AWS Costing Summary

Compliance AI Content Generation Platform (1-Page)

💰 Costing Objective

Provide a **realistic, transparent estimate** of AWS costs for running the Compliance AI POC at enterprise scale.

☁️ AWS Services Used & Why

- **Amazon Bedrock** – LLM inference (generation + review) and embeddings
- **Amazon OpenSearch (Serverless)** – Vector search for regulatory RAG
- **Amazon RDS (PostgreSQL)** – Rules, results, audit logs
- **AWS Lambda** – Prompt enhancer & orchestration
- **Amazon S3** – Documents, outputs, logs
- **AWS CloudWatch** – Monitoring & observability

🔢 Key Usage Assumptions

- Avg input tokens per request: ~1,500
- Avg output tokens per request: ~800
- Reviewer tokens: ~1,400 per request
- Region: ap-south-1
- No fine-tuning

📊 Cost per 1,000 Requests (Approx)

Component	Cost (USD)
Bedrock – Content Generation	$10 – $12
Bedrock – Compliance Review	$4 – $6
Titan Embeddings	$1 – $2
OpenSearch (Vector DB)	$0.5 – $1
Lambda + Backend	$1 – $2
RDS + S3 + Logs	$1 – $2
Total	**$18 – $25**

Monthly Scale Examples

- **10,000 requests / month:** ~$350 – $750
- **100,000 requests / month:** ~$2,000 – $5,000

Cost Optimization Levers

- Prompt enhancement reduces retries
- Rule-only short-circuiting for simple cases
- Smaller reviewer models
- Caching of embeddings and prompts

Business Interpretation

For a large fintech organization, this cost is: - Predictable - Controllable - Insignificant compared to compliance risk mitigation

Executive Takeaway

The platform delivers compliance-safe AI content generation at roughly ₹1.5–2 per request, making it operationally viable for enterprise adoption.

Figures are approximate and based on official AWS pricing. Actual costs depend on model choice and usage patterns.