Investigation of Large Language Models, GenAI, and Proprietary AI Systems: Digital Forensic Evidence, Readiness and Regulation

Mark Scanlon[1]

[1] *Forensics and Security Research Group, School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland*

Large language models have already been considered in digital forensics as possible investigative assistants. They may help to generate scripts, summarise material, translate communications, identify patterns, or provide a natural language interface to complex evidence sources [17]. Those uses remain relevant, particularly against the backdrop of long-standing workload and automation pressures in digital forensic practice. They also carry familiar concerns around hallucination, over-reliance, explainability, validation, and human oversight [13, 12, 8, 10, 2]. This editorial considers a related but separate problem. Large AI models and services are increasingly becoming systems that may need to be investigated themselves.

This is not an entirely new observation. Prior work has proposed AI forensics as a domain of inquiry [1] and has asked whether an AI system caused or contributed to a particular event [14]. More recent studies have begun to examine artefacts from conversational AI services [4], mobile LLM applications [5, 15], local LLM deployments [9], invocation logs [3], and multi-agent frameworks [16]. That work is necessary. It helps to identify recoverable traces from specific systems and provides a starting point for practical examination. The issue considered here is broader. The future investigative problem is not only whether a ChatGPT, Gemini, Claude, Copilot, or locally hosted LLM application left useful artefacts on a device. It is whether a proprietary, cloud-hosted, continuously updated AI system can be preserved, examined, explained, and challenged in a manner consistent with digital forensic principles.

A useful starting point is to avoid the all-encompassing phrase "the AI did it". A deployed AI service is rarely a single model operating in isolation. It may include a base model, fine-tuned components, system and developer instructions, safety layers, retrieval-augmented generation (RAG) components, vector databases, plugins, agents, orchestration frameworks, API gateways, cloud infrastructure, moderation systems, and multiple layers of logging. A visible set of prompts and responses may be the most accessible part of the interaction, but they are not the entire interaction. The output may have been shaped by hidden instructions, retrieved documents, safety policies, tool calls, cached context, access controls, and cloud/service-side configurations. From a forensic standpoint, the object of examination is therefore not simply the response, but the wider model-interaction environment.

That environment may contain many potential sources of digital evidence. Prompt and response histories, user sessions and access logs, account records, device identifiers, API calls, model identifiers, inference parameters, deployment versions, tool-use traces, plugin logs, agent actions, retrieval logs, vector-store metadata, uploaded files, generated artefacts (including interim generated artefacts), moderation events, and disclosure records may all be relevant. In some cases, these artefacts may be available from a local device or user account data export. In others, they will be held only by the provider, or they may not have been retained in the first instance. This creates a familiar but intensified digital forensic problem: evidence of high investigative value may be located in a system that the investigator cannot independently acquire.

The regulatory context gives these records increased significance. Under the EU AI Act, certain high-risk AI systems are required to support automatic logging over their lifetime for traceability, post-market monitoring, and monitoring of operations. Providers of general-purpose AI models are required to maintain technical documentation concerning the model, including training and testing processes as well as evaluation results, subject to qualifications regarding intellectual property, confidential business information, and trade secrets. High-risk AI providers must also establish post-market monitoring systems that collect, document, and analyse relevant performance data throughout the system lifecycle. Serious incident reporting obligations will, in some circumstances, require subsequent investigation, risk assessment, corrective action, and cooperation with competent authorities [7]. NIS2, in parallel, imposes cybersecurity risk-management and reporting obligations across a broad set of essential and important entities, including digital infrastructure and cloud-related services [6].

These obligations will generate a new class of investigative questions. For example, an investigator, regulator, court, insurance company, affected party, or defence expert may need to establish:

- Whether an AI system misclassified data,
- Whether the required logs existed,
- Whether the logs were complete and authentic,
- Whether a serious incident should have been reported,
- Whether a provider's incident report omitted relevant prompt, retrieval, tool-use, or model-version context,
- Whether post-market monitoring detected a known risk,

- Whether a third-party deployer's use of a system was consistent with the provider's instructions,
- Whether a cybersecurity incident affected the operation of a model, retrieval pipeline, or agentic workflow, or
- Whether the user or the AI system was at fault, or what the correct apportionment of blame or responsibility is.

In criminal justice settings, these questions also intersect with defence access to a clear evidence trail where AI-enabled systems collect, organise, visualise, or assess investigative data [11]. These are not abstract compliance questions. They are questions about digital evidence, provenance, integrity, traceability, and reconstruction.

The preservation problem follows directly from this. A screenshot of a generated answer or a copied transcript may be useful, but it is not the complete picture of an AI interaction. It may not record the full prompt context, model version, system instructions, safety-layer decisions, retrieval sources, tool calls, access logs, or provider-side modifications. Nor will it necessarily show whether an output was regenerated, edited, refused, truncated, or influenced by a temporary configuration change. In cloud-hosted systems, delay may determine whether the relevant evidence can be preserved at all. Logs may expire, model versions may be retired, retrieval indexes may be rebuilt, and the system state that produced the relevant output may no longer exist.

Repeatability is also more difficult than in many established areas of digital forensic examination. Replaying a prompt thread is not reproduction due to the non-deterministic nature of these systems. The same visible prompt(s) may generate different outputs because of probabilistic generation, altered system instructions, changed safety policies, updated model weights, modified retrieval corpora, tool availability, or provider-side deployment changes or resource throttling. Even where a service exposes a model name, that label may not be sufficiently granular for forensic reconstruction. A forensically meaningful record may need to identify the model's deployment build version, configuration, relevant instruction set, tool calls, and moderation path. Without that information, an examiner may only be able to demonstrate plausibility but not reproducibility.

The proprietary nature of many large AI systems compounds these issues. In a conventional digital forensic examination, the practitioner may encounter encryption, damaged media, or closed application formats. In AI investigations, the examiner may instead be asked to rely on a provider's account of what occurred inside a system that cannot be imaged, inspected, or re-executed independently. Model weights, system prompts, safety layers, telemetry, training data, fine-tuning data, and internal performance records may be unavailable. Some limitations on disclosure may be justified. They may protect security, privacy, intellectual property, or trade secrets. Nonetheless, commercial sensitivities cannot be allowed to become a substitute for forensic scrutiny. A provider-provided export is not the same thing as an independent acquisition, and a provider-provided compliance assertion is not the same thing as evidential validation.

This tension is likely to become more prominent in regulated AI investigations. Service providers may be under a duty to produce documentation, logs, incident reports, monitoring records, or explanations. Investigators will then need to assess not only the content of those materials, but also how they were created, retained, selected, redacted, and disclosed. Were the logs generated during the original execution or assembled after the fact? Were they protected against tampering? Can they be tied to a specific deployment version and associated deployed parameters? Were relevant retrieval, moderation, or agentic actions included? Is there an audit trail for later deletion, export, or provider-side change? These questions are familiar in digital forensics, but the scale, opacity, and service-dependence of AI systems make them more difficult to answer.

Privacy and privilege should not be treated as secondary concerns. Prompts, uploaded files, embeddings, retrieval stores, and interaction logs may contain sensitive personal or commercial data, privileged communications, trade secrets, or data pertaining to third-party entities. The investigative response cannot simply be to request "all AI logs". That may be disproportionate and may expose unrelated data. Equally, an overly narrow request may miss the context needed to interpret the evidence. AI forensic readiness should therefore support scoped, privacy-aware, and legally controlled preservation and disclosure.

Of course, human oversight remains necessary, but it is not the complete answer. A human-in-the-loop process can reduce the risk of reliance on AI outputs and can help ensure that conclusions are verified against the underlying evidence. However, human review at the end of a process does not make an opaque system at the beginning of the process forensically sound. If the examiner cannot establish what the system was instructed to do, what data it accessed, what version was deployed, what it retrieved, what it filtered, and what tools and scripts it executed, then the human reviewer is validating only part of the process. Oversight without traceability does not provide assurance.

The practical response should be AI forensic readiness. Service providers should assume that their systems may later become the subject of legal, regulatory, or internal investigation. At a minimum, a forensically useful AI interaction record should include:

- A stable user-model interaction identifier,
- Trusted timestamps,
- Account, device, and API metadata,
- Model and deployment version information,
- Relevant model parameters and configuration,
- System and developer instruction identifiers or hashes,
- User prompts, uploaded files, and visible responses,
- Moderation and safety-layer events,
- Retrieval source identifiers and index versions,
- Tool calls, plugin calls, and agent actions,

- Generated artefact identifiers and hashes,
- Human approval, override, or escalation events, and
- Retention, deletion, export, and disclosure history.

Not every investigation will require the disclosure of every element. The point is that the system should be capable of preserving and attesting to them when legally required.

There is a role here for the digital forensic community. If AI logging requirements are left entirely to providers, the resulting records may be optimised for product analytics, security monitoring, customer support, or regulatory minimum compliance rather than forensic examination. The field should help define what makes an AI-system record evidentially useful. This includes terminology, validation procedures, preservation request formats, minimum metadata requirements, tamper-evident logging, signed deployment records, versioning standards, export mechanisms, and test datasets for evaluating completeness and repeatability. Research should also continue into forensic artefacts from AI clients, local LLM environments, agentic frameworks, vector databases, retrieval pipelines, and provider disclosure workflows.

The question is no longer just whether AI can assist in digital forensic investigations. In constrained and validated settings, it can. The more difficult question is whether digital forensics can examine AI. Regulation may require logs, documentation, monitoring, and incident reports, but it will not automatically make those materials forensic. They must still be preserved, scoped, authenticated, interpreted, tested, and challenged. If large AI systems are allowed to make decisions, generate artefacts, operate tools, and record interactions without suitable forensic readiness, investigators will be left with evidence that is convenient but fragile. This will be further compounded by complex multi-layer AI-AI interactions. We need to ensure that AI systems are not merely compliant on paper but are examinable in practice.

References

[1] Baggili, I.M., Behzadan, V., 2020. Founding the Domain of AI Forensics, in: Proceedings of the Workshop on Artificial Intelligence Safety, co-located with the 34th AAAI Conference on Artificial Intelligence (SafeAI@AAAI 2020), pp. 1–5. URL: https://ceur-ws.org/Vol-2560/paper53.pdf.

[2] Chernyshev, M., Baig, Z., Syed, N., Doss, R., Shore, M., 2026. Large Language Models in Digital Forensics: Capabilities, Challenges and Future Directions. Forensic Science International: Digital Investigation 56, 302043. doi:10.1016/j.fsidi.2025.302043.

[3] Chernyshev, M., Baig, Z.A., Doss, R.R.M., 2024. Towards Large Language Model (LLM) Forensics Using LLM-Based Invocation Log Analysis, in: Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis, Association for Computing Machinery, New York, NY, USA. pp. 89–96. doi:10.1145/3689217.3690616.

[4] Cho, K., Park, Y., Kim, J., Kim, B., Jeong, D., 2025. Conversational AI Forensics: A Case Study on ChatGPT, Gemini, Copilot, and Claude. Forensic Science International: Digital Investigation 52, 301855. doi:10.1016/j.fsidi.2024.301855.

[5] Dragonas, E., Lambrinoudakis, C., Nakoutis, P., 2024. Forensic Analysis of OpenAI's ChatGPT Mobile Application. Forensic Science International: Digital Investigation 50, 301801. doi:10.1016/j.fsidi.2024.301801.

[6] European Parliament and Council of the European Union, 2022. Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on Measures for a High Common Level of Cybersecurity across the Union, Amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and Repealing Directive (EU) 2016/1148 (NIS2 Directive). Official Journal of the European Union, L 333, 27 December 2022, pp. 80–152. URL: http://data.europa.eu/eli/dir/2022/2555/oj.

[7] European Parliament and Council of the European Union, 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations and Directives. Official Journal of the European Union, L 2024/1689. URL: http://data.europa.eu/eli/reg/2024/1689/oj.

[8] Hargreaves, C., Breitinger, F., Dowthwaite, L., Webb, H., Scanlon, M., 2024. DFPulse: The 2024 Digital Forensic Practitioner Survey. Forensic Science International: Digital Investigation 51, 301844. doi:10.1016/j.fsidi.2024.301844.

[9] Jeong, S., Lee, S., Park, J., 2025. LangurTrace: Forensic Analysis of Local LLM Applications. Forensic Science International: Digital Investigation 54, 301987. doi:10.1016/j.fsidi.2025.301987.

[10] Michelet, G., Henseler, H., van Beek, H., Scanlon, M., Breitinger, F., 2025. Fine-Tuning Large Language Models for Digital Forensics: Case Study and General Recommendations. Digital Threats: Research and Practice 6, 1–18. doi:10.1145/3748264.

[11] Sachoulidou, A., 2026. A Typology of Automated Law Enforcement Systems under EU Law: What Standards for Defence Rights? European Papers – A Journal on Law and Integration 11, 595–630. doi:10.15166/2499-8249/884.

[12] Scanlon, M., Breitinger, F., Hargreaves, C., Hilgert, J.N., Sheppard, J., 2023a. ChatGPT for Digital Forensic Investigation: The Good, the Bad, and the Unknown. Forensic Science International: Digital Investigation 46, 301609. doi:10.1016/j.fsidi.2023.301609.

[13] Scanlon, M., Nikkel, B., Geradts, Z., 2023b. Digital Forensic Investigation in the Age of ChatGPT. Forensic Science International: Digital Investigation 44, 301543. doi:10.1016/j.fsidi.2023.301543.

[14] Schneider, J., Breitinger, F., 2023. Towards AI Forensics: Did the Artificial Intelligence System Do It? Journal of Information Security and Applications 76, 103517. doi:10.1016/j.jisa.2023.103517.

[15] Tyagi, S., Gong, Y., Karabiyik, U., 2025. Forensic Analysis and Privacy Implications of LLM Mobile Apps: A Case Study of ChatGPT, Copilot, and Gemini. Forensic Science International: Digital Investigation 54, 301974. doi:10.1016/j.fsidi.2025.301974.

[16] Walker, C., Gharaibeh, T., Alsmadi, R., Hall, C., Baggili, I.M., 2024. Forensic Analysis of Artifacts from Microsoft's Multi-Agent LLM Platform AutoGen, in: Proceedings of the 19th International Conference on Availability, Reliability and Security, Association for Computing Machinery, New York, NY, USA. pp. 198:1–198:9. doi:10.1145/3664476.3670908.

[17] Wickramasekara, A., Breitinger, F., Scanlon, M., 2025. Exploring the Potential of Large Language Models for Improving Digital Forensic Investigation Efficiency. Forensic Science International: Digital Investigation 52, 301859. doi:10.1016/j.fsidi.2024.301859.