

FREE RESOURCE — AUDIT CHECKLIST

AI Data Quality Audit Checklist

A thorough audit of your data quality readiness for AI and machine learning projects — covering completeness, accuracy, freshness, governance, integration and bias assessment across 50+ critical evaluation items. Score each section to surface the data risks that derail model training.

6

AUDIT
SECTIONS
COVERED

50+

AUDIT ITEMS
TO CHECK

/10

SCORE EACH
SECTION

Fillable

TICK & TYPE
IN ANY VIEWER

COMPLETENESS

ACCURACY

GOVERNANCE

BIAS & FAIRNESS

PREPARED FOR

Cloudswitched Knowledge Library
Cloudswitched Ltd.

PREPARED BY

VERSION

2026 Edition

FORMAT

Interactive PDF

00

How to use this checklist

Complete this audit **before beginning any AI or machine learning project**. Work with your data owners, database administrators and business analysts to assess each item honestly. Tick every checkpoint that is fully in place — not partially, not "we're working on it", but verifiable today against a dataset, a report or a policy.

Award each section a score out of 10 based on how many checkpoints you can confidently tick. Sections scoring **below 6** indicate significant data quality risks that should be remediated **before model training begins**. This is an **interactive PDF** — tick the boxes and type your scores and notes directly in any modern PDF viewer.

WHAT THIS CHECKLIST IS

A baseline readiness assessment for teams preparing data for AI and machine learning — covering whether the data exists, is accurate, is current, is well-governed, can be integrated and is free from harmful bias. It is not a substitute for a full MLOps maturity review, but it is a fast way to surface the data gaps that quietly sink model accuracy and fairness.

SCORING GUIDANCE

Award one point per checkpoint where you have evidence the control is in place — a profiling report, a data dictionary, a lineage diagram, a fairness metric, a signed-off DPIA. Half measures do not score. Round each section to the nearest whole number out of 10 and carry it forward to the summary on page 9.

The six sections

- **01 Data Completeness & Availability** — does the data exist, is it accessible, is the volume sufficient.
- **02 Data Accuracy & Consistency** — validation, deduplication, formats, units, outliers.
- **03 Data Freshness & Timeliness** — update cadence, latency, retraining triggers.
- **04 Data Governance & Documentation** — ownership, dictionary, lineage, RBAC, UK GDPR.
- **05 Data Integration & Accessibility** — APIs, common keys, joins, formats, sandbox.
- **06 Bias & Fairness Assessment** — representation, proxies, fairness metrics, monitoring.

01

Data Completeness & Availability

Assess whether the data you need for your AI initiative actually **exists**, is **accessible**, and has sufficient volume and coverage to train reliable models. The most expensive failures happen when a project starts before anyone confirms the data is there.

- ☐ All required data fields for your target AI use case have been **identified and documented** in a data requirements specification (*feature list*).
- ☐ The **percentage of missing values** in each required field has been measured — fields with more than 20% missing data require remediation before use in model training.
- ☐ Data covers a **sufficient time period** to capture seasonal patterns, cyclical trends and edge cases relevant to your domain (*minimum 12–24 months recommended*).
- ☐ The **volume of labelled training examples** is adequate for your chosen approach — simple classifiers need thousands; complex NLP or vision models need tens of thousands or more.
- ☐ Data from **all relevant business units and systems** has been identified and access permissions secured, including legacy systems that may hold critical historical records.
- ☐ There are **no critical data gaps** that would prevent the model from learning key patterns — e.g. customer churn data without the preceding behaviour signals is insufficient.
- ☐ A **data collection plan** is in place for any missing data, with realistic timelines for accumulating sufficient volume before model training begins.
- ☐ **Sample data** has been extracted and reviewed manually by domain experts to confirm it contains the information expected and is representative of real-world conditions.

WHAT GOOD LOOKS LIKE

A documented feature list maps every required field to a confirmed source, with a measured missingness rate per field and a domain expert sign-off that a real sample looks like the world it claims to describe.

SECTION 01 SCORE / 10

Aim for 8+. Below 6 = data availability gap.

02

Data Accuracy & Consistency

Evaluate whether your data is **truthful**, consistently formatted and free from errors that would cause AI models to learn incorrect patterns. Garbage in, garbage out is not a cliché here — it is the single most common cause of failed models.

- ☐ A **random sample** of records (*minimum 5%*) has been manually validated against source documents to measure the baseline accuracy rate across key fields.
- ☐ **Data entry standards and validation rules** are enforced at the point of capture — free-text fields where structured data is expected are a common source of inconsistency.
- ☐ **Duplicate detection** has been performed across all datasets, with a documented deduplication strategy for exact and fuzzy matches (*e.g. "Ltd" vs "Limited"*).
- ☐ **Date and time fields** use a consistent format across all sources — mixed formats (DD/MM/YYYY, MM/DD/YYYY, YYYY-MM-DD) cause silent errors in training.
- ☐ **Numerical fields** use consistent units and scales — revenue in pounds vs pence, distances in miles vs kilometres, temperatures in Celsius vs Fahrenheit must be standardised.
- ☐ **Categorical fields** use a controlled vocabulary rather than free text — "United Kingdom", "UK", "U.K." and "Great Britain" should all map to a single canonical value.
- ☐ Data from **merged systems or acquisitions** has been reconciled and harmonised — legacy data often uses different coding schemes, field names and business rules.
- ☐ **Outlier detection** has been performed on numerical fields to identify implausible values (*age 999, negative quantities, future dates*) that may distort training.
- ☐ A **data quality scorecard** has been produced summarising accuracy, completeness and consistency metrics for each dataset, with red/amber/green status indicators.

WATCH OUT FOR

Silent format drift across sources — two systems writing the same date field in different orders, or one storing revenue in pence and another in pounds. These errors never throw an exception; they just quietly poison the model.

SECTION 02 SCORE / 10

Aim for 8+. Below 6 = accuracy / consistency gap.

03

Data Freshness & Timeliness

AI models trained on stale data produce stale predictions. Assess whether your data is **current enough** to drive reliable AI outputs, and whether the pipeline that keeps it fresh is automated and monitored.

- ☐ The **last update timestamp** for each data source has been documented, confirming data is current enough for your use case (*real-time, daily, weekly or monthly refresh*).
- ☐ You have defined the **maximum acceptable data latency** — a fraud model needs near real-time data; a quarterly forecast can tolerate daily updates.
- ☐ Data refresh schedules are **automated and monitored** — manual exports that rely on individuals remembering to run them are unreliable and will eventually be forgotten.
- ☐ **Historical data has not been retroactively modified** in ways that distort training — backdated corrections should be flagged and handled appropriately.
- ☐ You have assessed whether the **business environment has changed** significantly since the training data was collected — post-pandemic behaviour may not match pre-pandemic data.
- ☐ A **data freshness monitoring system** alerts the team when data pipelines fail or data has not been updated within expected timeframes.
- ☐ The **training/serving data gap** has been measured — if the model trained on data up to last month but serves on today's data, understand what might have changed in between.
- ☐ There is a defined **retraining trigger** based on data freshness — when new data accumulates beyond a threshold or metrics degrade, retraining is initiated automatically.

CONCEPT DRIFT IS SILENT

A model does not announce when the world has moved on from its training data. A freshness monitor plus a defined retraining trigger is the only way to catch drift before it shows up as falling business outcomes.

SECTION 03 SCORE _____ / 10

Aim for 8+. Below 6 = freshness / drift risk.

04

Data Governance & Documentation

Well-governed data with clear documentation is essential for **reproducible, auditable and trustworthy** AI systems. If you cannot explain where a number came from, you cannot defend the decision a model made with it.

- ☐ Every dataset used for AI has a designated **data owner** accountable for its quality, accuracy and appropriate use within the organisation.
- ☐ A **data dictionary** documents every field, its definition, data type, allowed values and business meaning — essential for new team members and auditors.
- ☐ The **lineage** of each dataset is documented — where it originated, how it was transformed and what business rules were applied during processing.
- ☐ Data access is governed by **role-based access controls (RBAC)** ensuring only authorised personnel can read, modify or export data used for AI training.
- ☐ A **data classification scheme** categorises all datasets by sensitivity (*public, internal, confidential, restricted*) with corresponding handling requirements.
- ☐ Changes to data schemas, business rules or source systems are managed through a **formal change control process** with impact assessment and stakeholder notification.
- ☐ All datasets containing personal data have been assessed under **UK GDPR**, with lawful basis documented, retention periods defined and data subject rights procedures in place.
- ☐ A **metadata management system or catalogue** tracks dataset versions, quality metrics, usage history and dependencies across the organisation.
- ☐ There is a documented process for **decommissioning datasets** no longer needed, including secure deletion and removal from all downstream pipelines and models.

COMPLIANCE EXPOSURE

Training a model on personal data without a documented lawful basis under UK GDPR — and without a route to honour erasure requests — is a regulatory finding waiting to happen. Govern the data before it reaches the model, not after.

SECTION 04 SCORE / 10

Aim for 8+. Below 6 = governance gap.

05

Data Integration & Accessibility

AI projects often require combining data from multiple systems. Assess how easily your data can be **integrated and made available** for model training — data that cannot be extracted is data that cannot be used.

- ☐ All relevant data sources can be accessed **programmatically** via APIs, database connections or file exports — data locked in proprietary systems without export is a significant blocker.
- ☐ A **common identifier** (*customer ID, employee number, product code*) exists across systems to enable reliable joins and lookups between different data sources.
- ☐ Data from different systems has been **test-joined** to verify matching works — mismatched keys, orphaned records and many-to-many relationships have been identified and resolved.
- ☐ Your **data warehouse or lakehouse** can handle the query volumes required for training — extracting millions of rows should not cripple production database performance.
- ☐ **Real-time or near-real-time access** is available for use cases that require it — streaming pipelines (Kafka, Event Hubs) have been evaluated where batch is insufficient.
- ☐ Data is available in **formats compatible with your AI platform** — CSV, Parquet, JSON or direct connections, with appropriate character encoding (*UTF-8*) for international data.
- ☐ There is a **sandbox or development environment** where data scientists can explore and experiment without risking production systems or violating data protection policies.
- ☐ **Network connectivity** between data sources and the AI platform has been tested, with sufficient bandwidth and acceptable latency for transferring large training datasets.

THE JOIN IS WHERE IT BREAKS

Most integration projects discover too late that two systems share no reliable key, or that a "one-to-one" relationship is actually many-to-many. Test-join early on real data — before you build features on an assumption that does not hold.

SECTION 05 SCORE / 10

Aim for 8+. Below 6 = integration blocker.

06

Bias & Fairness Assessment

AI models amplify patterns in data — **including biases**. Proactively assess your data for sources of bias before training begins, because a model trained on biased history will faithfully reproduce and amplify it.

- ☐ You have analysed **demographic representation** in your training data to identify whether certain groups are over- or under-represented relative to the population the model will serve.
- ☐ **Historical decisions** encoded in your data have been reviewed for systemic bias — if past decisions were biased, training on them will perpetuate and amplify that bias.
- ☐ Data collection methods have been assessed for **selection bias** — data gathered only from certain channels, time periods or segments may not represent the full picture.
- ☐ **Proxy variables** that correlate with protected characteristics (*postcode for ethnicity, name for gender*) have been identified and their impact on fairness evaluated.
- ☐ A **fairness metric** has been selected for your use case — equal opportunity, demographic parity or predictive equality — and baseline measurements have been established.
- ☐ You have a plan for **ongoing bias monitoring** in production, including regular audits of model outputs across demographic groups to detect emerging disparities.
- ☐ The **ICO's guidance** on AI and data protection has been reviewed, with attention to automated decision-making provisions under UK GDPR Article 22.
- ☐ Your team includes individuals with **diverse perspectives** who can identify potential biases a homogeneous team might overlook during data review and model evaluation.
- ☐ There is a documented **remediation process** for when bias is detected — model retraining, data rebalancing, algorithmic constraints or removal of problematic features.

ARTICLE 22 APPLIES

Where a model makes automated decisions with legal or similarly significant effects on individuals, UK GDPR Article 22 grants rights to explanation and human review. A documented fairness metric and remediation process is not optional — it is part of being lawful.

SECTION 06 SCORE / 10

Aim for 8+. Below 6 = fairness risk.



Audit score summary

Transfer the score for each section into the table. Set a priority (H/M/L) based on the gap to target. Any section below 6 is a candidate for the top-3 actions on the next page.

#	AUDIT AREA	SCORE / 10	PRIORITY
01	Data Completeness & Availability	_____ / 10	☐ H ☐ M ☐ L
02	Data Accuracy & Consistency	_____ / 10	☐ H ☐ M ☐ L
03	Data Freshness & Timeliness	_____ / 10	☐ H ☐ M ☐ L
04	Data Governance & Documentation	_____ / 10	☐ H ☐ M ☐ L
05	Data Integration & Accessibility	_____ / 10	☐ H ☐ M ☐ L
06	Bias & Fairness Assessment	_____ / 10	☐ H ☐ M ☐ L
Σ	TOTAL SCORE	_____ / 60	—

Score interpretation

48–60

Excellent. Your data is AI-ready. Focus on monitoring, drift detection and ongoing fairness audits.

36–47

Good foundation, gaps exist. Prioritise any section scoring below 6 with owners and deadlines before training begins.

Below 36

Significant gaps. Remediate data quality before model training — consider a structured data readiness review with a specialist.



Priority actions & notes

Using the readiness band and section scores from the summary, commit to a small number of next steps. The goal is one page of decisions, not a wish list — pick the lowest-scoring sections you flagged High and turn them into owned, dated actions.

Top 3 priority actions

01

02

03

Additional notes

AUDIT COMPLETED BY

DATE

NEXT REVIEW DUE

Getting your data AI-ready?

Cloudswitched helps UK organisations prepare data for AI and machine learning — profiling, governance, integration and fairness assessment, with fixed-price remediation roadmaps.

info@cloudswitched.com

cloudswitched.com



CLOUDSWITCHED

IT SUPPORT & PROJECT SERVICES

info@cloudswitched.com

New London House, 6 London St
London EC3R 7LP · United Kingdom