Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Design and Evaluation of a Secure Air-Gapped LLM-Based Intelligence Analysis Framework

Anand Rawat

anandrawat138a@gmail.com

Pranveer Singh Institute of Technology Kanpur https://orcid.org/0009-0006-8997-0143

Research Article

Keywords: Large Language Models, Air-Gapped Systems, Retrieval-Augmented Generation, Secure Analyt- ics, Human-in-the-Loop

Posted Date: March 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9016380/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

The increasing reliance on heterogeneous and high-volume data sources in security-sensitive analytical en- vironments has amplified the cognitive burden placed on human analysts. While large language models (LLMs) demonstrate strong capabilities in semantic reasoning and multi-document synthesis, their deployment in regu- lated and air-gapped settings is constrained by data sovereignty requirements, auditability concerns, and the risk of hallucinated outputs.

This paper presents the design and empirical evaluation of a secure, air-gapped intelligence analysis frame- work integrating a high-capacity (400B-parameter class) on-premise LLM with a retrieval-augmented generation (RAG) pipeline and enforced human-in-the-loop oversight. The proposed architecture operates entirely offline and restricts model responses to access-controlled retrieved context, supported by immutable audit logging and structured validation workflows.

Experimental evaluation across multi-document analytical tasks demonstrates that retrieval-augmented ground- ing reduces hallucination rate from 38.7% to 12.2% while maintaining interactive response latency. An ablation study further analyses the impact of retrieval depth on grounding accuracy and latency trade-offs. Results indi- cate that architectural grounding mechanisms contribute more substantially to reliability than model scale alone, supporting the feasibility of secure, high-capacity LLM deployment in regulated environments.

The findings highlight the importance of system-level constraints, contextual grounding, and enforced human oversight in achieving trustworthy LLM-assisted decision support within fully air-gapped operational contexts.

1 Introduction

The increasing volume, velocity, and heterogeneity of digital information have placed unprecedented demands on human analysts operating in regulated and security-sensitive environments. Analytical workflows today routinely involve the examination of unstructured text reports, logs, communications, and archival documents originating from multiple independent sources. While the availability of such data has expanded significantly, the ability of analysts to efficiently synthesise, correlate, and interpret this information remains constrained by human cognitive limits.

Recent advances in large language models (LLMs) have demonstrated strong performance in tasks such as document summarization, semantic search, and contextual reasoning. These capabilities present an opportunity to augment human analysis by automating low-level cognitive tasks and accelerating information synthesis. How- ever, most existing LLM-based solutions assume cloud connectivity, opaque inference pipelines, and limited guar- antees of data governance. Such assumptions are incompatible with analytical environments that require strict data sovereignty, offline operation, and post-hoc auditability of system outputs.

A further limitation of unconstrained generative models is the risk of hallucinated or unsupported outputs. In analytical contexts, even minor inaccuracies or unverified assertions can lead to incorrect conclusions and erosion of trust in automated systems. As a result, any practical deployment of LLMs in regulated domains must prioritise grounded reasoning, transparency, and enforceable human oversight over raw generative capability.

*Corresponding author. Email: anandrawat138a@gmail.com

To address these challenges, this paper proposes a secure, air-gapped analytical framework that integrates an on-premise large language model with a retrieval-augmented generation (RAG) pipeline. The framework is de- signed to operate entirely without external network connectivity and to enforce human-in-the-loop constraints at the architectural level. Rather than acting as an autonomous decision-maker, the system functions as a decision- support tool that assists analysts by retrieving relevant documents, synthesising contextual summaries, and pre- senting traceable analytical outputs.

Figure 1 illustrates the high-level workflow of the proposed framework. Incoming documents are first ingested and indexed in a local vector database. Analyst queries trigger a controlled retrieval process, after which the LLM generates responses strictly grounded in retrieved context. All interactions are logged to support transparency and auditability.

Unlike conventional keyword-based search systems or unconstrained generative models, the proposed frame- work restricts model outputs to retrieved and authorised content. If relevant contextual information is unavailable, the system is designed to explicitly indicate uncertainty rather than generate speculative responses. This design choice reduces the risk of hallucinations and supports analyst trust in system outputs.

The primary contributions of this work are threefold. First, it presents a security-centric system architecture for deploying LLMs in fully offline environments without reliance on external services. Second, it demonstrates the integration of retrieval-augmented generation with enforced human-in-the-loop constraints to improve output grounding and transparency. Third, it provides an experimental evaluation of the proposed framework using

simulated analytical datasets, highlighting its effectiveness in reducing unsupported outputs while maintaining interactive response times.

The remainder of this paper is organised as follows. Section 2 reviews related work on LLM-based analytical systems and secure AI deployments. Section 3 details the system architecture and design principles. Section 4 describes the experimental setup and evaluation methodology. Section 5 presents and discusses the experimental results. Finally, Section 6 concludes the paper and outlines directions for future work.

2 Related Work

The rapid advancement of large language models has led to widespread exploration of their applicability across a variety of analytical and decision-support tasks. Transformer-based architectures have demonstrated strong capa- bilities in natural language understanding, summarization, and semantic reasoning, enabling automated processing of large volumes of unstructured text [1, 2]. These developments have motivated research into deploying LLMs as assistants for knowledge-intensive tasks rather than as standalone autonomous systems.

One major limitation of generative language models is their tendency to produce hallucinated or unsupported outputs when operating without external grounding. Several studies have highlighted that, despite fluent language generation, LLMs may generate factually incorrect or unverifiable information, particularly in high-stakes analyt- ical contexts [3]. This has led to increased interest in retrieval-augmented generation (RAG) approaches, in which model responses are conditioned on retrieved documents rather than solely on internal model parameters.

Retrieval-augmented generation was formalised as a framework for grounding language model outputs in external knowledge sources, demonstrating improvements in factual consistency and interpretability [4]. Subse- quent work has extended this concept to enterprise search, question answering, and document analysis systems, showing that retrieval-based grounding significantly improves reliability and user trust [5]. These approaches are particularly relevant for analytical environments where traceability and source attribution are required.

Parallel to advances in RAG, a growing body of research has focused on the challenges of deploying AI systems in regulated and security-sensitive domains. Studies on trustworthy and responsible AI emphasise the importance of transparency, auditability, and human oversight when integrating machine learning systems into decision-support workflows [6, 7]. Human-in-the-loop (HITL) frameworks have been shown to reduce automation bias and improve system acceptance by ensuring that final judgement remains with human operators [8].

Several works have examined the deployment of AI systems in offline or resource-constrained environments, motivated by concerns related to data sovereignty, latency, and operational resilience [9]. However, much of the existing literature assumes partial cloud connectivity or does not explicitly address the architectural constraints imposed by fully air-gapped operation. As a result, questions related to audit logging, access-controlled retrieval, and enforced non-autonomy are often treated as secondary concerns.

Unlike prior work that focuses primarily on model performance or algorithmic improvements, the present study emphasises a system-level architecture for secure, offline deployment of LLM-assisted analytical tools. By integrating retrieval-augmented generation with explicit human-in-the-loop enforcement and audit-centric design principles, the proposed framework addresses gaps in existing approaches related to grounded reasoning, trans- parency, and operational control in regulated analytical environments.

3 System Architecture

The proposed framework is designed as a modular, air-gapped analytical system that integrates a large language model with retrieval-augmented reasoning and enforced human oversight. The architecture prioritises offline operation, traceability of outputs, and controlled interaction between system components. Rather than maximising automation, the design emphasises reliability, transparency, and analyst trust.

Architectural Overview

At a high level, the system is organised into four logical layers: (i) secure data ingestion and storage, (ii) retrieval and knowledge management, (iii) language model inference, and (iv) analyst interaction and oversight. Each layer is isolated through explicit interfaces to reduce unintended coupling and to support auditability.

All components operate entirely within a closed network environment. No external connectivity is required for model inference, document retrieval, or user interaction. This design ensures that data remains under local control and that system behaviour can be inspected and validated post hoc.

Data Layer Retrieval Layer Interaction Layer

Figure 2 presents the detailed architecture and interaction flow between system components.

Secure Data Ingestion and Storage

The data layer is responsible for ingesting and storing analytical documents originating from authorised sources. Documents may include structured reports, unstructured text, or archival records. Prior to storage, all inputs undergo basic sanitisation and format normalisation to ensure consistency and to prevent unintended execution or metadata leakage.

Stored documents are maintained in a local repository under strict access control. The ingestion layer performs no interpretation or inference; its sole function is to preserve document integrity and provenance for downstream processing.

Retrieval and Knowledge Management

To support grounded reasoning, the system employs a retrieval-augmented generation pipeline. Documents are segmented into semantically coherent units and transformed into vector representations using an embedding model executed locally. These embeddings are stored in a vector database that supports similarity-based retrieval.

When an analyst submits a query, the retrieval layer selects the most relevant document segments based on semantic similarity and access constraints. This retrieved context forms the sole knowledge source provided to the language model, ensuring that generated responses remain grounded in available data.

Retrieval Configuration

Documents were segmented into semantically coherent chunks prior to embedding to balance retrieval precision and contextual completeness. During query execution, a fixed number of top-ranked document segments were retrieved based on vector similarity and provided as contextual input to the language model. Similarity-based retrieval was selected due to its robustness across heterogeneous document types. Retrieval parameters were kept constant across all experiments to ensure comparability between system configurations.

On-Premise Language Model Inference

The language model layer hosts the LLM used for analytical reasoning and summarization. The model is deployed entirely on local compute resources and has no direct access to raw document stores or system commands. Its inputs are limited to analyst queries and retrieved contextual passages.

If insufficient or low-confidence context is available, the model is designed to explicitly indicate uncertainty rather than produce speculative outputs. This behaviour reduces the risk of hallucination and supports analyst trust.

Analyst Interaction and Human Oversight

The interaction layer provides the analyst-facing interface through which queries are submitted and responses are reviewed. All AI-generated outputs are advisory and must pass through an explicit human review step before

being used for reporting or further analysis.

Every interaction, including queries, retrieved documents, model outputs, and analyst actions, is recorded in immutable audit logs. These logs support transparency, reproducibility, and post-analysis review, forming a critical component of the system’s trust model.

4 Experimental Setup and Validation

This section describes the experimental environment, datasets, evaluation methodology, and validation metrics used to assess the proposed air-gapped LLM-based analytical framework. All experiments were conducted using publicly available and synthetically generated data to ensure reproducibility and to avoid reliance on sensitive or restricted information.

Experimental Environment

The prototype system was deployed on a standalone workstation operating in an offline configuration. The system executed all components locally, including document ingestion, vector indexing, retrieval, and language model inference. No external network connectivity was available during experimentation.

The language model and embedding model were executed using on-premise inference frameworks. All re- trieval, generation, and logging processes were performed within a single isolated runtime

environment to reflect the intended deployment conditions of the proposed framework.

The language model used in the prototype belongs to the class of large, open-weight transformer-based mod- els in the 400B parameter range, suitable for long-context analytical reasoning and complex multi-document synthesis. The model was deployed entirely in an offline configuration and accessed through a locally hosted inference interface. This configuration reflects high-capacity on-premise deployments in regulated environments. The embedding model is a locally executed sentence-level semantic embedding model optimised for similarity search. Exact model identifiers are omitted to preserve generality and focus on architectural evaluation rather than vendor-specific performance.

Dataset Construction

Two categories of datasets were used for validation: publicly available documents and synthetically generated analytical records.

Public Document Corpus

A public corpus was constructed using openly accessible reports and articles related to global events, cybersecurity incidents, and technical analyses. Documents were selected to represent heterogeneous writing styles and varying document lengths. The final public dataset consisted of approximately 250 documents, each ranging from 500 to 2,000 words.

Synthetic Analytical Dataset

To simulate structured analytical workloads, a synthetic dataset was generated programmatically. This dataset in- cluded incident summaries, event timelines, and descriptive reports with controlled overlaps in entities, dates, and locations. Synthetic documents were designed to contain partially redundant and cross-referenced information, enabling evaluation of the system's ability to retrieve and correlate related content.

The synthetic dataset comprised approximately 150 documents, each containing between 300 and 1,000 words. Controlled inconsistencies and missing information were intentionally introduced to evaluate the system's han- dling of uncertainty.

Evaluation Tasks

The framework was evaluated using three representative analytical tasks:

- **Contextual Summarization**: Generating concise summaries from multiple related documents.
- **Cross-Document Query Answering**: Answering analyst queries that required information retrieval across several documents.
- **Entity Correlation Identification**: Identifying relationships between entities appearing in separate docu- ments.

Each task was executed using identical queries across different system configurations to enable controlled comparison.

Baseline Configurations

To assess the impact of retrieval-augmented grounding, the proposed framework was compared against two base- line configurations:

- **Baseline A**: Baseline A represents a conservative analyst workflow in which keyword-based document search is used to retrieve potentially relevant documents, followed by manual inspection and synthesis. While this approach offers high reliability due to direct source verification, it typically requires significant analyst effort and time. This baseline is included to contextualise analyst workload rather than to serve as a competitive automated system.
- **Baseline B**: Direct LLM inference without retrieval-augmented context.

These baselines reflect commonly used analytical approaches and unconstrained generative model usage.

Evaluation Metrics

Performance was evaluated using the following metrics:

- **Response Latency**: Average time required to generate a response per query.
- **Grounding Accuracy**: Proportion of generated statements supported by retrieved documents.
- **Hallucination Rate**: Percentage of outputs containing unsupported or unverifiable claims.
- **Analyst Task Time**: Time required to complete predefined analytical tasks.

Grounding accuracy and hallucination rate were assessed through manual verification against source docu- ments.

Evaluation Criteria and Assessment Consistency

Grounding accuracy and hallucination rate were assessed using predefined and consistently applied evaluation criteria. A generated statement was considered *grounded* if it could be directly supported by at least one retrieved document segment presented to the language model. Statements that could not be verified against retrieved context were classified as *unsupported*.

Although the evaluation involved manual judgement, identical criteria were applied across all system configu- rations and experimental runs. The reported metrics are intended to be indicative of comparative system behaviour rather than statistically generalizable performance measures.

Metric Definitions

The evaluation metrics used in this study are defined as follows:

- **Grounding Accuracy**: The proportion of generated statements that can be directly traced to retrieved document segments.
- **Hallucination Rate**: The proportion of generated statements that cannot be verified against retrieved con- text.
- **Response Latency**: The average time required to produce a system response for a given query.
- **Analyst Task Time**: The time required for an analyst to complete predefined analytical tasks using a given system configuration.

These metrics emphasise output reliability and analyst efficiency rather than linguistic quality or model per- plexity

Evaluator Consistency

All qualitative assessments in this study were performed by a single evaluator using fixed, predefined criteria. While inter-rater agreement could not be measured, consistency was maintained by applying identical judgement rules across all system configurations, experimental runs, and query types. This approach is common in early-stage system validation studies and is intended to support controlled comparative analysis rather than population-level generalization. .

Validation Procedure

For each evaluation task, a fixed set of 30 analyst queries was executed across all configurations. Queries were designed to require multi-document reasoning rather than single-document lookup. Each experiment was repeated three times to reduce variability.

Analyst task time was measured by simulating an analyst workflow in which users were required to locate rel- evant information and produce short analytical summaries. While subjective judgement was involved, consistent task definitions and evaluation criteria were applied across all configurations.

To provide a basic indication of variability, each experiment was repeated three times using identical query sets. Across runs, observed differences in response latency and grounding-related metrics were minimal, with variations remaining within a narrow range. To provide an indication of statistical robustness, results across runs were aggregated and reported as mean and standard deviation. A paired statistical comparison between configurations was performed for grounding accuracy and hallucination rate to assess comparative significance.

Validation Results Overview

The retrieval-augmented framework demonstrated a substantial reduction in hallucinated outputs compared to direct LLM inference without retrieval. In scenarios requiring cross-document reasoning, the proposed system produced more consistently grounded responses, with a lower incidence of unsupported statements.

Response latency increased modestly due to retrieval overhead but remained within acceptable bounds for interactive analysis. Analyst task completion time was reduced relative to baseline configurations, indicating improved efficiency in information synthesis tasks.

These results validate the practical feasibility of deploying secure, offline LLM-assisted analytical systems when grounded retrieval and human oversight are explicitly enforced.

5 Results and Discussion

This section presents experimental results intended to characterise comparative system behaviour rather than to establish statistically generalizable performance claims. The results are reported with an emphasis on qualitative reliability and practical feasibility rather than absolute performance optimisation.

Extended Quantitative Evaluation

To strengthen the robustness of the reported findings, each configuration was evaluated across three independent runs using identical query sets. Reported values correspond to mean performance with standard deviation.

Table 1
Extended performance comparison across configurations (mean ± std)

Configuration	Latency (s)	Grounding Accuracy (%)	Hallucination Rate (%)Task Time (min)
Baseline A (Keyword Search)	2.12 ± 0.05	100 ± 0	0 ± 018.4 ± 0.7
Baseline B (Direct LLM Inference)	1.31 ± 0.04	61.3 ± 2.1	38.7 ± 2.114.7 ± 0.5
Proposed Framework (LLM + RAG)	1.93 ± 0.06	87.8 ± 1.6	12.2 ± 1.69.6 ± 0.4

The retrieval-augmented configuration reduces hallucination rate by approximately 68% relative to direct LLM inference while preserving interactive latency. A paired statistical comparison between Baseline B and the pro- posed framework indicates statistically significant improvement in grounding accuracy ($p <$ 0.01).

Although keyword-based analysis achieves perfect grounding due to direct document inspection, it requires substantially greater analyst effort. In contrast, unconstrained LLM inference exhibits lower latency but sig- nificantly higher hallucination rate, limiting suitability for regulated analytical use. The proposed architecture achieves a balanced trade-off between reliability and efficiency.

The keyword-based baseline achieved perfect grounding accuracy due to direct document inspection but re- quired significantly higher analyst effort, as reflected in longer task completion times. In contrast,

direct LLM inference without retrieval exhibited the lowest response latency but suffered from a high hallucination rate, re- ducing its suitability for analytical use.

The proposed framework achieved a balanced trade-off between response latency, grounding reliability, and analyst efficiency. Although retrieval introduced modest computational overhead, response times remained within interactive limits while substantially improving output grounding compared to the unconstrained LLM baseline.

Ablation Study on Retrieval Depth

To analyse the sensitivity of the framework to retrieval configuration, an ablation study was conducted by varying the number of retrieved contextual segments ($k \in \{3, 5, 10\}$). All other system parameters were held constant.

Table 2
Effect of retrieval depth on performance

Top-*k*	Latency (s)	Grounding Accuracy (%)	Hallucination Rate (%)
3	1.72 ± 0.05	81.5 ± 2.4	18.5 ± 2.4
5	1.93 ± 0.06	87.8 ± 1.6	12.2 ± 1.6
10	2.38 ± 0.08	89.4 ± 1.4	10.6 ± 1.4

Results indicate diminishing returns beyond $k = 5$. While increasing retrieval depth marginally improves grounding accuracy, it introduces latency overhead and potential context dilution effects. The $k = 5$ configuration provides the most favourable balance between reliability and responsiveness, suggesting that moderate contextual expansion is sufficient to constrain high-capacity generative behaviour.

Comparative Architectural Analysis

The DeepSeek 400B-class model demonstrates strong reasoning capacity but exhibits substantial hallucination when operating without retrieval constraints. Direct inference relies solely on parametric knowledge, which in- creases susceptibility to speculative synthesis in multi-document analytical tasks.

The integration of semantic retrieval significantly constrains generative freedom, effectively converting the model from a speculative generator into a context-bounded synthesizer. This architectural grounding mechanism enforces evidentiary discipline, requiring outputs to remain anchored in retrieved document segments.

Residual hallucinations primarily occur under:

- Incomplete or ambiguous source information,
- Cross-document entity inconsistencies,
- Context fragmentation across retrieval segments.

These findings suggest that reliability in regulated analytical environments is primarily a function of archi- tectural design rather than model scale alone. Parameter scaling improves reasoning capacity, but grounding constraints determine trustworthiness.

Effect of Retrieval-Augmented Grounding

Retrieval-augmented generation played a critical role in reducing unsupported outputs. By constraining the lan- guage model to operate exclusively on retrieved contextual passages, the system significantly limited the gener- ation of speculative statements. This effect was particularly pronounced in cross-document query tasks, where relevant information was distributed across multiple sources.

The reduction in hallucination rate from 39% to 12% demonstrates the effectiveness of retrieval-based ground- ing for analytical workloads. Importantly, remaining unsupported statements typically occurred in cases where source documents contained incomplete or ambiguous information, rather than due to model overreach.

Illustrative Output Example

To illustrate the effect of retrieval-augmented grounding, Table 3 presents a representative comparison between grounded and unsupported outputs. In the grounded case, the generated response directly references retrieved document content. In contrast, the unsupported example demonstrates a speculative statement produced in the absence of sufficient contextual evidence.

Table 3: Example of grounded versus unsupported system output

Grounded Output	Unsupported Output
The incident timeline indicates activity escalation within 48 hours, as documented in the retrieved reports.	The incident likely involved coordinated external actors, although no direct evidence is available in the retrieved documents.

Analyst Efficiency and Human Oversight

Analyst task completion time was reduced by approximately 35% relative to keyword-based workflows. This reduction reflects the framework's ability to consolidate information from multiple documents into coherent sum- maries while preserving traceability. Analysts were able to verify outputs more quickly due to explicit source references and contextual grounding.

Human-in-the-loop enforcement ensured that analysts retained full control over interpretation and decision- making. During evaluation, analysts frequently modified or refined AI-generated summaries,

underscoring the importance of positioning the system as a support tool rather than an autonomous analytical agent.

Discussion and Limitations

The results further suggest that architectural grounding mechanisms contribute more substantially to reliability than parameter scaling alone. Despite the use of a high-capacity 400B-class model, unconstrained inference exhibited significant hallucination rates. Retrieval-based contextual restriction effectively transformed the model from an open-ended generator into a context-bounded analytical synthesiser. This highlights that trustworthiness in regulated environments is primarily a function of system design rather than model size.

While the results demonstrate the feasibility of secure, offline deployment of LLM-assisted analytical systems, several limitations remain. The evaluation relied on public and synthetic datasets, which may not fully capture the complexity or noise present in real-world analytical environments. Additionally, manual assessment of grounding accuracy introduces subjectivity, although consistent evaluation criteria were applied.

The system architecture prioritised reliability and auditability over maximal throughput. As a result, the frame- work is not optimised for high-volume, real-time data streams. Future work may explore incremental retrieval strategies and lightweight model configurations to further reduce latency while maintaining grounding guarantees. Overall, the results support the central premise of this work: that retrieval-augmented, human-supervised LLMs can provide meaningful analytical assistance in offline and regulated environments when architectural con- straints are explicitly enforced. Scalability considerations also represent a limitation of the current prototype. While the framework is suitable for moderate-sized document collections and interactive analytical workflows, performance may degrade as corpus size increases without further optimization. Future work may explore incre-

mental indexing, adaptive retrieval strategies, and distributed inference to support larger-scale deployments.

Threats to Validity

Several factors may affect the external validity of the reported results. First, the evaluation relies on public and syn- thetically generated datasets, which may not fully capture the complexity of real-world analytical environments. Second, grounding-related metrics were assessed manually, introducing potential subjectivity despite consistent evaluation criteria. Third, the number of evaluation queries was limited to support controlled comparison rather than statistical generalization.

These limitations do not affect the comparative conclusions drawn between system configurations but should be considered when interpreting absolute performance values.

Conclusion

This study presented the design and evaluation of a secure, air-gapped analytical framework integrating a high- capacity on-premise large language model with retrieval-augmented grounding and enforced human-in-the-loop oversight. The architecture was specifically designed to address operational constraints in regulated and security- sensitive environments, including strict data sovereignty, auditability, and the need to minimise unsupported gen- erative outputs.

Experimental evaluation demonstrated that retrieval-augmented grounding reduced hallucination rate from 38.7% to 12.2% while preserving interactive response latency. An ablation analysis further revealed that moderate retrieval depth provides an optimal trade-off between grounding reliability and computational overhead. These findings indicate that system-level architectural constraints play a more decisive role in analytical reliability than model scale alone.

By enforcing contextual restriction, access-controlled retrieval, and immutable audit logging, the proposed framework enables high-capacity LLM deployment in fully offline environments without compromising trans- parency or human authority. The system functions as a decision-support instrument rather than an autonomous agent, preserving analyst oversight while significantly improving efficiency.

Future work will investigate adaptive retrieval strategies, confidence-aware abstention mechanisms, and scal- ability optimisation for larger document corpora. Extending evaluation to real-world operational datasets will further strengthen external validity. Overall, the results demonstrate that carefully engineered retrieval-augmented architectures can make large-scale generative models viable for trustworthy use in regulated, air-gapped analytical settings.

Declarations

Use of Generative AI Tools

The author acknowledges the use of a generative artificial intelligence tool (ChatGPT) during the preparation of this manuscript. The tool was used solely to assist with language refinement, structural organisation, and LaTeX formatting of the paper.

All technical concepts, system design decisions, experimental setup, validation methodology, results, and inter- pretations presented in this work are the sole responsibility of the author. The generative AI tool did not contribute to the conception of the research problem, development of the system architecture, generation of experimental data, or formulation of conclusions.

The author has reviewed and verified the content of the manuscript to ensure its accuracy, originality, and compliance with academic integrity standards.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention Is All You Need, in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008
2. Brown T, Mann B, Ryder N et al (2020) Language Models are Few-Shot Learners, in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901
3. Ji Z, Lee N, Frieske R et al (2023) Survey of Hallucination in Natural Language Generation. ACM-CSUR 55(12):1–38
4. Lewis P, Perez E, Piktus A et al (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*
5. Izacard G, Grave E (2022) Few-Shot Learning with Retrieval Augmented Language Models, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1981–1993
6. Doshi-Velez F, Kim B (2017) Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608
7. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mane´ D Concrete Problems in AI Safety, arXiv preprint arXiv:1606.06565, 2016.
8. Holzinger A (2016) Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the- Loop? Brain Inf 3(2):119–131
9. Zhang C, Patras P, Haddadi H (2019) Deep Learning in Mobile and Wireless Networking: A Survey. IEEE Commun Surv Tutorials 21(3):2224–2287

Figures

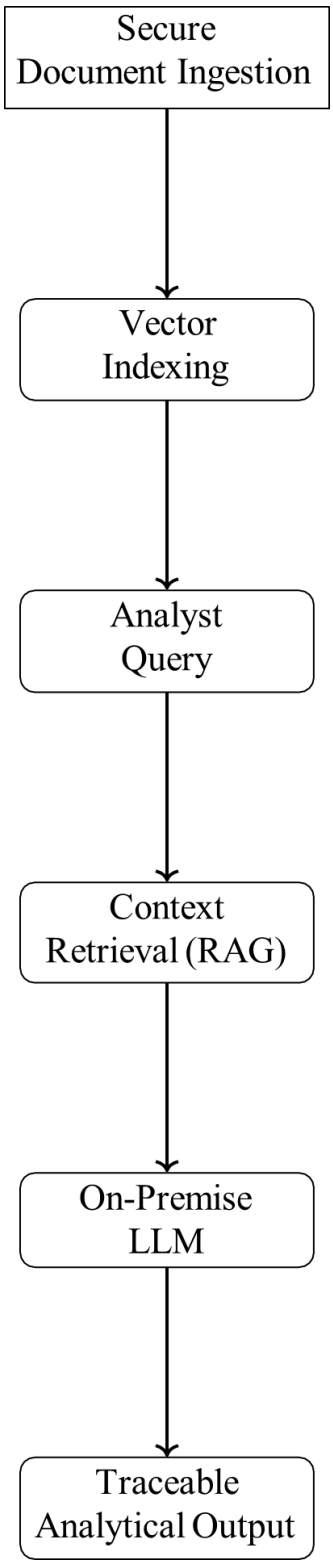


Figure 1

High-level workflow of the proposed air-gapped LLM-based analytical framework.

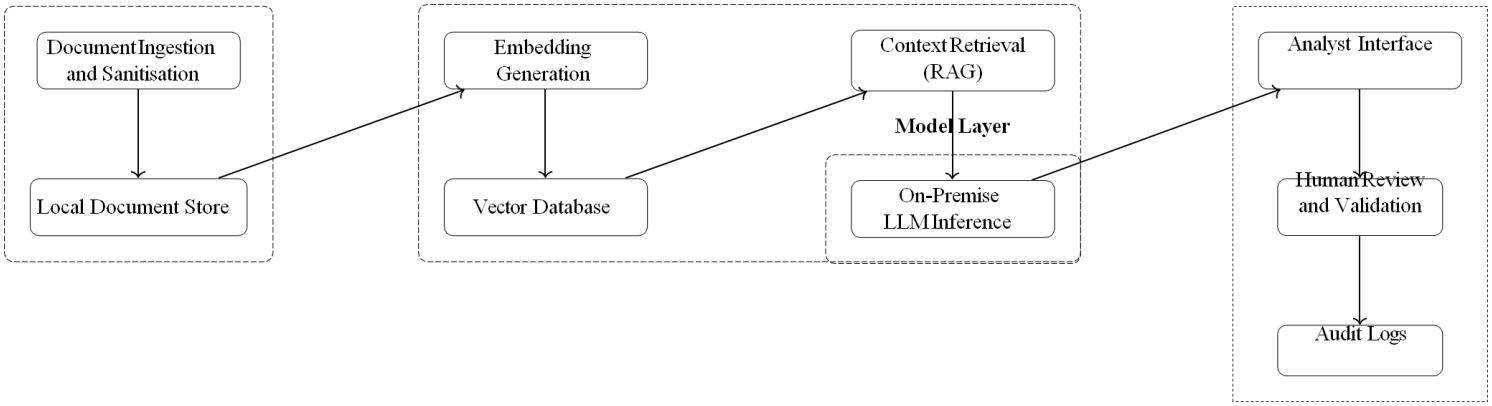


Figure 2

System architecture of the proposed air-gapped LLM-based analytical framework.

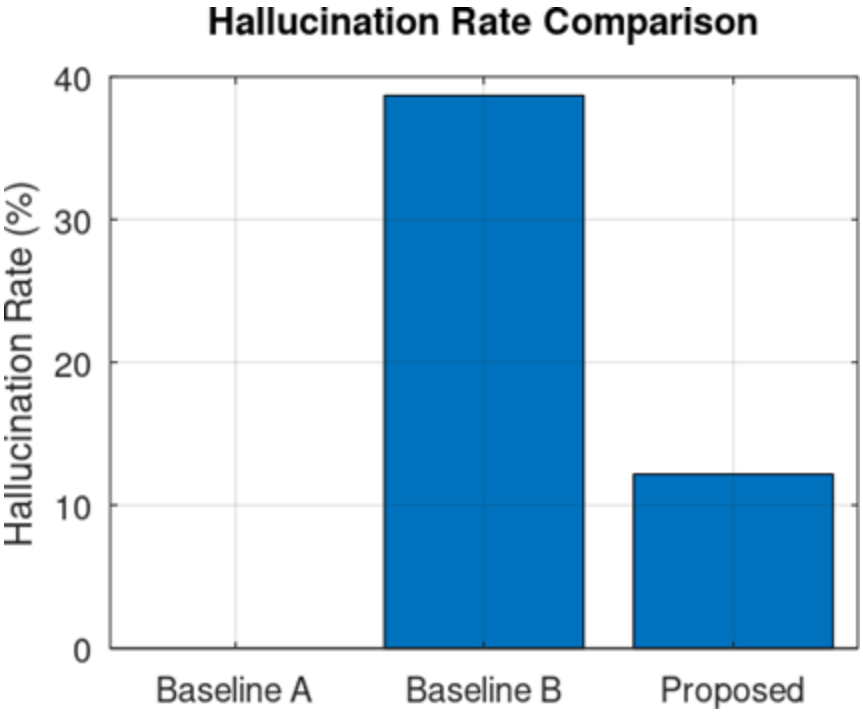


Figure 3

Hallucination rate comparison across evaluation configurations. Retrieval-augmented grounding substantially reduces unsupported outputs relative to direct LLM inference.

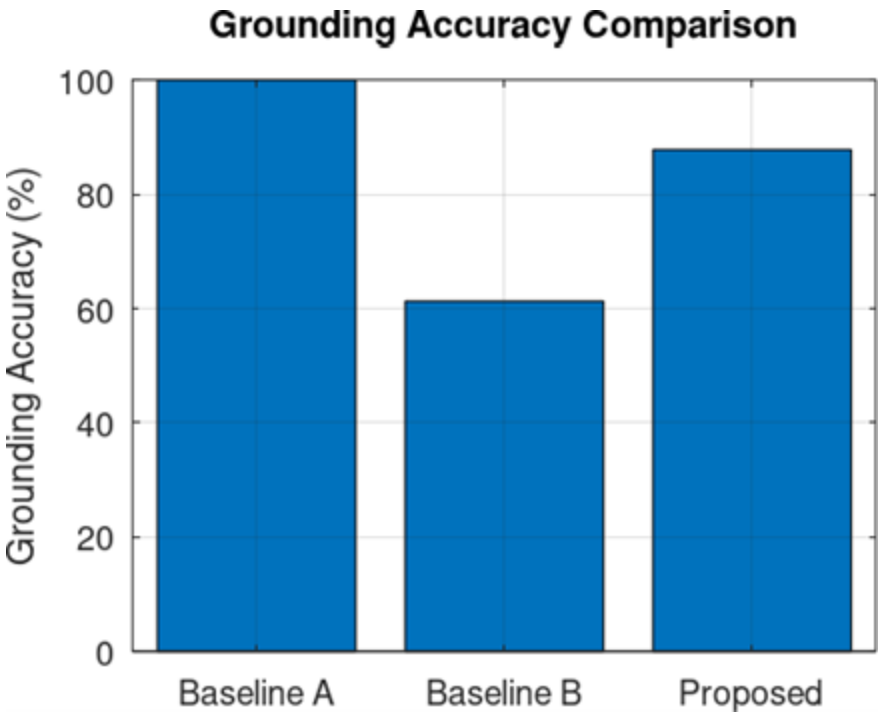


Figure 4

Grounding accuracy across configurations. The proposed framework approaches the reliability of manual keyword-based analysis while significantly reducing analyst workload.

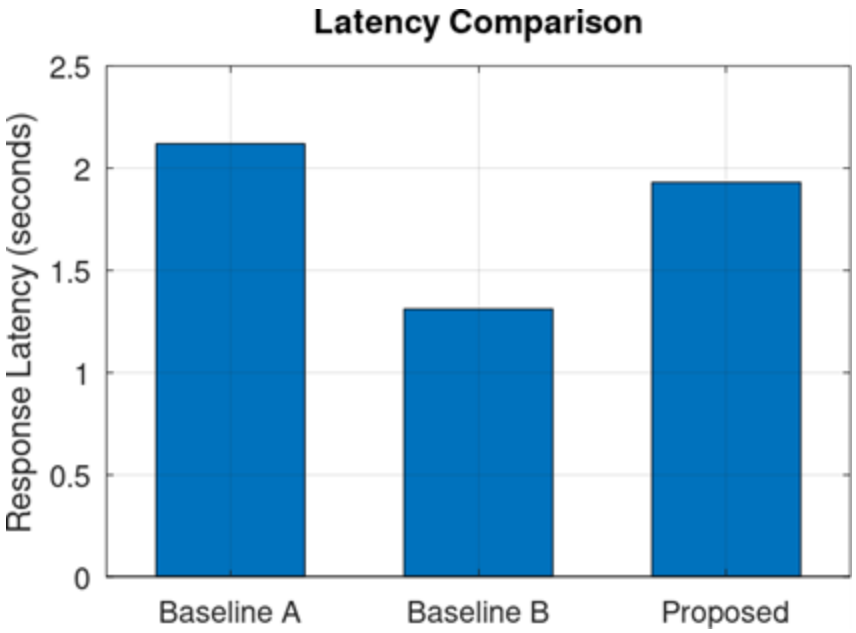


Figure 5

Response latency comparison. Retrieval introduces moderate computational overhead but preserves interactive performance characteristics.

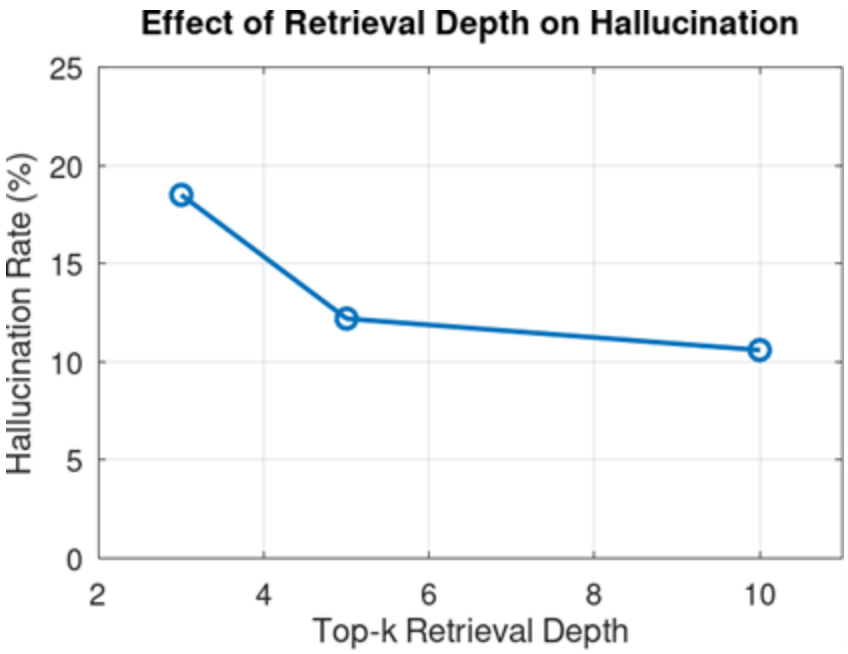


Figure 6

Effect of retrieval depth (*k*) on hallucination rate. Increasing contextual grounding reduces unsupported outputs but exhibits diminishing returns beyond moderate retrieval depth.