Enhancing Criminal Investigation Analysis with Summarization and Memory-based Retrieval-Augmented Generation: A Comprehensive Evaluation of Real Case Data

Mads Skipanes[*,1] mads.skipanes@politiet.no

Tollef Emil Jørgensen[*,2] tollef.jorgensen@ntnu.no

Kyle Porter[1] kyle.porter@ntnu.no

Gianluca Demartini[3] g.demartini@uq.edu.au

Sule Yildirim Yayilgan[1] sule.yildirim@ntnu.no

[1]Norwegian University of Science and Technology, Gjøvik, Norway
[2]Norwegian University of Science and Technology, Trondheim, Norway
[3]University of Queensland, Brisbane, Australia

Abstract

This study introduces KriRAG, a novel Retrieval-Augmented Generation (RAG) architecture designed to assist criminal investigators in analyzing information and overcoming the challenge of information overload. KriRAG structures and summarizes extensive document collections based on existing investigative queries, providing relevant document references and detailed answers for each query. Working with unstructured data from two homicide case files comprising approximately 3,700 documents and 13,000 pages, a comprehensive evaluation methodology is established, incorporating semantic retrieval, scoring, reasoning, and query response accuracy. The system's outputs are evaluated against queries and answers provided by criminal investigators, demonstrating promising performance with 97.5% accuracy in relevance assessment and 77.5% accuracy for query responses. These findings provide a rigorous foundation for other query-oriented and open-ended retrieval applications. KriRAG is designed to run offline on limited hardware, ensuring sensitive data handling and on-device availability.

1 Introduction

Criminal investigators face information overload due to the manual analyses of vast volumes of data (Gianola, 2020; Rossmo, 2021; Partridge and Zaghloul, 2023). Investigations are cognitively demanding, characterized by assessments of incomplete information that often lead to a broad set of open questions. Moreover, insights gained today may become inapplicable tomorrow, requiring constant reassessment of information.

Criminal investigators collect, store, and analyze information in an environment where the volume of information increases rapidly and new questions arise continuously. In this process, questions are conceptualized as *information-needs* and expressed as specific questions such as "Was the deceased involved in any conflicts?", or "Who resided at the address?". The analytical task of answering information-needs and organizing the answers into coherent themes, a process we refer to as thematization builds upon foundational work by Nissen (2018). This is a resource-intensive task that relies on dedicated roles such as the document reader and the indexer (NPCC, 2021). Retaining and recalling established themes when manually analyzing case files with thousands of documents is challenging enough; ensuring relevant information is not overlooked is an additional layer of complexity.

Currently, the use of large language models (LLMs) in criminal investigations has been researched for tasks such as generating police reports (Michelet and Breitinger, 2024), artifact identification, keyword searching, and programming within the digital forensics discipline (Scanlon et al., 2023) and as an investigative copilot (Henseler and van Beek, 2023). Europol has reported a prototype RAG-based system evaluated on nonsensitive documents for the potential use within operational and general support functions (Europol, 2024). However, to our knowledge, no solutions have been proposed to support criminal investigators in thematizing information. Thus, we develop KriRAG, a system using summarization and memory techniques with Retrieval-Augmented Generation (RAG) to enhance the process of thematization in criminal investigations. KriRAG's core is based on open-weight LLMs, sentence embedding models for retrieval, and vector databases. The RAG process is guided by a *memory* that builds up during a query, aiming to keep track of references to document IDs and involved entities to summarize the key details throughout. Finally, all acquired memories are combined to recreate a concise answer to the query. This process is guided by extensive test-

[*]These authors contributed equally to this work. Corresponding authors.

Proceedings of the 31st International Conference on Computational Linguistics, pages 4993–5010
January 19–24, 2025.

ing, and we meticulously evaluate all outputs and compare them with data from real investigations. Although RAG-based systems can reduce hallucinations and misinformation (Yu et al., 2024), it remains a critical challenge to minimize these occurrences, where the need for caution and human oversight in deploying such technologies has been emphasized (Scanlon et al., 2023; Yu et al., 2024), along with calls for interpretable artificial intelligence in forensics (Garrett and Rudin, 2023). Our study aims to provide a comprehensive evaluation of the potential advantages, and risks of integrating KriRAG into the criminal investigative workflow. We ask: *to what extent can KriRAG enhance the process of thematizing information?*

The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 introduces the data, while Section 4 provides a detailed description of the system architecture. The methodology and research approach are outlined in Section 5. Experiments and results are presented in Section 6, followed by a discussion in Section 7. Finally, Section 8 summarizes the findings and suggests directions for future research.

Note: This study discusses case files that contain unsettling information and language pertaining to violent crimes.

2 Related Work

Transformer-based models have proven effective in addressing resource-intensive tasks within the legal domain, such as entity extraction and entity linking (Zhong et al., 2020; Batini et al., 2021; Rodrigues et al., 2022; Barros et al., 2023). For law enforcement in particular, "Hansken", an open digital forensic platform by The Netherlands Forensic Institute, has published a wrapper for SBERT models (Reimers and Gurevych, 2019) for search applications, although with limited additional information.[1] SBERT models were also used by Skipanes et al. (2023) to retrieve relevant documents from user queries in investigative settings. Zhao et al. (2024) developed the framework *Diverse Legal Factor-enhanced Criminal Case Matching* (DLF-CCM). While effective for case-to-case comparisons, this approach does not extend to analyzing larger document collections connected to a single case, leaving a gap for applications requiring cross-document information.

LLMs One of the many benefits of LLMs is the increased sequence length from previous generative and discriminative language models, often limited to less than a thousand tokens (Devlin et al., 2019; Raffel et al., 2023). There have been massive improvements in context length both from provided APIs (e.g. OpenAI's GPT models[2]) and from open-weight models (Touvron et al., 2023; Dubey et al., 2024; Team et al., 2024; Abdin et al., 2024) – scaling up to 128K tokens. Moreover, there are several efforts to increase it even further, e.g., by modifying the positional encoding with Rotary Position Embeddings (RoPE) (Su et al., 2024) and YaRN (Peng et al., 2023). The development is promising, although the larger context will still not support case files spanning thousands of pages, especially in low-resource hardware settings. Applications of LLMs in criminal investigations have so far focused on immediate prompt-based outputs rather than integration into RAG systems. For instance, Henseler and van Beek (2023) explored ChatGPT (GPT-3.5) as a "copilot" for digital evidence, proving useful for tasks like summarizing conversations, identifying relationships, and cross-referencing information, given data from a fictional case. Michelet and Breitinger (2024) compared ChatGPT and a 4-bit Llama-2 13B model (Touvron et al., 2023) for writing digital forensics reports. While ChatGPT shows superior performance, there are still issues with inaccuracies and hallucinations. These findings indicate that LLMs are stronger for generating texts with a less analytical nature, e.g., summarization.

RAG Current implementations of RAG in related fields are mainly limited to legal documents and question-answering. Louis et al. (2024) created a question-and-answer dataset to fine-tune local LLMs to conduct RAG-supported answering of legal questions. In doing so, user questions could be answered with a degree of expertise, while referencing the legal provisions it was commenting on. In evaluating the system, they found the system to produce syntactically correct text, with some occurrences of incorrect answers. A similar study by Ryu et al. (2023) describes an evaluation method for LLM-generated texts for Korean Legal question-answering. Ryu et al. applied a RAG-based system to the legal domain to verify the authenticity of previously generated text regarding a legal query.

[1] https://github.com/NetherlandsForensicInstitute/bert-embeddings

[2] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

They show that RAG-based evaluation can provide more contextually accurate answers and better align them with the ground-truth evaluations.

3 Data

In this study, three data sources are used. Two are confidential case files from the Norwegian Police – referred to as *Case files* (*Case A* and *Case B*). For reproducibility, an openly available case is included – referred to as *open case*.

Open case The open case contains 135 documents and 586 sentences sourced from a court decision. To establish a ground truth, an investigator annotated the data by generating four information-needs and matched each information-need to sentences and documents they answered (if any). This reflects the investigative practice. The four information-needs maps to 44 sentences spanning 44 unique documents.

Case Files The confidential datasets contain 3729 PDFs comprising 202,568 sentences. Existing information analysis performed by multiple investigators is used as ground truth. For Case A and B investigators had priorly annotated data by identifying sentences and documents that answered specific information-needs (if any). For case A, 14 information-needs map to 371 sentences spanning 326 documents. For case B, 12 information-needs map to 356 sentences spanning 121 documents.

Case	Info. needs	Sents relevant	Sents total	Docs relevant	Docs total
Open	4	44	586	44	135
A	14	371	106748	326	1810
B	12	356	95820	121	1913

Table 1: Number of information-needs, sentences, and documents for all case files. For a complete list of the formulated information-needs please see Table 8, 9, 10.

Additionally, based on the information-needs established in the Open case, Case A and B, three investigators produced the final answers (overall summaries with concluding remarks) for all cases. This was done to establish a ground truth.

4 System Architecture

KriRAG is comprised of three main components.

1. Segmentation and encoding of sentence embeddings
2. Storage of the embeddings for fast lookup
3. Employing Summarization and Memory-based Retrieval-Augmented Generation

Figure 1 shows its components and interactions.

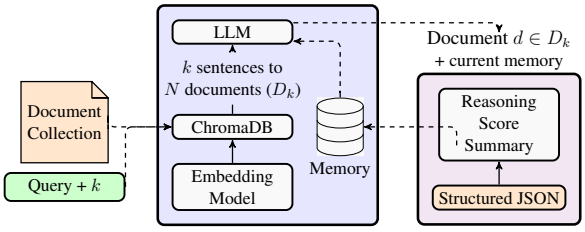

Figure 1: From a document collection and queries, KriRAG finds similar sentences and analyzes connected documents. A memory is built for each query as it loops through the documents.

4.1 Segmenting and Encoding

Before encoding sentences, the original data is filtered and validated to root out issues with the provided data due to OCR errors, repeated symbols, and more. A language-specific sentence encoder[3] is then used, which is trained on top of a Norwegian BERT model (Kummervold et al., 2021) with the SBERT library (Reimers and Gurevych, 2019), producing 768-dimensional embeddings. This model is chosen based on earlier experiments on the Open Case and its annotations (Skipanes et al., 2023). Figure 2 shows the averaged F_1 score when retrieving sentences from the provided information-needs (formulated in Norwegian). Any sentence-transformer model may be used for multilingual or language-specific applications.

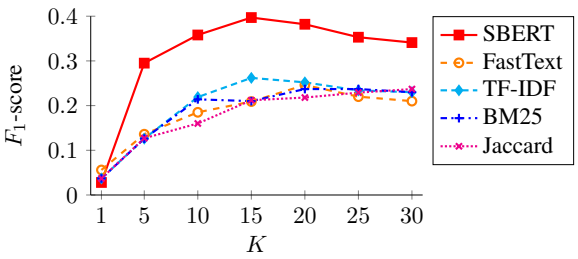

Figure 2: F_1 score for information-needs on the Open Case at top-K retrieved results, with different methods.

4.2 Vector Store and Queries

The Chroma library (Chroma, 2024) stores, queries, and filters embeddings on a sentence level. The sentence encoder (Section 4.1) is used for all interactions with Chroma. When querying, Chroma

[3] https://huggingface.co/NbAiLab/nb-sbert-base

retrieves the k nearest neighbors through the Hierarchical Navigable Small World (HNSW) algorithm (Malkov and Yashunin, 2018). The similarity between the query embedding and existing embeddings is calculated with the Euclidean (L2) distance. To support retrieval with references, sentence IDs and page numbers are added as metadata. Thus, any returned top-k items from the vector database will include these identifiers. Finally, as the entries in the database are *sentences*, the sentence IDs are linked to the corresponding document IDs to access the full texts, which the RAG system then batches and parses.

4.3 The RAG Architecture

The system draws inspiration from *self memory* (Cheng et al., 2024) and *data recycling* in RAG systems (Li et al., 2023), but is entirely based on in-context learning. Prompts are guided by stepwise explanations (Kojima et al., 2022; Thomas et al., 2024). Within the prompts, the system is tasked to generate structured JSON through explicit instructions, along with schemas to generate formal grammars with GBNF – a format based around Backer-Naur Form (McCracken and Reilly, 2003), as integrated into `llama.cpp`.[4] The generated JSON outputs are *questions*, *reasoning*, *relevance score*, *summary* and *memory*. Prompts and schemas are found in Appendix A.

KriRAG builds up its memory as it parses documents for each specific query. The memory is recursively defined as a meta-summary over previous document summaries. Let $D_{k,q}$ represent the set of the top-k relevant documents retrieved for query q, and $S_{d,q}$ denote the query-oriented summary of document $d \in D_{k,q}$. The memory update process is defined as `MemSumm`, such that for every document d, the memory for query q is updated from previously seen documents:

$$M_{d,q} = \texttt{MemSumm}(M_{d-1,q}, S_{d,q}),$$

After processing all documents, the final memory is obtained by applying this update over all documents for a query, in the same process for the remaining RAG operations – retrieving related questions, reasoning, and relevance score. All data is stored for each parsed document, and the *reasoning* and *memory* fields are later used in subsequent prompts, combining them into a concise answer/-summary to the query.

[4] https://github.com/ggerganov/llama.cpp/blob/9379d3c/grammars/README.md

4.4 LLM and Hardware

For the LLM, we opted for Google's Gemma-2 27B (Team et al., 2024), running in a post-training quantized setting – the `Q5_K_M` through llama.cpp.[5] This library allows layers to be offloaded to the CPU and/or GPU. Experiments were limited to an on-site server (Intel Xeon Gold 6448Y). Inference for the open case was done with the same library on an Apple Macbook M1 Max 32GB. We developed an evaluation tool using the open case samples to determine the quality of prompts and model configurations for assessments with the provided annotations described in Section 3. The results are merely a guideline and will differ based on the input data. Table 2 shows the scores across different LLMs where the model classified a document as *relevant*. Note that our model selection highly depends on the observed quality and less on the quantitative results in a single scenario. Gemma-2 27B returned the highest quality answers and was at the limit of what our hardware could serve (albeit slowly).[6] During experiments (May–June 2024), other available models did not provide satisfactory responses. Section 6 describes the evaluation process in detail, and full outputs are found in Appendix B.

Metric	Model	Q1	Q2	Q3	Q4
P	Gemma-2-9B	0.43	**0.50**	0.80	0.33
	Gemma-2-27b	**1.00**	0.44	**0.88**	**0.50**
	Llama-3.1-8b	0.67	**0.50**	0.62	0.17
	Phi-3-Medium	0.12	0.38	0.71	0.00
R	Gemma-2-9B	**0.60**	**0.29**	0.24	**0.40**
	Gemma-2-27b	0.40	0.24	0.41	**0.40**
	Llama-3.1-8b	0.40	**0.29**	**0.47**	**0.40**
	Phi-3-Medium	0.20	0.18	0.29	0.00
F_1	Gemma-2-9B	0.50	**0.37**	0.36	0.36
	Gemma-2-27b	**0.57**	0.31	**0.56**	**0.44**
	Llama-3.1-8b	0.50	**0.37**	0.53	0.24
	Phi-3-Medium	0.15	0.24	0.42	0.00

Table 2: Performance metrics (Precision, Recall, F1 Score) for various instruction-tuned LLMs across the four queries defined in the Open Case.

5 Methodology

We present our methodology by demonstrating how KriRAG aligns with the investigative workflow.

[5] https://github.com/ggerganov/llama.cpp/blob/c8ddce8/examples/quantize/README.md

[6] A GPU with $\geq$ 24GB VRAM is highly recommended for this model in its quantized configuration (approx. 20 GB).

For a visual representation, we direct readers to our overview in Appendix C. In the criminal investigative workflow, initial information is received, giving rise to *information-needs*.
To meet these needs, information is collected, stored in documents, and then analyzed and synthesized into final answers. Investigators manually structure case file documents into tabular formats before synthesizing them into conclusive summaries. Our study uses documents and structured tabular data from the investigative process as our ground truth for evaluating KriRAG. We apply the information-needs as queries and evaluate its outputs at different stages. It is crucial to recognize that our study distinguishes itself by using outcomes from *real criminal investigations* rather than traditional annotations. This approach allows us to perform evaluations within an authentic investigative context, directly comparing outputs to actual investigative results.

Cases A and B were only accessible through an offline virtual machine on the police intranet, with limited computational resources. All excerpts and queries related to these cases are anonymized in this paper while adhering to similar concepts.

5.1 Evaluation

Evaluation of RAG and LLMs is a challenging task due to the complex nature of natural language. This is reflected in the broad set of evaluation criteria discussed in the literature, e.g., bias, efficiency, hallucination, and omission, to mention a few (Liang et al., 2022; Bang et al., 2023; Schiller, 2024). Common criteria are accuracy, reasoning, and coherence (Chang et al., 2024; Yu et al., 2024; Deriu et al., 2021; Hamid et al., 2023; Bang et al., 2023). Our evaluation primarily considers these dimensions.

5.2 Research Questions (RQ)

To systematically evaluate the performance of KriRAG, we follow its core information processing pipeline. The system operates in three main stages:

1. Document retrieval, where relevant documents are identified (RQ1).

2. Relevance assessment, where the system analyzes the retrieved content's relation to the information-need (RQ2).

3. Answer generation, where findings are synthesized into comprehensive responses (RQ3).

A *purpose*, an *evaluation method*, and *evaluation criteria* are specified for each experiment tied to the research questions.

RQ 1 To what extent can KriRAG identify and retrieve documents relevant to information-needs?
Purpose: Information retrieval.
Evaluation Method: Automatic (metrics).
Evaluation Criteria: Precision, Recall, F1-score, and Mean Average Precision @k retrieved documents.

RQ 2 To what extent can KriRAG assess whether retrieved information answers the information-need, and what is the accuracy of these responses?
Purpose: Provide relevance reasoning.
Evaluation Method: Manual.
Evaluation Criteria:

- *Relevance Reasoning:* The ability to assess the relevance between the retrieved information and the query. Measured by summing instances of irrelevant reasoning.

- *Single Detail Fabrication*: Instances where one incorrect detail is fabricated.

- *Several Details Fabrication:* Instances where multiple incorrect details are fabricated.

- *Contextual Error:* Errors arising from misinterpreting the context of the information, or entirely off-topic.

- *Misinterpretation:* Instances of misinterpreting the meaning or content of the information, such as mixing up dates and names.

RQ 3 To what extent do KriRAG's answers to information-needs align with those of an investigator, and what is the accuracy of these responses?
Purpose: Provide answers to information-needs.
Evaluation Method: Manual.
Evaluation Criteria:

- *Highly Similar:* The system's ability to generate answers highly similar to the investigator's response.

- *Partial Overlap:* Instances where the system's response partially overlaps with the investigator's response.

- *Clear Divergence:* Instances of significant divergence between the system's and the investigator's responses.

- *Single Detail Fabrication*: Instances where one incorrect detail is fabricated.
- *Several Details Fabrication:* Instances where multiple incorrect details are fabricated.
- *Reasoning:* The logical conclusions of the system's generated output. Measured by summing irrational reasoning.
- *Coherence:* The consistency and flow of the generated output. Measured by summing incoherent responses.

6 Experiments and results

In this section, we present experiments designed to evaluate the performance of KriRAG in retrieving, reasoning, and generating answers in the context of real criminal investigations. Each experiment addresses a specific research question based on the provided information-needs formulated in Norwegian: 14 for case A, 12 for case B, and 4 for the open case. We evaluate retrieval metrics, relevance scoring, reasoning quality, levels of misinformation, and alignment of the generated answers with the investigators' conclusions. KriRAG's output is constrained to retrieve from a limited number of documents, k, due to the computing time and as a threshold for what is feasible with a thorough manual evaluation. We set $k = 100$ for the larger case files and $k = 17$ for the open case, the maximum number of relevant documents discovered by the investigator. We split longer documents into batches suitable for the preset context length, and the final number of documents (batches) may thus exceed 100 (denoted by N in tables). KriRAG's ability to set a relevance score is used in Experiments 1 and 3, denoted T, defined as a number from 0-3: irrelevant, somewhat relevant, relevant, and extremely relevant.[7]

6.1 Experiment 1 – Retrieval

We begin by setting the foundation for future experiments by measuring KriRAG's ability to retrieve relevant information. Retrieved documents are compared to those manually identified as answering information-needs by the investigators. We compute precision (P), recall (R), and F_1 scores, along with $P@k$, $R@k$, and Mean Average Precision (MAP@k) to assess performance at different k retrieved documents. MAP@k represents the mean precision of all queries, giving an idea of the general retrieval performance. Table 3 shows examples of considerable differences in the number of retrieved documents and metric results at varying k and thresholds for KriRAG's relevance score T in cases A and B. Without filtering ($T \geq 0$), recall is higher due to the number of documents, but $T \geq 3$ has a significant gain otherwise. For the *open case*, Table 4 shows superior MAP scores at $k = 1$ when applying $T \geq 3$ as a threshold. Full results are found in Appendix B.

Another example also highlights the effectiveness of T: Case B using the query "AA's involvement related to the murder". From the results in Table 5, we can observe that as T increases, precision drastically improves. With $T \geq 3$, KriRAG scores better at only $k = 5$ than an unconstrained k, returning 158 batches. There are only four relevant documents in the annotations.

6.2 Experiment 2 - Reasoning and Accuracy

In the second experiment, we evaluate *relevance reasoning* and the overall accuracy of the information generated in response to queries. These assessments are crucial for understanding how well the system supports investigative tasks.

A criminal investigator manually reviewed 518 outputs – 59 from the open case, 154 from Case A, and 305 from Case B. For Cases A and B, only outputs scoring *extremely relevant* ($T \geq 3$) were included in the evaluation to reduce the number of outputs to an amount suitable for manual evaluation. Each output was classified as either relevant or irrelevant based on its reasoning. KriRAG demonstrated strong performance in relevance reasoning, with only 13 out of 518 outputs marked as irrelevant, resulting in 97.5% accurate relevance assessments.

Using the same set of 518 outputs, a criminal investigator identified errors, on top of which an in-depth analysis was performed by counting the occurrence of errors based on the four categories described in Section 5.2: *single detail fabrication*, *several details fabrication*, *contextual errors* and *misinterpretations*. Table 6 provides examples and a breakdown of these errors. In total, there were 117 instances of misinformation, of which a majority are minor, e.g. misunderstanding vocabulary in cross-lingual settings. This results in an accuracy of 77.5%.

[7] The use of stronger adverbs (like *extremely*) resulted in a better distinction for scoring during testing.

Query	N	P@5	R@5	P@20	R@20	P@50	R@50	P	R	F_1
$T \geq 0$ (irrelevant) – Case A										
blue van	126	0.25	0.04	0.25	0.16	0.24	0.40	0.18	**0.64**	0.28
persons who were at [LOCATION-2] on the 12.02.14	99	0.00	0.00	0.06	0.17	0.04	0.33	0.04	**0.50**	0.07
persons with knowledge that FF was not supposed to attend [event]	119	**0.33**	**0.07**	0.33	0.27	0.17	0.40	0.12	**0.60**	0.19
$T \geq 3$ (extremely relevant) – Case A										
blue van	30	**0.60**	**0.12**	**0.40**	**0.32**	-	-	**0.30**	0.36	**0.33**
persons who were at [LOCATION-2] on the 12.02.14	5	**0.50**	**0.33**	-	-	-	-	**0.50**	0.33	**0.40**
persons with knowledge that FF was not supposed to attend [event]	15	0.25	**0.07**	-	-	-	-	**0.33**	0.27	**0.30**
$T \geq 0$ (irrelevant) – Case B										
GG's personality, his relationships, and social circle	192	0.00	0.00	0.17	0.07	0.21	0.18	0.14	**0.39**	0.21
rumors and stories about what happened to GG	196	0.00	0.00	0.20	0.08	0.13	0.12	0.15	**0.44**	**0.22**
what was the cause of death?	158	0.00	0.00	0.00	0.00	0.00	0.00	0.02	**0.33**	0.03
$T \geq 3$ (extremely relevant) – Case B										
GG's personality, his relationships, and social circle	48	**0.50**	**0.07**	**0.20**	**0.11**	-	-	**0.25**	0.29	**0.27**
rumors and stories about what happened to GG	34	**0.20**	**0.04**	**0.38**	**0.20**	-	-	**0.25**	0.20	**0.22**
what was the cause of death?	7	**0.25**	**0.33**	-	-	-	-	**0.17**	**0.33**	**0.22**

Table 3: Examples from Cases A and B, highlighting the change in performance as we filter by KriRAG's relevance score T. The highest values per query and case (for each k) are highlighted. While recall is often higher with more documents retrieved (N), relevant documents are retrieved at a much higher ratio per k after filtering.

KriRAG Relevance	MAP @1	MAP @3	MAP @5	MAP @8	MAP @12	MAP
$T \geq 0$	0.25	0.17	0.20	0.22	0.29	0.35
$T \geq 1$	0.25	0.17	0.25	0.25	0.34	0.37
$T \geq 2$	0.25	0.17	0.30	0.44	0.43	0.43
$T \geq 3$	**0.50**	**0.75**	**0.73**	**0.75**	**0.75**	**0.75**

Table 4: Mean Average Precision at k retrieved documents at varying KriRAG relevance thresholds for *Open case*. Highest values are highlighted.

KriRAG Relevance	N	P @5	R @5	P @20	R @20	P @50	R @50	P	R	F_1
$T \geq 0$	158	0.00	0.00	0.00	0.00	0.00	0.00	0.02	**0.50**	0.04
$T \geq 1$	89	0.00	0.00	0.00	0.00	0.07	**0.50**	0.04	**0.50**	0.07
$T \geq 2$	66	0.00	0.00	**0.20**	**0.50**	**0.08**	**0.50**	0.06	**0.50**	0.10
$T \geq 3$	9	**0.67**	**0.50**	-	-	-	-	**0.40**	**0.50**	**0.44**

Table 5: Results for various relevance thresholds for the query "AA's involvement related to the murder", enabling higher precision with fewer retrieved documents (N). Only four documents are annotated *relevant*.

6.3 Experiment 3 - Provide Answers

The third experiment validates the ability to provide overall answers to an information-need and the accuracy of these answers. This identifies the potential advantages and risks of implementing KriRAG. A criminal investigator manually compared KriRAG answers to investigators' answers. We received 30 answers from three dedicated investigators, which we categorized into *highly similar*, *partial overlap*, and *clear divergence* (described further in Section 5.2).

In parallel, we also performed an accuracy evaluation. Similar to experiment 3, we categorized errors into *single detail fabrication* and *several detail fabrication*. Moreover, we evaluated reasoning and coherence, as these answers are longer paragraphs. KriRAG returned 33 system-generated answers, whereas investigators did not answer the additional three. This experiment revealed substantial dissimilarity. The best results are from the open case, with 3 of 4 highly similar answers. For cases A and B, however, we see a much higher overlap rate and clear divergences, with only 2 *highly similar*. The expert-generated ground truth was characterized by their precision and depth of detail. In contrast, KriRAG tended towards more generalized responses. Table 7 shows an overview of the results. Finally, from the generated answers, we identified 13 outputs containing a single fabricated detail, 9 with *several*, but only 2 with irrational reasoning and 2 with incoherent responses.

7 Discussion

Investigations are often constrained by resource limitations, time pressure, and incomplete information. Consequently, by comparing KriRAG's outputs to human investigators, we assess its potential to complement current practices and pinpoint areas where it could support or enhance human efforts rather than outperform thorough investigative work.

Our experiments covered multiple tasks, from document retrieval to answering specific information-needs. Promising results were observed in the information retrieval phase, where KriRAG demonstrated 97.5% accuracy in

Type of misinformation	Occurrences	Example output	Explanation
Fabrication of a single detail	76/117	Stabbed with a knife	Document does not contain 'knife'
Fabrication of several details	13/117	He has a history of violent altercations, including a bar fight and a road rage incident.	Document does not mention these details
Contextual error	6/117	Document 10 is an autopsy report and would directly address the cause of death.	Document 10 is not an autopsy
Misinterpretation	22/117	Between 22:10 and 23:10.	Document mentions dates, not time.

Table 6: Types and occurrences of misinformation for Experiment 2. Excerpts from KriRAG summaries.

Case	Highly similar	Partial overlap	Clear Divergence
Open Case	3/4	1/4	0/4
A	0/14	6/14	8/14
B	2/12	4/12	6/12
Sum	5/30	11/30	14/30

Table 7: Result for similarity in experiment 4, showing the distribution of KriRAG answers in different similarity categories to the ground truth.

generating relevant reasoning. Additionally, 77.5% of query responses were deemed accurate (e.g., no fabrications). However, answering specific information-needs presented several challenges.

Sources of Error One notable issue was the replication of previously identified misinformation, leading to propagation in later stages due to the memory-based RAG architecture. One proposed improvement is incorporating self-correction in the memory module, e.g., by looking up generated information and verifying whether the seen documents contain the actions and events in memory. Reducing misinformation in this phase would significantly improve KriRAG's overall performance in answering the information-needs. In one case, objects such as a firearm and knife were mentioned despite not being part of the data. More concerning was an output such as "[...] in a dispute with the Black Skull gang over drug territory," a fabricated narrative with no supporting evidence in the documents. Such hallucinations must be addressed to ensure reliability and could be corrected in the memory module.

A potential cause of errors stems from inconsistent formatting of file IDs within the dataset (e.g., some using commas, others hyphens). Normalizing these IDs could help improve the accuracy of references and reduce errors during the memory retrieval process, not mistaking document identifiers as part of a date or another numeric value.

System Limitations in Criminal Investigations LLMs face difficulties when applied to the complex domain of criminal investigations, which requires dealing with a wide variety of data sources – ranging from reliable documents to unverified rumors and deliberate misinformation in interrogations. Investigators rely heavily on experience-based, tacit knowledge to evaluate the credibility of these sources (Ask, 2013; Fahsing, 2013). These factors are incredibly difficult for language models – especially engaging in "what if" reasoning (Adam and Carter, 2023) to identify the absence of critical information. KriRAG, in its current setup, is unable to discern the relative importance of different information sources. It failed to prioritize more credible sources in cases with conflicting information, such as suspects providing alibis that contradict evidence. Suspects claimed to be in one location, but evidence placed them elsewhere, and KriRAG treated both sources as equally reliable. With access to (or through generating) more metadata from the source documents, including separations on *who said what*, we can adjust the models' instructions accordingly, although with the chance of introducing bias.

We also often observe generic answers instead of detailed, case-specific information. Criminal investigations often require highly precise responses to information-needs, including personal identifiers of people, timestamps, and dates. Instead, we observe overly generalized answers, such as "several people had debt during different periods." We plan to refine prompts and perhaps introduce additional modules to control for graph-like knowledge of entities.

Languages and Model Performance While the datasets used in this study were entirely in Norwegian, we see better overall performance when managing prompts and outputs in English. This is consistent with findings from the ScandEval benchmark (Nielsen, 2023), where multilingual models perform better on Norwegian tasks (excerpts in Appendix D). However, we anticipate that models

trained specifically on Norwegian language data would yield even better results through improved vocabulary and understanding of cultural contexts. For now, however, language-specific models lag behind the massive efforts by larger corporations regarding instruction-following and other desired properties.

Finally, as generated outputs rely on the chosen LLM, the progress of model development plays a massive role in applying RAG for complex query-oriented tasks like criminal investigations. We have seen incredibly rapid development over the past year, with models from the Llama-3-, Phi-, and Gemma-series (Dubey et al., 2024; Abdin et al., 2024; Team et al., 2024) which have improved performance greatly over the alternatives just a year ago, like Llama-2 (Touvron et al., 2023). The system relies on an LLM through a local API, and we will continue evaluating new, emerging models.

8 Conclusion and Future Work

We have designed and evaluated KriRAG, a system that shows promise in enhancing the manual process of information analysis in criminal investigations. Our experiments reveal that KriRAG performs relevant reasoning in 97.5% of its outputs and achieves an accuracy of 77.5% for query responses. These results demonstrate its potential to reduce information overload and assist investigators in discovering information for thematization.

While KriRAG shows promise in addressing information-needs and queries, challenges concerning misinformation still need to be discussed. Even minor errors can seriously affect criminal investigations, where accuracy is a non-negotiable requirement. While often correct, our evaluation also identified several hallucinations, such as the example of a fabricated involvement of a gang in a case where no such information existed. These errors pose a threat not only to the investigation process but to fair justice. We aim to refine KriRAG by incorporating a validation or self-correcting layer based on a mixture of semantic textual similarity, keyword matching, and separate models for entity and coreference resolution. This layer would support the verification of outputs like "[...] in a dispute with the Black Skull gang [...]" by, e.g., validating entities against the source documents. This approach could provide a robust method for detecting and eliminating examples of misinformation before it impacts decision-making or propagates errors throughout further processing. We will address these challenges in future work. All code and evaluation outputs for the open case are available on an open-source repository, including a user interface and instructions for Docker images.[8]

Acknowledgements

We wish to express our appreciation to Robert Fleet and Mat Bettinson at the Digital Observatory at Queensland University of Technology. Your inspiration and knowledge laid the fundament for this study. This project is financed by the Norwegian Research Council (project numbers 333898, 322964, and 331893), Kripos, and NTNU. Thanks to the Center for Information Resilience (CIRES) at the University of Queensland. Thanks to The Norwegian Police IT Unit for technical facilitation and colleagues at Kripos for valuable feedback.

Limitations

In this study, we have primarily evaluated KriRAG using confidential data, which reduces transparency. This effort is to protect the persons involved and is part of the data protection agreement using confidential data. Thus, we also evaluate using an open case. Our experiments focused on Norwegian data, which may not fully represent the system's performance in other languages or domains. Some limitations may be addressed by integrating validation mechanisms in future work.

Ethical considerations

The Norwegian Director of Public Prosecutions and The Norwegian National Police Directorate have permitted access to data for this study. Further research is needed to understand potential risks when implementing KriRAG in criminal investigations; our study serves as a modest start, emphasizing the importance of considering both technological and investigative aspects. Thus, our study focuses on errors and limitations, and it is crucial to emphasize that KriRAG, like any AI-assisted tool, must not be considered inherently reliable and should always be subject to human oversight, verification, and rigorous testing before any operational deployment at Kripos or any law enforcement agency.

[8] https://github.com/tollefj/KriRAG

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

C Adam and Richard Carter. 2023. Large language models and intelligence analysis. *CETaS Expert Analysis*.

K Ask. 2013. Bias: Fejl og faldgruber i efterforskning. *Om at Opdage: Metodiske Refleksioner over Politiets Unders øgelsespraksis, pp. 149e169. Frederiksberg C: Samfundslitteratur*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Thierry S Barros, Carlos Eduardo S Pires, and Dimas Cassimiro Nascimento. 2023. Leveraging bert for extractive text summarization on federal police documents. *Knowledge and Information Systems*, 65(11):4873–4903.

Carlo Batini, Valerio Bellandi, Paolo Ceravolo, Federico Moiraghi, Matteo Palmonari, and Stefano Siccardi. 2021. Semantic data integration for investigations: Lessons learned and open challenges. In *2021 IEEE International Conference on Smart Data Services (SMDS)*, pages 173–183. IEEE.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2024. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36.

Chroma. 2024. Chroma. https://www.trychroma.com/. [Accessed 31-07-2024].

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Police Chief Online Europol. 2024. Policing in an ai-driven world. Available at https://www.policechiefmagazine.org/policing-ai-driven-world-europol/.

Ivar Fahsing. 2013. Tænkestile: Effektivitet, dyder og krydspress i efterforskninger. *Om at opdage–Metodiske refleksjoner over politiets undersøkelsespraksis*, pages 117–146.

Brandon L Garrett and Cynthia Rudin. 2023. Interpretable algorithmic forensics. *Proceedings of the National Academy of Sciences*, 120(41):e2301842120.

Lucie Gianola. 2020. *Aspects textuels de la procédure judiciaire exploitée en analyse criminelle et perspectives pour son traitement automatique*. Ph.D. thesis, Université de Cergy-Pontoise.

Aamir Hamid, Hemanth Reddy Samidi, Tim Finin, Primal Pappachan, and Roberto Yus. 2023. Genaipabench: A benchmark for generative ai-based privacy assistants. *arXiv preprint arXiv:2309.05138*.

Hans Henseler and Harm van Beek. 2023. Chatgpt as a copilot for investigating digital evidence.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a norwegian transformer model. *arXiv preprint arXiv:2104.09617*.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Heng Huang, Jiuxiang Gu, and Tianyi Zhou. 2023. Reflection-tuning: Data recycling improves llm instruction-tuning. *Preprint*, arXiv:2310.11716.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22266–22275.

Yu. A. Malkov and D. A. Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *Preprint*, arXiv:1603.09320.

Daniel D McCracken and Edwin D Reilly. 2003. Backus-naur form (bnf). In *Encyclopedia of Computer Science*, pages 129–131.

Gaëtan Michelet and Frank Breitinger. 2024. Chatgpt, llama, can you write my report? an experiment on assisted digital forensics reports written using (local) large language models. *Forensic Science International: Digital Investigation*, 48:301683.

Dan Saattrup Nielsen. 2023. Scandeval: A benchmark for scandinavian natural language processing. *Preprint*, arXiv:2304.00906.

Alf Nissen. 2018. *Informasjonsbehandling i store etterforskningssaker*.

Homicide Working Group NPCC. 2021. Major incident room standardised administrative procedures (mirsap 2021).

Justin Partridge and Fatema Zaghloul. 2023. Policing the data: Can data analytics help law enforcement? *Journal of Information Technology Teaching Cases*, page 20438869231212214.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Fillipe Barros Rodrigues, William Ferreira Giozza, Robson de Oliveira Albuquerque, and Luis Javier García Villalba. 2022. Natural language processing applied to forensics information extraction with transformers and graph visualization. *IEEE Transactions on Computational Social Systems*.

D Kim Rossmo. 2021. Dissecting a criminal investigation. *Journal of Police and Criminal Psychology*, 36(4):639–651.

Cheol Ryu, Seolhwa Lee, Subeen Pang, Chanyeol Choi, Hojun Choi, Myeonggee Min, and Jy-Yong Sohn. 2023. Retrieval-based evaluation for llms: A case study in korean legal qa. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 132–137.

Mark Scanlon, Frank Breitinger, Christopher Hargreaves, Jan-Niclas Hilgert, and John Sheppard. 2023. Chatgpt for digital forensic investigation: The good, the bad, and the unknown. *Forensic Science International: Digital Investigation*, 46:301609.

Christian A Schiller. 2024. The human factor in detecting errors of large language models: A systematic literature review and future research directions. *arXiv preprint arXiv:2403.09743*.

Mads Skipanes, Tollef Jørgensen, and Katrin Franke. 2023. *Advancing Knowledge Discoveries in Criminal Investigations with Semantic Textual Similarity*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. *arXiv preprint arXiv:2405.07437*.

Jie Zhao, Ziyu Guan, Wei Zhao, and Yue Jiang. 2024. Enhancing criminal case matching through diverse legal factors. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2379–2383.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*.

A Prompts

A.1 Main KriRAG prompt

```
prompt = """You are an AI assisting a
    ↪ criminal investigation,
    ↪ analyzing case files for
    ↪ knowledge discoveries. You
    ↪ follow strict logical and
    ↪ deductive reasoning, and will
    ↪ only present information for
    ↪ which you have a complete
    ↪ overview of. Do not make
    ↪ assumptions, or add any
    ↪ superfluous information.
You have info from previous
    ↪ interrogations: '{MEMORY}'. Use
    ↪ this info to guide your
    ↪ reasoning if relevant.

You receive a new document with ID
{DOC_ID}: '{DOC_TEXT}'.
Investigate document {DOC_ID} grounded
    ↪ in the QUERY: '{query}'.
Generate a JSON object with
1) questions: a list of investigative
    ↪ questions(based on e.g.,
    ↪ objects, actions, events,
    ↪ entities) that are directly
    ↪ related to the QUERY in {DOC_ID}.
2) reason: discuss whether document
    ↪ {DOC_ID} answers the QUERY.
3) score: if the document is 0
    ↪ irrelevant, 1 somewhat relevant,
    ↪ 2 relevant, or 3 extremely
    ↪ relevant.
4) a summary of vital details uncovered
    ↪ in {DOC_ID}."""

schema = {
  "type": "object",
  "properties": {
    "questions": {
        "type": "array",
        "minItems": 1,
        "maxItems": 3,
        "items": {
            "type": "object",
            "properties": {
                "question": {
                    "type": "string"
                }
            },
            "required": ["question"],
        },
    },
    "reason": {"type": "string"},
    "score": {
        "type": "integer",
        "enum": [0, 1, 2, 3]
    },
    "summary": {"type": "string"},
    },
    "required": [
        "questions", "reason",
        "score", "summary",
    ],
}
```

Different prompts are used in KriRAG's separate processes. We create JSON schemas which are converted to GBNF – a format based around Backer-Naur Form (McCracken and Reilly, 2003), supported by the llama.cpp-API.[9] This yields accurate and controllable prompting with no errors or deviations from the JSON format. The memory summarization is kept as text only.

A.2 Memory Summarization

```
memsumm_prompt = """You are an AI
    ↪ assisting a criminal
    ↪ investigation, analyzing case
    ↪ files for knowledge discoveries.
    ↪ You follow strict logical and
    ↪ deductive reasoning.
From the following data:
{CURRENT_MEMORY}
create a summary of vital information
    ↪ related to the query: '{query}'.
Make sure to reference the ID {DOC_ID},
    ↪ and keep all previous document
    ↪ references."""
```

A.3 Meta-Summary

```
meta_summary_prompt = """You are an AI
    ↪ assisting a criminal
    ↪ investigation, analyzing case
    ↪ files. You follow strict logical
    ↪ and deductive reasoning, and
    ↪ will only present information
    ↪ for which you have a complete
    ↪ overview of. Avoid assumptions
    ↪ and uncertainty. Do not repeat
    ↪ yourself.
You receive the following information:
'{MEMORIES_FROM_QUERY}'.
Assess the relevance of each document
    ↪ to the query '{QUERY}' and write
    ↪ a highly detailed summary
    ↪ (including involved persons,
    ↪ objects, locations and other
    ↪ entities), based on the most
    ↪ relevant documents. Return a
    ↪ JSON object with the summary and
    ↪ references to the most relevant
    ↪ documents."""

schema = {
    "type": "object",
    "properties": {
        "summary": {"type": "string"},
        "references": {
            "type": "array",
            "items": {"type": "string"},
        }
    },
    "required": ["summary",
    ↪ "references"],
}
```

[9] https://github.com/ggerganov/llama.cpp/blob/9379d3c/grammars/README.md

Query – Open Case	N	P@5	R@5	P@10	R@10	P	R	F_1
$T \geq 0$								
details about the murder weapon (what is the murder weapon?)	16	0.00	0.00	0.00	0.00	0.19	0.60	0.29
how did the victim die (what is the cause of death?)	15	0.20	0.06	0.40	0.24	0.33	0.29	0.31
persons with residence and connections to the address (the crime scene) as owner, tenant, visitor, etc.	14	0.60	0.18	0.70	0.41	0.79	0.65	0.71
the victim's involvement in conflict or argument prior to death	14	0.00	0.00	0.00	0.00	0.07	0.20	0.11
$T \geq 1$								
details about the murder weapon (what is the murder weapon?)	14	0.00	0.00	0.10	0.20	0.21	0.60	0.32
how did the victim die (what is the cause of death?)	14	0.40	0.12	0.40	0.24	0.36	0.29	0.32
persons with residence and connections to the address (the crime scene) as owner, tenant, visitor, etc.	14	0.60	0.18	0.70	0.41	0.79	0.65	0.71
the victim's involvement in conflict or argument prior to death	9	0.00	0.00	-	-	0.11	0.20	0.14
$T \geq 2$								
details about the murder weapon (what is the murder weapon?)	8	0.20	0.20	-	-	0.38	0.60	0.46
how did the victim die (what is the cause of death?)	12	0.40	0.12	0.40	0.24	0.42	0.29	0.34
persons with residence and connections to the address (the crime scene) as owner, tenant, visitor, etc.	11	0.60	0.18	0.80	0.47	0.82	0.53	0.64
the victim's involvement in conflict or argument prior to death	8	0.00	0.00	-	-	0.12	0.20	0.15
$T \geq 3$								
details about the murder weapon (what is the murder weapon?)	1	-	-	-	-	1.00	0.20	0.33
how did the victim die (what is the cause of death?)	6	0.60	0.18	-	-	0.67	0.24	0.35
persons with residence and connections to the address (the crime scene) as owner, tenant, visitor, etc.	2	-	-	-	-	1.00	0.12	0.21
the victim's involvement in conflict or argument prior to death	3	-	-	-	-	0.33	0.20	0.25

Table 8: Metrics @k retrieved for the Open Case, separated on scoring thresholds.

B Metrics

Metrics for all cases are presented in detail in Tables 8, 9 and 10. In each table, Precision (P), Recall (R), F_1-score (F_1) represent the average over retrieved documents, along with "@k", demonstrating KriRAGs' ability to retrieve and score documents accordingly in constrained settings. Furthermore, we show the results for each threshold T corresponding to the score provided by KriRAG.

B.1 Mean Average Precision

Tables 11, 12 and 13 shows MAP@k values for different KriRAG relevance score thresholds of a minimum T value. k is set to 100 for cases A and B and 17 for the open case (the maximum number of discovered documents relevant to a single information-need). The system may discover $> k$ documents due to batching documents exceeding the LLM context length. If all documents contained 10,000 tokens, they would be split into 10 with a context length of 1,000, thus increasing k by a multiple of 10.

Query – Case A	N	P@5	R@5	P@20	R@20	P@50	R@50	P	R	F_1
$T \geq 0$ (irrelevant)										
blue van	126	0.25	0.04	0.25	0.16	0.24	0.40	0.18	0.64	0.28
person with access to a reward card from Esso	119	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.29	0.04
persons in financial distress, money problems, debt	146	0.00	0.00	0.00	0.00	0.08	0.15	0.05	0.23	0.08
persons who passed by the water between 00:30 and 02:30	110	0.00	0.00	0.05	0.01	0.08	0.03	0.09	0.08	0.08
persons who were at [LOCATION-1] on 12.02.14	118	0.00	0.00	0.06	0.11	0.09	0.44	0.05	0.44	0.09
persons who were at [LOCATION-2] on the 12.02.14	99	0.00	0.00	0.06	0.17	0.04	0.33	0.04	0.50	0.07
persons who were at the event on 12.02.14	117	0.20	0.04	0.11	0.08	0.09	0.16	0.05	0.16	0.07
persons with [BRAND] shoes, size 42	120	0.00	0.00	0.33	0.06	0.49	0.23	0.26	0.31	0.28
persons with knowledge that FF was not supposed to attend [event]	119	0.33	0.07	0.33	0.27	0.17	0.40	0.12	0.60	0.19
planning to commit a robbery of ...	162	0.00	0.00	0.00	0.00	0.10	0.23	0.08	0.54	0.14
$T \geq 1$ (somewhat relevant)										
blue van	89	0.33	0.04	0.28	0.20	0.26	0.48	0.20	0.64	0.30
person with access to a reward card from Esso	92	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.29	0.05
persons in financial distress, money problems, debt	123	0.00	0.00	0.00	0.00	0.08	0.15	0.05	0.23	0.08
persons who passed by the water between 00:30 and 02:30	75	0.20	0.01	0.15	0.03	0.12	0.05	0.12	0.08	0.09
persons who were at [LOCATION-1] on 12.02.14	67	0.00	0.00	0.05	0.11	0.09	0.44	0.06	0.44	0.11
persons who were at [LOCATION-2] on the 12.02.14	60	0.33	0.17	0.06	0.17	0.04	0.33	0.05	0.50	0.10
persons who were at the event on 12.02.14	107	0.20	0.04	0.11	0.08	0.09	0.16	0.05	0.16	0.07
persons with [BRAND] shoes, size 42	97	0.00	0.00	0.42	0.10	0.48	0.28	0.27	0.31	0.29
persons with knowledge that FF was not supposed to attend [event]	90	0.50	0.07	0.31	0.33	0.15	0.40	0.14	0.60	0.23
planning to commit a robbery of ...	119	0.00	0.00	0.06	0.08	0.15	0.38	0.09	0.54	0.16
$T \geq 2$ (relevant)										
blue van	65	0.50	0.08	0.32	0.24	0.29	0.56	0.27	0.64	0.38
person with access to a reward card from Esso	42	0.00	0.00	0.00	0.00	-	-	0.05	0.29	0.09
persons in financial distress, money problems, debt	79	0.00	0.00	0.00	0.00	0.06	0.15	0.06	0.23	0.10
persons who passed by the water between 00:30 and 02:30	34	0.00	0.00	0.20	0.03	-	-	0.18	0.05	0.08
persons who were at [LOCATION-1] on 12.02.14	27	0.25	0.11	0.22	0.44	-	-	0.16	0.44	0.24
persons who were at [LOCATION-2] on the 12.02.14	25	0.25	0.17	0.11	0.33	-	-	0.13	0.50	0.21
persons who were at the event on 12.02.14	80	0.25	0.04	0.16	0.12	0.09	0.16	0.06	0.16	0.09
persons with [BRAND] shoes, size 42	63	0.20	0.01	0.55	0.14	0.40	0.26	0.34	0.27	0.30
persons with knowledge that FF was not supposed to attend [event]	73	0.50	0.07	0.29	0.33	0.21	0.53	0.17	0.60	0.26
planning to commit a robbery of ...	65	0.00	0.00	0.15	0.15	0.19	0.46	0.16	0.54	0.25
$T \geq 3$ (extremely relevant)										
blue van	30	0.60	0.12	0.40	0.32	-	-	0.30	0.36	0.33
person with access to a reward card from Esso	3	-	-	-	-	-	-	0.00	0.00	0.00
persons in financial distress, money problems, debt	26	0.00	0.00	0.20	0.23	-	-	0.15	0.23	0.18
persons who passed by the water between 00:30 and 02:30	2	-	-	-	-	-	-	0.50	0.01	0.02
persons who were at [LOCATION-1] on 12.02.14	4	-	-	-	-	-	-	0.25	0.11	0.15
persons who were at [LOCATION-2] on the 12.02.14	5	0.50	0.33	-	-	-	-	0.50	0.33	0.40
persons who were at the event on 12.02.14	17	0.20	0.04	-	-	-	-	0.12	0.08	0.10
persons with [BRAND] shoes, size 42	17	0.80	0.05	-	-	-	-	0.47	0.10	0.17
persons with knowledge that FF was not supposed to attend [event]	15	0.25	0.07	-	-	-	-	0.33	0.27	0.30
planning to commit a robbery of ...	13	0.50	0.15	-	-	-	-	0.38	0.23	0.29

Table 9: Metrics @k retrieved for Case A, separated on scoring thresholds. The last columns include all outputs.

Query – Case B	N	P@5	R@5	P@20	R@20	P@50	R@50	P	R	F_1
$T \geq 0$										
AA's involvement related to the murder	158	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.50	0.04
Descriptions of the murder weapon's functions, defects, modifications, and testing	115	0.00	0.00	0.00	0.00	0.04	0.25	0.02	0.25	0.03
G's involvement in conflicts that might shed light on why he was killed	146	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.14	0.09
GG's personality, his relationships, and social circle	192	0.00	0.00	0.17	0.07	0.21	0.18	0.14	0.39	0.21
HH's involvement with the murder weapon, modifications to the weapon, and test firing	130	0.00	0.00	0.00	0.00	0.00	0.00	0.03	1.00	0.06
HH's personality, his relationships, and social circle	153	0.00	0.00	0.14	0.05	0.20	0.21	0.17	0.58	0.26
NN's personality, his relationships, and social circle	162	0.00	0.00	0.00	0.00	0.17	0.22	0.19	0.72	0.31
What plans did HH, NN, and FF have on January 10?	138	0.00	0.00	0.00	0.00	0.03	0.25	0.03	0.50	0.06
information about searches for weapons and seized weapons	133	0.25	0.10	0.24	0.40	0.19	0.60	0.10	0.70	0.18
persons with access to firearms	140	0.00	0.00	0.00	0.00	0.23	0.14	0.33	0.49	0.40
rumors and stories about what happened to GG	196	0.00	0.00	0.20	0.08	0.13	0.12	0.15	0.44	0.22
the time of death of the deceased	161	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.25	0.02
what was the cause of death?	158	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.33	0.03
$T \geq 1$										
AA's involvement related to the murder	89	0.00	0.00	0.00	0.00	0.07	0.50	0.04	0.50	0.07
Descriptions of the murder weapon's functions, defects, modifications, and testing	80	0.00	0.00	0.00	0.00	0.03	0.25	0.02	0.25	0.03
G's involvement in conflicts that might shed light on why he was killed	130	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.14	0.09
GG's personality, his relationships, and social circle	175	0.20	0.04	0.17	0.07	0.20	0.18	0.14	0.39	0.21
HH's involvement with the murder weapon, modifications to the weapon, and test firing	106	0.00	0.00	0.00	0.00	0.00	0.00	0.04	1.00	0.07
HH's personality, his relationships, and social circle	143	0.00	0.00	0.14	0.05	0.20	0.21	0.17	0.58	0.27
NN's personality, his relationships, and social circle	134	0.00	0.00	0.00	0.00	0.18	0.28	0.19	0.72	0.31
What plans did HH, NN, and FF have on January 10?	65	0.00	0.00	0.06	0.25	0.06	0.50	0.05	0.50	0.08
information about searches for weapons and seized weapons	94	0.20	0.10	0.30	0.60	0.15	0.60	0.11	0.70	0.19
persons with access to firearms	121	0.00	0.00	0.09	0.03	0.21	0.14	0.33	0.49	0.40
rumors and stories about what happened to GG	167	0.00	0.00	0.17	0.08	0.12	0.12	0.15	0.44	0.22
the time of death of the deceased	109	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.25	0.03
what was the cause of death?	115	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.33	0.03
$T \geq 2$										
AA's involvement related to the murder	66	0.00	0.00	0.20	0.50	0.08	0.50	0.06	0.50	0.10
Descriptions of the murder weapon's functions, defects, modifications, and testing	55	0.00	0.00	0.00	0.00	0.03	0.25	0.03	0.25	0.05
G's involvement in conflicts that might shed light on why he was killed	97	0.00	0.00	0.00	0.00	0.05	0.05	0.08	0.14	0.10
GG's personality, his relationships, and social circle	152	0.20	0.04	0.17	0.07	0.19	0.18	0.15	0.39	0.22
HH's involvement with the murder weapon, modifications to the weapon, and test firing	82	0.00	0.00	0.00	0.00	0.09	1.00	0.06	1.00	0.11
HH's personality, his relationships, and social circle	126	0.00	0.00	0.22	0.11	0.17	0.21	0.18	0.58	0.27
NN's personality, his relationships, and social circle	101	0.00	0.00	0.20	0.17	0.20	0.33	0.24	0.72	0.36
What plans did HH, NN, and FF have on January 10?	13	0.50	0.25	-	-	-	-	0.11	0.25	0.15
information about searches for weapons and seized weapons	75	0.20	0.10	0.30	0.60	0.17	0.60	0.13	0.70	0.22
persons with access to firearms	87	0.00	0.00	0.15	0.06	0.39	0.34	0.35	0.49	0.41
rumors and stories about what happened to GG	121	0.00	0.00	0.20	0.12	0.21	0.24	0.19	0.44	0.26
the time of death of the deceased	51	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
what was the cause of death?	60	0.00	0.00	0.00	0.00	0.03	0.33	0.03	0.33	0.05
$T \geq 3$										
AA's involvement related to the murder	9	0.67	0.50	-	-	-	-	0.40	0.50	0.44
Descriptions of the murder weapon's functions, defects, modifications, and testing	12	0.00	0.00	-	-	-	-	0.00	0.00	0.00
G's involvement in conflicts that might shed light on why he was killed	9	0.50	0.05	-	-	-	-	0.29	0.05	0.08
GG's personality, his relationships, and social circle	48	0.50	0.07	0.20	0.11	-	-	0.25	0.29	0.27
HH's involvement with the murder weapon, modifications to the weapon, and test firing	46	0.00	0.00	0.09	0.50	-	-	0.10	1.00	0.18
HH's personality, his relationships, and social circle	33	0.50	0.11	0.25	0.16	-	-	0.30	0.32	0.31
NN's personality, his relationships, and social circle	16	0.40	0.11	-	-	-	-	0.31	0.22	0.26
information about searches for weapons and seized weapons	46	0.20	0.10	0.22	0.40	-	-	0.14	0.50	0.22
persons with access to firearms	42	0.00	0.00	0.38	0.17	-	-	0.32	0.26	0.29
rumors and stories about what happened to GG	34	0.20	0.04	0.38	0.20	-	-	0.25	0.20	0.22
the time of death of the deceased	3	-	-	-	-	-	-	0.00	0.00	0.00
what was the cause of death?	7	0.25	0.33	-	-	-	-	0.17	0.33	0.22

Table 10: Metrics @k retrieved for Case B, separated on scoring thresholds. The last columns include all outputs.

T	MAP@1	MAP@3	MAP@5	MAP@8	MAP@12	MAP
0	0.25	0.17	0.20	0.22	0.29	0.35
1	0.25	0.17	0.25	0.25	0.34	0.37
2	0.25	0.17	0.30	0.44	0.43	0.43
3	0.50	0.75	0.73	0.75	0.75	0.75

Table 11: MAP for the Open case for each relevance threshold (T), constrained to k retrieved documents.

T	MAP@1	MAP@3	MAP@5	MAP@8	MAP@13	MAP@21	MAP@35	MAP@55	MAP@89	MAP@100	MAP@144	MAP
0	0.00	0.05	0.08	0.10	0.12	0.13	0.15	0.14	0.11	0.11	0.10	0.09
1	0.00	0.10	0.16	0.18	0.15	0.16	0.17	0.14	0.11	0.11	0.11	0.11
2	0.00	0.13	0.20	0.19	0.21	0.21	0.20	0.17	0.16	0.16	0.16	0.16
3	0.30	0.35	0.36	0.38	0.34	0.31	0.30	0.30	0.30	0.30	0.30	0.30

Table 12: MAP for Case A for each relevance threshold (T), constrained to k retrieved documents.

T	MAP@1	MAP@3	MAP@5	MAP@8	MAP@13	MAP@21	MAP@35	MAP@55	MAP@89	MAP@100	MAP@144	MAP
0	0.00	0.00	0.02	0.02	0.04	0.07	0.09	0.09	0.11	0.11	0.11	0.10
1	0.00	0.00	0.03	0.02	0.05	0.08	0.10	0.10	0.11	0.12	0.10	0.10
2	0.08	0.04	0.07	0.06	0.06	0.12	0.13	0.13	0.13	0.13	0.12	0.12
3	0.08	0.22	0.27	0.21	0.20	0.23	0.22	0.21	0.21	0.21	0.21	0.21

Table 13: MAP for Case B for each relevance threshold (T), constrained to k retrieved documents.

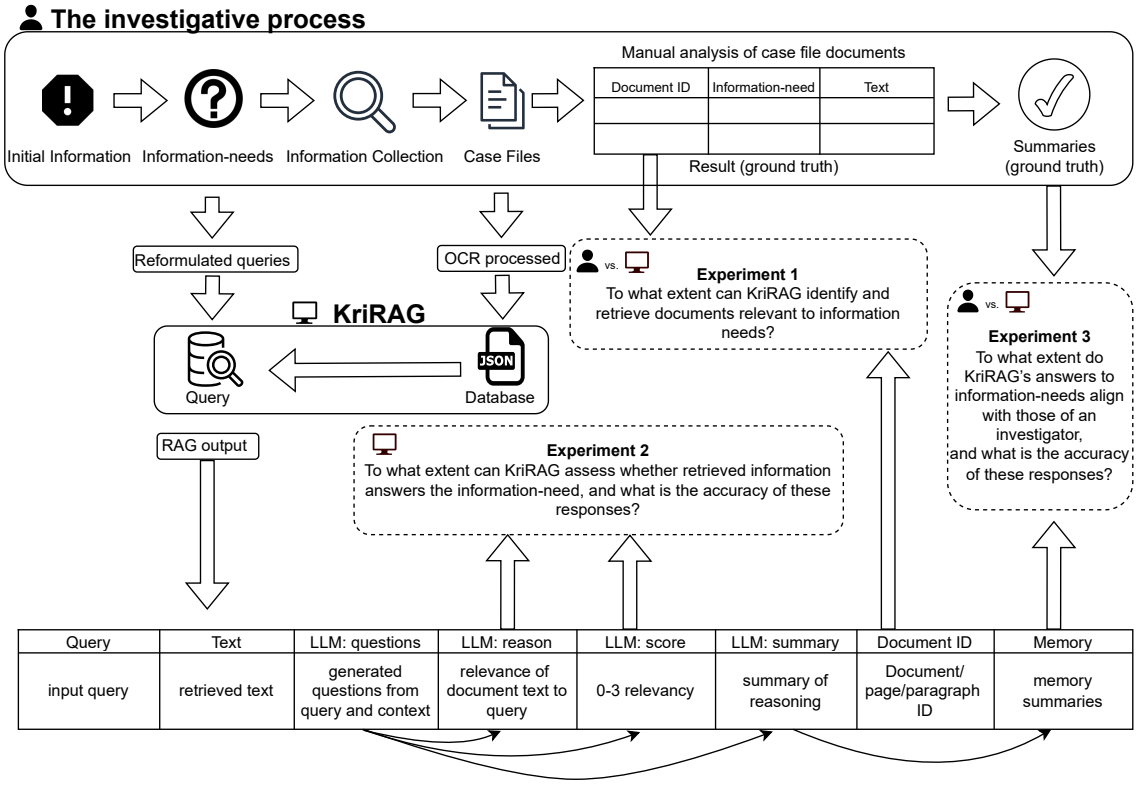

Figure 3: Overview of study. Shows the investigative process, KriRAG system outputs, and experiments.

C System Overview

Figure 3 shows the an overview of processes involved in our methodology and system architecture, including the investigative process from the initial information to how investigators combine information, resulting in the ground truth used to evaluate the three experiments described in Section 6.

Model ID	Param (M)	Vocab (k)	Context Length	Rank (score)
meta-llama/Meta-Llama-3-70B	70554	128	8192	1.45
google/gemma-2-27b-it	27227	256	8193	1.63
google/gemma-2-9b-it	9242	256	8193	1.86
meta-llama/Meta-Llama-3.1-8B-Instruct	8030	128	131073	2.56
mistralai/Mistral-7B-v0.3	7248	33	32768	2.81
google/gemma-7b	8538	256	8067	2.83
microsoft/Phi-3-mini-4k-instruct	3821	32	4096	3.47
NorwAI/NorwAI-Mistral-7B	7537	68	4096	3.52
norallm/normistral-7b-warm-instruct	7248	33	2048	3.58

Table 14: Excerpt from ScandEval for generative Norwegian tasks (as of November 2024), comparing various open-weight models, sorted from best to worst performance (lower rank is better). Models trained on Norwegian data are highlighted. Fine-tuned models on top of official models are omitted, as training data cannot be verified.

D ScandEval Excerpt

Table 14 shows various open-weight models as evaluated by the ScandEval benchmark (Nielsen, 2023). Full results available on the ScandEval website.[10]

[10] https://scandeval.com/norwegian-nlg/