AI Audit Testing Toolkit

Operational AI assurance guidance for internal auditors

1. AI Control Catalog

This section provides **fully written, control statements** for each AI assurance domain.

	Domain	*Sample Controls*
1	**Data Quality & Data Governance**	**1. Data Lineage Documentation:** Management documents and maintains end-to-end data lineage for all datasets feeding the AI model, including source systems, transformations, and loading processes. **2. Data Quality Rules:** Automated data quality checks (e.g., null checks, range checks, outlier detection) run before model execution; exceptions are logged and reviewed. **3. Restricted Data Access:** Access to training and production datasets is restricted based on least privilege and reviewed quarterly. **4. Sensitive Data Safeguarding:** Sensitive data is masked, sanitized, or excluded from training and inference datasets.
2	**Model Design & Development**	**1. Documented Requirements:** Model requirements, assumptions, performance expectations, and limitations are formally documented before development. **2. Algorithm Selection Justification:** Management documents the rationale for algorithm choice, including alternatives considered. **3. Version-Controlled Model Code:** All model code, parameters, and training pipelines are maintained in version control.
3	**Model Validation**	**1. Independent Model Validation:** Models undergo independent validation prior to deployment and at least annually. **2. Accuracy & Fairness Testing:** Validation includes accuracy KPIs, stress tests, fairness checks, and edge-case testing.
4	**Explainability & Transparency**	**1. Explainability Mechanisms:** Explainability techniques (e.g., SHAP, LIME) are maintained and available for audit. **2. User-Facing Explanations:** Outputs include clear, actionable explanations understandable to non-technical stakeholders.
5	**Model Drift & Performance Monitoring**	**1. Drift Monitoring:** Model performance and data drift are monitored continuously; alerts trigger review. **2. Retraining Procedures:** Models are retrained based on triggers (drift, seasonal patterns, accuracy decay).

	Domain	Sample Controls
6	**Access, Security, & Change Management**	**1. Role-Based Access Control:** Access to model artifacts, scripts, and compute resources requires authorization. **2. Change Control:** All model updates follow formal change management procedures.
7	**Responsible Use & Governance**	**1. AI Acceptable Use Policy:** Users follow an AI acceptable use policy including data handling and review expectations. **2. Prompt Governance (GenAI):** All prompts follow approved templates and avoid prohibited data types.
8	**Output Quality & Human Oversight**	**1. Human Review:** Critical AI outputs undergo mandatory human review with documented approval. **2. Output Accuracy Monitoring:** Inconsistencies, hallucinations, or anomalies are logged and investigated.

2. PASS/FAIL CRITERIA ("What Good Looks Like")

Below is a **clear, auditor-friendly** pass/fail table for each domain.

Domain 1: Data Quality

Test Objective	Pass Criteria	Fail Criteria
Validate data lineage	Complete lineage covering source → transformations → model	Missing nodes; undocumented manual steps
Data quality checks	Logs show checks run; exceptions reviewed	No log evidence; no review of exceptions
Data representativeness	Dataset balanced; bias tests performed	Skewed data; no bias tests

Domain 2: Model Development

Test	Pass	Fail
Model documentation	Business Requirement Document, assumptions, architecture clear	No documentation
Reproducibility	Versioning + fixed seeds	Manual runs, no version control

Domain 3: Validation

Test	Pass	Fail
Independent validation	Clear evidence of independence	Developer validated their own model
Metrics	Accuracy ≥ threshold	Below threshold, no remediation

Domain 4: Explainability

Test	*Pass*	*Fail*
Explanations accessible	SHAP/LIME available and meaningful	"Black box" output

Domain 5: Drift Monitoring

Test	*Pass*	*Fail*
Drift logs	Monitored monthly; alerts documented	No drift monitoring
Retraining	Evidence of retraining cycle	No retraining ever done

Domain 6: Access & Change Control

Test	*Pass*	*Fail*
Access	Role-based, reviewed quarterly	Shared accounts
Change tracking	Change tickets exist	Direct edits to code

Domain 7: Responsible Use

Test	*Pass*	*Fail*
Prompt use	Sanitized prompts used	Sensitive data in logs
Policy	Policy exists and is used	No AI use policy

Domain 8: Output Quality

Test	*Pass*	*Fail*
Reperformance	AI output matches manual	AI output incorrect
Review	Evidence of human review	No review documented

3. Audit Scoring Guidance

Provide auditors a consistent way to score AI controls.

Control Effectiveness Rating

- **Strong** — Fully documented, monitored, independently validated
- **Moderate** — Some gaps; documentation incomplete
- **Weak** — Major missing governance or controls
- **Unsatisfactory** — No controls exist

Residual Risk Rating

- **High** — Model affects financial, compliance, or customer decisions
- **Medium** — Model supports internal decisions
- **Low** — Model augments productivity (GenAI summaries)

4 Evidence Quality Standards

Clearly define what "*good evidence*" looks like.

✓ **Strong Evidence**

- Exported logs (timestamped)
- Documented validation reports
- Config files
- Reproducible outputs

✗ **Weak Evidence**

- Emails
- Verbal confirmation
- Screenshots without timestamps
- "We think it works fine" statements

3. AI AUDIT WORKPAPER EXAMPLE

SECTION 1 — Planning & Risk Understanding

Obtain AI Inventory & Model Overview

Procedures

1. Request the enterprise AI use-case inventory.
2. Verify each model has: owner, purpose, data sources, deployment environment, risk classification, and last validation date.
3. Identify the models in audit scope based on risk tiering.

Evidence

- AI inventory
- Model overviews / model cards

Red Flags

- Inventory incomplete
- No model owner assigned

SECTION 2 — Control Testing

DOMAIN 1 — Data Quality & Data Governance

1.1 - Data Lineage Documentation

Control Objective

Data lineage is documented end-to-end, including source systems, transformations, and model ingestion.

Audit Procedures

1. **Obtain lineage documentation** (diagrams, ETL flows, data dictionaries).
2. **Trace a sample of 5–10 data elements** from source → transformation → feature repository → model input.
3. **Reconcile lineage with actual system metadata** (e.g., SQL queries, pipeline scripts).
4. **Verify documentation is up-to-date** by checking last update timestamps.
5. **Confirm manual data handling steps** (if any) are documented and approved.

Evidence

- Lineage diagrams
- ETL code extracts
- Metadata repository exports

Common Failures

- Missing manual transformation steps
- Lineage last updated >12 months ago

1.2 - Data Quality Rules

Control Objective

Automated data quality rules run before model execution; exceptions are logged and reviewed.

Audit Procedures

1. Identify the DQ rule set (null checks, schema checks, range checks).
2. Request **DQ execution logs** for last 3 cycles.
3. Verify rules executed successfully (no silent failures).
4. Select **a sample of exceptions** and confirm documented review and closure.
5. Confirm whether unresolved exceptions halted or bypassed model runs.

Evidence

- DQ logs (timestamped)
- Exception reviews
- Ticketing system screenshots

Common Failures

- Exceptions logged but not reviewed
- DQ checks run manually

1.3 - Restricted Data Access & Sensitive Data Safeguarding

Control Objective
Access restrictions enforce least privilege; sensitive data is sanitized, masked, or excluded.

Audit Procedures

1. Obtain access lists for:
    - Training datasets
    - Production datasets
    - Feature stores
2. Compare user access against roles (least privilege).
3. Review quarterly access reviews for evidence of challenge/removal.
4. Evaluate masking/PII removal procedures for training data.
5. Confirm no sensitive attributes are used for model inputs unless justified (e.g., for fairness analysis).

Red Flags

- Shared service accounts
- No PII masking in training datasets
- No quarterly access review

DOMAIN 2 — Model Design & Development

2.1 - Documented Requirements

Audit Procedures

1. Obtain Business Requirements Document (BRD).
2. Confirm documentation includes:
    - Purpose
    - Assumptions
    - KPIs
    - Performance thresholds
    - Limitations
    - Ethical considerations
3. Confirm alignment between requirements and model implementation.

Evidence

- BRD
- Model cards
- Architecture diagrams

2.2 - Algorithm Selection Justification

Audit Procedures

1. Request documentation of algorithm selection (alternatives, tradeoffs).
2. Assess whether criteria included:
    - Accuracy
    - Complexity
    - Explainability
    - Fairness impacts
    - Computational cost
3. Check for management approval of the final selection.

Red Flags

- Only one algorithm considered
- No fairness vs. accuracy tradeoff analysis

2.3 - Version-Controlled Code

Audit Procedures

1. Inspect Git repository for branching strategy.
2. Confirm all model artifacts are versioned:
    - Code
    - Hyperparameters
    - Training data snapshots
    - Feature engineering scripts
3. Recreate a prior version using commit history to confirm reproducibility.

Evidence

- Git logs
- Tags/releases
- CI/CD Artifact

DOMAIN 3 — Model Validation

3.1 - Independent Model Validation

Audit Procedures

1. Obtain latest validation report.
2. Verify validator independence (separate reporting line).
3. Confirm validation coverage:
    - Accuracy testing
    - Stress testing
    - Fairness assessment
    - Edge-case testing
    - Overfitting/underfitting analysis
4. Review remediation actions and confirm closure.

Evidence

- Validation reports
- Test outputs
- Validation scripts

3.2- Accuracy & Fairness Testing

Audit Procedures

1. Review accuracy metrics vs. thresholds.
2. Inspect confusion matrices or business KPIs.
3. Review fairness metrics output:
    - Demographic parity
    - Error parity
    - Equalized odds
4. Validate sampling and representativeness.

Red Flags

- Disparate error rates across groups
- No fairness testing conducted

DOMAIN 4 — Explainability & Transparency

4.1 - Explainability Mechanisms

Audit Procedures

1. Request SHAP/LIME notebooks.
2. Select 10 random test cases and independently review SHAP outputs.
3. Confirm explanation consistency across similar inputs.
4. Validate documentation of limitations of explainability.

4.2 - User-Facing Explanations

Audit Procedures

1. Review end-user interfaces or reports.
2. Confirm explanations match model behavior and are understandable.
3. Interview users to validate comprehension.

DOMAIN 5 — Model Drift & Performance Monitoring

5.1 - Drift Monitoring

Audit Procedures

1. Request drift logs for last 6–12 months.
2. Verify:
    - Data drift
    - Concept drift
    - Feature stability
    - Performance decay
3. Select 3 drift alerts and confirm review and remediation.

5.2 - Retraining Procedures

Audit Procedures

1. Obtain retraining policy.
2. Confirm retraining triggers (data drift, schema change, threshold decline).
3. Review latest retraining cycle documentation.
4. Confirm revalidation after retraining.

DOMAIN 6 — Access, Security & Change Management

6.1 - Role-Based Access Control

Audit Procedures

1. Review access lists to dev, test, prod environments.
2. Confirm segregation of duties (dev ≠ prod).
3. Assess whether access reviews occur quarterly.

6.2 - Change Control

Audit Procedures

1. Obtain last 5 change tickets related to:
    - Model updates
    - Parameter changes
    - Pipeline changes
2. Confirm:
    - Approvals
    - Testing
    - Rollback plans
    - Deployment sign-off

DOMAIN 7 — Responsible Use & Governance

7.1 - AI Acceptable Use Policy

Audit Procedures

1. Review the AI Acceptable Use Policy.
2. Confirm:
    - Required training
    - Sign-off process
    - Monitoring of compliance
3. Inspect logs for violations.

7.2 - Prompt Governance (GenAI)

Audit Procedures

1. Request prompt templates and prohibited data guidelines.
2. Inspect prompt logs for sensitive data exposure.
3. Confirm user awareness and approval workflows.

DOMAIN 8 — Output Quality & Human Oversight

8.1 - Human Review

Audit Procedures

1. Select 25 model decisions flagged for human review.
2. Confirm approvals, comments, timestamps.
3. Assess escalation for conflicting or uncertain outputs.

8.2 - Output Accuracy Monitoring

Audit Procedures

1. Reperform a sample of 20 model predictions using known ground truth.
2. Compare differences and confirm investigation of discrepancies.
3. Review hallucination logs for GenAI models.