ML Pilot to Production

Six steps to assess whether an AI proof of concept is ready for further production investment. Built from the Devox Software article workflow; RAG-specific examples are labeled where relevant.

01 Review what the PoC actually proved

Separate model performance under controlled tests from claims about operational reliability. Passing an offline evaluation is evidence of feasibility, not a complete release decision.

Check: Evaluation coverage, untested behavior, and remaining assumptions.

Decision output: A documented evidence gap list.

02 Identify production failure modes

Examine how conditions can change outside the model. In RAG applications, stale retrieval indexes may produce plausible but outdated answers without causing application errors.

Check: Whether indexed data remains aligned with current authoritative records.

Decision output: A list of relevant failure mechanisms and their consequences.

03 Scope the supporting data engineering

Assess data ingestion, cleaning, preprocessing, document versioning, metadata synchronization, and index refresh where relevant. These may represent substantial work beyond the pilot.

Check: Which components exist, which are only experimental, and which remain unbuilt.

Decision output: A production engineering scope and dependency list.

Validate, operate, decide

04 Gather production-condition evidence

Use predeployment evaluation and runtime observability as distinct sources of evidence. Consider latency and cost when applying model-based scoring to production traffic.

Check: Which live behaviors are observable and which remain unverified.

Decision output: Evidence to inform a controlled expansion decision.

05 Define continuing maintenance responsibilities

Plan for ongoing model observation and maintenance of supporting information. For RAG, versioning and index refresh must prevent obsolete content from remaining retrievable.

Check: Who will maintain each production component and how issues will be investigated (assignments require project-specific validation).

Decision output: An operations responsibility and maintenance plan.

06 Make the investment decision

Compare remaining engineering requirements and recurring infrastructure/support costs with the organization's business case. Do not treat successful pilot results as automatic approval to scale.

Check: Cost assumptions, unresolved production evidence, and expected business value.

Decision output: Proceed, extend validation/redesign, or stop.

Final decision gate

PROCEED	Required evidence and costs are sufficiently understood for the specific business case.
VALIDATE / REDESIGN	The use case remains promising but material dependencies or behavior are still unverified.
STOP	Expected value does not justify remaining engineering, operating commitments, or risk.

Evidence note: This guide reflects the supplied article and its retained research. It is not a validated universal AI readiness checklist. General PoC exit thresholds, shadow-running procedures, and organizational ownership models were not established in the supplied sources.