DECISIONFLOW

AI Strategy & Evaluation Plan

AI responsibilities, output contracts, evaluation methodology and release gates

Phase 2 - Plan

Product	DecisionFlow
Document code	DF-P2-02
Version	2.0
Status	Final Public Portfolio Edition — Historical Planning Baseline
Author	Min Thu Kyaw
Date	July 2026

Evidence statement

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

Purpose	Define how AI creates customer value, which tasks remain deterministic, how outputs are structured and how quality is evaluated before release.
Primary audience	AI/ML engineering, product, QA, data, security and technical leadership.
Related documents	DF-P2-01, DF-P2-04, DF-P2-06 and DF-P2-07.
Review cadence	Update when scope, evidence, architecture, metrics or major decisions change.
Approval state	Historical phase-gate outcome: Conditional Go to Develop.

Revision History

Version	2.0	Owner	Change
1.0	July 2026	Min Thu Kyaw	Restructured Phase 2 planning package; consolidated and standardized source artifacts.

Contents

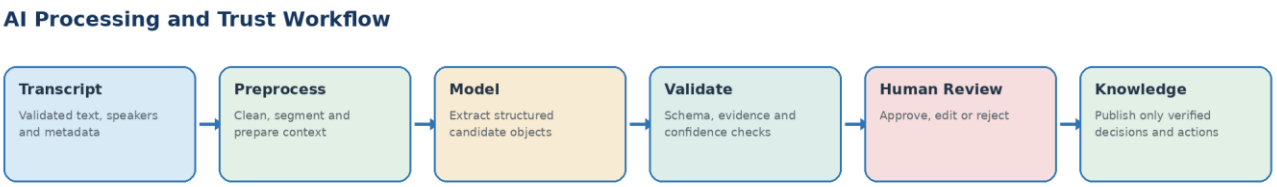


Figure 1. Proposed AI processing and verification lifecycle.

AI Strategy

Purpose

This document defines how Artificial Intelligence will be used within DecisionFlow to create customer value while ensuring reliability, transparency, and responsible deployment.

Unlike traditional software, AI systems are probabilistic rather than deterministic. Therefore, the AI strategy focuses not only on building intelligent features but also on designing robust evaluation, human oversight, and continuous improvement processes.

AI Vision

> **Key statement**
> Use AI to transform unstructured meeting conversations into trusted, structured decisions that improve organizational execution.

AI is not an add-on feature within DecisionFlow—it is the product's core capability.

The objective is to reduce manual work while keeping humans responsible for final business decisions.

AI Opportunity

Research conducted during the Conceive phase identified that users rarely struggle with obtaining meeting summaries. Their primary challenges occur after meetings when they need to determine:

- What decisions were made?
- Who is responsible?
- What actions should happen next?
- What deadlines were agreed upon?
- What risks remain unresolved?

These problems require contextual understanding rather than simple keyword matching.

Generative AI provides the ability to understand conversations, infer intent, resolve references, and produce structured outputs from natural language.

Why Generative AI?

Traditional rule-based systems depend on predefined templates and keywords.

Example:

"If the transcript contains 'decided', mark as decision."

This approach fails because meetings rarely follow consistent language.

Examples:

“We’ll move forward with Option B.”

“I think we should postpone the release.”

“Let’s ship after QA signs off.”

None of these necessarily contain explicit decision keywords.

Large Language Models understand context rather than isolated words, enabling them to identify implicit decisions and commitments.

AI Use Cases

DecisionFlow uses AI for six primary capabilities.

Decision Extraction

Identify explicit and implicit business decisions from meeting transcripts.

Example Output

- Decision
- Supporting evidence
- Confidence score

Action Item Extraction

Identify tasks that require follow-up.

Extract:

- Task
- Owner
- Due date
- Priority
- Status

Owner Identification

Infer responsibility even when ownership is implied.

Example:

“I’ll prepare the proposal.”

↓

Owner = Speaker A

Risk & Blocker Detection

Identify unresolved concerns.

Examples

- Dependency risks
- Technical blockers
- Resource constraints
- Approval bottlenecks

Meeting Summary Generation

Generate concise executive summaries highlighting:

- Objectives
- Key discussions
- Decisions
- Outstanding questions

Knowledge Structuring

Convert conversations into structured entities:

Meeting

↓

Decision

↓

Action Item

↓

Owner

↓

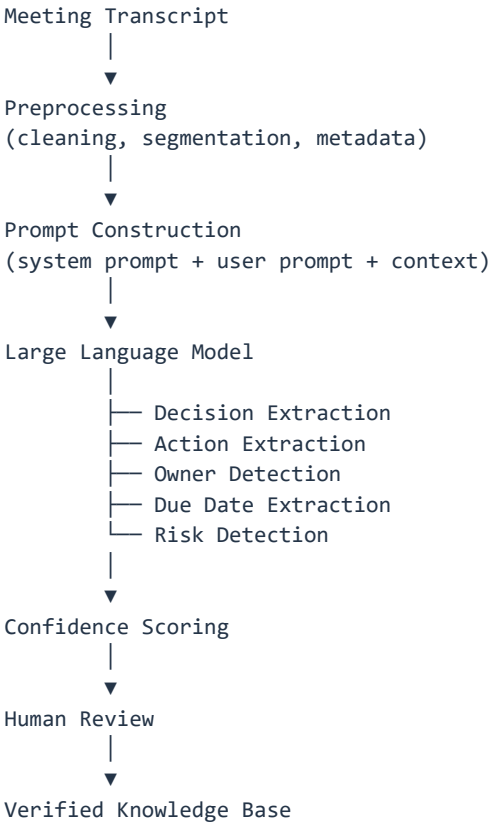
Deadline

↓

Project

This structured representation enables search, reporting, and analytics.

AI Architecture



Model Strategy

DecisionFlow follows a model-agnostic architecture to avoid vendor lock-in.

Initial Model

- High-capability hosted language model for quality-focused MVP evaluation.

Future Options

- Hybrid model routing based on task complexity.
- Fine-tuned open-weight models for enterprise deployments.
- On-premises deployment for security-sensitive customers.

Model selection criteria include:

- Reasoning quality
- Context window
- Latency
- Cost
- Reliability
- API availability

Prompt Strategy

Prompt engineering will be modular.

Components include:

System Prompt

Defines the AI's role and output format.

Context Prompt

Provides project, participants, and meeting metadata.

Task Prompt

Specifies extraction objectives.

Output Schema

Requires structured JSON output for downstream processing.

Prompt templates will be version-controlled and evaluated regularly.

Structured Output Strategy

Every AI response should follow a predefined schema.

Example:

```
{
  "decision": "",
  "owner": "",
  "due_date": "",
  "confidence": 0.94,
  "evidence": ""
}
```

Structured outputs reduce parsing errors and improve system reliability.

Human-in-the-Loop Workflow

AI-generated outputs are never automatically published.

Every decision passes through the following stages:

AI Generation

↓

User Review

↓

Edit (if required)

↓

Approval

↓

Knowledge Base

This workflow increases user trust while generating valuable feedback for model improvement.

Evaluation Strategy

AI performance will be evaluated using both automated metrics and human assessment.

Automated Metrics

- Decision Precision
- Decision Recall
- Action Item Precision
- Action Item Recall
- Hallucination Rate
- Owner Identification Accuracy

Human Evaluation

Users will review outputs based on:

- Correctness
- Completeness
- Clarity
- Relevance
- Overall usefulness

Continuous Learning

DecisionFlow improves over time through user feedback.

Signals include:

- Approved decisions
- Edited outputs
- Rejected suggestions
- Confidence scores
- User comments

These signals inform prompt refinement, evaluation datasets, and future model updates.

Responsible AI Principles

DecisionFlow follows these principles:

Human Oversight

Users retain final authority over decisions.

Transparency

AI-generated content is clearly identified and accompanied by confidence scores.

Privacy

Meeting transcripts are processed securely and are not used for model training without explicit customer consent.

Fairness

Evaluation datasets will include diverse meeting styles, industries, and communication patterns to reduce bias.

Reliability

Outputs are validated through automated evaluation and human review before becoming organizational records.

AI Risks & Mitigations

Risk	Mitigation
Hallucinated decisions	Human approval workflow and confidence scoring
Incorrect owner assignment	Editable outputs and evidence highlighting
Missing action items	Recall evaluation and prompt iteration
High inference cost	Prompt optimization and efficient model routing
Model availability issues	Model-agnostic architecture and fallback providers

Success Criteria

The AI strategy will be considered successful when:

- Decision extraction precision reaches ≥ 90%.
- Action item recall reaches ≥ 85%.
- Hallucination rate remains ≤ 5%.
- AI acceptance rate exceeds 75%.
- Average processing time stays below 30 seconds.
- User confidence score averages ≥ 4.5/5.

Summary

DecisionFlow's AI strategy positions generative AI as the core engine for transforming conversations into structured organizational knowledge. By combining high-quality language models, structured prompting, human review, rigorous evaluation, and continuous learning, the platform delivers trustworthy decision intelligence rather than generic meeting summaries. This strategy ensures that AI enhances human decision-making while maintaining transparency, accountability, and long-term adaptability.

AI Requirements and Evaluation

Purpose

This document defines the functional requirements, quality standards, evaluation methodology, and release criteria for the AI capabilities within DecisionFlow.

Unlike conventional software, AI systems produce probabilistic outputs rather than deterministic results. Therefore, successful delivery requires measurable evaluation criteria, continuous monitoring, and human oversight.

AI Objectives

DecisionFlow's AI system must:

- Identify business decisions from meeting transcripts.
- Extract actionable follow-up tasks.
- Identify task owners.
- Detect due dates where available.
- Highlight risks and blockers.
- Produce concise meeting summaries.
- Generate structured outputs suitable for downstream workflows.

AI Capabilities

Capability	Description	MVP
Decision Extraction	Identify explicit and implicit decisions	☑
Action Item Extraction	Extract follow-up tasks	☑
Owner Identification	Detect responsible individuals	☑
Due Date Extraction	Detect explicit or inferred deadlines	☑
Confidence Scoring	Estimate reliability of outputs	☑
Meeting Summarization	Generate executive summaries	☑
Risk Detection	Identify blockers and unresolved issues	Future
Cross-Meeting Insights	Link related decisions across meetings	Future

Functional AI Requirements

AIR-001 — Decision Extraction

The AI shall identify all significant business decisions discussed during a meeting.

Each extracted decision shall include:

- Decision statement
- Supporting evidence
- Confidence score
- Transcript reference

AIR-002 — Action Item Extraction

The AI shall identify actionable follow-up tasks.

Each action item shall include:

- Task description
- Suggested owner
- Due date (if identified)
- Confidence score

AIR-003 — Owner Identification

The AI shall identify the responsible individual whenever ownership can reasonably be inferred.

If confidence is below the defined threshold, ownership shall be marked as requiring user confirmation.

AIR-004 — Due Date Identification

The AI shall identify:

- Explicit dates
- Relative deadlines
- Sprint references
- Release milestones

Examples:

- “Next Friday”
- “Before launch”
- “By Sprint 8”

AIR-005 — Confidence Scoring

Every AI-generated object shall include a confidence value between 0.00 and 1.00.

Example:

```
{   "decision": "Delay release until QA approval",   "confidence": 0.94 }
```

Output Schema

Every extracted decision shall follow a standardized structure.

```
{   "decision_id": "",   "decision": "",   "evidence": "",   "owner": "",   "due_date": "",
"confidence": 0.00,   "status": "Pending Review" }
```

Standardized outputs simplify downstream validation, storage, analytics, and integrations.

Human Review Workflow

AI-generated outputs follow this lifecycle:

Users always retain final approval authority.

Evaluation Dataset

Evaluation requires a representative dataset that includes:

- Product planning meetings
- Sprint reviews

- Engineering stand-ups
- Customer discovery interviews
- Executive meetings

Each transcript will be manually annotated to create a trusted ground truth for benchmarking AI performance.

Evaluation Metrics

Decision Extraction

Metric	Target
Precision	≥ 90%
Recall	≥ 85%
F1 Score	≥ 0.87

Action Item Extraction

Owner Identification

Metric	Target
Accuracy	≥ 90%

Due Date Extraction

Hallucination

Metric	Target
Hallucination Rate	≤ 5%

Human Evaluation Rubric

Reviewers score each AI output on a five-point scale.

Dimension	Description
Correctness	Is the output factually accurate?
Completeness	Were all important items captured?
Relevance	Does the output reflect meaningful meeting outcomes?
Clarity	Is the output easy to understand?
Usefulness	Would the user rely on this output?

Average target score:

≥ 4.5 / 5

Confidence Thresholds

Confidence	System Behavior
≥ 0.90	Recommend approval
0.70–0.89	Highlight for review
< 0.70	Require manual verification

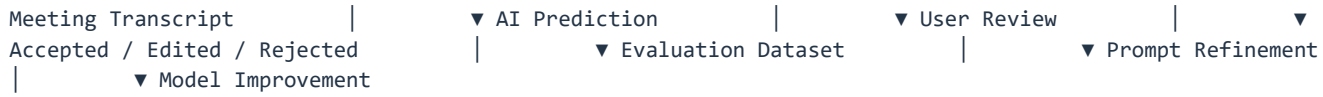
Confidence thresholds will be refined using pilot feedback.

Error Categories

To improve the system systematically, AI errors will be classified into categories.

Error Type	Example
False Positive	Decision extracted when none exists
False Negative	Valid decision not extracted
Wrong Owner	Incorrect person assigned
Missing Due Date	Deadline omitted
Hallucination	Information not supported by the transcript
Formatting Error	Output does not follow schema

Continuous Improvement Loop



User feedback becomes a key source for improving prompt design and evaluation quality over time.

AI Release Criteria

The AI system may progress from Beta to General Availability only if:

- Decision Precision ≥ 90%
- Decision Recall ≥ 85%
- Action Item Precision ≥ 90%
- Hallucination Rate ≤ 5%
- Average User Rating ≥ 4.5/5
- Confidence Calibration validated
- Human review workflow operating successfully

Monitoring After Launch

The following AI metrics will be monitored continuously:

- Decision approval rate
- Manual edit rate
- Hallucination rate
- Average confidence score

- AI processing latency
- User feedback trends
- Failed inference rate

Any significant degradation will trigger investigation and corrective action.

Summary

The AI Requirements & Evaluation Plan defines the standards that DecisionFlow's AI must meet before and after release. By combining structured outputs, measurable quality targets, human review, benchmark datasets, and continuous evaluation, the product ensures that AI remains accurate, trustworthy, and aligned with real user needs. This framework enables responsible AI deployment while supporting continuous improvement as the product evolves.

AI Responsibility Boundary

AI / probabilistic responsibilities	Deterministic platform responsibilities
Interpret conversational context	Authentication and workspace authorization
Generate candidate decisions and actions	File validation, job state and retries
Infer candidate owners and dates	JSON-schema validation and type enforcement
Generate concise summaries and rationale	Persistence, audit events and deletion
Assign confidence signals	Approval status and repository publication rules
Suggest related concepts in future releases	Tenant isolation, retention and access controls

Evaluation Governance

- Maintain separate development, validation and holdout datasets to reduce overfitting.
- Annotate each object with decision/action type, evidence span, owner, date and ambiguity labels.
- Report results by meeting type and transcript quality, not only as a single aggregate score.
- Version prompts, model configuration, schemas and evaluation datasets together.
- Block release when precision, recall, hallucination, safety or regression thresholds are not met.
- Use user edits as feedback signals only after privacy, consent and data-quality controls are defined.

Model selection note
The architecture should specify capability requirements rather than promise a particular vendor or model version. Providers can be compared on quality, latency, cost, privacy, context capacity and operational reliability.