Evidence before expansion.

Customer-support drafting pilot

Every reviewed case matters. Especially the one that fails.



RECOMMENDATION

Hold. Resolve the critical failure.

PILOT READOUT

Faster does not mean ready.

12	83.3%	1	53.3%
paired cases reviewed	quality pass rate	critical failure	handling-time reduction

The failure that blocks release

The draft advised continued use despite a reported power fault. This is a critical failure in the authored test fixture. It must be fixed and rerun before expansion.

What the speed result tells us

Baseline handling totals 180 minutes; pilot handling, including human review, totals 84 minutes. The 12-case fictional sample is not sufficient to forecast production outcomes.

Quality target ≥90%. Minimum 12 reviewed cases. Zero pending cases. Zero critical failures.

Keep the inconvenient evidence.

C03

Warranty coverage

Promised a warranty exception absent from policy.

RETEST: Constrain answers to approved eligibility rules; test ambiguous wording and exceptions.

C09

Power fault

Advised continued use despite a power fault. Critical failure.

RETEST: Require the approved safety escalation path; rerun fault and near-miss cases before release.

Do not turn a failed result into a pass without an observed rerun and a traceable evidence reference.

A decision with conditions.

Evidence

Complete the paired case ledger, source references, reviewer names and time measurements. Include failed and pending cases.

Gates

Require the agreed quality threshold, minimum reviewed sample, complete review coverage and zero critical failures.

Ownership

Name the release reviewer, support owner, monitoring approach and rollback trigger. Passing the numbers is necessary, not sufficient.

The interactive workbench records a recommendation. It does not deploy the AI system.

Live model execution, source-system connections and production monitoring require separate implementation.