HAW Kiel
Hochschule für Angewandte Wissenschaften Kiel
Kiel University of Applied Sciences

SH Schleswig-Holstein
Der echte Norden

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

Authors: B. Puhani, K. Brehmer, M. Prieß

Presenter: Benjamin Puhani
AI Research Unit of the State Police Schleswig-Holstein, Kiel, Germany
Contact email: **benjamin.puhani@polizei.landsh.de**

IARIA

SH Schleswig-Holstein
Der echte Norden

Benjamin Puhani

DevOps Engineer @ AI Research Unit, State Police SH

- Education:
  - B.Sc. Computer Science
  - Trained IT Specialist for Application Development
- Focus: K8s, OpenSearch, and AI Orchestration

XING

IARIA

AI Research Unit @ State Police Schleswig-Holstein

Main Goal: Bridging the gap between cutting-edge AI research and judicial/police requirements.

Current Projects:

- Sovereign LLM orchestration for sensitive data.
- Automated Transcription & Translation.
- Case-Related Mass Data Analysis Platform.

Connect with us:
projekt-ki@polizei.landsh.de

Dissecting the Title

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

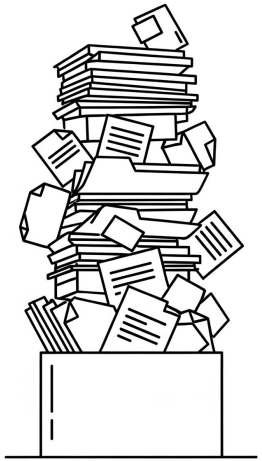
- **LLM-Assisted OpenSearch:** Using AI to generate complex queries, not to read the data.
- **Investigative Settings:** High-stakes scenarios requiring absolute precision.
- **Cloud Native:** Scalable, containerized microservices.
- **Human-in-Control:** The user decides, the AI only translates.

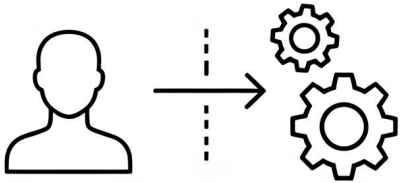
Related Work

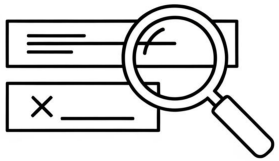
- **The Bottleneck:** Processing unstructured text is a critical barrier in forensics (Skipanes et al., 2025).
- **Current LLM Retrieval:**
  - **Text-to-SQL:** Struggles with the unstructured and fuzzy nature of forensic data (Shi et al., 2026).
  - **Standard RAG:** Risks hallucinations and lacks deterministic precision for legal filtering (Gao et al., 2024).
- **Cognitive Limits:** LLMs often fail to identify relevant information in long texts ("Lost-in-the-Middle", Liu et al., 2024).

The Core Problem: The "Semantic Gap"

- **Massive Scale:** Huge volumes of unstructured data.
- **The Gap:** Human Intent ≠ Technical DSL.
- **The Keyword Limit**: Standard search misses context and indirect phrasing.



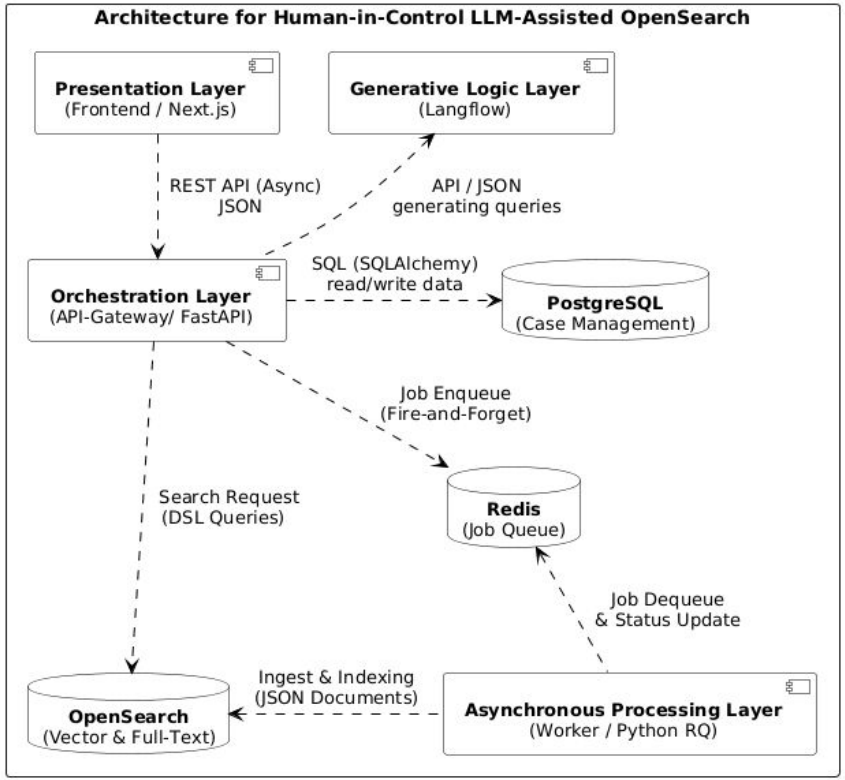




A closed Environment

- **Strict Privacy:** 100% On-Premise. No public APIs, no external clouds.
- **The Sovereign Stack:**
  - Self-hosted LLMs via local servers.
  - Private Cloud orchestration (Kubernetes) within high-security perimeters.
- **Resource Realties:** High hardware costs. Limited Supply of GPUs and RAM available.
- **Legal Constraint:** AI results are not reliable enough.

System Architecture

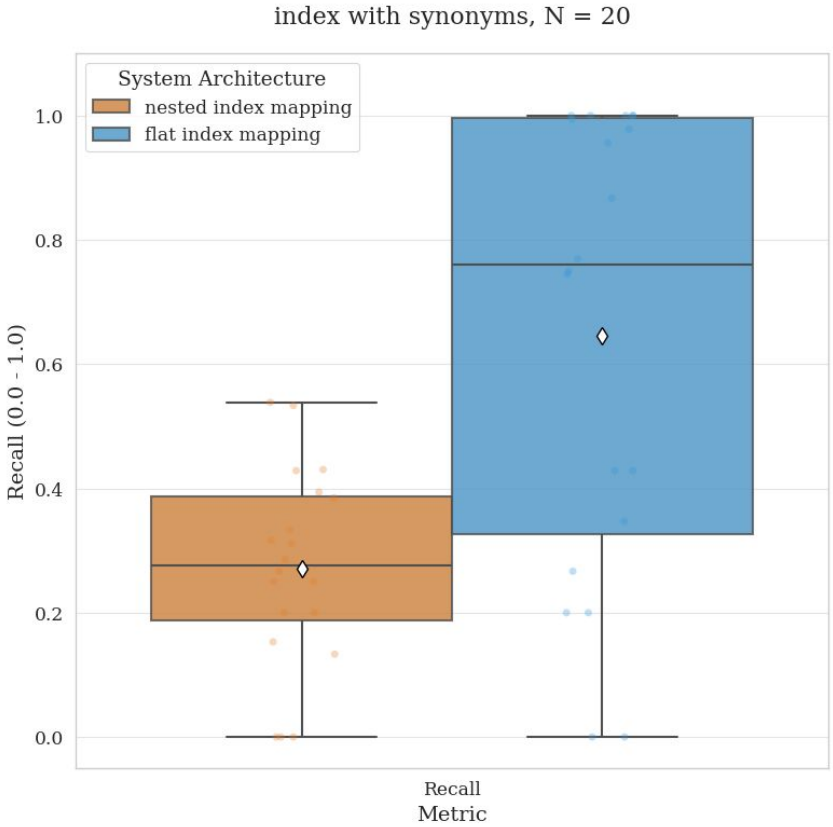


The "Human-in-Control" Pipeline

- **Input:** "Find conversations about financial distress."
- **Translation:** Multi-Agent Pipeline generates OpenSearch DSL.
- **Approval:** Human validates logic before execution.

```
{
  "query": {
    "hybrid": {
      "queries": [
        {
          "nested": {
            "path": "semantic_segments",
            "query": { "match":
              { <field_name>: "financial distress bankruptcy" }
            }
          }
        },
        {
          "neural": {
            "semantic_segments.embedding": {
              “query_text”: “conversations about ...”
              "model_id": "paraphrase-multilingual",
              "k": 100
            }
          }
        }
      ]
    }
  }
}
```

Preliminary Findings: Flat vs. Nested Indices



Goal: Solving the "Lost-in-the-Middle" problem in long documents.

Hypothesis: Strict segmentation leads to context fragmentation.

Conclusion & Future Work

- **The Bridge:** Successfully translated investigative intent into secure database logic.
- **The "Vibe Coding" Dilemma:** The human is in control, but forced to verify complex JSON they barely understand.
- **Future Work:**
  - Optimizing "Score Fusion" (Flat vs. Nested).
  - UX Layer: Translating the JSON back into a visual, readable summary before execution.

HAW Kiel
Hochschule für Angewandte Wissenschaften Kiel
Kiel University of Applied Sciences

SH Schleswig-Holstein
Der echte Norden

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

Authors: B. Puhani, K. Brehmer, M. Prieß

Presenter: Benjamin Puhani
AI Research Unit of the State Police Schleswig-Holstein, Kiel, Germany
Contact email: **benjamin.puhani@polizei.landsh.de**

IARIA

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: 2026-04-12], The Linux Foundation, 2025.

[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.

[9] Langflow AI, Langflow, version 1.6.8, https://github.com/langflow-ai/langflow [visited: 2026-04-12], 2025.