DECISIONFLOW

Pilot Validation & Measurement Plan

Real-user validation without fabricated results

Document code	DEV-04
Version	2.0
Status	Final Public Portfolio Edition — Implementation Baseline
Baseline date	14 July 2026
Author	Min Thu Kyaw
Role	AI Product Manager

Evidence transparency: Actual implementation evidence is separated from targets, proposed release actions, and the illustrative DEV-06 continuity bridge. Later commercial outcomes are not presented as actual results.

Document Control

Purpose	Document the Develop phase for Pilot Validation & Measurement Plan.
Primary audience	Hiring managers, product leaders, engineering stakeholders, and pilot reviewers.
Evidence classification	Verified implementation fact / implementation-derived assessment / target / proposed next action.
Source limitation	Implementation status is grounded in the authored implementation baseline and public repository references supplied for the case study.
Change control	Publication rule: retain evidence labels unless content is replaced by documented real-world observations.

Evidence Links

- Low-Fidelity Prototype
- High-Fidelity Prototype
- MVP Frontend Repository
- MVP Backend Repository

Contents

Pilot Purpose

The pilot tests whether an individual decision owner can use DecisionFlow to convert messy source material into a trusted, durable decision record with less cognitive and administrative effort. It is not a broad product launch and it does not validate team collaboration, enterprise administration or cross-meeting organizational memory.

No fabricated evidence — This document defines the study. It does not claim participant counts, task completion, AI accuracy, satisfaction, retention or business results that have not yet been measured.

Hypotheses and Questions

Hypothesis	Key question
H1: Structured extraction reduces setup effort.	Can users reach a reviewable structure without manually recreating every option and criterion?
H2: Human review creates sufficient trust.	Do users understand, edit and approve AI output rather than accepting blindly?
H3: Recommendation adds decision value.	Does the reasoning help users clarify trade-offs, not merely restate the source?
H4: Durable records matter.	Do users return to finalized decisions and find the rationale useful later?
H5: The workflow is worth repeating.	Will users use the product for another real decision within seven days?

Pilot Cohort

Dimension	Recommended pilot
Size	8–12 active participants; enough for workflow patterns, not statistical market proof.
Profile	Product managers, engineering leads, project owners, founders or senior knowledge workers who make documented trade-off decisions.
Eligibility	Has a real but non-sensitive decision source; accepts third-party AI processing disclosure; can complete one follow-up interview.
Exclusions	Regulated/highly confidential decisions, unsupported languages, team-admin expectations or required PDF/DOCX ingestion.
Duration	Two weeks of active use plus final interviews.

Onboarding and Consent

1. Explain supported inputs: pasted text and UTF-8 TXT only.

2. Explain that AI output is advisory and requires review.

3. Disclose model-provider processing and raw-source retention.

4. Ask users not to submit credentials, regulated data or highly sensitive material.

5. Confirm support channel and issue-reporting method.

6. Record consent to analytics that excludes source content.

Core Scenarios

Scenario	Task	Success evidence
S1 Create	Start from pasted decision notes.	Decision record created once; no inherited demo data.
S2 Extract	Run AI extraction.	Structured output appears or recoverable error is shown.
S3 Review	Correct at least one field and adjust criteria.	Edits persist after reload.
S4 Recommend	Generate and inspect recommendation.	User can identify rationale, risks and provenance.
S5 Compare	Review option scores and trade-offs.	User understands where scores came from.
S6 Finalize	Choose an outcome, including disagreement with AI.	Selected option and accepted recommendation version are durable.
S7 Retrieve	Open the final record later or in a new session.	Complete record loads independently of local draft.
S8 Resume	Resume an unfinished decision.	Correct decision state is hydrated by ID.

Instrumentation

Event	Safe properties
account_registered	Acquisition source, timestamp.
decision_started	New/duplicate, device class.
source_submitted	Paste/TXT, size bucket; never source text.
extraction_succeeded / failed / fallback	Latency, provider/model, error category.
review_completed	Edited-field count and field types; never values.
recommendation_generated / regenerated	Version, latency, tokens, fallback.
recommendation_accepted	Recommendation version.
comparison_viewed	Method version.
decision_finalized	Selected-vs-recommended match flag; never option text.
decision_resumed	Stage, time since last activity.
decision_archived	Time since finalization.

Success Thresholds

Metric	Pilot threshold	Decision use
Valid-input workflow completion	≥ 90% without manual admin recovery.	Release reliability.
Start-to-finalize conversion	≥ 60% of pilot decisions.	Core value/funnel.
Median active completion time	Baseline; target ≤ 20 minutes for typical source.	Usability.

Metric	Pilot threshold	Decision use
Extraction fallback rate	< 5% under configured provider.	AI reliability.
Critical extraction unsupported-claim rate	≤ 5% benchmark/pilot review.	AI safety.
Reviewed-field edit rate	Measure by field; no fixed “good” value initially.	Model improvement.
Recommendation regeneration	Investigate if > 35%.	Quality or trust signal.
Resume success	≥ 95% for saved in-progress records.	Data integrity.
Seven-day second-decision rate	≥ 35% among eligible participants.	Repeat value.
Trust rating	≥ 4.0/5 with qualitative explanation.	Adoption.

Qualitative Research

- What work did you do before DecisionFlow to structure this decision?
- Which extracted fields were wrong, missing or unnecessary?
- Did confidence and evidence change how carefully you reviewed the result?
- Was the recommendation genuinely useful or merely plausible-sounding?
- Did the comparison reflect your actual priorities?
- Would you keep the final record and return to it? Why?
- What would stop you from using this for another decision?

Weekly Learning Cadence

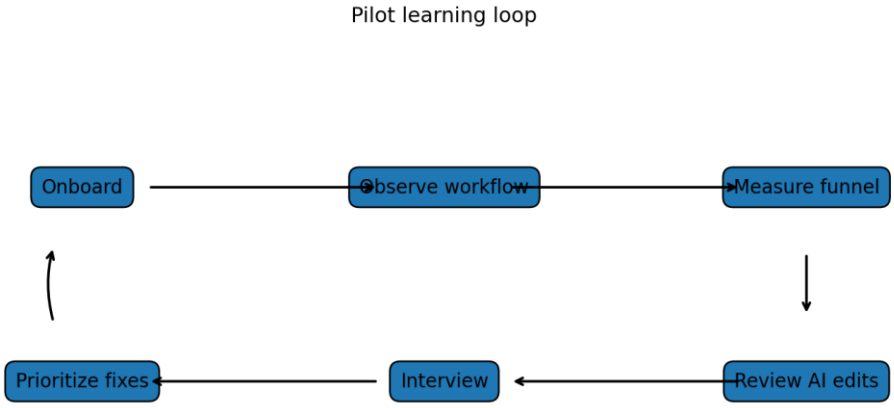

Figure 1. Pilot evidence and prioritization loop.

Cadence	Activity
Daily	Monitor failures, security events, latency, fallback and critical defects.
Twice weekly	Review funnel drop-off, edited fields and user support notes.

Cadence	Activity
Weekly	Product/engineering/AI evidence review; rank fixes by user impact and risk.
End of pilot	Synthesize metrics, interviews and defect severity; write Scale/Iterate/Stop decision.

Issue Triage

Severity	Definition	Response
Critical	Cross-user access, data loss, wrong final decision record or secret exposure.	Pause pilot and contain immediately.
High	Workflow cannot complete, edits disappear, AI fallback is hidden or final record is incomplete.	Fix before affected participant continues.
Medium	Recoverable error, confusing scoring or significant usability friction.	Prioritize in weekly review.
Low	Cosmetic issue or minor copy inconsistency.	Backlog unless it blocks understanding.

Scale / Iterate / Stop

Decision	Condition
Scale controlled pilot	Release gates pass, no critical risks, repeated usage emerges, and users describe meaningful decision value.
Iterate	Core value is promising but workflow, AI quality or trust misses thresholds with fixable causes.
Stop / reposition	Users do not repeat the workflow, recommendation adds no useful value, or the problem is materially weaker than expected.

Pilot Outputs

- Pilot scorecard and funnel dashboard.
- AI evaluation report with actual, not target, results.
- Usability finding log with severity and evidence.
- Updated P0/P1 backlog.
- Development decision updates.
- Scale / Iterate / Stop recommendation with rationale.

Evidence and References

Source	Use in this document
Implementation-derived baseline	DECISIONFLOW_MVP_DEVELOPMENT_PHASE.md, baseline date 14 July 2026.
Low-fidelity prototype	https://minthukyawdev-droid.github.io/decision-flow/
High-fidelity prototype	https://minthukyawdev-droid.github.io/decisionflow-hifi-prototype/

Source	Use in this document
Frontend repository reference	https://github.com/minthukyawdev-droid/decisionflow-webapp
Backend repository reference	https://github.com/minthukyawdev-droid/decisionflow-api
Plan-phase package	DecisionFlow Plan Phase Document Set, version 1.0.

Interpretation rule — Targets, proposed metrics, future releases, and pilot thresholds are not reported as achieved results. Actual results must be added only after executing the relevant test or study.