A Cloud-Native Architecture for Human-in-Control LLM-Assisted OpenSearch in Investigative Settings

Benjamin Puhani*, Kai Brehmer*, Malte Prieß†

*AI Research Unit of the
State Police of Schleswig-Holstein, Kiel, Germany
e-mail: {benjamin.puhani | kai.brehmer}@polizei.landsh.de

†Faculty of Computer Science and Electrical Engineering
Kiel University of Applied Sciences, Germany
e-mail: malte.priess@haw-kiel.de

***Abstract*—Complex criminal investigations are often hindered by large volumes of unstructured evidence and by the semantic gap between natural language investigative intent and technical search logic. To address this challenge, we present a design and feasibility study of a cloud-native microservice architecture tailored to private-cloud deployments, contributing to research in secure cloud computing and leveraging modern cloud paradigms under high security and scalability requirements. The proposed system integrates Large Language Models into a "Human-in-Control" workflow that translates natural-language queries into syntactically valid OpenSearch Domain-Specific Language expressions. We describe the implementation of a hybrid retrieval strategy within OpenSearch that combines BM25-based lexical search with nested semantic vector embeddings. The paper focuses on system design and preliminary functional validation, establishing an architectural baseline for future empirical evaluation. Technical feasibility is demonstrated through a functional prototype, and a rigorous evaluation methodology is outlined using the Enron Email Dataset as a structural proxy for restricted investigative corpora.**

Keywords-OpenSearch; Information Retrieval; Semantic Search; Investigative Settings; Enron Dataset.

I. INTRODUCTION

Investigations into international and transnational crimes, such as genocide, crimes against humanity, and related violations of international criminal and humanitarian law, require the analysis of large volumes of unstructured textual evidence, including witness statements, interview transcripts, and communication records. Proceedings under national universal jurisdiction frameworks, such as the German Code of Crimes against International Law (Völkerstrafgesetzbuch, VStGB), serve as a representative example of this broader class of investigations. Across such contexts, investigators face a recurring challenge: relevant evidence is often present in the data but remains difficult to access because of the gap between natural language investigative intent and the technical logic of search systems. Recent scholarship underscores this urgency: Skipanes et al. [1] identify the processing of unstructured text as a critical bottleneck in digital forensics and highlight that current methodologies largely fail to bridge the divide between computational opportunities and practical investigative reasoning.

Although modern search engines, such as OpenSearch [2], provide scalable indexing and retrieval capabilities, they inherently require input in a rigid and structured Query Domain-Specific Language (DSL) rather than in natural language. Most investigators and legal practitioners lack this expertise, leading them to rely on manual review or simple keyword searches. These approaches are poorly suited to capturing semantic variation, indirect references, variant spellings, and translation artifacts, and they limit recall - the ability to retrieve all relevant information - precisely in those cases where exploratory and hypothesis-driven search is required.

This paper addresses this semantic gap by presenting an exploratory proof-of-concept system that integrates Large Language Models (LLMs) as a translation layer between investigative intent and open search query logic. Natural language questions are mapped to syntactically valid OpenSearch DSL queries within a Human-in-Control architecture, in which the LLM functions as a supervised assistant rather than as an autonomous agent. The contribution of this paper is to outline a principal system design and methodological foundation, along with a novel architectural integration that provides a basis for future empirical evaluation. Because of legal and ethical constraints associated with real investigative data, the approach is demonstrated using the Enron Email Dataset [3] as a structural proxy that exhibits key characteristics of investigative corpora, including unstructured text, noisy data, and complex communication networks. Architecturally, the system is positioned within cloud-native computing paradigms, addressing challenges in private-cloud orchestration, horizontally scalable components, and secure cloud environments to ensure strict data sovereignty.

The remainder of this paper contextualizes this architecture within existing research (Section II), details the system design and hybrid retrieval strategy (Section III), demonstrates its functional feasibility (Section IV), and outlines the roadmap for empirical evaluation (Section V).

II. Related Work

To address the semantic and technical challenges of forensic data analysis, our work builds upon and integrates research from three primary domains.

A. Bridging the Semantic Gap in Digital Forensics:

In a recent comprehensive analysis, Skipanes et al. [1] identify the processing of unstructured text as a critical bottleneck in contemporary criminal investigations. They highlight that, while computational opportunities exist, current methods largely fail to bridge the gap between technical retrieval logic and the qualitative reasoning required by investigators. Our work directly addresses this architectural gap by operationalizing these opportunities within a secure on-premises environment.

B. LLM-Assisted Retrieval:

The integration of LLMs into Information Retrieval systems has evolved rapidly from simple re-ranking tasks to complex query generation [4]. Current approaches often focus on *Text-to-SQL* paradigms, in which LLMs translate natural language into structured SQL queries for relational databases [5]. However, these methods are inherently constrained by the unstructured and fuzzy nature of forensic text data. Conversely, Retrieval-Augmented Generation (RAG) grounds LLM responses in retrieved search results but may introduce hallucinations or lack the deterministic precision required for rigorous investigative filtering [6].

C. Cognitive Architectures and Prompting:

Our work builds upon the findings of Liu et al. [7] concerning the "Lost-in-the-Middle" phenomenon, which posits that LLMs often fail to identify relevant information in long contexts. We address this limitation by segmenting documents into semantic units rather than processing entire texts. Furthermore, we adopt the Chain-of-Thought (CoT) prompting strategy proposed by Wei et al. [8] to improve the logical consistency of generated OpenSearch queries. Unlike autonomous agents, our architecture enforces a "Human-in-Control" design that prioritizes the investigator's control over search logic to ensure procedural accountability within legal domains.

III. System Architecture and Methodology

To meet the high security and scalability requirements of law enforcement agencies, the proposed system is designed as a cloud-native microservice architecture. While leveraging modern cloud paradigms such as containerization and orchestration, the system is intended for deployment within a restricted private-cloud environment (e.g., on-premise Kubernetes) to ensure strict data sovereignty. At the same time, the architecture remains deployment-agnostic: the identical microservice stack can operate either in a fully orchestrated Kubernetes environment for large-scale installations or in a lightweight Docker Compose configuration for resource-constrained agencies. This flexibility stems from consistent containerization of all services and a clear separation between application logic and infrastructure orchestration, enabling the system to scale operational complexity according to organizational needs.

A. Cloud-Native Service Orchestration

The system follows a microservice architectural pattern composed of four distinct layers that communicate via RESTful APIs and asynchronous message queues (see Figure 1):

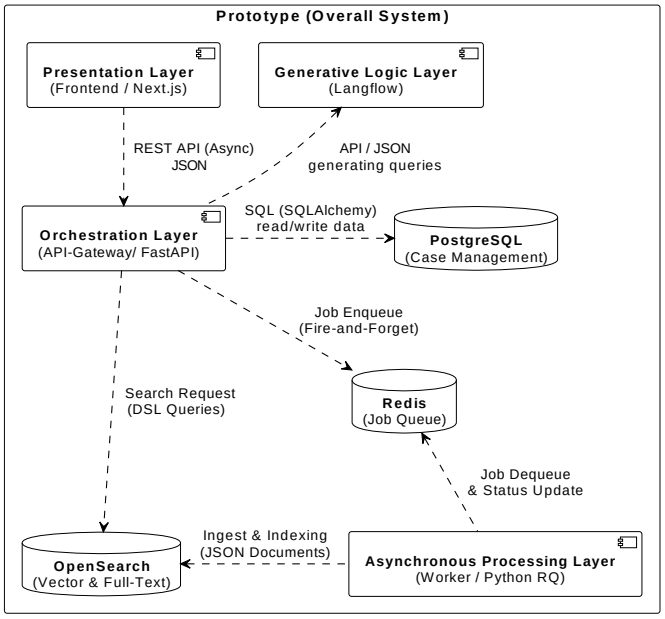

Figure 1. Schematic representation of the modular architecture and data flow.

Presentation Layer: A Single Page Application built with Next.js provides the investigative user interface. It uses Server-Side Rendering to optimize initial load performance and communicates asynchronously with the backend to ensure a non-blocking user experience, which is essential for reviewing large document sets.

Orchestration Layer (API Gateway): The core application logic is handled by a FastAPI backend. Unlike monolithic frameworks, FastAPI implements the ASGI standard, enabling native asynchronous request handling. This capability is critical for maintaining high throughput when coordinating I/O-intensive operations across the database, search engine, and LLM services. State management is offloaded to persistent stores - PostgreSQL for case management and Redis for job queues. Crucially, this design transforms the system from a stateless search engine into a case-management workspace. The backend tracks the "reviewed status" of each retrieved document, supporting a coverage-oriented workflow that enables investigators to systematically examine evidence by relevance.

Asynchronous Processing Layer (Worker): To handle large-scale data ingestion, we implemented a producer-consumer pattern using Redis. Python-based workers perform CPU-intensive heuristic parsing and disentanglement. However, computationally expensive vectorization is offloaded to the OpenSearch cluster thorugh a dedicated ingest pipeline that runs on specialized machine learning nodes. This design separates the cleaning logic from the inference workload, enabling independent scaling of ingestion workers and neural inference resources.

Generative Logic Layer: Instead of hardcoding prompt logic, the system integrates Langflow [9] as a visual low-code environment to manage interaction chains. To enforce data sovereignty, the architecture deliberately avoids reliance on public APIs. Instead, it uses locally hosted open-weight models (e.g., Llama 3 or Mixtral) running on on-premise inference servers via vLLM. This design ensures that no sensitive investigative intent leaves the secure private-cloud perimeter.

B. Semantic Segmentation and Ingestion Strategy

A major challenge in processing large-scale forensic data is the "Lost-in-the-Middle" phenomenon, in which LLMs fail to retrieve relevant information embedded in long, unstructured contexts [7]. Indexing a typical investigative document (e.g., a 50-page witness statement or an extended email thread) as a monolithic block degrades vector search performance because of the architectural token limitations of the underlying transformer models [10].

To address this limitation, we developed a configurable, modular adapter pattern within the asynchronous worker nodes. Unlike generic chunking strategies (e.g., fixed-size splitting), our system supports custom parsing profiles explicitly designed for each input type to preserve semantic boundaries in forensic data:

Heuristic Chunking (Legacy Documents): For unstandardized text documents, we implemented a custom regex-based heuristic to detect semantic shifts (e.g., speaker changes or timestamps). This approach allows the system to generate atomic segments even in the absence of structured separators.

Disentanglement (Communication Data): For the Enron dataset, the ingestion logic was tailored to disentangle forwarded message chains. By selectively stripping technical headers (e.g., `X-UID`) from the semantic payload (Level 2) while preserving them for structured filtering (Level 1), we minimize noise in the vector space.

This pre-processing step ensures that embeddings represent specific statements rather than diluted document-level averages.

C. Hybrid Data Modeling and Indexing

A core architectural decision was to implement a hybrid search index within OpenSearch. The current approach focuses exclusively on text-based data. Image and audio-based data would need to be converted into text. The hybrid search index addresses the fundamental linguistic limitations of purely lexical retrieval, specifically synonymy and polysemy [11]. By integrating vector-based semantic retrieval, our schema combines the strengths of both paradigms, a pattern applicable to both interview protocols and digital communication:

Unstructured Baseline (Level 1): The document's full text and metadata are indexed using standard BM25 lexical search [12]. This configuration ensures high recall for specific keywords, such as *names, case numbers, or senders/receivers.*

Structured Nested Embeddings (Level 2): To mitigate context dilution, we avoid embedding long documents as a single vectors. Instead, the atomic semantic units identified in Section III-B (e.g., specific paragraphs, message bodies, or individual statements) are stored as nested objects containing the segment text and a 384-dimensional vector embedding. We use the HNSW algorithm [13] with Cosine Similarity, adhering to the optimization objective of the underlying `paraphrase-multilingual-MiniLM-L12-v2` model [10]. To minimize architectural complexity, this model is provisioned through the OpenSearch ML Commons framework and executed directly on the cluster's internal ML nodes. This configuration ensures that ranking is determined by semantic alignment (vector orientation) rather than by vector magnitude. Additionally, a search-time synonym graph expands queries, e.g., mapping "detention" to "imprisonment", without increasing the physical index size.

This nested structure is designed to prevent "cross-object matching" errors in which a query matches unrelated parts of a document (e.g., matching a suspect's name from page 1 with an action described on page 10), and to enable the LLM to generate queries that target specific semantic segments.

Hybrid Fusion for Prioritized Review: During the exploratory evidence-review phase, minimizing *False Negatives* is paramount. Investigators require a ranking mechanism that surfaces the most relevant documents first to support the coverage-oriented workflow described in Section III-A. Our system employs a score normalization approach to fuse unbounded lexical scores (BM25) with normalized semantic scores (Cosine Similarity). This design ensures that a document describing "off-balance sheet debt" (semantic match) is ranked competitively with documents containing the specific project code "Raptor" (lexical match), even when their vocabularies do not overlap.

D. The "Human-in-Control" Generative Pipeline

The translation of natural language into the complex OpenSearch Query DSL is handled by a multi-agent LLM pipeline orchestrated via Langflow. To ensure domain agility without code modifications, the system employs schema-aware prompting: the current index definition is injected into the prompt context at runtime. Crucially, this design enforces a strict separation of concerns: the LLM operates exclusively on the abstract index schema, never on the actual evidentiary content. Unlike standard RAG workflows [14], which require feeding retrieved text into the model's context window, our architecture ensures that sensitive document payloads remain confined within the OpenSearch cluster and are never exposed to the inference context. Rather than relying on opaque "black box" logic, we implement a CoT workflow:

Reasoning & Generation: The first agent acts as a "Query Architect". It analyzes the user's intent and the injected schema structure, generating an intermediate reflection before constructing the JSON query.

Auditing (Quality Assurance): The second agent, the "Auditor", validates each generated query against known error patterns. For instance, it detects when the model attempts to search for structured entities (e.g., specific dates or person names) within the semantic vector field, which typically yields

lower precision. The Auditor enforces correct mapping to structured fields (e.g., moving a name search to the `sender` or `people` field) before execution.

Transparent Execution & Deterministic Retrieval: The validated query is then executed to provide immediate feedback. However, unlike fully autonomous agents, the system enforces transparency: the generated search logic (the "translation") is displayed alongside the results. Since all presented results are deterministic database retrievals rather than LLM-generated text, the risk of evidence hallucination is eliminated. The investigator retains full authority to evaluate the relevance of retrieved documents and iteratively refine the generated query logic, ensuring that the final assessment of evidence remains a human decision.

This pipeline ensures that the resulting DSL query is syntactically valid and semantically aligned with the investigator's intent prior to execution.

IV. Implementation Status and Preliminary Feasibility

The architecture described in Section III has been implemented as a fully functional prototype. The system successfully orchestrates interactions among the React frontend, the FastAPI gateway, and the asynchronous worker nodes. Initial functional tests confirm that the Langflow-based "Human-in-Control" pipeline is capable of generating syntactically valid OpenSearch DSL queries from natural language input. Specifically, the multi-agent setup (Generator and Auditor) demonstrated the ability to detect and correct basic logical errors, such as mapping named entities to incorrect fields, before execution. This technical readiness establishes the necessary baseline for the empirical evaluation strategy outlined in Section V, which will be the focus of subsequent research.

V. Conclusion and Future Work

This paper presented a cloud-native, microservices-based architecture designed to reduce technical barriers in accessing mass data for criminal investigations. By combining a hybrid OpenSearch index with a supervised LLM pipeline, we established a framework to translate investigative intent into high-precision database queries without compromising data sovereignty.

The next phase focuses on the quantitative calibration of the system. As real-world investigative data is legally restricted to operational use under strict purpose limitation regulations, it cannot be utilized for public academic benchmarking. Therefore, we will utilize the Enron email dataset [3] as a reproducible Ground Truth to simulate forensic retrieval tasks.

The primary objective is not merely to compare algorithms, but to determine the optimal hybrid configuration for forensic workflows, which prioritize high recall (coverage) over precision. The evaluation will specifically investigate two core architectural decisions:

Segmentation Granularity: Comparing retrieval performance when indexing monolithic documents versus the proposed granular segmentation (e.g., heuristic chunking or thread-splitting). We hypothesize that granular segments yield higher relevance scores for specific details but require aggregation to preserve context.

Score Fusion Tuning: Evaluating different weighting strategies for the Score Normalization (Lexical vs. Semantic weights) to maximize Recall@100. This metric, which measures the proportion of relevant documents retrieved within the top 100 results [11], is chosen to reflect the operational reality of investigators, who require the most critical evidence to appear within the first few pages of results.

To test these hypotheses, the evaluation design involves:

Adversarial Scenario Design: We will define 5–10 distinct search scenarios designed to simulate the semantic gap. Instead of searching for known identifiers (e.g., "Project Raptor"), queries will formulate abstract investigative intents (e.g., "conversations expressing anxiety about the company's financial stability" or "instructions to destroy documents"). These scenarios are specifically chosen to challenge lexical search engines, as they rely on sentiment and context rather than unique keywords.

Semantic Ground Truth Construction: We utilize the publicly available CMU Enron Corpus [3] as the structural baseline. For the evaluation of retrieval performance (Recall/Precision), we utilize established relevance judgments from the TREC Legal Track or comparable academic annotation sets (e.g., UC Berkeley Enron Analysis). This ensures that the Ground Truth represents the semantic reality of the documents, independent of the specific vocabulary used in the query.

Comparative Ablation Study: To quantify the added value of the hybrid architecture, we will conduct an ablation study comparing three configurations:

(a) Purely Lexical (Level 1 BM25),
(b) Purely Semantic (Level 2 Vector-only), and
(c) Hybrid Score Fusion (Level 1 + Level 2).

We hypothesize that the hybrid system yields the highest Recall@100, demonstrating superior robustness in scenarios where suspects employ obfuscated language or indirect phrasing that evades purely lexical detection.

References

[1] M. Skipanes, G. Demartini, K. Franke, and A. B. Nissen, "Information analysis in criminal investigations: Methods, challenges, and computational opportunities processing unstructured text," *Policing: A Journal of Policy and Practice*, vol. 19, paaf005, Mar. 2025, ISSN: 1752-4520. DOI: 10.1093/police/paaf005.

[2] OpenSearch Project, *OpenSearch*, version 3.4, https://opensearch.org/ [visited: 2026-03-13], The Linux Foundation, 2025.

[3] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *Machine Learning: ECML 2004*, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 217–226, ISBN: 978-3-540-30115-8. DOI: 10.1007/978-3-540-30115-8_22.

[4] Y. Zhu et al., "Large language models for information retrieval: A survey," *ACM Transactions on Information Systems*, vol. 44, no. 1, pp. 1–54, 2026, ISSN: 1046-8188. DOI: 10.1145/3748304.

[5] L. Shi, Z. Tang, N. Zhang, X. Zhang, and Z. Yang, "A survey on employing large language models for text-to-sql tasks," *ACM Computing Surveys*, vol. 58, no. 2, pp. 1–37, 2026, ISSN: 0360-0300. DOI: 10.1145/3737873.

[6] Y. Gao et al., *Retrieval-augmented generation for large language models: A survey*, 2024. DOI: 10.48550/arXiv.2312.10997.

[7] N. F. Liu et al., "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024. DOI: 10.1162/tacl_a_00638.

[8] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, S. Koyejo et al., Eds., vol. 35, Curran Associates, Inc., 2022, pp. 24 824–24 837.

[9] Langflow AI, *Langflow*, version 1.6.8, https://github.com/langflow-ai/langflow [visited: 2026-03-13], 2025.

[10] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2019. DOI: 10.48550/arXiv.1908.10084.

[11] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008, ISBN: 9780521865715. DOI: 10.1017/CBO9780511809071.

[12] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at trec-3," in *Overview of the Third Text REtrieval Conference (TREC-3)*, NIST, 1995, pp. 109–126.

[13] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 824–836, 2020. DOI: 10.1109/TPAMI.2018.2889473.

[14] P. Lewis et al., "Retrieval-Augmented Generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 9459–9474.