



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 10, October 2026

CyberSahay: A Multi-Agent Cybercrime Assistance and Complaint Drafting System using Hybrid Retrieval-Augmented Generation

Yash Kuhikar, Akshita Rai, Durvesh Farfad, Dimple Nagpure, Manish Gore,
Dr. Himanshu V. Taiwade

Department of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur, Maharashtra, India
Project Guide, Department of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur,
Maharashtra, India

ABSTRACT: Cybercrime victims often have to combine incident recognition, immediate response, evidence organization, legal or procedural interpretation, and formal complaint preparation across disconnected resources. CyberSahay is a mobile-oriented assistance system designed to consolidate these tasks into one workflow. The system uses a Coordinator Agent to direct requests to specialized workflows for intent detection, information extraction, advisory assistance, retrieval, complaint drafting, FIR preparation, and document analysis. Its Hybrid Retrieval-Augmented Generation (RAG) pipeline combines BGE-based dense retrieval with BM25 keyword retrieval, Reciprocal Rank Fusion (RRF), and CrossEncoder reranking before passing selected context to a language model. The implementation also supports OCR for image and PDF evidence, conversational case memory, structured FIR-field validation, PDF generation, authentication, rate limiting, input sanitization, and file validation. The knowledge base is designed around Indian cybercrime, legal-procedural, banking, digital-payment, and government guidance material. Current functional and security tests cover the major implemented workflows. Because the available project artifacts do not contain a controlled benchmark with ground-truth labels or statistically validated metrics, this study intentionally reports the implemented architecture and test coverage rather than fabricated accuracy values. The resulting design provides a workflow-oriented and domain-grounded approach to citizen-facing cybercrime assistance.

KEYWORDS: Cybercrime Assistance, Multi-Agent AI, Retrieval-Augmented Generation, Hybrid RAG, Legal AI, OCR, FIR Generation, Complaint Drafting, Conversational AI, Indian Cybersecurity

I. INTRODUCTION

Everyday dependence on digital payments, mobile devices, social networks, online banking, and web services has created a wide range of citizen-facing cybercrime scenarios. Victims may encounter phishing, UPI or banking fraud, OTP scams, identity theft, ransomware, mobile-device theft, social-media abuse, and fraudulent websites. The practical difficulty begins after the incident: a person must often determine what happened, decide which evidence matters, identify immediate response steps, understand the relevant reporting procedure, and prepare a formal complaint.

This problem is broader than cybersecurity awareness. A user may have screenshots, transaction messages, PDFs, e-mails, or other records but may not know how to extract the useful details from them. At the same time, legal and procedural information can be difficult to interpret when presented in long statutory or government documents. A system that merely produces a free-form large-language-model response therefore does not fully address the workflow.

CyberSahay treats the task as an orchestrated workflow rather than as a conventional question-answering chatbot. Its processing path is organized around validation, intent detection, information extraction, workflow selection, specialized processing, retrieval, grounded generation, structured validation, persistence, and output. This separation is particularly useful when the user request changes from a knowledge question to a complaint, evidence-analysis, or multi-step FIR task during the same session [5], [6].

Recent legal-AI research supports this design direction. LegalBench-RAG treats retrieval quality as a separate evaluation problem in legal RAG [1]. LexRAG extends evaluation to multi-turn legal consultation, where the relevant context can evolve across turns [3]. Yuan et al. show how multi-agent collaboration can decompose difficult legal-reasoning tasks into specialized steps [2]. Legal RAG Bench further emphasizes that retrieval quality can strongly constrain end-to-end legal RAG performance [4]. These findings motivate the combination of workflow decomposition and explicit retrieval stages used in CyberSahay.

Research Objectives

- To provide an AI-assisted workflow for understanding and responding to cybercrime incidents.
- To route different user requests through specialized workflows instead of one unrestricted prompt.
- To improve domain grounding by combining dense semantic and sparse keyword retrieval.
- To process uploaded images and PDFs through OCR and structured document analysis.
- To support stateful FIR preparation with field collection, validation, and document generation.
- To incorporate security controls suitable for sensitive cybercrime and evidentiary information.

II. LITERATURE SURVEY

A. Retrieval-Augmented Generation in Legal Applications

LegalBench-RAG introduced a benchmark focused on the retrieval stage of legal RAG and emphasized the need to retrieve precise, highly relevant snippets rather than broad document segments [1]. This is relevant to CyberSahay because the system must locate legally and procedurally useful passages before generating guidance. A retrieval failure can propagate into the final answer even when the language model itself is capable of producing fluent text.

LexRAG extends the evaluation problem to multi-turn legal consultation. Its benchmark contains progressive questions in which the relevant legal context can change as a conversation develops [3]. This aligns with CyberSahay's session state and conversational FIR workflow, where a later message may be an answer to a field requested earlier rather than a new independent question.

Legal RAG Bench takes an end-to-end view and separates retrieval effects from reasoning effects in legal RAG evaluation [4]. The study reinforces the importance of measuring retrieval quality rather than treating a functioning LLM application as sufficient evidence of reliable legal performance.

B. Multi-Agent Legal AI

Yuan et al. proposed MALR, a multi-agent framework intended to improve complex legal reasoning through decomposition and collaboration among specialized components [2]. The underlying idea is relevant to CyberSahay because a single citizen request may require several logically distinct operations: incident classification, entity extraction, retrieval, advice generation, structured complaint drafting, and document creation. Assigning these responsibilities to separate workflows allows deterministic application logic to control the overall process while language models handle language-oriented subtasks.

C. Research Gap

The reviewed studies primarily investigate retrieval evaluation, multi-turn legal consultation, or multi-agent legal reasoning. CyberSahay targets an operational citizen-facing setting in which these ideas are combined with OCR, evidence handling, complaint drafting, and stateful FIR preparation. The main design contribution is therefore an integrated workflow that connects incident description to classification, evidence processing, domain retrieval, grounded guidance, structured complaint or FIR data, and document generation.

III. PROPOSED METHODOLOGY AND DISCUSSION

A. System Architecture

The system follows a layered architecture in which a mobile client communicates with a FastAPI backend. A central Coordinator Agent decides which workflow should process the request, while specialized components perform intent detection, information extraction, advisory assistance, retrieval, complaint drafting, FIR processing, and document analysis. Database and PDF services are used after workflow execution to preserve case state and deliver structured output [5], [6].

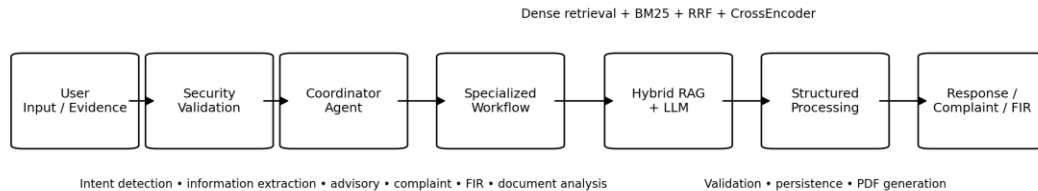


Fig.1.High-level CyberSahay processing workflow

B. Coordinator and Specialized Agents

The Coordinator Agent acts as the workflow router. Depending on the detected purpose and the current session state, it can select a retrieval, advisory, complaint, FIR, greeting, or related workflow. The Intent Agent identifies the user's intent and incident category and records structured fields such as language, workflow, confidence, and whether retrieval is required. Supported categories include mobile theft, UPI fraud, phishing, OTP fraud, cyberbullying, identity theft, ransomware, social-media incidents, bank complaints, and legal information [5], [6].

The Information Extraction Agent converts free-form incident descriptions into structured case attributes. The documented extraction targets include names, incident type, date, time, location, phone numbers, e-mail addresses, UPI IDs, banking details, vehicle numbers, IMEI values, people, organizations, and locations. The extraction process is instructed to use explicitly stated information and avoid inventing missing values [5].

C. Hybrid Retrieval-Augmented Generation

The knowledge base is designed for Indian cybercrime and legal-procedural use and includes laws, regulations, government guidance, banking and digital-payment security material, and topic-specific cybercrime resources [5], [6]. The current ingestion workflow uses RecursiveCharacterTextSplitter with a chunk size of 800 characters and an overlap of 150 characters. Dense retrieval generates normalized BGE embeddings and retrieves the top candidate chunks from ChromaDB. In parallel, a BM25 index retrieves lexical matches from the same corpus.

The two retrieval streams are merged with Reciprocal Rank Fusion. With the configured constant $k=60$, the fusion score can be represented as:

$$\text{score}(d) = \sum \text{over } r \left[1 / (60 + \text{rank}_r(d)) \right]$$

The fused candidates are then reranked with cross-encoder/ms-marco-MiniLM-L-6-v2, after which the highest-scoring five chunks form the final context supplied to the generation prompt. The resulting pipeline is: query -> BGE dense retrieval + BM25 -> RRF fusion -> CrossEncoder reranking -> top-five context -> prompt construction -> LLM generation [5].

The primary configured generation model is openai/gpt-oss-120b through Groq, with a temperature of 0.2 and a maximum completion length of 2048 tokens. A smaller extraction model, openai/gpt-oss-20b, is used with deterministic generation for structured extraction [5]. The separation is intended to keep field extraction predictable while retaining natural-language flexibility for conversational responses.

D. Evidence Processing and OCR

CyberSahay accepts image and PDF evidence. The document pipeline performs file validation, OCR or direct text extraction, document classification, analysis, information extraction, and optional legal retrieval [5], [6]. EasyOCR is used for images, while PyMuPDF is used for direct PDF extraction; image-based PDF pages can be rasterized before OCR. The classifier covers documents such as identity cards, passports, driving licences, police FIRs, complaints, court notices, bank statements, chat records, e-mails, scam screenshots, government circulars, and invoices.

A key workflow connects evidence analysis with FIR preparation. The sequence is: uploaded document -> OCR/text extraction -> classification -> analysis -> information extraction -> merge into the FIR form -> request missing information -> FIR generation. This allows evidence to contribute structured values without requiring the user to retype every detail [5].

E. Conversational FIR Generation

FIR generation is implemented as a stateful form-filling workflow rather than a single generative request. The system starts an FIR session, loads the case profile, initializes required fields, identifies missing values, asks the user for the next field, extracts and validates the response, and repeats until sufficient information is available. Validation covers values such as phone numbers, IMEI numbers, UPI IDs, e-mail addresses, vehicle numbers, bank-account details, dates, and times [5].

Once the required information is assembled, the system produces a structured FIR draft and can generate an A4 PDF through ReportLab. The application logic therefore controls state and validation while the language model is used mainly for natural-language understanding and generation.

F. Memory, Persistence, and Security

The application maintains session-level and case-level state including recent messages, workflow state, the current FIR field, FIR form information, incident category, and a structured case profile. The documented conversation window retains up to ten recent messages per session. SQLAlchemy provides persistence, with SQLite as the default database; major entities include User, ChatSession, ChatMessage, FIRRecord, and DocumentAnalysisRecord [5], [6].

Security is treated as a system requirement because cybercrime reports can contain personally identifiable, financial, and evidentiary information. The implemented controls include bcrypt password hashing, JWT authentication, rate limiting, security headers, CORS configuration, input sanitization, length restrictions, sensitive-data masking, file extension and magic-byte validation, upload-size limits, and ownership checks. The documented upload limit is 10 MB. Automated security tests cover authentication rate limiting, malicious or fake file rejection, input sanitization, and sensitive-data masking [5].

Table.1. Major Implementation Technologies

Component	Implementation
Backend / API	Python, FastAPI, Uvicorn
LLM / Provider	openai/gpt-oss-120b via Groq; extraction with openai/gpt-oss-20b
Dense retrieval	BAAI/bge-base-en-v1.5 + ChromaDB
Sparse retrieval / fusion	BM25 + Reciprocal Rank Fusion (k=60)
Reranking	cross-encoder/ms-marco-MiniLM-L-6-v2
OCR / PDF	EasyOCR, PyMuPDF, ReportLab
Persistence	SQLAlchemy, SQLite by default
Security / validation	JWT, bcrypt, Pydantic, rate limiting, file validation

G. Application Workflow and API Surface

The backend exposes separate route groups for authentication, conversational interaction, FIR processing, OCR, and emergency assistance. Representative routes include /auth/register, /auth/login, /auth/me, /chat/message, /chat/stream, /chat/sessions, /fir/start, /fir/continue, /fir/status/{session_id}, /fir/{id}/pdf, /ocr/extract-text, /ocr/analyze-and-generate-fir, /emergency/helplines, /emergency/checklists, and /emergency/panic-triage [5]. The separation reflects distinct system capabilities rather than a single generic chatbot endpoint.

IV. RESULTS

A. Functional Verification

The current implementation includes an automated functional test framework covering the principal application paths. The integration tests verify root and health endpoints, user registration and login, profile operations, emergency helplines, emergency checklists, panic triage, chat responses, chat persistence, FIR initiation, FIR continuation, FIR retrieval, and PDF generation [5], [6]. Additional quality-oriented checks cover phishing questions, phishing-victim advice, ransomware explanations, and simple-message fast paths.

B. Security Verification

The security test suite exercises the application controls used around authentication and uploads. The documented checks include rate limiting for repeated authentication attempts, security-header validation, malicious or fake-file rejection, input sanitization, upload validation, and masking of sensitive values [5]. These tests provide implementation-level evidence that the security controls are connected to the running workflow; they do not constitute a full third-party security audit or penetration test.

C. Result Interpretation and Scope

The principal result of the present work is a functioning integrated workflow rather than a single prediction score. The system demonstrates how multi-agent routing, hybrid retrieval, OCR, deterministic field validation, persistence, and document generation can be assembled into one citizen-facing process. Table 2 summarizes the currently verified workflow areas.

Workflow area	Current verification status
Core API and health checks	Verified in integration test suite
Authentication and profile flows	Verified in integration test suite
Emergency assistance routes	Verified in integration test suite
Chat and session persistence	Verified in integration test suite
FIR start / continue / retrieval	Verified in integration test suite
FIR PDF generation	Verified in integration test suite
Phishing / ransomware scenario checks	Covered by quality-oriented tests
Security controls	Covered by automated security tests

No quantitative accuracy, precision, recall, F1, OCR accuracy, retrieval precision/recall, or human-evaluation score is reported because such a controlled benchmark is not present in the supplied project artifacts [5], [6]. Reporting invented values would give a misleading impression of system reliability. A formal evaluation dataset is therefore a necessary next step before making quantitative performance claims.

D. Limitations

- The current project does not provide a statistically controlled benchmark for intent classification, entity extraction, retrieval, OCR, or answer groundedness.
- The supplied project archive contains a manifest of knowledge resources but not all underlying knowledge-base PDFs, so complete ingestion cannot be independently reconstructed from the archive alone.
- The documented session memory is process-local, which limits reliability when multiple workers or restarts are introduced.
- The deployment configuration does not by itself establish a production-scale deployment under load.
- The current authentication configuration includes a fallback JWT secret and should use mandatory external secret management in production [5].
- CyberSahay is an assistance and information-processing system, not a substitute for legal counsel, police authorities, or official legal procedures; human verification remains necessary.

V. CONCLUSION

CyberSahay presents a workflow-oriented architecture for citizen-facing cybercrime assistance in which multiple AI and application components cooperate instead of relying on one unrestricted chatbot prompt. The Coordinator Agent routes requests to specialized workflows, while the Hybrid RAG layer combines BGE dense retrieval, BM25 sparse retrieval, Reciprocal Rank Fusion, and CrossEncoder reranking before generation. OCR and document analysis allow image and PDF evidence to participate in the workflow, while conversational FIR preparation adds structured field collection and validation around generative language processing.

The current implementation has been functionally and security tested across its principal modules. The work deliberately avoids unsupported quantitative claims and instead documents the architecture, implemented controls, and test coverage. This makes the contribution a reproducible system design and implementation for domain-grounded cybercrime assistance rather than a benchmark study.

Future development should introduce a curated evaluation benchmark for Indian cybercrime scenarios, version the knowledge base so that legal and procedural updates can be tracked, replace process-local memory with scalable external storage, and perform expert evaluation of retrieval quality, groundedness, OCR, and FIR field completion. Multilingual evaluation should also be expanded for Indian-language interactions. These additions would strengthen both reproducibility and empirical validation of the system.

VI. ACKNOWLEDGMENT

The authors acknowledge the guidance of Dr. Himanshu V. Taiwade and the support of the Department of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur, during the development and evaluation of the CyberSahay project.

REFERENCES

- [1] N. Pipitone and G. Hourir Alami, "LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain," arXiv preprint arXiv:2408.10343, 2024. [Online]. Available: <https://arxiv.org/abs/2408.10343>.
- [2] W. Yuan, J. Cao, Z. Jiang, Y. Kang, J. Lin, K. Song, T. Lin, P. Yan, C. Sun, and X. Liu, "Can Large Language Models Grasp Legal Theories? Enhance Legal Reasoning with Insights from Multi-Agent Collaboration," in Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 7577-7597, 2024. doi: 10.18653/v1/2024.findings-emnlp.445.
- [3] H. Li, Y. Chen, Y. Hu, Q. Ai, J. Chen, X. Yang, J. Yang, Y. Wu, Z. Liu, and Y. Liu, "LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation," in Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025), 2025. doi: 10.1145/3726302.3730340.
- [4] A. Butler and U. Butler, "Legal RAG Bench: An End-to-End Benchmark for Legal RAG," arXiv preprint arXiv:2603.01710, 2026. [Online]. Available: <https://arxiv.org/abs/2603.01710>.
- [5] CyberSahay Project Documentation, "Implementation Architecture and Backend Analysis," project source documentation supplied by the authors, 2026.
- [6] CyberSahay Project Documentation, "Features, Methodology, Technology Stack, Dataset Manifest, and Testing Analysis," project source documentation supplied by the authors, 2026.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details