

DECISIONFLOW

AI Scale, Quality and Model Operations Plan

Maintaining grounded, attributable, human-controlled AI behavior as volume and model-routing complexity increase

CONTROLLED SCALE PHASE | Version 1.0 | Illustrative 12-month bounded-growth program

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	CS-04
Version / status	2.0 — Final Public Portfolio Edition — Simulated Controlled-Scale Scenario
Baseline / author	14 July 2026 - Min Thu Kyaw
Scenario horizon	Illustrative 12-month period after Deliver; outcome data is simulated.
Lifecycle continuity	Actual MVP -> assumed productionization -> simulated Launch -> Qualify -> Deliver -> Controlled Scale.
Scale decision	Grow the individual workflow within tested customer, AI, commercial, support, and capacity boundaries.
Scope boundary	No broad collaboration, integrations, richer ingestion, multilingual promise, SSO, or enterprise-platform claim.
Evidence rule	Actual, assumed, simulated, target, and recommendation statements are separated.

Actual Project Artifacts

Artifact	Reference
Low-Fidelity Prototype	View prototype
High-Fidelity Prototype	View prototype
MVP Frontend Repository	Open repository
MVP Backend Repository	Open repository

Contents

01	AI scale objective
02	Quality contracts and evaluation
03	Routing and cost policy
04	Monitoring, drift and provenance
05	Scaled results
06	Incident and change governance
07	AI scale decision

AI Scale Objective

AI scale means that higher request volume and cost optimization do not weaken DecisionFlow's product principle: AI assists; the user decides. Every generation remains grounded in reviewed inputs, schema-validated, attributable to a model/prompt route, and subject to explicit human finalization.

Non-negotiable boundary - No model, prompt, routing, or fallback change can silently alter accountability fields or bypass human confirmation.

Scale risk	Required control
Cheaper route reduces quality	Fixed evaluation and accountable-field gates before routing change.
Provider output changes over time	Versioned evaluations, drift monitoring, and canary comparison.
Fallback appears authoritative	Visible fallback state and provenance in UI and records.
Higher volume hides rare failures	Stratified sampling, incident intake, and high-risk fixture expansion.
Prompt injection / sensitive content	Untrusted-input instructions, provider disclosure, user guidance, and evaluation fixtures.

Quality Contracts and Evaluation

Metric	Approved gate	Purpose
Owner accuracy	>=92%	Protect accountability assignment.
Due-date accuracy	>=90%	Protect action timing and avoid invented commitments.
Hallucination control	>=97%	Limit unsupported factual or decision claims.
Confidence calibration ECE	<=0.08	Displayed confidence should track observed correctness.
Fallback rate	<=1%	Fallback remains exceptional and visible.
Human finalization	100%	No autonomous final decision.
Evaluation slice	Required coverage	
Source quality	Complete transcript, sparse notes, conflicting statements, no options.	
Accountability	Named/unnamed owners, relative dates, ambiguous commitments.	
Safety	Prompt injection, sensitive data, role-changing instructions, secret requests.	
Language	English production claims only; other languages measured separately and not marketed broadly.	
Provider / route	Primary, alternate, fallback, and cost-aware route comparison.	
Regression	Known failures converted into immutable fixtures.	

Routing and Cost Policy

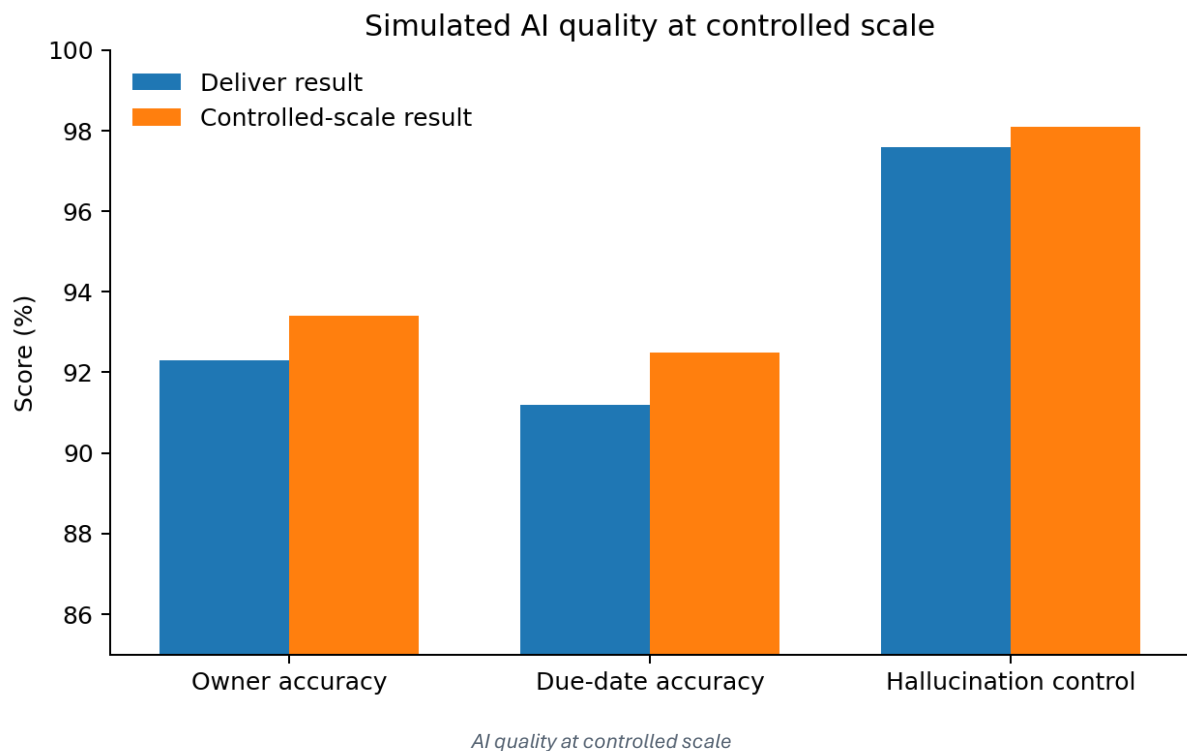
Task class	Routing policy	Guardrail
Structured extraction	Use qualified route selected by complexity and size.	Schema, field accuracy, injection, and fallback gates.
Recommendation	Use quality-focused qualified route.	Groundedness, option consistency, rationale, and confidence gates.
Regeneration	Same governance as first generation; user confirmation and version history.	No hidden downgrade due to cost.
Provider failure	Retry within policy, pause, or approved alternate route.	Provenance shown; no silent provider switch.
Heuristic fallback	Only for defined failure classes.	Clearly labeled and requires review.

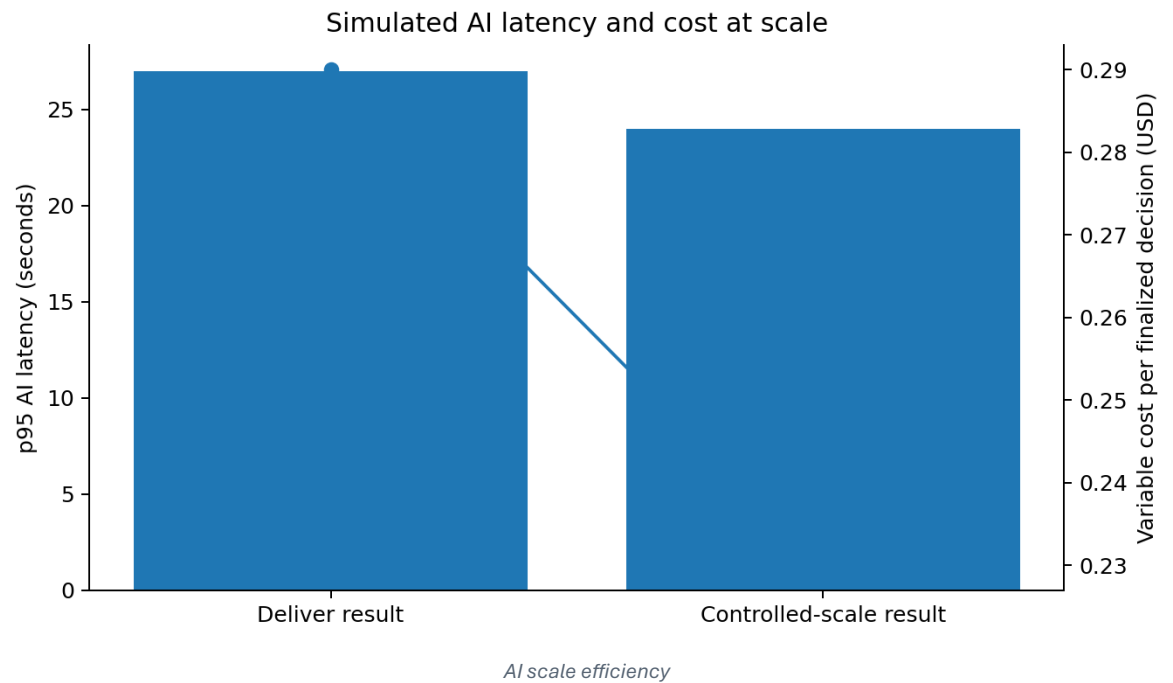
- Routing decisions use task metadata and content size, not sensitive semantic classification stored in analytics.
- Cost optimization is accepted only when quality confidence intervals and high-risk slices remain within gates.
- Model names alone are insufficient; provider, route, prompt version, evaluation version, and timestamp are retained.

Monitoring, Drift and Provenance

Signal	Frequency	Action
Online latency, failure, usage and fallback	Continuous / daily review	Operational alert and route/provider investigation.
User edit rate by field	Weekly	Detect extraction drift or unclear UI; no content in analytics.
Fixed evaluation suite	Every model/prompt/routing change and scheduled monthly run	Block, canary, approve, or rollback.
Human-reported unsupported output	Immediate	Incident triage and fixture creation.
Calibration and confidence bins	Monthly and after material route change	Adjust prompt/display or block route.
Cost per finalized decision	Weekly	Routing, prompt, and packaging review.
Provenance field	Required value	
Provider / model	Exact route used.	
Prompt version	Immutable identifier.	
Evaluation version	Suite used for approval.	
Fallback state	Provider, alternate, or heuristic; reason recorded.	
Usage / latency	Tokens where available, duration, retries.	
Human outcome	Accepted recommendation version and explicit selected option.	

Scaled Results





Measure	Deliver result	Scaled result	Status
Owner accuracy	92.3%	93.4%	Pass
Due-date accuracy	91.2%	92.5%	Pass
Hallucination control	97.6%	98.1%	Pass
Confidence ECE	0.08	0.065	Pass
p95 generation latency	27s	24s	Pass
Fallback rate	0.9%	0.6%	Pass
Variable cost/finalized decision	\$0.29	\$0.23	Pass

Incident and Change Governance

Event	Response
Quality gate regression	Block route/prompt/model deployment and retain qualified version.
Provider outage	Pause, retry, or use approved alternate with visible provenance.
Unsupported high-impact output	Contain, contact affected user when appropriate, investigate, add fixture, and review disclosure.
Calibration drift	Review confidence display and route; do not present misleading certainty.
Cost spike	Check prompt/input size, retry behavior, provider pricing, and regeneration misuse; quality gates remain fixed.
Sensitive-data incident	Activate security/privacy response; model operations does not treat it as a normal quality defect.
Change class	Approval
Prompt wording without contract change	AI owner + automated regression + sampled review.
Model or provider route	AI owner, product, engineering; fixed evaluation, cost/latency, canary, rollback.
Confidence or fallback UX	Product, AI, UX, and risk review.
New language or content format	Separate evaluation and launch gate.
Autonomy increase	Not within current strategy; requires new product/safety qualification.

AI Scale Decision

Decision - AI OPERATIONS ARE QUALIFIED FOR CONTINUED CONTROLLED SCALE within the current English-language, human-finalized workflow. Broader language, richer-ingestion, or autonomy claims require separate evidence.

- Keep model routing and prompt changes evidence-gated.
- Maintain visible provenance and fallback status.
- Use rare-failure discovery to expand the evaluation suite continuously.
- Preserve the AI/evaluation assets as strategic option value regardless of the standalone product's future.

Evidence discipline: replace all simulated controlled-scale values with observed, reproducible results before presenting them as actual product performance.