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.

Multimodal LLM-Driven Forensic Framework for Criminal Intent Detection in Chat Histories

Aashi Manker
Chaitanya Bharathi Insitute of Technology
Juweriah Abdul Raheem Mohammed
Chaitanya Bharathi Insitute of Technology
Sairam Utukuru
sairam.cbit@gmail.com

Chaitanya Bharathi Insitute of Technology https://orcid.org/0000-0001-9639-2923
Siddhartha Mourya
Chaitanya Bharathi Insitute of Technology
Vivek Yelleti
SRM University AP

Research Article

Keywords: Digital Forensics, Multimodal Analysis, Large Language Models (LLM), Criminal Investigation, Hidden Intent Detection

Posted Date: April 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9360101/v1

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

Additional Declarations: The authors declare no competing interests.

Multimodal LLM-Driven Forensic Framework for Criminal Intent Detection in Chat Histories

Aashi Manker[1], Juweriah Abdul Raheem Mohammed[1], Sairam Utukuru[1*], Siddhartha Mourya[1], and Vivek Yelleti[2]

[1] Department of Information Technology,
Chaitanya Bharathi Institute of Technology,
Hyderabad, Telangana – 500075, India
[2] Department of Computer Science & Engineering,
SRM University, Andhra Pradesh, India.
*

Abstract. This paper proposes a multimodal forensic framework that processes audio, text, and image or video captions from suspected chats. Chat history play vital role in criminal case analysis to capture the hidden intent and slang of the people. Using the graph constructor model, we extract the entities involved in the discussion for building a social graph by identifying key influencers and the flow of discussion. A large language model analyzes the generated graphs to identify behavioral insights, motives, and individual influences on the discussion. A structured forensic report is generated to uncover the hidden intent, suspicious terms, and rank the influencers and key entities with confidence scores to support the investigators. Experimental analysis demonstrates that the generated final report is more accurate for the conclusion of the chat history.

Keywords: Digital Forensics, Multimodal Analysis, Large Language Models (LLM), Criminal Investigation, Hidden Intent Detection

*Corresponding author(s). E-mail(s): sairam.cbit@gmail.com

1 Introduction

Digital communication platforms, such as messaging apps and social media channels, offer a platform for content generation and the sharing of information. Conversations today serve as illicit coordination and legitimate collaboration [5]. In the domain of digital forensics, chat histories often contain the planning and coordination among offenders. In the process of the investigation, the investigator needs to manually analyze these and face challenges in terms of colloquial slang, multilingual exchanges, unstructured text, and multimodal elements such as shared images and voice notes, which alter the interpretation of the meaning. Traditional forensic investigations rely on structural mapping for keyword spotting, basic social network analysis, and metadata spotting. These techniques find the details, such as who interacts with whom, but fail to capture what those interactions truly signify. For example, a conversation between two people saying "deliver the goods to our friend" may appear non-harmful until the latent intent or command structures within criminal networks are not considered.

This paper proposes a multimodal LLM-driven forensic framework for Criminal Influence Ranking and Mastermind Identification in chat histories. There are two stages involved in the framework: i) Using the Hyperlink-Induced Topic Search (HITS) algorithm [15] to build a semantically weighted communication network; and ii) Generating a structured forensic report by interpreting the network content using an LLM. In stage I, we identify the roles in the conversation, such as mastermind, middleman, and follower, based on hub-authority dynamics, which collectively map the hierarchical network structure of the criminal conversation. In stage II, LLM analyzes the message content, infers behavioral intent, extracts suspicious phrases, and quantifies confidence. Though graph based method offer the structural impact in communication networks, they normally do not provide the semantic intent on the backdrop of interactions. To overcome this shortcoming, the proposed work includes multimodal LLM reasoning that interprets conversational text and combines network measurements for explicable forensic knowledge.

By combining qualitative language understanding with quantitative graph representation, the proposed framework enables the investigators to go beyond descriptive data to explanatory insights into the data. The generated report explains critical interactions and the hierarchies, includes textual reasoning for suspicious context and content, enabling human-level interpretability in digital forensics for ethical judgment.

Figure 1 shows the process of the proposed framework. It receives chat history input and passes it through two major components. The Graph Constructor is first employed to convert the conversation data into a graph representing the relationships and interaction patterns between participants. The graph representation enables the detection of influential positions and roles in the interaction. After producing the graph, the LLM Analyzer analyzes it to infer behavioral intent from messages based on what is present, recognize suspicious phrases and decipher communication context. An investigation report is then structured

as output with role of the key actors involved, their interaction hierarchy that indicated criminal intent to create a report.

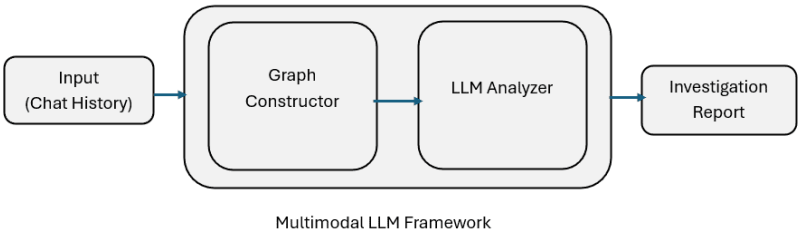


Fig. 1: Overview of the Multimodal LLM-based forensic investigation framework.

The major contributions of this work are:

- To design a framework to interpret the conversational context in the chat history to detect master mind.
- To visualize the critical interactions and hierarchies for suspicious text.

1.1 Background and Motivation

Forensic analysis in digital communications is not a new field, but due to the increasing use of digital communication in recent times, it brings volume and complexity in modern chat history, which presents a significant challenge for the investigators [12]. Traditional methods, such as social network analysis is effective in identifying the connection between the people but often struggle to interpret the actual meaning when they only map who-talks-with-whom, without considering the context [10].Natural Language Processing (NLP) and graph-based algorithms such as HITS usually operate on single modality called text data. But real time communication includes text, audio, image and video and the insights from the conversation is more meaningful only when the context from the heterogeneous data is understood [11]. Recent developments in large language models have demonstrated promising capabilities in analyzing digital evidence and assisting investigators in extracting meaningful insights from communication data [16]. However, these models are not widely used in digital forensic intelligence domain for crime investigation. In this paper we combine a graph based network analysis with multimodal reasoning to capture the semantic and syntax of the criminal patterns in communication history.

The rest of the paper is organized as follows: Section II presents the related work, and the proposed methodology is presented in Section III. Results and discussion are presented in Section IV. The conclusion is given in section V.

2 Related Work

Social Network Analysis (SNA) is an approach that allows viewing and measuring the invisible hierarchies of secret groups in a mathematical perspective. Although criminal groups strive to turn a blind eye to their chain of command, their communication patterns unwillingly give them away. By representing the suspects as nodes and the interactions as edges, an apparently chaotic system of chats is turned into an organized graph.

Conventional SNA tends to equate the importance with activity, yet the frequency of the messages tends to obey Zipf law and is seldom semantically rich [10]. A leader may just be a high-frequency user who is a casual chatter. To get a deeper understanding of influence hierarchy, we need to go beyond the mere metrics to recognize the positions, differentiating the high-authority mastermind from the operational middlemen who facilitate the criminal enterprise. Although the initial schemes identified criminal gangs by examining the co-occurrence of certain entities [12], they used the exact match of keywords. Moreover, although there exist systematic approaches to finding networks in unstructured text [3], they do not always take non-textual evidence into consideration.

These are improved in our framework by weighting the graph edges on the semantic weight of content as opposed to volume. Though our work is based on the mathematically scoring the criminal importance of the messages included in the threads of the internet-based chats, though there is a foundation of the previous methods that provided an opportunity to identify and extract topic threads in the internet-based chats [1]. SentenceTransformers are helpful at processing informal language and noisy data better than traditional Latent Dirichlet Allocation [19], [32]. This will enable us to determine whether a small-world cluster is indeed organizing a crime or it is a mere casual social group and this will considerably decrease false positive [33].

In the literature, few works have utilized the HITS algorithm as a method to distinguish the kinds of importance by dividing the nodes into a "hub" (which directs or curates attention), and an authority (trusted source of information). This is the nature of credibility circulating within real systems, the influence itself is never linear and it is passed through channels and repeated cycles of reference.

Previous uses of n-gram frequency patterns in HITS on chat logs to recognize key users [6] did not take into account the context of a conversation, but such approaches had limited success. The other customized versions, like SCHITS, were specialized in text data only [25]. We have broadened this by making our framework include Whisper as the audio and CLIP as the image, so that suspicious behavior is recorded in any format.

Although there are studies [18], [2] by researchers who have integrated HITS with sentiment scores (S-HITS) to identify influential nodes, high sentiment is not necessarily criminal influence. We have a model that substitutes general sentiment with Zero-shot classification to address criminal commands. Moreover,

the modern Graph Neural Networks (GNNs) are able to identify the key people, however, it frequently works as a black box [36]. Our weighted HITS algorithm yields a mathematically transparent and explainable output, which gives a list of ranked hubs and authorities that is easy to interpret by legal reporting, as well as forensic accountability [22]. Finally, through applying the concept of semantic similarity in vectors rather than on the lexicon at rest [9], our model brings a solution to all of this and consequently utilizes the content scoring methodology to determine the weight of the network structure [24].

Large Language Models (LLMs) have recently emerged as powerful tools for understanding complex and unstructured textual data. Their ability to capture semantic, contextual, and syntactic relationships makes them well-suited for analyzing conversational data in digital forensic investigations. A systematic review of intention recognition methods [13] in digital forensics shows the shift from rule-based and traditional machine learning approaches to deep learning and transfer learning models. These models can uncover hidden intent in online communications. Such models enable investigators to identify patterns of deception, manipulation, or coordination that may be embedded within everyday language. Furthermore, the incorporation of generative models like ChatGPT into this field brings new opportunities along with unverified threats to automated digital forensic investigations [28].

BERT and RoBERTa-based systems have been employed to detect suspicious intent in voice and chat communications by analyzing contextual cues and coded language patterns [7]. Similarly, research on cyber threat intelligence [8] demonstrates that LLMs such as GPT-based models can analyze discussions within cybercrime forums, summarize conversations, and extract relevant intelligence variables with high accuracy . These capabilities help automated systems process large volumes of communication data and flag potential threats that manual analysis often misses. In cases related to public security, large language models have been used to extract important factors about incidents from unstructured sources of information, making the process even more efficient [35].

LLMs have also been applied to detect malicious patterns in emails and online conversations. Comparative analysis of LLM-based models [23] for criminal email detection show that transformer architectures can improve classification accuracy by a lot when dealing with imbalanced or noisy textual datasets. According to the paper [34], one approach involves interacting directly with cybercriminals in chats to gather intelligence on fraud schemes. These solutions highlight the capabilities of LLMs not just to analyze but to proactively counter cybercriminal activity.

Incorporating LLMs into multimodal and real-time systems is something that has further gained traction within the realm of security and forensics. For example, WatchOverGPT [29] utilizes sensors worn by individuals, cameras, and LLMs to process signals from the environment and determine if there is any activity indicating a criminal event. On the other hand, digital forensics [16] has leveraged

the capabilities of LLMs to automatically generate summaries of conversations collected from messaging applications such as Telegram and Signal.

Methods like prompt engineering, retrieval augmented generation, and supervised fine-tuning [4] have been used in order to improve accuracy in threat detection with regards to multilingual and unstructured social media posts. Additionally, generative AI models including RoBERTa, ALBERT, LLaMA, and Gemini [31] have been used in the detection of malicious behavior in online environments, particularly predatory activities and abuse, and have proved effective in this task. These methods represent important tools in combating cybercriminals hiding behind abbreviations and code words in chat rooms.

In addition, investigated the application of AI-driven NLP systems in detecting threatening or aggressive communicational patterns. Systems based on automation [30] have been used for the analysis of threatening messages and prediction of the associated risk of criminal behavior with a high level of precision. Domain-specific fine-tuned transformers [21] like LLaMA2 have been deployed to identify chat interactions that could be considered predatory or abusive in nature, illustrating how powerful domain-specific large language models can be in predicting crimes on the internet. Following this trend, another research conducted in this area has involved hybrid models [26] incorporating BERT into recurrent architectures to predict cyber threats from online forums on the dark web.

Beyond just detecting crimes, LLMs have found applications in larger models aimed at categorization and prediction of crimes. For instance, investigations done [27] on the ability of transformers such as BART, BERT, and GPT in classifying crimes via zero-shot and few-shot prompting indicate the ability of the algorithms to perform flexible analyses in various crime data sets. On the other hand, in forensic analysis procedures [11], generative AI models such as ChatGPT have been used by investigators to mine important information from text and imagery data. The practical application of these abilities has been discussed in recent research that suggested a complete framework for utilizing LLMs for digital forensics in crime investigations [14].

However, there are still a few challenges to address. For instance, most of the current research works have focused on content analysis or intent classification in isolation, without considering the structural relationship between participants in a communication network. Although LLMs have proven to be highly effective in interpreting the underlying semantic meaning of a given text, they are often found to be lacking in terms of their ability to explicitly process interaction hierarchies or influence between participants. On the other hand, graph-based approaches are found to be highly effective in capturing the underlying structure of a communication network, although they are often found to be lacking in terms of interpreting the underlying semantic meaning. Hence, there still remains a need to fill the gap in terms of integrating graph-based communication analysis with LLM-based semantic reasoning.

3 Methodology

Combining graph-based analysis (providing quantitative analysis) with large language model (providing qualitative analysis) offers a new approach for detecting the major people involved in the crime along with the analysis made by LLM for better interpretation will be helpful to the process of investigation. Through graph-based analysis the invisible hierarchies and communication patterns can be recognized. Proposed approach converts web of chats into a structured graph by detecting the suspects as nodes and their interactions as edges. Most of the time people make a potentially misleading assumption as equating the activity with importance to identify the most active communicators. But the goal is to find the network gravitational center [22] by identifying not only the key people but also the specific roles they play such as gatekeepers using centrality measures such as link analysis algorithms.

Our proposed framework follows a two-stage approach that combines structural network analysis with large language model–based reasoning:

- **Stage I: HITS-Based Network Analysis**
- **Stage II: LLM-Based Forensic Reasoning**

Figure 2 illustrates the overall architecture of the proposed framework for multimodal crime analysis. As illustrated, the framework accepts multimodal information, including textual information, audio information, and visual information. This information is then preprocessed by standardizing it. After this, relevant information regarding entities is extracted, and the interaction between individuals is modeled by the graph construction module. In this module, individuals suspected of crime are represented as nodes, and communication between them is represented by edges. This information is then analyzed by network analysis techniques, such as identifying roles and influence between individuals. After this, the information is passed on to the large language model (LLM) analyzer, along with the prompt and other multimodal information. Based on this information, a detailed interpretation of the network behavior is generated by the LLM analyzer. Ultimately, a structured report is generated by the framework, including suspected crime activities, roles of individuals, suspicious behavior, and an overall analysis of the crime scenario. The system uses a special Knowledge Base (KB) to turn raw, unstructured data into actions that are against the law. This KB is the semantic anchoring, which is the expert logic of the framework that is needed to understand intent in different ways.

Before we used the two-stage method, we performed preprocessing. The preprocessing steps are given as follows:

Step A: CLIP-Based Analysis
Image processing module is based on the Contrastive Language-Image Pretraining (CLIP), which is used to transform visual information into verbal words. The system uses specific list of crime-relevant visual concepts (e.g., firearms, illegal substances or large amounts of money) in place of generic object detection.

This module has a critical design element in the form of incorporation of Neutral Labels like domestic animals or landscapes. Given that the CLIP model is a maximum likelihood estimate, which is a softmax based model, it must pick the most relevant label among the given list. With such neutral categories, the system will guarantee that the non-criminal images are classified as safe categories instead of being misclassified by being classified as threats.

Step B: Prototype Sentences
Criminal communication does not literally use specific keywords; rather it is similar to coded language and trying to use indirect directives. In order to address that we have installed a Prototype-Based Semantic Engine.

The system does not attempt to decode a message given by a suspect to exactly the same string but rather, compares an embedding of the text of the suspect to a structured library of Prototype Sentences. These models are the ideal prototypes of various criminal intentions (e.g. "Proceed with the plan" in Commands or "Settle the accounts" in Financial Transactions). Even in case the vocabulary is distorted or altered, the engine is able to identify criminal intent by matching a real-world message to these prototypes.

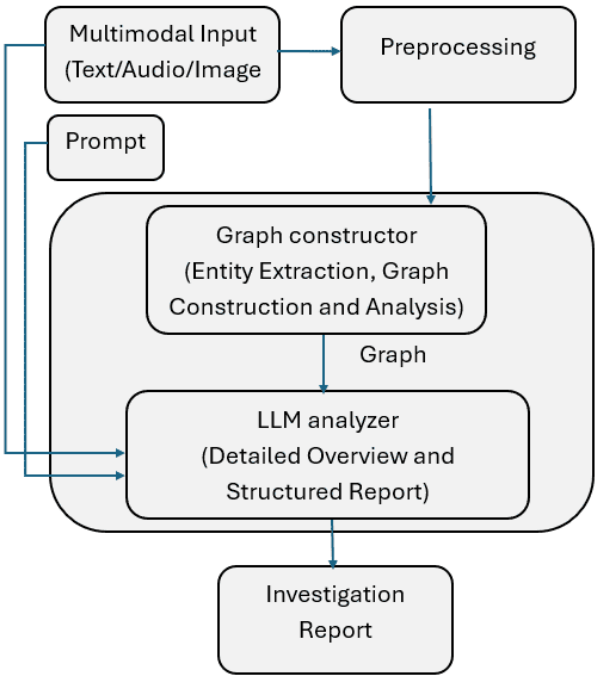


Fig. 2: Proposed Multimodal Crime Analysis Framework Integrating Graph-Based Network Analysis and LLM Reasoning

Step C: Hierarchical Weighting Logic
The system uses a Weighted Category Hierarchy to distinguish roles in a network (e.g. distinguishing between a leader and a courier). These weights are not arbitrary, but they indicate the extent of influence or criminal gravity conveyed by a particular type of communication. Table 1 refers to edge weights on the communication graph are configured with a slightly adapted Likert-scale system especially to analyze criminal networks, along the lines of Weighted HITS (W -HITS)[17].

Table 1: Category-wise Weight Assignment and Justification

Category	**Weight (W)**	**Logic & Justification**
Commands & Directives	1.8	Highest weight; these are the "Hubs" or leaders who move the actions of the network.
Financial Transactions	1.6	Good indicator of advanced criminal logistics and traffic.
Threats & Enforcement	1.5	High-risk behavior, but a little less frequent than everyday orders.
Logistics & Secrecy	0.8–1.2	Tactical communication that is more supportive of operations and has less hierarchical emphasis.
Reporting & Updates	0.5	This is mainly informational in nature; subordinates use this form of reporting to report to their superiors.

We continued to adjust the numbers by using a number of empirical tests until we achieved the optimal role differentiation. In effect we place greater importance on authoritative types such as Commands & Directives (W=1.8) and down-grade noise of Reporting (W=0.5), thus the HITS algorithm is actually able to discern the true power structure and leadership functions. In this capacity the model is able to categorically divide the top leaders and the bottom-level members without being clouded by all the high-frequency chatter [20].

Step D: Pre-Embedding of Prototype Sentences
All KB prototypes are pre-computed into the form of vector embeddings to make sure that they are scaled in real-time during the global initialization stage. By converting the prototypes into high-dimensional tensors once, it means that the system will not repeat any calculations in the course of the iterative chat history processing. This enables quick similarity searches and as a result the framework is able to support large scale datasets with low latency. The following sections provide a comprehensive explanation of the proposed two-stage approach.

3.1 Stage I: HITS-Based Network Analysis

In order to enable multimodal parity, the non-textual information undergoes the transformation to descriptive text before analysis: acoustic data is run

through the Whisper speech-to-text model to extract literal transcriptions, and visual data is run through CLIP to zero-shot recognise against a pre-established criminal knowledge base. The framework allows an integrated assessment of the communication network by bringing audio, pictures, and raw text into one stream of text. After the data has been uniformed, it is processed with a Dual-Engine Scoring Pipeline, which assesses communication in two different perspectives; explicit intent and implicit context.

1. Zero-Shot Intent Classification Engine (grounded on BART-Large-MNLI model) tests the face value of the message to understand the likelihood of explicit labels, e.g. Criminal Conspiracy or Issuing a Command.
2. Semantic Similarity Engine (using the all-MiniLM-L6-v2 model) evaluates whether the input message is similar to the knowledge base prototypes of various forms of communication of criminals.

Both outputs, Zero-Shot Classification score and Semantic Similarity Engine score, are further used to generate a hybrid score, which makes the prediction of interaction strength more accurate prior to constructing the social network graph. After the computation of such Hybrid Scores, a multidimensional social graph is built in which the Hybrid Scores are used as custom edge weights, on which HITS algorithm is applied to determine the key hubs and authorities of the network. The complete process is explained below in a detailed and systematic step-by-step manner.

Step 1: Zero-Shot Intent Classification Engine
We follow a Large Language Model (LLM) that follows the BART-Large-MNLI architecture to detect explicit criminal intent without necessarily having explicitly labeled data. According to this model, there is a likelihood of a certain message falling into any one of three candidate classes, which include: Criminal Conspiracy, Issuing a Command and Normal Conversation. This phase is important as it takes up the face value of the communication. Where the semantic engine (which is discussed below) tries to find camouflaged meanings, the Zero-Shot engine tries to detect blatant criminal behaviour. The raw probability scores of the classifier are converted to a weighted Classification Score as shown below:

$$C_{score} = 1.5P_{conspiracy} + 1.0P_{command} \tag{1}$$

The numerical constants in the C_{score} formula have been specially tuned to give the network graph stable and the role identification accurate. The Multiplier of 1.5 of the Criminal Conspiracy to be 1.5, is to be sure that the conspiracy-related text prevails over the "Normal Conversation" without generating a plethora of false positives. Had this weight been below 1.0 criminal intent would be statistically the same as normal chat. On the other hand, when this weight was more than 2.0, the system would be hyper-sensitive and even minimal variance would be considered a serious threat. The 1.5 number falls within a safe zone that gives distinct distinction between illegal and lawful action. The 1.0 Multiplier of issuing a Command is a Neutral Amplifier. It accepts that commands are needed to

determine leaders (Hubs) within a network, but it does not make a command enough to make a verdict of a Criminal. It supports the score and does not take over it, so the administrative or neutral tasks are not wrongly identified as crimes on high levels.

Step 2: Semantic Similarity Engine
This engine is based on the all-MiniLM-L6-v2 Sentence Transformer model. It is a model that maps all messages to a 384 dimensional space of vectors in which meanings are expressed in terms of proximity to each other. The system compares the similarity of the currently being embedded message (E_{msg}) to the received embeddings of the Knowledge Base prototypes (E_{proto}). S_{score}(Semantic Score) of any message is calculated by the maximum similarity match of the message to any of the criminal categories, with special weight to the hierarchical weights of the categories:

$$S_{score} = \max\left(\cos(E_{msg}, E_{proto}) \times W_{cat}\right) \tag{2}$$

Step 3: Hybrid Score Integration
The ultimate System Confidence Weight (W_{final}) is a combination of the Zero-Shot Classification as well as the Semantic Engine. This is the most important part of the detection logic because it is a hybrid one.

The formula for the hybrid score is:

$$W_{final} = (0.3 \times C_{score} + 0.7 \times S_{score}) \times 2.0 \tag{3}$$

The end weight of every communication edge is decided as a result of hybrid computation that balances the intentionality and the unintentional meaning. We use a 70/30 distribution between the Semantic Score, that is, with semantic similarity having a weight of 0.7 and zero-shot classification having a weight of 0.3. The system emphasizes the underlying context and intent as opposed to being limited to literal word choices and thus concentrating on the semantic engine. In order to make these results compatible with the phase of network analysis, a scaling constant of 2.0 is used on the combined score. In the absence of this multiplier, the aggregate of the raw scores would normally cluster in a smaller range (say between 0.2 to 0.6) that would not be able to pass the system 1.2 threshold of a graph. This scaling has the advantage of ensuring that only the interactions that have high statistical confidence are mapped, which helps to screen out the noise and ensures that the eventual influence network contains strong, high-probability criminal associations.

Step 4: Graph Construction and HITS Centrality Analysis
After the computation of the hybrid score from equation 3 the framework builds the directed, weighted graph G = (V, E), where V is defined as the set of the suspected nodes and E is defined as the communication edges weighted by their corresponding W_{final} value. In order to measure the invisible hierarchy in this network, a Hyperlink-Induced Topic Search (HITS) algorithm is used. In contrast to pure degree measures, HITS is an iterative calculation to provide two reinforcing scores to each node:

1. Authority Score: Determines the significance of the node, through the weight of incoming edges to the highly-ranked hubs, and determines important targets or subject matter experts of the cell.
2. Hub Score : The score indicating the strength of the node in terms of its links to other nodes with high authority, in other words, it shows the masterminds controlling the flow of the processes.

The scores are updated iteratively using the following power iteration logic:

$$a_i = \sum_{j:(j,i)\in E} w_{ji} \cdot h_j \tag{4}$$

$$h_i = \sum_{j:(i,j)\in E} w_{ij} \cdot a_j \tag{5}$$

The system achieves this by using HITS on a graph weighted by deep semantic intent as opposed to simple message frequency to make sure that the rankings that come out are actually the ones that depict real criminal dominance. This enables the investigators to differentiate between the high volume chatters and the actual gravitational centers of the secret organization.

Step 5: Hierarchical Role Identification and Threshold Calibration To form the hierarchy of the network, the system employs HITS (Hyperlink-Induced Topic Search) algorithm that creates Hubs (leaders that issue directives and are known as nodes) and Authorities (carries them out). The determination of role identification involve certain variance factors to the average scores: a node is declared a Mastermind when the Hub score is greater than the average score in the network by a factor of 1.2 or more and a Middleman must have an Authority score a factor of 1.5 or higher than the average. These constants are carefully selected to provide statistical significance; a 1.2 constant with Masterminds is enough to filter out a primary leader and not too restrictive to ignore a first-among-equals dynamic but the higher 1.5 constant with Middlemen is more of a filter and makes it hard to find high-level coordinators instead of common followers. Reducing these constants would mean over-identification, with all the active members identified as leaders, and increasing them would lead to an empty-graph problem as it would discover a hierarchical structure with no valid structure. Evaluation and performance of the proposed HITS approach have been discussed in section 3.2 in terms of confidence and accuracy measures.

3.2 System Confidence and Accuracy Measures

In order to determine the reliability of the graph generated by Stage I process, an evaluation approach is considered which analyzes the distinctiveness of the network structure in terms of the discovered roles.

3.2.1 Mastermind Identification Confidence ($Conf_{MM}$)

For determining the "top hub" in a crime syndicate, it is important to select a leader that makes him stand out from the other possible leaders. The system measures the distance between the Hub and the most important secondary node.

$$Conf_{MM} = \frac{Hub_{max} - Hub_{runner_up}}{Hub_{max}} \tag{6}$$

If Hub_{max} increases by a little more than the position of the secondary node, then it is due to a big difference, which refers to a Mastermind, meaning one person giving out most of the commands. Normalization is done so that when we divide by Hub_{max} we normalize our score to be between 0 and 1 hence reporting consistency in accuracy across datasets of varying sizes.

3.2.2 Structural Authority Confidence ($Conf_{Mid}$)

This index indicates the clarity of the hierarchy between the (high authority receivers) known as Middlemen and the (end-node receivers) known as Followers.

$$Conf_{Mid} = \min\left(\frac{Avg(Auth_{mid}) - Avg(Auth_{fol})}{Avg(Auth_{mid})}, 1.0\right) \tag{7}$$

To have a network that qualifies as a structured hierarchy, the network needs to have mathematical distinctions between those who have authority and those who do not. When the average authority of a Middleman is almost similar to that of a Follower, the Structural Confidence collapses, indicating that the network is a flat peer-to-peer group, and not a criminal cell.

3.2.3 Final System Confidence Score ($Score_{Sys}$)

The leadership clarity and the structural clarity are weighted to give the total accuracy of the detection.

$$Score_{Sys} = (0.6 \times Conf_{MM}) + (0.4 \times Conf_{Mid}) \tag{8}$$

The chief objective of criminal intelligence is to find out who is the Mastermind. Thus the clarity of the top hub (60%) is given a bigger weight than that of the subordinates.The identification of the middlemen gives sufficient evidence of an ordered operation.Without a clear middle-tier, the system remains cautious, even if a leader is identified.

3.2.4 Threshold-Based Accuracy Filtering

A High-Pass Filter (Threshold = 1.2) is used to guard the whole pipeline. Communication edges with a hybrid score of less than 1.2 are removed prior to the

commencement of the HITS algorithm. This will make sure that only strong evidence edges are used to calculate the accuracy metrics. This type of noise rejection maintains the final confidence report of the system since the low-confidence noise in the system is rejected before it even has the opportunity to dilute it.

3.3 Stage II: Multimodal LLM-Based Forensic Reasoning

While Stage I focuses on extracting structural relationships and identifying influential actors using network analysis, Stage II performs high-level semantic reasoning on the structured outputs generated in the previous stage. The objective of this stage is to combine the structural insights derived from graph analysis with multimodal communication evidence in order to generate a comprehensive forensic interpretation of suspected criminal activity. Through the incorporation of chat data, centrality measures, and diverse media content—such as texts, audio, and imagery—the system creates an organized narrative that helps to clarify the role of participants, highlight any suspicious communications, and deduce the criminal behind the network of interactions.

Step I: Model Configuration
Implementation of LLM analyzer is done through the use of the Gemini 2.5 Pro large language model, available through the Google Generative AI API. The parameters used for the model are 0.3 for temperature and 3,500 tokens for max output. Temperature is kept relatively low since output stability and determinism are needed when analyzing data forensically. The reason for this is that results need to be reproducible. With a high limit on token count, the model can create a comprehensive report of analysis

Step II: Evidence Aggregation
Prior to drawing conclusions, the first step is to investigate the communication data to find out who communicates with whom, and to count the number of people participating in this network. These people provide the background necessary for reasoning about this data. Centrality measures and actor roles determined during the Stage I are given for better understanding by the LLM.

Next, the model takes as input the table of centrality measures along with the visualization graph that were produced in the previous stage, which gives an opportunity to analyze the relationship and interaction dynamics as shown in the network visualization graph. Also, the communication data is classified according to communication modes: audio, text, and image messages. Data is sent to the model according to the modality selected.

Step III: Prompt Design
An extensive prompt is used to direct the model towards outputs that are organized and coherent. The prompt forces the model to take on the persona of a forensic investigator and analyze the information in this capacity. It compels the model to strictly follow the JSON format by mandating that the model generate an overview of the crime, detailed descriptions of roles played by each individual, suspicious terminologies used and their interpretations, criminal intent behind

the crime, and confidence scores. Other formatting conditions make sure that the language model generates only valid JSON and not markdown or other types of explanation texts. The example of the prompt format is provided below:

```
You are a forensic analyst AI, given participants A,B,C...
Analyze the evidence and RETURN ONLY valid JSON:
{
"model_used":"gemini-2.5-pro",
"summary":"High-level narrative of the criminal operation",
"roles":{"A":"role description","B":"role description"},
"suspicious_terms":{"term":"why suspicious"},
"intent_classification":"crime type",
"confidence_score":0.0
}
```

This well-defined prompt guarantees that the output generated by the model is consistent with the schema, with evidence from structure and multimodality used to facilitate forensic inference.

Step IV: Multimodal Inference Process
During inference, the system constructs a multimodal input consisting of the instruction prompt, structural evidence obtained from graph analysis, the network graph visualization, and the available communication media. The evidence modalities included in the input are controlled through predefined experimental configurations, which allow the system to perform unimodal analysis (text-only, audio-only, or image-only) as well as multimodal analysis combining multiple data sources.

The LLM receives multimodal inputs which are well-structured and include components such as centrality measurements, chat logs, and media assets. All the multimodal inputs are received in parallel to make an evaluation of human interaction, analyze relationships between individuals, and integrate structural evidence from the network graph into the conversation and media context.

Step V: Structured Forensic Report Generation
After the inference process, the response goes through a process where formatting elements like the use of code blocks are stripped out. Afterward, the response is transformed into a valid JSON object to ensure that the output adheres to the predefined schema. The final product is a structured forensic investigation report that includes a narrative summary of the suspected criminal operation, detailed descriptions of the roles of the people involved, explanations of any suspicious terms, a classification of the likely criminal intent, and a numerical confidence score. The report also records the communication modalities used during the analysis to maintain traceability across different experimental configurations.

4 Experimental results and discussion

4.1 Dataset Generation

We have created a Synthetic Weighted Communication Network to assess the performance of the framework in order to replicate real-world criminal hierarchies. In contrast to random datasets, this is an engineered environment that aimed to determine the interaction between NLP-learned weights and graph-theoretic role discovery. We have used a custom data generator to generate a structured stream of 65 to 80 messages in a centralized criminal activity in the form of the JSON format. The semantic validation of the Masterminds and Middlemen links was achieved by aligning the message text strategically with our semantic knowledge base specifically making sure that the links between the Masterminds and Middlemen were of high-value directives (weight 1.8) thus providing ground-truth validation by its own.

The simulation imposed a clear role segregation by means of definite patterns of communication. One Mastermind was the main source of command (high Hub score) and three to five Middlemen were the coordinators with high authority that further distributed tasks and a higher number of Followers were the end-node focused on reporting (low Hub/Authority scores). This design was to provide the HITS algorithm with a structure that can be mathematically proved. This would allow us to determine the validity of the model in distinguishing between leaders and followers. Table 2 represents eight customized datasets that were used for testing our framework.

Table 2: Summary of Synthetic Datasets for Clandestine Network Analysis

Dataset	**Description**
Dataset 1	International trafficking and money laundering involving a strict command hierarchy and enforcement actions.
Dataset 2	Industrial espionage and data theft focused on illegal asset acquisition and rapid financial conversion.
Dataset 3	Coordinated ransomware attack involving system encryption, digital currency extortion, and technical coordination.
Dataset 4	International arms trafficking utilizing forged documentation, untraceable logistics, and hidden financial channels.
Dataset 5	Illegal antiquities smuggling involving secret excavation, international buyers, and high-value artifact transport.
Dataset 6	Cross-border arms smuggling supported by official corruption, decoy logistics, and digital evidence erasure.
Dataset 7	Organized museum heist targeting art and antiques with black-market distribution and evidence destruction.
Dataset 8	Large-scale currency counterfeiting operation involving specialized manufacturing and a hidden distribution network.

The creation of each of the dataset was done with a specific design, which would test the ability of the system to deal with different kinds of crime and

organizational hierarchies. The full dataset collection, including the actual JSON chat data used in this research, can be found at: https://github.com/aashimanker/Major-Project/tree/main/Datasets.

4.2 Implementation Details

The graph analysis model was implemented in the Python programming language, leveraging a suite of robust, open-source libraries to perform the complex analytical tasks. The core of the natural language understanding module was built using the Hugging Face ecosystem; specifically, the `transformers` library was employed to access the `facebook/bart-large-mnli` model for zero-shot classification, while the `sentence-transformers` library provided the `all-MiniLM-L6-v2` model for generating semantically rich sentence embeddings. For the structural analysis, the `NetworkX` library was instrumental in constructing, manipulating, and analyzing the directed communication graph, including the execution of the HITS algorithm. The resulting data, such as node scores and role assignments, were organized and managed using the `pandas` library for clear, tabular presentation. Finally, all network visualizations were generated using the `Matplotlib` library to create clear and interpretable graphical representations of the influence network. These NLP and data science libraries operate on top of the `PyTorch` deep learning framework, which provides the foundational tensor operations.

4.3 Influence Analysis and Network Visualization

In order to test the framework and confirm that it could deconstruct a communication hierarchy, our system processed the experimental datasets. The analysis produced two main outputs, a table of the influence measures in quantitative form, and a visualization of the network structure in the qualitative form. These findings are introduced to prove that the framework is effective in determining the key actors and their functional roles in terms of the flow of influential communication. Table 3 denotes the various roles according to the movement of the influence in the network.

Table 3: HITS Algorithm Results for Dataset 1: Hub and Authority Scores by Role

Person Name	**Hub Score**	**Authority Score**	**Assigned Role**
Alpha	0.490099	0.029439	Mastermind
Bravo	0.035064	0.344892	Middleman
Charlie	0.041182	0.340868	Middleman
Delta	0.135239	0.014626	Follower
Echo	0.137753	0.018150	Follower
Foxtrot	0.077820	0.104600	Follower
Golf	0.082843	0.147426	Follower

The Mastermind scores high on Hub Score since he/she is the central source of guidance, giving orders to other subordinates. A hub is a node that does point

to a lot of authorities in terms of graph theory. The opposite end of the list is Middlemen and the Followers who are concerned with Authority Scores since they take the orders. When the Authority Score is high, it indicates that the node is referred to by powerful hubs. Middlemen sort of fill in the gap - they have a hub and an authority score that are mixed as they receive the orders and relay them. The followers are largely the receiver of information and hence they are high in authority and low in hub influence.

In order to demonstrate how the HITS algorithm updates in an iterative process in our forensic analysis, we have followed the computation of a sample node, Alpha. As an example, the Alpha node receives directed edges to four adjacent hub nodes: Foxtrot, Golf, Bravo, and Charlie. The details are explained in the following steps.

Step 1: Initialization

The way we began was that we initially gave all the hub (h) and authority (a) scores a uniform value in order to have an unbiased start in the process.

Step 2: Iterative Updates

To update the score of Alpha, we add the value of the hubs of its predecessors. Table 4 provides the variation of the authority score of Alpha with three manual iterations as defined by equation 4. The hub score of a node is calculated similarly by summing the authority scores of its successors as represented in equation 5.

Table 4: Manual Authority Score Update for Alpha Across Iterations

Iteration	**Calculation Logic:** $a(\alpha) = \sum h(\textbf{predecessors})$	**Intermediate Hub Sum**	**Normalized Authority** $a(\alpha)$
Initial	N/A	1.000000	1.000000
Iter 1	$h(F) + h(G) + h(B) + h(C)$	$1 + 1 + 1 + 1$	0.485071
Iter 2	$0.4358 + 0.4358 + 0.3995 + 0.3995$	1.670600	0.499587
Iter 3	$0.4218 + 0.4218 + 0.4186 + 0.4186$	1.680800	0.502197

Step 3: Global Convergence

Though the manual iterations provide an idea of the update mechanisms, the computed values of the final authority were calculated using NetworkX library. In the complete dataset, the algorithm was able to converge in about 25–30 iterations. The variation between the hand-rolled version (0.502197) and the final result (0.029439) is due to the global normalization of the entire multi-modal network, to ensure that all authority scores add to a limited sum.

The analysis of the network under different inputs and brings out clear results of node relationships and edge weights change. We discussed these various graph outputs in the following sections to show the way criminal structure is mapped. Figure 3 shows that the system's ability to map the criminal hierarchy changes with the type of input data.

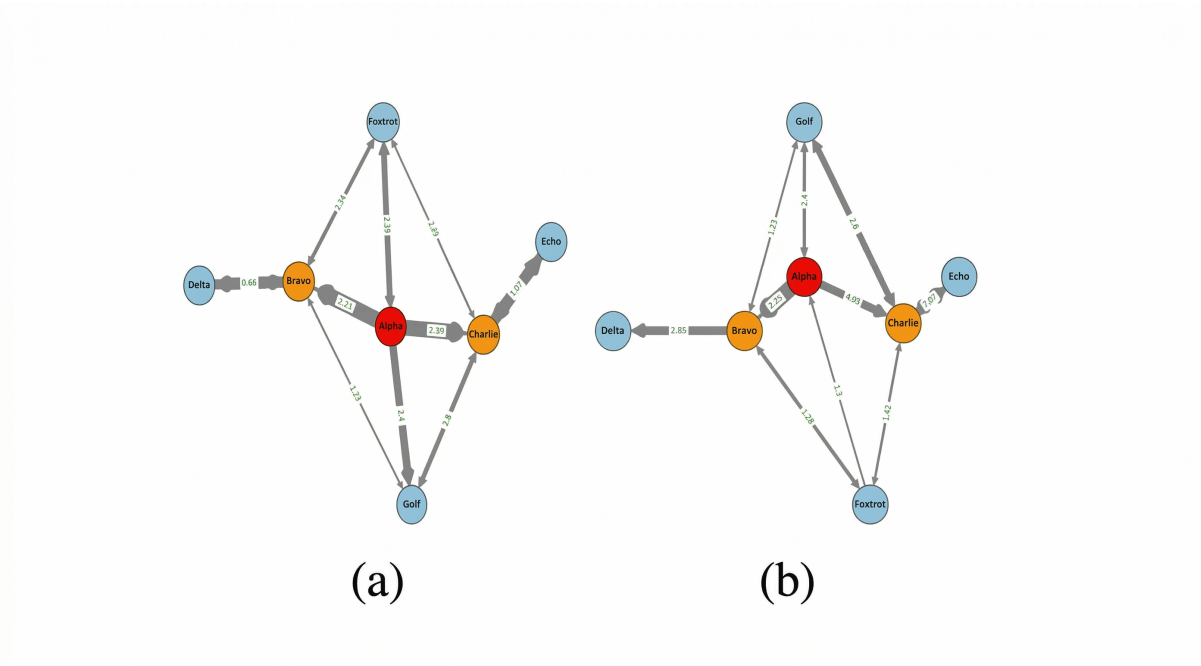


Fig. 3: Influence network diagrams for two different input cases in Dataset 1. (a) Graph generated with only audio input, and (b) Graph generated with only text input

Figure 3(a) is the network constructed entirely out of audio data and does identify Alpha as the key central Mastermind, although the weights on all the edges are weaker due to all the complexity and possible information loss during the speech-to-text translator. Figure 3(b), by comparison, indicates the outcome of the text analysis in which the influence scores are significantly higher. This is because typed communication has more salient semantic features such as explicit instructions and particular criminal lexicon that are more consistent with weighted prototypes of the system. The image modality provided no graph at all; visual evidence provides situational context, but it does not meet the requirement of the directive intent necessary to generate a network hierarchy or graph. The effect of dual inputs fusion on the influence network is discussed as depicted in Figures 4(c)-4(e).

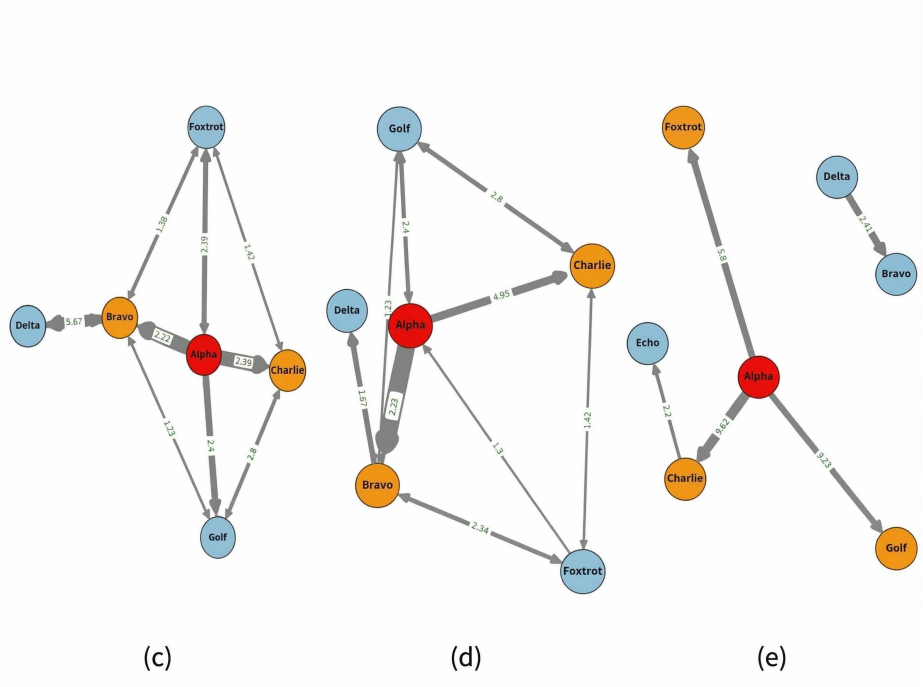


Fig. 4: Influence network diagrams for different combinations of input for Dataset 1

Fig. 4(c) is the most comprehensive network structure. By combining text with audio, the system uses the clarity of written messages to validate the speech data, resulting in a much sharper organizational map.

Fig. 4 (d) shows that there is contextual reinforcement in between the text and image data combination. The pictorial data supports the verbal assertions, and thus the most crucial edges are assigned very high weights in comparison to a one-modality analysis.

Lastly, Fig. 4(e) emphasizes the combination of image and audio. It has the highest weight in individual edges, like the Alpha-Charles edge of 9.62, but the fact that there is no textual connective tissue causes the topology to be fragmented. Therefore, audio-visual data is a great testimony, but the text is considered to be the most dominant medium of preserving the structural integrity of the entire criminal hierarchy. Figure 5 depicts the obtained graphical view of the influence network. This graph gives an intuitive, visualization of the topology of the network in a fast manner.

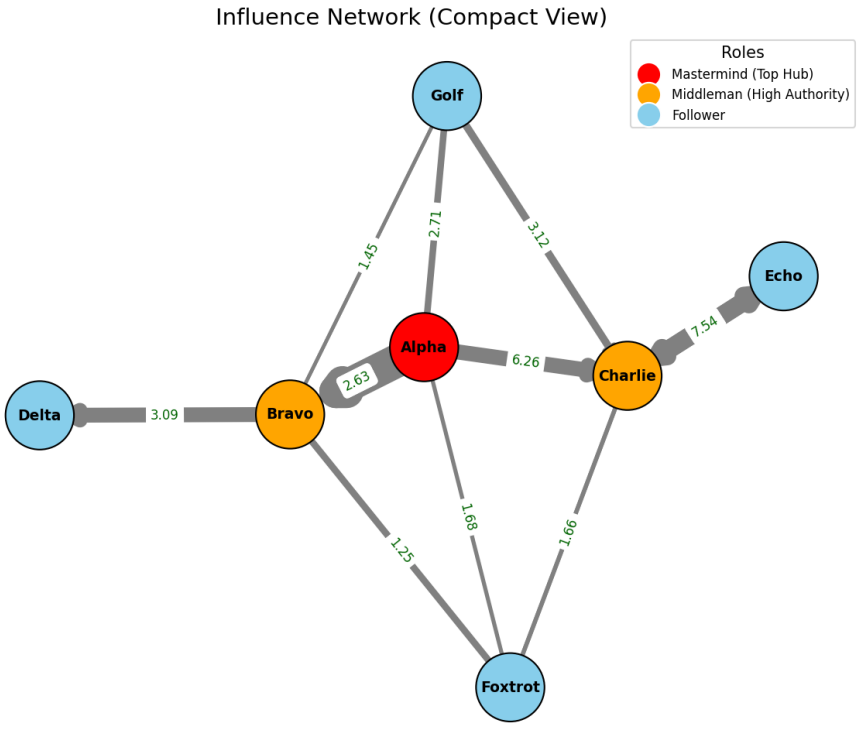


Fig. 5: Graph generated when all modalities are considered for Dataset 1

The assigned roles are used as the node colors, the directed edges are used to indicate the main communication pathways with the weighted ones being the ones with the calculated influence scores. The combination of these findings can give a complete and explainable forensic overview of the power structure of the network.

4.4 LLM-Based Forensic Report Evaluation

The graphs generated from Stage I were analyzed using a multimodal large language model. The Gemini 2.5 Pro model was used to generate structured forensic reports by jointly reasoning over graph-based interaction patterns and multimodal conversational evidence. The model was provided with multiple inputs, including graph (from Stage I) centrality measures and raw conversation data. Depending on the experimental condition, audio and image evidence were also incorporated. This multimodal setup enabled the model to correlate structural influence patterns with semantic cues, improving the understanding of participant roles and behaviors.

The output consists of a JSON report containing a summary of the conversation, role descriptions, detected suspicious phrases, and their confidence scores. Strict

requirements were set for the output schema to ensure uniformity and ease of evaluation.

As shown in Figure 6, this is the forensic report output for the Dataset 1, which considers all the modalities (i.e., text, voice, and image). The rest of the reports will be provided in the supplementary material (Appendix A).


```
{
  "summary": "The evidence indicates a hierarchical criminal enterprise led by 'Alpha', the mastermind. Alpha issues directives concerning financial crimes, such as money laundering via offshore accounts, and orders acts of intimidation or violence. Lieutenants 'Bravo' and 'Charlie' act as middlemen, managing the operational logistics and disseminating orders to a team of followers. This team ('Delta', 'Echo', 'Foxtrot', 'Golf') executes field tasks including surveillance, sourcing untraceable vehicles, and transporting illicit 'packages' or 'assets'. The group employs counter-surveillance techniques like burner phones and secure lines, and there is a clear intent to use coercion and eliminate potential risks ('loose ends').",
  "roles": {
    "Alpha": "Mastermind. The central hub and leader of the operation. Issues commands, authorizes actions, manages the financial aspects (money laundering), and enforces control through intimidation.",
    "Bravo": "Primary Middleman/Lieutenant. Reports directly to Alpha. Manages logistics, procures equipment and vehicles, and directs followers Delta, Foxtrot, and Golf in carrying out physical tasks.",
    "Charlie": "Secondary Middleman/Lieutenant. Reports directly to Alpha. Relays threats and instructions, manages operational security, and directs followers Echo, Foxtrot, and Golf.",
    "Delta": "Follower/Logistics Operator. Works under Bravo's command to handle vehicle acquisition and coordinate pickups at designated rendezvous points.",
    "Echo": "Follower/Field Operative. Executes specific tasks for Charlie, including deliveries and evidence destruction ('wipe the hardware clean').",
    "Foxtrot": "Follower/Transporter. Responsible for the transportation and delivery of 'packages' and 'assets', taking orders from both Bravo and Charlie.",
    "Golf": "Follower/Surveillance Specialist. Conducts surveillance on targets, provides status reports, and assists with deliveries."
  },
  "suspicious_terms": {
    "launder the money": "Direct evidence of financial crime, specifically the act of concealing the illegal origin of funds.",
    "offshore account": "A common tool for money laundering and tax evasion, used to hide assets from authorities.",
    "untraceable vehicle": "Indicates the use of stolen or unregistered vehicles to evade identification and tracking by law enforcement during criminal acts.",
    "burner phone": "A disposable, anonymous phone used to prevent communications from being traced back to the user.",
    "loose end": "A euphemism for a person who knows too much and poses a threat to the operation, implying they need to be silenced, possibly through violence.",
    "neutralize the threat": "Coded language for eliminating a problem or a person, suggesting violent intent.",
    "the goods / package / asset": "Coded terms used to refer to illicit items being transported, such as narcotics, weapons, or stolen property.",
    "clean documents": "Refers to forged or fraudulent identity documents, likely for creating new identities or smuggling.",
    "wipe the hardware": "An instruction to destroy digital evidence on electronic devices to obstruct a potential investigation."
  },
  "intent_classification": "Organized Crime / Criminal Enterprise",
  "confidence_score": 0.98,
  "evaluation_metrics": {
    "bleu_score": 0.0,
    "rougeL_f1": 0.079,
    "central_actor": "alpha",
    "central_actor_detected": 1
  }
}
```


Fig. 6: Structured Forensic Report Generated by Multimodal LLM Analyzer for Dataset 1 considering all modalities

Based on the analysis of the generated reports, it can be concluded that the developed model successfully combines structural data with contextual evidence. Individuals who possess high centrality scores in the network are viewed as dominating players and are allocated positions like coordinators and masterminds. On the other hand, low connectivity is associated with the designation of individuals as support players.

From these observations, it can be concluded that the model assimilates the influence dynamics derived from the graph and converts them into behavioral descriptions. Besides, the model detects relevant or suspicious context-based terminology and explains its significance.

4.5 Evaluation Metrics

There are two major factors for the assessment, namely: (i) the degree of similarity between the generated reports and the chat dialogues – in terms of the BLEU score and ROUGE-L score, and (ii) the accuracy of identifying the central actors.

4.5.1 BLEU Score

BLEU is a metric used to assess the overlap of n-grams in the generated output and the reference text:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{9}$$

The BLEU scores are found to be low in all scenarios in this study. This is not unexpected since the generated sentences can be considered free-form natural language output, while the reference sentences are shorter and may use different wording. Even when the meaning is semantically accurate, different formulations result in a low n-gram match, which causes low BLEU scores. Thus, BLEU mainly indicates lexical similarity.

4.5.2 ROUGE-L Score

ROUGE-L scores for LCS (longest common subsequence) between generated and reference texts:

$$ROUGE\text{-}L = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{Recall + \beta^2 \cdot Precision} \tag{10}$$

Unlike BLEU, ROUGE-L offers an improved measure of structural and semantic similarity. Stable ROUGE-L scores under different conditions show consistency in the structural and semantic aspects of the generated texts despite different wordings.

4.5.3 Central Actor Detection

Central Actor Detection checks the capability of the proposed framework to detect the most influential individual in the network of communication. This measure compares the detected central actor based on the generated report and the real central actor using graph metrics. High scores for this measure show successful integration of the network structure information by the model.

4.6 Discussion

4.6.1 Comparison of Model Architectures

A Base model was also employed for comparison purposes. The Base model adopts a straightforward prompt-based method with Gemini 2.5 Pro, which offers multimodal inputs directly into the language model in their raw form from chat conversations, including textual, auditory, and visual inputs. No structural transformation occurs here, and the entire operation is based purely on reasoning by the model.

However, the proposed model employs another level of analysis, which is based on the concept of graph-based network analysis. The chat data will be converted into structured outputs, whereby nodes and weighted edges will represent the participants and interactions, respectively. Centrality analyses through Hub and Authority scores will be used to reveal influence networks and role hierarchies. These structured outputs will be incorporated into the input prompt alongside the multimodal evidence. Therefore, this design makes it possible for the LLM to reason with the help of semantic context and relational graph structure information simultaneously.

4.6.2 Comparison of Generated Forensic Reports

As shown in Figure 7, it provides an example of a structured forensic report produced by Base model. The other reports will be presented in supplementary material - Appendix A. While a structured response is induced with prompting, the lack of any graph-based information makes the identification of the roles inconsistent and inaccurate at times. The model uses only conversational hints to establish importance, causing mistakes in identifying the most frequent individuals as the main role players. However, the proposed model receives explicit centrality markers from graphs, making it easier for it to identify different roles, namely the Masterminds, Middlemen, and Followers.

```
{
  "summary": "The evidence points to a well-structured criminal organization involved in multiple illicit activities. The group, led by 'Alpha', uses a clear hierarchy and coded language to coordinate operations. Activities include money laundering through offshore accounts, smuggling of 'packages' or 'assets' using 'untraceable vehicles', and coercion/intimidation ('firm reminder', 'by any means necessary'). The operation involves surveillance of a 'target', procurement of 'tools' (likely weapons or criminal equipment), and sophisticated counter-surveillance measures like using burner phones and wiping hardware. The presence of a firearm in the evidence suggests the group is armed and potentially violent. A 'complication' with a vehicle, possibly captured by a traffic camera, indicates a potential operational failure.",
  "roles": {
    "Alpha": "The leader or mastermind. Issues commands, authorizes operations, manages high-level finances (money laundering), and enforces strict security protocols.",
    "Bravo": "A high-ranking lieutenant and operational manager. Acts as a direct subordinate to Alpha, relaying orders and managing logistics, procurement of 'tools' and 'documents', and financial transactions.",
    "Charlie": "An enforcer and mid-level manager. Responsible for carrying out intimidation and threats. Manages field operatives and coordinates logistics like vehicle sourcing and hardware destruction.",
    "Delta": "A logistics field operative. Executes tasks assigned by Bravo, such as coordinating pickups, acquiring vehicles, and managing equipment at rendezvous points.",
    "Echo": "A specialist field operative. Handles the hands-on aspects of the operation, including 'delivery of goods', securing 'assets', and wiping digital evidence. Reports directly to Charlie.",
    "Foxtrot": "A logistics and transport operative. Arranges schedules, transports 'assets' and 'packages', and sources vehicles and 'clean documents'.",
    "Golf": "A surveillance operative. Conducts surveillance on the 'target', reports status updates to leadership, and assists in logistical tasks like deliveries."
  },
  "suspicious_terms": {
    "Asset": "A person or high-value item being transported, possibly a kidnap victim or contraband.",
    "Package / Goods / Merchandise": "Euphemisms for illicit items being moved, such as drugs, weapons, or stolen property.",
    "Untraceable vehicle": "A stolen vehicle or one with falsified plates intended to be used in a crime to evade identification.",
    "Clean documents": "Forged or illicitly obtained identification, passports, or other official papers.",
    "Launder the money": "The process of concealing the origins of illegally obtained money.",
    "Loose end": "A potential witness or liability to the operation who needs to be silenced or eliminated.",
    "Tools / Equipment": "Coded language for weapons, breaking-and-entering implements, or other gear needed for a criminal act.",
    "Wipe the hardware": "The act of destroying digital evidence on computers, phones, or other electronic devices.",
    "Rendezvous point": "A pre-arranged, secret meeting location for operatives to exchange goods, information, or vehicles.",
    "Firm reminder / Send a clear message": "Coded instructions for acts of intimidation, threats, or violence against an individual."
  },
  "intent_classification": "Organized Crime (Smuggling, Money Laundering, Coercion)",
  "confidence_score": 0.95,
  "evaluation_metrics": {
    "bleu_score": 0.0,
    "rougeL_f1": 0.096,
    "central_actor": "alpha",
    "central_actor_detected": 1
  }
}
```


Fig. 7: Structured forensic report generated by the base model for Dataset 1 (Text + Audio + Image)

4.7 Findings Across Datasets

4.7.1 Results of Proposed Model

The outcomes of the Multimodal LLM-Based Forensic Model are summarized as follows:

Table 5 summarizes the overall performance of the Multimodal LLM-Based Forensic Reasoning model across various modality conditions. The BLEU and ROUGE-L scores indicate the extent of textual alignment, while the Central Actor Accuracy metric evaluates the model's capability to detect key participants in chat conversations.

Table 5: Overall Performance of Multimodal LLM-Based Forensic Reasoning Model

Condition	BLEU	ROUGE-L	Central Actor Detection
Audio	0.0000	0.0811	0.875
Image	0.0000	0.0570	1.000
Text	0.0001	0.0878	0.875
Image+Audio	0.0000	0.0804	1.000
Text+Audio	0.0000	0.0873	1.000
Text+Image	0.0001	0.0895	0.750
Text+Image+Audio	0.0001	0.0884	1.000

Table 6 provides an individual performance analysis of the Multimodal LLM-Based Forensic Reasoning model in eight datasets (D1-D8). The BLEU and ROUGE-L scores evaluate the quality of the generated text, whereas the Central Actor Detection metric reflects the model's capability to identify prominent individuals based on each dataset.

Table 6: Dataset-wise Performance of Multimodal LLM-Based Forensic Reasoning Model

Condition	D1	D2	D3	D4	D5	D6	D7	D8
BLEU Score								
Audio	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Image	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Text	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
Image+Audio	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Text+Audio	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Text+Image	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000
Text+Image+Audio	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
ROUGE-L Score								
Audio	0.069	0.106	0.075	0.074	0.077	0.072	0.092	0.084
Image	0.000	0.081	0.052	0.054	0.043	0.051	0.000	0.061
Text	0.073	0.101	0.086	0.075	0.084	0.087	0.103	0.093
Image+Audio	0.077	0.099	0.090	0.077	0.054	0.077	0.078	0.091
Text+Audio	0.082	0.099	0.108	0.077	0.072	0.062	0.102	0.096
Text+Image	0.090	0.096	0.103	0.079	0.079	0.075	0.098	0.096
Text+Image+Audio	0.079	0.096	0.081	0.102	0.084	0.098	0.078	0.089
Central Actor Detection								
Audio	1	1	1	1	1	0	1	1
Image	0	1	1	1	1	1	0	1
Text	1	0	1	1	1	1	1	1
Image+Audio	1	1	1	1	1	1	1	1
Text+Audio	1	1	1	1	1	1	1	1
Text+Image	1	1	1	1	1	1	0	0
Text+Image+Audio	1	1	1	1	1	1	1	1

4.7.2 Results of Base Model

Table 7 shows the overall performance of the Base Multimodal LLM model under varied modality conditions. The BLEU and ROUGE-L metrics gauge the textual similarity between the generated reports and the reference report, while the Central Actor Accuracy metric evaluates the model's efficacy in detecting central actors without graph-based reasoning.

Table 7: Overall Performance of Base Multimodal LLM

Condition	**BLEU**	**ROUGE-L**	**Central Actor Detection**
Audio	0.000250	0.088625	0.000
Image	0.000000	0.054750	0.000
Text	0.000125	0.077500	0.625
Image+Audio	0.000125	0.089750	0.000
Text+Audio	0.000000	0.080375	0.750
Text+Image	0.000000	0.089375	1.000
Text+Image+Audio	0.000000	0.089875	0.875

In Table 8 Dataset-wise performance of the Base Multimodal LLM across eight datasets is displayed. The table highlights ROUGE-L scores for text alignment and Central Actor Detection accuracy, illustrating variability and limitations of the Base model in the absence of graph-informed reasoning.

Table 8: Dataset-wise Performance of Base Multimodal LLM

Condition	D1	D2	D3	D4	D5	D6	D7	D8
BLEU Score								
Audio	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Image	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Text	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000
Image+Audio	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Text+Audio	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Text+Image	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Text+Image+Audio	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ROUGE-L Score								
Audio	0.090	0.130	0.091	0.064	0.067	0.081	0.096	0.090
Image	0.050	0.063	0.066	0.072	0.036	0.045	0.047	0.059
Text	0.075	0.093	0.096	0.092	0.060	0.078	0.049	0.077
Image+Audio	0.099	0.104	0.099	0.089	0.064	0.072	0.096	0.095
Text+Audio	0.078	0.104	0.053	0.084	0.055	0.093	0.082	0.094
Text+Image	0.105	0.115	0.090	0.083	0.060	0.086	0.063	0.113
Text+Image+Audio	0.096	0.090	0.106	0.098	0.061	0.084	0.093	0.091
Central Actor Detection								
Audio	0	0	0	0	0	0	0	0
Image	0	0	0	0	0	0	0	0
Text	0	1	0	1	1	1	0	1
Image+Audio	0	0	0	0	0	0	0	0
Text+Audio	1	1	0	1	1	1	0	1
Text+Image	1	1	1	1	1	1	1	1
Text+Image+Audio	1	1	1	1	1	1	1	0

4.7.3 BLEU Score Analysis

In both cases, BLEU scores continue to be low. It is a normal occurrence and not something that reflects poorly on the performance of the model. In Base, descriptions of images are created without any constraints and can end up being paraphrases or summaries, which do not match the vocabulary of the chat.

The Multimodal LLM-Based Forensic Model exhibits similarly low BLEU scores. However, this is due to the same limitation of BLEU: it penalizes variation in phrasing even when semantic meaning is preserved. Since both models generate descriptive forensic narratives rather than extractive summaries, lexical overlap is minimal. This means that BLEU is not a good measure of semantic accuracy for this task.

4.7.4 ROUGE-L Score Analysis

The ROUGE-L metrics give a more sensible comparison between the two systems. The Base system gives moderate ROUGE-L metrics, reflecting partial structural

coherence with the reference texts. Nevertheless, its results fluctuate based on dataset and environment changes. The Multimodal LLM-Based Forensic Reasoning system gives higher and stable ROUGE-L metrics for the multimodal settings. This is because the system gets structural inputs from the graph analysis, which facilitates the production of more coherent summaries by the LLM. The use of relational data cuts down ambiguity and enables coherence of the narration with the reference texts.

4.7.5 Central Actor Detection

One main difference between the two approaches can be found in the determination of the central actors in the communication process. This is because the Base approach determines the role of actors only based on the content of the conversation without knowing about the pattern of interaction among actors.

The problem here can be observed when the Base approach does not manage to determine the right central actors, despite being equipped with all the modalities. This happens because there is no clear information regarding the centrality of certain actors. Conversely, the Multimodal LLM-Based Forensic Model uses centrality metrics derived from graphs of communications. With its use of hub metrics and graph structure, the model clearly understands the importance of influential nodes, which helps in achieving substantially improved accuracy in detecting central nodes. It is worth mentioning that even in a unimodal scenario, the Multimodal LLM-Based Forensic Model continues to perform impressively, while the Base model suffers more degradation.

4.7.6 Overall Findings

The comparative analysis highlights that incorporating graph-based structural information significantly improves interpretability and reasoning. The Multimodal LLM-Based Forensic Reasoning model achieves more consistent performance across different datasets and conditions, demonstrating its robustness and adaptability. Central actor detection is substantially more reliable in the Multimodal LLM-Based Forensic Reasoning model due to the explicit use of centrality measures, which enhances the accuracy of identifying key entities in forensic investigations. Furthermore, the analysis shows that multimodal inputs alone are insufficient without structured reasoning, as evidenced by the performance of the Base model, which lacks integrated graph-based and reasoning mechanisms and therefore produces less reliable results.

Overall, the results demonstrate that the proposed Multimodal LLM-Based Forensic Reasoning framework effectively bridges the gap between raw multimodal perception and structured analytical reasoning, making it more suitable for forensic investigation tasks compared to Base LLM approaches.

5 Conclusion

This paper introduces a two-step multimodal forensic protocol that combines the graph-based structural analysis with the LLM-based behavioral interpretation of the criminal intent in chat history. Module 1 builds an influence-weighted directed graph in which the HITs algorithm is used to infer node roles providing a measurable perspective of communication hierarchy. Module 2 builds on this to interpret the semantic, psychological dimensions of the same network with a Large Language Model, and generate an explainable forensic report with highlight suspicious terms and intent classification and confidence scores. The synthetic datasets used to evaluate the system experimentally reveal that it is capable of detecting organizational hierarchies, implicit threats, and generating coherent investigation summaries. In contrast to the conventional SNA or keyword-based systems, our system offers a twofold benefit, i.e. numerical traceability based on hub-authority measurements, and linguistic explanations based on transformer-based reasoning.

Further research will further expand the framework to real-world encrypted messaging databanks and be multilingual in nature and adaptive to fine-tuning emotion and sentiment analysis. More modules on tracking temporal evolution, correlation of cross-chat, multimedia evidence linking will give more strength to its use by digital-forensic experts and law-enforcement agencies.

Declarations

– Funding : The author(s) received no financial support for the research, authorship, and/or publication of this article.
– Competing Interests : The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
– Ethical and informed consent for data used : This article does not contain any studies with animals performed by any of the authors.
– CRediT authorship contribution statement: All authors contributed equally. In particular, Aashi Manker, Juweriah Abdul Raheem Mohammed, and Siddhartha Mourya: Investigation, Data curation, design, and experimentation. Sairam Utukuru: Methodology, Validation, Analysis, Supervising, and Data interpretation
Yelleti Vivek: Conceptualization, Methodology, Formal verification of the model
– Availability of data and materials: The datasets used in this study are taken from public domain and the appropriate URLs have been cited in the text.

Bibliography

[1] Paige H Adams and Craig H Martell. Topic detection and extraction in chat. In *2008 IEEE international conference on Semantic computing*, pages 581–588. IEEE, 2008.

[2] Khaoula Ait Rai, Mustapha Machkour, and Jilali Antari. Influential nodes identification in complex networks: a comprehensive literature review. *Beni-Suef University Journal of Basic and Applied Sciences*, 12(1):18, 2023.

[3] Rabeah Al-Zaidy, Benjamin CM Fung, Amr M Youssef, and Francis Fortin. Mining criminal networks from unstructured text documents. *Digital Investigation*, 8(3-4):147–160, 2012.

[4] Ashwaq Almorjan, Mohammed Basheri, Miadah Almasre, and Mona Alkhozae. Cyber threat detection and mitigation on social media using large language models (llms): A systematic review. In *2025 10th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, pages 1–10. IEEE, 2025.

[5] Musaad Alzahrani. Robust and efficient log anomaly detection: A hybrid id-semantic approach for evolving systems. *Information and Software Technology*, page 108105, 2026.

[6] Tarique Anwar and Muhammad Abulaish. A social graph based text mining framework for chat log investigation. *Digital Investigation*, 11(4):349–362, 2014.

[7] Amiya Biju, Aparna Sajeev, Arunya Bimal, Muhammed Fais, and Sreela Chandran. Ai surveillance system for detecting criminal intent in voice or chat communication. *INTERNATIONAL JOURNAL OF ENGINEERING DEVELOPMENT AND RESEARCH*, 13(4):710–716, 2025.

[8] Vanessa Clairoux-Trepanier, Isa-May Beauchamp, Estelle Ruellan, Masarah Paquet-Clouston, Serge-Olivier Paquette, and Eric Clay. The use of large language models (llm) for cyber threat intelligence (cti) in cybercrime forums. *arXiv preprint arXiv:2408.03354*, 2024.

[9] Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. Learning to detect conversation focus of threaded discussions. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 208–215, 2006.

[10] Vladimir Gligorijevic, Marcin Skowron, and Bosiljka Tadic. Directed networks of online chats: Content-based linking and social structure. In *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems*, pages 725–730. IEEE, 2012.

[11] Marah Radi Hawa, Majdi Owda, and Amani Yousef Owda. Enhancing digital investigation: the role of generative ai (chatgpt) in evidence identification and analysis in digital forensics. In *2025 12th International Conference on Information Technology (ICIT)*, pages 19–26. IEEE, 2025.

[12] Farkhund Iqbal, Benjamin CM Fung, and Mourad Debbabi. Mining criminal networks from chat log. In *2012 IEEE/WIC/ACM International Conferences*

on Web Intelligence and Intelligent Agent Technology, volume 1, pages 332–337. IEEE, 2012.

[13] Yidnekachew Worku Kassa, Joshua Isaac James, and Elefelious Getachew Belay. Intention recognition in digital forensics: systematic review. *Preprints*, 2024.

[14] Kyung-Jong Kim, Chan-Hwi Lee, So-Eun Bae, Ju-Hyun Choi, and Wook Kang. Large language model-based digital forensic framework for crime investigation. *Forensic Science International: Digital Investigation*, 54:301939, 2025.

[15] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[16] Yi-Anson Lam, Kam-Pui Chow, and Siu-Ming Yiu. Enhancing digital evidence collection and analysis using built-in instant messaging solutions and large language models. In *IFIP International Conference on Digital Forensics*, pages 61–76. Springer, 2025.

[17] R. Li, Y. Shang, and W. Zhang. Improvement of hits algorithms based on weights. In *Proceedings First IEEE International Conference on Cognitive Informatics*, pages 169–174. IEEE, 2002.

[18] Snigdha Luthra, Rakesh Sharma, and Meenu Gupta. Hubs and authorities in social network analysis using hits algorithm combined with sentiment score. In *International Conference on MAchine inTelligence for Research & Innovations*, pages 145–155. Springer, 2023.

[19] Abdur Rahman MA Basher and Benjamin CM Fung. Analyzing topics and authors in chat logs for crime investigation. *Knowledge and information systems*, 39(2):351–381, 2014.

[20] V. Miller and C. Morselli. Structuring criminal networks: The role of communication in the organization of crime. *Journal of Contemporary Criminal Justice*, 27(3):300–316, 2011.

[21] Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. *arXiv preprint arXiv:2308.14683*, 2023.

[22] Ermelinda Oro, Clara Pizzuti, Nicola Procopio, and Massimo Ruffolo. Detecting topic authoritative social media users: a multilayer network approach. *IEEE Transactions on Multimedia*, 20(5):1195–1208, 2017.

[23] Levi Jeremiah Pittman. Analysis of the application of large language models in criminal email detection. Master's thesis, University of Central Arkansas, 2025.

[24] Roshan Rabade, Nishchol Mishra, and Sanjeev Sharma. Survey of influential user identification techniques in online social networks. In *Recent Advances in Intelligent Informatics: Proceedings of the Second International Symposium on Intelligent Informatics (ISI'13), August 23-24 2013, Mysore, India*, pages 359–370. Springer, 2014.

[25] Khan Sameera and Pinki Vishwakarma. Cybercrime: To detect suspected user's chat using text mining. In *Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018, Volume 1*, pages 381–390. Springer, 2018.

[26] Kanti Singh Sangher, Archana Singh, and Hari Mohan Pandey. Lstm and bert based transformers models for cyber threat intelligence for intent identification of social media platforms exploitation from darknet forums. *International Journal of Information Technology*, 16(8):5277–5292, 2024.

[27] Paria Sarzaeim. *Design and development of an LLM-based framework for crime classification and prediction*. PhD thesis, University of Nevada, Las Vegas, 2024.

[28] Mark Scanlon, Frank Breitinger, Christopher Hargreaves, Jan-Niclas Hilgert, and John Sheppard. Chatgpt for digital forensic investigation: The good, the bad, and the unknown. *Forensic Science International: Digital Investigation*, 46:301609, 2023.

[29] Abdur R Shahid, Syed Mhamudul Hasan, Malithi Wanniarachchi Kankanamge, Md Zarif Hossain, and Ahmed Imteaj. Watchovergpt: A framework for real-time crime detection and response using wearable camera and large language model. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 2189–2194. IEEE, 2024.

[30] Viacheslav Shevchenko, Oleh Zvonarov, Leonid Shcherbyna, Yana Lutsenko, and Volodymyr Pyvovarov. Using ai to counter criminal offences against the will, honour, and dignity of a person. *Journal of Theoretical and Applied Information Technology*, 103(1), 2025.

[31] Filipe Silva, Rodrigo Rocha Silva, and Jorge Bernardino. A survey of generative ai for detecting pedophilia crimes. *Applied Sciences*, 15(13):7105, 2025.

[32] Ville H Tuulos and Henry Tirri. Combining topic models and social networks for chat data mining. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, pages 206–213. IEEE, 2004.

[33] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 71–80, 2012.

[34] Yifan Yao, Baojuan Wang, Jinhao Duan, Kaidi Xu, ChuanKai Guo, Zhibo Eric Sun, and Yue Zhang. The imitation game: Using large language models as chatbots to combat chat-based cybercrimes. *arXiv preprint arXiv:2512.21371*, 2025.

[35] Lei Zhang et al. Application of large language models in public security incident element extraction. *Applied Sciences*, 14(15):7819, 2024.

[36] Yiming Zhang, Pengxiao Li, Sanchuan Guo, Ming Zhang, and Xi Zhang. A key person extraction method for chat groups based on structure-aware graph neural network. In *International Conference on Autonomous Unmanned Systems*, pages 223–236. Springer, 2023.