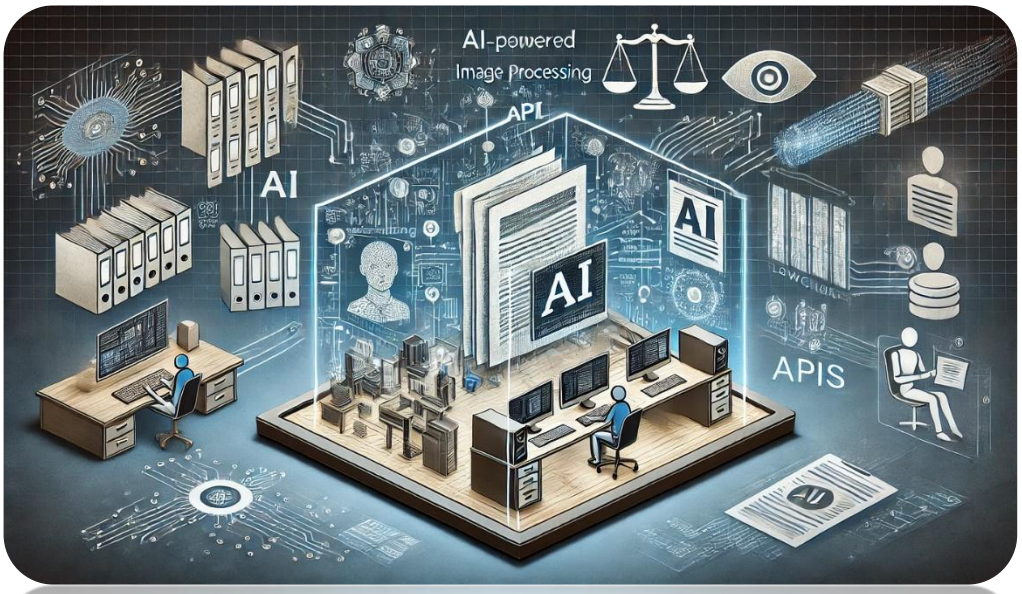
DICE ANALYTICS



Generative AI & Agentic AI – An 8-Week Intensive Training

Overview

- Intensive 8-week program moving from Python fundamentals → advanced AI agent development.
- Covers: Python, data science, deep learning, LLMs, agent frameworks, RAG, fine-tuning, and alignment.
- Tools & Environment:
  - Python for coding and ML development.
  - VS Code as the primary IDE.
  - Google Colab for GPU-powered experiments and tutorials.
  - Vector Databases (Chroma, Pinecone, FAISS) for RAG and semantic search.
- Emphasis: hands-on projects, current frameworks, and industry resources (Jay Alammar's Hands-On LLMs, DeepLearning.AI, PyTorch tutorials).
- Weekly workload: 12–16 hours (two intensive weekend sessions).
- Outcome: Build and deploy LLM-based agents, mastering prompt chaining, model fine-tuning, and agent orchestration.

Week 1: Python & Data Science Foundations

- Python basics: variables, types, loops, functions, modules.
- Data structures: lists, dictionaries, NumPy arrays.
- Data science libraries: Pandas (tabular data), Matplotlib/Seaborn (plotting).
- Intro ML: train/test split, linear/logistic regression, evaluation metrics.
- Hands-on: scripting, data loading/cleaning, visualization.

Week 2: PyTorch & Deep Learning Basics

- Tensors and operations in PyTorch; GPU acceleration.
- Building neural nets with `torch.nn` (layers, activations).
- Autograd: backpropagation and gradients.
- Training loops: loss functions, optimizers (SGD/Adam).
- Model evaluation, saving/loading.
- Hands-on: FashionMNIST workflow.

Week 3: Generative AI, Large Language Models & Multimodal AI

- Introduction to Generative AI
  - What is Generative AI?
  - Key applications: text generation, code, images, audio, video.
  - Difference between discriminative vs. generative models.
  - Examples of modern generative systems (ChatGPT, Stable Diffusion, Whisper).
- Role of LLMs as the backbone of agentic AI.
- Transformers: attention mechanism, encoder/decoder, seq2seq architecture.
- Tokenization & embeddings: word vs. subword tokens.
- Using Hugging Face models: GPT, BERT, CLIP, BLIP-2.
- Text tasks: generation, summarization, translation, code generation.
- Embeddings: semantic search, Retrieval-Augmented Generation (RAG).
- Multimodal models:

- Whisper (ASR – speech-to-text).
- ElevenLabs Scribe (TTS/ASR, multilingual transcription).
- Image captioning and multimodal prompts.

Week 4: AI Agents & RAG

- LangChain basics: prompt templates, tools, LLM wrappers.
- LangGraph: multi-agent workflows (graph state, tool calls).
- RAG systems: embeddings + vector DB + LLM retriever.
- Agentic search: multi-answer retrieval for context.
- Frameworks: AutoGen, crewAI, StackAI (no-code).
- Workflow automation: Botpress, n8n AI agents.

Week 5: Web APIs & Docker Deployment

- Flask basics: routes, request handling, JSON responses.
- FastAPI: async APIs, path/query params, Pydantic, auto-docs.
- Dockerization: writing Dockerfile, building images, running containers.
- Deployment concepts: env variables, volumes, Docker Compose.
- Hands-on: exposing ML model as API, containerized deployment.

Week 6: Fine-Tuning & Efficiency

- LoRA/PEFT: low-rank adapters for lightweight fine-tuning.
- QLoRA: 4-bit quantization + LoRA for large models.
- Merging adapters: `merge_and_unload()` for inference.
- Quantization: 8-bit/4-bit with BitsAndBytes, GPTQ, AWQ.
- Training tips: memory management (gradient checkpointing, mixed precision).
- Hands-on: fine-tuning with PEFT + quantization.

Week 7: Alignment (RLHF & DPO)

- RLHF pipeline: human preference data → reward model → fine-tuned LLM.
- Direct Preference Optimization (DPO): align with classification loss.
- Comparison: RLHF = complex, DPO = simpler, often as effective.
- Alignment topics: bias mitigation, safety, responsible deployment.

Week 8: Final Project

- Project:
  - RAG-based company knowledge assistant.
  - Multimodal customer-support agent (text+audio+image).
  - Autonomous multi-agent deal-spotting system.
- Deliverables: code, containerized app, model report.
- Integration: APIs (Flask/FastAPI), Docker, Hugging Face models, LangChain agents, LoRA/DPO alignment.