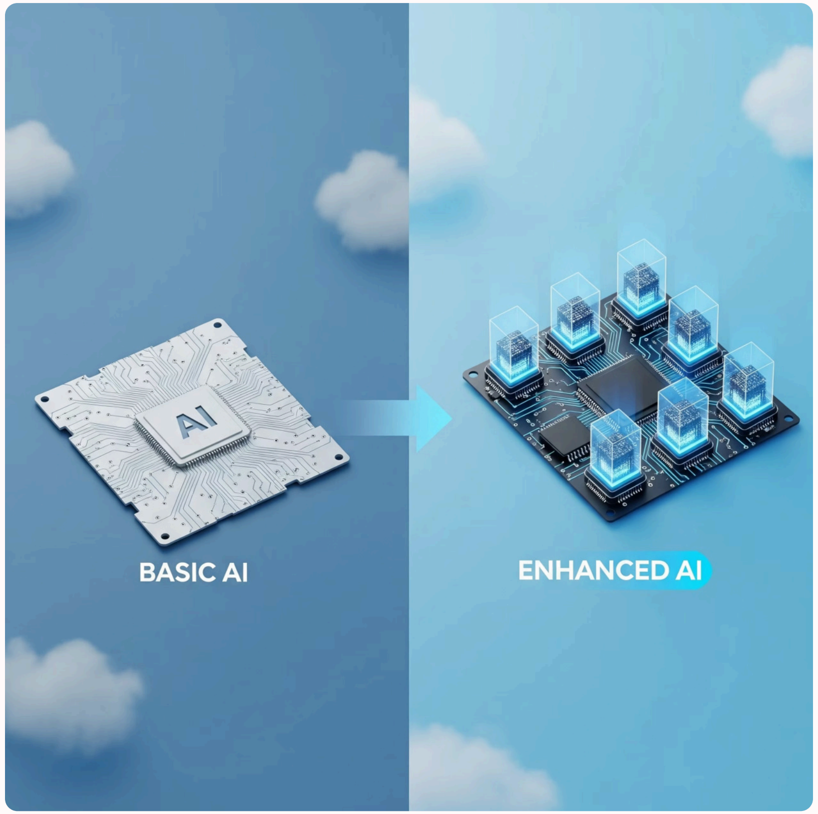
Generic training data means limited expertise in specialized domains

Domain-Specific Needs

Agents require targeted knowledge and behaviors for specific tasks

Performance Gains

Customized models show significant improvements in accuracy and reliability



Fine-Tuning Models

What is Fine-Tuning?

The process of further training a pre-trained LLM on a curated dataset to enhance its capabilities in specific domains or tasks.

Best for: Jargon-heavy industries like law, medicine, finance

Resource needs: Substantial computing power, large datasets

Updates: Requires complete retraining for new information

ⓘ **Agent Example:** A legal assistant agent fine-tuned on case law and regulations can draft specialized documents with proper terminology and citation formats.

Lightweight Adapters

Parameter-Efficient Fine-Tuning (PEFT)

Efficiency

Train only small adapter layers (0.1-1% of parameters) while keeping base model frozen

Cost-Effective

Requires fraction of compute resources compared to full fine-tuning

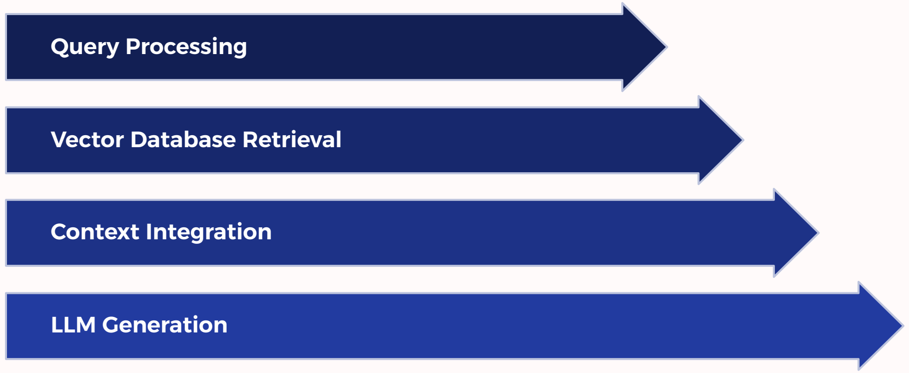
Flexibility

Can swap different adapters for various tasks using same base model

LoRA (Low-Rank Adaptation) is particularly popular for agent development, allowing quick iterations and specialized behavior adjustment without retraining the entire model.

Retrieval-Augmented Generation (RAG)

RAG connects language models to external knowledge sources, enabling agents to access up-to-date information without retraining.



Query Processing

Vector Database Retrieval

Context Integration

LLM Generation

Key Benefits for Agents:

- Dynamically incorporates fresh information
- Reduces hallucinations with evidence-based responses
- Scales knowledge without model size increases
- Enables verifiable citations and sources

Choosing the Right Approach

Fine-Tuning

When to use: Deep domain specialization needed

Trade-offs: Higher cost, less flexibility, better performance

Example: Medical diagnosis agent requiring comprehensive terminology

Adapters

When to use: Quick customization with limited resources

Trade-offs: Balanced approach, good for rapid iteration

Example: Tuning agent personality or instruction-following

RAG

When to use: Dynamic knowledge needs, frequent updates

Trade-offs: Retrieval latency, but always current

Example: Customer support agent needing latest product specs

> **Most effective agents combine multiple approaches**

Consider your agent's specific requirements around domain expertise, update frequency, resource constraints, and performance needs when designing your extension strategy.