Here’s a tight, shippable **Final Sprint Outline** aligned with the **Priority Goals (source of truth)** and **RoE v1.1 — Inmigración a España**. It’s sequenced for parallel work, each item has clear acceptance criteria (AC) and a definition of done (DoD).

**0) Release Criteria (gates to ship)**

* **RC-1**: UI at ai.olivogalarza.com renders a working chat, streaming responses from the API.
* **RC-2**: KB-first answers; when confidence is low, server does web fallback; **no fabricated citations**.
* **RC-3**: Responses follow RoE style: Spanish, 3–6 sentences, **bold key terms**, bullets for procedures; **one-time disclaimer** per session; **“Respuesta incierta. Contáctenos.”** on uncertainty.
* **RC-4**: Basic observability + rollback plan.
* **RC-5**: Deterministic health checks pass (curl + DevTools).

**1) Backend (API: kb-legal-assistant)**

**1.1 Route contracts & health**

* Implement/confirm endpoints:
  + GET /api/health → { ok: true, version, runtime }
  + POST /api/chat (SSE) and POST /api/rag/answer (JSON)
  + GET /api/debug/cache?key=… (read-only)
* **AC**: Curl returns 200 for health; SSE streams tokens; JSON returns full answer meta.
* **DoD**: Typed responses, Zod validation, CORS enabled for ai.olivogalarza.com.

**1.2 RAG core (KB-first)**

* Load data/kb/embeddings.json; top-K retrieval; MIN\_SCORE threshold.
* Confidence gate decides: **answer from KB** vs **trigger web fallback**.
* **AC**: For a known KB query (e.g., *artículo 123*), web fallback not used; citations point to KB.
* **DoD**: Unit tests for retriever; log retrieval scores (debug only).

**1.3 Web fallback (server-side)**

* On low confidence, fetch recent trusted sources; attach hidden citations (reveal on request).
* **AC**: For an OOD query, returns web-grounded Spanish answer; hidden citations available.
* **DoD**: Domain allow-list, timeout & error handling; never expose API keys.

**1.4 RoE Finalizer (policy enforcement)**

* Final pass to guarantee: Spanish, 3–6 sentences, **bold key terms**, procedures as bullets; inject **session-scoped disclaimer** once.
* Insert **uncertainty phrase** when confidence < floor.
* **AC**: Snapshot tests for style/length/disclaimer.
* **DoD**: Idempotent finalizer; toggle via feature flag ROE\_FINALIZER\_ENABLED.

**1.5 Redis cache MISS fix (priority)**

* Normalize key schema: rag:chat:{sha256(query)} and rag:kbfiles:v1.
* Force **Node runtime**, align Upstash region; add explicit EX TTL; await writes; immediate get-after-set assert in logs.
* /api/debug/cache to verify values.
* **AC**: Consecutive identical requests show **HIT** in logs and measurable latency drop.
* **DoD**: Structured logs for read/write; redaction for PII.

**1.6 Deterministic test mode**

* ?test=1 returns canned retrieval + canned answer; SSE emits stable token sequence.
* **AC**: DevTools Console/Network checks pass 100% deterministically.
* **DoD**: Guard behind TEST\_MODE\_ENABLED.

**2) Frontend (UI: kb-legal-assistant-ui)**

**2.1 API integration & envs**

* Use NEXT\_PUBLIC\_API\_BASE=https://api.olivogalarza.com.
* SSE client with exponential backoff; surface network errors gracefully.
* **AC**: Round-trip prompt → streamed tokens → rendered transcript.
* **DoD**: Type-safe fetcher; CORS verified.

**2.2 RoE UX elements**

* Show one-time **disclaimer** per session.
* Render answer in Spanish; preserve **bold key terms** and **bulleted procedures**.
* Citations **hidden by default**, toggle to reveal.
* **AC**: Manual test confirms RoE formatting; toggle reveals citations.
* **DoD**: Minimal UI polish; mobile responsive.

**2.3 DevTools test harness**

* Console prints PASS/FAIL for:
  + SSE open, first token < 1s from headers received (in test mode).
  + Presence of disclaimer on first answer only.
  + Cache HIT observed on 2nd identical prompt (from API debug flag).
* **AC**: All tests PASS consistently.
* **DoD**: Docs in /docs/tests/ui.md.

**3) Query Rewriter v2 (minimal, safe)**

* Pipeline: lowercase → accent fold → trim → expand legal synonyms (art., artículo, arts.) → normalize Extranjería/visado variants.
* No risky aggressive stemming; maintain semantic precision.
* **AC**: 30-query benchmark reaches ≥90% rewrite match to expected forms.
* **DoD**: Unit tests; can be toggled off with REWRITE\_ENABLED.

**4) Knowledge Base hygiene**

* Verify paths, embeddings checksum, and **top-K** configuration.
* Add script to recompute embeddings (documented, not needed to ship).
* **AC**: Sanity run shows stable top-K for canonical queries.
* **DoD**: README with steps & gotchas.

**5) Security & Compliance**

* Server-side model keys only; never in the UI.
* CORS: allow only ai.olivogalarza.com.
* Basic rate limiting (per IP) at API; log anomalies.
* **AC**: Scan repo for secrets; envs documented.
* **DoD**: 429 on abuse scenarios in test.

**6) Observability & Ops**

* Structured logs for: request id, mode (test/normal), cache (HIT/MISS), latency p50/p95.
* Uptime check for /api/health.
* **AC**: Dash or log view shows p95 < configured target under light load.
* **DoD**: Runbook with common errors + mitigations.

**7) Deployment Plan**

**7.1 Staging cut**

* Branch cut release/v1. Deploy API → staging URL; UI → staging with NEXT\_PUBLIC\_API\_BASE pointing to staging API.
* **AC**: All health & deterministic tests PASS on staging.

**7.2 Production rollout**

* Deploy API first, then UI (env points to prod API).
* Feature flags on: ROE\_FINALIZER\_ENABLED=on, REWRITE\_ENABLED=on, TEST\_MODE\_ENABLED=off.
* **AC**: Smoke tests (below) PASS on prod.

**7.3 Rollback plan**

* Revert to last stable image; disable new flags; invalidate CDN if needed.
* **AC**: Documented and tested on staging prior to prod push.

**8) Smoke Tests (copy-paste)**

# Health

curl -s https://api.olivogalarza.com/api/health | jq .

# Deterministic JSON (test mode)

curl -s "https://api.olivogalarza.com/api/rag/answer?test=1" -X POST \

-H "Content-Type: application/json" \

-d '{"query":"ping"}' | jq .

# SSE path (inspect headers + first chunk)

curl -N -s -H "Accept: text/event-stream" \

-X POST "https://api.olivogalarza.com/api/chat?test=1" \

-H "Content-Type: application/json" \

-d '{"query":"ping"}'

# Cache probe (after one normal request)

curl -s [https://api.olivogalarza.com/api/debug/cache?key=<known-key>](https://api.olivogalarza.com/api/debug/cache?key=%3cknown-key%3e)

**9) Staffing & Timeline (aggressive but realistic)**

* **Day 0**: Branch cut, health/test mode, RoE finalizer scaffold.
* **Day 1**: Redis MISS fix + debug endpoint; UI SSE + disclaimer.
* **Day 2**: Web fallback + citations toggle; Query Rewriter v2 + tests.
* **Day 3**: DevTools harness; security checks; staging deploy.
* **Day 4**: Staging validation, perf pass, prod deploy + smoke tests.

**10) Definition of Done (per workstream)**

* Code merged to release/v1, feature flags configured, tests green, docs updated, smoke tests pass in prod, rollback verified.