Compliance AI Content Generation Platform

Master Overview Document (POC)

1. Executive Summary

This document presents the **end-to-end overview** of the **Compliance AI Content Generation Platform**, a Proof of Concept (POC) designed for **fintech and insurance organizations** (e.g., Bajaj Finserv).

The platform automatically generates **marketing and sales content that is compliant by design**, embedding regulatory intelligence directly into the AI generation workflow. It eliminates manual compliance review loops, reduces regulatory risk, and provides predictable AI costs while remaining fully explainable and audit-ready.

This master document is intended to give reviewers, interviewers, and judges a **complete understanding of the POC vision, architecture, workflow, governance model, and cost feasibility** before diving into technical modules.

2. Problem Statement

In large fintech and insurance organizations:

- Marketing teams require fast content creation
- Compliance teams enforce strict regulatory checks (IRDAI, brand, SEO)
- Manual review cycles slow down campaigns
- Existing AI tools generate content first and fix compliance later, creating risk

There is a need for a system that **generates compliant content from the start**, not as an afterthought.

3. POC Objective

The objective of this POC is to demonstrate that:

- AI can generate high-quality marketing content **without violating regulatory rules**
- Compliance logic can remain **deterministic and auditable**
- LLMs can be safely used as language engines, not decision makers
- The system can scale economically for enterprise usage

This POC focuses on **architecture correctness, governance, and cost realism**, not UI polish.

---

4. Target Users & Roles

4.1 Agents (Business / Marketing Users)

- Create marketing prompts
- Upload existing content for compliance validation
- View final generated content and compliance status

4.2 Admins (Compliance Team)

- Manage rule activation and severity
- Monitor violations and rule hit frequency
- Ensure regulatory alignment

4.3 Super Admins (System Owners)

- Create and update compliance rules
- Upload regulatory documents (IRDAI)
- Control rule versions and system governance

---

5. High-Level System Philosophy

The system is built on the following non-negotiable principles:

1. **Rule-First Architecture** – Rules always override AI
2. **Compliance-by-Design** – Compliance embedded before generation
3. **Separation of Authority** – AI assists, rules decide
4. **Explainability** – Every decision is traceable
5. **Cost Predictability** – Token usage is controlled

---

6. End-to-End Workflow Overview

```
User Prompt / Uploaded Document
        ↓
Preprocessing & Intent Classification
        ↓
Prompt Enhancer (Compliance-Aware)
        ↓
Load Active Rules (SQL Database)
        ↓
Retrieve Regulatory Context (Vector DB – RAG)
```

```
        ↓
(Optional) Web Search Wrapper (Language Enrichment)
        ↓
LLM Content Generation (AWS Bedrock)
        ↓
Strict AI Compliance Review (AWS Bedrock)
        ↓
Deterministic Rule Validation
        ↓
Decision Aggregation Engine
        ↓
Final Compliant Content + Compliance Report
```

---

7. Document Ingestion, Chunking & Tokenization (High-Level)

- Supports PDF, DOCX, Markdown, and HTML
- Uses **structure-aware parsing** (headers → paragraphs → clauses)
- Applies **compliance-safe normalization** (no loss of legal meaning)
- Final chunks are token-based with overlap
- Metadata preserved for auditability

This ensures no disclaimers, financial terms, or legal clauses are lost.

---

8. Compliance Rules & Knowledge Management

8.1 Rule Storage Strategy

- **SQL Database (Amazon RDS)**: Source of truth for rules
- **Vector Database (Amazon OpenSearch)**: Semantic retrieval only

Rules are versioned, auditable, and never silently overwritten.

8.2 Regulatory Grounding (RAG)

- Titan Text Embeddings (AWS Bedrock)
- Clause-level retrieval of IRDAI regulations
- Used only to ground AI language, not to decide compliance

---

9. AI Architecture (Safe Multi-Model Design)

9.1 Content Generator Model

- AWS Bedrock (Claude / Llama)
- Responsible only for language generation

9.2 Strict Compliance Reviewer Model

- AWS Bedrock
- Conservative validation of generated content
- Outputs structured risk and violation signals

No model has final authority.

---

10. Prompt Enhancer (Why It Exists)

Users often submit vague or ineffective prompts, leading to: - Multiple retries - Increased token usage - Higher AI costs

The Prompt Enhancer: - Standardizes user intent - Injects compliance constraints - Reduces retries - Controls billing

---

11. Optional Web Search Wrapper (Perplexity-Style)

- Used only for language quality improvement
- Read-only, ephemeral, and sanitized
- Never updates rules or databases
- Compliance rules always override

---

12. Data Storage & Audit Strategy

Structured Data (Amazon RDS)

- Users
- Rules
- Compliance results
- Audit logs

Unstructured Data (Amazon S3)

- Uploaded documents
- Generated content
- Logs and artifacts

Vector Data (Amazon OpenSearch)

- Regulatory clause embeddings only

---

13. AWS Services Used

- Amazon Bedrock (LLMs + Titan Embeddings)
- Amazon OpenSearch (Vector DB)
- Amazon RDS (PostgreSQL)
- Amazon S3
- AWS Lambda
- Amazon EC2 / ECS
- AWS CloudWatch
- AWS IAM

---

14. Cost Overview (Executive Summary)

- Approx cost per 1,000 requests: **$20–21**
- Approx cost per request: **$0.02 (~₹1.6–1.7)**
- Optimized range: **$14–16 per 1,000 requests**

For a large fintech organization, this cost is **operationally insignificant** relative to the value delivered.

---

15. What This POC Deliberately Does NOT Include

- Fine-tuning of models
- Human-in-the-loop approvals
- Production UI polish
- Auto-updating rules via AI

These are conscious design choices to keep the POC safe and focused.

---

16. POC Success Criteria

The POC is successful if it demonstrates:

- Safe AI usage in a regulated domain
- Deterministic compliance enforcement
- Explainable AI outputs
- Realistic AWS cost modeling
- Scalable, enterprise-ready architecture

---

17. Final Takeaway

This POC demonstrates that compliant fintech content can be generated automatically by embedding regulatory intelligence into every stage of the AI workflow, delivering speed, safety, and governance without compromising auditability.

This document serves as the master reference. Detailed technical modules will follow.