Work Plan Document (Revised)

Project: seda.fm

Document Type: WORK PLAN

Version: 2

Created by: TPM

Created: Sat Sep 06 2025

Executive summary — program view This document is the implementation work plan to deliver seda.fm’s Recommendation Engine MVP (Rooms, Playlists, Artists) with the operational, privacy and EU residency controls required for compliance and canary release. It converts architecture into deliverable epics, features and user stories; defines technical dependencies; identifies required resources and owners; lays out a milestone timeline (24 weeks / ~6 months) with acceptance gates; and enumerates main risks and mitigations.

Key program decisions (canonical)

- Data purge: Option A (backup/PITR retention ≤ 30 days) is the canonical production purge strategy for all projects (EU and non EU). Option B (per user DEK / crypto erase via KMS) is supported only as a documented, enterprise exception requiring Legal/DPO + Finance + Product approval and Platform provisioning.
- EU embeddings: Do not treat hosted provider region options as sufficient for EU residency guarantees. For MVP, the canonical EU path is Platform owned EU self hosted embedding runner + EU Supabase project. Provider EU endpoints may be considered later only after signed contractual guarantees and QA/legal sign off.
- Admin tooling (MVP): A secured admin API surface + CLI (SRE/operator workflow) is acceptable and recommended for MVP. A hardened admin web console is a post MVP deliverable.
- Notification pre permission flow: Server authoritative pre permission shown when (session_count >= 2) OR (first meaningful engagement); OS prompt invoked only after user taps in app “Enable”; in app Push toggle default = OFF even if OS permission is granted.
- Provenance UI placement: Single provenance badge per discovery module header + provenance field in Settings Privacy + on demand provenance in item detail modals. No per card badge by default.

Primary objectives (MVP)

- Provide a first meaningful personalized feed in <60s after genre onboarding.
- Preserve EU residency and DSAR/purge guarantees with auditable proof for Option A.
- Build EmbeddingsService with provider abstraction, spend caps ($1,500/mo), and EU routing to Platform owned runner.
- Surface recommendations in feed and controlled notifications with daily cap.

Target customers (unchanged)

- Fans, Artists/DJs, New Users.

Underserved needs (unchanged)

- Discovery friction, fragmented cross platform activity, friend presence awareness.

Value proposition (unchanged)

- Automatic, personalized, social discovery across linked music profiles with privacy and residency assurances.

Scope In Scope (MVP)

- Profile linking (Spotify + Apple Music session import; other providers later).
- Data ingestion: top artists, top genres, top playlists (summary signals).
- Seda.fm activity signals and social signals for friend presence recs.
- Always on engine: hybrid cadence (real time social triggers + daily batch re rank).
- Delivery channels: in app notifications (capped), Home feed modules.
- EU residency: Platform owned EU Supabase + EU self hosted embedding runner for users.data_residency == 'eu'.
- Privacy: Option A canonical (backup/PITR ≤30d), immediate exclusion on delete, privacy_job purge.
- Admin controls: protected admin APIs + CLI (two approver overrides, spend enforcement).

Out of Scope (Future Phases)

- Paid/premium boosted recs, third party ad recs, cross platform playlist auto sync, provider EU endpoints as canonical until legal/QA sign off.

Resolved decisions (new / consolidated)

- Option A is canonical for production purge strategy.
- EU self hosted embedding runner + EU Supabase is canonical for EU users in MVP.

- Admin UI is not required for MVP; protected APIs + secure CLI are required.
- Embeddings provider: OpenAI text-embedding-3-small (1536d) primary; Cohere adapter available; provider usage subject to spend cap enforcement.
- Embedding spend cap programmatic enforcement: $1,500/mo baseline; 80% projected triggers soft actions; 100% triggers hard stop for non essential jobs.
- Real time transport: Socket.IO + Redis adapter (coalescing) remains planned for presence; scale plan retained but will be validated in Event Replayer.

Technical Architecture (draft, refined) Core components (unchanged with clarifications)

- Data Ingestion APIs and Connectors: Spotify Web API (persist refresh_token_encrypted via KMS), Apple Music session import (transient token pattern), and Mock Provider for sandbox.
- Data Normalization Layer: canonical models (Artist, Playlist, Genre, Room), summary signal storage (top N).
- EmbeddingsService (provider abstraction): OpenAI adapter primary; EU runner used for EU users by default; batching, circuit-breaker, cost-metering hooks.
- Vector store: Postgres + pgvector (Supabase projects partitioned by region).
- Recommendation Engine: content-based embeddings + collaborative signals + social friend boost; daily batch + real-time triggers.
- Real-time layer: Socket.IO cluster + Redis adapter for presence and toasts.
- Notification Service: event-driven, relevance thresholds + daily cap enforcement.
- Admin & Enforcement Engine: protected admin APIs, CLI, enforcement automation for embedding spend, override workflows, audit logs.
- Privacy Jobs worker: immediate exclusion + background physical purge + audit bundle generation.

Data flow (unchanged high-level)

1. User links profile via OAuth ingest summary signals.
2. Normalize and embed embeddings stored regionally (EU vs non EU).
3. Recommendation engine runs hybrid jobs recommendations written to recommendations table.
4. Notification service + feed modules surface recs; admin APIs control enforcement.

Data schema draft (unchanged but with provenance)

- users (data_residency, session_count, meaningful_engagement flags)
- linked_accounts
- artists (embeddings, external IDs, genres)

- playlists (embeddings, provider, external_id)
- rooms (embeddings, created_by, genre)
- user_activity (summary-level events)
- recommendations (user_id, type, entity_id, score, expires_at, provenance)
- embeddings_meta / embedding_audit (provider, model, region, job_id, cost, deleted_at)
- privacy_jobs
- friends (social graph) Notes: pgvector for embeddings; embedding_meta.region required; store summary signals, not raw per-track logs.

Top-level delivery phases (24 weeks) — refined with Q&A commitments Phases remain 0–6 as originally defined. Adjustments and SLAs noted below.

Phase 0 (Weeks 0–2) — Infra & Scaffolding (revised) Deliverables (must have for MVP timeline to hold)

- Platform provisions sandbox-us and sandbox-eu Supabase projects (pgvector enabled) with backup/PITR retention ≤30 days. SLA: sandbox eu available within 7 business days; hardened sandbox within 14 days.
- KMS (EU + US) provisioned; envelope encryption pattern implemented; test encrypt/decrypt validated.
- EU self hosted embedding runner (SentenceTransformers container) deployed in sandbox eu as canonical EU compute path: initial instance available within 7 business days; autoscale baseline and admin test hooks by Day 14.
- LaunchDarkly integration; feature flags created.
- Mock Provider and Mock Embeddings images. Acceptance: sandbox eu + KMS + EU runner accessible; retention policies verified.

Phase 1 (Weeks 2–6) — Core APIs & Onboarding Deliver:

- users.data_residency, residency_changes audit; RLS policies; pre write validators preventing cross region writes.
- Onboarding genre selection (seed first feed <60s).
- Feed GET /feed with caching, MV templates for genre top rooms.
- Acceptance: feed P95 <250ms; first feed <60s.

Phase 2 (Weeks 6–10) — Provider Links & Embeddings Pipeline Deliver:

- Spotify OAuth (persist refresh_token_encrypted via KMS), Apple session import (transient, 202 async).
- EmbeddingsService adapter (OpenAI primary), batching, cost hooks and shadow-write capability for auditing.

- User taste embedding compute, per-user embeddings persisted with embedding_meta.region mapping to users.data_residency.
- PGVector indexes (HNSW) created and tuned. Acceptance: links working; EU users routed to EU runner; embedding_meta.region == users.data_residency.

Phase 3 (Weeks 10–14) — Candidate generation, Re-ranker & Real-time Deliver:

- ANN candidate generation and linear re-ranker (configurable weights).
- Recommendations table with TTL and caching.
- Socket.IO presence with Redis adapter; presence coalescing and ephemeral toasts with server-side caps. Acceptance: re-rank NDCG baseline met; presence P95 latency within planned bounds.

Phase 4 (Weeks 14–18) — Privacy, Admin APIs & Enforcement Deliver:

- Export / Delete flows with privacy_job worker, embedding_audit.deleted_at, immediate exclusion semantics.
- Admin protected APIs and secure CLI (embedctl) for override lifecycle (request/approve/execute), spend simulation hooks, and enforcement automation for 80%/100% spend triggers.
- Backup/PITR retention verification artifacts for Option A. Acceptance: delete yields immediate exclusion; privacy_job created; sandbox accelerated purge verified; admin APIs + CLI tested with two approver enforcement.

Phase 5 (Weeks 18–22) — QA, Load, EU Validation Deliver:

- Sandbox load tests and Event Replayer scenarios for real-time spikes.
- EU runner reliability tests, failover exercises, and residency enforcement verification (no cross-region writes).
- Embedding spend simulation and enforcement automation tests. Acceptance: ANN P95 <120ms target after tuning; real-time P99 <500ms medium profile; residency gating tests pass.

Phase 6 (Weeks 22–24) — Hardening & Canary Rollout Deliver:

- Monitoring, alerts, SLO dashboards, runbooks, production provisioning, canary flags in LaunchDarkly.
- Canary plan and roll/fallback criteria. Acceptance: operational & product gates satisfied; canary passes ramp to full.

Resource plan (updated commitments) Core startup team (required commitments)

- Backend engineers: 4 FTE (NestJS/workers/integrations) — must be dedicated Weeks 2–10 for critical path.
- Frontend engineers: 2 FTE (React/mobile) — onboarding & feed UX.
- Data Scientist/ML Eng: 1 (re-ranker) — part-time for Phase 2/5 tuning.
- Platform/Infra: 2 FTE (provision sandbox eu, KMS, EU runner) — must be dedicated Weeks 0–6.
- SRE/Operations: 1 FTE — owns EU runner ops & admin executor role; on call.
- QA/Test: 2 FTE — automation & load testing across phases.
- UX Designer: 1 (privacy, onboarding, provenance).
- Product Manager: 1 — PO/PM oversight. Estimate: core dev capacity = 4 backend, 2 frontend x 6 months. If Platform or Backend availability is reduced, rebaseline immediately — see Staffing & rebaseline section.

Stakeholders and RACI (updated)

- Product (PO) — R: product acceptance, experimentation targets.
- Engineering (Backend Lead, Frontend Lead) — A: implement APIs, workers, EmbeddingsService, privacy APIs.
- Data Science — C/R: ranking weights, offline CF, model validation.
- Platform/Infra — A: provision Supabase projects (EU/US), KMS, EU runner, quotas, sandbox.
- SRE/Operations — A/R: runbooks, on call, operate EU runner, execute overrides.
- UX — C/R: onboarding, provenance UX.
- QA/Test Lead — R: test plans, sandbox execution, DSAR tests.
- Legal / DPO / Finance — C/A: privacy & enterprise exception approvals.
- PM — A: program management, escalations.

Epics Features User Stories (for estimation and planning) Epics A–I preserved from original plan. Key clarifications and mandatory inclusions surfaced by Q&A are integrated into feature acceptance criteria below (only highlights of updated/added stories are shown — full story templates remain in the backlog).

Epic A: Infra & Sandbox (Phase 0)

- Feature A1: Supabase projects + pgvector provisioning
    - Story A1.1: Provision sandbox-us and sandbox-eu Supabase projects with pgvector enabled and backup/PITR retention ≤30 days. (AC: retention policy verifiable via provider API / IaC; size: M)

  - Story A1.2: Provide admin/test API to return backup snapshot metadata (snapshot IDs, expiry). (AC: test returns snapshot metadata for sample backups; size: S)
- Feature A4: Sandbox tooling
  - Story A4.1: Mock Provider & Mock EmbeddingsService with deterministic embeddings; admin knobs to simulate EU runner outage/latency. (AC: mock endpoints available within 3 business days; size: M)

Epic B: Core API, Auth, Residency & Onboarding (Phase 1)

- Feature B1: Auth + Users + Residency
  - Story B1.1: Implement users.data_residency, residency_changes audit, and pre write validators rejecting cross region writes for EU users. (AC: test attempts to write EU user vector to non EU DB rejected; size: L)
  - Story B1.2: RLS policies for EU project; server enforcement for residency. (AC: automated tests pass; size: M)

Epic C: Provider Integrations & Embeddings (Phase 2)

- Feature C1: Spotify OAuth + linked_accounts
  - Story C1.1: Persist refresh_token_encrypted using KMS envelope encryption; implement background sync toggle (default OFF). (AC: link/unlink + decrypt test; size: M)
- Feature C3: EmbeddingsService adapter
  - Story C3.1: EmbeddingsService abstraction with provider adapters (OpenAI primary) and region selector routing per users.data_residency. (AC: EU users evaluated against EU runner; embedding_meta.region recorded; size: L)
  - Story C3.2: Spend metering + enforcement hooks (80% soft actions + 100% hard stop for non essential jobs). (AC: enforcement sim triggers actions in sandbox; size: L)

Epic D: Vectorized Candidate Gen & Re-ranker (Phases 2/3)

- Story D1.1: ANN query path (user taste nearest entities) with region-specific vector stores. (AC: recall tests; size: M)
- Story D2.1: Linear re-ranker with friend boost and diversity penalty; config weights. (AC: AB testable; size: L)

Epic E: Real time presence & toasts (Phase 3)

- Story E1.1: Socket.IO auth + Redis adapter + presence coalescing. (AC: presence aggregation P95 < 500ms; size: L)
- Story E2.1: In app toasts rules server enforced (3/session, 10min cooldown) with server counts. (AC: counts enforced server-side; size: M)

Epic F: Privacy, Export/Delete & Backup (Phase 4)

- Story F1.1: Export data job; generate signed URL for DSAR ZIP (privacy_job audit). (AC: sample DSAR bundle produced; size: M)
- Story F2.1: Delete imports workflow + immediate exclusion + enqueue privacy_job for physical purge. (AC: deletion results in immediate exclusion; privacy_job created; size: L)
- Story F3.1: Backup/PITR retention ≤30 days configured and verified for prod & EU projects. (AC: provider API proof; size: S)

Epic G: Admin tooling & Spend enforcement (Phase 4/5)

- Story G1.1: Embedding spend telemetry + projected run rate alerts at 80/100 and enforcement automation for 80%/100%. (AC: simulate-spend triggers enforcement; size: L)
- Story G2.1: Override request/approve/execute API; two distinct approvers required; append-only audit logs. (AC: cannot execute without two approvals; size: L)
- Story G2.2: Secure CLI (embedctl) wrapping admin APIs for SRE operations; requires SSO+MFA. (AC: CLI calls produce audit records; size: M)

Epic H: QA, Performance & EU Validation (Phase 5)

- Story H2.1: EU-runner outage simulation tests: verify fallback rules, audit override_id annotations, and purge reconciliation. (AC: fallback requires two-approver override; all fallback uses logged; size: M)
- Story H3.1: DSAR verification tests producing audit bundle. (AC: DSAR ZIP contains privacy_job, embedding_audit entries, backup metadata; size: M)

Epic I: Observability, SLOs and Runbooks (Phase 5/6)

- Story I1.1: Metrics/tracing for embedding latency, ANN query P95, queue depths; dashboards. (AC: dashboards + alerts; size: M)
- Story I2.1: Runbooks for embedding provider outage, EU runner outage, KMS compromise, index rebuild; conduct tabletop. (AC: tabletop executed; size: S)

Acceptance & Go/No Go gates (updated operational hard gates) Operational hard gates (required for canary)

- Rec API: P95 < 250ms (fail if >350ms).

- ANN-only queries: P95 < 120ms target; degrade acceptable up to 200ms.
- Real-time notifications: P99 < 500ms (MVP acceptable); fail if >1s sustained.
- Error rates: rec endpoints 5xx <0.1%.
- Privacy: RLS & deletion flow tests pass; no cross-region writes for EU users (automated tests).
- EU runner provisioning SLA: sandbox eu initial instance available within 7 business days; hardened sandbox within 14 days. Business/quality gates
- Recommendation CTR: no >10% relative drop in canary.
- listen_30s: no >5pp drop and >=60% maintained.
- Embedding spend enforcement: 80%/100% enforcement simulated and validated in sandbox.

Timeline (24 weeks / 12 sprints of 2 weeks) — sprint highlights with Q&A-driven SLAs Sprints 0–11 follow the original plan but with tighter infra SLAs and admin/API priorities in Phase 4. See above Phase descriptions for specific sprint contents and acceptance criteria.

Cross team coordination & communication plan (unchanged but actionable)

- Weekly program meeting (Product/Eng/SRE/Platform/DS/QA) — 30–60m.
- Bi weekly demos to stakeholders.
- Slack channels: #rec mvp, #rec infra, #rec qa. Named Platform contact & Backend lead must be present in weekly infra sync during Weeks 0–6.
- Dependency tracker published; Platform must nominate named contacts and SLA for sandbox provisioning within 24 hours of plan approval.

Risk register & mitigations (updated with Q&A clarifications)

1. Embedding spend overrun

- Risk: OpenAI usage exceeds $1,500/mo.
- Mitigation: programmatic spend metering; enforcement automation at 80% (throttle/switch model) and 100% (hard stop for non essential jobs); admin override workflow (two approvers); sandbox simulate-spend; alerting.

2. EU compliance / cross region writes

- Risk: accidental storage or processing of EU user vectors outside EU.
- Mitigation: users.data_residency canonical field; Platform owned EU runner for MVP; pre write validators rejecting non EU writes; per write provenance; automated tests and CI checks; two approver override for emergency non EU fallbacks (time boxed & audited).

3. Real time fanout scale

- Risk: presence floods & latency.
- Mitigation: server aggregation/coalescing, client coalescing, counts-only fallback for large rooms, Event Replayer load tests; broker migration plan if necessary.

4. Provider outages and residency posture

- Risk: OpenAI/Spotify outages or provider-side cross-border logs.
- Mitigation: circuit-breaker, cached fallback, Cohere standby adapter, EU runner canonical for EU users; Legal to negotiate provider guarantees for future adoption.

5. Privacy Delete / backup proofs

- Risk: inability to prove physical purge.
- Mitigation: Option A canonical; Platform to enforce retention ≤30 days; audit bundle generation; sandbox acceleration hooks for QA; Option B available only for enterprise exceptions and validated with KMS proof.

6. Operational complexity & resource shortage

- Risk: Platform/back-end staffing shortages delay critical infra.
- Mitigation: require written commitment for Platform 2 FTE Weeks 0–6 and Backend 4 FTE Weeks 2–10. If unavailable, rebaseline immediately, provide mocks, or engage contractors.

Instrumentation & metrics (minimum)

- Business: rec_impression, rec_click, listen_30s, rec_dismiss, follow_artist.
- Provenance telemetry: embedding_meta.region on rec_impression events.
- System: rec_api_latency_p50/p95/p99, ann_query_latency, embedding_latency per provider/runner, embedding_cost_runrate, queue_depths, socket_fanout_latency.
- Alerts: embedding projected >80% (email+slack), rec_api_p95 >350ms for 5m (page), real-time P99 >500ms for 5m (page).

Testing strategy (QA view) — augmented

- Unit tests: RLS & residency validators.
- Integration tests: provider adapters mocking; KMS encrypt/decrypt tests.
- Contract tests: EmbeddingsService adapter.
- E2E: onboarding   feed   link providers   import   delete   export flows; executed in sandbox eu for residency.

- Load: k6 + Event Replayer validating ANN P95 and real-time P99 targets.
- Privacy compliance: automated DSAR verification script produces audit bundle.
- Sandbox test hooks: simulate-spend, accelerate-purge, kmstest (for Option B sandbox validation).

Deliverables & artifacts (by milestone)

- DB migrations, OpenAPI specs for admin APIs, EmbeddingsService adapter, admin CLI, runbooks, SLO dashboards, DSAR audit artifacts, compliance proof of backup retention.

Budget & procurement notes (unchanged with emphasis)

- Immediate procurement: OpenAI keys, LaunchDarkly, EU runner hosting, KMS budget. Embedding cost cap enforcement remains primary control; Finance to sign off emergency override budget process.

Staffing & availability — mandatory confirmations and contingencies Required confirmations (must receive to keep schedule)

- Platform/Infra: commit 2 FTE (dedicated Weeks 0–6) — responsible for sandbox eu, KMS, EU runner. If not available, rebaseline Phase 0 (extend by 1–2 weeks) or provide mocked EU runner within 3 business days.
- Backend: commit 4 FTE (dedicated Weeks 2–10). Missing FTEs will add ~1 sprint delay per FTE shortfall.
- SRE: 1 FTE on call to operate EU runner and execute overrides. Without SRE, go/no go blocked for canary. If any role is constrained, PM will rebaseline and propose scope de scoping or contracting.

Admin tooling design decision (final)

- MVP: Protected admin APIs + secure CLI (embedctl) that enforce two approver overrides, audit trails, and spend enforcement automation. APIs require SSO + enforced MFA and RBAC. Audit bundles exportable for DSARs.
- Post MVP: Hardened admin web console built on same APIs.

EU embeddings residency posture (final)

- Platform owned EU self hosted runner + EU Supabase project are canonical for users.data_residency == 'eu' from day one.
- Provider EU endpoints may be evaluated and used only if Legal obtains signed contractual guarantees for EU only processing/storage and QA verifies parity and auditability. Until then, route EU users to the Platform runner.
- EU runner capacity SLA for planning:

  - Baseline sustained throughput: 2,000 items/min (≈33 items/sec).
  - Burst capability: up to 10,000 items/min short bursts (60–120s) via autoscale/queueing.
  - Latency SLO (compute-only): P50 ≤ 100ms; P95 ≤ 250ms (target); P95 ≤ 500ms acceptable for MVP under baseline.
  - Provisioning SLA: initial sandbox instance within 7 business days; hardened within 14 days.

Provenance UI placement (final)

- Module header badge: single "Processed in EU" badge per discovery module header when content computed/stored in EU for the current user.
- Settings Privacy: full provenance details and exportable audit bundle (authoritative record).
- Item detail modal: on demand provenance chip (provider/model/region/computed_at).
- Feed cards: no per card badge by default; hero/promoted cards may show at most 1–2 compact chips per viewport if justified.
- Toasts & push: no provenance in normal toasts; emergency non EU fallback receives one concise notification linking to Settings.

Push pre permission gating & toggle behavior (final)

- Show in app pre permission when server show_pre_permission == true (session_count >= 2 OR first meaningful_engagement).
- Only invoke OS-level prompt after user taps in app “Enable”.
- Keep in app Push toggle = OFF by default, even if OS permission granted; user must toggle ON to receive FCM pushes.
- Server must enforce both push_os_granted == true AND push_in_app_enabled == true before sending pushes.
- Backoff rules: dismiss prevents re prompt for 7 days or until new conditions are met.
- Cohort exceptions: teens and EU/high-privacy cohorts require stricter gating (both conditions).

Risks revisited & mitigations (summarized)

- Provider contract risk (EU processing): mitigate by Platform EU runner canonical; negotiate provider contracts in parallel.
- Staffing risk: Platform and Backend availability are hard blockers — require written commitments. Mitigations: mock runners, contractors, or rebaseline.
- Spend risk: enforcement automation + admin override gating (two approvers) + alerting.

- Privacy proof risk: Option A canonical; Platform to provide retention config + QA to validate accelerated purge in sandbox.

Testing & acceptance matrix (high level)

- Residency: automated tests that verify users.data_residency -> embedding_meta.region == 'eu' and no cross-region writes.
- Delete/DSAR: immediate exclusion validated; privacy_job completes; audit bundle produced with backup metadata.
- Spend enforcement: simulate 80%/100% in sandbox; validate throttle/hard stop and logging.
- Real-time: presence P95 target validated under Event Replayer medium profile.
- Recommendation quality: via NDCG tests and listen_30s metrics in canary.

Runbooks & operational playbooks (deliverables)

- Runbook: EU runner outage, embedding provider outage, KMS compromise, privacy_job failures, override execution procedure.
- Operator playbook: embedctl CLI usage, two-approver override process, emergency fallback steps, audit artifact generation.

Open questions & clarifications required (actionable items for stakeholders)

1. Confirm platform/infra resource commitment: please confirm 2 FTE (Platform) dedicated for Weeks 0–6 and named contacts. If not available, state alternate plan (mock runner vs rebaseline) within 24 hours.
2. Confirm backend staffing: please confirm 4 backend FTE Weeks 2–10 and named owners. If less, indicate which backend priorities can be de-scoped.
3. Confirm Option A acceptance (canonical): Legal/Compliance to acknowledge Option A is canonical and sign off on sandbox proof requirements (backup metadata export). If any tenant requires Option B now, provide tenant list.
4. Confirm admin tooling decision: Product/Legal to confirm CLI + APIs acceptable for MVP (if some approvers require UI to approve, we will provide read-only approval links or emailed artifacts for non-SRE approvers).
5. Confirm LaunchDarkly flag names and initial canary cohorts so we can encode gates in the backlog.

Immediate next steps (first 7 calendar days)

- PM to request written commit from Platform and Backend leads for declared FTE availability and named contacts.

- Platform to begin provisioning sandbox eu, KMS keys, and EU runner; deliver initial access within 7 business days or provide mock runner within 3 business days.
- Backend to lock API contracts (OpenAPI) for admin override endpoints and EmbeddingsService routing; deliver draft by end of Sprint 0.
- Legal to prioritize provider contract review for OpenAI/Cohere EU guarantees; status update in 1 week.
- QA to prepare sandbox DSAR/accelerated purge test plan; start validation as infra becomes available.
- DS to prepare re rank baseline experiments and NDCG tests for Sprint 5 acceptance.

Appendix — quick estimates & planning units (updated)

- Infra Phase (A): 4–6 engineer-weeks (Platform + SRE + 1 backend).
- Core API & Onboarding (B): 6–8 backend eng-weeks + 2 FE-weeks.
- Embeddings & Provider (C): 8–12 backend eng-weeks + 2 DS-weeks.
- Vector & ANN (D): 6–8 backend eng-weeks + 1 DS-week.
- Real-time (E): 6–10 backend eng-weeks + 2 FE-weeks + SRE tuning.
- Privacy & Admin (F/G): 8–12 backend eng-weeks + 2 Platform-weeks + QA.
- QA & Load (H): 6–10 QA-weeks + Platform booking.
- Observability/Runbooks (I): 3–4 weeks cross-team. Buffer: 15–25% for integration complexity and legal reviews.

Closing & required approvals Please confirm the requested staffing commitments, Option A acceptance, and admin tooling decision within 48 hours so I can:

- Finalize sprint-by-sprint backlog with ticket-level owners and estimates.
- Produce acceptance-test matrices, CI gating scripts, and canary rollout plan.
- Publish the exact test calendar (EU runner staging windows) and schedule first tabletop run.

If preferred, I can immediately produce:

- A 12 week detailed sprint backlog with JIRA ready tickets.
- A list of required cloud IAM roles & RBAC mappings.
- A go to production checklist (operational & compliance).

Which deliverable would you like first?

Generated by Feature Forge • 9/10/2025, 1:15:41 PM