Large Language Models in Modern Forensic Investigations: Harnessing the Power of Generative Artificial Intelligence in Crime Resolution and Suspect Identification

Anastasios Nikolakopoulos [1], Spyridon Evangelatos [1,2], Eleni Veroni [1,2],
Konstantinos Chasapas [1], Nikolaos Gousetis [1], Apostolos Apostolaras [3,4],
Christos D. Nikolopoulos [2] and Thanasis Korakis [3,4]
[1] *Research & Innovation Development Department, Netcompany-Intrasoft S.A.*, Luxembourg, Luxembourg
[2] *Department of Electronics Engineering, Hellenic Mediterranean University*, Crete, Greece
[3] *Department of Electrical and Computer Engineering, University of Thessaly*, Volos, Greece
[4] *Centre for Research & Technology, Hellas,* Volos, Greece
Emails: {Anastasios.Nikolakopoulos@netcompany.com, tasosnikolakop@mail.ntua.gr},
Spyros.Evangelatos@netcompany.com, Eleni.Veroni@netcompany.com, Konstantinos.Chasapas@netcompany.com,
Nikolaos.Gousetis@netcompany.com, apaposto@iti.gr, cnikolo@hmu.gr, korakis@iti.gr

***Abstract*—Large Language Models (LLMs) have recently captured the attention of the scientific community. Since the global launch of LLM-based chatbots in late 2022, the field has witnessed a rapid increase in interest from researchers, technology providers and citizens alike. With its wide-ranging applicability, Generative Artificial Intelligence (GenAI) has the potential to impact various aspects of society, from improving communication and accessibility to transforming industries such as healthcare, education and security. More specifically, in the field of Forensic Science, LLMs could offer significant benefits assisting Law Enforcement Agencies (LEAs) and Forensic Practitioners in crime investigations. This paper proposes the implementation of a Retrieval Augmented Generation (RAG) LLM, trained with criminology data, to provide swift and actionable insights into specific incidents, thereby enhancing Forensic Data Analysis and facilitating the daily operations of LEAs.**

***Index Terms*—Generative Artificial Intelligence, Large Language Models, Security, Forensics, Law Enforcement Agencies, Forensic Data Analysis, Biometric Data**

I. Introduction

Over the last few years, GenAI has become one of the most popular subjects within the research community. At the same time, technology providers across leading markets have been investing in GenAI to deliver innovative solutions tailored to the needs of their customers. Several definitions have been proposed for GenAI, with the most prominent one describing it as a branch of AI that analyzes training examples, learns their patterns and distributions from vast datasets, and then autonomously generates synthetic artifacts (e.g., text, images, video and audio) based on those patterns, leveraging the advances in Deep Learning (DL) [1], [2].

A sub-category of GenAI is that of Large Language Models (LLMs), artificial neural networks based on the transformer architecture. Transformer models, a key component of LLMs, are neural networks designed to comprehend context and semantics by analyzing relationships within sequential data, such as the words within a sentence [3]. LLMs are probabilistic natural language models proficient in general-purpose language comprehension and generation. When it comes to text generation, LLMs process natural language input and iteratively forecast the subsequent token or word, thereby generating output that corresponds to the original context [4]. Few of the most well-known LLMs include OpenAI's GPT versions, Google AI's PaLM, Meta (formerly Facebook) AI's LLaMA versions 1 and 2, as well as Mistral AI's Mistral. These are only few of the LLMs that have been published over the past years but it is safe to assume that, despite the open research challenges [5], more are to arrive, especially from prominent technology providers targeting various domains.

Within the LLM architecture, there lies a specific technique of interacting with a given model, the so-called Retrieval Augmented Generation (RAG). RAG is a technique that aims to augment the knowledge of an LLM by incorporating supplementary data using the following process. First, a Large Language Model is enriched through document retrieval. When presented with a query, the system is invoked to fetch the most relevant information. This is typically determined by initially encoding the query and documents into vectors and subsequently identifying the documents with vectors closest to the query vector (in Euclidean norm). Finally, the LLM generates an output, considering both the query and the retrieved documents [6]. The RAG technique offers several key advantages, including the ability to streamline large generative models, accommodate extensive contextual inputs, and streamline the generation process by eliminating certain steps [7].

To develop an AI model capable of reasoning about proprietary data or information introduced after the model's cutoff date, it is essential to enhance the LLM's knowledge by incorporating specific information required for such tasks. As this information is typically restricted to in-house access only, RAG is often leveraged to expand the knowledge of the LLM. An organization can train an LLM on-premises, using proprietary and/or restricted data, and interact with it accordingly. Such a scenario has a high impact, as it can be applied in several use cases.

Forensic science is the application of scientific principles and techniques that analyze evidence in legal contexts, often used in criminal investigations to establish facts and support legal proceedings [12]. A branch of Forensic science, criminalistics, is dedicated to the meticulous collection and analysis of physical evidence obtained from crime scenes. The employment of these scientific methodologies can heavily assist in the investigation and resolution of criminal cases, facilitating the establishment of connections between evidence and potential suspects.

The application of a RAG LLM, trained with data collected from criminal incidents and therefore capable of generating a narration of the crime scene, can have a significant impact on the way forensic investigations are performed today. This paper presents the Investigation Enhancement Model (IEM), a RAG LLM that has been customized to be used by the LEAs during the investigation of crimes. It can provide them with information and updates regarding a given incident, using natural language and reason with them for specific details. Furthermore, it can generate a complete report for this incident and narrate the events to a user (i.e., forensic investigator) in chronological order, enhancing situational awareness and facilitating efficient decision-making.

The paper is organised as follows: Section II summarizes the most recent research on LLMs and GenAI used in forensics. Section III constitutes the core part of this work, presenting an overview of the proposed framework, its architecture, the functionality of the recommender, and the enhancements brought by the incorporation of the GenAI. Section IV lists the implementation challenges and Section V concludes this paper, summarizing the key findings and proposing future work that can advance the overall functionality of the framework.

II. Related Work

While ChatGPT by OpenAI was not the first LLM created, it was the first one to be made accessible by the general public, marking the beginning of a new era for Generative AI. Unlike its predecessors, ChatGPT exhibits remarkable versatility across diverse subjects, rather than offering in-depth knowledge for a single area. Trained comprehensively across multiple knowledge domains, ChatGPT has been demonstrating considerable competence in engaging in both general and specialized discussions [8]. LLMs have progressively become faster and smarter ever since, having emerged as transformative technology with the potential to revolutionize various industries including security and more specifically, forensics. Nowadays, there are many research efforts on the application of LLMs in digital forensics [9] extracted from various electronic devices such as computers, servers, smartphones and IoT-enabled devices. In [10], the impact of ChatGPT (its latest pre-trained LLM, GPT-4) on the field of digital forensics was examined, assessing its capability across several digital forensic use cases including artifact understanding, evidence searching, anomaly detection and incident response.In [11] the potential value of ChatGPT was also assessed but this time in locating evidence for digital investigation, focusing on various identification tasks such as evidence evaluation, language interpretation, conceptual comprehension, phrase and word extraction, and criminal examination of communication patterns. Its findings highlighted the enormous potential that LLMs have in the realm of digital forensics. However, while this technology shows promise in digital investigations, its application in traditional forensics, where the collection, analysis, and presentation of physical evidence are crucial, remains largely unexplored. The proposed solution outlined in this paper represents an initial step towards bridging this gap.

III. Description of the Framework

A. Overview

This paper introduces the Investigation Enhancement Model (IEM), a software component designed to enhance LEAs investigative capabilities in combating terrorism and serious crime through informed decision-making. The IEM integrates a suite of biometric identification technologies to extract actionable intelligence, leveraging data collected from crime scenes. The scope of data collection may vary depending on available sources, as illustrated in Figure 1. For instance, within areas covered by CCTV surveillance, modern physiological and behavioral biometric technologies can analyze footage to conduct facial and gait recognition on suspects [12], [13]. This process also provides insights into additional features of the suspect crucial to investigations, such as clothing details, brand logos, the use of facial coverings, glasses, as well as identifying scars and tattoos. Voice recognition [14] can further aid investigations when audio recordings are available, revealing details about the suspect's identity and spoken language. These biometric traits can then be used to deduce soft biometric attributes such as gender, age, height, race, and ethnicity [15].

In cases involving mobile or personal devices found at crime scenes, behavioral analysis techniques [16] can be deployed to derive intelligence on the device owner's habits, past locations, and recent financial activities. Additionally, pattern recognition algorithms [17] can be used to inspect media files from these devices to identify pertinent patterns, such as photos related to the victim or the crime scene. The resulting metadata (e.g., biometric matching scores, soft biometric descriptors, potential suspect identities, criminal histories, and incident timestamps and locations) are then integrated into the IEM through the Data Fusion layer, together with relevant data from existing biometric databases such as AFIS (Automated Fingerprint Identification System) [18] or ABIS (Automated

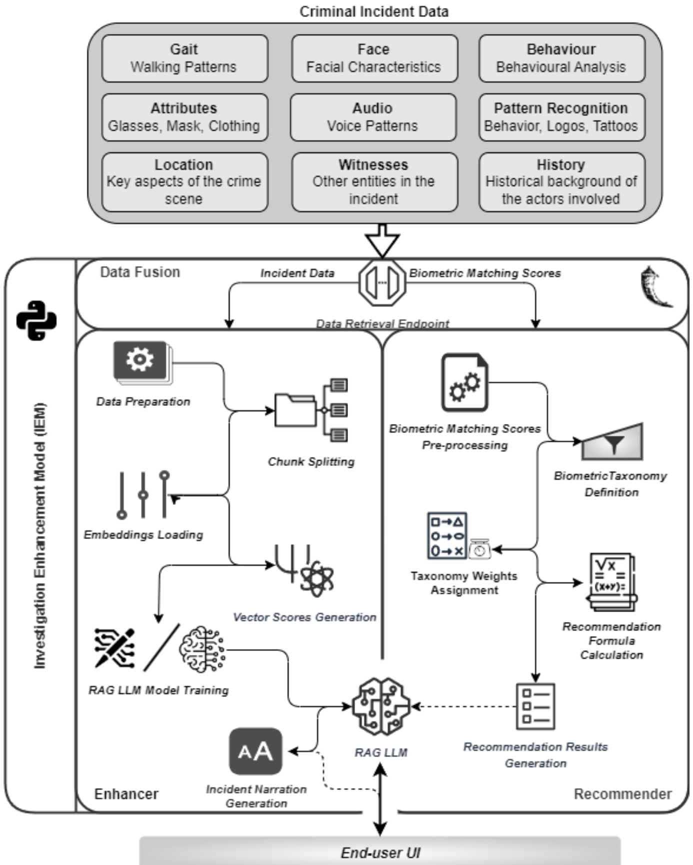

Fig. 1. Investigation Enhancement Model's architecture

Biometric Identification System) [19], [20]. Within this layer, data undergo fusion and classification, and are accordingly routed to the respective component of the IEM.

As depicted in Figure 1, the IEM is composed of two main components, the "Enhancer" and the "Recommender". Through the Data Fusion layer, any incident information is seamlessly transmitted to the "*Enhancer*", initiating the pre-processing stage within the IEM. The creation of GenAI LLM and subsequent narration generation are crafted. Then, if biometric matching scores are obtained, these are directed to the IEM's second internal component, the "*Recommender*". Within this component, the received matching scores are fused considering a predefined biometric taxonomy used for weight assignment based on the reliability of the examined biometric traits, producing a set of recommendations for the identities of the potential suspects. Finally, the generated incident narration, along with any accompanying suspect recommendations, are sent to the framework's user interface (UI). Following the retrieval of the narration, the investigator (e.g., Police Officer or forensic practitioner) can interact with the LLM through prompt engineering to elicit further information or details concerning specific aspects of the incident.

B. Internal Components

1) Enhancer: The *Enhancer* houses a GenAI model that produces a complete narration of the crime, enabling the investigators to request additional information or details regarding the event. This process unfolds through a series of sequential steps, outlined as follows:

- Data Preparation: Process the incoming data using primary techniques such as converting all characters to lowercase, eliminating whitespace, and other relevant adjustments tailored to the specific data types received. Furthermore, each piece of information is classified based on context and prepares the collected data for the subsequent stages. A new dataset is created, based on the incoming data. For the RAG LLM training to succeed, this session has to be thoroughly performed.
- Split text into chunks: Divide the incoming information of text into smaller segments or tokens to efficiently process and analyze information. These chunks enable the model to handle large inputs effectively while mitigating computational constraints.
- Embeddings loading: Convert words or tokens into dense numerical representations, capturing semantic and contextual information. These embeddings enable the model to understand relationships between words and improve its ability to generate coherent and contextually relevant text.
- Vector scores generation: Assign numerical values to different elements of the input text (or tokens), often representing probabilities or relevance scores. The chunks created represent smaller segments of text, and are transformed into the embeddings, dense numerical representations capturing semantic information. These embeddings are then utilized to compute vector scores, which quantify the relevance or probability of different elements within the chunks, enabling the model to generate contextually relevant responses or predictions.
- RAG LLM training: Integrate a retrieval mechanism with the language model during training, enabling it to access external knowledge sources (from the vector scores generated), in order to enhance text generation. This approach enables the model to produce more contextually relevant and informative responses by incorporating information retrieved from the criminal incident data.
- Narration generation: Use the newly trained LLM to produce a narration of the investigation process, combining all the data collected (along with the potential recommendations from the Recommender).

Once the LLM has undergone training using the RAG approach and has finalized the narration generation process, it proceeds to transmit the narration to the frontend, making it accessible to the investigators. In case future data associated with a specific incident emerge, the LLM can undergo further training to incorporate this new information into its knowledge base. This adaptive learning process can be repeated across various cases or incidents, provided that they are properly fed to the LLM as separate and distinct events.

2) Recommender: The second internal component of the Investigation Enhancement Model is the *Recommender*, which is only triggered if the Data Fusion layer has collected suspect biometric matching scores. Biometric matching scores refer to numerical values (expressed as a percentage) generated by a biometric-based identification system, such as facial recognition or fingerprint matching. These scores signify a potential correlation between the biometric sample under analysis and a stored reference sample within a suspect database, after assessing the degree of similarity between the two samples.

Following the fusion of biometric matching scores through a weighted average formula, which dynamically assigns varying weights based on the number and reliability of analyzed biometric modalities, the IEM's Recommender generates a list of potential suspects. The operational steps of the Recommender are as follows:

- Retrieval and pre-processing of matching scores: Proceed to retrieve and pre-process the biometric matching scores from the Data Fusion layer.
- Biometric taxonomy definition: Create a taxonomy of the different biometric modalities (e.g., face, gait, voice, etc.), based on the biometric nature of the incoming matching scores.
- Assignment of taxonomy weights: Assign biometric reliability scores (weights) to the matching scores retrieved, in order to enhance the accuracy of the recommendation process.
- Recommendation calculation: Apply a formula to the matching score/taxonomy weight pairs and proceed to calculate the final recommendations.
- Suspect recommendation with confidence level estimation: Recommend potential suspects and provide their corresponding confidence level, feeding them to the RAG LLM for the incident narration generation.

The biometric taxonomy used for the weights assignment currently relies on [21] and [22]). Reliability scores are subject to dynamic adjustments based on several factors, including the quantity of analyzed biometric modalities and the values of the matching scores generated. These adjustments occur particularly if the received scores are below a predefined threshold.

Finally, the Recommender will feed the results to the RAG LLM enabling the investigators to query the model further about the suspect recommendations, the confidence level and the reasoning behind the produced results, allowing them to gain insight into the identities of the actors of the incident.

Regarding the LLM to use, IEM will be based on a local large language model, which will be trained with the RAG approach, using all available criminal incident data as inputs. Although research for the optimal model is underway, current progress indicates Mistral AI's [23] Mistral 7B as the preferable choice, due to its ability to be deployed fast and be easily customizable, whilst taking a relatively normal amount of disk space. However, extensive research will be further conducted, in order to provide the most suitable LLM for the needs of IEM.

IV. Implementation Challenges

Two main challenges have been identified and need to be addressed, before and during IEM's development. The first is about the computational resources needed for the proper training and final operation of a RAG LLM. Training and fine-tuning a large language model with additional knowledge sources requires considerable computational resources. Acquiring high-quality training data and efficient data storage further adds to the resource challenge, highlighting the need for careful resource management to ensure effective model training and performance.

The second challenge is about mitigating any potential bias in the LLM that may appear during its responses to the end-users. For this purpose, it is highly important to filter the training data, so that they may not contain any kind of information that can be considered as bias or societal prejudices (e.g., for a suspect or group of suspects), for the model not to inadvertently perpetuate these biases in its generated responses. For instance, if the retrieved knowledge contains gender stereotypes or racial prejudices, the model might produce biased or discriminatory outputs, reinforcing harmful societal norms.

Moreover, insufficient training of RAG LLMs can result in inaccurate or nonsensical responses. Without exposure to a well-defined and properly processed dataset, the model may struggle to generate coherent or relevant text. For example, if the model misinterprets retrieved knowledge or fails to understand specific relationships between concepts, it may produce responses that are incorrect or fail to address the user's query effectively. As future work, research on the optimal resource allocation of the RAG LLM during the training phase can be conducted, as well as the proper data preparation, in order to avoid biases and false responses by the model. It should also be noted that the process of fine-tuning will be taken under consideration. However, since fine-tuning is a more complex approach, requiring a deeper examination of the underlying Machine Learning model, model configuration and natural language processing (NLP) will be examined after the evaluation of the RAG method.

V. Conclusion

The IEM is a proposed software tool that can prove to be an important assistant to LEAs and police authorities around the world. Its novel adoption of an LLM, further trained with the RAG approach using crime data, in combination with a method to provide recommendations of potential criminals (where biometric matching scores are available), will allow end-users to easily and seamlessly obtain incident information without the need of querying databases. IEM can inform law enforcement and police officers about specific aspects of a crime using natural language, making the data retrieval process easier. IEM exploits a fraction of GenAI's vast capabilities and bridges the gap between current forensic/criminalistics methodologies and trending software solutions. The development and implementation of the IEM can provide a definitive answer on its feasibility question, taking into consideration the amount of computational resources needed for such a task. The performance analysis of the system can also highlight how an LLM understands and further handles criminal-related information, as well as relevant questions from the end-users. The replies it provides will be analyzed for biases or misinformation.

VI. Acknowledgment

The research leading to these results has been partially funded by the European Union's Horizon Europe research and innovation programme under the project "TENSOR" (Grant Agreement No. 101073920). The content reflects only the authors' views and the European Commission is not responsible for any use that may be made of the information it contains.

References

[1] K. S. Kaswan, *et al.*, "Generative AI: A Review on Models and Applications," *2023 IEEE International Conference on Communication, Security and Artificial Intelligence (ICCSAI)*, Greater Noida, India, pp. 699-704, 2023.

[2] M. Jovanović and M. Campbell, "Generative Artificial Intelligence: Trends and Prospects," in *IEEE Computer*, vol. 55, no. 10, pp. 107-112, Oct. 2022.

[3] K. T. Chitty-Venkata, *et al.*, "Neural Architecture Search for Transformers: A Survey," in *IEEE Access*, vol. 10, pp. 108374-108412, 2022.

[4] Y. Chang, *et al.*, "A Survey on Evaluation of Large Language Models," Association for Computing Machinery, *ACM Transactions on Intelligent Systems and Technology*, New York, NY, USA, 2024.

[5] A. Fan, *et al.*, "Large Language Models for Software Engineering: Survey and Open Problems," *2023 IEEE/ACM International Conference on Software Engineering: Future of Software Engineering (ICSE-FoSE)*, Melbourne, Australia, pp. 31-53, 2023.

[6] P. Lewis, *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 9459–9474, 2020.

[7] P. Zhao, *et al.*, "Retrieval-Augmented Generation for AI-Generated Content: A Survey," arXiv preprint arXiv:2402.19473, 2024.

[8] T. Wu, *et al.*, "A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development," *in IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122-1136, 2023.

[9] X. Du, *et al.*, "SoK: Exploring the State of the Art and the Future Potential of Artificial Intelligence in Digital Fforensic Investigation, " in *Proceedings of the 15th International Conference on Availability, Reliability and Security (ARES '20)*, Association for Computing Machinery, New York, NY, USA, Article 46, 1–10, 2020.

[10] M. Scanlon, *et al.*, "ChatGPT for Digital Forensic Investigation: The Good, the Bad, and the Unknown," *in Forensic Science International: Digital Investigation*, Vol. 46, Supplement, 301609, ISSN 2666-2817, 2023.

[11] S. Sakshi and M. Wadhwa, "Enhancing Digital Investigation: Leveraging ChatGPT for Evidence Identification and Analysis in Digital Forensics," *in 2023 IEEE International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Greater Noida, India, pp. 733-738, 2023.

[12] M. Jacquet and C. Champod, "Automated face recognition in forensic science: Review and perspectives", in *Forensic science international*, 307, 110124, 2020.

[13] A. Sepas-Moghaddam and A. Etemad, "Deep Gait Recognition: A Survey," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 264-284, 2023.

[14] A. Neustein and H. Patil, "Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism," Springer Publishing Company, Incorporated, 2011.

[15] F. Alonso-Fernandez, *et al.*, "Soft-Biometrics Estimation In the Era of Facial Masks," *IEEE International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, pp. 1-6, 2020.

[16] M. Papaioannou, *et al.*, "Behavioral Biometrics for Mobile User Authentication: Benefits and Limitations," *2023 IFIP Networking Conference (IFIP Networking)*, Barcelona, Spain, pp. 1-6, 2023.

[17] K. Barmpatsalou, T. Cruz, E. Monteiro and P. Simoes, "Mobile Forensic Data Analysis: Suspicious Pattern Detection in Mobile Evidence," in *IEEE Access*, vol. 6, pp. 59705-59727, 2018.

[18] Z-H. Usmani, *et al.*, "How to Build an Automated Fingerprint Identification System," *in 2013 IEEE International Symposium on Biometrics and Security Technologies*, Chengdu, China, pp. 40-47, 2013.

[19] V. Rai, *et al.*, "Automated Biometric Personal Identification-Techniques and Applications," *in 2020 IEEE 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, pp. 1023-1030, 2020.

[20] Vandana and N. Kaur, "A Study of Biometric Identification and Verification System," *in 2021 IEEE International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India, pp. 60-64, 2021.

[21] A. Dube, *et al.*, "A Framework for Evaluation of Biometric Based Authentication System," *2020 IEEE 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, 2020, pp. 925-932, 2020.

[22] D. Y. Hwang, *et al.*, "Evaluation of the Time Stability and Uniqueness in PPG-Based Biometric System," in *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 116-130, 2021.

[23] A. Q. Jiang *et al.*, "Mistral 7B," arXiv eprint 2310.06825, 2023.