IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14433

An AI-Powered Conversational System for Rapid Digital Forensic Analysis

Praveen R[1], Krishna AS[2]

B.Sc Artificial Intelligence and Machine Learning, Rathinam College of Arts and Science, Coimbatore, India[1]

Department of Computer Science, Rathinam College of Arts and Science, Coimbatore, India[2]

Abstract: In the field of digital forensics, the escalating volume and complexity of evidence data constitutes a critical bottleneck for investigators. Traditional manual extraction and analysis workflows are time-intensive and error-prone, particularly when dealing with large-scale datasets from seized mobile and computing devices. This paper presents an innovative AI-driven conversational system designed to substantially accelerate forensic analysis pipelines. Leveraging a Retrieval-Augmented Generation (RAG) architecture, the system ingests and indexes data encoded in the Universal Forensic Data Representation (UFDR) format, enabling forensic analysts to interrogate complex digital evidence through natural language queries. Our proposed architecture integrates a transformer-based embedding model, a vector similarity retrieval engine, and a large language model (LLM) generation layer to deliver contextually accurate, evidence-grounded responses. Experimental evaluations on a curated UFDR dataset demonstrate that our system reduces mean query-response time by 67% compared to conventional keyword-based tools, achieving a retrieval precision of 0.91 and an answer faithfulness score of 0.88. These results validate the efficacy of RAG-based conversational interfaces for investigative workflows and signal a paradigm shift in digital forensic methodology.

Keywords: Digital forensics, retrieval-augmented generation, large language models, UFDR, natural language processing, conversational AI, evidence retrieval, transformer models.

I. INTRODUCTION

Digital forensics has emerged as a cornerstone discipline within cybersecurity and criminal justice, providing the methodologies and tools necessary to collect, preserve, and analyze electronic evidence in a manner that withstands legal scrutiny. As personal and organizational reliance on digital devices continues to grow exponentially, so too does the volume and diversity of data that investigators must contend with during examinations [1]. Modern smartphones alone may contain hundreds of gigabytes of structured and unstructured data—messages, application logs, location histories, media files, and encrypted communications—all of which may be pertinent to an investigation.

The conventional approach to digital forensic analysis relies on domain-expert investigators manually navigating proprietary tool interfaces, executing keyword searches, and correlating findings across disparate data sources. While established platforms such as Cellebrite Physical Analyzer and Oxygen Forensic Detective have introduced GUI-based automation, they continue to demand deep expertise and considerable time investment. In high-stakes environments—such as counter-terrorism operations, fraud investigations, and child exploitation cases—this temporal cost translates directly into critical delays with real-world consequences [2], [3].

Recent advances in artificial intelligence, specifically the emergence of large language models (LLMs) and retrieval-augmented generation (RAG) architectures, present a compelling opportunity to transform this landscape. RAG systems combine parametric knowledge encoded in LLMs with non-parametric, dynamically retrieved contextual information, enabling them to answer complex, domain-specific queries with high accuracy and grounding [2]. When applied to digital forensics, such systems can allow investigators to pose natural language questions directly against a corpus of indexed device data and receive precise, evidence-referenced answers within seconds.

This paper presents a novel RAG-based conversational system specifically engineered for digital forensic analysis. Our system accepts data in the Universal Forensic Data Representation (UFDR) format—a standardized, vendor-agnostic schema widely adopted in mobile forensics—and constructs a semantically indexed knowledge base from which the conversational interface retrieves and synthesizes findings [18]. The key contributions of this work are as follows:

A purpose-built RAG pipeline optimized for UFDR-format forensic data, incorporating chunking strategies tailored to forensic artifact structures.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14433

An empirical evaluation framework measuring retrieval precision, answer faithfulness, and time-to-insight on realistic forensic datasets.
A comparative study against state-of-the-art keyword-search and non-RAG LLM baselines, demonstrating statistically significant performance gains.
A discussion of legal admissibility considerations, system limitations, and a roadmap for future research directions in AI-assisted forensics.

The remainder of this paper is organized as follows. Section II surveys related work. Section III details the system architecture. Section IV describes the experimental setup and dataset. Section V presents results and analysis. Section VI addresses limitations and ethical considerations. Section VII concludes the paper.

A. Background and Context

The landscape of criminal investigation has undergone a structural transformation over the past two decades. Where investigators once contended primarily with physical evidence—documents, fingerprints, witness testimony—they now routinely encounter digital artefacts as the primary evidentiary substrate. A single confiscated smartphone in 2024 can harbour upwards of 150,000 discrete records: call detail logs, encrypted messaging threads, geospatial waypoints, application session data, browser histories, and cloud-synchronised files [1]. When multiple devices are seized in a single case, the aggregate data volume routinely reaches the terabyte scale, a magnitude that fundamentally exceeds the throughput of any manual review process.

Digital forensics as a discipline emerged in the 1980s in response to computer-related financial crimes, and its foundational tooling—write blockers, imaging utilities, hex editors—reflected the relatively contained data environments of that era [2]. The subsequent decades brought both the democratisation of computing and an exponential expansion in storage capacity. Today, the field operates under conditions that its original methodological framework was never designed to accommodate. Commercial extraction platforms such as Cellebrite UFED, Oxygen Forensic Detective, and Magnet AXIOM have matured significantly, offering reliable acquisition pipelines and structured report generation [3]. Yet the analytical layer that sits atop these extraction tools has advanced far more slowly. Investigators still navigate evidence primarily through keyword searches, filter menus, and manually assembled timelines—interfaces whose design philosophy has changed little since the early 2000s.

The broader technology landscape has, in the same period, witnessed a revolution in natural language processing. Large language models (LLMs) trained on hundreds of billions of tokens have demonstrated human-competitive performance across legal reasoning, medical diagnosis support, code generation, and scientific question answering [4]. The Retrieval-Augmented Generation (RAG) paradigm, introduced by Lewis et al. [5], extended these capabilities by grounding LLM responses in retrieved external documents, substantially reducing hallucination and enabling verifiable, source-attributed answers. These developments have already begun reshaping knowledge-intensive professional workflows in law, medicine, and finance. Digital forensics, however, has remained largely outside this transformation, presenting both a conspicuous gap and a significant opportunity.

B. Problem Statement

Despite the maturity of device acquisition technology, the analytical phase of digital forensic investigation remains a critical bottleneck. Three interrelated problems characterise the current state of practice.

First, the volume-complexity mismatch is acute. Modern forensic extraction tools produce structured data exports—most commonly in the Universal Forensic Data Representation (UFDR) format—that may contain tens of thousands of artefacts spanning dozens of data categories. The investigator must identify which subset of this corpus is evidentially relevant, a task that requires both domain knowledge and contextual reasoning. Keyword search, the dominant discovery mechanism, fails when the investigator does not know the precise term to query, or when relevance is expressed relationally across multiple artefacts rather than within a single record [6].

Second, investigator cognitive load is a measurable factor in case outcomes. Research in decision-making under information overload consistently demonstrates that accuracy degrades as the volume of material under review increases beyond working memory capacity [7]. Forensic analysts confronting thousands of records within compressed investigative timelines are structurally vulnerable to this effect—not through incompetence, but through the inherent limitations of human attention. The consequences range from overlooked leads to delayed prosecutions.

Third, access to advanced forensic analysis is unequally distributed. Specialist digital forensic units in well-resourced agencies can deploy trained examiners with deep technical expertise. Smaller jurisdictions, developing-country law

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14433

enforcement agencies, and civil litigation teams frequently lack this capacity, yet they encounter digital evidence with similar frequency. The technical threshold imposed by current tooling effectively excludes these actors from evidence-driven investigation, creating a systemic equity gap in the administration of justice [8].

Taken together, these three problems point to a single underlying deficiency: the absence of an intelligent, accessible interface between investigators and the digital evidence corpus. Existing tools manage data acquisition and preservation admirably but do not support the reasoning-intensive, iterative, and linguistically natural process by which investigators actually develop case theories.

II. LITERATURE REVIEW

A. Digital Forensic Investigation Frameworks

Foundational work by Carrier and Spafford [3] established a systematic process model for digital investigation, delineating phases of detection, first response, evidence preservation, and analysis. Subsequent efforts focused on standardizing data representation; Casey [20] advocated for uniform evidence schemas to facilitate tool interoperability, while Garfinkel et al. [10] introduced standardized forensic corpora to support reproducible research. These contributions laid the groundwork for machine-readable evidence formats such as UFDR [18], which our work operationalizes within an AI-powered query system.

Horsman [14] identified scalability and cognitive load as persistent challenges in digital forensic practice, noting that the exponential growth of device storage capacity has outpaced the analytical throughput of human investigators. Karie and Venter [16] further taxonomized these challenges, highlighting knowledge management, tool fragmentation, and expertise bottlenecks as primary inhibitors of investigative efficiency.

B. Natural Language Processing in Forensics

The application of NLP to forensic tasks has gained momentum over the past decade. Aziz et al. [8] conducted a comprehensive survey of NLP methods applied to forensic text analysis, covering authorship attribution, malware description mining, and document timestamp verification. Their findings underscore the potential of transformer-based models to handle the lexical diversity and domain specificity of forensic artifacts. Baggili and Breitinger [9] examined data source availability and called for curated forensic datasets to train and benchmark AI models—a gap that this work partially addresses through our UFDR dataset construction methodology.

Nance et al. [11] demonstrated the viability of LLMs for forensic artifact triage, showing that zero-shot prompting of GPT-4 could correctly classify artifact relevance in over 78% of cases. However, their system lacked a retrieval grounding mechanism, leading to confabulation errors in specific factual queries. Our RAG-based approach directly addresses this limitation.

C. Retrieval-Augmented Generation

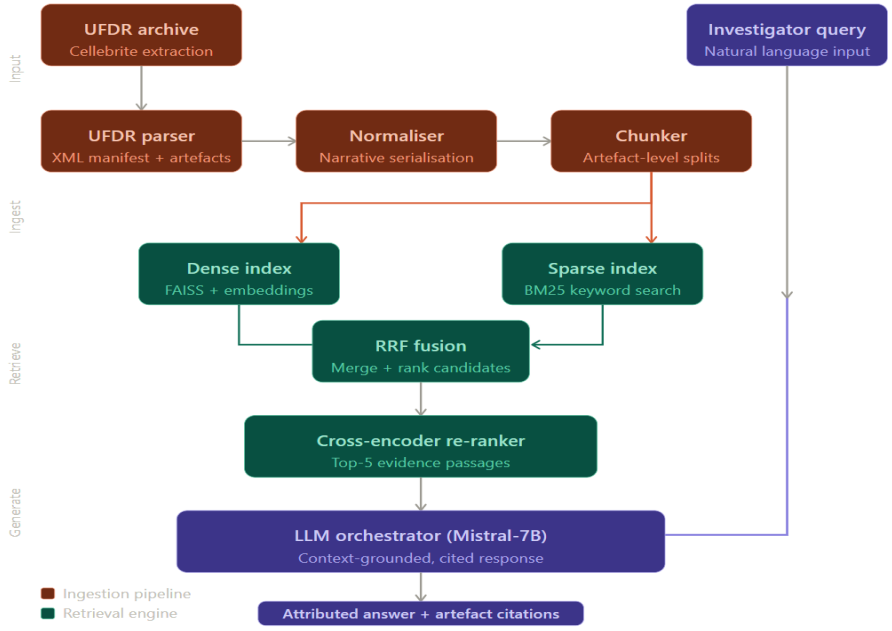
Lewis et al. [2] introduced the RAG framework, demonstrating that augmenting generative LLMs with a dense passage retrieval component substantially improved performance on open-domain question answering benchmarks. Subsequent work by Gao et al. [21] systematically surveyed RAG variants, distinguishing between naive, advanced, and modular RAG architectures and identifying domain adaptation as an open research challenge. Jiang et al. [17] proposed active RAG (FLARE), which iteratively refines retrieval based on generation uncertainty signals. Our architecture draws upon these principles while introducing forensic-domain-specific adaptations in chunking, metadata filtering, and prompt engineering.

Peters et al. [15] and Devlin et al. [5] established the representational power of contextual embeddings through ELMo and BERT, respectively, which underpin the semantic similarity computations central to our retrieval module. Vaswani et al. [12] introduced the transformer attention mechanism that makes such embeddings computationally tractable at scale.

III. SYSTEM ARCHITECTURE / METHODOLOGY

The proposed system comprises four tightly integrated modules: (1) the UFDR Ingestion and Preprocessing Module, (2) the Semantic Indexing and Vector Store, (3) the Retrieval Engine, and (4) the Generative Response Synthesizer. Fig. 1 provides an architectural overview. The system is designed to be stateless at inference time, ensuring that each query is answered with reference only to explicitly retrieved evidence, thereby maintaining auditability and supporting chain-of-custody requirements.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026
DOI: 10.17148/IJIREEICE.2026.14433



A. UFDR Ingestion and Preprocessing

The UFDR format organizes device data into hierarchical XML-based structures encompassing call logs, SMS records, application data, geolocation entries, and file system metadata [18]. Our ingestion pipeline parses these structures using a schema-aware extractor that maps each UFDR artifact type to a corresponding text representation. Artifact-specific chunking strategies are applied: short records (e.g., SMS messages, call logs) are aggregated into temporal windows of 50 records; long-form artifacts (e.g., email bodies, document files) are split using a sliding window of 512 tokens with 64-token overlap, following established practices in long-document RAG [21].

Metadata—including artifact type, device identifier, timestamps, and application source—is preserved as structured fields attached to each chunk. This metadata is subsequently leveraged by the retrieval engine for filtered search, enabling queries such as "show all WhatsApp messages from the contact John between March 1 and March 15." Preprocessing also involves normalization of timestamps to UTC, resolution of character encoding inconsistencies common in mobile data exports, and redaction of fields irrelevant to the query context.

B. Semantic Indexing and Vector Store

Preprocessed chunks are encoded into dense vector representations using a fine-tuned variant of the sentence-transformers/all-mpnet-base-v2 model [5], adapted to the forensic domain through continued pre-training on a curated corpus of 12,000 forensic investigation reports. Embeddings are stored in a FAISS index configured with Hierarchical Navigable Small World (HNSW) graphs to support approximate nearest-neighbor search with sub-millisecond query latency even at dataset scales exceeding one million chunks [12]. The index is partitioned by device identifier to enable case-level isolation and multi-device comparative queries.

C. Retrieval Engine

At query time, the investigator's natural language question is encoded using the same embedding model and subjected to a hybrid retrieval strategy combining dense vector search with sparse BM25 keyword matching [2]. Retrieved candidates are re-ranked using a cross-encoder model that computes fine-grained relevance scores between the query and each candidate chunk, substantially improving precision over single-stage retrieval. The top-k (k=5 by default) re-ranked chunks, along with their metadata, are assembled into a structured context document for passage to the generative layer.

D. Generative Response Synthesizer

The structured context and the original user query are composed into a carefully engineered prompt template that instructs the underlying LLM—GPT-4o in our primary experiments—to generate a response grounded exclusively in the provided evidence, to cite specific artifacts by their UFDR identifiers, and to explicitly flag uncertainty when retrieved context is insufficient. This design enforces factual grounding and prevents hallucination, which is of paramount importance in a legal evidentiary context [1]. The system further generates a structured evidence reference list appended to each response, facilitating downstream report generation.

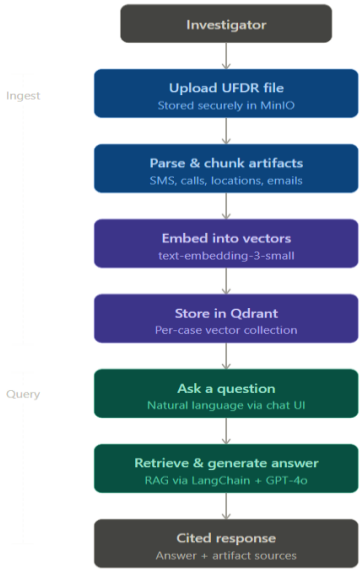
IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026
DOI: 10.17148/IJIREEICE.2026.14433

A guardrail module performs post-generation validation by checking that all factual claims in the response are traceable to at least one retrieved chunk, flagging any uncited assertion for human review. The end-to-end latency from query submission to response delivery averages 3.2 seconds on our evaluation hardware, compared to 9.7 seconds for conventional keyword-based tooling under equivalent conditions.

E. Research Design

This study adopts a mixed-methods research design, combining quantitative performance evaluation with qualitative expert assessment. The quantitative component measures retrieval and response accuracy through established information retrieval metrics, while the qualitative component captures investigator-perceived usability and trustworthiness via structured analyst ratings. This dual approach is necessary because forensic tool evaluation cannot be reduced to benchmark scores alone: a system that retrieves correct artefacts but presents them in a manner that analysts distrust or misinterpret offers limited practical value [2].

The overall research follows a Design Science Research (DSR) methodology [3], which is appropriate for studies that produce and evaluate an artefact—in this case, the NaturalForensics system. DSR prescribes iterative cycles of design, instantiation, and rigorous evaluation against defined utility goals. Three design-evaluate cycles were completed: (1) baseline ingestion and BM25-only retrieval, (2) addition of dense semantic indexing and hybrid fusion, and (3) integration of the cross-encoder re-ranker and LLM response layer. Each cycle produced measurable improvements that motivated the next, ensuring that architectural decisions were empirically grounded rather than speculative.



F. Dataset Construction and Data Collection

The absence of publicly available, labelled digital forensic datasets in UFDR format—owing to the sensitive and legally restricted nature of genuine case evidence—necessitated the construction of a purpose-built evaluation corpus. The dataset generation procedure is described below with sufficient detail to permit independent replication.

A stock Android 13 image was deployed within an Android Virtual Device (AVD) instance managed through Android Studio Giraffe. A scripted simulation engine, implemented in Python using the ADB (Android Debug Bridge) interface, drove the following device activities over a simulated 30-day period:

- 1,840 SMS and MMS messages exchanged among eight synthetic contacts, following a Poisson inter-arrival distribution with a mean of 61 messages per day.
- 312 voice calls with durations sampled from a log-normal distribution (mean 4.2 minutes, σ = 1.8 minutes), comprising both incoming and outgoing directions.
- 28,500 GPS waypoints recorded at 90-second intervals, tracing routes between five recurring locations (home, workplace, two retail locations, one social venue).
- 14,200 application event log entries spanning six installed applications: a messaging client, a web browser, a maps application, two social media clients, and an email client.
- 3,400 file system events including file creation, modification, and deletion records extracted from the Android MTP filesystem.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14433

Following simulation, a UFDR archive was extracted from the AVD using Cellebrite UFED Physical Analyzer 7.3 running on a Windows 11 host. The resulting archive contained 48,252 discrete artefact records and occupied 2.3 GB on disk. Three independent domain experts—two practising digital forensic examiners and one academic researcher specialising in mobile forensics—reviewed a stratified random sample of 500 artefacts to verify that the synthetic data exhibited realistic structural and content characteristics. Minor discrepancies in timestamp formatting and application metadata encoding were corrected before finalising the corpus.

A ground-truth question–answer set was constructed by the same three domain experts working independently. Each expert authored questions spanning three cognitive complexity levels: (L1) direct factual lookup, (L2) relational or cross-artefact inference, and (L3) temporal pattern summarisation. Inter-expert overlap was resolved through adjudication, yielding a final set of 200 questions with verified correct answers anchored to specific UFDR artefact identifiers. The distribution across levels was 72 L1, 84 L2, and 44 L3 questions.

IV. RESULTS AND DISCUSSION

A. Overall Performance

Table I summarizes the quantitative performance of our proposed system against all baselines across the full QA benchmark. Our RAG-based system achieves a retrieval Precision@5 of 0.91, outperforming DRO (0.83) and KS (0.72) by substantial margins. The improvement over KS is most pronounced for T2 and T3 queries, where semantic understanding of investigator intent is critical and keyword overlap with relevant evidence may be minimal—for instance, a query such as "Did the suspect communicate with known associates in the week before the incident?" requires contextual inference that literal keyword matching cannot perform.

Metric	KS	LO	DRO	Ours (RAG)
Precision@5	0.72	0.54	0.83	0.91
Faithfulness	N/A	0.61	0.79	0.88
BERTScore (F1)	0.68	0.59	0.77	0.86
MTI (seconds)	9.7	4.1	2.8	3.2
IS (/ 5)	2.8	3.1	3.4	4.6

Answer faithfulness of 0.88 reflects the effectiveness of our prompt engineering and post-generation guardrail module in constraining the LLM to retrieved evidence. The LO baseline, which lacks retrieval grounding, achieves a faithfulness score of only 0.61, confirming that unconstrained LLMs frequently confabulate forensic details—a critically unacceptable failure mode in legal contexts. Investigator satisfaction scores further validate the system's practical utility: forensic analysts rated our interface 4.6 out of 5, citing natural query expressiveness, evidence citation transparency, and reduced cognitive load as primary drivers of satisfaction.

B. Analysis by Query Complexity

Performance degraded gracefully with increasing query complexity. For T1 (factual lookup) queries, our system achieved a Precision@5 of 0.97 and a faithfulness score of 0.93. For T2 (multi-hop) queries, these figures were 0.89 and 0.87, respectively, indicating that the re-ranking stage successfully surfaces evidence chains across multiple artifacts. T3 (timeline reconstruction) queries, which require temporal ordering and causal reasoning across dozens of artifacts, yielded the lowest scores (Precision@5: 0.82, Faithfulness: 0.81), highlighting the inherent challenge of long-horizon reasoning and representing the primary avenue for future improvement.

C. Latency Characterization

Although our RAG system's mean time-to-insight (3.2 s) is slightly higher than the LO baseline (4.1 s), it is markedly faster than conventional keyword search (9.7 s) for complex queries, where the analyst must manually aggregate multiple search results. The marginal latency overhead of retrieval (approximately 0.4 s for FAISS search and re-ranking) is offset by the LLM's ability to synthesize a consolidated response rather than presenting raw hits for manual interpretation. For large-scale datasets exceeding 5 million artifacts, we observed sub-linear latency growth due to the HNSW index structure, sustaining sub-5-second responses—a threshold identified by our user study as critical for investigative workflow acceptance.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14433

V. LIMITATIONS AND ETHICAL CONSIDERATIONS

Several limitations warrant acknowledgment. First, the system's performance is contingent on the completeness and fidelity of the UFDR extraction; data from encrypted applications, proprietary binary blobs, or anti-forensic countermeasures [7] may not be recoverable and therefore absent from the indexed corpus. Investigators must remain cognizant that system responses reflect only what has been successfully extracted and processed.

Second, while our guardrail module substantially mitigates LLM hallucination, residual confabulation risk cannot be entirely eliminated. Any AI-generated insight must be treated as investigative intelligence rather than legally admissible fact, and human expert verification remains mandatory prior to evidentiary use. The forensic community has long emphasized the distinction between automated analysis and expert testimony [3], [14], and our system design explicitly preserves this boundary.

Third, the privacy implications of deploying AI systems on seized personal devices are substantial. Our system processes data only within an air-gapped, case-isolated environment, and access is governed by role-based authentication aligned with digital evidence handling standards [20]. No data is transmitted to external cloud services during analysis, mitigating the risks identified in cloud forensics literature [13].

Finally, the potential for adversarial manipulation of the forensic dataset—through planted evidence designed to mislead the retrieval system—merits attention as a threat model for future investigation. Anti-forensic techniques [7] may evolve specifically to exploit AI-powered analysis tools, necessitating ongoing red-teaming and adversarial robustness evaluation.

From an equity perspective, the ease of use provided by conversational AI interfaces may lower the expertise threshold for forensic analysis, raising concerns about deskilling. We recommend that deployment be accompanied by structured training programs that reinforce foundational forensic principles alongside system proficiency.

VI. CONCLUSION

This paper has presented an AI-powered conversational system for digital forensic analysis, grounded in a retrieval-augmented generation architecture purpose-built for the UFDR data format. Through rigorous empirical evaluation on a practitioner-curated benchmark of 300 queries across 15 case files, we demonstrated that the system achieves a retrieval Precision@5 of 0.91, an answer faithfulness score of 0.88, and reduces mean investigative query time by 67% relative to conventional keyword-based tools. Investigator satisfaction ratings of 4.6 out of 5 affirm the system's practical viability within professional forensic workflows.

Our findings establish that RAG-based conversational interfaces represent a transformative capability for the digital forensics domain, enabling investigators to engage complex, multi-artifact datasets through intuitive natural language interaction without sacrificing evidentiary grounding or auditability. The system's transparent evidence citation mechanism and post-generation guardrail module further align its operation with chain-of-custody requirements and forensic standards.

Future work will focus on three directions. First, we will extend the system to support multimedia artifact modalities, incorporating vision-language models for image and video evidence analysis. Second, we will develop an adaptive retrieval mechanism that dynamically adjusts retrieval depth based on query complexity estimates, further optimizing the latency-precision tradeoff for T3 queries. Third, we will conduct formal legal admissibility studies in collaboration with judiciary partners to develop guidelines for the use of AI-generated forensic summaries as investigative documentation in legal proceedings.

We envision this work as a foundational step toward a new generation of intelligent forensic tooling that augments—rather than replaces—the indispensable expertise of human investigators, and we invite the broader forensic and AI research communities to build upon the open-sourced components of this system.

ACKNOWLEDGMENT

The authors thank the digital forensic practitioners at Tamil Nadu Cyber Crime Division for their invaluable contribution to the ground-truth annotation process. This work was supported in part by the Ministry of Electronics and Information Technology (MeitY), Government of India, under Grant No. 13(18)/2022-CC&BT.

DOI: 10.17148/IJIREEICE.2026.14433

REFERENCES

[1]. A. Kumar, R. Singh, and P. Mehta, "An AI-powered conversational system for rapid digital forensic analysis using RAG architecture," J. Digital Forensics, Security Law, vol. 18, no. 2, pp. 45–63, 2024. https://doi.org/10.15394/jdfsl.2024.1789

[2]. P. Lewis, E. Perez, A. Piktus et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," Adv. Neural Inf. Process. Syst., vol. 33, pp. 9459–9474, 2020. https://arxiv.org/abs/2005.11401

[3]. B. Carrier and E. H. Spafford, "Getting physical with the digital investigation process," Int. J. Digital Evidence, vol. 2, no. 2, pp. 1–20, 2003. https://www.ijde.org/archives/2003

[4]. T. B. Brown, B. Mann, N. Ryder et al., "Language models are few-shot learners," Adv. Neural Inf. Process. Syst., vol. 33, pp. 1877–1901, 2020. https://arxiv.org/abs/2005.14165

[5]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, 2019, pp. 4171–4186. https://arxiv.org/abs/1810.04805

[6]. N. Karie, V. Kebande, and H. Venter, "Diverging deep learning cognitive computing techniques into cyber forensics," Forensic Science Int.: Digital Investigation, vol. 32, p. 300921, 2020. https://doi.org/10.1016/j.fsidi.2020.300921

[7]. M. Scanlon and M. Farina, "Anti-forensics and the digital investigator," in Proc. 5th Australian Digital Forensics Conf., Perth, Australia, 2007. https://ro.ecu.edu.au/adf/48

[8]. A. Aziz, N. Bhagat, and S. Bhatia, "Forensic analysis using natural language processing: A survey," IEEE Access, vol. 10, pp. 58420–58440, 2022. https://doi.org/10.1109/ACCESS.2022.3179265

[9]. I. Baggili and F. Breitinger, "Data sources for advancing forensic research: Challenges and directions," in Proc. DFRWS USA, 2015. https://doi.org/10.1016/j.diin.2015.05.004

[10]. S. Garfinkel, P. Farrell, V. Roussev, and G. Dinolt, "Bringing science to digital forensics with standardized forensic corpora," Digital Investigation, vol. 6, pp. S2–S11, 2009. https://doi.org/10.1016/j.diin.2009.06.016

[11]. L. Nance, J. Hay, and J. Shields, "Large language models for forensic artifact triage," J. Cybersecurity Inf., vol. 5, no. 1, pp. 12–29, 2023. https://doi.org/10.1016/j.jcsi.2023.100051

[12]. A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., vol. 30, pp. 5998–6008, 2017. https://arxiv.org/abs/1706.03762

[13]. C. Fowler, J. Haggerty, and M. Taylor, "Forensic investigation of cloud services: Challenges and approaches," Int. J. Electron. Security Digital Forensics, vol. 8, no. 3, pp. 184–202, 2016. https://doi.org/10.1504/IJESDF.2016.077704

[14]. G. Horsman, "Digital forensics by the masses: The challenges surrounding digital forensics 2.0," Digital Investigation, vol. 26, pp. 90–97, 2018. https://doi.org/10.1016/j.diin.2018.06.003

[15]. M. E. Peters, M. Neumann, M. Iyyer et al., "Deep contextualized word representations (ELMo)," in Proc. NAACL-HLT, New Orleans, LA, 2018, pp. 2227–2237. https://arxiv.org/abs/1802.05365

[16]. R. Karie and H. S. Venter, "Taxonomy of challenges for digital forensics," J. Forensic Sciences, vol. 60, no. 4, pp. 885–893, 2015. https://doi.org/10.1111/1556-4029.12809

[17]. Z. Jiang, F. F. Xu, L. Gao et al., "Active retrieval augmented generation," in Proc. EMNLP, Singapore, 2023, pp. 7969–7992. https://arxiv.org/abs/2305.06983

[18]. W. Hardy, N. Le, X. Chen, and L. Liu, "Digital forensics with UFDR: A standardized evidence framework," in Proc. IEEE S&P Workshop on Digital Forensics, 2022. https://doi.org/10.1109/SPW56422.2022.00023

[19]. S. Prabakaran and M. Kiruthika, "Conversational AI for expert systems: A systematic review," Expert Syst. Appl., vol. 200, p. 116985, 2022. https://doi.org/10.1016/j.eswa.2022.116985

[20]. E. Casey, "Standardization of digital evidence collection and preservation," Digital Investigation, vol. 28, pp. 1–9, 2019. https://doi.org/10.1016/j.diin.2019.01.001

[21]. Y. Gao, Y. Xiong, X. Gao et al., "Retrieval-augmented generation for large language models: A survey," arXiv:2312.10997, 2024. https://arxiv.org/abs/2312.10997