FLO — Client Proposal and Concurrent Roadmap

▯ Table of Contents

Executive Summary

Flo delivers immediate, compounding value by prioritizing critical communications and drafting high-quality replies in seconds. While V1 focuses on Gmail for speed and safety, Flo is an input-centric platform designed to connect multiple sources (e.g., Slack, WhatsApp, email, and more) into a single prioritization and drafting engine. We structure delivery into short sprints with parallel workstreams across DevOps, Backend, Frontend and QA to minimize time-to-value while reducing risk.

Business Outcomes (V1)

Metric	Target	Notes
Time savings	~50% reduction	Average time to respond to priority emails within 2 weeks of use (baseline captured in Sprint 0)
Automated triage coverage	≥30% with ≥85% precision	Incoming emails auto-classified with thumbs up/down feedback loop
Reply speed	≤3s P50 / ≤7s P95	First draft for threads ≤10 messages
Security & reliability	0 P0 incidents	OAuth/OIDC; PII encrypted in transit and at rest

1) Scope Overview (Must / Should / Could)

Must (strict V1)

1. **Google Login (OAuth 2.0)** with minimal scopes (Gmail read/draft/send) + secure token storage.
2. **Initial ingestion** of last *N* emails (e.g., 500) with normalization (headers, body, thread, attachments → metadata).
3. **Priority triage**: rules + lightweight model (scoring) + explanation “why it is priority”.

4. **Assisted draft** for *reply / forward / new*, with **thread context** + configurable tone + safe citation of facts.
5. **Minimal UI**: priority inbox, thread view, side context panel, “Generate draft” button.
6. **Telemetry & feedback**: events (latency, draft acceptance, edits), thumbs-up/down per draft.
7. **Observability**: structured logs, traces, metrics; feature flags.

Should (parallelizable / next)

- Advanced rules (people, SLAs, keywords, VIP clients).
- **RAG** with knowledge base (FAQ, proposals, contracts) and **pgvector** for factual context.
- **Reply templates** by label.
- **Admin / Settings** (team management, tones, limits).

Could (experiments)

- Slack / Teams integration (priority notifications, “reply from Slack”).
- Intelligent daily summary.
- Prompt A/B tests.

No-Goals (for now)

- Deep CRM, revenue analytics, complex marketing automations.

Estimation & Teaming

Aspect	Details
Total effort	320 points ≈ 832 hours across H0–H7
Team composition	1 DevOps, 1 Backend, 1 Frontend, 1 QA (can scale up/down; plan adjusts accordingly)
Availability	~4 hours per role per business day
Calendar duration	Approximately 3 months
Delivery model	Six 2-week sprints

> ***Note:*** *Dates below are placeholders for visualization, not hard commitments.*

Architecture (service-agnostic and extensible)

Core domain (stable): Threads, Messages, Contacts, Prioritization, Drafts, Traceability.
Adapters (pluggable): Gmail (V1), Slack/Teams (later), Webhooks.
Key services:

- **Auth**: OIDC/OAuth, token rotation, secret management.
- **Ingestion**: Incremental pull + deduplication + MIME parse → canonical JSON.
- **Prioritizer**: Rules + model (light GBM/LogReg) + explanation (top-N features).
- **Drafts**: prompt orchestrator, guardrails (do not invent facts), source citation; *human-in-the-loop* mode.

- **Store**: Postgres + **pgvector** (for RAG when enabled).
- **Queues**: Redis/BullMQ for jobs (ingestion, scoring, generation).
- **Observability**: OpenTelemetry, metrics (latency, success rate, token usage). **Security & data**: encryption in transit/at rest; minimized PII; minimal scopes; least privilege; basic **RBAC**.

Logical diagram (text) User → UI (Next.js) → BFF/API (Node/TS) → Services (Auth · Ingestion · Prioritizer · Drafts) → DB (Postgres/pgvector) → Queues (Redis) → Observability.

Concurrent Roadmap by Sprints (America/New_York)

> ***Note:*** *Dates assume a Dec 1, 2025 start and are provided for timeline visualization only.*

Sprint 1 — Dec 1–Dec 14, 2025 (2 weeks)

Role	Activities
DevOps	H0 Infrastructure (CI/CD, minimal IaC, secrets, environments), H7 foundation (structured logging & tracing)
Backend	API/BFF skeleton, H1 Google Auth, data model, H2 bootstrap (list threads)
Frontend	Login screen (H1), Minimal UI prototypes (H5), base navigation
QA	DoR/DoD, test plan, seed data, smoke tests

> ***Deliverable:*** *Functional login, partial ingestion with end-to-end trace, basic logs/metrics dashboard; internal demo.*

Sprint 2 — Dec 15–Dec 28, 2025 (2 weeks)

Role	Activities
DevOps	Extend H7 (metrics, per-request tracing, feature flags), baseline alerts
Backend	H2 complete (dedupe/backoff/indexing), H3 core (rules/score + explanation), H5 endpoints
Frontend	H5 navigable Minimal UI (priority list, thread view, context panel, "Generate draft")
QA	Validate H2/H3/H5; instrument acceptance telemetry

> ***Deliverable:*** *Priority Inbox 1.0 with explanations, navigable UI, complete ingestion; candidate pilot demo (≤5 users).*

Sprint 3 — Dec 29, 2025–Jan 11, 2026 (2 weeks)

Role	Activities
DevOps	Observability dashboards and SLO baselines; expand feature flags
Backend	Stabilize H2/H3, refine explanations and ranking consistency

Frontend	UX refinements for priority views; loading/progress for ingestion
QA	Regression on H2/H3/H5; data quality checks and error budgets

> ***Deliverable:*** *Stable Prioritizer 1.0 with clear explanations; improved UX and ops readiness.*

Sprint 4 — Jan 12–Jan 25, 2026 (2 weeks)

Role	Activities
DevOps	Security posture reviews; secrets rotation routines
Backend	H4 Assisted Drafts core (orchestration + guardrails + citation)
Frontend	H4 UI (tone selection, undo/regenerate) scaffolding
QA	Draft quality evaluation framework; human-in-the-loop validation

> ***Deliverable:*** *First assisted drafts on real threads with guardrails and citations.*

Sprint 5 — Jan 26–Feb 8, 2026 (2 weeks)

Role	Activities
DevOps	Observability maturity (dashboards, alerts), runbooks
Backend	H6 Telemetry (latency, acceptance, edit events); performance tuning
Frontend	H6 UI (thumbs up/down, minimal internal metrics panel)
QA	Performance (P50/P95), canary test flows, rollback drills

> ***Deliverable:*** *Live telemetry/feedback loops; improved performance and visibility.*

Sprint 6 — Feb 9–Feb 22, 2026 (2 weeks)

Role	Activities
DevOps	Hardening, scaling tests, cost monitoring
Backend	Feature hardening, polish on H4/H6; readiness for multi-input adapters
Frontend	UX polish, edge-case handling, accessibility
QA	Full regression, business metrics validation, release readiness

> ***Deliverable:*** *V1 complete and hardened; ready to expand pilot and connect additional input sources.*

Parallel Workstreams (Gantt-style overview)

Team	Sprint 0	Sprint 1	Sprint 2
DevOps	H0/H7-base	H7+	H7 close + hardening
Backend	H1/H2-base	H2/H3	H4/H6
Frontend	H1/H5-mock	H5	H4/H6

QA	plan/smoke	validate H2/H3/H5	performance/regression/DoD

Acceptance & Success Criteria

Category	Criteria
Functional	Priority listing with explanations; assisted drafts with correct thread grounding and tone options; login/logout/session refresh; feature flags per user
Performance	First draft ≤3s P50 / ≤7s P95 on ≤10-message threads
Quality & learning	≥30% auto-triage with ≥85% perceived precision (measured via thumbs up/down), iterative improvement via feedback
Security/Compliance	OAuth/OIDC, encrypted PII at rest/in transit, audited token handling, least-privilege access

---

Risks & Mitigations

Risk	Mitigation Strategy
Gmail quotas/APIs	Backoff, pagination, per-user limits, caching; early load testing
Draft hallucinations	Strict grounding to thread context, mandatory citation, no-send guardrail without human review
Latency spikes	Job queues, token streaming, context caching; progressive rendering
Privacy/PII	Data minimization, encryption, log masking; role-based access controls

---

Collaboration Model (aligned to CN)

Aspect	Description
Ownership	CN owns domain models, business logic and UI decisions. UKLOK provides reusable infrastructure patterns, knowledge systems and orchestration
Cadence	Weekly demos, telemetry reviews, backlog refinement; ad-hoc product syncs as needed
Deliverables per sprint	Running build in staging, demoable flows, short tech note, rollback plan

---

Effort Summary (informational)

Metric	Value
Total story points	320 points (H0–H7)
Total hours	≈ 832 hours across roles

Daily availability	~4h/day per role
Calendar duration	~3 calendar months
Sprint structure	Six 2-week sprints starting Dec 1, 2025

> ***Note:*** *Concurrency in sprints limits calendar time while preserving quality gates. Dates are indicative only.*

Annexes

Future Possibilities after V1

High-impact, low-friction expansions

Feature Category	Description
Omnichannel expansion	Connect Slack, Microsoft Teams, WhatsApp, Telegram, and more—one intelligent inbox for everything
Cross-channel prioritization	A unified view that surfaces what truly needs your attention, regardless of source
Assisted replies everywhere	Draft, refine, and send with your voice and tone across channels; reusable templates for speed
Smart follow-ups	Automatic nudges and reminders to keep conversations moving after meetings or time-sensitive threads
Summaries that matter	Daily digests and topic snapshots for fast catch-up on busy channels
Knowledge-grounded answers	Replies reinforced with your company context and cited sources to build trust
Rules and automations	Personal and team workflows to triage, route, label, and assign—no-code where possible
Controls and compliance	Fine-grained permissions, audit trails, and privacy-by-design for enterprise readiness
Insights and performance	Live dashboards for time saved, response SLAs, and quality signals from feedback
Everywhere you work	Mobile-friendly experience (PWA), notifications, and flexible APIs to embed Flo into your tools

---