
# DF-graph: Structured and explainable analysis of communication data for digital forensics

By:
Jeongin Lee, Chaejin Lim, Beomjin Jin, Moohong Min, Hyoungshick Kim

From the proceedings of
The Digital Forensic Research Conference
**DFRWS APAC 2025**
Nov 10-12, 2025

**DFRWS** is dedicated to the sharing of knowledge and ideas about digital forensics research. Ever since it organized the first open workshop devoted to digital forensics in 2001, DFRWS continues to bring academics and practitioners together in an informal environment.
As a non-profit, volunteer organization, DFRWS sponsors technical working groups, annual conferences and challenges to help drive the direction of research and development.
**https://dfrws.org**

Forensic Science International: Digital Investigation 54 (2025) 301981

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS APAC 2025 - Selected Papers from the 5th Annual Digital Forensics Research Conference APAC

# DF-graph: Structured and explainable analysis of communication data for digital forensics

Jeongin Lee [a], Chaejin Lim [b], Beomjin Jin [b], Moohong Min [c,*], Hyoungshick Kim [b,**]

[a] Department of Forensic Sciences, Sungkyunkwan University, Republic of Korea
[b] Department of Electrical and Computer Engineering, Sungkyunkwan University, Republic of Korea
[c] Department of Computer Education, Sungkyunkwan University, Republic of Korea

ARTICLE INFO

*Keywords:*
Digital forensics
AI communication
Retrieval-augmented generation
Explainable NLP

ABSTRACT

Communication data, such as instant messenger exchanges, SMS records, and emails, plays a critical role in digital forensic investigations by revealing criminal intent, interpersonal dynamics, and the temporal structure of events. However, existing AI-based forensic tools frequently hallucinate unverifiable content, obscure their reasoning paths, and ultimately fail to meet the traceability and legal admissibility standards required in criminal investigations. To overcome these challenges, we propose DF-GRAPH, a graph-based retrieval-augmented generation (Graph-RAG) framework designed for forensic question answering over communication data. DF-GRAPH constructs structured knowledge graphs from message logs, retrieves query-relevant subgraphs based on semantic and structural cues, and generates answers guided by forensic-specific prompts. It further enhances legal transparency through rule-based reasoning traces and citation of message-level evidence. We comprehensively evaluate DF-GRAPH across real-world, public, and synthetic datasets, including a narrative dataset adapted from *Crime and Punishment*. Our evaluation compares four approaches: (1) a direct generation approach using only a language model without retrieval; (2) a BERT embedding-based selective retrieval approach that identifies relevant messages before generation; (3) a conventional text-based retrieval approach; and (4) our proposed graph-based retrieval approach (DF-GRAPH). Empirical results show that DF-GRAPH consistently outperforms all baseline approaches in exact match accuracy (57.23 %), semantic similarity (BERTScore F1: 0.8597), and contextual faithfulness. A user study with eight forensic experts confirms that DF-GRAPH delivers more explainable, accurate, and legally defensible outputs, making it a practical solution for AI-assisted forensic investigations.

## 1. Introduction

In high-stakes criminal investigations, digital communication records such as emails, chat logs, and social media interactions serve as critical evidence for revealing intent, planning, and interpersonal relationships essential to reconstructing events and establishing timelines (Mehta et al., 2024). However, their unstructured and context-dependent nature creates significant analytical challenges that extend beyond what conventional methods can address (Sun et al., 2021).

Traditional approaches, such as keyword searches and tabular representations, fail to capture the deeper semantic coherence and speaker dynamics embedded within these communications. The evidentiary value depends not merely on individual message content, but on broader contextual elements including temporal sequencing, interaction patterns, and psychological signals—all of which remain difficult to extract through standard techniques (Shahbazi and Byun, 2022). Furthermore, to ensure admissibility in legal proceedings, any analytical approach must provide traceable and explainable results (Palmer, 2001).

Recent advances in large language models (LLMs) have introduced new possibilities for analyzing text-based digital evidence. These models can interpret communication context, infer implicit relationships, and generate coherent summaries from unstructured input. However, these benefits come with critical limitations: generative models like GPT are prone to hallucination and lack verifiability, classifiers often rely on surface-level lexical cues, and dense retrievers typically fail to capture

* Corresponding author.
** Corresponding author.
*E-mail addresses:* munich1984@skku.edu (J. Lee), chaejin98@skku.edu (C. Lim), jinbumjin@skku.edu (B. Jin), iceo@skku.edu (M. Min), hyoung@skku.edu (H. Kim).

https://doi.org/10.1016/j.fsidi.2025.301981

Available online 3 November 2025
2666-2817/© 2025 The Author(s). Published by Elsevier Ltd on behalf of DFRWS. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

communication data structure or participant interactions. Most importantly, these systems operate as black boxes, making it difficult to audit their reasoning—a crucial requirement in legal contexts (Khalid et al., 2024). In response, global AI governance bodies, including the European Union's AI Act (European Parliament and Council of the European Union, 2024), UNESCO (UNESCO, 2024), and the U.S. Federal Judicial Center (Baker et al., 2023), emphasize explainability and accountability as legal prerequisites for high-risk AI systems.

To address the limitations of generic LLMs in forensic communication data analysis, we propose DF-GRAPH, a graph-based retrieval-augmented generation (Graph-RAG, https://github.com/microsoft/graphrag) framework specifically designed for digital forensic investigations. DF-GRAPH enhances traditional RAG architectures by incorporating a structured knowledge graph that models message sequences, participant interactions, and temporal relationships. This structured representation enables more accurate context modeling, question understanding, and evidence grounding.

Unlike conventional models that treat communication data logs as flat text, DF-GRAPH transforms them into a graph of messages and edges, enabling graph-based subgraph retrieval and rule-based reasoning. This design allows the system to explicitly reconstruct how a conclusion is derived, offering traceable explanations that align with legal standards of accountability and procedural fairness.

Through its multi-stage pipeline—data acquisition and preprocessing, graph construction from communication logs, query-relevant subgraph retrieval, evidence-guided answer generation, and explainable reasoning traces—DF-GRAPH delivers high-fidelity forensic reasoning with legal transparency. Comprehensive evaluations across real, public, and synthetic datasets combine quantitative metrics with expert user studies to validate its effectiveness.

Our key contributions are as follows.

- We introduce DF-GRAPH, a Graph-RAG-based framework designed for digital forensic question answering (QA) over communication data. By integrating discourse–aware graph construction, semantic and structural subgraph retrieval, and rule-based trace generation, DF-GRAPH supports interpretable, legally defensible responses in high-stakes investigative contexts (see Section 4).
- We develop a rigorous quantitative evaluation pipeline, comparing four representative models—GPT only, Hybrid (BERT + GPT), Naive RAG, and DF-GRAPH—across automatic metrics including exact match accuracy, BERTScore-F1, and SummaC-based faithfulness. This benchmarking confirms the superiority of DF-GRAPH in contextual alignment and factual grounding across diverse forensic datasets (see Section 5).
- We conduct a controlled user study with eight experienced digital forensic professionals affiliated with government investigative units. Participants evaluated model outputs across multiple forensic scenarios, assessing factual accuracy, reasoning transparency, interpretability, and task efficiency. The results show that DF-GRAPH consistently delivers more reliable, explainable, and faster responses than baseline models, demonstrating its operational suitability for real-world forensic workflows (see Section 6).

## 2. Related work

### 2.1. Principles and legal requirements in digital forensics

Digital forensics has traditionally relied on foundational principles such as *reliability*, *integrity*, and *verifiability* (Palmer, 2001; Stoykova and Franke, 2023). These principles ensure that digital evidence is both scientifically valid and legally admissible by requiring analytical processes to be consistent, reproducible, and methodologically transparent.

Modern legal and ethical frameworks have expanded these requirements to address emerging technologies in judicial contexts. The European Union's 2024 AI Act (European Parliament and Council of the European Union, 2024) mandates that high-risk systems deployed in judicial and law enforcement settings provide explainable outputs, transparent documentation of decision logic, and meaningful human oversight (Cabrera et al., 2025). Similarly, UNESCO's 2023 Recommendation on AI Ethics emphasizes fairness, transparency, and contestability (UNESCO, 2021), while Article 22 of the GDPR affirms individuals' rights to understand and contest decisions based solely on automated processing (European Parliament and Council of the European Union, 2016).

These developments establish that modern forensic systems must uphold not only traditional scientific rigor but also legal accountability and interpretability to be deemed credible and admissible in legal proceedings (Grimm et al., 2021).

### 2.2. AI integration in digital forensic analysis

The integration of AI into forensic workflows presents both substantial opportunities and critical challenges. While AI offers efficiency gains in analyzing large-scale, heterogeneous datasets, it often operates as a black-box system, raising concerns about verifiability and interpretability that directly conflict with established forensic principles.

Existing AI applications include convolutional networks for detecting illicit imagery (Rondeau et al., 2022; Roopak et al., 2023), NLP methods for suspect identification and semantic extraction from communication logs (Adkins et al., 2024), and linguistic models for authorship attribution (Huang et al., 2025). Additionally, AI-assisted triage tools help prioritize evidence and detect anomalies (Dunsin et al., 2024).

Building on these foundational applications, recent studies have begun integrating NLP with graph-based modeling to better capture discourse structure, temporal context, and relational dynamics. Yin et al. (2025) demonstrate that combining LLMs with knowledge graphs facilitates the reconstruction of fragmented messages into coherent investigative narratives. Similarly, Zhang et al. (2021) model sentence-level semantic relations using hierarchical document graphs, while Zhao et al. (Zhao and Gao, 2024) incorporate emotional dynamics and topic transitions in multi-party dialogues using graph neural networks.

However, these approaches face critical limitations in forensic contexts. Technical constraints include insufficient modeling of speaker roles (DialogueGCN (Ghosal et al., 2019)), temporal progression (MuserGCN (Zhang et al., 2021)), and causal inference capabilities (RAMAS (Barradas et al., 2019)). More fundamentally, concerns about model bias, hallucination, and opaque decision-making processes (Tynan, 2024) limit their legal defensibility. Without transparent reasoning paths and verifiable sources (Garrett and Rudin, 2023), AI-generated outputs may fail to meet the evidentiary standards required for judicial proceedings.

### 2.3. Graph-RAG and explainability in forensic analysis

To address these technical and legal limitations, recent work has explored RAG methods with structural enhancements designed to improve forensic traceability and interpretability. RAG enhances LLMs by grounding their responses in external documents, thereby improving factual accuracy and contextual coherence. However, conventional RAG systems typically rely on vector similarity over chunked texts, which often fails to preserve essential structural semantics, such as temporal event sequences, inter-actor interactions, and causal relations, that are essential for forensic analysis (Fang et al., 2024).

Existing RAG variants attempt to address these limitations but remain inadequate for forensic requirements. While approaches like Mindful-RAG (Agrawal et al., 2024), Hybrid-RAG (Sarmah et al., 2024), and Modular RAG (Gao et al., 2024) offer improvements in specific areas, they lack full support for structural reasoning, discourse modeling, and evidentiary traceability demanded by forensic

applications.

In contrast, graph-structured retrieval and reasoning better support forensic requirements by encoding message relationships, participant roles, and temporal progression. Graph-RAG structures domain knowledge into graphs where nodes represent entities such as messages, actors, and timestamps, and edges encode relations like 'replies-to' or 'temporally-after.' This enables retrieval of structurally coherent subgraphs that maintain narrative flow (Han et al., 2024) and evidence alignment (Larson and Truitt, 2024), while supporting human verification of inference steps (Wu et al., 2024).

Explainable Artificial Intelligence (XAI) is particularly crucial in forensic contexts, where legal standards demand interpretability and defensibility (Hall et al., 2022). Recent XAI techniques include adapting SHAP to graph components, attention-based subgraph visualization (Du et al., 2019), and rule-based tracing of inference paths. Building on these insights, DF-GRAPH advances Graph-RAG by integrating actor-aware temporal reasoning, message-level citation tracking, and rule-based traceability within a unified framework, producing structured, interpretable, legally defensible outputs tailored for forensic communication analysis.

## 3. Motivation and Problem Statement

Transformer-based models, such as BERT, GPT, and RAG, are gaining attention in digital forensics but still fall short of meeting the strict legal and evidentiary demands of the field. In Section 2, we discuss how these models have been applied to tasks like keyword filtering, semantic retrieval, and triage classification. However, forensic settings demand more than raw performance; they require traceability, reproducibility, and structured reasoning aligned with investigative logic and legal standards.

LLMs such as GPT-3 and GPT-4 demonstrate strong linguistic fluency and contextual understanding, but frequently produce hallucinated or unverifiable content (Rudin, 2019), undermining their admissibility in court. BERT-based classifiers are effective for short-text classification (Devlin et al., 2019), but suffer from narrow context windows and opaque reasoning processes (Kelly et al., 2020). RAG models improve factual grounding by retrieving external documents (Zhang and Zhang, 2025), yet lack support for modeling the structural elements of communication data, such as conversational flow, speaker identity, or temporal causality (Han et al., 2024). Furthermore, their inference paths are often non-transparent, failing to meet legal explainability standards.

Recent studies confirm that these AI models fall short of forensic expectations for structured inference, traceable evidence paths, and explainable logic (Bokolo and Liu, 2024). These limitations are especially acute when dealing with communication records (e.g., emails, chats, or messages) that are linguistically complex and structurally rich. Accordingly, new methods are needed that incorporate both the semantic and structural dimensions of forensic communication data.

**Problem Statement.** In digital forensic investigations involving communication records, it is essential to generate answers that are accurate, legally traceable, and grounded in the communication structure. Existing AI systems often overlook speaker dynamics and temporal flow, which limits their interpretability and legal defensibility.

We ask: *How can we design an AI framework that* supports *accurate, structure-aware, and legally transparent reasoning over communication data in forensic settings?*

To address this, we propose DF-GRAPH, a graph-based retrieval-augmented generation framework that encodes message relations, speaker roles, and temporal links into a knowledge graph. This structure enables transparent, evidence-grounded reasoning for forensic question answering.

## 4. System architecture and implementation

### 4.1. Overview

To address the structural and legal limitations of existing LLM-based approaches in forensic QA, we propose DF-GRAPH, a graph-based RAG framework tailored to communication data. DF-GRAPH operates through a multi-stage pipeline that integrates data structuring, graph-based retrieval, and explainable answer generation, as illustrated in Fig. 1.

The pipeline begins with the acquisition and preprocessing of raw message data from sources such as chat applications or SMS records, which are transformed into a structured schema suitable for downstream processing (see Section 4.2). A knowledge graph is then constructed from these normalized communication logs, capturing both temporal and communication structure to support interaction-aware reasoning (see Section 4.3). Next, given a forensic query, DF-GRAPH retrieves a semantically and structurally relevant subgraph by combining embedding-based filtering with topological graph expansion. This process ensures that the retrieved context maintains narrative continuity and role-aware coherence (see Section 4.4). The selected messages are then linearized and combined with system-generated instructions to construct a structured input prompt, which guides an LLM (GPT-4o) to generate an answer grounded in the retrieved evidence (see Section 4.5). Finally, DF-GRAPH extracts rule-based reasoning traces from the retrieved subgraph, identifying interpretable paths between evidence-bearing messages and the final conclusion. These human-readable explanations improve forensic transparency and support legal admissibility of AI-generated responses (see Section 4.6).

### 4.2. Data acquisition and preprocessing

In realistic investigative scenarios, communication records are typically collected from smartphone apps such as instant messengers (e.g., WhatsApp and Line) or SMS via digital forensic tools. These raw data sources contain heterogeneous formats, incomplete timestamps, and personal identifiers. To ensure compatibility with downstream processing, DF-GRAPH applies a standardized preprocessing pipeline:

First, all records are transformed into a structured schema containing sender, receiver, timestamp, and message content. For real-case datasets, an anonymization process is applied by semantically replacing personally identifiable information (PII) to comply with legal obligations mandated by national forensic data handling regulations. Public and synthetic datasets do not require anonymization but are subjected to the same structural normalization for consistency.

For the synthetic dataset adapted from Dostoevsky's *Crime and Punishment*, we convert internal monologues into messenger-style dialogues by introducing a fictional psychiatrist character who acts as

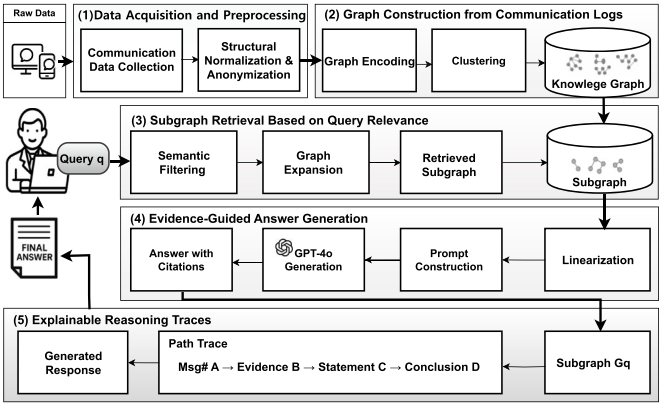


**Fig. 1.** Overview of DF-GRAPH: A multi-stage framework for forensic QA over communication data.

Raskolnikov's interlocutor, and reorder the sequence chronologically. This design simulates high-stakes investigative communication flows in a legally constrained setting.

### 4.3. Graph construction from communication logs

After preprocessing, each communication dataset is transformed into a directed graph $G = (V, E)$, where each node $v \in V$ corresponds to a message annotated with metadata such as speaker, timestamp, and dialog thread. Edges $e \in E$ represent structural or semantic relationships between messages and support downstream QA and reasoning.

We define two base edge types: *temporal edges*, capturing chronological order, and *communication edges*, indicating reply links or same-speaker continuity. To enable higher-level reasoning and improve interpretability, we additionally incorporate semantic edges representing abstract relations. These include `CAUSES`, SUPPORTS, `MENTIONS`, and `CONTRADICTS`, each inferred via prompt-based LLM reasoning.

Relations are extracted via pairwise prompting with GPT-4o, then refined through semantic clustering of the model's explanations. Instead of relying on fixed thresholds, each candidate relation is mapped to one of four predefined types using similarity-based consensus, ensuring structural consistency while preserving LLM expressiveness and mitigating the impact of outlier generations. For example, the following pairs illustrate typical semantic edges:

*Message A:* "She was coming back unexpectedly."

*Message B:* "This compelled him to act quickly."

**Relation:** `CAUSES`

*Message A:* "I only wanted to have the means of living."

*Message B:* "Yes, poverty drives a man to madness and crime."

**Relation:** `SUPPORTS`

*Message A:* "She mentioned something about the library meeting."

*Message B:* "Yeah, the event she texted me about."

**Relation:** `MENTIONS`

*Message A:* "I never killed the old woman. There's no evidence."

*Message B:* "But you knelt before me and confessed everything."

**Relation:** `CONTRADICTS`

*Message A:* "He wasn't making any sense during the conversation."

*Message B:* "Yeah, he sounded completely detached from reality."

**Relation:** `SUPPORTS`

To improve retrieval granularity, the graph is clustered using the Leiden algorithm (Traag et al., 2019), grouping semantically coherent messages into subgraphs. This supports efficient selection of context for each query and preserves thematic unity.

While the semantic edge extraction pipeline showed high consistency across multiple datasets, reproducibility remains a challenge. Inference results may vary due to LLM nondeterminism and ambiguity in message content. We acknowledge this limitation and suggest future work exploring chain-of-thought prompting and domain adaptation to improve stability and generalization.

This graph-based representation enables structured, explainable reasoning over digital conversations and supports forensic QA by modeling both surface structure and latent semantics.

### 4.4. Subgraph retrieval based on query relevance

Given a forensic query $q$, DF-GRAPH retrieves a focused and contextually coherent subgraph $G_q = (V_q, E_q)$ from the full communication graph $G$ to serve as the answer context. The retrieval process consists of two sequential stages designed to balance semantic precision with discourse continuity.

In the first stage, semantic filtering is performed by embedding both the query $q$ and all message nodes using SentenceTransformers. Cosine similarity is then computed between the query and each node embedding to identify top-$k$ semantically relevant nodes, which serve as anchors for subgraph construction. In the second stage, graph expansion is applied to recover local discourse structure around the selected anchors. Using a radius-limited breadth-first search, we include temporally and conversationally adjacent nodes, such as replies and sequential utterances from the same speaker. This step ensures that the retrieved subgraph maintains narrative coherence, speaker-role continuity, and causal flow, all of which are critical for accurate forensic reasoning.

The resulting subgraph $G_q$ contains a concise yet informative set of message nodes that are both semantically aligned with the query and structurally situated within the broader communication context. This graph serves as the foundation for the subsequent prompt construction and answer generation phases.

### 4.5. Evidence-guided answer generation

To generate context-grounded forensic answers, DF-GRAPH leverages the subgraph $G_q$ retrieved in the previous stage. This subgraph captures the temporally and semantically relevant message nodes, which are then linearized into a chronologically ordered sequence $\mathscr{C}_q$. Each message retains its metadata, including speaker, timestamp, and textual content, to preserve evidentiary fidelity and dialog continuity.

The linearized context $\mathscr{C}_q$ is paired with a forensic question $q$ to construct an input prompt $P = \texttt{concat}(\mathscr{C}_q, q)$, which is then passed to a pre-trained generative language model (GPT-4o). To ensure consistency with forensic reasoning standards, DF-GRAPH prepends structured system instructions that constrain the model's behavior in three key ways. First, the model is required to generate answers strictly based on the provided message context, without relying on external or speculative information. Second, it must explicitly cite identifiable evidence sources, including message IDs (e.g., Msg#A), and when applicable, structured references such as report IDs, entity IDs, and relationship IDs. Third, the answer must be presented in a concise and interpretable format that meets the standards of legal admissibility.

This citation scheme enables traceable reasoning, allowing forensic analysts to verify each inference step against underlying data. By grounding outputs in explicit message and knowledge references, DF-GRAPH strengthens the auditability and legal defensibility of its responses. An example of the complete prompt structure used to guide GPT-4o is shown in Fig. 2.

Unlike naive RAG architectures that retrieve semantically related text without structural awareness, DF-GRAPH constrains generation to causally coherent, role-aware, and temporally grounded contexts. This architecture supports both high semantic fidelity and explainability, properties critical for real-world digital forensic deployments.

### 4.6. Explainable reasoning traces

To enhance transparency and legal defensibility, DF-GRAPH generates a reasoning trace for each answer. This trace is constructed by identifying paths within the retrieved subgraph $G_q$ that link premise-bearing nodes to the node corresponding to the final answer. Each path is expressed as a human-readable explanation, such as: `Message A` → `Evidence B` → `Statement C` → `Conclusion D`.

These traces are stored alongside the generated response and serve as supporting evidence in forensic workflows. They can be used for auditing, court presentation, or model verification. Unlike attention-based explainability methods that offer soft, probabilistic cues, rule-based traces provide discrete, logically ordered steps that align more closely with the procedural rigor required in digital forensics.

**Prompt for Graph-RAG forensic answer generation**

```
Role
You are a digital forensic analyst using a
knowledge graph-based AI system.  Your role
is to analyze structured communication data
from a digital investigation and provide
evidence-grounded answers to forensic questions.
Below is a communication context extracted from
case evidence.  Based on this context, answer the
forensic question.
Context (chronologically ordered message nodes
from the knowledge graph):

  • Msg#A: RASKOLNIKOV to DC at 2025-07-05 18:00
    | "I've decided to visit Alyona tonight.
    She'll be alone."
  • Msg#B: RASKOLNIKOV to DC at 2025-07-05 18:20
    | "I wrapped the axe inside my coat."
  • Msg#C: RASKOLNIKOV to DC at 2025-07-05 19:10
    | "It's done.  There's no turning back."

Instruction

  1. Base your answer strictly on the provided
     message context.
  2. Cite not only the message IDs(e.g., Msg#A),
     but also any identifiable sources such as
     reports, entities, relationships, and raw
     data entries whenever possible.
  3. Consider temporal relationship and
     communication flows.
  4. Format your response as either "Yes" or
     "No" followed by a brief, evidence-based
     explanation.
  5. Do not use external knowledge or assumptions
     beyond the provided text.
  6. Ensure your answer supports legal
     defensibility by grounding it in
     verifiable message evidence and maintaining
     transparency in reasoning.

Forensic QA Example
- Question:  Did Raskolnikov plan the murder in
advance?
- Your answer:  Yes.  Msg#A shows that
Raskolnikov intentionally chose a time when
Alyona would be alone, while Msg#B confirms that
he had concealed a weapon in advance.  Msg#C
further indicates that he was aware the act
had already been carried out.  Taken together,
these messages provide verifiable evidence of
premeditation and preparation, thereby supporting
a legally defensible conclusion.
[Source Information:  Reports (7, 21); Entities
(3157, 2726, 2962); Relationships (1106, 1218,
999); Sources (Msg#A, Msg#B, +more)]
```

**Fig. 2.** Prompt format for evidence-grounded forensic answers in DF-GRAPH.

### 4.7. Implementation details

DF-GRAPH is designed as a modular system to ensure flexibility and scalability in forensic QA. It integrates three main components: graph-based knowledge representation, semantic retrieval, and LLM-based response generation.

**Graph Engine and Storage.** The knowledge graph is implemented using `Microsoft GraphRAG` (v0.5.0) as the backend. Each message is modeled as a node containing metadata such as sender, receiver, timestamp, and message ID. Directed edges encode structural relationships such as `REPLY_TO` and `BEFORE`. Additional edges such as `CAUSES` can be defined through prompt-based entity extraction. Graph construction and traversal utilize the built-in pipeline, supporting contextual subgraph reconstruction for downstream retrieval and analysis.

**Semantic Embeddings and Retrieval.** The framework interfaces with external embedding models, such as OpenAI's `text-embedding-3-small`, to map both queries and message chunks into a shared vector space. Input texts are chunked and embedded asynchronously. The embeddings are stored in a LanceDB index and queried using cosine similarity to retrieve the top-$k$ relevant chunks. These are expanded via radius-1 breadth-first search using relational edges to form a contextual subgraph.

**Prompt Generation.** GraphRAG analyzes input data using the `prompt-tune` command to automatically generate structured prompt templates. During this process, the input text is chunked and processed by an LLM to extract entities, infer relationships, and generate summaries. The resulting templates are aligned with *local*, *global*, and *drift* retrieval strategies (as shown in Fig. 3), enabling context-optimized QA tailored to the target domain.

**Language Model Backend.** All generations were performed using OpenAI's `GPT-4o` via the Chat Completion API (model = `gpt-4o-2024-05-13`, temperature = 0.2, max tokens = 512). A low temperature was chosen to promote deterministic and legally defensible outputs, while the token limit ensured concise, audit-friendly responses consistent with forensic reporting practices.

**Traceability Module.** For each generated answer, a path-tracing module reconstructs the shortest evidence path within the subgraph that logically supports the response. This trace is rendered in both graph-based and textual formats, and stored alongside the output for evaluation or audit purposes.

**Experimental Environment.** Experiments were conducted using the Google Colab Pro + environment and conda. We utilized an NVIDIA A100 GPU for accelerated computation and implemented all components in Python (v.3.12). For semantic embedding, we used the `text-embedding-3-small` model provided by OpenAI. Answer generation was performed via OpenAI's GPT-4o model using the Chat Completion API. Graph construction and traversal were handled using the official Microsoft GraphRAG framework.

## 5. Evaluation: Quantitative results

### 5.1. Evaluation methodology

We conduct a quantitative evaluation to compare the forensic reasoning capabilities of four representative model architectures: GPT only, Hybrid (BERT + GPT), Naive RAG, and DF-GRAPH. All models are given the same communication-based input and are evaluated on their

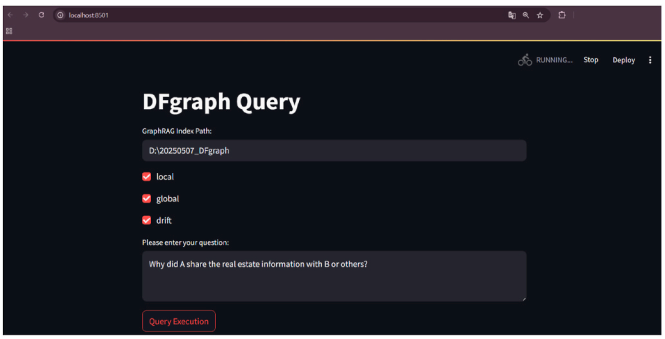


**Fig. 3.** DF-Graph query interface. The user enters a forensic question and selects the relevant subgraph search scopes (local, global, and drift) before executing the query.

ability to generate contextually grounded and factually correct answers to forensic questions.

The evaluation focuses on measurable accuracy-related dimensions. We apply three metrics: Exact Match Accuracy (for binary correctness), BERTScore-F1 (for semantic similarity), and Faithfulness via SummaC-Conv (for logical support from retrieved evidence). These metrics capture complementary aspects of answer quality relevant to legal and forensic standards.

### 5.2. Approaches compared

We evaluate four modeling approaches for forensic QA.

#### 5.2.1. GPT only

A baseline using GPT-4o with no retrieval or structural augmentation. The model receives 35 messages as input and generates responses based only on this limited context. Despite prompt tuning and few-shot examples, the model frequently produces hallucinated or ungrounded answers, limiting its forensic reliability.

#### 5.2.2. Hybrid (BERT + GPT)

This pipeline uses a BERT classifier to infer the intent category of each forensic question (e.g., motive, planning), which guides both message filtering and prompt construction for GPT-4o. While this improves precision over GPT, errors in classification can propagate, and the lack of structural awareness limits deeper reasoning.

#### 5.2.3. Naive RAG

This model performs dense retrieval using FAISS over MiniLM-based embeddings. Retrieved message chunks are prepended to the input question and sent to GPT-4o. Though more contextually informed than GPT or hybrid, this method is vulnerable to irrelevant or tangential retrievals due to a lack of structural constraints.

#### 5.2.4. DF-graph (Graph-RAG)

DF-GRAPH builds a communication graph encoding message order, reply structure, and speaker roles. Given a question, it identifies a subgraph aligned both semantically and structurally, constructs a context-aware prompt, and generates responses using GPT-4o. By grounding generation in the graph, DF-GRAPH achieves higher fidelity, traceability, and legal defensibility.

### 5.3. Data sources and preparation

To evaluate our forensic QA framework, we curated three types of datasets.

- **Real-Case Datasets:** Five anonymized datasets were reconstructed from prior digital forensic investigations. These contain authentic communication patterns—including planning, justification, and concealment—used in criminal contexts.
- **Public Datasets:** (1) The NIST Messenger Dataset (Stockholm Stealer) simulates an organized crime case involving multi-user mobile messages with timestamped interactions. (2) The Cornell Movie Dialog Corpus provides multi-party scripted communication, useful for evaluating dialog structure and temporal flow.
- **Synthetic Dataset:** Adapted from Dostoevsky's *Crime and Punishment*, this dataset reimagines the novel's narrative as a digital communication log by converting internal monologues and plot events into chronological messages, emails, and chat dialogues between characters. The simulated interactions preserve core story elements while reflecting realistic digital formats. The dataset is annotated with forensic reasoning tasks designed to test temporal, relational, and causal inference, mirroring challenges in real-world digital investigations.

All datasets were preprocessed to ensure structural consistency, legal compliance, and model compatibility. Real-world data underwent two-stage anonymization: semantic-preserving substitution of PII and noise injection in embeddings, following the guidelines of GDPR Article 25 (European Parliament and Council of the European Union, 2016) and ISO/IEC 30141 (International Organization for Standardization, 2024). Messages from all datasets were converted into a standardized schema (sender, receiver, timestamp, content) to enable unified parsing. For the synthetic dataset, we preserved psychological continuity by mapping introspective narratives to realistic dialogic exchanges. Table 1 summarizes the dataset types, sizes, and message counts used in our experiments.

### 5.4. Evaluation metrics

To assess the forensic reasoning capabilities of the compared models, we constructed 35 structured questions across five categories commonly encountered in investigative settings: *Motive*, *Execution*, *Concealment*, *Relationship*, and *Confession*. Each question is paired with a ground-truth answer based on domain knowledge or case annotations. For example, a question under the "Motive" category might ask, "*Did [Person A] share the investment plan with [B] in order to mislead?*"—requiring the model to infer intent based on contextual messages.

We evaluate model performance using three quantitative metrics and one structural interpretability indicator, chosen to reflect the unique demands of forensic QA: factual correctness, semantic fidelity, contextual justification, and traceability.

#### 5.4.1. Exact match accuracy

This metric evaluates whether the model's binary response (`Yes/No`) exactly matches the gold-standard label:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i)$$

where $N$ is the number of evaluation instances, $\hat{y}_i$ is the predicted answer for instance $i$, $y_i$ is the corresponding ground truth, and $\mathbf{1}(\cdot)$ is the indicator function returning 1 when the prediction is correct and 0 otherwise.

This metric is critical in forensic analysis where binary determinations (e.g., "Was the act premeditated?") have direct legal implications. High accuracy indicates alignment with expert-labeled truth.

#### 5.4.2. BERTScore-F1

To measure semantic similarity between the model's generated answer and the gold reference, we compute BERTScore-F1:

$$\text{BERTScore} - \text{F1} = \frac{1}{N}\sum_{i=1}^{N} \frac{2 \cdot P_i \cdot R_i}{P_i + R_i}$$

where $P_i$ and $R_i$ denote the precision and recall for instance $i$, derived from cosine similarity between contextual token embeddings using a pre-trained BERT model.

This metric reflects whether the model's response captures the

**Table 1**
Overview of real, public, and synthetic datasets used in the study.

| Dataset | Size | Message Count |
|---|---|---|
| Real-Case Dataset 1 | 2.79 MB | 16,653 |
| Real-Case Dataset 2 | 58.3 KB | 500 |
| Real-Case Dataset 3 | 38.9 KB | 500 |
| Real-Case Dataset 4 | 32.3 KB | 500 |
| Real-Case Dataset 5 | 26.7 KB | 500 |
| Public Dataset 1 (Stockholm Stealer) | 872 KB | 6,302 |
| Public Dataset 2 (Cornell Movie Dialogs) | 41.4 KB | 365 |
| Synthetic Dataset (Crime and Punishment) | 345 KB | 2,367 |

intended meaning of the correct answer, even when wording differs. In forensic contexts, semantic misalignment can lead to misinterpretation or ambiguity in expert reports.

#### 5.4.3. Faithfulness (SummaCConv)

Faithfulness assesses whether the generated answer is logically entailed by the retrieved message context. For each instance $i$, it is computed as:

$$\text{Faithfulness}_i = \frac{1}{k}\sum_{j=1}^{k} s_{i,j}$$

where $k$ is the number of retrieved context sentences, and $s_{i,j}$ is the entailment confidence score (ranging from 0 to 1) between the model's answer and the $j$-th context sentence, as predicted by the SummaCConv natural language inference model.

This metric is essential for legal admissibility, as it quantifies whether a model's output is justifiable based on available evidence. Faithfulness ensures that the answer is not only plausible but grounded in retrievable, case-specific information.

### 5.5. Results

Table 2 presents the performance of all four models on exact match accuracy, semantic similarity (BERTScore-F1), and contextual faithfulness. DF-GRAPH outperforms all baselines across all metrics.

DF-GRAPH achieved the highest exact match accuracy (57.23 %), substantially outperforming Naive RAG (40.51 %), Hybrid (33.06 %), and GPT (21.04 %). However, errors were observed in cases involving ambiguous speakers, temporally disordered evidence, or fragmented conversational contexts.

In semantic similarity, DF-GRAPH again led with a BERTScore-F1 of 0.859, indicating a strong match with reference answers. GPT produced more speculative content, while Naive RAG occasionally overlooked relevant message flow.

Contextual faithfulness was also highest for DF-GRAPH (0.561), confirming that its answers were most consistently grounded in retrieved evidence. This is particularly critical for forensic QA tasks requiring traceable justification.

We further tested statistical significance using the Shapiro–Wilk and Friedman tests. With non-normal distributions in Hybrid and DF-GRAPH scores, we applied the Friedman test, revealing significant differences across all metrics ($p < 10^{-18}$). Bonferroni-corrected post-hoc comparisons confirmed that all model pairs differed significantly ($p < 0.05$), reinforcing the reliability of DF-GRAPH's performance.

DF-GRAPH consistently outperforms existing LLM-based and retrieval-augmented baselines across all evaluation dimensions. Its structured retrieval, graph-based reasoning, and evidence-grounded prompting collectively support more accurate, interpretable, and legally defensible outputs for digital forensic analysis.

## 6. Evaluation: User study

### 6.1. Study design and evaluation protocol

We conducted a within-subjects user study to evaluate four QA models—GPT only, Hybrid (BERT + GPT), Naive RAG, and DF-GRAPH—on realistic forensic tasks. Model order was counterbalanced using a Latin Square method (Keppel and Wickens, 1992).

Each participant completed four tasks, each corresponding to a different model and forensic category: *motive*, *action*, *relationship*, or *confession*. Tasks were constructed from real-world-inspired message logs (chat, SMS, email) and reviewed by forensic experts to ensure comparable evidentiary complexity, ambiguity, and message length.

Each task presented a short, timestamped communication exchange (around 500 messages) and a forensic question requiring contextual interpretation (e.g., "*Why did Person A attempt to share the sales information with Person B or others?*"). Participants received one model-generated answer per task and were asked to write a brief forensic conclusion based solely on the answer and highlight supporting evidence in the model output.

For example, one task presented the exchange shown in Fig. 4, where Person A suggests alternative packaging methods and Person B expresses concern about name traceability.

Participants were expected to infer that Person A aimed to mask ownership and reduce legal exposure. For each task, participants.

1. Wrote their own answer (Q1), later scored for **Accuracy** against expert-curated gold standards.
2. Rated the model's answer on four 5-point Likert items:
   - **Faithfulness** (Q2): Grounded in message content?
   - **Explainability** (Q3): Clear reasoning?
   - **Clarity** (Q4): Specific and unambiguous?
   - **Interpretability** (Q5): Easy to understand?
3. Optionally submitted clarification questions or comments.

We also measured **Efficiency** by recording task completion time and the number of clarification queries, which indirectly indicates each model's cognitive load and practical usability in forensic workflows.

### 6.2. Participants

Eight digital forensic professionals from government-affiliated investigative units participated in the study. All participants specialized in communication log analysis and technical digital forensics, routinely examining messaging data (SMS, messenger logs) as part of their official duties. Table 3 provides detailed participant demographics.

The participants had an average of 8.7 years of professional experience (SD = 3.8), with a range from 5 to 16 years. Among them, 5 were female and 3 were male, representing a diverse and experienced practitioner pool. All participants reported regular exposure to communication data, with six indicating daily handling and two indicating weekly analysis, consistent with high operational relevance for the study tasks.

### 6.3. Results

We report model performance across the evaluation dimensions defined in Section 6.1. Each dimension corresponds to either structured participant responses (Q2–Q5) or behavioral indicators (task time,

**Table 2**
Comparison of Forensic QA Performance across Four Models. Each value represents the mean ± standard deviation over 30 runs.

| Model | Exact Match (%) | BERTScore-F1 | Faithfulness |
|---|---|---|---|
| GPT | 21.04 ± 4.65 | 0.823 ± 0.001 | 0.430 ± 0.018 |
| Hybrid | 33.06 ± 2.78 | 0.841 ± 0.000 | 0.361 ± 0.000 |
| Naive RAG | 40.51 ± 0.88 | 0.809 ± 0.001 | 0.542 ± 0.005 |
| DF-GRAPH | **57.23 ± 1.95** | **0.859 ± 0.005** | **0.561 ± 0.005** |

**Person A → Person F** *(2025-01-04 00:40)*
*"These days, just mentioning digital assets gets people hooked. Dress it up nicely, and no one will question it."*

**Person F → Person A** *(2025-01-04 00:52)*
*"Yeah, B's getting suspicious again. Let's bring E in as a buffer. We'll make it look like it's coming from a third party."*

**Fig. 4.** Example suspicious messenger conversation from a task in the *motive* category, showing a possible attempt to conceal financial activities.

**Table 3**
Participant demographics (N = 8).

| ID | Gender | Experience (yrs) | Data Handling Frequency |
|---|---|---|---|
| P1 | Male | 9 | daily |
| P2 | Male | 16 | weakly |
| P3 | Female | 5 | daily |
| P4 | Male | 5 | daily |
| P5 | Female | 8 | daily |
| P6 | Female | 13 | weakly |
| P7 | Female | 7 | daily |
| P8 | Female | 7 | daily |

clarification queries), offering a multifaceted view of each model's forensic utility in realistic investigation settings.

#### 6.3.1. Accuracy (Q1)

DF-GRAPH achieved the highest task accuracy at 78.1 %, followed by Naive RAG (56.2 %), Hybrid (43.8 %), and GPT (40.6 %). This metric reflects whether participants could infer the correct forensic conclusion using only the model's response. As forensic conclusions must be not only plausible but also evidentially defensible, these results underscore the advantage of structure-aware context retrieval for high-confidence decision-making.

#### 6.3.2. Faithfulness (Q2)

As shown in Fig. 5 (top), 90.6 % of participants "Strongly Agreed" or "Agreed" that DF-GRAPH's output was grounded in actual message content. In contrast, 40.6 % of responses for GPT fell into the "Disagree" range, citing unsupported or hallucinated assertions. Hybrid and Naive RAG performed moderately, but frequently failed to provide verifiable citations or message references.

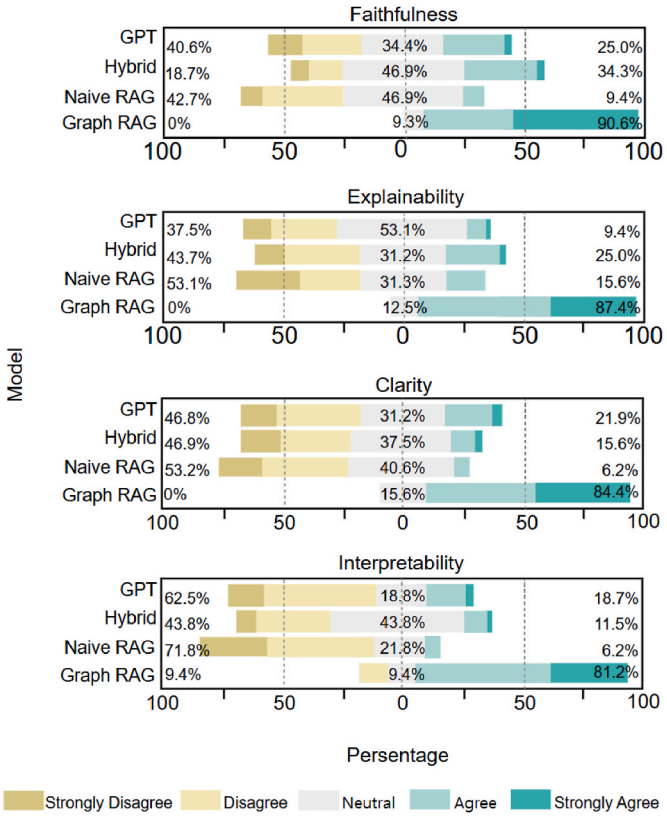


**Fig. 5.** Distribution of Likert-scale responses across three evaluation dimensions. Ratings were collected per task on a 5-point scale.

#### 6.3.3. Explainability (Q3)

Fig. 5 (second top) shows that 87.4 % of participants "Strongly Agreed" or "Agreed" that DF-GRAPH's answers provided traceable, logically structured justifications. In comparison, GPT (37.5 %) and Naive RAG (53.1 %) received "Disagree" or "Strongly Disagree" ratings, indicating opaque reasoning or insufficient causal explanation.

#### 6.3.4. Clarity (Q4)

Clarity evaluates whether the response is specific and unambiguous. Fig. 5 (third top) indicates that 84.4 % of DF-GRAPH outputs were rated as "Strongly Agree" or "Agree" for clarity. Naive RAG scored the lowest, with 53.2 % of responses rated "Strongly Disagree," largely due to vague language and verbose answers. GPT and Hybrid performed better, but often included redundant phrasing that reduced interpretive precision.

#### 6.3.5. Interpretability (Q5)

Interpretability assesses how easily users can extract meaning and rationale from the model output. As shown in Fig. 5 (bottom), DF-GRAPH again led with 81.2 % positive agreement. GPT followed at 18.7 %, with Hybrid (11.5 %) and Naive RAG (6.2 %) trailing. Participants noted that DF-GRAPH's use of message IDs and temporal structure made the explanation path easier to follow and verify.

#### 6.3.6. Efficiency (Behavioral metrics)

Efficiency was measured through task duration and clarification frequency. DF-GRAPH enabled the shortest average task time (41 s) and the fewest clarification queries (1.3 per task), suggesting minimal ambiguity and faster decision support. In contrast, GPT required the most time (58 s) and the highest number of clarification queries (3.3), pointing to a heavier cognitive burden during interpretation. Hybrid and Naive RAG showed intermediate performance.

#### 6.3.7. Qualitative feedback

The participants described DF-GRAPH as "*methodical*," "*justifiable*," and "*aligned with real forensic workflows*" (P2, P5). Specifically, P8 noted, "*The structure helped me see the logic path immediately.*" However, P2 noticed that the model occasionally generated overly detailed outputs that obscured the main point. In contrast, GPT was described as "*imprecise*" and "*detached*," with several responses failing to cite relevant context. Naive RAG was often perceived as "*repetitive*" or "*incomplete*," particularly in motive and relationship inference tasks.

DF-GRAPH consistently outperforms baseline models across subjective evaluations and behavioral metrics. Its integration of graph-based context and source-grounded prompting enables more accurate, interpretable, and operationally reliable outputs, making it a strong candidate for AI-assisted digital forensic investigation.

## 7. Discussion

### 7.1. Alignment with forensic reasoning

DF-GRAPH was designed to reflect how forensic analysts reason—by constructing structured chains of evidence grounded in temporal sequence, speaker roles, and causal relationships. Unlike generative-only or retrieval-based models that treat text as a flat sequence, DF-GRAPH encodes discourse structure using a graph representation and retrieves subgraphs that preserve conversational and evidentiary context.

Findings from the user study confirm that this structural alignment translates into improved interpretability. In tasks such as motive reconstruction and relationship inference, participants consistently rated DF-GRAPH's responses as more faithful and explainable. Notably, 87.4 % of answers received "Strongly Agree" or "Agree" ratings for explainability, highlighting the model's ability to support traceable and coherent reasoning. In contrast, baseline models like GPT and Naive RAG often failed to establish clear cause-and-effect links, resulting in fragmented or ambiguous outputs.

### 7.2. Traceability and legal transparency

Legal admissibility in forensic analysis requires not only factual accuracy but also justification of how conclusions were reached. DF-GRAPH addresses this through rule-based trace generation, which maps the answer to a sequence of evidence-bearing message nodes (e.g., *Message A → Statement B → Motive C → Conclusion D*). This format aligns with documentation practices of forensic reports and court testimony.

Participants strongly endorsed this transparency. As shown in Fig. 5, DF-GRAPH achieved a 90.6 % "Strongly Agree" or "Agree" rating for faithfulness, with no negative responses. Qualitative feedback echoed this trust: "Since the sources are clear, it can be used as evidence, and I hope it will be implemented in actual practice" (P5). These findings highlight that traceability is not merely a desirable property but a critical prerequisite for operational deployment in legal contexts.

### 7.3. Deployment considerations

While DF-GRAPH achieves strong performance under controlled conditions, real-world deployment presents critical challenges for forensic practice.

#### 7.3.1. Scalability

Investigative cases often involve massive, unstructured datasets such as full-device extractions or month-long chat histories, where efficient retrieval becomes essential. Although our implementation supports graph clustering, operational deployment requires advances in hierarchical partitioning, dynamic subgraph caching, and real-time forensic filtering to handle large-scale data processing.

#### 7.3.2. Uncertainty handling

Forensic environments often involve incomplete records, ambiguous language, and multiple plausible interpretations. Future versions of DF-GRAPH should incorporate human-in-the-loop workflows, probabilistic reasoning, and multi-hypothesis generation. Each interpretation must be grounded in distinct, traceable evidence subsets, enabling analysts to evaluate alternative explanations systematically.

#### 7.3.3. Infrastructure constraints

GPT-4o's reliance on commercial cloud APIs raises concerns about data sovereignty, chain-of-custody compliance, and cost-effectiveness in high-throughput forensic environments. To address these limitations, we propose integrating self-hosted lightweight language models (sLLMs) into DF-GRAPH. While sLLMs may underperform on open-ended reasoning, they excel at structured forensic QA tasks and offer superior local deployment, privacy assurance, and cost control. A hybrid architecture where sLLMs handle routine queries and selectively invoke commercial models under strict anonymization protocols would optimize the balance between performance and compliance requirements.

## 8. Conclusion

DF-GRAPH is a graph-based RAG framework for forensic QA. Integrating temporal, structural, and semantic edges into a dynamic message graph and subgraph-guided prompting, DF-GRAPH enables interpretable, traceable reasoning over complex dialogues.

Empirical results on real-case, public, and synthetic datasets show that DF-GRAPH outperforms GPT, hybrid, and naive RAG baselines in all major metrics—achieving 57.2 % exact match accuracy, 0.859 BERTScore-F1, and the highest contextual faithfulness. Statistical analysis confirmed the robustness of these gains across 30 trials. Investigators in a user study rated DF-GRAPH highest in clarity of reasoning, evidentiary alignment, and overall decision-making support.

While DF-GRAPH targets text-based communication, real investigations often involve multimodal evidence, such as call logs, images, and audio. Future work will extend DF-GRAPH by redesigning evidence representation, temporal alignment, and causal reasoning to support multimodal inputs while preserving interpretability and legal validity.

## Acknowledgments

We thank our shepherd, Jewan Bang, for his guidance. We also appreciate the anonymous reviewers for their valuable feedback.

This research was supported by the MSIT, Korea (RS-2024-00459638) supervised by IITP, and the NRF grant funded by MSIT (RS-2024-00451909).

## References

Adkins, J., Al Bataineh, A., Khalaf, M., 2024. Identifying persons of interest in digital forensics using NLP-based AI. Future Internet 16, 426.

Agrawal, G., Kumarage, T., Alghamdi, Z., Liu, H., 2024. Mindful-RAG: a study of points of failure in retrieval augmented generation. In: Proceedings of the International Conference on Foundation and Large Language Models (FLLM).

Baker, J., Hobart, L., Mittelsteadt, M., 2023. An introduction to artificial intelligence for federal judges. Technical report. Federal judicial center. URL: https://www.fjc.gov/sites/default/files/materials/47/An_Introduction_to_Artificial_Intelligence_for_Federal_Judges.pdf. (Accessed 8 May 2025).

Barradas, D., Brito, T., Duarte, D., Santos, N., Rodrigues, L., 2019. Forensic analysis of communication records of messaging applications from physical memory. Comput. Secur. 86, 484–497.

Bokolo, B.G., Liu, Q., 2024. Artificial intelligence in social media forensics: a comprehensive survey and analysis. Electronics 13, 1671.

Cabrera, B.M., Luiz, L.E., Teixeira, J.a.P., 2025. The artificial intelligence act: insights regarding its application and implications. Procedia Comput. Sci. 256, 230–237.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).

Du, M., Liu, N., Hu, X., 2019. Techniques for interpretable machine learning. Commun. ACM 63, 68–77.

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 Sci. Int.: Digit. Invest. 48, 301675.

European Parliament and Council of the European Union, 2016. Regulation (EU) 2016/679 of the European parliament and of the Council of 27 April 2016 (General data protection regulation) URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj. (Accessed 8 May 2025).

European Parliament and Council of the European Union, 2024. Regulation (EU) 2024/1689 of the European parliament and of the Council of 13 June 2024 (Artificial intelligence act). URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj. (Accessed 8 May 2025).

Fang, J., Meng, Z., Macdonald, C., 2024. TRACE the evidence: constructing knowledge-grounded reasoning chains for retrieval-augmented generation. Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 8472–8494.

Gao, Y., Xiong, Y., Wang, M., Wang, H., 2024. Modular RAG: transforming RAG systems into LEGO-like reconfigurable frameworks. arXiv preprint arXiv:2407.21059.

Garrett, B.L., Rudin, C., 2023. The right to a glass box: rethinking the use of artificial intelligence in criminal justice. Cornell Law Rev. 109, 561–628.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., Gelbukh, A., 2019. DialogueGCN: a graph convolutional neural network for emotion recognition in conversation. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), pp. 154–164.

Grimm, P.W., Grossman, M.R., Cormack, G.V., 2021. Artificial intelligence as evidence. Northwest. J. Technol. Intellect. Property 19, 9–106.

Hall, S.W., Sakzad, A., Choo, K.K.R., 2022. Explainable Artificial Intelligence for Digital Forensics, vol. 4. Wiley Interdisciplinary Reviews: Forensic Science, e1434.

Han, H., Wang, Y., Shomer, H., Guo, K., Ding, J., Lei, Y., Halappanavar, M., Rossi, R.A., Mukherjee, S., Tang, X., et al., 2024. Retrieval-augmented generation with graphs (GraphRAG). arXiv preprint arXiv:2501.00309.

Huang, B., Chen, C., Shu, K., 2025. Authorship attribution in the era of LLMs: problems, methodologies, and challenges. ACM SIGKDD Explorations Newsletter 26, 21–43.

International Organization for Standardization, 2024. ISO/IEC 30141:2024 internet of things (IoT) – Reference architecture. URL: https://www.iso.org/standard/88800.html. (Accessed 8 May 2025).

Kelly, L., Sachan, S., Ni, L., Almaghrabi, F., Allmendinger, R., Chen, Y.W., 2020. Explainable artificial intelligence for digital forensics: opportunities, challenges and a drug testing case study. In: Digital Forensic Science.

Keppel, G., Wickens, T.D., 1992. Design and Analysis: A Researcher's Handbook. Prentice Hall.

Khalid, Z., Iqbal, F., Fung, B.C., 2024. Towards a unified XAI-based framework for digital forensic investigations. Forensic Sci. Int.: Digit. Invest. 50, 301806.

Larson, J., Truitt, S., 2024. GraphRAG: unlocking LLM discovery on narrative private data. URL: https://www.microsoft.com/en-us/research/blog/graphrag-unlocking-llm-discovery-on-narrative-private-data/. (Accessed 8 May 2025).

Mehta, U., Chougule, S., Mulla, R., Alone, V., Borate, V.K., Mali, Y.K., 2024. Instant messenger forensic system. In: Proceedings of the International Conference on Computing Communication and Networking Technologies (ICCCNT).
Palmer, G., 2001. A road map for digital forensic research. In: Proceedings of the Digital Forensic Research Workshop (DFRWS).
Rondeau, J., Deslauriers, D., Howard III, T., Alvarez, M., 2022. A deep learning framework for finding illicit images/videos of children. Mach. Vis. Appl. 33, 66.
Roopak, M., Khan, S., Parkinson, S., Armitage, R., 2023. Comparison of deep learning classification models for facial image age estimation in digital forensic investigations. Forensic Sci. Int.: Digit. Invest. 47, 301637.
Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215.
Sarmah, B., Mehta, D., Hall, B., Rao, R., Patel, S., Pasquali, S., 2024. HybridRAG: integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. In: Proceedings of the ACM International Conference on AI in Finance (ICAIF).
Shahbazi, Z., Byun, Y.C., 2022. NLP-based digital forensic analysis for online social network based on system security. Int. J. Environ. Res. Publ. Health 19, 7027.
Stoykova, R., Franke, K., 2023. Reliability validation enabling framework (RVEF) for digital forensics in criminal investigations. Forensic Sci. Int.: Digit. Invest. 45, 301554.
Sun, D., Zhang, X., Choo, K.K.R., Hu, L., Wang, F., 2021. NLP-Based digital forensic investigation platform for online communications. Comput. Secur. 104.
Traag, V.A., Waltman, L., van Eck, N.J., 2019. From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. 9, 1–12.
Tynan, P., 2024. The integration and implications of artificial intelligence in forensic science. Forensic Sci. Med. Pathol. 20, 1103–1105.
UNESCO, 2021. Recommendation on the Ethics of artificial intelligence. URL: https://unesdoc.unesco.org/ark:/48223/pf0000381137. (Accessed 8 May 2025).
UNESCO, 2024. Document for consultation: draft UNESCO guidelines for the use of AI systems in courts and tribunals. URL: https://unesdoc.unesco.org/ark:/48223/pf0000390781. (Accessed 8 May 2025).
Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., Grau, V., 2024. Medical graph RAG: towards safe medical large language model via graph retrieval-augmented generation. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 28443–28467.
Yin, Z., Wang, Z., Xu, W., Zhuang, J., Mozumder, P., Smith, A., Zhang, W., 2025. Digital forensics in the age of large language models. arXiv preprint arXiv:2504.02963.
Zhang, H., Wang, C., Wang, Z., Duan, Z., Chen, B., Zhou, M., Henao, R., Carin, L., 2021. Learning hierarchical document graphs from multilevel sentence relations. IEEE Transact. Neural Networks Learn. Syst. 34, 4273–4285.
Zhang, W., Zhang, J., 2025. Hallucination mitigation for retrieval-augmented large language models: a review. Mathematics 13, 856.
Zhao, J., Gao, W., 2024. A semantic-enhanced heterogeneous dialogue graph network for sentiment analysis in conversations. In: Proceedings of the International Conference on Electronic Technology, Communication and Information (ICETCI).