COURSE TABLE OF CONTENTS

Generative AI in Production

OEM: Koenig Original • Duration: 3 Days (24 hrs)

COURSE MODULES & TOPICS

Module 1: Production GenAI Landscape and Architecture

- Differences between proof of concept and production
- Common production use cases and patterns
- Core architecture components for GenAI applications
- Model-hosted vs self-hosted deployment options
- Architecture choices: quality, latency, privacy, cost
- Environment separation and deployment topology

Module 2: Foundation Models and Application Integration

- Foundation model capabilities and selection criteria
- Inference patterns: request-response, streaming, batch, async
- Model APIs, SDKs, authentication and response handling
- Inference parameters: temperature, token limits, sampling
- Model routing, fallback, version management, provider abstraction
- API design, input validation, output parsing, retries, rate limiting

Module 3: Prompt Engineering and Lifecycle Management

- Instruction patterns and prompt structure
- Role, context, task, constraints, examples, response format
- Few-shot prompting and schema-constrained responses
- Prompt templates, parameterization, versioning, change management
- Prompt injection awareness and trusted instruction separation
- Prompt experimentation, refinement and regression testing

Module 4: Retrieval-Augmented Generation (RAG) for Production

- RAG architecture and retrieval lifecycle
- Document ingestion, parsing, normalization, metadata extraction
- Chunking, overlap, metadata filtering, source traceability
- Embedding models, vector stores, similarity search, hybrid retrieval
- Re-ranking, query rewriting, contextual retrieval, relevance tuning
- Grounded generation, citations, abstention, handling missing evidence
- Knowledge base refresh and data deletion workflows

Module 5: Agentic AI, Tool Use and Workflow Orchestration

- Agent patterns and bounded autonomy
- Tool/function calling, schemas, validation, permission boundaries
- Planning, task decomposition, routing, state management, memory
- Multi-step workflows, agent handoffs, human approval checkpoints
- Error handling, retries, timeouts, idempotency, loop prevention
- Agent testing, traceability, controls for external actions

© Koenig Solutions

Module 6: Production Application Engineering and Deployment

- Integration patterns for web, mobile, enterprise systems
- Backend orchestration and service-layer design
- Session handling, conversation history, state persistence, context management
- Containerization and deployment: managed services, VMs, Kubernetes
- Serverless and event-driven integration patterns
- API gateways, identity integration, network boundaries, private connectivity
- Release strategies: development, test, staging, production

Module 7: Evaluation, Testing and Quality Engineering

- Quality dimensions: relevance, correctness, groundedness, safety
- Golden datasets, representative test cases, evaluation set management
- Offline and online evaluation, human review, feedback loops
- RAG evaluation: retrieval precision, recall, context relevance
- Agent evaluation: task success, tool selection, escalation behavior
- Regression testing across model, prompt, data, application
- Red teaming, adversarial inputs, edge cases, failure-mode testing

Module 8: Security, Privacy, Governance and Responsible AI

- GenAI threat landscape: prompt injection, data leakage, excessive agency
- Authentication, authorization, least privilege, secrets management
- Sensitive data protection, PII handling, retention, minimization
- Guardrails, content moderation, policy enforcement, safe refusal
- Grounding, hallucination risk controls, human oversight
- Audit trails, model and prompt governance, approval processes
- Compliance, responsible AI principles, risk-based controls

Module 9: Observability, Reliability and Production Operations

- Logging, metrics, distributed tracing, prompt/response traceability
- Operational metrics: latency, throughput, token usage, error rates
- Monitoring retrieval, model calls, tool execution, workflow outcomes
- Quality monitoring, user feedback, drift signals, issue triage
- Resiliency: retries, circuit breakers, fallbacks, graceful degradation
- Incident management, runbooks, service-level objectives, readiness

Module 10: Performance, Scalability and Cost Optimization

- Capacity planning, concurrency, throughput, workload characterization
- Latency optimization: retrieval, orchestration, inference, application
- Token optimization, context management, caching, response streaming
- Model selection and routing by task complexity and objectives
- Batch processing, asynchronous workloads, quota management
- Cost attribution, usage budgets, chargeback/showback, optimization reviews