Session 8 – Agentic RAG

Learning Objectives

- Understand how RAG and agents combine into Agentic RAG.
- Learn the workflow: Plan → Retrieve → Act → Reflect → Answer.
- See how an orchestrator routes queries between LLM, RAG, and MCP tools.
- Explore enterprise scenarios where Agentic RAG adds value.
- Run a live demo with RAG + MCP integration.

Beginner-Level Explanation

- RAG (Retrieval-Augmented Generation) lets AI look up knowledge from documents.
- Agents let AI act by calling tools like databases, file systems, or APIs.
- Agentic RAG combines both: the AI not only finds information but also uses tools.
- For example, it can retrieve a policy from documents and then grep a repo for matching code.
- This makes the AI more useful in real enterprise workflows.

Expert-Level Explanation

- Agentic RAG extends the standard RAG pipeline by adding agent reasoning loops.
- Instead of always retrieving text, the agent decides whether to call a retriever or another tool.
- The workflow is typically Plan → Retrieve (RAG) → Act (MCP tool) → Reflect → Answer.
- This requires orchestration: routing, error handling, synthesis of multiple results.
- Challenges include retrieval accuracy, latency, and balancing tool use with cost/security.

Demo Flow

- 1. User asks: 'Summarize onboarding policy' → routed to RAG service.
- 2. RAG retrieves top chunks from stored docs.
- 3. Orchestrator synthesizes answer from snippets.
- 4. User asks: 'Search for login in repo' → routed to MCP grep_text tool.
- 5. Orchestrator shows trace of plan, tool call, and final answer.
- 6. General question (e.g., 'Capital of France') → fallback to LLM-only response.

Enterprise Implications

- Agentic RAG enables AI copilots that combine knowledge retrieval with system actions.
- Example: GitLab MR reviewer that retrieves code context and greps for patterns.
- Example: Merchant onboarding assistant that retrieves KYC docs and calls APIs.
- Enterprise benefit: more autonomy while retaining auditability (via MCP logs).
- Scales to multi-agent systems with specialized retrievers and toolkits.

Wrap-Up

- Agentic RAG = RAG (knowledge) + Agents (action).
- Workflow: Plan → Retrieve → Act → Reflect → Answer.
- Demo showed how orchestrator routes queries to LLM, RAG, or MCP tools.
- Sets the stage for orchestration frameworks (LangGraph, Prefect) in production.