GSDC
Global Skill Development Council

AI TESTING PROFESSIONAL PLAYBOOK

www.gsdcouncil.org

CHAPTER 1: WHY AI TESTING IS DIFFERENT FROM TRADITIONAL TESTING

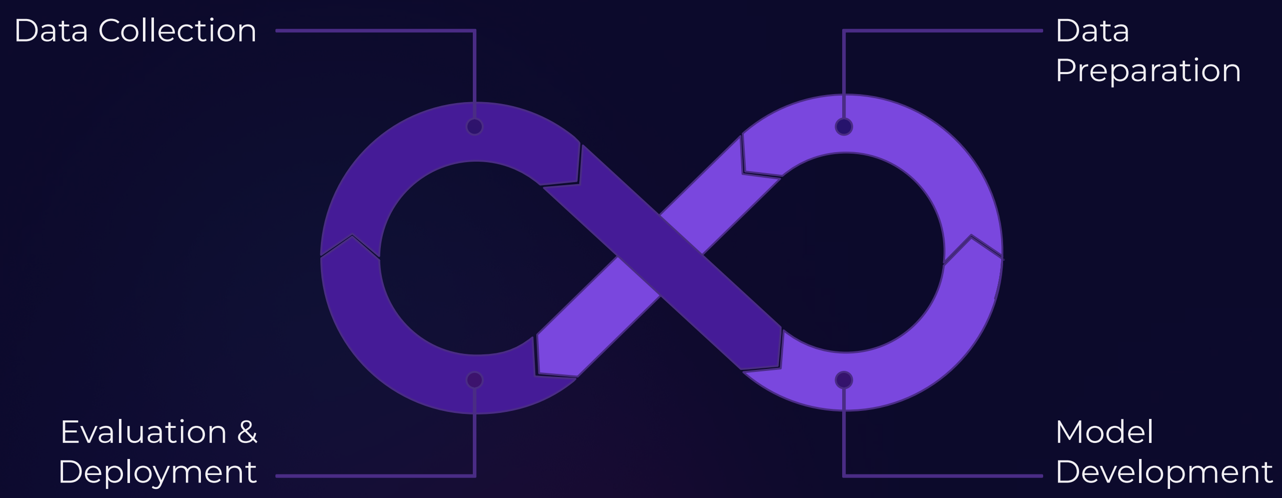
Traditional software testing assumes deterministic behavior — same input, same output, every time. AI breaks that assumption completely. Here is what changes:

Traditional Software Testing	AI System Testing
Fixed rules produce predictable outputs	Models learn from data — outputs can vary
Pass/Fail is binary and clear	"Correct" output is often probabilistic or contextual
Requirements are explicit	AI behavior emerges from training — not always documented
Bugs are code defects	AI failures can be data defects, model drift, or bias
Test once, ship	AI systems require continuous monitoring post-deployment
Single version testing	Models retrain — every version is potentially a new system

The core challenge of AI testing: You are not testing logic — you are testing learned behavior across an infinite input space.

CHAPTER 2: THE AI SYSTEM LIFECYCLE — WHERE TESTING FITS

Testing in AI is not a phase — it is embedded across the entire lifecycle.



Testing touchpoints at each stage span the full pipeline, from raw data to live production systems.

Data Collection

- Is the data representative of real-world use?
- Are protected groups fairly represented?
- Is the data source trustworthy and legal to use?

Data Preparation

- Are there missing values, duplicates, or outliers that will skew the model?
- Has data been correctly labeled? Is labeling consistent?
- Are train/validation/test splits properly separated?

Model Development

- Is the model architecture appropriate for the task?
- Are evaluation metrics aligned with business outcomes?
- Is the model overfitting to training data?

Lifecycle Testing Touchpoints (Continued)

Model Evaluation

- Does the model perform equitably across all user groups?
- Is accuracy, precision, recall, and F1 score within acceptable thresholds?
- Is the model explainable — can you understand why it made a decision?

Deployment

- Does the model behave consistently in the production environment?
- Are latency and resource consumption within acceptable bounds?
- Are there guardrails for edge cases and unexpected inputs?

Monitoring

- Is model performance degrading over time? (Model Drift)
- Is the incoming data distribution shifting? (Data Drift)
- Are there new bias patterns emerging in production?

CHAPTER 3: THE 8 CORE AI TESTING TYPES — WHAT TO TEST AND HOW

TYPE 1

FUNCTIONAL TESTING

What: Does the AI do what it was designed to do?

How to approach

- Define expected behavior for representative inputs
- Test typical inputs, boundary inputs, and edge cases
- Test with inputs the model has never seen (out-of-distribution inputs)
- Verify output format, range, and structure are as specified

Key questions

- Does the model produce outputs in the expected format?
- Does it handle null, empty, or malformed inputs gracefully?
- Does it fail safely when confidence is low?

TYPE 2

DATA QUALITY TESTING

What: Is the data used to train and run the model trustworthy?

How to approach

- Profile data for completeness, consistency, accuracy, and timeliness
- Check label quality — are annotations accurate and consistent across labelers?
- Test for data leakage — is test data accidentally contaminating training?
- Verify data provenance — where did this data come from, is it ethically sourced?

Key metrics

% Missing Values

Per feature across the dataset

Inter-Annotator Agreement

Rate for labeled data consistency

Distribution Similarity

Between training and production data

TYPE 3

BIAS & FAIRNESS TESTING

What: Does the model treat all groups equitably?

How to approach: Define protected attributes relevant to your use case (gender, race, age, disability status). Test model performance separately across demographic groups. Check for disparate impact — does the model produce different outcomes for similarly situated people from different groups? Run counterfactual fairness tests — change only a protected attribute and observe output change.

Fairness Metrics to Know

Metric	What It Measures
Demographic Parity	Equal positive prediction rates across groups
Equal Opportunity	Equal true positive rates across groups
Predictive Parity	Equal precision across groups
Individual Fairness	Similar individuals receive similar outcomes
Counterfactual Fairness	Outcome unchanged when protected attribute is altered

Tools: IBM AI Fairness 360, Microsoft Fairlearn, Google What-If Tool

TYPE 4

MODEL PERFORMANCE TESTING

What: Is the model accurate enough for real-world use?

Core metrics by task type

Task Type	Key Metrics
Classification	Accuracy, Precision, Recall, F1 Score, AUC-ROC
Regression	MAE, RMSE, R^2 Score
NLP / Generative	BLEU, ROUGE, Perplexity, Human Evaluation
Ranking	NDCG, MAP, MRR
Object Detection	mAP, IoU

Thresholds matter: Define acceptable performance ranges before testing — not after seeing the results.

TYPE 5

ROBUSTNESS & ADVERSARIAL TESTING

What: Does the model behave reliably when inputs are noisy, unusual, or deliberately manipulated?

→ Noise Injection

Add random noise to inputs and test stability of outputs

→ Adversarial Examples

Craft inputs designed to fool the model (slight pixel changes that flip image classification)

→ Boundary Testing

Test at decision boundaries where the model is most uncertain

→ Input Mutation

Systematically alter inputs to find breaking points

→ Stress Testing

Test under high load, high data volume, or rapid input changes

Key question: Can an attacker manipulate this model's output by carefully crafting their input?

TYPE 6

EXPLAINABILITY & TRANSPARENCY TESTING

What: Can you understand and explain why the model made a specific decision?

Why it matters: Regulators (EU AI Act, GDPR) and business stakeholders increasingly require AI decisions to be explainable, especially in high-stakes domains.

Explainability Approaches

Approach	Tool	What It Shows
Feature Importance	SHAP, LIME	Which features most influenced the prediction
Counterfactual Explanations	DiCE	What would need to change to get a different outcome
Attention Maps	Built into transformers	Where the model "looked" in the input
Model Cards	Google's framework	Standardized model behavior documentation
Saliency Maps	Used in image models	Which pixels drove the decision

Test: Can a non-technical stakeholder understand the explanation produced? If not, it is not explainable enough.

TYPE 7

SECURITY & PRIVACY TESTING

What: Is the model resistant to attacks and does it protect sensitive data?

Attack Types to Test For

Attack	What It Is
Model Inversion	Can an attacker reconstruct training data from model outputs?
Membership Inference	Can an attacker determine if a specific record was in training data?
Data Poisoning	Can malicious training data corrupt the model's behavior?
Model Extraction	Can an attacker replicate the model through repeated queries?
Prompt Injection	For LLMs — can malicious prompts override system instructions?

Privacy Checks

- Does the model memorize and leak PII from training data?
- Is differential privacy applied where required?
- Are outputs compliant with GDPR/CCPA data minimization requirements?

TYPE 8

MODEL DRIFT MONITORING

What: Is the model's performance degrading over time as the world changes?

Types of Drift

Drift Type	What Happens	Example
Data Drift	Input distribution shifts	User demographics change after product launch
Concept Drift	The relationship between inputs and outputs changes	"Fraud pattern" changes as fraudsters adapt
Label Drift	What counts as the correct answer changes	Medical guidelines update
Model Degradation	Performance declines without obvious drift cause	Infrastructure or dependency changes

Monitoring Approach

01

Set performance baselines at deployment

02

Define alert thresholds for acceptable deviation

03

Monitor continuously with automated alerting

04

Establish retraining triggers and cadence

CHAPTER 4: AI TESTING STRATEGY — HOW TO BUILD ONE

A good AI testing strategy answers 6 questions:

1	**What are we testing?** Define the AI system's purpose, inputs, outputs, and decision context.
2	**What does "good" look like?** Set explicit performance thresholds, fairness criteria, and reliability benchmarks before testing begins.
3	**Who could be harmed?** Identify all user groups and consider which groups face higher risk of harm from model errors or bias.
4	**What are the highest-risk areas?** Use risk-based testing — prioritize test effort where model failures have the greatest potential impact.
5	**How will we test?** Define the mix of automated testing, human evaluation, adversarial testing, and monitoring.
6	**How will we know when it is good enough?** Define clear go/no-go criteria. An AI model does not need to be perfect — it needs to meet defined thresholds across all critical dimensions.

CHAPTER 5: THE NON-DETERMINISM PROBLEM — TESTING WHEN OUTPUTS VARY

One of the most fundamental challenges in AI testing is that the same input can produce different outputs — especially in generative AI. Traditional "expected output" testing fails here.

Strategies for Testing Non-Deterministic Systems

Range-based assertions

Output must fall within an acceptable range, not match a fixed value

Property-based testing

Test that outputs satisfy properties (e.g., "always returns a valid JSON," "never contains PII") rather than specific values

Semantic similarity scoring

For text outputs, measure meaning similarity rather than exact match (using cosine similarity or embedding distance)

Human evaluation panels

For high-stakes or creative outputs, use structured human judgment with defined rubrics

Statistical testing

Run the same input many times and test that the distribution of outputs is acceptable

Consistency testing

Same input should produce outputs that are at least semantically consistent across runs

CHAPTER 6: ETHICAL AI TESTING — THE EMERGING FRONTIER

AI testing is no longer purely technical. Ethical dimensions are now a core part of any serious AI quality assurance program.

The 5 Ethical Testing Dimensions

Dimension	What You Test
Fairness	Does the model treat people equitably regardless of protected characteristics?
Transparency	Can the model's decisions be understood and explained?
Accountability	Is there a clear chain of responsibility for model decisions?
Privacy	Does the model protect personal data and prevent re-identification?
Safety	Does the model fail safely, especially in high-stakes contexts?

Regulatory Landscape to Know

EU AI Act

Risk-based regulation; high-risk AI systems require mandatory testing and documentation

GDPR / CCPA

Data privacy requirements affect training data and output handling

US Executive Order on AI (2023)

Safety and security testing requirements for frontier models

IEEE Ethically Aligned Design

Voluntary framework for ethical AI development

GSDC
Global Skill Development Council

CERTIFIED AI TESTING PROFESSIONAL (CAITP)

GSDC
Global Skill Development Council
Ethical Hacking Foundation
CERTIFIED

ABOUT GSDC CERTIFICATION

EBOOK

Extensive and exclusive Ebook created by world's experts to help you with understanding core concepts.

LEARNING MATERIALS

Get access to learning materials such as videos, ebooks, templates, and practice exams, which will help you clear the certification exam.

CREATED BY EXPERTS

GSDC certifications are created and authored by world's leading experts in the field.

LEARNING OBJECTIVE

- Gain insights into autonomous decision-making processes
- Apply knowledge using ready-to-implement templates
- Demonstrate ability to work with Agentic AI models
- Validate your skills wit

Enroll now with the code **LEARN20** To avail **20%** discount

Enroll Now

www.gsdcouncil.org