

AI AGENT TESTING & EVALUATION

From AI Agent Fundamentals to Evals, Golden Datasets, Python Frameworks, CI/CD, Red Teaming and AI Governance





Course Overview

This course is designed for QA and automation professionals who want to move into AI Agent Testing and AI Evaluation. It starts with AI and agent fundamentals, moves into no-code agent development and evaluation, and then progresses to Golden Datasets, evaluation lifecycles, LLM-as-a-Judge, Python evaluation frameworks, CI/CD, red teaming, governance and compliance.

Target Audience : QA Professionals, Test Engineers, Automation Engineers and Test Managers

Learning Approach : Concepts + No-Code Labs + Python Evals + CI/CD

Core Focus : Testing, Evaluation, Regression, Governance and Compliance

Final Outcome : Build a complete AI Agent Quality & Evaluation solution

CONTACT US

info@ittalenthub.co.uk

+44 7453 289655

https://ai-testing-hub.sriram-angajala.chatgpt.site/

COURSE MODULES

Module 1

AI Engineering & AI Agent Fundamentals

Module 2

No-Code AI Agent Development, Testing & Evaluation

Module 3

AI Evals & Evaluation Metrics

Module 4

Golden Dataset Preparation & AI-Generated Test Data

Module 5

AI Agent Evaluation Lifecycle

Module 6

Python AI Evaluation Frameworks

Module 7

AI Regression Testing & CI/CD

Module 8

AI Governance, Risk & Compliance

Optional Track: Using AI in Software Testing

1: Using AI to Generate Testing Artifacts

2: Developing Test Cases & Test Data Using AI

3: AI-Assisted Automation with JavaScript & Playwright

COURSE LEARNING JOURNEY

AI FUNDAMENTALS → BUILD & TEST NO-CODE AI AGENTS→ AI EVALS & METRICS → GOLDEN DATASET CREATION → AI AGENT EVALUATION LIFECYCLE → PYTHON + DEEPEVAL → REGRESSION TESTING & CI/CD → AI GOVERNANCE, RISK & COMPLIANCE



Course Outcome

On completion, learners will understand how to build, test, evaluate, automate and govern AI Agents. The course supports progression toward roles such as AI Test Engineer, AI QA Engineer, AI Evaluation Engineer, AI Quality Engineer and Agentic AI Tester.

COURSE MODULES

Module 1: AI Engineering & AI Agent Fundamentals

- Introduction to Artificial Intelligence and Generative AI
- Large Language Models (LLMs): tokens, context windows and temperature
- Prompting and system instructions
- Tools and function calling
- Retrieval-Augmented Generation (RAG)
- Model Context Protocol (MCP)
- AI Agents and Agentic AI
- Planning, reasoning, memory and multi-step execution
- Open-source vs. closed AI models
- Understanding a typical AI Agent architecture

Hands-on Practical :

Trace a request through: User → LLM → RAG → Tool → External System → LLM → Response.

Deliverable :

Annotated AI Agent architecture diagram.

Module 2: No-Code AI Agent Development, Testing & Evaluation

- Introduction to no-code/low-code AI Agent platforms such as n8n
- Triggers, prompts, instructions, tools/actions and knowledge sources
- Connecting external applications, tools and RAG knowledge bases
- Creating test datasets with inputs and reference answers
- Running multiple test cases against an AI Agent
- Reviewing execution traces and logs

- Identifying incorrect tool calls and ungrounded responses
- Using the Set Metrics operation for evaluation
- Correctness (AI-based): meaning consistent with the supplied reference answer; score 1–5
- Helpfulness (AI-based): whether the response answers the query; score 1–5
- String Similarity: character-level/edit-distance similarity; score 0–1
- Categorization: exact/category match against the reference; result 0 or 1
- Tools Used: whether expected tool usage occurred; score 0–1
- Comparing AI-based metrics with deterministic metrics
- Understanding why low string similarity can still have high semantic correctness

Hands-on Practical :

Build a no-code AI Agent and execute at least 10 test cases, recording outputs, pass/fail decisions and metric scores.

Deliverable :

No-code AI Agent + test dataset + evaluation metrics report.

Module 3: AI Evals & Evaluation Metrics

- What is an AI Eval?
- AI testing vs. traditional software testing
- Why traditional pass/fail assertions are not enough for open-ended AI output
- Deterministic vs. non-deterministic testing
- Reference-based vs. reference-free evaluation
- Correctness, relevance, helpfulness, coherence, fluency, completeness, groundedness and safety
- BLEU, ROUGE and F1: uses and limitations
- Introduction to LLM-as-a-Judge
- Defining acceptance thresholds and interpreting scores
- Defining 'good' before measuring it

Hands-on Practical

Evaluate sample AI responses against multiple quality dimensions and compare scoring differences.

Deliverable

Evaluation scorecard and good-vs-bad response examples.

Module 4: Golden Dataset Preparation & AI-Generated Test Data

- What is a Golden Dataset and why it matters
- Designing Golden Test Cases
- User input and expected/reference output
- Expected tools and expected tool-call sequences
- Expected RAG/context sources
- Must-have requirements and must-not-do behaviours
- Happy-path, negative, edge and adversarial scenarios
- Creating the first Golden Dataset manually
- Using AI to generate additional test cases
- Using AI to generate synthetic test data
- Human review and validation of AI-generated cases
- Detecting duplicate and near-duplicate test cases
- Separating human-approved Golden Data from synthetic coverage data
- Tagging, categorisation and dataset versioning
- Adding production failures back into the regression dataset

Hands-on Practical :

Create 10 human-curated Golden Test Cases, generate 20–30 additional cases using AI, then review, deduplicate, tag and version them.

Deliverable :

Versioned Golden Dataset ready for automated evaluation.

Module 5: AI Agent Evaluation Lifecycle

- **Level 1 –** Assertions and unit testing: schema, structure, tool calls, arguments, negative/safety assertions
- Handling non-determinism using soft assertions, repeat runs and controlled settings

- **Level 2 –** Model-based evaluation / LLM-as-a-Judge
- Rubric design, weighted criteria and hard-fail conditions
- Judge biases: verbosity, position and self-preference

- **Level 3 –** Live experiments and A/B testing
- Tool selection accuracy and tool argument correctness
- Agent trajectory and multi-step behaviour testing
- Error handling and recovery
- Red teaming: prompt injection, indirect prompt injection, harmful output, leakage, ungrounded content and copyright risk
- Introduction to PyRIT and automated red-team approaches
- Production monitoring and the production-failure feedback loop

Hands-on Practical

Map checks to PR time, nightly runs, pre-release and production; show how a production failure becomes a new regression case.

Deliverable

End-to-end AI Agent evaluation lifecycle diagram.

Module 6: Python AI Evaluation Frameworks

- Introduction to Python-based AI evaluation
- The judge pattern: agent output → judge model → structured score and reasoning
- DeepEval and LLMTestCase
- GEval for custom evaluation criteria
- ToolCorrectnessMetric for tool-call evaluation

- Agent-specific evaluation metrics
- Creating custom weighted rubrics
- Hard-fail caps and pass/fail thresholds
- Reference-based evaluation for Golden Datasets
- Reference-free evaluation for sampled production responses
- Evaluation result reporting
- Response caching, batching and managing evaluation cost

Hands-on Practical

Build a Python evaluation script using DeepEval and run it against at least 10 sample AI Agent outputs.

Deliverable

Working Python evaluation framework producing scores, reasoning and pass/fail results.

Module 7: AI Regression Testing & CI/CD

- Running Golden Datasets as regression suites
- pytest for AI testing
- Integrating DeepEval with pytest
- Level 1 checks on every Pull Request
- Scheduled Level 2 evaluations
- Nightly and pre-release AI evaluations
- Quality gates and evaluation thresholds
- Baseline comparison and score drift
- Detecting prompt drift, model changes and performance degradation
- Handling flaky/non-deterministic failures
- Caching judge responses in CI
- Failure analysis and triage
- Automated red-team pipeline stages
- GitHub Actions or equivalent CI/CD integration

.

Hands-on Practical

Create PR-gate, nightly-evaluation and pre-release red-team stages in a sample CI/CD pipeline.

Deliverable

CI/CD pipeline definition with AI quality gates

Module 8: AI Governance, Risk & Compliance

- Introduction to AI Governance and AI risk management
- AI system documentation, evidence and traceability
- Data privacy, personal data and sensitive data
- GDPR considerations
- Bias and fairness testing
- Explainability, transparency and human oversight
- Prompt, model and dataset versioning
- Logging and AI audit trails
- AI security testing
- Prompt injection and indirect prompt injection risks
- RAG security and data leakage
- Excessive agency and tool-access risks
- EU AI Act overview
- UK AI regulatory approach
- NIST AI Risk Management Framework
- ISO/IEC 42001 overview
- Mapping Requirement → Risk → Test → Evaluation → Result → Evidence

Hands-on Practical

Create an AI risk register and map selected risks to tests, evaluation results and evidence.

Deliverable

AI Governance testing checklist + Risk/Test/Evidence matrix.

Optional Track: Using AI in Software Testing

1: Using AI to Generate Testing Artifacts

- Generate test plans, test scenarios, test cases and traceability matrices using AI
- Prompting patterns for diverse test coverage
- Reviewing AI-generated artifacts for missing coverage, false coverage and duplicates

Practical: Generate a test plan and at least 15 scenarios from sample requirements, then review and deduplicate them.

2: Developing Test Cases & Test Data Using AI

- Generate happy-path, negative, edge and adversarial test cases
- Synthetic test data generation
- Human review and validation
- Combining AI-generated coverage with the approved Golden Dataset
- Tagging and versioning generated data

Practical: Generate and review a 30-case synthetic testing dataset.

3: AI-Assisted Automation with JavaScript & Playwright

- Using AI to scaffold a Playwright framework
- Page Objects, fixtures and configuration
- Generating Playwright scripts from natural-language test cases
- Reviewing locator reliability, waits and flakiness
- Integrating Playwright into CI/CD

Practical: Use AI to scaffold a Playwright framework and automate at least 5 test scenarios.

Final Course Project :

- Learners will build or use an AI Agent and create a complete AI Quality & Evaluation solution.
- AI Agent
- Golden Dataset
- No-code evaluation and metrics
- LLM-as-a-Judge evaluation
- Python/DeepEval framework
- Tool-call and trajectory testing
- Regression test suite
- CI/CD integration
- Red-team testing
- Governance and compliance evidence

BOOK YOUR FREE CAREER CONSULTATION TODAY

+44 7453 289655

https://ai-testing-hub.sriram-angajala.chatgpt.site/

info@ittalenthub.co.uk

SCAN TO VISIT OUR WEBSITE