

Testing AI in the Real World

A practical guide for testing AI applications

Bianca Mihalcea
QA Engineering Manager, Zitec

Certified



Corporation

Agenda

01**Classic vs. AI Testing**

Mindset shifts & paradigm comparison

02**Risk Assessment**

Personas, worst-case scenarios & baselines

03**Tracing & Observability**

Root cause isolation with Langfuse

04**Test Data Generation**

Dimension-based synthetic data techniques

05**Evaluation Methodology**

Taxonomy, code evaluators & LLM-as-a-Judge

06**Non-Functional Safeguards**

Security, performance & guardrails

What is actually different?

Classic application

- Deterministic verification (expected == actual).

EXAMPLE

Classic Form

- Structured POST request (order_id: 12345)
- Deterministic database query

AI application

- Probabilistic evaluation.
- Was the output text accurate, polite, and relevant?

EXAMPLE

AI Chatbot

- Free-form natural language query
- Intent classification
- Dynamic API selection

Identify what matters most

01

User Personas & Goals

- Identify who is using the application and their expertise level.
- Determine how users benefit and what their primary goal is.
- Align feature scope with target user workflows and expectations.

02

Worst-Case Risk Matrix

- Determine critical failure risks across primary user scenarios.
- Map true/false positive and negative outcomes based on business impact.
- Prioritize high-severity edge cases for rigorous testing and mitigations.

03

Baseline Criteria

- Agree on application-specific success criteria with key team stakeholders.
- Set quality and performance thresholds before launching evaluation pipelines.
- Establish objective benchmarks for continuous automated testing.

Opening the Black Box

AI System Observability

End-to-End Execution Tracing

Record complete execution context across all component layers:

- **Autonomous Agents:** Track decision chains, tool routing, and agent interactions.
- **Vector Databases & Retrieval:** Log query embeddings, similarity scores, and retrieved contexts.
- **API Tools & LLM Generations:** Capture payload inputs, external API tool calls, raw prompts, and LLM text completions.

Operational Telemetry & Metrics

Track real-time operational metadata to optimize performance and efficiency:

End-to-End Latency

Measures total response time and latency breakdown across every call step.

Token Usage

Monitors exact input prompt and output completion token counts.

API Cost Control

Calculates real-time estimated API costs across model providers.

Dimension-Based Synthetic Data Generation

01

Define Categories

Define 3+ categories of variation to establish core data dimensions.

02

Seed Tuples

Create ~20 manual seed tuples to serve as baseline anchors.

03

LLM Expansion

Use LLM tuple expansion and query conversion for high-volume generation.

04

Quality Traces

Aim for ~100 diverse, high-quality traces ready for evaluation.

Error analysis

A mandatory step before
automatic evaluation

01

Gather traces and manually analyse 30–50 of them. “**Journal**” about what you identify

02

Cluster similar failures into distinct **categories**

03

Label and measure. Tag every trace in the sample against the taxonomy and compute failure rates per category

04

Decide and act. For each category, choose between a prompt or code fix or an evaluator.

Evaluation Approaches

Code Evaluators

Deterministic & Algorithmic

- Fast, cheap deterministic rules
- Output structure verification
- Regex pattern matching
- Tool execution state

LLM-as-a-Judge

Probabilistic & Qualitative

- Narrow, single-purpose prompts
- Subjective probabilistic quality
- Faithfulness assessment
- Tone & style evaluation

Evaluation Approaches

Reference-Based evaluation

Example

An invoice-processing test where the extracted total (\$450.00) is verified via exact string match against the pre-labeled golden answer (\$450.00).

Reference-Free evaluation

Example

A marketing copy rewriter evaluated live by an LLM-as-a-Judge to score whether the response matches brand guidelines, with no reference output available

01

Security

OWASP Top 10 for LLMs

- Prompt injection
- Sensitive data exposure

02

Performance

Measure latency across multi-step agent chains, large retrieval operations and unexpected user input.

03

Cost Management

Set strict token/financial limits for graceful degradation.

“Testing AI isn't about proving deterministic perfection—it's about building probabilistic confidence.”

ZITEC

Thank you