Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DOLPHIN: A Privacy-Preserving Digital Investigation Readiness Assessment Framework based on Offline Large Language Models and Retrieval-Augmented Generation

Divya K

divya98research@gmail.com

SRM Institute of Science and Technology

Article

Keywords: Digital forensics, forensic readiness, offline Large Language Models, Retrieval-Augmented Generation, evidence-gap taxonomy, investigation readiness

Posted Date: September 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10489777/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DOLPHIN: A Privacy-Preserving Digital Investigation Readiness Assessment Framework based on Offline Large Language Models and Retrieval-Augmented Generation

Divya K[a, *]

Assistant Professor, Department of Computer Applications,

[a] SRM Institute of Science and Technology, Ramapuram, Chennai, 600089, Tamilnadu, India

email-divya98research@gmail.com

Abstract

Increasingly, digital forensic investigations require a quick, reliable, and privacy-preserving assessment of case materials before formal forensic examination. Existing approaches to forensic readiness are, however, infrastructure-based and inadequately support the completeness of evidence, the identification of evidence gaps, and preparation of a structured investigation at the case level. This study introduces a privacy-preserving framework for assessing digital investigation readiness called DOLPHIN, which integrates offline large language models with retrieval-augmented generation to provide evidence-based investigation support. Thirty sets of 30 cases of structured investigation scenarios were created under five evidence-completeness conditions: complete-evidence, partial-evidence, and low-evidence cases. A Forensic Readiness Index was scored for each case across six categories of evidence: victim statement, FIR/CSR, CCTV evidence, witness statement, ownership documents, and metadata. The deficiencies were categorized using a taxonomy of evidence gaps. Evidence gaps were classified into four categories: documentation gaps, corroboration gaps, visual gaps, and time gaps. To ensure that critical investigation data is not processed in the cloud, the framework was deployed locally with the help of LM Studio, AnythingLLM, and Qwen 3. The experimental results indicated that DOLPHIN obtained a readiness classification accuracy of 92.0% and excellent evidence-gap detection performance with F1 scores ranging from 90.27% to 94.60%. The retrieval performance with metadata-filtered RAG was found to be superior, with 90.33% top-1 accuracy, 99.67% top-3 accuracy, and 94.06% mean reciprocal rank. The unsupported output rate went down from 16.33% to 2.00% when using metadata-filtered RAG. DOLPHIN also significantly reduced the mean preparation time, from 15.46 to 4.41 minutes per case, and increased the completion rate for

reports. The results show the potential of offline LLM-RAG systems to aid in investigation-readiness assessment in a human-centered and interpretable way while preserving privacy.

Keywords: Digital forensics; forensic readiness; offline Large Language Models; Retrieval-Augmented Generation; evidence-gap taxonomy; investigation readiness

1. Introduction

A major challenge in digital forensics investigations is the complexity of the investigation, which often contains substantial amounts of diverse digital and documentary evidence in criminal, civil, and administrative cases. Records of complaints, victim statements, CCTV footage, victim information, ownership records, metadata, device records, communication logs, and other supporting documents may need to be considered by investigation teams prior to formal forensic examination. Manual review of evidence becomes tedious and inconsistent as the volume and variety of evidence grow, especially in a resource-poor investigative setting. The workload and amount of data for digital forensic investigators are growing, and automated evidence processing is a critical trend for the future of forensic practice, according to Du et al. (2020). In the same way, Nayerifard et al. (2023) noted that machine learning has received a lot of attention in digital forensics due to the ever-increasing volume and complexity of digital evidence.

While historically digital forensic strategies have been centered on evidence acquisition, preservation, examination, and reporting, many investigations will also need some degree of “preparation” in which the extent to which existing case materials are complete and usable is evaluated. The significance of this stage is that incomplete complaint documentation, missing corroborative evidence, visual evidence, and evidence metadata may hinder the investigation and affect the interpretation of the evidence. Preparing for digital forensics has thus come to be known as a proactive strategy for ensuring that the relevant evidence can be identified, preserved, and effectively used. More recent research on forensic readiness in software-defined networks, database systems, cloud-based systems, and big data infrastructures has confirmed that readiness has become a fundamental requirement for effective and successful investigations (Karie and Valli, 2021; Albugmi, 2024; Akilal and Kechadi, 2025). Many of the current readiness frameworks, however, are infrastructure-based and lack any meaningful mechanism to evaluate a specific investigation file to see if it is complete enough to continue forensic processing.

The potential for using artificial intelligence to ease the strain on forensic analysis, including evidence triage, classification, entity extraction, pattern recognition, and report generation, has been explored. Machine learning and deep learning models have been used in various fields, including image forensics, malware detection, memory forensics, file classification, and network analysis. But there are many AI-based forensic instruments for post-acquisition analyses, not for the early stages of investigation preparation. Further, forensic AI systems must meet the requirements of transparency, reliability, reproducibility, and human verification that are higher than the normal standards due to the fact that they may be used in legal or investigative decisions. To achieve these requirements, there will need to be a shift from black-box automation to interpretable decision-support frameworks that can clearly communicate the role of evidence availability in investigative readiness.

The advent of large language models has opened new frontiers for supporting digital forensics in processing unstructured text, summarizing case materials, extracting information from case materials, answering natural language questions, and creating structured reports. Wickramasekara and Scanlon (2024) proposed that LLMs can enhance the efficiency of digital forensic investigations but also highlighted issues of bias, explainability, infrastructure, and legal and ethical concerns. Similarly, LLMs can potentially revolutionize the digital forensic processes, but they need to be deployed in a way that ensures the handling of hallucination, interpretability, bias, and accountability (Yin et al., 2025). Chernyshev et al. (2026) also identified LLMs as a new research trend within digital forensics but noted that it is important to use them carefully and in a controlled manner.

While multi-purpose LLMs can be powerful tools, they are not well-equipped for sensitive forensic uses. They can make unrefuted claims, use pre-existing information as proof, and leak sensitive information if they are being used through cloud-based APIs. The restrictions are especially significant in investigative situations where case documents can include private, confidential, and legally relevant information. For this reason, recent studies, including ForensicLLM, have focused on the deployment of local LLMs in the digital forensic domain to minimize reliance on cloud infrastructure and enhance the domain's suitability (Sharma et al., 2025). However, local deployment alone cannot fully address the issue of evidence grounding. A forensic LLM must ensure its output is based on the correct case materials and has no unsupported claims.

Retrieval-Augmented Generation (RAG) is a potential solution to this issue because it uses retrieved documents to guide the LLM responses instead of just relying on pre-trained knowledge. Gao et al. (2023) explained that RAG can improve the accuracy, credibility, and traceability of generated outputs by incorporating information from external knowledge sources. Chen et al. (2023) also demonstrated that RAG can reduce hallucination, but there are challenges with noise, information integration, and false context handling. These are critical challenges in the forensic investigation context, where there can be several similar files within a single case, and finding the "wrong" file could lead to the misleading assessment of a case's readiness. Thus, forensic RAG systems must not only be able to perform semantic retrieval but also be able to separate cases, filter for metadata, and verify humanly.

To meet these challenges, this study introduces DOLPHIN, a privacy-preserving framework for assessing digital investigation readiness that leverages offline large language models and retrieval-augmented generation. DOLPHIN is developed to assist with the preparation of investigations prior to formal forensic examination. The framework is composed of four main elements: an offline LLM for local reasoning, a RAG pipeline for retrieving evidence grounded in the document, an Evidence Gap Taxonomy for assessing the lack or incompleteness of evidence at the case level, and a Forensic Readiness Index for providing a quantitative score of case-level evidence. DOLPHIN does not generate an independent forensic opinion but rather acts as a human-centered decision-support system for investigators to help them organize case materials, calculate readiness level, highlight evidence gaps, and create structured readiness reports.

The most significant result of this study is defining the preparation of an investigation as a measurable and privacy-preserving step in the digital forensic lifecycle. First, the study suggests an index of forensic readiness that transforms the availability of the various evidence types into a clear readiness score. Second, it introduces an evidence gap taxonomy that categorizes the gaps in evidence into documentation, corroboration, visual evidence, and timeline gaps. Third, it has a full offline LLM-RAG workflow to ensure confidentiality and minimize reliance on cloud AI services. Fourth, it analyzes metadata-filtered RAG as a way to enhance the reliability of retrievals and minimize outputs not supported by the metadata. Lastly, it shows that, with consumer-grade hardware, AI-based investigation-readiness assessment is feasible and can be applied in educational, investigative, and resource-constrained forensic settings.

The rest of this paper is organized as follows. The second section summarizes the relevant literature on digital forensic readiness, including the role of AI in digital forensics, LLM-assisted investigations, the application of RAG in LLM-assisted investigations, privacy-preserving AI, and evidence-gap assessments. The methodology and experimental setup, such as the construction of the benchmarks, evidence-completeness coding, FRI scoring, RAG configuration, evaluation metrics, and ethical considerations, are described in Section 3. The results of the experiments with respect to readiness classification, detection of evidence gaps, retrieval performance, unsupported-output reduction, efficiency of preparation time, and report completeness are reported in Section 4. Section 5 discusses the implications, limitations, and future directions for research. Section 6 concludes the study.

2. Literature Review

2.1 Digital Forensic Readiness & Investigation Preparedness

The concept of digital forensic readiness has become an important topic in the cybercrime investigation of today as organizations and investigative agencies are now expected to gather, preserve, and organize digital evidence prior to a formal investigation. A traditional approach to digital forensics practice tends to be reactive, starting after an incident has taken place. Recent studies also indicate that there are potential adverse outcomes of the reactive approach, including evidence loss, delayed investigation, incomplete evidence documentation, and higher investigative cost. Forensic readiness thus seeks to assure the availability of the relevant evidence sources, logs, metadata, documentation, and procedural controls in a legally and technically usable format prior to or early in the investigation.

Over the last few years, forensic capabilities have been extended from traditional computing devices to the cloud, databases, software-defined networks, industrial systems, and the IoT environment. The challenges in implementing digital forensic readiness in software-defined networks have been pointed out by Karie and Valli (2021) due to the changes in the new network architectures from the old evidence collection and monitoring methods. In a proposed Digital Forensics Readiness Framework by Albugmi et al. (2024), the authors highlighted the following elements: risk assessment, data classification, monitoring and logging, access control, encryption, backup, recovery, and incident response. Likewise, Akilal and Kechadi (2025) discussed cloud

readiness in digital forensics, an open-source approach to law-enforcement request management designed to facilitate cloud service providers and consumers in the process of meeting cross-border evidence access and law-enforcement requests.

These studies show that forensic readiness extends beyond the technical level of acquiring evidence. This also encompasses organizational readiness, evidence documentation, privacy protection, retrieval, compliance, and decision support. But most of the forensic readiness frameworks are infrastructure-centric and are based on system logs, databases, cloud environments, or network architectures. There has been less focus on investigation readiness for cases, where the presence, or absence, of victim statements, formal complaints, CCTV footage, witness statements, ownership evidence, and metadata would need to be evaluated before more in-depth investigation. This provides a methodological opportunity to develop a structured readiness-scoring model that would be able to rate a case file as ready for investigation.

2.2 Artificial Intelligence in Digital Forensics

With the ever-increasing volume and complexity of digital evidence, artificial intelligence (AI) has become an area of interest for overcoming these challenges. Digital forensic laboratories are frequently burdened with high volumes of data, heterogeneous evidence formats, and high case backlogs and are time-consuming when manually analyzing the evidence. Du et al. (2020) noted that AI evidence processing tools could enhance forensic capacity by categorizing, prioritizing, and analyzing the large amounts of evidence. More recent reviews have supported this argument by highlighting the use of machine learning/deep learning techniques in image forensics, malware analysis, memory forensics, network forensics, file classification, and evidence triage.

In 2023, Nayerifard et al. performed a systematic review of the application of machine learning in digital forensics, concluding that machine learning techniques have been applied in various digital forensic fields to detect and analyze digital evidence. Tageldin et al. (2023) also highlighted that automated evidence processing can enhance the capacity of forensic analysis and case processing, especially through classification, clustering, and triage techniques. AI has also been mentioned as a crucial solution for managing diverse devices, distributed data sources, and vast amounts of evidence extraction in smart and IoT environments.

Despite these advancements, issues of explainability, reliability, reproducibility, bias, and legal admissibility continue to hinder the use of AI in digital forensics. Many AI models are considered "black boxes," which are models that lack explainability. Many AI models are black boxes, which are models that lack explainability. Furthermore, many systems do not address investigation-readiness assessment but rather only classify artifacts or extract evidence. Thus, although AI can speed up the forensic analysis process, there is a need for decision-support frameworks that are transparent, interpretable, and verifiable by humans as part of the early stages of preparing for investigations.

2.3 Large Language Models for Digital Forensic Investigation

Recently, LLMs have become popular in the field of digital forensics because they can understand unstructured text, summarize long documents, identify entities, make reports, and let people ask natural language questions about case materials. Wickramasekara et al. (2025) investigated the use of LLMs to enhance the effectiveness of digital forensic investigations and found opportunities to use LLMs for traceability, workload reduction, and assistance in investigations. Yin et al. (2025) presented a comprehensive systematic review of the application of LLMs in digital forensics, concluding that LLMs have the ability to transform digital forensics processes, helping with evidence analysis, case summarization, and decision support. They also pointed out the hazards of hallucination, bias, interpretability, ethical use, and standardization, however.

There have been recent empirical studies that started to shift from a general discussion to domain-specific implementation. Fozia Sharma et al. (2025) introduced ForensicLLM, a local large language model fine-tuned with question-answer data from forensic literature and curated artifacts. Their study is significant as it directly tackles two important issues that have been raised in the field of forensics. AI: the lack of domain specialization in general-purpose LLMs and the privacy concerns of cloud-based APIs. Other recent studies have suggested the use of a forensic chatbot based on LLMs and NER to examine the system logs, emails, and metadata. The results indicate the potential of LLMs to aid investigators in structuring and analyzing evidence, while emphasizing the necessity of grounding, source attribution, validation, and human oversight in forensic applications.

The focus of the present study is similar but not the same as this young stream of research. DOLPHIN adopts offline LLM reasoning for investigation-readiness assessment, rather than just LLM forensic artifact extraction or general question answering. The framework is based on the completeness of the case materials, the missing evidence, how readiness can be measured, and how readiness reports can be created prior to the formal forensic analysis.

2.4 Retrieval-Augmented Generation for Evidence-Grounded Reasoning

One major challenge of LLMs is that they sometimes produce fluent, but unsupported and factually inaccurate, output. In forensic and legal settings, where decisions rely on the truth and accuracy of statements, this issue can be particularly dire, potentially leading to incorrect conclusions and compromising the integrity of evidence. To mitigate hallucination, the idea of Retrieval-Augmented Generation (RA-Gen) with external documents or domain-specific knowledge bases has been proposed. Gao et al. (2024) introduced RAG and summarized its characteristics. RAG enhances both the accuracy and credibility of the generated results while also providing traceability, as it is based on the LLM's reasoning and retrieved knowledge. Chen et al. (2024) evaluated LLMs for RAG tasks and demonstrated that, although retrieval boosts performance, LLMs can also be susceptible to negative rejection, noisy context, information integration, and false information.

These results have significant implications for digital forensics. Forensic records may contain identical words in varying cases, and misinterpretation of an incorrect record can result in errors. Therefore, it is not advisable to use semantic similarity as the sole criterion for RAG in forensic applications. It also needs to be separated at the case level, filtered by metadata, and carefully validated for context retrieval. There is also recent research focusing on forensic LLMs, which reinforces the significance of retrieval and source attribution since the results of an LLM cannot be separated from evidence records. To realize this principle, DOLPHIN combines offline RAG with metadata-based filtering. The purpose is not just to produce a fluent report but to have the "readiness" scores, evidence gaps, and recommendations supported by the case materials that are uploaded. This approach is significant because it must demonstrate with evidence that the investigation employed traceable reasoning for readiness classification.

2.5 Privacy-Preserving and Offline AI Systems

The privacy issue is one of the major concerns in the field of forensic investigation since case files may contain personally identifiable information, victim information, witness statements, sensitive information in metadata, and evidence that is restricted under the law. There could be security issues related to data transmission, storage, third-party access, and the uncertain nature of the data. Recent forensic-AI research has focused on the deployment of LLM services locally and offline, given these concerns. The use of APIs and powerful infrastructure that reside in the cloud poses a limitation in forensic applicability, especially when dealing with resource-constrained and privacy-protective settings, as pointed out by Sharma et al. (2025).

Advantages in terms of forensic readiness in the offline deployment. It helps keep case material in the investigator's local system instead of on the internet, reduces the need for internet connectivity, and provides the ability for the organization to control the prompts, embeddings, retrieved context, and reports generated. This advantage holds particularly true during the initial stages of investigation preparation, when documents may remain unprocessed or un anonymized. Offline systems, however, need to be functional as well. Small police stations, legal offices, or academic forensic laboratories may limit enterprise-level hardware. To tackle this challenge, the present study demonstrates DOLPHIN on consumer-grade hardware with a local inference platform and a local knowledge/RAG environment. This is a step toward the possibility of providing privacy-preserving aids for investigations using AI without cloud-based services or enterprise computing infrastructure.

2.6 Scoring Models, Checklists, and Evidence-Gap Assessment

There are multiple forensic-ready studies that have suggested frameworks, maturity models, and checklists to measure the preparedness of organizations. These models can be beneficial in determining if an organization has policies, controls, logging systems, incident-response systems, and evidence handling systems. However, these models tend to focus on the organizational or infrastructure level. They typically provide a qualitative assessment of an investigation file rather than a numerical score, indicating its completeness for further investigation. In practice, readiness is dependent on the presence and quality of case-specific evidence in the practical investigation environment. In cases of motorcycle theft, for instance, a victim statement, FIR/CSR record, CCTV footage, witness statement, ownership documentation, and metadata can play a crucial role in establishing a consistent timeline and proving the case. Many of these items may be missing,

and even when the organization has a general forensic-readiness policy, the case may not be investigation-ready.

The findings in Table 1 indicate an improvement in digital forensic readiness and machine learning-based forensic analysis, retrieval-augmented generation, and LLM-assisted investigation support, all of which are recent and ongoing research. Current research primarily entails infrastructure readiness, forensic artifact analysis, general LLM applications, or retrieval-augmented reasoning, however. There is a knowledge gap in creating an offline framework that is privacy-preserving, integrates case-level readiness scoring, evidence-gap taxonomy, metadata-filtered RAG, unsupported-output reduction, and structured readiness report generation. The current work focuses on this aspect by introducing the proposed DOLPHIN framework. This gap supports the need to use a Forensic Readiness Index (FRI) and Evidence Gap Taxonomy (EGAT). A numerical readiness index can transform the availability of heterogeneous evidence into an interpretable score, and a taxonomy can provide an understanding of the specific reasons for lowered readiness. The Evidence Gap Taxonomy is used in this study to categorize missing or incomplete evidence into documentation gap, corroboration gap, visual evidence gap, and timeline gap. It provides a more concrete result than a yes/no answer.

Table 1. Summary of recent literature and research gap related to DOLPHIN

Author/year	Research focus	Main contribution	Limitation / gap	Relevance to DOLPHIN
Du et al. (2020)	AI in digital forensic investigation	Reviewed how AI can support digital forensic analysis, triage, and evidence processing.	Focused mainly on AI-assisted forensic analysis rather than case-level readiness scoring.	Supports the need for AI-assisted investigation preparation.

Karie and Valli (2021)	Digital forensic readiness in SDN	Discussed challenges and countermeasures for forensic readiness in software-defined networking.	Infrastructure-oriented; does not provide LLM-based case-readiness assessment.	Supports the importance of proactive forensic readiness.
Li et al. (2022)	Retrieval-augmented text generation	Reviewed retrieval-augmented generation methods for improving generation quality.	Not forensic-specific and does not address investigation evidence gaps.	Supports the use of RAG for evidence-grounded generation.
Nayerifard et al. (2023)	Machine learning in digital forensics	Systematically reviewed ML applications across digital forensic domains.	Focuses on ML-based detection and analysis, not readiness scoring or offline LLMs.	Justifies the wider role of AI in digital forensic workflows.
Gao et al. (2023/2024)	RAG for LLMs	Explained how RAG reduces hallucination and improves credibility by grounding LLM output in retrieved knowledge.	General RAG survey; not designed for forensic case preparation.	Supports DOLPHIN's metadata-filtered RAG design.
Albugmi (2024)	Database forensic readiness	Proposed a digital forensic readiness framework for	System-level readiness; lacks case-level evidence scoring	Shows need for structured readiness frameworks.

		securing database systems.	and LLM-based reporting.	
Wickramasekara et al. (/2025)	LLMs for digital forensic efficiency	Reviewed the potential of LLMs to improve investigation efficiency, traceability, and workload management.	Highlights challenges such as explainability, legal issues, and resource demands.	Supports the use of constrained offline LLMs in DOLPHIN.
Sharma et al. (2025)	Local forensic LLM	Introduced ForensicLLM, a local forensic LLM designed to reduce reliance on cloud APIs and address hallucination concerns.	Focuses on forensic Q&A/domain adaptation rather than FRI scoring and evidence-gap taxonomy.	Strongly supports the offline and privacy-preserving direction of DOLPHIN.
Yin et al. (2025)	LLMs in digital forensics	Discussed opportunities and limitations of LLMs in digital forensic processes, including hallucination, interpretability, bias, and ethics.	Does not propose a measurable readiness index or experimental evidence-gap framework.	Supports the need for human-verified LLM decision support.

Akilal and Kechadi (2025)	Cloud digital forensic readiness	Proposed an open-source approach to managing law-enforcement requests in cloud forensic readiness.	Cloud-focused and not aimed at local offline investigation preparation.	Supports privacy and readiness concerns in modern forensic environments.
Chernyshev et al. (2026)	LLMs in digital forensics	Provided a recent survey on the intersection of generative AI and digital forensics.	Highlights emerging opportunities but does not operationalize readiness scoring.	Positions DOLPHIN within the newest LLM-forensics research direction.

2.7 Research Gap and Positioning of the Present Study

The literature review reveals four main trends. First, there is an emerging recognition that preparing for digital evidence is vital in the interests of proactive evidence preservation and efficient investigation. Second, AI and machine learning are used to minimize forensic workloads and evidence triage. Third, LLMs are becoming promising technologies for forensic summarization, answering questions, extracting entities, and report generation. Fourth, the prospects for deploying RAG and local LLM for enhanced grounding, traceability, and privacy are being explored.

There are still significant unmet needs, however. Current forensic-ready frameworks are mostly based on organizational readiness or cloud, databases, IoT, or network infrastructure. The bulk of the research on AI and the forensics field is related to artifact extraction, classification, triage, or analysis once the evidence has been collected. Although recent LLM-based forensic studies show the potential of LLMs, most LLMs do not offer a structured readiness score, evidence-gap taxonomy, offline RAG workflow, and a quantitative assessment of unsupported forensic outputs, retrieval accuracy, time efficiency, and report completeness within a single structure.

To meet the aforementioned gaps, the present study introduces DOLPHIN, a framework for investigating readiness assessment in a privacy-preserving manner with offline large language models and retrieval-augmented generation. The framework standardizes investigation preparation into a measurable process with the Forensic Readiness Index; defines missing evidence with the Evidence Gap Taxonomy; formalizes outputs with local RAG; and has measures for performance evaluation with readiness classification, evidence-gap detection, evidence retrieval measures, unsupported-output reduction, preparation time, and report completeness measures. The study aims to shift the role of DOLPHIN from an independent forensic decision-making tool to a human-centered decision-support system for preparing digital investigations at an early stage.

3. Methodology and Experimental Setup

3.1 Research Design

This study adopted a controlled experimental design to evaluate the proposed DOLPHIN framework for privacy-preserving digital investigation readiness assessment. The evaluation examined whether an offline Large Language Model supported by Retrieval-Augmented Generation could assist in evidence organization, case-level readiness scoring, evidence-gap identification, retrieval accuracy, unsupported-output reduction, and structured readiness-report generation.

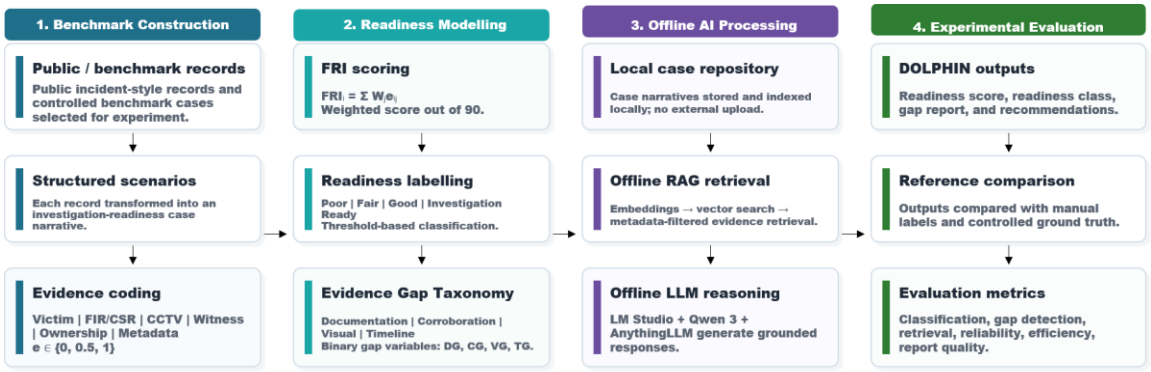


Figure 1. Methodological workflow for offline investigation-readiness assessment using DOLPHIN

The focus of the experiment was investigation preparation rather than forensic examination. Therefore, DOLPHIN was evaluated as a decision-support framework intended to assess whether

available case materials were sufficiently complete and organized before formal forensic analysis. Figure 1 illustrates the complete experimental methodology followed in the study. Public incident records and controlled benchmark cases were first transformed into structured investigation scenarios. Each scenario was coded using six evidence-completeness indicators and evaluated using the Forensic Readiness Index. Evidence deficiencies were then labelled using the Evidence Gap Taxonomy. The structured case narratives were indexed in a local RAG environment and processed by an offline LLM through LM Studio and AnythingLLM. The generated outputs were compared with reference labels using readiness classification accuracy, evidence-gap detection metrics, retrieval performance, unsupported-output rate, preparation time, and report-completeness score.

3.2 Benchmark Dataset Construction

Base public records were downloaded from publicly available vehicle-theft public records on the Kaggle datasets, such as Motor Vehicle Thefts and Stolen Vehicle. These datasets were set to provide public incident style data on vehicle theft that was converted to structured investigation readiness scenarios. Such public datasets do not have full forensic case files, so controlled evidence-completeness variants were created to mimic full, partial, and low evidence investigations. A controlled benchmark of 300 structured investigation scenarios was developed for evaluation. The benchmark consisted of five evidence-completeness groups: base public incident records, complete-evidence scenarios, missing witness/metadata scenarios, missing CCTV/ownership scenarios, and low-evidence scenarios. Each group contained 60 cases.

Because publicly available crime datasets do not usually contain complete forensic case files such as FIR/CSR records, witness statements, CCTV documentation, ownership documents, and metadata, each public incident record was transformed into a structured investigation-readiness scenario. Controlled evidence-completeness conditions were then applied to simulate complete, partial, and weakly documented investigations.

3.3 Evidence-Completeness Coding

Each investigation case was represented using six evidence categories: victim statement, FIR/CSR, CCTV evidence, witness statement, ownership documentation, and metadata. For a

given case (i), the evidence vector was defined as (1):

$$E_i = [v_i, f_i, c_i, w_i, o_i, m_i] \quad (1)$$

where v_i, f_i, c_i, w_i, o_i, and m_i represent the availability of victim statement, FIR/CSR, CCTV evidence, witness statement, ownership documentation, and metadata, respectively.

Each evidence component was coded using a three-level completeness scale (2):

$$e_{ij} = \begin{cases} 1, & \text{if the evidence is available} \\ 0.5, & \text{if the evidence is partially available} \\ 0, & \text{if the evidence is missing} \end{cases} \quad (2)$$

This coding scheme allowed partially available evidence to be represented more realistically than a simple binary present/absent model.

3.4 Forensic Readiness Index Scoring

The Forensic Readiness Index was used to quantify investigation preparedness at the case level. Each evidence category was assigned a predefined weight according to its relevance to preliminary investigation readiness. The FRI for case i was calculated as (3):

$$FRI_i = \sum_{j=1}^{n} W_j \, e_{ij} \quad (3)$$

where W_j represents the weight assigned to the j^{th} evidence category and e_{ij}) represents the evidence-completeness value for case (i). In the proposed framework, the FRI was operationalized as (4):

$$F \qquad RI_i = 15v_i + 20f_i + 15c_i + 15w_i + 15o_i + 10m_i \quad (4)$$

The maximum possible score was 90. Based on the resulting FRI score, each investigation case was classified into one of four readiness levels (5):

$$R_i = \begin{cases} \text{Poor}, & 0 \le FRI_i \le 50 \\ \text{Fair}, & 51 \le FRI_i \le 70 \\ \text{Good}, & 71 \le FRI_i \le 85 \\ \text{Investigation Ready}, & 86 \le FRI_i \le 90 \end{cases} \quad (5)$$

3.5 Evidence Gap Taxonomy Labelling

The Evidence Gap Taxonomy was used to classify the specific type of deficiency affecting each investigation scenario. Four binary gap variables were defined: Documentation Gap DG_i, Corroboration Gap CG_i, Visual Evidence Gap VG_i, and Timeline Gap TG_i.

A Documentation Gap was assigned (6) when FIR/CSR or ownership documentation was incomplete:

$$DG_i = \begin{cases} 1, & \text{if } f_i < 1 \text{ or } o_i < 1 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

A Corroboration Gap was assigned (7) when witness evidence was missing or partial:

$$CG_i = \begin{cases} 1, & \text{if } w_i < 1 \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

A Visual Evidence Gap was assigned (8) when CCTV evidence was missing or partial:

$$VG_i = \begin{cases} 1, & \text{if } c_i < 1 \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

A Timeline Gap was assigned (9) when metadata was missing or partial:

$$TG_i = \begin{cases} 1, & \text{if } m_i < 1 \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

This taxonomy enabled the framework to explain not only the readiness level of a case but also the specific evidence deficiency responsible for reduced readiness.

3.6 Offline LLM and RAG Configuration

DOLPHIN was configured as a fully offline investigation-support workflow to ensure local processing, privacy preservation, and independence from cloud-based AI services. The local language model environment was implemented using LM Studio, while Retrieval-Augmented Generation was managed through AnythingLLM. The offline language model was connected to a local RAG workspace so that case-specific documents could be retrieved and supplied as contextual evidence before response generation.

The experimental implementation was deployed on consumer-grade hardware to evaluate the practical feasibility of running AI-assisted investigation-readiness assessment without enterprise-level infrastructure. The hardware and software configuration used in the experiment is presented in Table 2.

Table 2. Experimental hardware and software configuration

Component	Specification
Laptop	Acer Nitro V
Processor	AMD Ryzen 7000 Series
GPU	NVIDIA RTX 4050, 6 GB VRAM
RAM	16 GB
Operating system	Windows 11
Inference platform	LM Studio v0.4.15
Knowledge/RAG platform	AnythingLLM v1.13.0
Language model	Qwen 3

For each user query q_i, the retrieval system searched the locally indexed case documents and ranked them according to relevance. The retrieved context was then provided to the offline LLM to support evidence-grounded reasoning, FRI scoring, evidence-gap identification, and readiness-report generation. Retrieval performance was evaluated using Top-(k) accuracy, which measures whether the correct case document appears within the top (k) retrieved results as (10):

$$Top\text{-}k = \frac{1}{N}\sum_{i=1}^{N} I\,(r_i \leq k) \tag{10}$$

where N represents the total number of queries,r_i represents the rank position of the correct case document for query q_i, and $I(\cdot)$ is an indicator function that returns 1 when the correct document appears within the top (k) retrieved results and 0 otherwise.

Mean Reciprocal Rank was also calculated to evaluate the average ranking quality of the retrieval system using (11):

$$MRR = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{r_i} \tag{11}$$

where r_i denotes the rank position of the first relevant document for the i^{th} query. These metrics were used to compare keyword search, semantic RAG, and metadata-filtered RAG

configurations. The offline configuration ensured that all case narratives, embeddings, prompts, retrieved context, and generated outputs remained within the local system during the experimental workflow.

3.7 Experimental Procedure

The experimental procedure was conducted in sequential stages to ensure consistency across all investigation scenarios. First, each benchmark case was converted into a structured case narrative containing the available, partially available, and missing evidence categories. These narratives included victim statement, FIR/CSR status, CCTV availability, witness statement, ownership documentation, and metadata information.

Second, the structured case narratives were uploaded and indexed within the local RAG workspace. During indexing, case documents were organized using case-level identifiers to reduce the risk of cross-case retrieval errors. Third, a fixed prompt template was used across all scenarios to instruct the offline LLM to identify available evidence, calculate the Forensic Readiness Index, assign a readiness category, detect evidence gaps, and generate a structured readiness report.

Fourth, DOLPHIN-generated outputs were compared with manually verified reference labels. The comparison included readiness classification, evidence-gap detection, retrieval accuracy, unsupported-output rate, preparation time, and report completeness. Finally, the results were aggregated across the five evidence-completeness groups to evaluate the performance of the proposed framework under complete, partial, and low-evidence investigation conditions.

This procedure ensured that all cases were processed using the same experimental workflow and that the performance of DOLPHIN was evaluated under controlled and repeatable conditions.

3.8 Evaluation Metrics

The performance of DOLPHIN was evaluated using multiple quantitative metrics covering readiness classification, evidence-gap detection, retrieval performance, unsupported-output reduction, processing efficiency, and report quality. These metrics were selected to assess both the technical reliability and practical usefulness of the proposed framework.

Readiness classification performance was evaluated by comparing DOLPHIN-generated readiness labels with manually verified reference labels. Overall classification accuracy was calculated as (12):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

where (TP), (TN), (FP), and (FN) represent true positives, true negatives, false positives, and false negatives, respectively.

Evidence-gap detection was evaluated separately for each Evidence Gap Taxonomy category, including documentation, corroboration, visual evidence, and timeline gaps. Precision, recall, and F1-score were calculated as follows using (13) (14) (15):

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

$$Recall = \frac{TP}{TP+FN} \quad (14)$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (15)$$

Precision measured the proportion of detected evidence gaps that were correct, while recall measured the proportion of actual evidence gaps successfully detected by the framework. The F1-score represented the harmonic mean of precision and recall.

Retrieval performance was evaluated using Top-(k) accuracy and Mean Reciprocal Rank. Top-(k) accuracy measured whether the correct case document appeared within the top (k) retrieved results as (16):

$$Top\text{-}k = \frac{1}{N}\sum_{i=1}^{N} I\,(r_i \le k) \quad (16)$$

where (N) represents the total number of queries, r_i represents the rank position of the correct case document for query q_i, and $I(\cdot)$ is an indicator function.

Mean Reciprocal Rank (17) was used to measure the average ranking quality of the retrieval system:

$$MRR = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{r_i} \quad (17)$$

where r_i denotes the rank position of the first relevant document for the i^{th} query.

Unsupported-output rate was used to evaluate the reliability of generated responses. An output was considered unsupported when it included a claim that could not be verified from the retrieved case context or structured evidence record. The unsupported-output rate was calculated as (18):

$$Unsupported\ Output\ Rate = \frac{N_{unsupported}}{N_{total}} \times 100 \quad (18)$$

where $N_{unsupported}$ represents the number of unsupported generated outputs and N_{total} represents the total number of generated outputs.

Processing efficiency was assessed using preparation-time reduction. The percentage reduction in case preparation time was calculated as (19):

$$Time\ Reduction = \frac{T_{manual} - T_{DOLPHIN}}{T_{manual}} \times 100 \quad (19)$$

where T_{manual} represents the mean time required for manual checklist review and $T_{DOLPHIN}$ represents the mean time required for DOLPHIN-assisted review.

Report quality was evaluated using a five-point report-completeness score. Each generated report was assessed for the presence of five required components: evidence summary, missing evidence identification, FRI score, Evidence Gap Taxonomy classification, and recommended investigative actions. The report-completeness score was calculated as (20):

$$RC_i = S_i + M_i + F_i + G_i + A_i \quad (20)$$

where S_i represents inclusion of an evidence summary, M_i represents missing evidence identification, F_i represents FRI score reporting, G_i represents Evidence Gap Taxonomy classification, and A_i represents recommended investigative actions. Each component was assigned a value of 1 if present and 0 if absent, producing a maximum report-completeness score of 5. Together, these metrics enabled a comprehensive evaluation of DOLPHIN in terms of classification reliability, evidence-gap identification, retrieval quality, hallucination control, time efficiency, and structured reporting performance.

3.9 Ethical and Privacy Considerations

The present study was conducted using a controlled investigation-readiness benchmark and did not involve identifiable real police case files, personal victim records, or legally active investigation data. Public incident-style records were transformed into structured benchmark scenarios, and controlled evidence-completeness variants were created only for experimental evaluation. Therefore, the benchmark was used to assess investigation-readiness modelling rather than to make legal, criminal, or forensic conclusions about real individuals. To preserve privacy, the entire DOLPHIN workflow was configured in a local offline environment. Case narratives, evidence indicators, prompts, retrieved context, embeddings, model responses, FRI scores, and generated readiness reports were processed within the local system. No data were transmitted to external cloud-based LLM services or third-party APIs during the experimental workflow.

The framework was designed as a human-centered decision-support system. Its purpose is to assist investigators in organizing case materials, identifying missing evidence, estimating readiness level, and generating preliminary readiness reports. The system does not replace forensic experts, investigating officers, legal authorities, or judicial decision-making. All AI-generated outputs require human verification before any operational, forensic, or legal use. Responsible use of the framework also requires that unsupported or uncertain outputs be reviewed carefully. Although RAG and metadata filtering were used to reduce unsupported generation, the system may still produce incomplete or incorrect interpretations when evidence records are ambiguous, inconsistent, or insufficient. Therefore, DOLPHIN should be used as an assistive screening tool for investigation preparation, not as an autonomous decision-making system.

4. Results and Analysis

4.1 Experimental Evaluation

A controlled investigation-readiness benchmark was used to measure and evaluate the capability of DOLPHIN to quantify case-level preparedness, detect missing evidence, search for relevant case information, filter out unsupported LLM outputs, and produce structured case-preparedness reports. The six key dimensions assessed were forensic readiness index (FRI) scoring, agreement between the evaluators on readiness classification, evidence gap detection, retrieval performance of RAG, contribution of the components by ablation, reduction of unsupported output, efficiency in preparation time, and report completeness quality.

The benchmark was structured into evidence-completeness groups based on the different investigation scenarios used, such as a base public incident record, a full evidence case, witnesses missing evidence, metadata missing evidence, CCTV missing evidence, evidence missing ownership documentation or evidence missing case ownership documentation, and evidence missing case record evidence. The design was created so that the framework could be tested with different degrees of investigation readiness and not just with fully documented investigations.

4.2 FRI-Based Investigation Readiness Scoring

Figure 2 presents the distribution of Forensic Readiness Index scores across the evidence-completeness groups. The complete-evidence group achieved the highest FRI scores, with most cases concentrated near the upper end of the 0–90 scale. In contrast, the low-evidence group demonstrated the lowest readiness distribution, confirming that the FRI model is sensitive to missing or incomplete evidence.

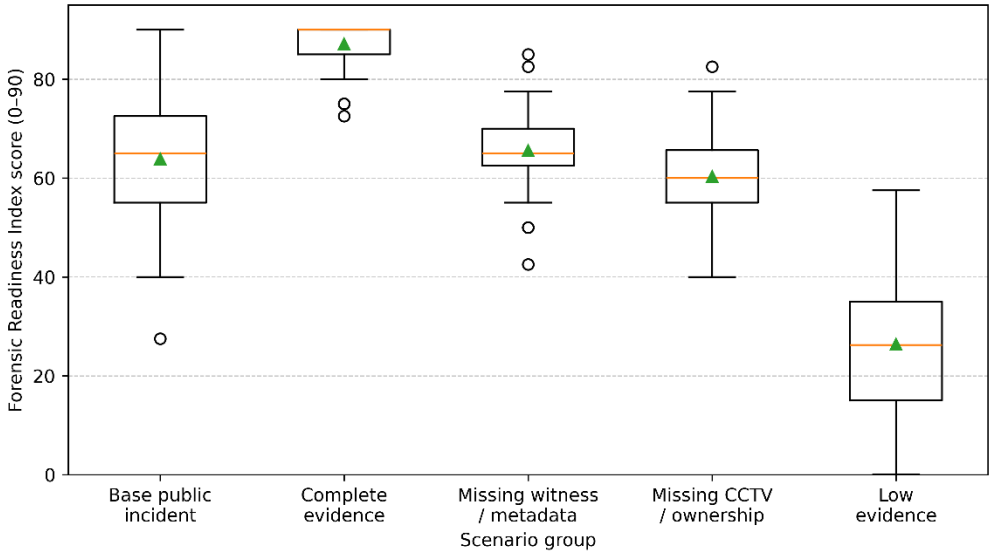


Figure 2. Distribution of Forensic Readiness Index scores across evidence-completeness groups

The base public-incident group had a greater variation, indicating that it is well known that the level of documentation of public crime records varies widely. Cases that lacked witness statements or metadata had moderate FRI values, while cases that lacked CCTV or ownership evidence had reduced readiness scores based on the weight that was given to the missing evidence category. In

sum, the results of the FRI distribution suggest that the proposed scoring model has the potential to measure and interpret investigations that are completed, partially completed, and poorly prepared. Table 3 shows the descriptive statistics of the FRI scores for the five groups of evidence completeness.

Table 3. Descriptive statistics of FRI scores across evidence-completeness groups

Scenario group	n	Mean ± SD	Median	IQR	Min–Max
Base public incident	60	63.83 ± 13.32	65.00	55.00–72.50	27.50–90.00
Complete evidence	60	87.08 ± 4.72	90.00	85.00–90.00	72.50–90.00
Missing witness/metadata	60	65.54 ± 7.84	65.00	62.50–70.00	42.50–85.00
Missing CCTV/ownership	60	60.29 ± 9.43	60.00	55.00–65.63	40.00–82.50
Low evidence	60	26.38 ± 14.81	26.25	15.00–35.00	0.00–57.50

4.3 Readiness Classification Agreement

Figure 3 shows the confusion matrix comparing expert/manual readiness labels with DOLPHIN-generated readiness classifications. The results demonstrate strong agreement between the reference labels and framework outputs, with most classifications concentrated along the diagonal of the matrix. The overall classification accuracy was approximately 92.0%, indicating that DOLPHIN was able to reproduce manual readiness judgments with high consistency.

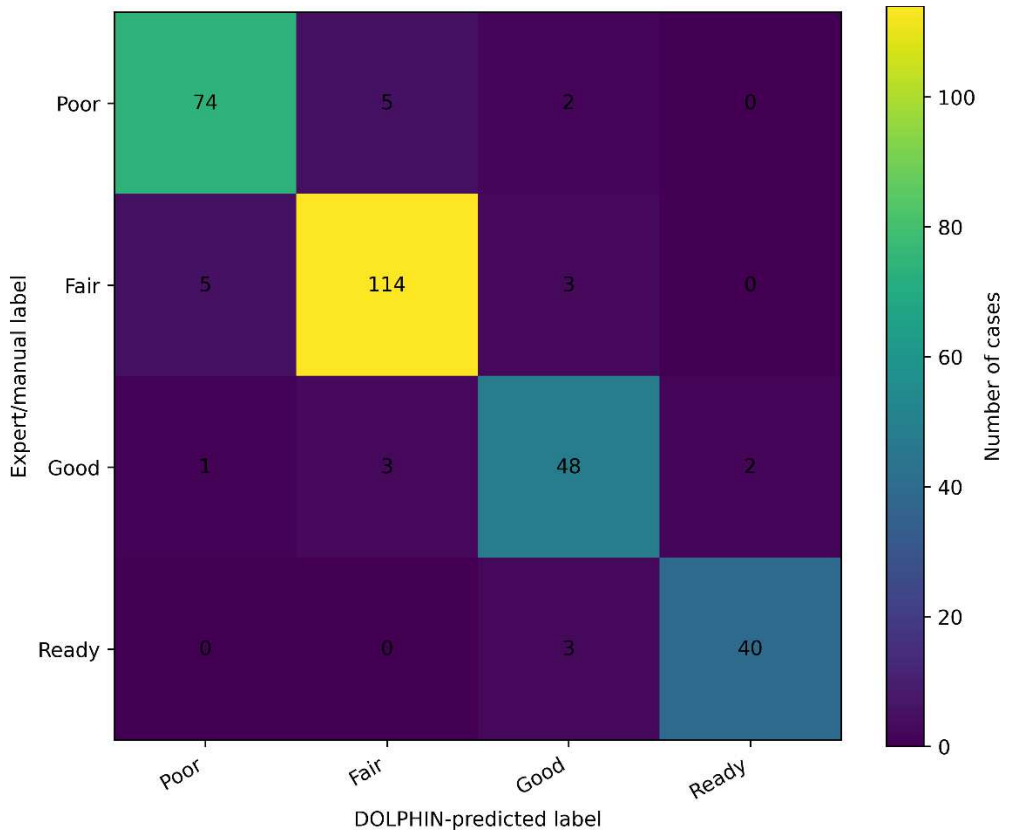


Figure 3. Readiness classification agreement between expert/manual labels and DOLPHIN predictions

The readiness classification performance of DOLPHIN against expert/manual reference labels is presented in Table 4. Most classification errors occurred between adjacent readiness categories, such as Poor and Fair or Fair and Good. This is expected because borderline cases may contain partial evidence, making the distinction between neighboring readiness levels less rigid. Importantly, the framework rarely misclassified low-readiness cases as fully ready or high-readiness cases as poor. This supports the use of DOLPHIN as a decision-support tool for preliminary investigation preparation while preserving the need for final investigator verification.

Table 4. Readiness classification performance

Readiness class	Support	Precision (%)	Recall (%)	F1-score (%)
Poor	81	92.50	91.36	91.93
Fair	122	93.44	93.44	93.44
Good	54	85.71	88.89	87.27
Investigation Ready	43	95.24	93.02	94.12
Macro average	300	91.72	91.68	91.69
Overall accuracy	300	—	—	92.00

4.4 Evidence-Gap Detection Performance

Figure 4 presents the performance of the Evidence Gap Taxonomy across four gap categories: documentation, corroboration, visual evidence, and timeline gaps. The framework achieved consistently high precision, recall, and F1-scores across all categories. The F1-scores were approximately 94.6% for documentation gaps, 90.3% for corroboration gaps, 91.2% for visual evidence gaps, and 92.5% for timeline gaps.

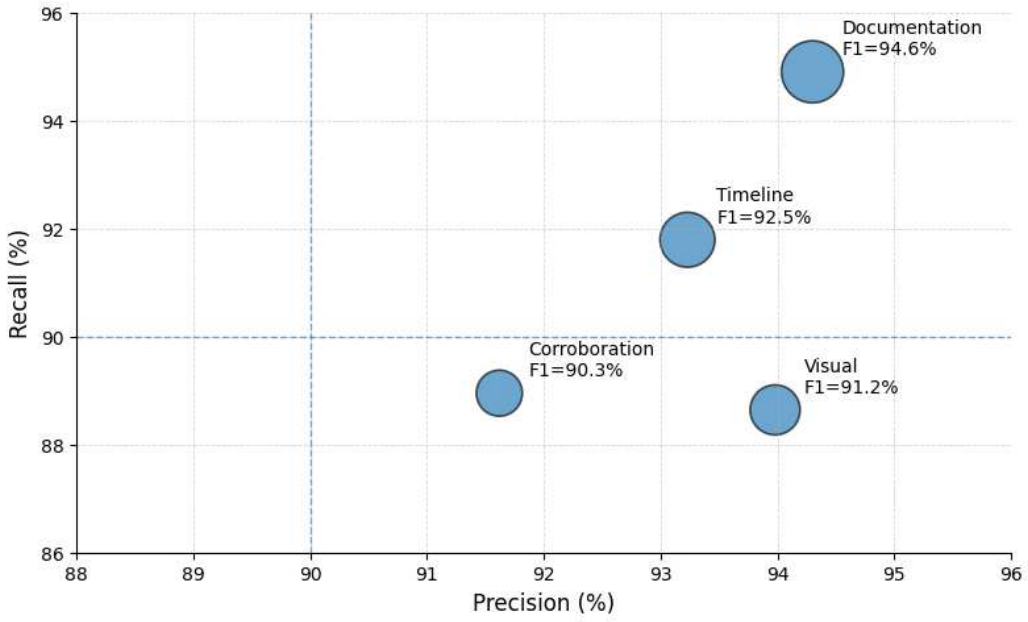


Figure 4. Evidence-gap detection performance across Evidence Gap Taxonomy categories

These results indicate that the proposed taxonomy can reliably identify the specific type of evidence deficiency affecting investigation readiness. The evidence-gap detection performance across the four Evidence Gap Taxonomy categories is presented in Table 5. Documentation gaps were detected with the strongest performance because missing FIR/CSR records or ownership documentation are relatively explicit indicators. Corroboration and timeline gaps were slightly more challenging because witness and metadata evidence may be partially available rather than completely absent. Nevertheless, the consistently high F1-scores demonstrate that DOLPHIN can convert evidence availability assessment into actionable gap identification.

Table 5. Evidence-gap detection performance

Evidence-gap category	TP	FP	FN	TN	Precision (%)	Recall (%)	F1-score (%)
Documentation gap	149	9	8	134	94.30	94.90	94.60
Corroboration gap	153	14	19	114	91.62	88.95	90.27
Visual evidence gap	156	10	20	114	93.98	88.64	91.23
Timeline gap	179	13	16	92	93.23	91.79	92.51

4.5 RAG Retrieval Performance

Figure 5 compares retrieval performance across keyword search, semantic RAG, and metadata-filtered RAG. Keyword search produced the weakest Top-1 retrieval performance because it depends heavily on exact term overlap. Semantic RAG improved retrieval by identifying conceptually similar evidence records, while metadata-filtered RAG achieved the strongest performance across Top-1 accuracy, Top-3 accuracy, and mean reciprocal rank.

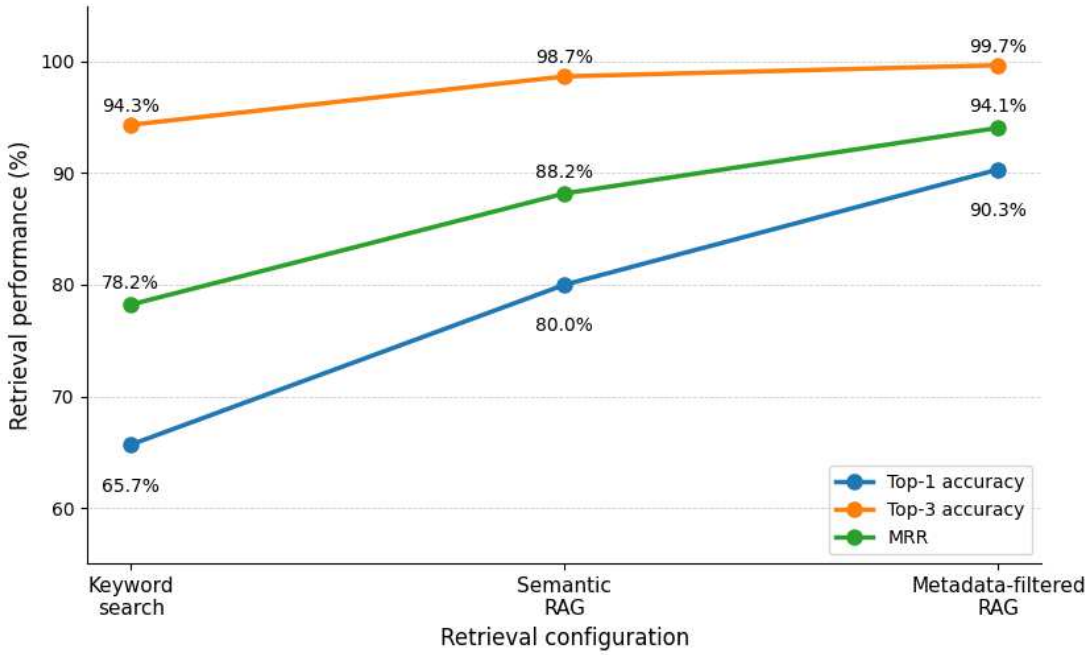


Figure 5. Retrieval performance across keyword search, semantic RAG, and metadata-filtered RAG settings

The improvement obtained through metadata-filtered RAG is especially important in forensic settings. Investigation records may contain similar terminology across multiple cases, and retrieving a semantically similar but incorrect case can compromise readiness assessment. The retrieval accuracy and reliability performance of keyword search, semantic RAG, metadata-filtered RAG, and LLM-only configurations are presented in Table 6. The results therefore show that case-level identifiers and metadata-aware filtering are essential for reliable forensic RAG implementation. This finding supports the architectural decision to integrate retrieval grounding with structured case separation in DOLPHIN.

Table 6. RAG Retrieval and Reliability Performance

Configuration	Top-1 accuracy (%)	Top-3 accuracy (%)	MRR (%)	Unsupported output rate (%)
Keyword search	65.67	94.33	78.22	—
Semantic RAG	80.00	98.67	88.18	6.33

Metadata-filtered RAG	90.33	99.67	94.06	2.00
LLM only	—	—	—	16.33

4.6 Ablation Analysis of DOLPHIN Components

Figure 6 presents an ablation analysis comparing LLM-only reasoning, LLM with RAG, LLM with RAG and metadata filtering, the full DOLPHIN framework, and manual checklist review. The LLM-only configuration produced the weakest performance because it lacked evidence grounding, structured scoring, and taxonomy-based gap analysis. Adding RAG improved performance by grounding generated responses in retrieved case evidence. The addition of metadata filtering further improved consistency by reducing cross-case retrieval errors.

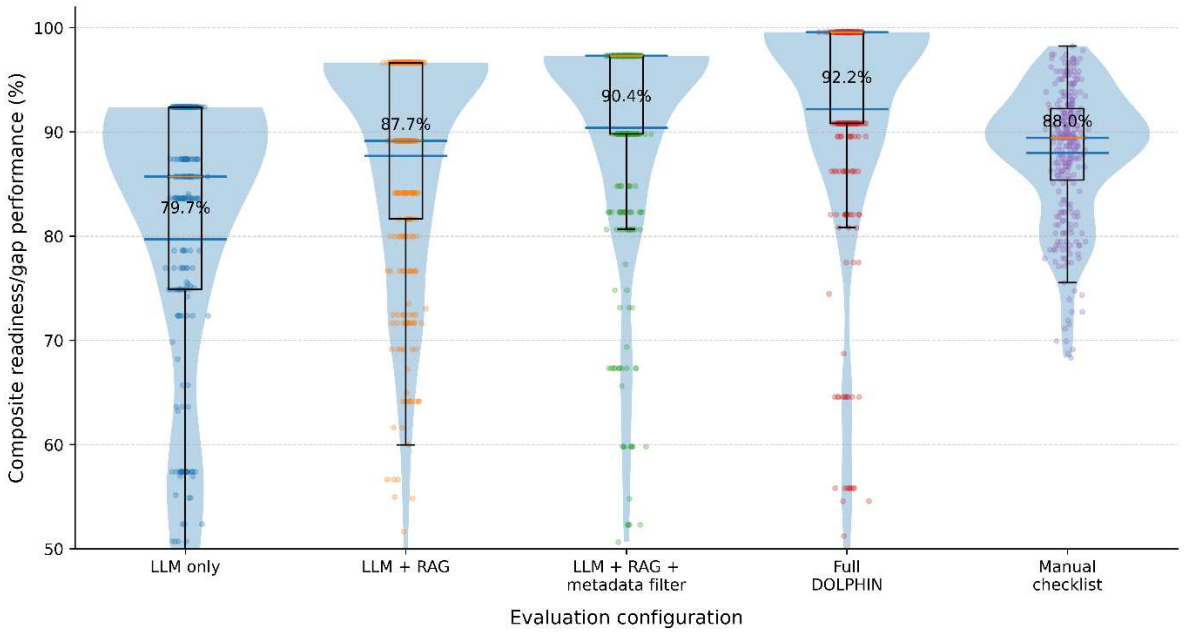


Figure 6. Ablation performance distribution across DOLPHIN configurations

The full DOLPHIN framework achieved the highest composite performance at approximately 92.2%, exceeding the LLM-only, RAG-only, and manual checklist configurations. This demonstrates that the proposed framework is not merely a standalone LLM application. Its performance is achieved through the combined contribution of offline LLM reasoning, RAG-based

evidence retrieval, metadata filtering, FRI scoring, Evidence Gap Taxonomy, and structured report generation.

4.7 Unsupported Output Reduction

Figure 7 evaluates unsupported output rates across different LLM configurations. The LLM-only configuration produced the highest unsupported output rate at approximately 16.3%, indicating that an unguided language model may generate plausible but ungrounded statements when handling incomplete investigation records. When RAG was introduced, the unsupported output rate decreased to approximately 6.3%, demonstrating the benefit of evidence-grounded response generation.

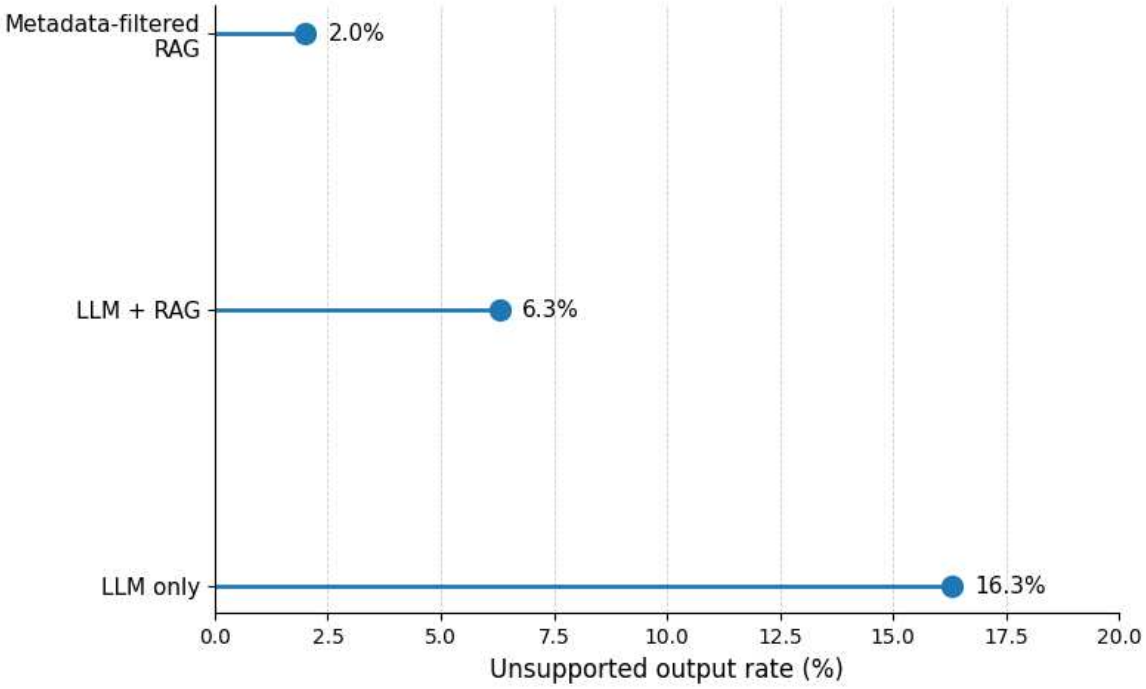


Figure 7. Unsupported output reduction across LLM configurations

The lowest unsupported output rate was observed in the metadata-filtered RAG configuration, at approximately 2.0%. This result confirms that retrieval grounding becomes more reliable when retrieved evidence is constrained to the correct caseThe efficiency and report-quality comparison across manual review, keyword-based review, LLM-only review, and full DOLPHIN-assisted review is presented in Table 7. In privacy-sensitive forensic contexts, this reduction is important because unsupported statements may mislead investigators or weaken confidence in AI-assisted

outputs. Therefore, the results support the use of metadata-filtered offline RAG as a reliability-enhancing mechanism in investigation preparation.

Table 7. Efficiency and report-quality comparison

Method	Mean preparation time ± SD, min/case	Report completeness score ± SD
Manual checklist review	15.46 ± 2.15	2.34 ± 0.79
Keyword-based review	9.17 ± 1.23	—
LLM-only review	—	2.63 ± 0.81
Full DOLPHIN-assisted review	4.41 ± 0.87	3.14 ± 0.78

4.8 Investigation Preparation Time

Figure 8 compares the mean time required for investigation preparation using manual review, keyword-based review, and DOLPHIN-assisted review. Manual checklist review required the longest time, with an average of approximately 15.46 minutes per case. Keyword-based review reduced the mean time to approximately 9.17 minutes per case but still required substantial manual interpretation.

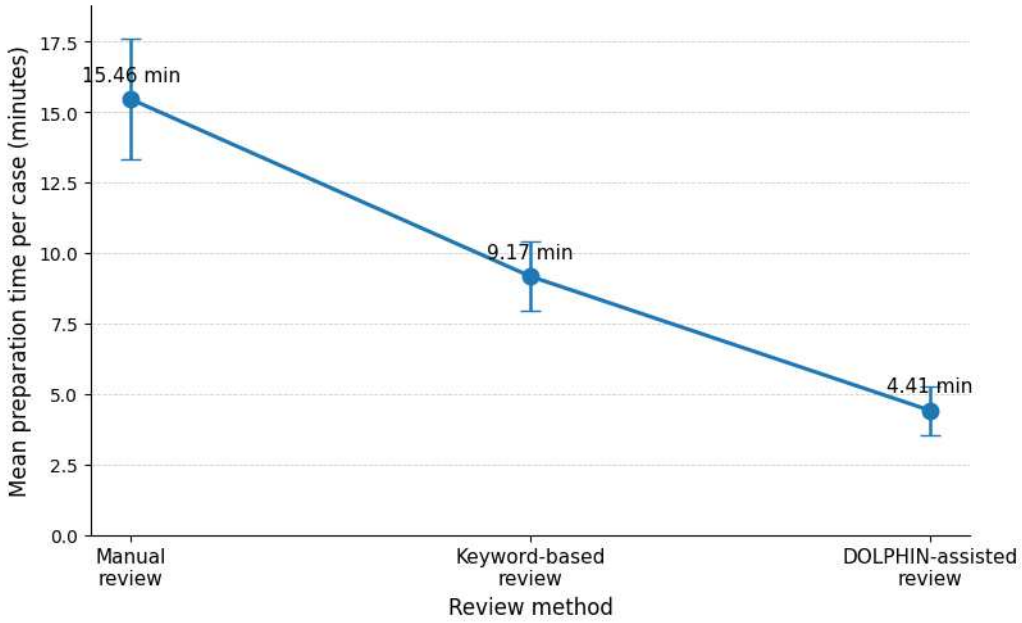


Figure 8. Investigation preparation time reduction across review methods

DOLPHIN-assisted review produced the shortest preparation time, averaging approximately 4.41 minutes per case. This indicates that the framework can improve investigative efficiency by automating evidence organization, FRI scoring, evidence-gap detection, and readiness report generation. The reduction in preparation time suggests that DOLPHIN may be particularly useful for screening multiple investigation files before formal forensic examination.

4.9 Readiness-Report Completeness

Figure 9 compares the completeness of readiness reports generated through manual checklist review, LLM-only output, and the full DOLPHIN framework. The full DOLPHIN configuration achieved the highest report-completeness score because it systematically included evidence summary, missing evidence identification, FRI score, Evidence Gap Taxonomy classification, and recommended investigative actions.

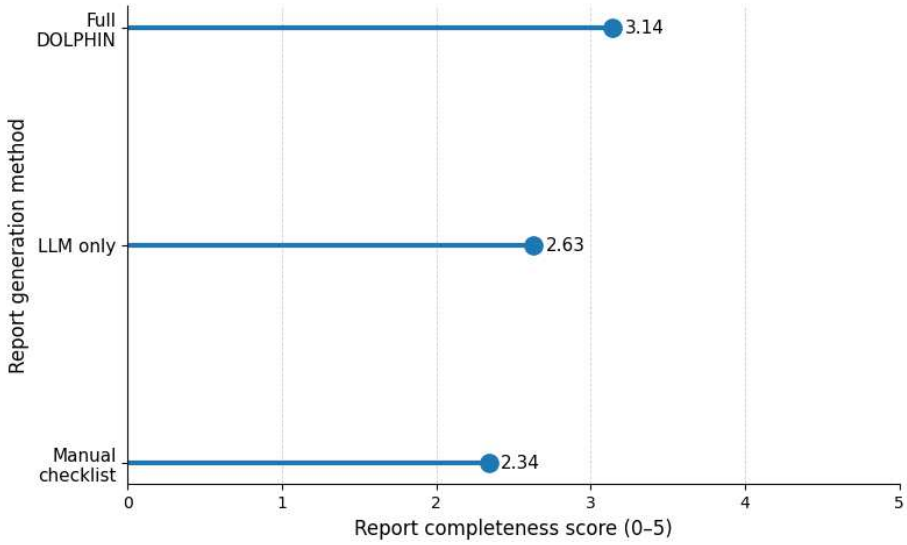


Figure 9. Readiness-report completeness across generation methods

The manual checklist produced more limited outputs because it mainly captured evidence presence or absence without detailed explanation. The LLM-only configuration generated narrative responses but lacked consistent scoring and taxonomy-based structure. In contrast, DOLPHIN combined narrative explanation with transparent quantitative scoring and structured gap analysis. This resulted in more complete and operationally useful readiness reports.

4.10 Discussion

The results demonstrate that DOLPHIN can support privacy-preserving digital investigation readiness assessment using offline Large Language Models and Retrieval-Augmented Generation. The FRI score distribution confirmed that the framework can distinguish between complete, partial, and low-evidence investigations. The confusion matrix showed strong agreement with expert/manual readiness labels, while evidence-gap detection metrics demonstrated reliable identification of documentation, corroboration, visual evidence, and timeline deficiencies.

The RAG retrieval and ablation results further show that metadata-filtered retrieval is essential for reliable forensic use of LLMs. The unsupported output analysis confirms that evidence-grounded RAG substantially reduces unverified model statements. Finally, the preparation-time and report-completeness results indicate that DOLPHIN improves both efficiency and documentation quality. Collectively, these findings support the feasibility of formalizing investigation preparation as a

measurable, AI-assisted, and privacy-preserving stage within the digital forensic lifecycle. However, because the present evaluation is based on a controlled benchmark, future work should validate the framework using larger real-world forensic case collections, multiple crime categories, expert investigator panels, and operational deployment environments.

6. Conclusion and Future Work

This study introduced a privacy-preserving framework for assessing digital investigation readiness, DOLPHIN, which combines offline large language models and retrieval-augmented generation for preparing investigations based on the evidence. The framework aimed to assist with the early stages of digital investigations by evaluating the completeness of the evidence within a case, calculating a forensic readiness index, outlining evidence gaps using an evidence gap taxonomy, and generating structured readiness reports. Unlike cloud-based AI models, DOLPHIN was built in a fully local setting with LM Studio, AnythingLLM, and Qwen 3, allowing for the processing of sensitive investigation data without privacy concerns. The 300 cases of structured investigations tested with a controlled benchmark proved that DOLPHIN is effective at differentiating between cases with complete, partial, and low evidence. The framework's performance in readiness classification, evidence gap detection, retrieval reliability (with metadata-filtered RAG), and a significant decrease in unsupported LLM output was promising. The findings also indicated that DOLPHIN led to decreased investigation preparation time and more comprehensive readiness reports than manual checklists and LLM-only methods. Based on these results, offline LLM-RAG systems can be used in practical, interpretable, and human-centered support for investigation-readiness assessment. The most valuable part of this study is to formally define "preparation to investigate" as a measurable part of the digital forensic lifecycle. The proposed Forensic Readiness Index offers a clear quantitative indication of the case's preparedness, and the Evidence Gap Taxonomy clarifies the type of evidence gap that exists in each case. With metadata filtered RAG and offline LLM reasoning, DOLPHIN provides a structured approach for making decisions, mitigates reliance on ungrounded model outputs, and enhances the traceability of AI-assisted investigation reports. The study has some limitations, however. A controlled benchmark was used to evaluate, rather than operational police case files. This design allowed for systematic testing over varying degrees of evidence completeness, but in

a real investigation, evidence records could be more complex, ambiguous, multilingual, incomplete, and legally significant. Moreover, the present study focused on the investigation-readiness assessment and failed to consider the framework for the final forensic interpretation, legal admission, and autonomously made decisions.

Future research should confirm DOLPHIN with larger and more varied real forensic case sets in several crime types, including the areas of cyber fraud, missing devices, financial crime, online crime harassment, and vehicle theft. Additional research is needed to evaluate the validity, accessibility, readability, and utility of the readiness reports generated by digital forensic experts, investigating officers, and the legal community. The future study should also include the multilingual processing of cases, the integration of a digital evidence management system, the improvement of audit trail generation, explainable retrieval, and standardized validation protocols for forensic LLM applications.

Future work could involve fine-tuning the model for specific tasks, using adaptive evidence weighting, deployment in institutional forensic laboratories, and comparative testing of the LLM and embedding model with other offline versions. Multimodal evidence support such as CCTV frames, scanned documents, audio transcripts, and device metadata may also be included in future versions of DOLPHIN. In summary, the results indicate that privacy-preserving offline LLM-RAG systems have the potential to enhance the preparation of investigations as a valuable assistive technology but must be transparent, evidence-based, and verifiable by humans.

Declarations

Disclosure of potential conflicts of interest: There is no potential conflicts of interest among the authors.

Research involving human participants and/or animals: This article does not contain any studies with human participants or animals performed by any of the authors.

Funding: There is no funding available.

Competing Interests: There are no competing interests among the authors

Author Contributions: All the authors have contributed in an equal manner

Data Availability: The dataset generated and analysed during this study is publicly available at https://www.kaggle.com/datasets/kdivya98/investigation-readiness-benchmark/data . The dataset includes 300 structured investigation readiness scenarios for assessing the DOLPHIN framework, evidence completeness information, Forensic Readiness Index (FRI) scores, readiness classifications, evidence-gap labels, and RAG-related data for the experiments.

Use of Artificial Intelligence Tools

Offline Large Language Models and Retrieval-Augmented Generation were used as part of the proposed DOLPHIN framework and experimental system. The manuscript was reviewed and verified by the authors. The authors remain responsible for the accuracy, integrity, and interpretation of the work.

References

1. Akilal, A., & Kechadi, M. (2025). Cloud Digital Forensic Readiness: An Open Source Approach to Law Enforcement Request Management. *arXiv preprint arXiv:2507.04174*.
2. Alam, S., & Altiparmak, Z. (2026). XAI-CF—Examining the role of explainable artificial intelligence in cyber forensics. *Engineering Applications of Artificial Intelligence*, *167*, 113892.
3. Albugmi, A. (2024). Digital Forensics Readiness Framework (DFRF) to Secure Database Systems. *Engineering, Technology & Applied Science Research*, *14*(2), 13732-13740.
4. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024, May). Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations* (Vol. 2024, pp. 9112-9141).
5. Chen, J., Lin, H., Han, X., & Sun, L. (2024, March). Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 16, pp. 17754-17762).
6. Cherif, B., Bisztray, T., Dubniczky, R. A., Aldahmani, A., Alshehhi, S., & Tihanyi, N. (2025, November). Dfir-metric: A benchmark dataset for evaluating large language models in digital forensics and incident response. In *International Conference on Neural Information Processing* (pp. 17-31). Singapore: Springer Nature Singapore.

7. Chernyshev, M., Baig, Z., Syed, N., Doss, R., & Shore, M. (2026). Large language models in digital forensics: capabilities, challenges and future directions. *Forensic Science International: Digital Investigation*, *56*, 302043.
8. Du, X., Hargreaves, C., Sheppard, J., Anda, F., Sayakkara, A., Le-Khac, N. A., & Scanlon, M. (2020, August). SoK: Exploring the state of the art and the future potential of artificial intelligence in digital forensic investigation. In *Proceedings of the 15th international conference on availability, reliability and security* (pp. 1-10).
9. Dunsin, D., Ghanem, M. C., Ouazzane, K., & Vassilev, V. (2024). A comprehensive analysis of the role of artificial intelligence and machine learning in modern digital forensics and incident response. *Forensic Science International: Digital Investigation*, *48*, 301675.
10. Feng, Z., Feng, X., Zhao, D., Yang, M., & Qin, B. (2024, April). Retrieval-generation synergy augmented large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 11661-11665). IEEE.
11. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
12. Huang, Y., & Huang, J. X. (2026). A survey on retrieval-augmented text generation for large language models. *ACM Computing Surveys*, *58*(12), 1-38.
13. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys*, *55*(12), 1-38.
14. Karie, N. M., & Valli, C. (2021). Digital forensic readiness implementation in SDN: Issues and challenges. *arXiv preprint arXiv:2107.13759*.
15. Khatiwala, J. P., Addai, D. K. N., & Xu, W. (2025, July). Evaluating the Reliability of Digital Forensic Evidence Discovered by Large Language Model: A Case Study. In *2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)* (pp. 1067-1076). IEEE.
16. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, 9459-9474.

17. Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.
18. Loumachi, F. Y., Ghanem, M. C., & Ferrag, M. A. (2024). GenDFIR: Advancing Cyber Incident Timeline Analysis Through Retrieval Augmented Generation and Large Language Models. *arXiv preprint arXiv:2409.02572*.
19. Mekala, S. H., Baig, Z., Anwar, A., Syed, N., & Zeadally, S. (2025). Securing the Industrial Internet of Things: A Comprehensive Digital Forensic Readiness Framework and Cybersecurity Approach. *IT Professional*, *27*(3), 29-37.
20. Mpungu, C., George, C., & Mapp, G. (2024). Digital Forensics Readiness in Big Data Networks: A Novel Framework and Incident Response Script for Linux–Hadoop Environments. *Applied System Innovation*, *7*(5), 90.
21. Nayerifard, T., Amintoosi, H., Bafghi, A. G., & Dehghantanha, A. (2023). Machine learning in digital forensics: a systematic literature review. *arXiv preprint arXiv:2306.04965*.
22. Sanna, S. L., Regano, L., Maiorca, D., & Giacinto, G. (2024). Exploring the robustness of AI-driven tools in digital forensics: a preliminary study. *arXiv preprint arXiv:2412.01363*.
23. Scanlon, M., Breitinger, F., Hargreaves, C., Hilgert, J. N., & Sheppard, J. (2023). ChatGPT for digital forensic investigation: The good, the bad, and the unknown. *Forensic Science International: Digital Investigation*, *46*, 301609.
24. Scanlon, M., Nikkel, B., & Geradts, Z. (2023). Digital forensic investigation in the age of ChatGPT. *Forensic Science International: Digital Investigation*, *44*.
25. Sharma, B., Ghawaly, J., McCleary, K., Webb, A. M., & Baggili, I. (2025). ForensicLLM: A local large language model for digital forensics. *Forensic Science International: Digital Investigation*, *52*, 301872.
26. Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., ... & Zhou, D. (2023, July). Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning* (pp. 31210-31227). PMLR.
27. Tageldin, L., & Venter, H. (2023). Machine-learning forensics: state of the art in the use of machine-learning techniques for digital forensic investigations within smart environments. *Applied Sciences*, *13*(18), 10169.

28. Wickramasekara, A., Breitinger, F., & Scanlon, M. (2025). Exploring the potential of large language models for improving digital forensic investigation efficiency. *Forensic Science International: Digital Investigation*, *52*, 301859.
29. Yin, Z., Wang, Z., Xu, W., Zhuang, J., Mozumder, P., Smith, A., & Zhang, W. (2025). Digital forensics in the age of large language models. In *Artificial Intelligence Driven Forensics* (pp. 55-82). Cham: Springer Nature Switzerland.
30. Zhang, Q., Chen, S., Bei, Y., Yuan, Z., Zhou, H., Hong, Z., ... & Huang, X. (2025). A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.