The Quality Assessment of LLM in Digital Forensics

Federico Costantini[1,*,†], Fausto Galvan[1,†], Pier Luca Montessoro[3,†], Francesco Crisci[2,†] and Luca Baron[1,†]

[1]Dipartimento di Scienze Giuridiche, Università degli Studi di Udine, Via Tomadini 3, 33100 Udine, Italy
[2]Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di Udine, Via Tomadini 30/a, 33100 Udine, Italy
[3]Dipartimento Politecnico di Ingegneria e Architettura, Università degli Studi di Udine, Via delle Scienze 206, 33100 Udine, Italy

Abstract

The utilisation of Artificial Intelligence in investigative contexts facilitates the optimisation of the analysis of digital artefacts, which can be collected in a variety of formats, including text, printouts, images, and videos, as well as from diverse sources, such as datasets, networks, and devices. More recently, while the adoption of LLM (Large Language Models) appears to hold significant potentials, it is also accompanied by emerging concerns relating to its technological limitations, including distortions and hallucinations, which can hinder law enforcement and compromise the balance between the rights of the accused and the power of the judicial authorities. Notwithstanding the activities prohibited by Art. 5 of the AI ACT, the fact is that LLM technologies are already assuming a pivotal role in investigations, yet are still inadequate the principles, methodologies and tools allowing courts to assess accuracy and precision of the language models before implementing them. The objective of this research project is to develop a methodology that can be adopted for the assessment and comparison of LLMs in performing analytical processes. To this end, a synthetic dataset of electronic evidence is generated from a fictitious criminal scenario and fed to the agent with engineered prompting techniques.

Keywords

Artificial intelligence, Digital Forensics, Large Language Models, Quality of information

1. Towards an AIQA (Artificial Intelligence Quality Assessment)

The profound transformations brought about by the development and diffusion of algorithmic systems and artificial intelligence are affecting an ever-increasing number of areas of human activity[1]. It is therefore only natural that these systems are assuming an increasingly central role in the administration of justice, raising formidable questions and necessitating a reconsideration of some of the criminal justice system's fundamental pillars — both substantive and procedural. At the heart of this issue is the potential use of AI-based tools for extracting and managing evidences. The stakes are exceedingly high, as results from academic discussion. Indeed, since the adoption of AI system in the judicial sector is included in Annex III as "High Risk" activity of the AI ACT, experts insist in restricting the role of AI to limited activities [2], while others highlight the potentials of such tools [3].

In scientific literature, we find the expression "AI-driven", "AI-enhanced" or "AI-powered", meaning that a wide scope of tasks can be performed by AI, yet not completely substituting the human operator. Instead, the role assigned to the artificial agent, although sensitive, remains one of support and assistance such as to identify offensive or harmful language [4], to scan images [5] or the dark web [6], or networks [7], perform other complex analysis [8, 9], or assist in legal drafting [10].

In our previous research, we investigated the problem of "Information Quality" in the field of Digital Forensics, developing a formal approach with the aim of developing a theoretically grounded operational

AI4EVIR: Workshop on AI for evidential reasoning, December 9, 2025, Turin, Italy.
*Corresponding author
†This contribution was carried out within the Project "Cross Border Digital Forensics", Departmemt of Law, University of Udine. Although it is the result of a joint effort, the single paragraphs can be attributed as follows: Costantini: 1; Galvan 2.1; Crisci: 4; Montessoro: 3; Baron: 2.2
federico.costantini@uniud.it (F. Costantini); galvanfausto14@gmail.com (F. Galvan); pierluca.montessoro@uniud.it (P. L. Montessoro); francesco.crisci@uniud.it (F. Crisci); luca.baron@uniud.it (L. Baron)
0000-0003-2168-5523 (F. Costantini); 0000-0002-3412-2837 (F. Galvan); 0000-0003-2536-0603 (P. L. Montessoro); 0000-0003-2563-9612 (F. Crisci); 0000-0002-5428-4948 (L. Baron)

model [11]. If we consider the concept of “quality” in a technical sense, defined as the capacity of a given object to meet a certain purpose, then it is possible to identify a set of requirements that it - in our case, the artifact generated by forensic acquisition - must possess. In this way, it is possible to determine whether and to what extent the expectations of the actors involved – primarily the parties of the legal proceedings – are met. In this sense, a general formula was defined that allowed one to measure the reliability of the general information contained in a heterogeneous set of devices [12].

On this regard, it is important to remember that international standards have been developed for the sharing of digital evidence, as well as legal ontologies specifically dedicated to this sector, such as CASE[1], which contribute to spreading the adoption of AI in the forensic field and increasing the level of reliability.

However, if it is widely accepted that “data” and “algorithms” are two sides of the same coin, which we can generically name "information", their “quality” must be assessed using different criteria. First of all – but we will see this later on – very often assessments can only be made indirectly, because the technologies used do not offer transparency in their processing. Secondly, it should be noted that AI is, on the one hand, an “object”, similar to the digital data being observed, and on the other hand, an “agent”, like the human being who observes it. The automatism of the activity that AI is called upon to perform reveals not only the essence of its instrumentality – the fact that it is used to make data processing more efficient – but also the seed of issues concerning the balance between the operation of the artificial agent and the surveillance of the human being. In other words, AI can perform its tasks with "varying degrees of autonomy" – as also found in the regulatory definition of Art. 3(1) of the “AI ACT” Regulation[2] – and, consequently, external expectations – in terms of “quality”, which is what interests us most specifically here – depend on a very heterogeneous set of factors, including, in particular, the level of “trustworthiness” established by the human being with whom it interacts or in the reference context. In this sense, it becomes crucial to establish criteria for determining also the "quality" of algorithms.

In this work, we intend to present a research proposal that addresses the "Information Quality" in the implementation of Artificial Intelligence in Digital Forensics, with the aim of establishing an AIQA (Artificial Intelligence Quality Assessment) framework. To this end, we created a pilot experiment in which a "criminal scenario" is created based on official reports, and it is used to generate a synthetic dataset, which has been fed to LLM agents in order to assess their performances according to given benchmarks.

2. Experimental methodology

Recent contributions have explored the possibility of using LLMs in the forensic field for purposes other than strictly evidentiary or investigative ones, especially for for training and educational purposes [13]. In some cases, they have been used for creating quasi-real criminal scenarios to be used as a “test bed” to evaluate artificial agent performances [14]. In our research, we intend to follow similar paths.

2.1. Scenario description

The narrative used as a starting point for this work refers to online fraud. This is a constantly growing phenomenon in which fraud is perpetrated by offering the opportunity to invest in funds or virtual currencies [15, 16]. In the most common *modus operandi*, victims are “lured”, for example, by believing the promises of lucrative investments, via physical contacts, phone calls, direct messages on social

[1]The CASE (Cyber-investigation Analysis Standard Expression) ontology (https://caseontology.org) allows artifacts and their properties to be thoroughly described, facilitating sharing, transmission, and automatic processing.

[2]Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), OJ L, 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj.

networks, sometimes with aggressive campaigns involving "deep fake" videos that can feature the appearances of celebrities or even politicians, as happened recently with the Italian Prime Minister[3].

Fraudulent activities are perpetrated using a variety of methods, such as fake phone numbers (phone spoofing), forged identity documents or professional qualification certificates (e.g. “financial advisor”), sometimes issued by non-existent organizations, as well as web browsing techniques that allow them to mask their IP connection while interacting with their victims and each other(TOR, VPN). In addition, victims are often persuaded to grant remote access to their "consultants" to their computers, allowing them to transfer funds overseas very easily. In similar cases, law enforcement officers are expected to perform several tasks, such as:

1. **Forensic acquisition of digital evidence**. Cloning the victim’s devices, acquiring chats, emails, screenshots and files), while preserving data in accordance with international forensic standards (e.g. ISO/IEC 27037)[4].
2. **Verification of the integrity of communications**. Extracting and decoding data, verifying links to the fraudulent platform, analysing metadata and recovering deleted items).
3. **Tracing cryptocurrency flows**. Following the path of cryptocurrency funds and identify any wallets linked to exchanges or mixers, using blockchain forensic tools and specialised software (e.g. Chainalysis, Elliptic).
4. **OSINT**. Linking suspicious wallets, web domains, and digital identities.

2.2. Narrative, conventions adopted and dataset availability

The fictitious criminal organization we created is composed of four individuals located in the European Union (Italy, Austria, Croatia, Slovenia) with high operational capabilities and advanced technological skills, capable of creating shell companies and counterfeited ID cards, or to use real documents stolen from previous victims, managing spoofing techniques (both for emails and for telephone numbers), as well as of setting up fraudulent websites, producing "deep fakes", and generating international financial flows through digital platforms and cryptocurrencies. They supposedly operated in a specific time period (from 1 January 2024 to 30 June 2025), luring one hundred victims and recruiting five “straw men” – some unaware, others accomplices – spread in their respective Countries, generating a total revenue of at least five million euros.

The LLM was instructed to generate various kinds of digital artefacts produced by this activity, i.e. records of telephone conversations, texts (in different languages), and metadata from electronic communications (among criminals, or towards victims and with "straw men"), as well as financial transactions.

In creating the dataset, some simplifications were introduced to optimize Prompt and output lengths in accordance with the limitations of current LLM providers: telephone numbers are not falsified; the IP addresses of the connections are not masked; it was assumed that the exchanges complied with KYC (Know Your Customer) regulation and cooperated with the police authorities.

The prompt used, the dataset containing the “synthetic data” generated by the LLM and the outcomes are published in open format on a freely accessible dataset[5].

3. Workflow overview

As shown in Figure 1, our workflow can be divided into three phases: (1) Dataset preparation; (2) LLMs testing; (3) LLMs benchmarking. In the following paragraphs, we offer a few clarifications for each of them.

[3]https://www.milanofinanza.it/news/truffe-online-falsi-video-di-giorgia-meloni-promettono-un-guadagno-di-50-000-euro-con-un-investimento-202412121252369462.

[4]https://www.iso.org/standard/44381.html.

[5]the LLM adopted was ChatGPT5, the dataset is available at https://zenodo.org/records/16927274.

3.1. Dataset preparation

In the training phase, the choice of examples that are as realistic as possible to provide to the model is a critical aspect, since the quality of the outcome will depend on how the data are selected, classified, labeled, and processed. Biases can also present themselves in the forensic field and are very dangerous.

The methodology proposed here takes into account that inputs can be biased, and tries to reduce such risks by adopting separate measures for each step.

1. **Criminal scenario design**. As explained above, we created a simulation. In other worlds, our intention is not to incriminate real individuals using predictive policing approaches but to generate fictitious criminals enacting consolidated strategies.
2. **Prompt design**. The purpose is to generate a synthetic data set that can represent - as a whole - the narrative behind the data.
3. **LLM1 prompting**. Using an AI tool (preferably different from the one being evaluated), synthetic data are generated. For example, for an investigation into a cryptocurrency scam conducted via email and WhatsApp, sequences of messages will be produced that are consistent, appropriately distributed over time, and similar to those actually received by real victims.
4. **Criminal scenario data set**. The data set based on the criminal scenario designed. Being fictitious, it is by design anonymous.
5. **Reference data set**. This data set contains public domain, hence neutral, references[6]. This ensures also that the format of the evidences generated by the agent (e.g. bank account reports, telephone call logs, Whatspp messages) is consistent with that used in real-life law enforcement (e.g. ETSI Charging Data Records).
6. **Synthetic data set**. The synthetic data is added to the reference dataset to construct what we will call the synthetic dataset. This represents a realistic and consistent example of a set of data that could be found on one or more devices belonging to a possible perpetrator.

3.2. LLMs Testing

Detection, classification and analytical functions are applied to our datasets by the LLM to be evaluated, and the results obtained are compared with those expected (i.e. with the information available *a priori*).

1. **Prompt design.** By applying appropriate metrics, the reliability of the tool in the specific context is then determined objectively.
2. **LLM2 - LLMn prompting.** By applying the same procedure to different LLMs, it is possible to compare them, to evaluate the best performer, or to obtain confidence indices with which to measure the results, or even to deploy multiple LLMs for the same case.
3. **LLM2-LLMn extracted criminal dataset.** This sequence of operations can be repeated with different synthetic datasets to average the measurements and obtain a higher level of confidence.

3.3. LLMs benchmarking

Inevitably, discrepancies will emerge between the results obtained and the original synthetic data; therefore, the metrics described below are applied to quantitatively measure the performance of the tool in the context under consideration.

To define the metrics, it is necessary to consider the objectives of using the AI tool. Since, in the course of judicial investigations, law enforcers cannot avoid directly analysing the original data, the AI tool can be used to identify data of interest within device memories, which may contain – as typically do – amounts of information that make manual searching and selection difficult or impossible.

The metrics for evaluating the effectiveness of the AI tool in question must be defined in such a way as to measure how close the automatic processing is to the ideal result.

Some criteria for their definition are as follows:

[6]E.g. Computer Forensic Reference DataSet Portal, NIST, https://cfreds.nist.gov.

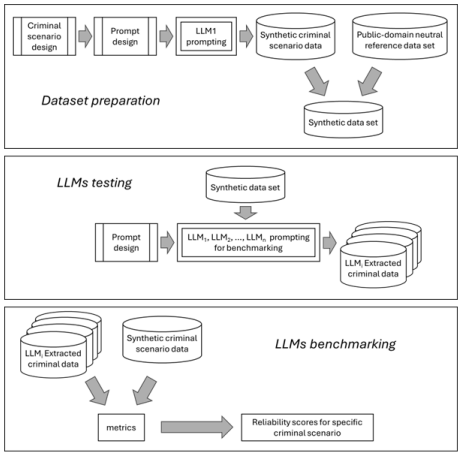

Figure 1: Description of the workflow

1. **Efficiency.** For each type of message, the ratio between the number of relevant messages found and the total number of messages of the same type in the synthetic data.
2. **False negatives.** For each type of message, the ratio between the number of relevant messages not found and the total number of messages of the same type in the synthetic data.
3. **False positives.** For each type of message, the ratio between the number of irrelevant messages classified as relevant and the number of messages of the same type in the summary data.
4. **Summary quality.** The LLM can provide a summary of the content of messages identified as relevant and recognise regular patterns, timing, operating modes, etc. The evaluation in this case will be represented by a score, for example obtained as the average of the scores assigned by the participants in the evaluation.

4. Concluding remarks and research prospects

This methodology does not aim in solving the explainability issues, but just to offer a tool for assessing the trustworthiness of artificial agents. We believe that AIQA, possibly in synergy with IQA and sector ontologies such as CASE, can be a useful tool with a variety of possible applications. For example, it could be possible not only to train a single artificial agent – optimising its model – but also to compare the quality of different agents used within a proceeding – for example, those of the prosecution, the defence and the civil party – in an easily understandable and communicable way.

Declaration on Generative AI

During the preparation of this work, the authors employed Generative AI tools (ChatGPT 5.1 Plus, Overleaf Writefull and Deepl) in supporting text polishing and translation. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] N. Saba, W. K. Balwan, Artificial intelligence in forensic science: Can it be a revolution or else?, Scholars Academic Journal of Biosciences 13 (2025) 335–339. doi:10.36347/sajb.2025.v13i03.005.

[2] R. Brighi, Informatica forense, algoritmi e garanzie processuali, Ars interpretandi 1 (2021) 153–164. URL: https://www.rivisteweb.it/doi/10.7382/100798. doi:10.7382/100798.

[3] K.-J. Kim, C.-H. Lee, S.-E. Bae, J.-H. Choi, W. Kang, Digital forensics in law enforcement: A case study of llm-driven evidence analysis, Forensic Science International: Digital Investigation 54 (2025) 301939. doi:10.1016/j.fsidi.2025.301939.

[4] D. Vyas, M. Shah, A. Kothari, J. Golakia, V. Parikh, Enhancing digital forensics: Machine learning techniques for social media investigation 258 (2025-01-01) 2290–2301. URL: https://www.sciencedirect.com/science/article/pii/S1877050925015868. doi:10.1016/j.procs.2025.04.483.

[5] J. Hendrix, D. Morozoff, Media forensics in the age of disinformation, in: H. T. Sencar, L. Verdoliva, N. Memon (Eds.), Multimedia Forensics, Springer Singapore, 2022, pp. 7–40. URL: https://link.springer.com/10.1007/978-981-16-7621-5_2. doi:10.1007/978-981-16-7621-5_2, series Title: Advances in Computer Vision and Pattern Recognition.

[6] H. Vaghela, N. Varshney, R. Jain, Leveraging AI and ML to innovate forensic frameworks for the identification of illicit operations and extraction of digital artifacts within deep web and dark web environments 2 (2025-05-16) 20–35. URL: https://www.digitalsecurityforensics.org/digisecforensics/article/view/43. doi:10.29121/digisecforensics.v2.i1.2025.43, number: 1.

[7] S. Rizvi, M. Scanlon, J. Mcgibney, J. Sheppard, Application of artificial intelligence to network forensics: Survey, challenges and future directions 10 (2022) 110362–110384. URL: https://ieeexplore.ieee.org/abstract/document/9919162. doi:10.1109/ACCESS.2022.3214506.

[8] S. Costantini, G. De Gasperis, R. Olivieri, Digital forensics and investigations meet artificial intelligence 86 (2019-07) 193–229. URL: http://link.springer.com/10.1007/s10472-019-09632-y. doi:10.1007/s10472-019-09632-y.

[9] R. S. A. Faqir, Digital criminal investigations in the era of artificial intelligence: A comprehensive overview 17 (2023) 77–94. URL: https://cybercrimejournal.com/menuscript/index.php/cybercrimejournal/article/view/189.

[10] R. Liepiņa, F. Lagioia, M. Lippi, P. Pałka, H.-W. Micklitz, G. Sartor, Automating legal tasks: LLMs, legal documents, and the AI act, in: M. Zou, C. Poncibò, M. Ebers, R. Calo (Eds.), The Cambridge Handbook of Generative AI and the Law, 1 ed., Cambridge University Press, 2025-08-07, pp. 407–424. URL: https://www.cambridge.org/core/product/identifier/9781009492553%23c23/type/book_part. doi:10.1017/9781009492553.029.

[11] L. Floridi, P. Illari, The Philosophy of Information Quality, Synthese library, Springer, 2014.

[12] F. Costantini, F. Galvan, M. A. De Stefani, S. Battiato, Assessing "information quality" in iot forensics: Theoretical framework and model implementation, Journal of Applied Logics – IfCoLog Journal of Logics and their Applications 8 (2021) 2373–2406.

[13] M. Scanlon, F. Breitinger, C. Hargreaves, J.-N. Hilgert, J. Sheppard, ChatGPT for digital forensic investigation: The good, the bad, and the unknown 46 (2023) 301609. URL: https://www.sciencedirect.com/science/article/pii/S266628172300121X. doi:10.1016/j.fsidi.2023.301609.

[14] B. Sharma, J. Ghawaly, K. McCleary, A. M. Webb, I. Baggili, ForensicLLM: A local large language model for digital forensics 52 (2025-03-01) 301872. URL: https://www.sciencedirect.com/science/article/pii/S2666281725000113. doi:10.1016/j.fsidi.2025.301872.

[15] European Union Agency for Law Enforcement Cooperation., Europol (2023), Online fraud schemes: a web of deceit, Technical Report, Europol, 2023.

[16] European Union Agency for Law Enforcement Cooperation., IOCTA, internet organised crime threat assessment 2024., Publications Office, 2024. URL: https://data.europa.eu/doi/10.2813/442713.

Online Resources

The dataset generated is available at: https:zenodo.org/records/16927274 .