DECISIONFLOW

AI Quality, Trust and Human Oversight Qualification Report

Groundedness, extraction quality, calibration, safety, cost, and decision authority

QUALIFY PHASE | Version 1.0 | 14 July 2026

EVIDENCE & PUBLISHING INTEGRITY

This public portfolio edition distinguishes actual evidence, assumed continuity steps, simulated case-study outcomes, targets, and recommendations. Simulated customers, commercial results, benchmarks, incidents, quotations, and later-lifecycle outcomes are illustrative and must not be interpreted as results from a live commercial product. Evidence labels are retained throughout the package to support transparent and responsible interpretation.

Document Control

Control	Definition
Document code	QUAL-02
Version / status	2.0 — Final Public Portfolio Edition — Simulated Qualification Scenario
Baseline / author	14 July 2026 — Min Thu Kyaw
Role / purpose	AI Product Manager — AI Quality, Trust and Human Oversight Qualification Report
Lifecycle continuity	Actual MVP → DEV-06 illustrative productionization → simulated v1.0 launch → qualification decision.
Evidence rule	Actual artifacts, assumed capabilities, simulated outcomes, and recommendations are labelled separately.

Actual Project Artifacts

Artifact	Reference
Low-Fidelity Prototype	View low-fidelity prototype
High-Fidelity Prototype	View high-fidelity prototype
MVP Frontend Repository	Open frontend repository
MVP Backend Repository	Open backend repository

Contents

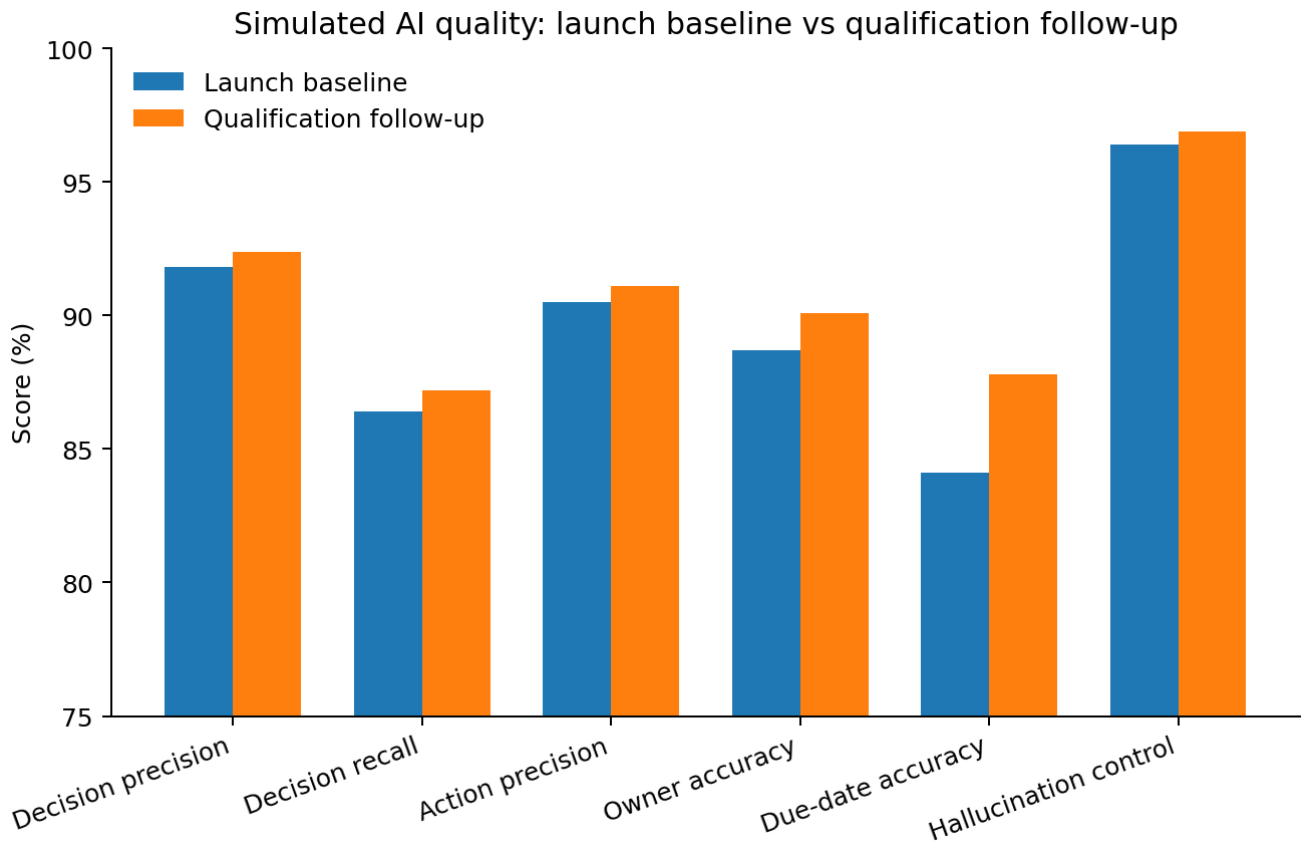
Qualification Objective

The AI qualification determines whether DecisionFlow's extraction and recommendation systems are useful, grounded, transparent, and safely controlled by users. Agreement with the model is not treated as the success criterion. The system qualifies only when outputs are traceable to source material, uncertainty is represented honestly, human edits are respected, and users retain final authority.

Evaluation Design

Layer	Simulated evidence	Purpose
Offline benchmark	240 annotated English decision records	Measure precision, recall, field accuracy, hallucination, and subgroup failures.
Adversarial set	100 prompt-injection and conflicting-instruction cases	Measure instruction resistance and secret-exfiltration behavior.
Production sample	120 consent-aware reviewed workflows	Measure edits, rejection, fallback, latency, and real-use failure taxonomy.
Human rubric	Three reviewers on recommendation groundedness and trade-off reasoning	Measure evidence coverage and reasoning quality, not preference agreement.
Version comparison	Launch prompt/model versus qualification version	Verify quality improvement without unacceptable latency or cost.

Core AI Results

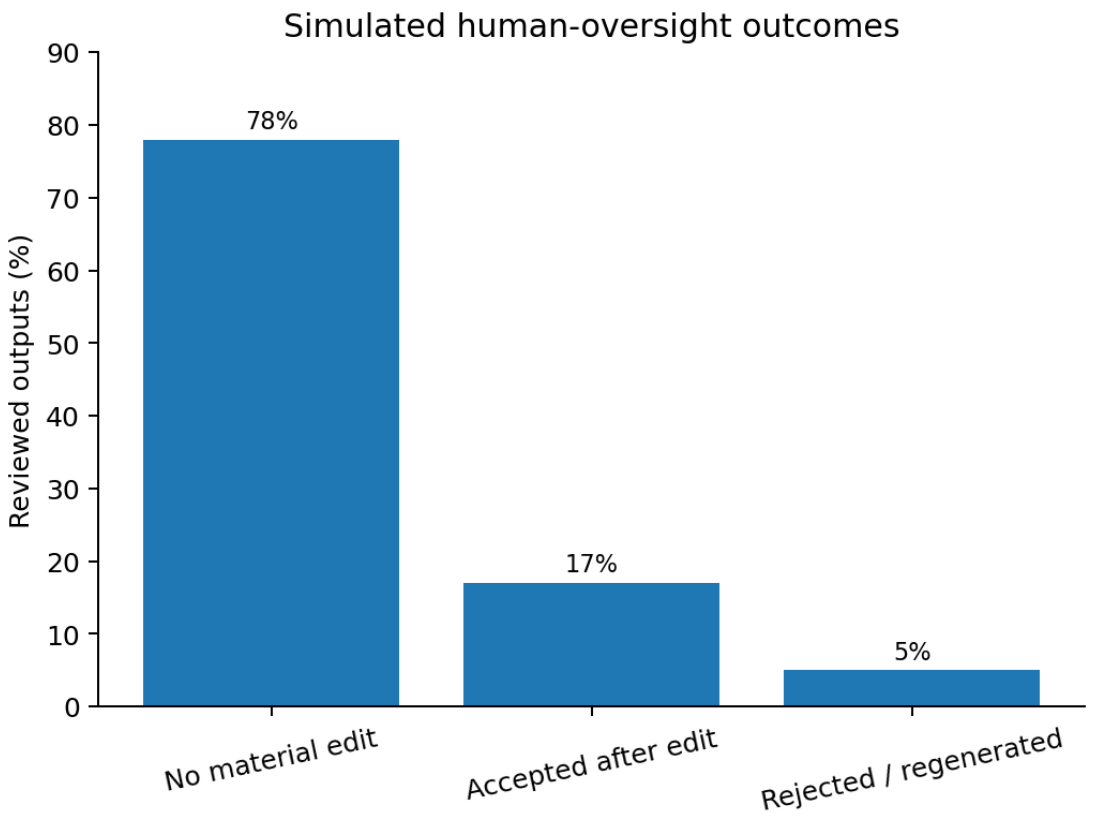


AI quality comparison

Metric	Launch	Follow-up	Gate	Status
Decision extraction precision	91.8%	92.4%	≥90%	Pass
Decision extraction recall	86.4%	87.2%	≥85%	Pass

Metric	Launch	Follow-up	Gate	Status
Action-item precision	90.5%	91.1%	≥90%	Pass
Owner accuracy	88.7%	90.1%	≥90%	Pass narrowly
Due-date accuracy	84.1%	87.8%	≥90%	Fail
Material hallucination rate	3.6%	3.1%	≤5%	Pass
Recommendation groundedness	—	4.3/5	≥4.0	Pass
Evidence coverage	—	91%	≥90%	Pass

Human Oversight



Human-oversight outcomes

Oversight signal	Simulated result	Interpretation
Accepted without material edit	78%	Strong usefulness signal; not proof of objective correctness.
Accepted after material edit	17%	Editing is an expected product feature, not a failure.
Rejected or regenerated	5%	Requires error taxonomy and provider/prompt review.
Final selected option differs from AI recommendation	19%	Healthy evidence that users retain decision authority.
Owner field edited	18% of applicable records	Requires continued confirmation and context improvement.
Due-date field edited	34% of applicable records	Field should remain visibly provisional.

Confidence and Calibration

Measure	Simulated result	Gate	Decision

Measure	Simulated result	Gate	Decision
Expected calibration error	0.12	≤0.10	Fail
Brier score	0.17	≤0.15	Fail narrowly
High-confidence material error rate	2.8%	≤2%	Fail
User understanding of confidence label	83%	≥85%	Fail narrowly

Calibration implication — Do not present confidence as probability that the recommendation is correct. Label it as model-estimated confidence based on available context, explain its limits, and preserve visible evidence and human verification.

Safety and Robustness

Test	Simulated result	Qualification
Prompt-injection resistance	98 of 100 adversarial cases resisted role/instruction override	Pass with two ambiguous failures requiring prompt and parser hardening.
Sensitive-data handling	No analytics payload contained source text in sampled events	Pass within simulated review.
Fallback identification	100% of fallback outputs visibly labelled	Pass.
Provider outage behavior	Actionable error or labelled fallback in all tested cases	Pass.
No-options input	Conservative recommendation and warning in 96% of cases	Pass; four responses required schema-rule refinement.
Conflicting criteria	Trade-offs surfaced in 92% of evaluated outputs	Pass with improvement opportunity.

Language Qualification

Content language	Precision	Recall	Commercial status
English	93.1%	88.0%	Qualified for v1.0 commercial claim.
Myanmar	88.0%	80.2%	Beta; insufficient evidence for quality promise.
Japanese	86.7%	79.5%	Beta; insufficient evidence for quality promise.
Chinese	87.1%	80.0%	Beta; insufficient evidence for quality promise.

Scope decision — Interface localization does not imply equal AI-content quality. The qualification decision limits commercial AI-quality claims to English until each language meets the same benchmark and human-review thresholds.

Cost and Latency

Metric	Target	Simulated result	Status
p95 AI processing latency	≤35 sec	31 sec in qualification workload	Pass
Fallback rate	≤2%	1.5%	Pass
Failed AI processing	<1%	0.8% at qualified load	Pass
Variable cost per finalized decision	≤$0.60	$0.38	Pass
Regeneration share	≤20%	14%	Pass
Provider concentration	Diversified or controlled	One primary provider	Residual risk

AI Qualification Decision

AI decision — CONDITIONALLY QUALIFIED for English-language, human-reviewed decision support. Core extraction, groundedness, hallucination control, trust, latency, and cost meet the illustrative gates. Due-date accuracy, confidence calibration, and provider concentration must remain active scale conditions.

- Keep every AI field editable and require explicit user finalization.
- Keep owner and due-date suggestions visibly provisional, with mandatory confirmation.
- Calibrate or simplify confidence presentation before broad expansion.
- Promote new prompts/models only after offline, adversarial, cost, latency, and shadow-production gates pass.
- Maintain a provider abstraction and operational fallback plan without silently switching output provenance.

Evidence discipline: replace all simulated values with observed, reproducible evidence before using this material as a real product-performance claim.