Deep Reasoning and Large Context Windows: Next-Generation AI in Digital Forensic Investigations

Hans Henseler[a,b,*], Timo Meconi[b]

[a] *University of Applied Sciences Leiden, The Netherlands*
[b] *Netherlands Forensic Institute, The Hague, The Netherlands*

Abstract

Recent advances in large language models (LLMs) allow the processing of entire case dossiers "in one breath" while employing deeper, step-by-step reasoning, heralding a new era in digital forensics. Where existing retrieval-augmented approaches can miss crucial clues by splitting long documents into fragments, larger context windows enable a more holistic view of all text, significantly reducing the risk of overlooking pivotal evidence. Coupled with explicit multi-step reasoning, these models can propose hypotheses, link scattered evidence, and illuminate investigative scenarios otherwise hidden from manual review.

This short paper reports on experiences gained from a proof-of-concept experiment involving the so-called "Crystal Clear-case," a fictitious yet realistic digital forensics scenario. We also draw parallels to Steven Johnson's "Long Context" experiment, illustrating how entire books or voluminous dossiers can be read and reasoned about cohesively. We discuss the potential for AI to automate tasks such as privilege filtering and data relevance assessments, thus helping protect fundamental rights. Finally, we present the Hansken Copilot Prototype—currently under development—a local LLM-based assistant designed to streamline investigative workflows in real forensic environments.

This document is intended as a long abstract for a presentation at DFRWS EU 2025 in which we will elaborate on the technical and procedural aspects of harnessing extended context windows, stepwise AI reasoning, and local deployments in forensic labs.

Keywords: Digital forensics, Large Language Models, Extended context windows, Stepwise reasoning, Evidence analysis, Hansken Copilot, Forensic AI

1. Introduction

Digital forensic practitioners routinely grapple with large, varied collections of electronic data—for instance, smartphone backups, chat logs, or emails. As the European Council emphasizes in its e-evidence publication, "The digital revolution is redefining every aspect of society, and crime is no exception" (European Council, 2024). In many investigations, the challenge is akin to searching for a "needle in a haystack." Tools based on Retrieval-Augmented Generation (RAG) attempt to mitigate this by splitting documents into chunks, but subtle or cross-referenced clues may be missed if the relevant fragments do not appear together (Henseler, 2023).

Emerging large language models (LLMs) with bigger context windows promise a new solution: feeding entire dossiers "in one breath," thus reducing reliance on fragmented retrieval methods. In tandem, explicit multi-step reasoning (sometimes referred to as "slow reasoning") provides deeper analytical detail, enabling investigators to trace the logic behind AI inferences (OpenAI, 2024b; DeepMind, 2024).

2. Background and Motivation

This section provides an overview of how extended context windows and transparent AI reasoning can benefit forensic investigations.

2.1. Steven Johnson's "Long Context" Example

Steven Johnson's project *You Exist In The Long Context* (Johnson, 2025) vividly demonstrates how an AI with a large context window can parse an entire book in one go, maintain narrative consistency, and improvise seamlessly with user-driven prompts. His text, *The Infernal Machine*, involves historical facts and a detective storyline around Joseph Faurot. By giving the AI an "integral view" of all chapters at once, the system can remain faithful to key events, respond to divergent user choices, and enrich the narrative with consistent details.

For forensics, this same capacity to hold an entire corpus "in focus" is crucial for correlating scattered references across suspects, chat conversations, or other evidentiary sources. A small snippet overlooked in chunked retrieval might hold the key to tying a suspect to a location or proving a contradictory statement. In essence, Johnson's experiment—though literary—illustrates the power of a single large context bubble for comprehensive reasoning.

[*] Corresponding author.
Email addresses: henseler.h@hsleiden.nl (Hans Henseler), timo.meconi@nfi.nl (Timo Meconi)

2.2. Fragmentation vs. Integral Context

When investigating real forensic cases, chunk-based retrieval remains common but can miss subtle cross-references (Henseler, 2025). An LLM with an extended context window can ingest entire chat histories or multiple interview transcripts simultaneously, reducing the risk of overlooking crucial connections. This holistic approach parallels Johnson's example, where the model does not rely on partially retrieved segments but instead holds the "whole text" in memory.

2.3. Stepwise Reasoning: More Transparency

Legal and forensic processes require every inference to be auditable. "Slow reasoning" or stepwise elaboration prompts the model to explain each step linking evidence to a conclusion (Kahneman, 2012; OpenAI, 2024a). Investigators can then examine the logic for plausibility, inconsistencies, or biases—a crucial feature when translating AI outputs into legal scrutiny. Furthermore, chain-of-thought models (Deepseek R1 family) offer investigators the possibility of evaluating the reasoning of the AI-output Liu et al. (2024).

3. Related Work on AI Copilots in Digital Forensics

The surge in digital evidence has prompted researchers and developers to explore AI copilots that enhance forensic investigations. These copilots, often LLM-driven, focus on tasks like anomaly detection, data summarization, or investigative support within secure environments.

3.1. Academic Research on AI Copilots

Scanlon et al. (2023) evaluated ChatGPT's capabilities in digital forensics, identifying use cases for artifact interpretation, incident response, and training. Nevertheless, the authors emphasized risks of inaccurate AI outputs and potential privacy breaches if sensitive data were uploaded to cloud-based models. Henseler & van Beek (2023) similarly studied ChatGPT within Hansken, demonstrating efficiency gains when data is carefully managed.

Ensuring the confidentiality of digital evidence remains a significant concern, prompting a shift toward offline or on-premise AI solutions in research labs. Such approaches aim to mitigate risks tied to cloud-based model queries. Wickramasekara et al. (2024) propose the use of a DFaaS like Hansken to reduce the high infrastructure costs to deploy a LLM on-premise. Furthermore, the authors acknowledge the high potential of LLM's in assisting investigators in the examination, analysis, and reporting phase. The LLM can handle certain subtasks such as parsing or converting data. However, human oversight is still needed to verify the results based on the analysis of digital evidence.

3.2. Commercial Tools and Developments

Numerous vendors have introduced AI copilots into forensic tools. Belkasoft (2024) describes BelkaGPT, an on-premise DFIR assistant that automates artifact recognition and timeline generation. An offline variant of Magnet Copilot was likewise released to address investigators' privacy constraints (Magnet Forensics, 2024). Moreover, Relativity has integrated generative AI (ChatGPT) into its platform, extending its existing machine-learning capabilities toward e-discovery and privilege filtering (Relativity, 2024).

While some forensic professionals remain skeptical of AI-based analysis due to concerns over chain-of-custody and legal admissibility, there is growing consensus that carefully designed copilot systems can improve investigative throughput without compromising data security.

4. Proof-of-Concept: The "Crystal Clear-case"

Before discussing the findings, it is important to note that the Crystal Clear dataset is used to provide a realistic training environment accompanying the e-learnings offered by the Hansken Academy[1]. Through interactive exercises, users can practice typical digital forensics tasks on this fictitious—yet representative—scenario.

4.1. Scenario and Data

Adapted from Henseler (2025), the Crystal Clear-case scenario involves multiple suspects implicated in a fictitious drug-lab operation. Investigators possess:

- Extended chat logs from various apps (WhatsApp, Telegram, Signal, Snapchat).
- Police interviews with each suspect, noting contradictory accounts.
- Location data (addresses like Bronsstraat 8 in IJmuiden).
- Ancillary records on finances and alleged "lab" references.

This dataset can easily exceed typical model token limits if chunked improperly.

4.2. Method and Findings

The OpenAI o1-preview model(OpenAI, 2024b) was tested with multiple complete chat conversations, prompting it to:

1. Construct an integrated timeline of suspect activities.
2. Cross-reference statements and location logs, flagging inconsistencies.
3. Identify privacy-sensitive content (medical or attorney-client).
4. Hypothesize about each suspect's role in the operation.

[1] For more information about Hansken and the Hansken Academy see https://hansken.org/

Within about a minute of inference time, the model produced a chronological narrative, linking ephemeral Snapchat references to more explicit WhatsApp messages, verifying time zones, and highlighting contradictions in suspect testimony. Had the data been chunked, certain ephemeral messages might not have been retrieved or correlated. These results underscore the efficiency and thoroughness of extended-context analysis.

5. Hansken Copilot Prototype (Under Development)

While the Crystal Clear-case demonstrates the potential of large context windows, we have also begun developing a Hansken Copilot prototype to explore how local LLM integration can streamline everyday forensic workflows. In this early phase, the prototype uses a locally installed LLM and a local database on each investigator's workstation. This setup allows on-the-fly summarization, translation, entity detection, and Hansken Query Language (HQL) generation without sending data to external services, thereby preserving data sensitivity.

We decided to start with a local prototype to accelerate development and clarify user requirements. Implementing an LLM and a new database layer within the Hansken backend is time-consuming, and we currently lack a precise blueprint for how such an integration should be structured. By opting for a local setup, we can rapidly iterate on design choices, functionality, and prompt engineering, gathering feedback from a small cohort of pilot users before committing to significant backend changes. However, because most Hansken users do not have GPU-equipped workstations, this local approach will necessarily be restricted to a limited number of participants during the prototype evaluation.

If the local prototype proves useful, we envision moving the LLM, storage, and associated APIs to the Hansken backend as a final phase. In that scenario, the LLM would be containerized and run on a GPU-powered node within Hansken's Kubernetes cluster, while the local database would transition to Hansken's PostgreSQL environment. This shift would enable large-scale deployments, robust session management, and multi-user collaboration—features essential for widespread adoption of AI-driven workflows. Nonetheless, early indicators suggest that investigators value how the Copilot can parse large data volumes, generate structured summaries, and propose relevant queries; the iterative approach aims to address remaining hardware, privacy, and scalability concerns as we progress toward full integration.

6. Challenges and Future Directions

We face a rapidly evolving AI landscape with a constant influx of newly published LLMs, each offering different strengths. Below is an overview of key challenges and future directions:

Resource Constraints. Running extended context windows locally requires significant memory and GPU/CPU power (DeepMind, 2024). Optimizations such as quantization may be vital for widespread adoption.

Privacy and Security. Keeping data on local machines enhances security but complicates multi-user collaboration. Chain-of-custody also remains a pressing concern.

Error Propagation & Bias. While stepwise reasoning improves transparency, it does not eliminate potential biases or flawed intermediate logic. Careful auditing and domain adaptation remain crucial.

Backend Integration. Tools like Hansken have strict versioning and accreditation requirements. Maintaining these standards with an AI "copilot" necessitates cautious planning and user acceptance testing.

Handling Extremely Large Cases. Despite larger context windows, practical token limits persist. For multi-terabyte investigations, a layered or iterative approach may be necessary, blending integral analysis with targeted chunking.

LLM Model Selection. With new LLMs published almost weekly, relying on a single "best model" is risky. We aim to develop benchmarks for specific forensic tasks to evaluate candidate models effectively. Different tasks (e.g., reasoning, summarization, translation) may benefit from different LLM architectures that balance speed, quality, and resource usage.

Prompt Development and Evaluation. Constructing and refining effective prompts is a specialized skill. In the Hansken Copilot, we plan to offer predefined prompts for summarization, translation, reporting, and query generation. Ongoing experiments will measure accuracy, performance, and reliability to refine these task-specific prompts.

7. Proposed DFRWS EU 2025 Presentation

At DFRWS EU 2025, the presentation will:

1. **Demonstrate Extended Context Models:** Show how integral analysis reveals hidden links, referencing both the Crystal Clear-case and Steven Johnson's "Long Context" experiment.
2. **Highlight Stepwise Reasoning:** Illustrate how an auditable chain of thought can reinforce the reliability of AI-aided investigations.
3. **Present Hansken Copilot PoC:** Outline the local LLM architecture, user feedback, and the phased roadmap for ongoing development toward a centralized deployment.
4. **Discuss Ethical/Legal Aspects:** Emphasize automated privilege filtering and synergy with user privacy demands (Baron & Others, 2023; Relativity, 2024).
5. **Solicit Community Insights:** Seek best practices for domain-tailored prompt engineering, offline LLM integration, and standardized benchmarks in forensic toolchains.

8. Conclusion

Large language models with extended context windows promise transformative gains for forensic analysis, enabling holistic reading of entire corpora without the fragmentation that often undermines RAG-based methods. Steven Johnson's "Long Context" demonstration, though focused on a literary narrative, shows how consistent logic and comprehensive detail can emerge when an AI references all text at once.

Our proof-of-concept *Crystal Clear-case* highlights how stepwise reasoning can bolster the forensic legitimacy of AI-driven insights. Complementing these efforts, the Hansken Copilot prototype—currently under development—shows the feasibility of local LLM deployment in real investigations, offering on-demand summaries, translations, and advanced query support. Challenges include hardware demands, privacy safeguards, multi-user scaling, the art of prompt engineering, and a constantly evolving landscape of new models. Nonetheless, initial practitioner feedback has been positive.

We look forward to presenting these findings at DFRWS EU 2025, engaging with peers on how best to deploy advanced AI responsibly for faster, higher-quality forensic investigations.

Acknowledgments

The authors thank the teams at the University of Applied Sciences Leiden and the Netherlands Forensic Institute for their support and collaboration, as well as the participants who tested the Hansken Copilot prototype. Portions of this paper were adapted from Dutch texts and refined with AI-based drafting support under human supervision.

References

Baron, D., & Others (2023). Can generative ai effectively handle foia exceptions? ArXiv preprint. ArXiv:2304.12345.

Belkasoft (2024). BelkaGPT—The First Offline AI Assistant for DFIR Investigations. URL: https://belkasoft.com/belkagpt last accessed: February 4, 2025.

DeepMind, G. (2024). Gemini 2.0 update december 2024. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/. Accessed: 8 January 2025.

European Council (2024). Better access to e-evidence to fight crime. https://www.consilium.europa.eu/en/policies/e-evidence/. Accessed: 2025-02-04.

Henseler, H. (2023). De opkomst van kunstmatige intelligentie in expertise en recht: de invloed van chatgpt en grote taalmodellen. Expertise & Recht. [In Dutch] Issue 5, p. 135-138.

Henseler, H. (2025). Dieper denken met ai als taalmodellen met een ruimer kortetermijngeheugen gaan redeneren. Expertise & Recht. [In Dutch] Issue 1, 2025.

Henseler, H., & van Beek, H. (2023). ChatGPT as a Copilot for Investigating Digital Evidence. *CEUR-WS*, *3423*. URL: https://ceur-ws.org/Vol-3423/paper6.pdf. Last accessed: February 4, 2025.

Johnson, S. (2025). You exist in the long context. https://thelongcontext.com/. Published on 20 November 2024, accessed on 6 January 2025.

Kahneman, D. (2012). *Thinking, Fast and Slow*. Penguin Books.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C. et al. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, .

Magnet Forensics (2024). Making Magnet Copilot's AI capabilities available offline. URL: https://www.magnetforensics.com/blog/making-magnet-copilots-ai-capabilities-available-offline/ last accessed: February 4, 2025.

OpenAI (2024a). Introducing openai o1. https://openai.com/o1. Accessed: 27 January 2025.

OpenAI (2024b). Introducing openai o1-preview. https://openai.com/index/introducing-openai-o1-preview. Accessed: 7 January 2025.

Relativity (2024). Relativity ai for review (air). https://www.relativity.com/data-solutions/air/privilege/. Accessed: 9 January 2025.

Scanlon, M., Breitinger, F., Hargreaves, C., Hilgert, J.-N., & Sheppard, J. (2023). ChatGPT for digital forensic investigation: The good, the bad, and the unknown. *Forensic Science International: Digital Investigation*, *46*, 301609. URL: https://www.sciencedirect.com/science/article/pii/S266628172300121X. Last accessed: February 4, 2025.

Wickramasekara, A., Breitinger, F., & Scanlon, M. (2024). Exploring the potential of large language models for improving digital forensic investigation efficiency. *arXiv preprint arXiv:2402.19366*, .