

BEST-PRACTICE GUIDE · TRANSFORM

Phase 06 — Test

Building trust in classical software and AI behaviour

The Test phase assures the quality of the target solution on two levels: classical, risk-based testing per ISTQB/ISO 29119 for the deterministic parts — and systematic evaluation (evals) for the behaviour of the AI agents. Test cases are generated agentically from the requirements and stay linked to them.

Goal of the phase

The phase delivers robust quality evidence: fulfilled test coverage along the test pyramid, passed acceptance tests with the business, and documented eval results for all agentic components.

Process model

- 01

Test strategy & risk analysis
Derive the test strategy from the Design phase's quality scenarios; risk-based prioritization of test depth per component and process by damage potential and failure probability.
- 02

Agentic test case generation
Automatic decomposition of the specification into atomic Gherkin requirements and derivation of complete test cases in Signoff; coverage analysis against the specification clauses.
- 03

Test automation along the pyramid
Unit, integration and end-to-end tests in the CI track; contract tests for the agents' API and tool interfaces; mutation testing for critical logic.
- 04

Agent evaluation (evals)
Scenario-based assessment of agent behaviour: task success, robustness against prompt variation, guardrail compliance, hallucination rate; regression evals on every model or prompt change.
- 05

Acceptance & quality evidence
Business acceptance tests with the client (UAT), consolidation of all results into the quality evidence report; release recommendation for the Deploy phase.

Methodology mix

METHOD	PURPOSE	REFERENCE
ISTQB test process	Standardized test process from planning to completion assessment; shared vocabulary across the engagement.	ISTQB
ISO/IEC/IEEE 29119	International standard series for test processes, documentation and techniques; frame of the quality evidence.	ISO/IEC/IEEE 29119-1:2022
Test pyramid	Distribution of test automation across unit, integration and UI levels; cost efficiency and fast feedback.	Vocke / Fowler (2018)
Risk-based testing	Steering test depth and order via product risks; focuses the budget on critical paths.	ISTQB Advanced

METHOD	PURPOSE	REFERENCE
Mutation testing	Empirical assessment of test suite quality through injected faults; mandatory for migrated business logic.	Jia & Harman (2011)
LLM/agent evaluation	Systematic, scenario-based assessment of AI behaviour (holistic evaluation); foundation of the ReqPOOL eval suites.	Liang et al. (2022), HELM

Platform support

ReqPOOL Suite: **Signoff** (AI-based acceptance)

Signoff — specification analysis	The specification (PDF, DOCX, text) is decomposed into atomic, verifiable requirements in Gherkin format — each with a unique ID and a verbatim reference to the source clause.
Signoff — test case generation	Complete test cases including test data, boundary and negative cases — automatically linked in a traceability matrix; coverage is reported transparently.
Signoff — agentic test execution	An AI agent tests the target system directly and collects evidence (screenshots, logs); alternatively the user works through a guided playbook whose screen recording is evaluated multimodally.
Signoff — independent assessment	An independent AI verification model assesses every test case (passed / failed / not verifiable) and suggests bugfixes.
Signoff — revision-proof quality evidence	Acceptance report (PDF) with full traceability from specification clause to test result; Excel export and audit log.

Best-practice checklist

	Have test cases and test runs approved by the project manager before execution — no unapproved runs against client systems.
	Treat generated test cases like generated code: business review of the critical cases, no unchecked adoption.
	Derive test scenarios from real process data and edge cases of the Analyse phase — synthetic happy-path tests create false confidence.
	Report coverage against the specification; unverified clauses block the test-exit gate.
	Accept every assessment only with evidence (screenshot, log); do not override the verification model — clarify deviations with the business side.
	After bugfixes, re-run the affected test cases and document the regression evidence in the report.

Artifacts & outcomes

- Test strategy with risk assessment
- Automated test suites along the pyramid
- Eval reports per agentic component
- Equivalence evidence for migrated systems
- Audit-ready quality evidence report (ISO 29119)

Quality gate criteria

☐	Risk-based coverage targets reached and proven
☐	All acceptance tests confirmed by the business
☐	Eval thresholds of all agents met
☐	No open priority 1 and 2 defects

Scientific deep dives

ISTQB: syllabi and glossary — International standard of tester education and terminology.
<https://www.istqb.org/>

ISO/IEC/IEEE 29119-1:2022 — Software Testing — Standard series for test processes and documentation.
<https://www.iso.org/standard/81291.html>

Vocke: The Practical Test Pyramid (martinfowler.com) — Practice reference on distributing test automation.
<https://martinfowler.com/articles/practical-test-pyramid.html>

Jia & Harman: An Analysis and Survey of Mutation Testing (IEEE TSE 2011) — Comprehensive research survey on mutation testing.
<https://doi.org/10.1109/TSE.2010.62>

Liang et al.: Holistic Evaluation of Language Models (HELM) — Scientific frame for multi-dimensional LLM evaluation.
<https://crfm.stanford.edu/helm/>

Anthropic: Statistical Approach to Model Evals — Statistical grounding of eval comparisons.
<https://www.anthropic.com/research/statistical-approach-to-model-evals>