RORI

Research On Regulatory for Industry(s)

Project Proposal

A General-Purpose Regulatory Research & Analysis Engine

Version 0.1 | Draft for Review

February 2026

Authors: Frank & Candi

Table of Contents

1 Executive Summary

RORI is a general-purpose regulatory research and analysis engine designed to ingest, curate, and reason over large corpuses of regulatory, compliance, and policy data across industries and jurisdictions. The platform provides agent-based retrieval with tunable depth and completeness, enabling users and downstream AI systems to accurately retrieve, compare, and synthesize regulatory intelligence without reliance on web scraping.

"Regulatory data is vast, fragmented, constantly evolving, and buried across thousands of sources with wildly different formats and semantics. No single AI model can reliably internalize all of it. RORI solves this by building a curated, auditable knowledge layer that agents can query with precision — and that humans can trust."

C

Curated Ingestion

Ingest and structure heterogeneous regulatory data with semantic indexing and metadata enrichment.

A

Agent-Based Retrieval

Tunable depth and completeness with audit-grade accuracy and citation provenance.

S

Cross-Corpus Synthesis

Compare, gap-analyze, and synthesize across regulatory corpuses into actionable plans.

2 Problem Statement

Organizations across industries face a common challenge: regulatory and compliance information is scattered across federal, state, and municipal sources in disparate formats — PDFs, legal text, guidance documents, seller/servicer guides, directives, and standards. The cost of manual research is high, the risk of missing applicable regulations is real, and the pace of regulatory change makes point-in-time snapshots unreliable.

Why Existing Approaches Fall Short

Web Scraping

Brittle, legally gray, and produces noisy results that degrade AI inference quality.

General LLMs

Hallucinate regulatory specifics, lack citation provenance, and cannot be audited.

Static Repos

Require manual curation and don't support semantic reasoning or cross-corpus comparison.

RORI addresses these gaps by creating a purpose-built regulatory knowledge platform with curated ingestion, structured indexing, agent-based retrieval, and auditability at every layer. The platform is designed from the ground up to deliver results that can withstand audit and regulatory scrutiny.

3 Vision & Product Scope

Three Primary Modalities

A Agent-as-Platform

RORI is an agent-based platform first. The core value is the agent's ability to reason over curated regulatory data and return accurate, citable, auditable answers. The agent supports tunable depth — from quick applicability checks to exhaustive

D Developer Integration

Developers incorporate RORI's agent into their own AI agentic workflows to retrieve regulations specific to an industry, region, and circumstance. RORI serves as a reliable context source within broader AI inference pipelines — a structured

P Packaged Applications

The agent is paired with a front-end and pre-configured for specific use cases: document-to-document comparison (e.g., a new Cyber Directive vs. NIST Standards), compliance gap analysis, rubric-based quality checks, or industry-specific

Use Case Examples

Use Case	Description	Modality
Regulatory Research	Query all applicable regulations for a mortgage product in a specific state	Agent / API
Gap Analysis	Compare a new Cyber Directive against NIST Standards and identify gaps	Packaged App
Quality Check	Validate work product against a requirements rubric for completeness	Packaged App
Context Manifold	Provide regulatory context to an external AI inference engine without web scraping	Developer API
Coverage Proposal	Generate comprehensive insurance coverage proposals with all applicable statutes	Packaged App

4 Core Principles

These five non-negotiable principles govern every architectural and product decision in RORI. They are the standard against which all design choices are measured.

Accuracy

The agent must be correct. Regulatory information cannot be 'mostly right.' Every response must be traceable to its source material. Hallucination is a disqualifying failure mode.

Consistency

The same query against the same corpus must produce the same result. Stochastic variation is unacceptable. Results must be repeatable and deterministic enough to withstand audit.

Completeness

The agent must know what it knows and what it doesn't. Partial answers must be labeled as such. When asked for all applicable regulations, it must confidently enumerate them or flag gaps.

Auditability

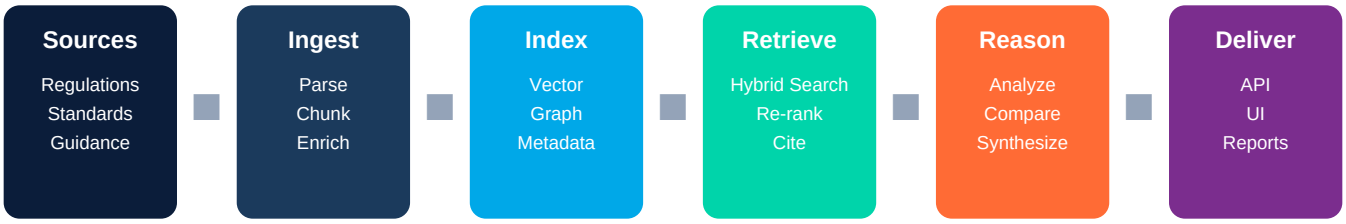
Every answer carries provenance: source documents, sections, versions, ingestion dates, verification timestamps. The platform must be defensible under regulatory scrutiny.

Tunability

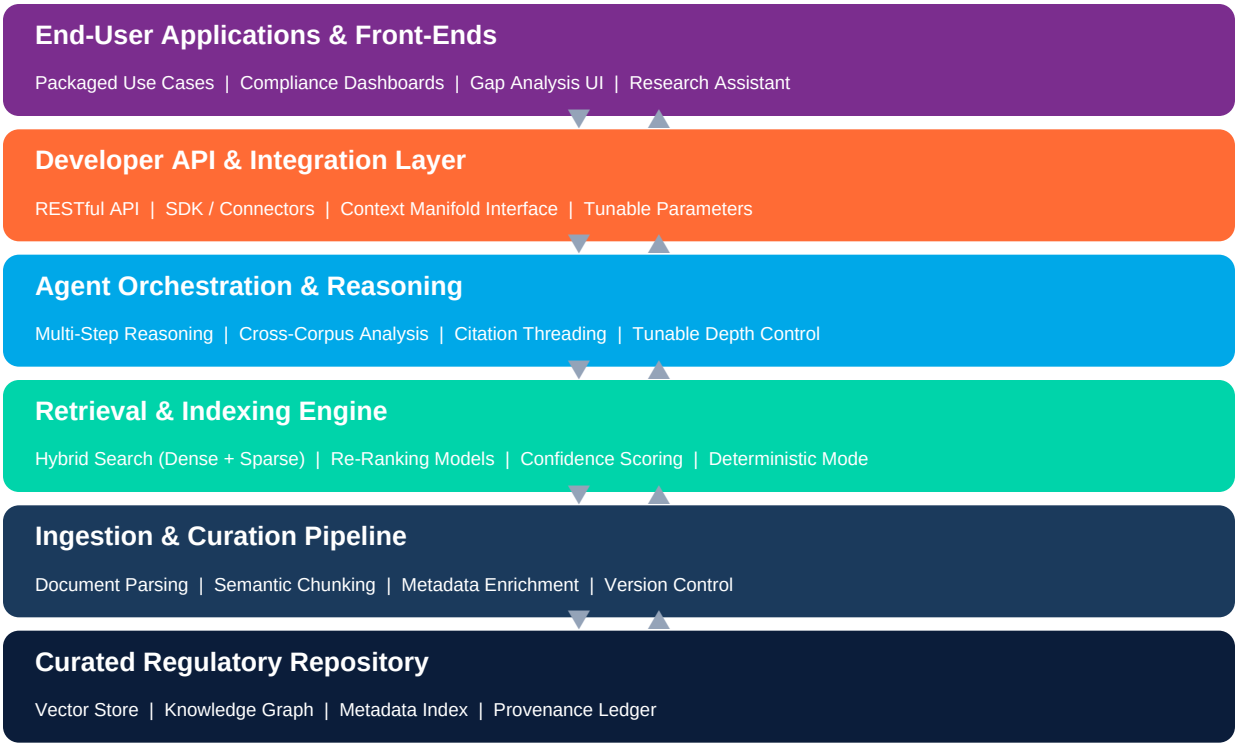
Not every query requires the same depth. The platform supports a spectrum from lightweight applicability checks to exhaustive regulatory audits, controlled by the caller.

5 Architecture & Data Flow

End-to-End Data Flow



Layered Platform Architecture



The architecture is layered to separate concerns and allow independent evolution of each tier. The curated repository forms the foundation; the ingestion pipeline feeds it; the retrieval engine queries it; the agent reasons over results; and the API and UI layers deliver value to end users and downstream systems. Provenance and citation thread through every layer.

6 Critical Focus Areas

6.1 Ingestion & Curation Pipeline

- Handling disparate document types (PDFs, HTML, legal XML, structured guides, unstructured guidance)
- Semantic chunking that preserves regulatory context, hierarchy, and cross-references
- Metadata extraction and enrichment (jurisdiction, effective dates, applicability, supersession chains)
- Version control and change tracking for living regulatory documents
- Quality gates and validation for ingested data

6.2 Indexing & Retrieval Architecture

- Hybrid retrieval (dense vector + sparse/lexical + structured metadata filtering)
- Advanced chunking optimized for legal and regulatory text (hierarchical, section-aware)
- Re-ranking models tuned for regulatory precision
- Citation-grounded generation — no claim without a cite
- Deterministic retrieval modes for audit repeatability
- Graph-based representations for regulatory relationships

6.3 Agent Architecture

- Agentic orchestration framework (tool-use, planning, self-correction)
- Tunable response depth (quick lookup vs. exhaustive analysis)
- Multi-step reasoning for gap analysis and cross-corpus comparison
- Guardrails against hallucination and unsupported assertions
- Context window management for large regulatory corpuses

6.4 Cross-Corpus Analysis

- Document-to-document comparison (directive vs. standard, policy vs. requirement)
- Gap analysis engines (required vs. covered)
- Rubric-based quality assessment
- Synthesis into structured output (compliance plans, gap reports, requirement matrices)

6.5 Developer API Surface

- Clean API for agent invocation with tunable parameters
- Context manifold integration for external AI engines
- SDK and integration patterns for common agentic frameworks
- Rate limiting, access control, and usage metering

6.6 Multilingual Support (Stage 2)

- Multilingual ingestion and indexing
- Cross-language retrieval (query in one language, retrieve from another)
- Translation-aware semantic matching
- Jurisdiction-language mapping

7 Industry Verticals

The platform is industry-agnostic by design, but will be validated through progressively complex verticals. Each stage proves out the engine's capabilities at increasing regulatory complexity and jurisdictional breadth.

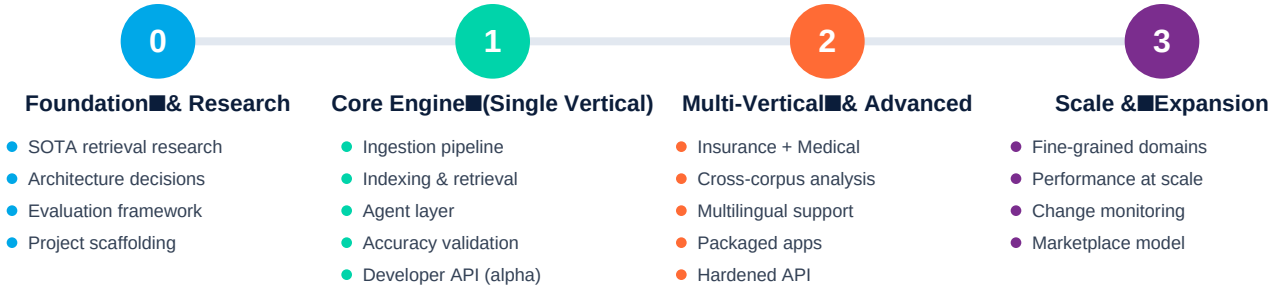
1. **Mortgage / First-Time Homebuyers**
   Federal/state mortgage regs, CFPB guidance, GSE seller/servicer guides, educational materials.

2. **Insurance (Broker/Agent Focus)**
   Insurance regulation across US states for brokers and agents generating coverage proposals. Highly jurisdictional.

3. **Medical / Gig Platforms for Clinicians**
   Regulations at the intersection of labor law, medical licensing, telehealth, and platform compliance.

4. **Fine-Grained Expansion**
   Municipal building codes, local zoning ordinances, industry sub-regulatory guidance. The deepest test of scale.

8 Staged Delivery Roadmap



Each stage builds on the foundation of the previous one. Stage 0 is critical — the research and architectural decisions made here will determine the ceiling of what the platform can achieve. We do not shortcut Stage 0.

9 Open Questions & Research Agenda

These critical questions must be resolved during Stage 0 before architectural commitments are made. Each represents a research area that directly impacts the platform's ability to meet its core principles.

1. **Repository Architecture**
   Vector DB, knowledge graph, hybrid? What combination best serves regulatory data with complex hierarchical relationships and cross-references?

2. **Chunking Strategy**
   What chunking approaches preserve the semantic integrity of regulatory text, which is heavily cross-referential and hierarchical?

3. **Retrieval Algorithms**
   What is the current state of the art for high-precision retrieval over legal/regulatory text? How do we achieve audit-grade repeatability?

4. **Evaluation Methodology**
   How do we rigorously measure accuracy, completeness, and consistency? What benchmarks exist for regulatory retrieval?

5. **Agent Framework**
   Which agentic orchestration patterns best support tunable depth and multi-step regulatory reasoning?

6. **Provenance & Citation**
   What are best practices for maintaining end-to-end citation provenance from source document through to agent response?

7. **Regulatory Change Mgmt**
   How do we handle continuous evolution of regulatory documents? Version tracking, supersession, effective date management?

10 Success Criteria

RORI will be measured against these concrete success criteria. Each maps directly back to the core principles and validates that the platform delivers on its promise.

Accuracy Exceeds Baselines
Agent retrieval accuracy measurably exceeds general-purpose LLM baselines on regulatory queries.

Repeatable Results
Same query, same corpus, same answer. Results are deterministic enough for audit.

Full Provenance
Every claim is traceable to a specific source, section, and version.

Actionable Gap Analysis
Cross-corpus gap analysis produces structured, actionable output.

Clean Developer Integration
API surface is clean enough to serve as context in third-party AI pipelines.

Audit Defensible
The system can withstand a simulated audit of its retrieval provenance and reasoning chain.

What This Document Is Not

This is not a technical architecture document, a sprint plan, or a detailed requirements specification. Those will follow. This is the project proposal — the *why*, the *what*, and the boundaries of the problem space. It is intended to align stakeholders, frame the research agenda, and provide a foundation for the architectural decisions that come next.