AWS Costing & Billing Analysis

Compliance AI Content Generation Platform (POC)

> **Purpose:** This document provides a **realistic, reference-backed AWS costing analysis** for the Compliance AI Content Generation POC. It explains **why each AWS service is used**, **how AWS bills it**, and provides **minimum / typical / maximum cost ranges** grounded in official AWS pricing pages and reputable analyses. This document is suitable for conversion to **PDF, Markdown, or Word**.

---

1. Costing Philosophy (How to Read This Document)

AWS bills based on **actual usage**, not architecture diagrams. Costs here are driven by: - **Token usage** for LLM inference (input + output) - **Vector search compute & storage** for RAG - **Compute & storage** for backend services - **Observability data volume** (logs/metrics)

This document avoids blind numbers and instead: - Anchors ranges to **official AWS pricing** - Explains **what increases or decreases cost** - Separates **one-time** vs **recurring** charges

---

2. AWS Services Used – Why & How They Are Billed

2.1 Amazon Bedrock (LLMs + Embeddings)

Why used - Managed access to multiple foundation models - No model hosting or training required - IAM-controlled, enterprise-ready inference

What we use in Bedrock - Text generation model (e.g., Claude / Llama) - Strict reviewer model (same/different FM) - Titan Text Embeddings (for RAG)

How AWS bills - **Per 1,000 input tokens** - **Per 1,000 output tokens** - Embeddings billed per 1,000 tokens embedded

Official pricing - AWS Bedrock Pricing: https://aws.amazon.com/bedrock/pricing/

Practical ranges (model-dependent) - Input tokens: ~$0.0001 – $0.001 per 1k tokens - Output tokens: ~$0.0004 – $0.003 per 1k tokens

> Note: Exact rates depend on model and region. This POC intentionally keeps models interchangeable.

2.2 Amazon OpenSearch Service (Vector Database)

Why used - AWS-native vector search for RAG - Metadata filtering + semantic similarity - Integrates cleanly with Bedrock

What is stored - Embeddings of regulatory clauses only - No user data, no decisions

How AWS bills (Serverless) - **OpenSearch Compute Units (OCUs) per hour** - **Vector storage (GB/month)**

Official references - OpenSearch Serverless Overview: https://docs.aws.amazon.com/opensearch-service/latest/developerguide/serverless-overview.html - OpenSearch Pricing: https://aws.amazon.com/opensearch-service/pricing/

Important note (real-world behavior) - Serverless collections may incur **baseline OCU cost even when idle** - This makes OpenSearch a fixed component at low traffic

---

2.3 AWS Lambda (Orchestration)

Why used - Prompt enhancer - Rule checks - Decision aggregation

How AWS bills - Per invocation - Per GB-second of execution time

Official pricing - https://aws.amazon.com/lambda/pricing/

Cost behavior - Negligible at low to medium scale - Linear with invocations

---

2.4 Amazon RDS (PostgreSQL)

Why used - Source of truth for rules - Compliance results - Audit logs

How AWS bills - Instance-hours - Storage (GB/month) - I/O operations

Official pricing - https://aws.amazon.com/rds/postgresql/pricing/

POC guidance - Small instance (e.g., db.t3.small) sufficient - Can be reserved later for cost savings

---

2.5 Amazon S3 (Object Storage)

Why used - Uploaded documents - Generated content artifacts - Long-term audit storage

How AWS bills - Storage per GB/month - PUT/GET request counts

Official pricing - https://aws.amazon.com/s3/pricing/

2.6 CloudWatch & IAM

CloudWatch - Logs and metrics billed by ingestion & retention - https://aws.amazon.com/cloudwatch/pricing/

IAM - No direct cost - Required for enterprise security

---

3. Usage Assumptions (POC Baseline)

- 1 request = 1 content generation OR 1 document check
- Avg prompt (after enhancement): ~1,500 tokens
- Avg generated output: ~800 tokens
- Reviewer input/output: ~1,200 / 200 tokens
- Regulatory documents embedded once (amortized)
- Region: ap-south-1 (India)

---

4. Cost Scenarios (Minimum / Typical / Maximum)

4.1 Minimum (Low Usage / Early POC)

- Requests: ~1,000 / month
- Bedrock tokens: Low
- OpenSearch: Baseline serverless cost

Approx monthly range - Bedrock: $5 – $15 - OpenSearch: $50 – $150 - RDS + Lambda + S3 + Logs: $20 – $40

Estimated total: $75 – $200 / month

4.2 Typical (Pilot Usage)

- Requests: ~10,000 / month
- Regular generation + review

Approx monthly range - Bedrock: $150 – $350 - OpenSearch: $150 – $300 - Other infra: $50 – $100

Estimated total: $350 – $750 / month

4.3 Maximum (High Usage / Production-like)

- Requests: ~100,000 / month
- Heavy Bedrock usage dominates

Approx monthly range - Bedrock: $1,500 – $3,500 - OpenSearch: $300 – $800 - Other infra: $100 – $200

Estimated total: $2,000 – $5,000 / month

---

5. Cost Control & Optimization Levers

- Prompt enhancer reduces retries and token waste
- Rule-only short-circuiting for simple cases
- Smaller reviewer models
- Cache embeddings and enhanced prompts
- Periodic cleanup / cold storage in S3

---

6. Business Interpretation

For a large fintech organization, even the **upper-bound production cost** is: - Predictable - Controllable - Negligible relative to compliance risk reduction

The architecture prioritizes **governance first**, then optimizes cost.

---

7. Credits & References

- AWS Bedrock Pricing: https://aws.amazon.com/bedrock/pricing/
- OpenSearch Serverless Overview: https://docs.aws.amazon.com/opensearch-service/latest/developerguide/serverless-overview.html
- OpenSearch Pricing: https://aws.amazon.com/opensearch-service/pricing/
- AWS Lambda Pricing: https://aws.amazon.com/lambda/pricing/
- Amazon RDS PostgreSQL Pricing: https://aws.amazon.com/rds/postgresql/pricing/
- Amazon S3 Pricing: https://aws.amazon.com/s3/pricing/
- AWS CloudWatch Pricing: https://aws.amazon.com/cloudwatch/pricing/

This document is intended for technical and architectural review. Figures are approximate and depend on region, model choice, and workload patterns.