Cloud Contraptions
TRAINING & CONSULTING

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

Mastering Generative AI: From Transformers to Agent Swarms

Class Duration

35 hours of live training delivered over 5 days.

Student Prerequisites

- Professional software development experience
- Familiarity with Python or another mainstream language
- No prior generative AI experience required

Target Audience

Software engineers, architects, and technical teams who want a practical, hands-on survey of modern generative AI: how language models work, frontier and small language models, retrieval-augmented generation, the Model Context Protocol, single and multi-agent systems, and evaluation. Relevant for teams deciding where to apply AI in their products and wanting a shared vocabulary before specializing. For deeper dives, see *Production RAG Systems for Engineering Teams*, *Model Context Protocol (MCP) for Developers*, *Designing Multi-Agent Systems*, and *Evaluating AI Coding Assistants and LLM Apps*.

Description

This course is a practitioner survey of generative AI as it stands in 2026, from how language models work to building agentic applications. It covers frontier and small language models and multimodal inputs, techniques for extending models with tool use, code generation, and fine-tuning, retrieval-augmented generation over your own documents, the Model Context Protocol for connecting models to services and data, building single agents and orchestrating multi-agent systems, advanced agent techniques for scale and interoperability, and bringing engineering discipline to evaluation. The course is a broad map of the field that cross-links to deeper dedicated courses for teams that want to specialize, and labs reinforce each module with runnable examples.

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

Learning Outcomes

- Explain how large language models work, from embeddings and attention to the transformer architecture and Mixture-of-Experts.
- Compare frontier and small language models and choose appropriately, including local and multimodal options.
- Extend models with tool use, code generation, and fine-tuning.
- Build retrieval-augmented generation pipelines over your own documents with vector databases and reranking.
- Explain the Model Context Protocol and write MCP clients and servers.
- Build single agents and orchestrate multi-agent systems, including agent skills and agentic frameworks.
- Apply advanced agent techniques for scale and cross-framework interoperability, including the Agent2Agent protocol.
- Bring engineering discipline to evaluation with DeepEval, quality metrics, and the LLM-as-a-Judge pattern.

Training Materials

Comprehensive courseware is distributed online at the start of class. All students receive a downloadable MP4 recording of the training.

Software Requirements

Python 3.12+, API keys for at least one frontier model, a vector database for the retrieval topics, optional local-model tooling (Ollama or vLLM), and Git.

Training Topics

Large Language Models

- The frontier model families from OpenAI, Anthropic, and Google
- What differentiates one model from another
- Small Language Models (SLMs) and running them locally
- Data sovereignty and avoiding per-token fees
- Multimodal models and their uses

How LLMs Work

- From word embeddings and text classifiers to RNNs and transformers
- The transformer architecture and attention
- Static versus context-aware embeddings
- Context windows and why GPUs matter

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

- Mixture-of-Experts (MoE) in frontier models

Making LLMs Smarter

- Why a base model is limited to text
- Tool use and calling external APIs
- Code generation as a capability
- Fine-tuning and customization
- Accessing documents, databases, and the web

Retrieval Augmented Generation

- What RAG is and when to use it
- Vector databases, embeddings, and cross encoders
- Building RAG pipelines over PDFs and documents
- Grounding answers and reducing hallucination
- Evaluating retrieval quality

The Model Context Protocol

- What MCP is and why it matters
- Writing MCP clients and servers
- Securing MCP servers
- Deciding whether to expose services through MCP
- Integrating MCP into AI applications

AI Agents

- What AI agents are and how they work
- Agentic frameworks: Google Agent Development Kit, the Microsoft Agent Framework, and Agno
- Building and using agent skills
- Orchestrating workflows beyond prescriptive interfaces
- Putting agents to work in the business

Multi-Agent Orchestration

- Dividing responsibility across multiple agents
- Isolating failures and improving reliability
- Reducing cost, latency, tool bloat, and context bloat
- Parallelism, composability, and reuse
- Best practices for teams of agents

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

Advanced Topics in Agentic AI

- Building agents that scale and share large volumes of data
- CodeGen agents
- Vector-based tool selection
- Streaming view fragments for responsive experiences
- The Agent2Agent (A2A) protocol across frameworks and hosts

Testing and Evaluation

- Why an evaluation strategy matters
- DeepEval and pytest-integrated evals in CI/CD
- Metrics: answer relevancy, faithfulness, and tool correctness
- The LLM-as-a-Judge pattern without labeled data
- Catching regressions before users do

