SHIPSHAPE DATA

AI/ML Data Readiness Checklist

7 Bell Yard, London, WC2A 2JR

+44 20 3143 1421

www.shipshapedata.com

A practical guide to building reliable AI systems

Artificial intelligence is becoming part of everyday business operations, influencing analysis, reporting, and decision-making across organisations. As adoption increases, so does the need for confidence in how these systems are designed, deployed, and maintained.

This checklist has been created to help teams assess whether the foundations supporting their AI and data initiatives are ready for production use. It focuses on the practical areas that most directly influence reliability, trust, and long-term performance.

Rather than concentrating on models alone, this guide looks at the broader environment around AI, including data quality, system stability, governance, and operational readiness. These elements play a significant role in determining whether AI performs consistently once it moves beyond experimentation and into real-world use.

The checklist is designed to be:

- Clear and practical
- Relevant to both technical and business stakeholders
- Applicable across different industries and use cases
- Focused on outcomes, not theory

It can be used as a self-assessment tool, a discussion framework, or a starting point for improving existing processes. Not every organisation will be at the same stage, and not every section will apply equally. The value comes from gaining visibility into strengths, gaps, and areas for improvement.

Reliable AI is not the result of a single decision. It is built over time, through structure, clarity, and good operational practice.

This checklist is intended to support that journey.

Data Ownership & Accountability

Understanding who is responsible for what when things break helps get services back up and running quickly. Clear ownership removes uncertainty, reduces downtime, and makes issue resolution more efficient.

☐	**Every dataset has a named owner**	Each dataset should have a clearly identified owner who takes responsibility for its accuracy, structure, and upkeep. This creates a reliable point of contact when questions arise or changes are required.
☐	**Data SLAs are defined (freshness, completeness, uptime)**	Setting expectations around how current, complete, and available data should be helps teams align on what “good” looks like. Clear standards reduce ambiguity and support more proactive management of data quality.
☐	**A clear escalation process exists**	When data issues occur, it should be obvious who to contact and how to raise concerns. A well-defined escalation path improves response times and helps issues get resolved efficiently.

Data Understanding

A strong understanding of data allows teams to work with confidence and make better decisions. When the origin, structure, and meaning of data are clear, issues are easier to identify and outcomes become more predictable.

☐	**Data sources are clearly identified**	Each dataset should have a known origin, whether it comes from an application, external provider, or internal system. Clear source visibility makes it easier to investigate issues and improves trust in downstream analysis.
☐	**Update frequency is explicitly known**	Teams should understand how often data is refreshed and how current it is expected to be. Clarity around freshness helps set realistic expectations and supports better decision-making.
☐	**Transformation steps are understood**	It should be clear how data changes as it moves through systems before reaching the model. Understanding this journey makes it easier to trace issues and maintain consistency.
☐	**Critical fields are clearly defined**	Teams should agree on which fields are essential and which are supplementary. This helps prioritise attention and ensures focus is placed on what matters most.
☐	**“Bad data” has a shared definition**	There should be a common understanding of what poor-quality data looks like. Agreeing on this early supports faster identification and resolution of potential issues.
☐	**Business meaning is documented**	Key fields should be explained in plain language. Clear definitions ensure that models reflect the intended business logic and reduce misunderstandings across teams.

Data Quality Baselines

Establishing quality benchmarks makes it easier to understand when data is performing as expected and when attention is needed. These baselines provide a practical reference point for maintaining consistency over time.

☐	**Expected row counts are defined**	Teams should have a clear view of the volume of data they normally receive. This makes unusual increases or drops easier to notice and quicker to investigate.
☐	**Schema and data types are enforced**	Keeping data structures consistent ensures systems can process information reliably. Clear definitions reduce the risk of unexpected behaviour.
☐	**Null thresholds are clear**	Knowing how much missing data is acceptable helps teams maintain consistent standards and focus effort where it has the greatest impact.
☐	**Uniqueness rules exist**	Where records are expected to be unique, this should be clearly defined. Consistency here supports more accurate reporting and analysis.
☐	**Valid ranges are set**	Defining acceptable values for numerical fields improves reliability and helps highlight unusual patterns early.

Pipeline Resilience

Data pipelines and ML models usually run in quiet harmony. When something does go wrong, knowing where to focus attention is key to restoring services quickly and minimising disruption.

☐	**Ingestion is automated**	Automating how data enters systems supports consistency and reduces avoidable delays or errors.
☐	**Failures are immediately visible**	Clear visibility into issues helps teams respond quickly and reduces disruption across dependent services.
☐	**Systems and dashboards provide oversight**	Dashboards and monitoring systems should make pipeline health easy to understand at a glance. This helps teams identify issues quickly and prioritise the right actions.
☐	**Retry logic is in place**	Allowing systems to recover automatically from temporary issues helps maintain continuity and improves operational stability.
☐	**Safe fallback behaviour exists**	Planning for missing or invalid data enables systems to continue operating and reduces wider impact during incidents.
☐	**Latency is actively monitored**	Understanding where delays occur helps teams focus improvement efforts and maintain predictable performance.
☐	**Upstream dependencies are known**	Knowing which systems influence data flow makes it easier to assess impact and coordinate resolution efficiently.

Training vs Production Parity

Most AI failures don't come from bad models, they come from using a model on data it was never trained to understand. Consistency between training and live data is mandatory, not optional.

☐	**Training and production schemas match**	AI and data failures are often the result of discrepancies between development, UAT and production environments. Keeping structures aligned across these systems helps reduce unexpected behaviour when models go live.
☐	**Feature logic is identical**	Using the same transformations in training and production reduces variation and improves confidence in outputs.
☐	**Data distributions are compared**	Monitoring how live data compares to training data helps teams recognise meaningful change and respond appropriately.
☐	**Category changes are monitored**	Tracking new or missing values supports accurate interpretation and ongoing performance.
☐	**Live data is tested before launch**	Testing with production-like data ahead of release helps teams validate assumptions and reduce surprises.
☐	**Model assumptions are documented**	Recording key assumptions supports clarity, communication, and long-term maintainability.

Dataset Versioning

Versioning helps prevent major failures when changes are introduced. When data, features, or models are updated, versioning makes it easier to understand what changed and reduces the risk of a new release breaking existing behaviour.

☐	**Training data is versioned**	Keeping identifiable snapshots of training data allows teams to understand what a model learned from at any point in time.
☐	**Feature definitions are controlled**	Managing changes to feature logic helps teams track how data evolves and ensures models behave as expected.
☐	**Model-to-data mapping exists**	Linking models to the data they were trained on improves transparency and supports investigation when results change.
☐	**Historical datasets are stored**	Retaining past versions of data supports comparison, review and analysis as systems mature.
☐	**Change history is visible**	Clear records of what changed and when support accountability and smoother collaboration.

Observability & Alerts

Clear visibility into systems and data helps teams act with confidence. When performance and quality are easy to see, issues can be addressed before they affect outcomes.

☐	**Pipeline health is visible**	Teams should have a clear view of whether data is flowing as expected, delayed, or interrupted. This supports timely and informed responses.
☐	**Anomalies are logged and reviewed**	Recording unexpected events creates a reliable history that helps teams understand patterns and improve over time.
☐	**Model confidence is observed**	Tracking confidence alongside performance metrics provides a fuller picture of reliability
☐	**Dashboards support decision-making**	Information should be accessible and easy to interpret. Well-designed dashboards help teams focus on what matters most.
☐	**Alerts reach the right people**	Notifications should be directed to those best placed to respond, improving resolution times.
☐	**Response steps are clear**	When alerts occur, teams should know what actions to take and where responsibility sits.

Security & Compliance

Strong governance protects both your data and your reputation. Effective security and compliance processes reduce risk and support responsible use of AI across the organisation.

☐	**Access is role-controlled**	Access to data should be based on role and responsibility. This helps ensure sensitive information is only handled by those who need it.
☐	**Sensitive data is protected**	Personal and confidential data should be encrypted and handled appropriately. This supports trust and reduces exposure.
☐	**Audit trails are in place**	Clear records of data access and changes support accountability and help resolve issues quickly.
☐	**Data retention is defined**	Knowing what data to keep, and for how long, supports compliance and reduces unnecessary risk.
☐	**Regulatory requirements are understood**	Teams should be aware of applicable obligations and reflect them in how data is handled and stored.
☐	**Credentials are managed securely**	Passwords, keys and tokens should be stored and managed safely. This reduces the risk of accidental exposure.

Go-Live Gate

Releasing a model into production is a significant operational step. A clear go-live process helps teams move forward with confidence and reduces the risk of disruption.

☐	**Data owner approval is given**	Formal sign-off confirms that the data meets agreed expectations and is ready for use.
☐	**Rollback is prepared**	Having a rollback approach in place makes it easier to restore services quickly if unexpected issues arise.
☐	**Fallback behaviour exists**	When predictions cannot be delivered as planned, alternative processes should support continuity.
☐	**Monitoring is live**	Performance, data quality and drift checks should be active from the outset to provide early insight.
☐	**Documentation is available**	Clear documentation helps teams understand how systems work and how to respond when changes are needed.
☐	**Ownership continues after launch**	Responsibility remains in place beyond release, supporting reliability as the system evolves.

Building reliable AI is not about reaching a final destination. It is about establishing practices that allow systems to evolve while remaining stable, trustworthy, and well governed.

No organisation completes this checklist perfectly on day one. Each section highlights areas where greater clarity, structure and oversight can strengthen outcomes. The aim is not perfection, but progress.

Used regularly, this checklist can support:

- Better decision-making
- Faster issue resolution
- Stronger collaboration between business and technical teams
- Greater confidence in AI-driven outcomes

As AI systems become more embedded in everyday operations, the foundations that support them make an increasing difference to long-term success.

About Shipshape Data

Shipshape Data supports organisations in building dependable data foundations for AI and analytics. We work with teams to improve data quality, strengthen pipelines, and increase operational visibility, helping AI systems perform more consistently in real-world environments.

If you would like to discuss any areas highlighted in this checklist, you can find more information about our services at:

www.shipshapedata.com