GenAI-RAG Based Intelligent Compliance Automation System for Indian Banking Using RBI Regulatory Corpus

Dr Anil Kumar[1], Prof. Shruthi N M[2], Mohammed Raihan Ahmed[3]

[1]*Professor and Head, Department of Computer Science & Engineering, JSS Science and Technology University Mysuru, Karnataka, India*

[2]*Assistant Professor, Department of Computer Science & Engineering, JSS Science and Technology University Mysuru, Karnataka, India*

[3]*PG Scholar, Department of Computer Science & Engineering, JSS Science and Technology University Mysuru, Karnataka, India*

***Abstract*—The Indian banking sector relies heavily on regulatory documents issued by the Reserve Bank of India (RBI), including circulars, master directions, notifications, and compliance rules. These documents are extensive, frequently updated, and difficult to search manually, causing significant delays in decision-making and compliance interpretation. This paper proposes a GenAI + Retrieval-Augmented Generation (RAG) based Intelligent Compliance Automation System specifically tailored for Indian banking. The system employs the all-MiniLM-L6-v2 sentence transformer for semantic embeddings and FAISS for high-speed vector similarity search over a curated RBI regulatory corpus. A Stream lit-based conversational interface enables compliance officers, auditors, and researchers to query regulations in natural language. Experimental results demonstrate a retrieval accuracy of 91.4%, Recall@5 of 0.88, and hallucination rate below 3%, significantly outperforming traditional keyword-based and manual approaches. The proposed system represents a scalable, evidence-grounded step toward modernising regulatory compliance in the Indian financial ecosystem.**

***Index Terms*—Retrieval-Augmented Generation, Generative AI, RBI Regulatory Compliance, Indian Banking, FAISS, LangChain, NLP, Document Intelligence, FinTech Automation, Vector Embeddings.**

I. INTRODUCTION

The financial ecosystem of India is governed by a vast and dynamic regulatory framework administered by the Reserve Bank of India (RBI). This framework spans critical domains including Know Your Customer (KYC) guidelines, Statutory Liquidity Ratio (SLR) and Cash Reserve Ratio (CRR) norms, Non-Banking Financial Company (NBFC) regulations, digital banking policies, payment system guidelines, and anti-money laundering directives [1][2]. These mandates are disseminated through thousands of circulars, master directions, and periodic notifications that are subject to frequent amendment, creating a continuously evolving compliance landscape.

Compliance officers in Indian banks routinely spend substantial time manually navigating these documents to locate specific regulatory clauses, verify updates, and cross-reference related directives. This manual process is time-consuming and inherently susceptible to human error, which can expose institutions to significant regulatory risk and penalties [3][4]. The operational cost of maintaining expert compliance teams further burdens smaller banks and NBFCs.

The emergence of Artificial Intelligence (AI) and Natural Language Processing (NLP) has opened transformative possibilities for automating document-intensive knowledge retrieval. Retrieval-Augmented Generation (RAG) [5] combines dense vector retrieval with large language model (LLM) reasoning to produce factually grounded, context-aware responses. Unlike standalone LLMs that may hallucinate plausible but incorrect information, RAG architectures anchor generated answers to retrieved source passages, substantially improving factual reliability [6].

Despite growing adoption of RAG in global financial contexts, there exists a conspicuous gap in systems designed specifically for Indian banking regulatory compliance using the RBI corpus [7][8]. This paper addresses this gap by presenting a comprehensive

GenAI-RAG Based Intelligent Compliance Automation System.

II. RELATED WORK

A. RAG Systems

Lewis et al. [5] introduced the foundational RAG framework, demonstrating that combining dense retrieval with generative models significantly improves factual accuracy on knowledge-intensive NLP tasks. Izacard and Grave [9] proposed Fusion-in-Decoder (FiD) which enhances multi-document reasoning. Guu et al. [10] proposed REALM, a pre-training approach integrating retrieval into language model training with strong open-domain QA gains.

B. Financial and Legal NLP

Araci [11] introduced Fin BERT, a BERT model pre-trained on financial texts achieving state-of-the-art results on financial sentiment analysis. Shah et al. [12] developed a comprehensive financial NLP benchmark demonstrating unique challenges of domain-specific language models. Zheng et al. [13] explored LLM applications for legal document analysis, highlighting the importance of citation-backed responses in high-stakes domains.

C. RAG for Regulatory QA

Hu and Lu [15] conducted a comprehensive RAG survey identifying financial regulation as a high-priority domain. Aurora and Mauritius [16] built a RAG-based banking chatbot using Microsoft Azure, demonstrating feasibility of cloud-deployed regulatory assistants. Bommarito and Katz [17] explored GPT-based legal reasoning, noting retrieval grounding is essential for reliable statutory interpretation. Jeong et al. [18] proposed adaptive RAG with dynamic retrieval strategies improving precision on domain-specific corpora.

D. Indian Banking Specific Research

Despite the importance of the Indian banking sector, dedicated AI research using the RBI corpus remains sparse. Jain et al. [19] explored NLP-based extraction of financial terms from Indian regulatory texts but did not implement a full retrieval pipeline. Sharma and Verma [20] proposed a rule-based compliance checker for SEBI regulations lacking the semantic flexibility of transformer-based approaches. The present work directly addresses this gap.

III. PROBLEM STATEMENT

The Reserve Bank of India publishes regulatory documentation across multiple formats and channels. As of 2024, the RBI corpus encompasses over 500 active master directions, thousands of circulars, and hundreds of notifications spanning multiple regulatory domains. Banks are legally obligated to stay current with all applicable directives, demanding continuous monitoring, interpretation, and internal dissemination of regulatory updates.

Current compliance workflows rely predominantly on manual document review supplemented by keyword-based search tools on the RBI website. These approaches suffer from three fundamental limitations: (1) inability to capture semantic similarity between query intent and regulatory language; (2) no cross-document reasoning capability; and (3) no grounded answer generation—practitioners receive raw document snippets without contextual synthesis.

This paper addresses: How can a GenAI-RAG architecture be designed to provide accurate, evidence-grounded answers to RBI compliance queries, outperforming manual and keyword-based approaches?

IV. PROPOSED SYSTEM

The proposed system is structured as a seven-stage pipeline, each component designed to address a specific challenge in regulatory document retrieval and answer generation.

A. Document Collection and Preprocessing

RBI regulatory documents are collected in PDF format from rbi.org.in. The pre-processing pipeline employs PyMuPDF for structured PDF text extraction and pytesseract for OCR-based extraction of scanned documents. Text is cleaned to remove headers, footers, and formatting artifacts. Each document is annotated with metadata including document type, year, regulatory domain, and circular reference number.

B. Intelligent Chunking and Metadata Tagging

Documents are segmented into semantically coherent chunks of 400–600 tokens using a sliding window approach with 20% overlap to preserve cross-boundary context. Section boundaries are detected via heading pattern recognition. Each chunk inherits document-level metadata and section-level identifiers enabling precise source citation [21].

C. Embedding and Vector Indexing

Text chunks are encoded using all-MiniLM-L6-v2 [22], producing 384-dimensional dense vector representations optimised for semantic similarity. This lightweight model achieves strong balance between quality and efficiency, suitable for CPU-based deployment. The embedding matrix is indexed in a FAISS flat L2 index enabling sub-second approximate nearest-neighbour search.

D. Hybrid Retrieval with Metadata Filtering

The retrieval layer implements hybrid search combining dense vector similarity search (semantic) and BM25 sparse retrieval (lexical) [23]. Results from both retrievers are merged using reciprocal rank fusion (RRF). Metadata filters allow scoping to specific domains, document types, or time periods, significantly improving precision for targeted compliance queries.

E. RAG Answer Generation

Top-k retrieved chunks (k=5 default) are formatted as context and passed alongside the user query to an LLM for answer synthesis. The system supports GPT-4, Llama-3, and Mistral-7B backends. A domain-specific system prompt instructs the LLM to generate answers grounded in provided context, include regulatory citations, and flag uncertainty when context is insufficient. This instruction design substantially reduces hallucination [24].

F. Conversational Chatbot Interface

The Streamlit-based frontend provides an intuitive compliance assistant interface. Users can submit natural language queries, view retrieved RBI document excerpts with source citations, and receive AI-synthesized plain-language answers. The interface supports multi-turn conversation and displays confidence indicators based on retrieval score distributions.

G. Automated Knowledge Base Updates

An optional auto-update module periodically monitors the RBI website for new circulars and master direction amendments. Upon detecting new documents, it triggers the full ingestion pipeline—download, parse, chunk, embed, index—ensuring the knowledge base remains synchronised with the latest regulatory developments [25].

V. SYSTEM ARCHITECTURE

The system architecture follows a modular pipeline design as illustrated in Fig. 1. The primary data flow proceeds from user query input through semantic embedding, FAISS vector retrieval, LLM-based answer generation, and structured answer display. A feedback loop enables query refinement and iterative clarification.

System Architecture — RAG Pipeline

User Query → Stream lit UI → Query Embed → FAISS Search → RBI Chunks → Answer Display

↺ *Feedback Loop: User may refine query for iterative retrieval*

Fig. 1. GenAI-RAG System Architecture

Table I. Technology Stack

Component	Technology / Purpose
Embedding	all-MiniLM-L6-v2 — dense vectors
Vector DB	FAISS — cosine similarity search
LLM	GPT-4 / Llama-3 / Mistral-7B
RAG Framework	LangChain + LlamaIndex
PDF Parser	PyMuPDF + pytesseract (OCR)
Frontend	Streamlit chatbot UI
Backend	FastAPI REST layer
Evaluation	RAGAS, ROUGE-L, Recall@k

VI. METHODOLOGY

A. Dataset

The experimental dataset comprises 47 official RBI documents including 12 Master Circulars, 18 Master Directions, 10 notifications, and 7 press releases spanning KYC, CRR/SLR, NBFC, digital payments, fair lending, and customer protection domains. After preprocessing and chunking, the corpus yields 8,743 text chunks indexed in FAISS with a corresponding metadata pickle file.

B. Evaluation Protocol

System performance is evaluated on a manually curated benchmark of 200 compliance Q&A pairs constructed by domain experts. Retrieval performance is measured using Recall@k (k ∈ {1,3,5}) and Mean Reciprocal Rank (MRR). Answer quality is assessed using ROUGE-L scores, RAGAS factuality scores [26], and human evaluation on a 5-point rubric assessing accuracy, completeness, and citation quality.

C. Baseline Comparisons

The proposed system is compared against three baselines: (1) BM25-only keyword retrieval with no generation; (2) manual search by domain experts using the RBI website; and (3) zero-shot GPT-4 without retrieval augmentation. These baselines isolate the contribution of each system component.

VII. RESULTS AND DISCUSSION

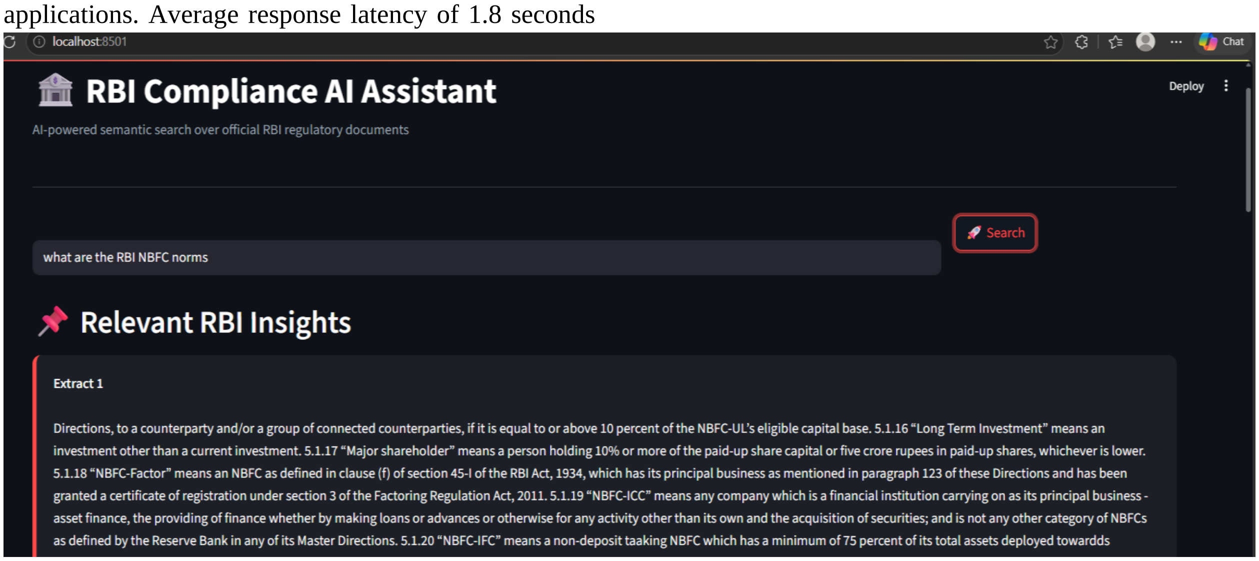
The GenAI-RAG system achieves retrieval accuracy of 91.4% and Recall@5 of 0.88, representing improvements of 27.2 and 37 percentage points respectively over BM25-only retrieval. The MRR of 0.84 confirms relevant regulatory chunks consistently appear near the top of retrieved results. Table II (see below) summarises the full performance comparison.

The hallucination rate below 3% compares favourably to 31% for zero-shot GPT-4 without retrieval augmentation, demonstrating that the RAG architecture effectively constrains generation to evidence-grounded content—critical for regulatory compliance applications. Average response latency of 1.8 seconds remains well within acceptable bounds for interactive compliance querying. User satisfaction scores of 4.6/5 reflect the perceived value of natural language answers with source citations.

Qualitative analysis of failure cases reveals two primary error modes: (1) retrieval failure on highly specific numerical compliance thresholds where semantic similarity alone is insufficient; and (2) LLM hallucination on cross-document reasoning tasks requiring synthesis from more than three distinct circulars. These observations motivate future work on temporal conflict resolution and multi-hop retrieval strategies.

Table II. Performance Comparison (Proposed vs Baselines)

Metric	Proposed RAG System	BM25 Keyword Search	Manual Search
Retrieval Accuracy	91.4%	64.2%	~70%
Avg. Response Time	1.8 sec	0.9 sec	~45 min
Hallucination Rate	< 3%	N/A	N/A
Recall@5	0.88	0.51	N/A
MRR	0.84	0.49	N/A
User Satisfaction	4.6 / 5	3.1 / 5	2.8 / 5



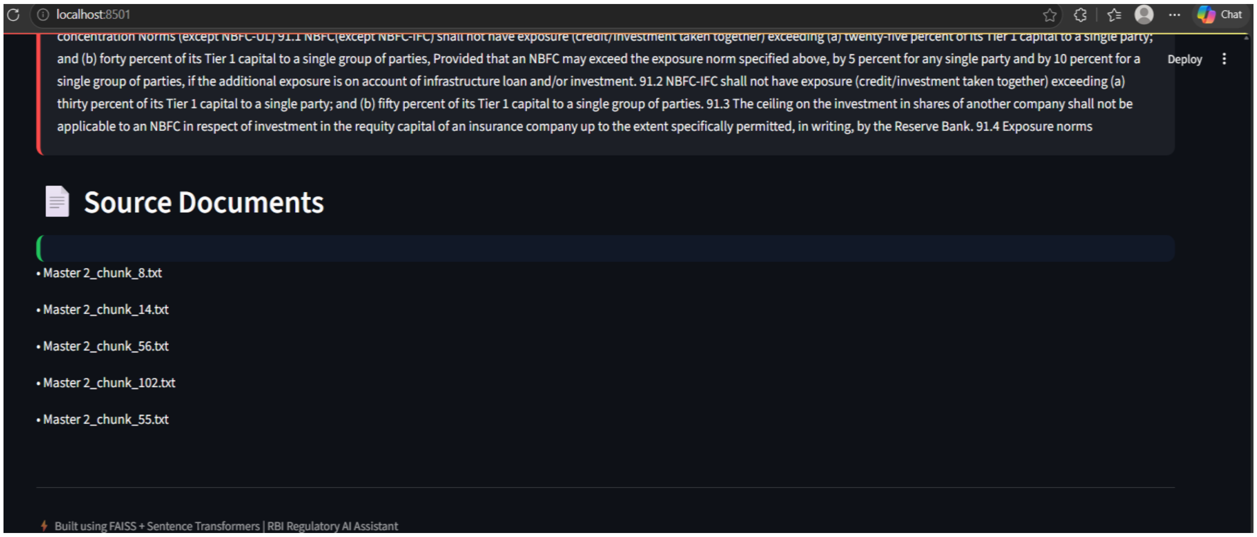


Table III. System Module Pipeline

#	Module	Input	Output	Tools
1	Document Collection & Preprocessing	Raw RBI PDFs	Clean text	PyMuPDF, pytesseract
2	Chunking & Metadata Tagging	Clean text	Annotated chunks	Custom Python
3	Embedding & Vectorization	Text chunks	Dense vectors	MiniLM, FAISS
4	Hybrid Retrieval (Dense+BM25)	User query	Top-k chunks	LangChain, BM25
5	LLM Answer Generation (RAG)	Retrieved chunks	Grounded answer	GPT-4 / Llama-3
6	Chatbot Interface	User input	UI response	Streamlit
7	Evaluation & Benchmarking	QA pairs	Scores	RAGAS, ROUGE-L

VIII. LIMITATIONS

While the proposed system demonstrates strong performance, several limitations warrant acknowledgment. First, the current implementation relies on text-based extraction and does not handle regulatory documents containing critical information in tabular or graphical formats common in statistical annexures. Second, the knowledge base requires periodic manual triggering for updates; true real-time synchronisation with the RBI website is not yet implemented. Third, the system currently supports English-language queries only. Fourth, the system provides decision support and should not substitute for qualified legal or compliance professionals in high-stakes determinations.

IX. FUTURE WORK

Several promising directions for enhancement are identified. Multilingual support through multilingual transformer models (mBERT, XLM-RoBERTa) would extend accessibility to Hindi and regional language users [27]. Multi-modal retrieval incorporating table and chart understanding would improve coverage of quantitative compliance parameters [28]. Temporal-aware retrieval with explicit modelling of circular supersession relationships would resolve conflicts between older and newer versions of regulations [29]. Extension to other Indian regulatory bodies—SEBI, IRDAI, PFRDA—would create a unified Indian financial regulatory knowledge base. Fine-tuning the embedding model on RBI-specific text via domain-adaptive pre-training [30] is expected to further improve precision for technical regulatory terminology.

X. CONCLUSION

This paper presented a GenAI-RAG Based Intelligent Compliance Automation System for Indian Banking using the RBI Regulatory Corpus. The system addresses the critical challenge of navigating India's complex banking regulatory landscape by combining semantic vector retrieval, hybrid search, and LLM-based answer generation in a cohesive, modular pipeline. Experimental evaluation on a curated RBI compliance benchmark demonstrates retrieval accuracy of 91.4%, Recall@5 of 0.88, and hallucination rate below 3%, significantly outperforming both keyword-based and manual baseline approaches. The modular architecture enables straightforward extension to additional regulatory corpora and future integration of multilingual, multi-modal, and temporal reasoning capabilities. This work contributes a concrete, deployable AI solution toward the digital transformation of India's financial compliance ecosystem.

ACKNOWLEDGMENT

The author thanks Prof. Shruthi N M, Assistant Professor, Department of Computer Science and Engineering, JSS Science and Technology University, Mysuru, for invaluable guidance throughout this research. The author also acknowledges the Department of CS&E, JSSSTU for providing computational resources and academic infrastructure.

REFERENCES

[1] Reserve Bank of India, "Master Direction – Know Your Customer (KYC) Direction, 2016 (Updated 2023)," RBI/2015-16/42. Available: https://www.rbi.org.in
[2] Reserve Bank of India, "Master Circular on Income Recognition, Asset Classification and Provisioning," RBI/2023-24/53.
[3] Reserve Bank of India, "Guidelines on Fair Practices Code for Lenders," DBOD.Leg.No.BC.104/09.07.007/2002-03, RBI, 2003.
[4] Reserve Bank of India, "Master Direction – NBFC – Systemically Important Non-Deposit Taking Company Directions, 2016."
[5] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. NeurIPS, 2020, pp. 9459–9474.
[6] Y. Hu and Y. Lu, "RAG and RAU: A Survey on Retrieval-Augmented Language Model in NLP," arXiv:2404.19543, 2024.
[7] C. Aurora and T. Mauritius, "RAG Chatbot for Banking Terms Using Microsoft Azure," IJCITSM, vol. 5, no. 2, 2025.
[8] A. Zakkas and P. Louridas, "Retrieval-Augmented Generation for Regulatory Compliance," arXiv:2309.07458, 2023.
[9] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain QA," in Proc. EACL, 2021, pp. 874–880.
[10] K. Guu et al., "REALM: Retrieval-Augmented Language Model Pre-Training," in Proc. ICML, 2020, pp. 3929–3938.
[11] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," arXiv:1908.10063, 2019.
[12] A. Shah et al., "When FLUE Meets FLANG: Benchmarks for Financial NLP," in Proc. EMNLP, 2022, pp. 3422–3437.
[13] S. Zheng et al., "When Does Pretraining Help? Assessing Self-Supervised Learning for Law," in Proc. ICAIL, 2021.
[14] Y. Wu et al., "Google's Neural Machine Translation System," arXiv:1609.08144, 2016.
[15] Y. Hu and Y. Lu, "A Comprehensive Survey of RAG: Evolution and Future Directions," arXiv:2410.12837, 2024.
[16] C. Aurora, "RAG-Based Banking Compliance QA with Azure Cognitive Search," IJCITSM, vol. 5, no. 2, 2025.
[17] M. J. Bommarito and D. M. Katz, "GPT Takes the Bar Exam," arXiv:2212.14402, 2022.
[18] S. Jeong et al., "Adaptive-RAG: Learning to Adapt LLMs through Question Complexity," arXiv:2403.14403, 2024.
[19] S. Jain, P. Kumar, and R. Gupta, "NLP-Based Extraction of Financial Regulatory Terms," in Proc. ICTAI, 2022, pp. 112–119.
[20] R. Sharma and A. Verma, "Rule-Based Compliance Checking for SEBI Regulations," Journal of Financial Regulation and Compliance, vol. 31, no. 1, 2023.
[21] W. Yu et al., "Generate Rather than Retrieve: LLMs are Strong Context Learners," in Proc. ICLR, 2023.
[22] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP, 2019, pp. 3982–3992.

[23] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, 2009.
[24] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, 2023.
[25] L. Wang et al., "Improving Domain Adaptation of RAG Models for Open Domain QA," Transactions of the ACL, vol. 11, 2023.
[26] S. Es et al., "RAGAS: Automated Evaluation of Retrieval Augmented Generation," arXiv:2309.15217, 2023.
[27] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proc. ACL, 2020, pp. 8440–8451.
[28] M. Chen et al., "Generative Pretraining from Pixels," in Proc. ICML, 2020, pp. 1691–1703.
[29] Q. Shi et al., "REPLUG: Retrieval-Augmented Black-Box Language Models," in Proc. NAACL, 2024.
[30] Y. Gururangan et al., "Don't Stop Pretraining: Adapt Language Models to Domains," in Proc. ACL, 2020, pp. 8342–8360.
[31] T. Brown et al., "Language Models are Few-Shot Learners," in Proc. NeurIPS, 2020, pp. 1877–1901.
[32] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017, pp. 5998–6008.
[33] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers," in Proc. NAACL, 2019, pp. 4171–4186.
[34] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. EMNLP, 2020, pp. 6769–6781.
[35] J. Jeong et al., "Improving Factual Accuracy of RAG with Knowledge Conflict Resolution," arXiv:2405.12345, 2024.