VAULT: Verified Access Control for LLM-Based Knowledge Graph Querying

Maximilian Stäbler
Institute for AI Safety & Security
German Aerospace Center (DLR)
Ulm, Germany
maximilian.staebler@dlr.de

Tobias Müller
Industry-University Collaboration
SAP SE
Walldorf, Germany
tobias.mueller15@sap.com

Frank Köster
Institute for AI Safety & Security
German Aerospace Center (DLR)
Ulm, Germany
frank.koester@dlr.de

Chris Langdon
Drucker School of Business
Claremont Graduate University
Claremont (CA), USA
chris.langdon@cgu.edu

***Abstract*—The exponential growth of unstructured textual data in enterprise environments has made automated knowledge graph generation essential for efficient information management. Although recent advances in natural language processing have enabled automated knowledge extraction, organizations face two critical challenges: maintaining domain specificity in knowledge representation and ensuring secure, role-based access to sensitive information. VAULT (Verified Access Control for Large Language Model (LLM)-Based Knowledge Graph Querying) presents a novel framework that combines ontology-driven knowledge extraction with dynamic access control mechanisms. The framework introduces three key innovations: (1) a configurable domain-driven node structure that enforces domain-specific knowledge organization through semantic validation, (2) a multitiered access control mechanism that implements both document-level restrictions and node-level visibility patterns, and (3) an LLM-powered inference engine that dynamically filters knowledge graph traversal based on user authorization levels. We implement our approach using a prototype system that demonstrates the automated conversion of natural language text into structured knowledge graphs while maintaining security constraints. Our experimental evaluation encompasses comprehensive testing across 16 different open-source LLMs, analyzing their performance under varying access control conditions and authorization levels. The results demonstrate the framework's effectiveness in maintaining information security while preserving query response quality across different access tiers. The adaptability of the framework makes it particularly valuable for industries handling sensitive information, such as healthcare, finance, and intellectual property management, where both domain specificity and information security are paramount. This paper contributes to the field by bridging the gap between generic knowledge graph generation and domain-specific requirements while providing empirical evidence for the effectiveness of multilevel access control in LLM-based knowledge systems.**

***Keywords*-Knowledge Graphs; Verified Roled-Based Access; LLM; Semantic Interoperability.**

I. Introduction

The exponential growth of unstructured textual data in modern enterprises has created unprecedented challenges in the management of information and the extraction of knowledge [1][2][3]. Organizations face the complex task of transforming large amounts of unstructured documents into actionable structured knowledge while maintaining strict security protocols and access controls [4][5]. This challenge is particularly acute as enterprises increasingly rely on automated systems to process and analyze their data repositories [1] or automate their business processes [6]. In modern enterprises, Enterprise Resource Planning (ERP) systems usually provide an integrated and continuously updated view of the core business processes, while Enterprise Knowledge Management (EKM) systems refer to the systematic handling of an organization's information assets, ensuring that employees can efficiently access, share, and utilize knowledge. With the rise of Natural Language Processing (NLP), companies are integrating AI-driven tools into their ERP and EKM systems to improve knowledge retrieval, automate document processing, support decision-making, and process automation [6][7][8][9]. However, maintaining domain and business process specificity, as well as implementing secure, role-based access, are critical challenges, which we explore in more detail in the following.

A. Maintaining Domain Specificity

Comprising domain-specific information is especially challenging for general-purpose NLP models that are trained on broad, openly available datasets, which may not adequately capture the nuances, such as distinct terms or abbreviations of specialized domains [10]. Without proper customization to local domain-specific data and business process knowledge, these models risk generating inaccurate, misleading, or overly generic knowledge representations that do not align with domain-specific terminology, ontologies, or reasoning frameworks [11]. Given an enterprise context, context awareness of NLP-based systems is especially relevant to ensure accurate interpretation of ERP-specific business processes, data, and terminology, as generic models may introduce inaccuracies that could disrupt operations, decision-making, or even lead to harmful consequences for the business.

Recent works explored various ways to ensure domain specificity, such as fine-tuning local data, prompt engineering, and few-shot learning, Knowledge Graph (KG) integration, or building Retrieval-Augmented-Generation (RAG) pipelines

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-272-2

[12]. Each technique has its own specific challenges. Although fine-tuning requires substantial computing resources and can be costly [13], simple prompt engineering with few-shot learning may not generalize well and requires careful prompt design and testing, which need to be revisited when new documents are introduced to the data repositories [14]. Curating KGs requires domain experts and may be complex to maintain and update dynamically [15]. RAG retrieves unstructured text that may contain conflicting or imprecise domain knowledge and lacks the reasoning ability to connect concepts.

The choice of the customization approach depends on the underlying use case, needs, and domain [12]. In our work, we are focusing on *EKM*, in which employees require accurate domain-specific responses from corporate knowledge bases. Hence, it is crucial to reduce misinformation and systems need to accommodate fast-changing and growing data repositories. In the context of EKM, we leverage a combination of structured KGs and RAG with the rationale of combining the precision of structured KG-based retrieval with the low-cost, real-time adaptability of RAG pipelines. To ensure appropriate knowledge representation across various domains, we present a novel configurable ontology-driven node structure.

B. Secure, Role-Based Access Control

While recent advances in Large Language Models (LLMs) have revolutionized knowledge extraction capabilities, they have simultaneously introduced critical security concerns regarding information access and distribution [2][16]. Traditional RAG systems, while efficient at knowledge extraction, rarely address the crucial aspects of user permissions and access restrictions, creating significant security risks and compliance challenges [1][4]. This limitation becomes particularly problematic in the context of EKM. Given the example in which a company uses an internal NLP-based search engine for corporate documents, it is essential to prevent unauthorized access and unintentional return of restricted sensitive information. Restricting information access policies should be dynamically changeable based on varying roles, since, for example, an executive should have access to strategic reports for their responsible domain, while employees usually have a more restricted view due to compliance or other company policies. For example, if an employee prompts the internal NLP-based system to *"Show the latest NDA template"*, the system should retrieve only the *template* without showing any information regarding any related confidential legal disputes. Hence, to ensure each employee's access to knowledge relevant to their role without unnecessary noise, the underlying NLP systems should integrate predefined enterprise identity and access management policies to enforce appropriate access control. To solve this challenge, our aim is to use explicit KG rules to store the relationships between users, roles, and access permissions. More specifically, we propose a novel multitiered access control mechanism with document-level restrictions and node-level visibility patterns that allows dynamic filtering of KG traversals based on user authorization levels.

C. Research Contribution

Existing solutions can be categorized into two distinct types: those that focus on the extraction of generic knowledge without considering security implications, and those that implement rigid access control mechanisms that lack the flexibility required for domain-specific knowledge management. The absence of a unified framework that combines robust security measures with sophisticated knowledge extraction capabilities represents a significant gap in current EKM systems. To address these challenges, we present VAULT (Verified Access Control for LLM-Based KG Querying), a novel framework that integrates three key innovations.

- A configurable domain-driven node structure that enforces domain-specific knowledge organization through semantic validation, ensuring consistent and contextually appropriate knowledge representation across various enterprise domains.
- A sophisticated multi-tiered access control mechanism that implements both document-level restrictions and node-level visibility patterns, providing granular control over information access while maintaining system flexibility.
- An innovative inference engine powered by open-source LLMs that dynamically filters KG traversal based on user authorization levels, demonstrating the framework's effectiveness across eleven different open-source language models.

This research addresses the critical gap between automated knowledge extraction and security requirements by providing a comprehensive solution that maintains both domain specificity and information security. The framework particularly addresses the challenges of managing permissions across multiple integrated data sources while ensuring zero margin of error in access control implementation. Empirical validation across multiple open-source LLMs demonstrates the framework's robustness and adaptability, establishing a foundation for secure, domain-aware knowledge management systems in enterprise environments.

The remainder of the paper is structured as follows. In Section II, we present the related work, reviewing approaches in knowledge graph generation, role-based access control, integration of large language models in knowledge management, ontology-driven knowledge extraction, and existing limitations. Section III describes the system architecture and implementation of the VAULT framework, covering the knowledge extraction layer, the access control layer, and the query processing layer. Section IV provides the results of our experimental evaluation, detailing the setup and methodologies used, including human expert evaluation and automated metrics. Finally, we conclude our work and discuss future research directions in Section V.

II. Related Work

Recent advances in knowledge management systems have highlighted the importance of integrating structured knowledge with flexible access control mechanisms. This section examines key approaches across several critical areas relevant to secure KG generation and management.

A. KG Generation from Unstructured Text

The transformation of unstructured textual data into KGs has become increasingly vital for enterprise information management [2][3][17]. Current approaches typically employ a three-stage process: entity extraction, relationship identification, and graph construction. While traditional methods like OpenNRE [18] achieve only 61.4% accuracy, modern techniques like REBEL [19] have demonstrated success rates of up to 87%. A significant challenge remains in verifying whether the extracted information actually exists in the source documents, leading to the development of hybrid approaches that combine traditional extraction methods with LLM capabilities.

B. Role-based Access Control in KGs

Role-Based Access Control (RBAC) has emerged as a critical component in KG systems, particularly in enterprise environments [4]. Modern implementations simplify security management by grouping users into roles based on their tasks rather than assigning individual permissions. Recent research has expanded this concept to include multilevel access control mechanisms that implement document-level restrictions and node-level visibility patterns [20][21]. A notable advancement is the development of graph-based access control patterns that enable both open and closed security policies.

C. LLM Integration in Knowledge Management Systems

The integration of LLMs into knowledge management systems represents a transformative development in organizational knowledge management [22][23]. Current approaches focus on automating content creation, improving knowledge retrieval, and improving system efficiency. However, implementation presents significant challenges, particularly regarding customization requirements and system integration. Although LLM integration has shown potential to improve knowledge discovery and automated summarization capabilities, concerns persist about the reliability and accountability of LLM-generated content.

D. Ontology-driven Knowledge Extraction

Ontology-driven knowledge extraction has been identified as a crucial method to maintain domain specificity in knowledge representation [24]. Current systems employ ontologies as formal knowledge sources that can unambiguously represent task specifications and domain knowledge. This approach has been particularly effective in specialized domains where maintaining semantic accuracy is paramount.

E. Limitations in Existing Solutions

Several key limitations persist in current approaches:

1) Verification Challenges: Existing systems face difficulties in verifying the accuracy of LLM-extracted information, particularly in maintaining the clear provenance of extracted knowledge [25].
2) Access Control Granularity: Although RBAC systems provide fundamental security mechanisms, they often lack the flexibility required for complex organizational hierarchies and dynamic access requirements [4].
3) Integration Complexity: The integration of LLMs with existing knowledge management systems often requires extensive customization, which can disrupt established workflows [22].

Domain adaptation is another key challenge. Current ontology-driven approaches often require significant manual effort to adapt to new domains, limiting their scalability across different business contexts. These limitations underscore the need for a more integrated approach that combines the strengths of LLM-based extraction, robust access control mechanisms, and domain-specific ontological validation.

III. System Architecture and Implementation

VAULT employs a three-layer architecture designed to ensure secure and efficient knowledge extraction and management. Each layer serves a specific purpose in the pipeline, from raw text processing to secure knowledge delivery. An overview of the architecture is shown in Figure 1.

A. Knowledge Extraction Layer

The knowledge extraction layer implements a sophisticated pipeline to transform unstructured text into structured KG. The resulting KG is shown in Figure 3. This process occurs in several distinct stages:

- *Document Processing:* Source documents are initially segmented into manageable text chunks, with an optimal chunk size of 600 tokens to maximize the extraction efficiency of the entities.
- *Entity and Relationship Extraction:* The system performs entity and relationship extraction using LLM-based processing (either ChatGPT or local Ollama models) through multiple "gleaning" rounds for comprehensive coverage. Users can define domain-specific entities for mapping, ensuring relevance to their application area. The extraction uses a multipart prompt to identify entities (with name, type, and description) and their relationships, which can be customized through few-shot examples for specialized domains. The summarisation of community detection results is facilitated by LLM-based abstractive summarisation, thereby enabling both hierarchical data exploration and focused querying.
- *Community Detection:* In contrast with related work that exploits the structured retrieval and traversal affordances of graph indexes, the focus here is on a previously unexplored quality of graphs in this context: their inherent modularity [26] and the ability of community detection algorithms to partition graphs into modular communities of closely-related nodes (e.g., Leiden [27]). LLM-generated summaries of these community descriptions provide comprehensive coverage of the underlying graph

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-272-2

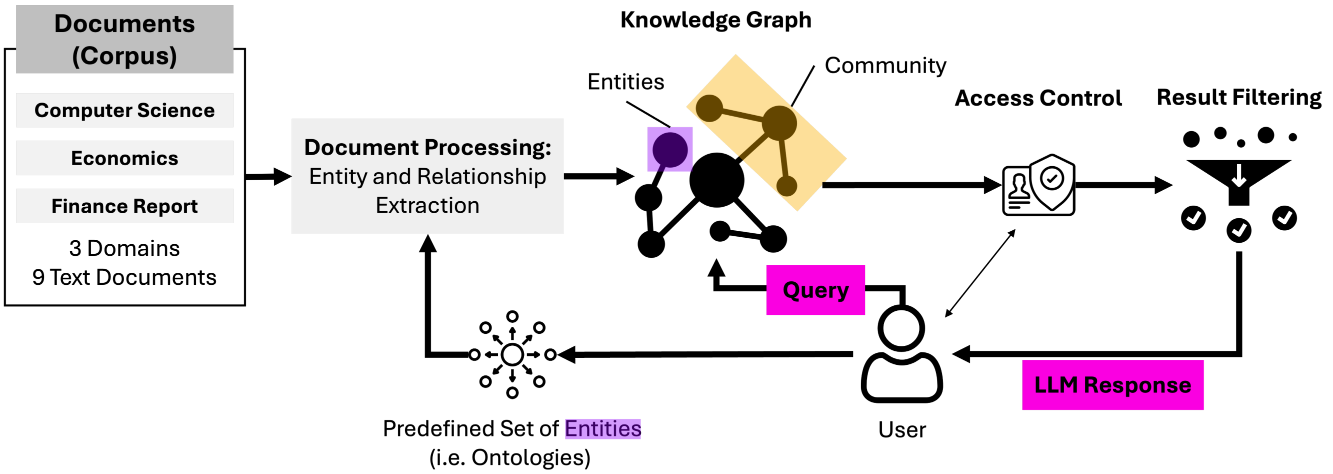


Figure 1. VAULT architecture overview: process from reading the input data, to building the KG, to generating a personalised access-controlled response to a user query. The user can specify a selection of entities to be used to build the KG.

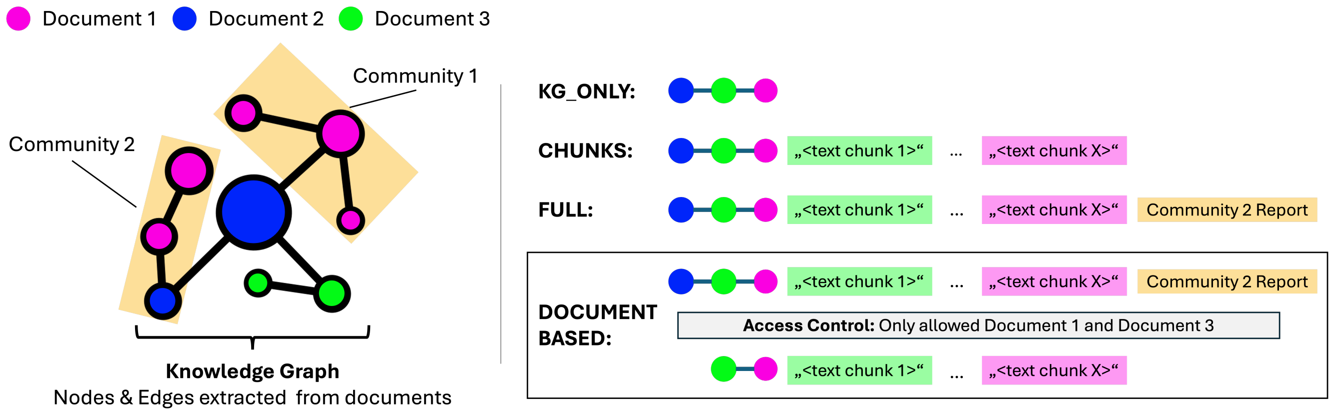


Figure 2. Overview of access levels: KG structure with predefined entities and Leiden-extracted communities. Responses are filtered based on user authorization, especially in the *DOCUMENT_BASED* mode.

index and the input documents it represents. Query-focused summarisation of an entire corpus is then facilitated through a map-reduce approach. This approach involves the use of each community summary to answer the query independently.

B. Access Control Layer

The Access Control Layer implements a sophisticated four-tier security system that provides granular control over knowledge access and retrieval. This hierarchical approach ensures precise information delivery while maintaining security boundaries across different user-authorization levels. An overview is given in Figure 2.

The system implements four distinct access levels:

1) *KG_ONLY*: Provides access exclusively to authorized nodes and edges within the KG that match the query parameters. This most restrictive level ensures visibility of the basic knowledge structure while maintaining strict information control.
2) *CHUNKS*: Extends the KG_ONLY access by including referenced text chunks from the original documents, enabling users to verify KG assertions through source material while maintaining security constraints.
3) *FULL*: Augments the CHUNKS level by incorporating community summaries derived from the KG structure. These summaries provide a contextual understanding of node clusters while preserving access control boundaries.
4) *DOCUMENT_BASED*: Implements a distinct approach where document access is determined by node-level permissions. The system first performs a FULL-level search, but then filters results based on user authorization for specific nodes associated with the extracted text.

This multitiered approach operates independently of the query processing layer, ensuring consistent security enforcement regardless of the underlying LLM implementation. The system validates access permissions before any content reaches the query processing stage, effectively creating a security boundary that prevents unauthorized information disclosure.

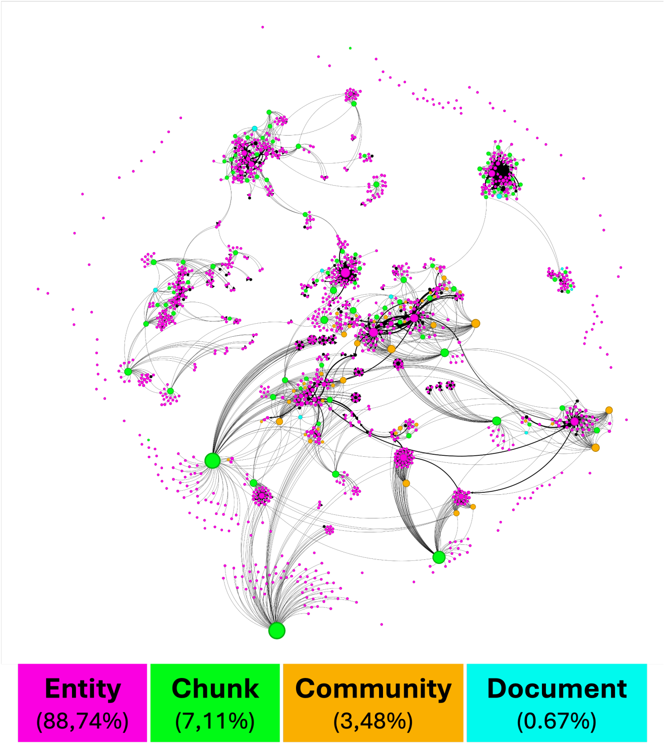


Figure 3. Visualisation of the KG extracted from the nine input documents and the predetermined entities.

The DOCUMENT_BASED level represents a particularly innovative approach, as it combines traditional document-level access control with node-based permissions, enabling fine-grained control over information flow while maintaining document context. This method ensures that users can access only the information from documents where they have appropriate authorization for the referenced KG nodes.

Each access level builds upon the previous one, creating a hierarchical security model that can accommodate various organizational security requirements while maintaining system flexibility. This layered approach enables organizations to implement precise access control policies while maximizing the utility of knowledge within authorized boundaries.

C. Query Processing Layer

The query processing layer utilizes a flexible LLM integration architecture that supports multiple open-source models through *Ollama*[28]. The implementation includes support for various models ranging from lightweight (1.5b parameters) to large-scale (32b parameters) architectures, including:

- *DeepSeek* models (1.5b and 32b variants)
- *Llama3.2* (1b and 3b)
- *Mistral-small* (24b)
- *Phi* variants (3.5b and 4b)
- *Qwen2.5* variants (0.5b to 7b)
- *SmolLM* series (135m to 1.7b)

The query processing implements a sophisticated retrieval and generation pipeline that leverages both the hierarchical community structure and the underlying KG. The system first identifies relevant entities through a semantic search, which serve as entry points for graph traversal. From these entry points, the system explores connected text chunks, community reports, and entity relationships, with all retrieved data being filtered according to the user's access level. The system employs a map-reduce approach to handle broad thematic queries. Retrieval of relevant community node reports from specified hierarchical levels, which are then shuffled and chunked. Each segment generates points with associated importance scores that are subsequently ranked and filtered to maintain the most significant information. This filtered intermediate response serves as a context for the final LLM-generated answer. This approach combines structured KG data with unstructured document content, enabling comprehensive responses that incorporate both specific entity information and broader thematic understanding. The community-based retrieval strategy has been shown to be particularly effective in addressing queries about broad themes and ideas, thus overcoming the limitations of traditional RAG methodologies in handling corpus-wide analysis.

The modular design of the system facilitates the seamless integration of new models while ensuring the consistent application of security controls across all configurations. Query processing is only initiated after access control validation, ensuring that responses are generated using only authorized information. This architecture enables VAULT to maintain strict security boundaries while leveraging the capabilities of modern LLMs for knowledge extraction and query processing. The implementation demonstrates both scalability and flexibility, accommodating various organizational security requirements while maintaining efficient knowledge management capabilities. The Knowledge Extraction Layer and the Query Processing Layer have been inspired by the Microsoft GraphRAG approach [29]. The complete implementation is available here [30].

IV. Results

This section presents the comprehensive evaluation of VAULT's performance across different access control configurations and LLM models.

A. Experimental Setup

We conducted an extensive evaluation using a diverse dataset comprising computer science papers and financial documents. The experiment included 20 carefully crafted questions derived from two distinct documents: a computer science research paper and Apple's SEC 8K report for 2024. The evaluation framework encompassed 16 open source language models deployed through Ollama, tested across four access control configurations with two different user roles, resulting in 2,240 unique question-response pairs.

B. Evaluation Methodology

The evaluation process consisted of two complementary approaches:

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-272-2

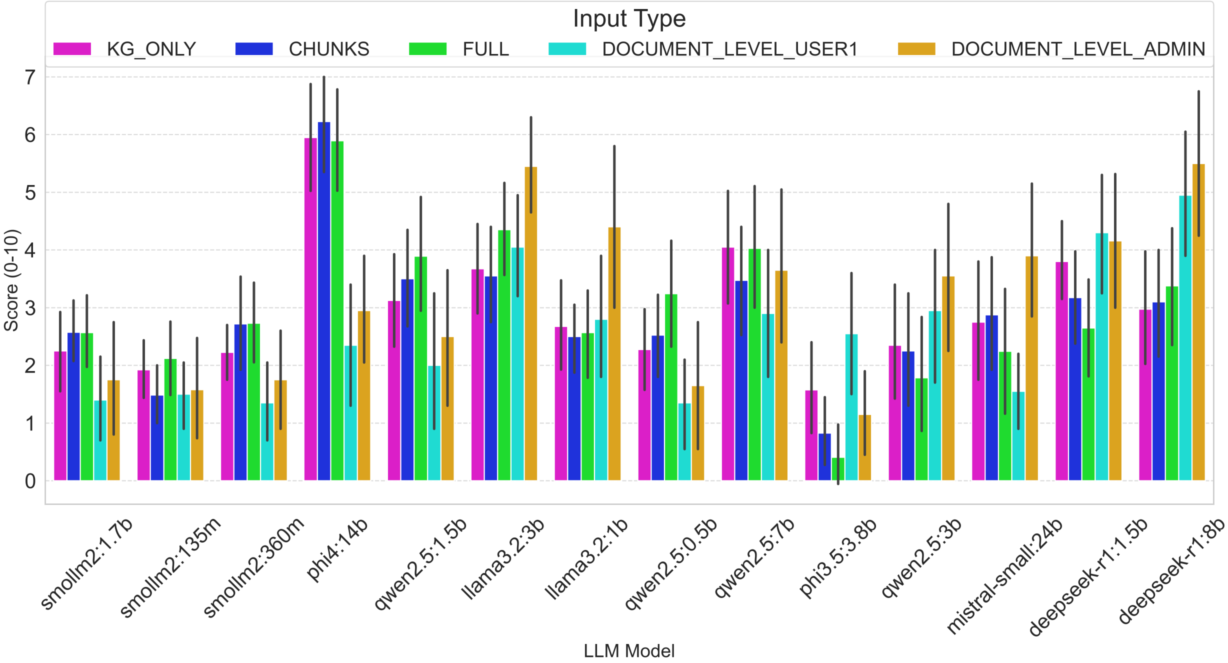


Figure 4. Visualization of the performance comparison across 16 different open-source LLM models with varying parameter sizes (135m to 8b) under different access control configurations. The y-axis represents the manual quality score (0-10), while the x-axis lists the different LLM models.

Human Expert Evaluation Eight researchers conducted systematic evaluations of responses, the evaluation corpus being equally divided between them. Each expert:

- Assigned a quality score (1-10 scale - higher is better)
- Provided qualitative justification for their scoring
- Verified response correctness within access control constraints

Automated Metric Analysis We used the *OPIK* framework by *CometML*[31] to calculate six key metrics:

- *LevenshteinRatio*: Quantifies response validity through string similarity comparison against reference text, identifying structural and content deviations
- *Answer Relevance*: Measures response alignment with query intent and appropriateness, independent of factual accuracy
- *Context Precision*: Evaluates the accuracy of context usage in responses, identifying information misalignment with the provided context
- *Context Recall*: Assesses completeness of context utilization, measuring inclusion of critical information from available context
- *Usefulness*: Scores practical value (0.0-1.0) based on completeness, clarity, and applicability of responses
- *Hallucination*: Identifies and quantifies information generation that is not supported by input context or access permissions.

Automated metrics provided objective measurements, while human evaluation offered nuanced qualitative insights into response effectiveness. The following sections present detailed analyses of the results in these evaluation dimensions, examining the effectiveness of different levels of access control and the performance variations among the LLM models tested. The experimental results in Figure 4 demonstrate varying performance across different LLM models and access control configurations. The *phi4* model emerges as the top performer, achieving scores between 6 and 7 across all access levels, significantly outperforming other models in the evaluation. This performance suggests that model size does not necessarily correlate directly with effectiveness in access-controlled knowledge retrieval tasks. Across the access control spectrum, *KG_ONLY*, *CHUNKS*, and *FULL* access levels exhibit relatively consistent performance patterns within each model, although with notable variations in absolute scores. The *DOCUMENT_LEVEL* access shows a clear differentiation between *USER1* and *ADMIN* permissions, with *ADMIN* consistently achieving higher scores. This pattern validates the effectiveness of the access control mechanisms implemented. The larger models, including *deepseek-rp (1.8b)* and *mistral-small (24b)*, demonstrate more stable performance at different access levels compared to their smaller counterparts. However, smaller models like *smollm2* variants show more pronounced variations in performance in different access configurations. The error bars indicate considerable variance in performance, particularly in models with larger parameter counts, suggesting that the size of the model may influence response consistency. The results also reveal that the *DOCUMENT_LEVEL_ADMIN* configuration generally achieves higher scores compared to *USER1* access, particularly evident in models like *llama3.2*

TABLE I. PERFORMANCE METRICS BY ACCESS TYPE

Access Type	RT	Score	Hall.	Rel.	Use.	Prec.	Rec.
DOCUMENT_LEVEL_ADMIN	**1.90**	**3.14**	**0.79**	**0.50**	0.47	**0.22**	**0.24**
FULL	3.24	3.00	0.91	0.46	0.48	0.17	0.18
KG_ONLY	3.18	2.97	0.90	0.48	**0.50**	0.19	0.20
CHUNKS	3.16	2.92	0.91	0.46	0.48	0.17	0.18
DOCUMENT_LEVEL_USER1	1.92	2.57	0.89	0.40	0.42	0.14	0.16

(3b) and *deepseek-r1*, indicating successful implementation of hierarchical access control mechanisms while maintaining response quality. A detailed analysis of performance metrics across access types, as shown in Table I, provides additional information on the effectiveness of the system. *DOCUMENT_LEVEL_ADMIN* configuration achieves the highest overall performance with a score of 3.14 and the lowest hallucination rate (0.79), indicating more reliable information retrieval. This configuration also demonstrates better precision (0.22) and recall (0.24) compared to other access levels, suggesting more accurate and comprehensive information extraction. Notably, while *KG_ONLY* access shows slightly lower overall scores (2.97), it achieves the highest usefulness metric (0.50), indicating that despite restricted access, responses remain practically valuable. *FULL* and *CHUNKS* access levels show similar performance patterns across the metrics, with scores of 3.00 and 2.92, respectively, suggesting that additional context beyond basic fragments may not significantly improve response quality. *DOCUMENT_LEVEL_USER1* consistently shows lower performance across all metrics, with the lowest overall score (2.57) and reduced relevance (0.40), confirming the effectiveness of access control mechanisms in restricting unauthorized information access. The response time (RT) metrics indicate that the $DOCUMENT\_LEVEL$ configurations (both *ADMIN* and *USER1*) process queries significantly faster (1.90 and 1.92 seconds, respectively) compared to other types of access, suggesting more efficient information retrieval when operating at the document level. These findings demonstrate that, while stricter access controls may limit overall information availability, they can lead to more precise and efficient information retrieval when properly implemented. The results also validate the system's ability to maintain security boundaries while preserving response quality within authorized access levels. Figure 5 presents a detailed analysis of the performance of the model in two key dimensions. The upper plot reveals a positive correlation between answer relevance and usefulness metrics, with most models clustering in the $0.4-0.7$ range for relevance and $0.3-0.6$ for usefulness. Notably, larger models like *qwen2.5:14b* and *mistral-small:24b* achieve higher scores on both metrics, while smaller models such as *deepseek-r1:1.5b* show lower performance. The lower plot examines the precision-recall relationship, where a distinct cluster of better performing models emerges in the upper right quadrant (precision: 0.25 – 0.30, recall: 0.22 – 0.32). This cluster, highlighted in the plot, predominantly consists of larger parameter models, suggesting that increased model size contributes to both higher precision and greater recall in knowledge retrieval tasks. Response times, indicated by dot sizes, remain relatively consistent across models, with no significant performance penalties for larger architectures. The visualization effectively demonstrates that while model size correlates with improved performance metrics, even smaller models can achieve competitive results, particularly in the midrange of the performance spectrum.

The effectiveness of VAULT's access control mechanisms is particularly evident when examining specific query responses - shown in Figure 6. Consider the question *"Who is Apple's new Chief Financial Officer?"* posed to the *mistral-small:24b* model under different access levels. When queried with *USER1* permissions, the model correctly responded with *"I don't have the information about who Apple's new Chief Financial Officer is,"* demonstrating appropriate handling of access restrictions, as the Apple SEC report was restricted to admin access only. In contrast, under *ADMIN* privileges, the same model provided a comprehensive response detailing Kevan Parekh's appointment as CFO, including contextual information about the transition and its implications for corporate governance. This stark contrast in response quality and content accuracy directly validates the effectiveness of access control implementation. Human evaluators noted this distinction, observing that *USER1* responses appropriately acknowledged information limitations, while *ADMIN* responses provided accurate and detailed information about Kevan Parekh's appointment. The *ADMIN* response not only identified the new CFO, but also provided valuable context about the leadership transition and its implications for Apple's financial management structure. This example effectively demonstrates VAULT's ability to:

- Maintain strict access control boundaries
- Prevent unauthorized information disclosure
- Provide comprehensive responses when appropriate access is granted
- Generate contextually appropriate responses based on access level

The significant difference in response quality and content between the *USER1* and *ADMIN* access levels validates the effectiveness of the framework in implementing secure, role-based access control while maintaining response quality within authorized boundaries.

V. Conclusion and Future Work

VAULT demonstrates effective integration of secure access control mechanisms with LLM-based knowledge graph generation and querying. The framework successfully addresses two critical challenges in enterprise knowledge management: maintaining domain specificity and implementing flexible access control. Through a comprehensive evaluation across

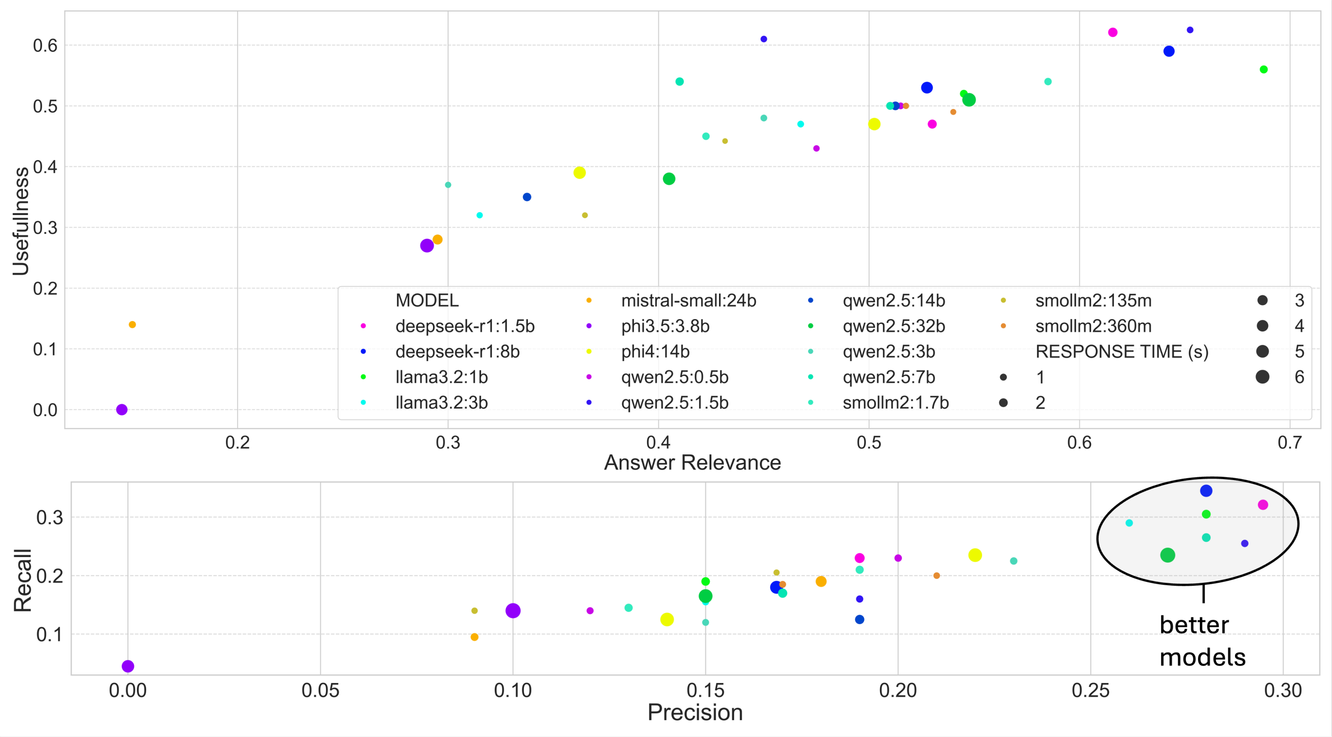
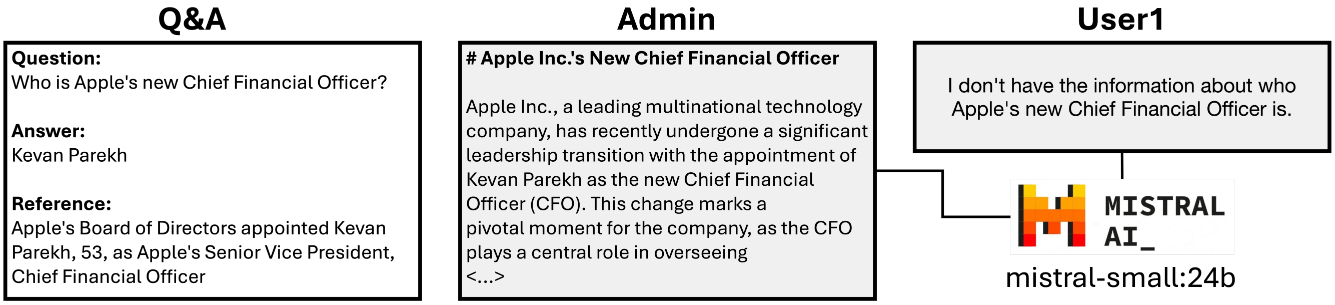


Figure 5. Open-Source-LLM Model comparison.



Question:
Who is Apple's new Chief Financial Officer?

Answer:
Kevan Parekh

Reference:
Apple's Board of Directors appointed Kevan Parekh, 53, as Apple's Senior Vice President, Chief Financial Officer

# Apple Inc.'s New Chief Financial Officer

Apple Inc., a leading multinational technology company, has recently undergone a significant leadership transition with the appointment of Kevan Parekh as the new Chief Financial Officer (CFO). This change marks a pivotal moment for the company, as the CFO plays a central role in overseeing
<...>

I don't have the information about who Apple's new Chief Financial Officer is.

Figure 6. Comparison of VAULT responses across different access levels using the *mistral-small:24b model*. The figure shows the ground truth (left), admin-level response with admin access (center), and user1-level response with restricted access (right), demonstrating effective access control implementation.

16 different open-source LLMs and multiple access control configurations, we have demonstrated the system's ability to maintain information security while preserving query response quality. Key contributions of this work include the following.

- A configurable ontology-driven architecture that enables domain-specific knowledge organization
- A multi-tiered access control system that provides granular information access management
- An LLM-powered inference engine that effectively filters knowledge graph traversal based on authorization levels

The results show that the *DOCUMENT_LEVEL_ADMIN* setup performs best, with the highest score (3.14) and lowest hallucination rate (0.79), effectively balancing response quality and strict access control.

A. Future Work

Several promising directions for future research emerge from this work:

- *Dynamic Access Control*: Developing mechanisms for real-time adaptation of access control policies based on user behavior and organizational changes.
- *Cross-Domain Integration*: Extending the framework to handle multiple domain ontologies simultaneously, enabling more flexible knowledge integration across different business units.
- *Performance Optimization*: Investigating techniques to reduce response times

These future directions aim to enhance VAULT's practical applicability while maintaining its core strengths in secure, domain-specific knowledge management.

References

[1] W. Fan *et al.*, "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Barcelona Spain: ACM, Aug. 2024, pp. 6491–6501, isbn: 9798400704901. doi: 10.1145/3637528.3671470.

[2] B. Peng *et al.*, "Graph Retrieval-Augmented Generation: A Survey," Sep. 2024. doi: 10.48550/arXiv.2408.08921. arXiv: 2408.08921 [cs].

[3] K. Pichai, "A Retrieval-Augmented Generation Based Large Language Model Benchmarked On a Novel Dataset," *Journal of Student Research*, vol. 12, no. 4, Nov. 2023, issn: 2167-1907. doi: 10.47611/jsrhs.v12i4.6213.

[4] K. Jagannath, "Enhancing Retrieval-Augmented Generation with Permissions Awareness," *Defensive Publications Series*, May 2024.

[5] K. Sawarkar, A. Mangal, and S. R. Solanki, *Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers*, Aug. 2024. doi: 10.48550/arXiv.2404.07220. arXiv: 2404.07220 [cs].

[6] J. Schnepf, T. Engin, S. Anderer, and B. Scheuermann, "Studies on the Use of Large Language Models for the Automation of Business Processes in Enterprise Resource Planning Systems," in *Natural Language Processing and Information Systems*, A. Rapp, L. Di Caro, F. Meziane, and V. Sugumaran, Eds., Cham: Springer Nature Switzerland, 2024, pp. 16–31, isbn: 978-3-031-70239-6. doi: 10.1007/978-3-031-70239-6_2.

[7] P. M. Mah, I. Skalna, and J. Muzam, "Natural Language Processing and Artificial Intelligence for Enterprise Management in the Era of Industry 4.0," *Applied Sciences*, vol. 12, no. 18, p. 9207, Jan. 2022, issn: 2076-3417. doi: 10.3390/app12189207.

[8] M. V. Godbole, "Revolutionizing Enterprise Resource Planning (ERP) Systems through Artificial Intelligence," *International Numeric Journal of Machine Learning and Robots*, vol. 7, no. 7, pp. 1–15, Dec. 2023.

[9] P. Pokala, *The Integration And Impact Of Artificial Intelligence In Modern Enterprise Resource Planning Systems: A Comprehensive Review*, SSRN Scholarly Paper, Rochester, NY, Nov. 2024. doi: 10.2139/ssrn.5069295. Social Science Research Network: 5069295.

[10] K. Pakhale, *Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges*, Sep. 2023. doi: 10.48550/arXiv.2309.14084. arXiv: 2309.14084 [cs].

[11] Y. Li *et al.*, *Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security*, May 2024. doi: 10.48550/arXiv.2401.05459. arXiv: 2401.05459 [cs].

[12] C. Ling *et al.*, *Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey*, Mar. 2024. doi: 10.48550/arXiv.2305.18703. arXiv: 2305.18703 [cs].

[13] A. Balaguer *et al.*, *RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture*, Jan. 2024. doi: 10.48550/arXiv.2401.08406. arXiv: 2401.08406 [cs].

[14] Z. Zhang *et al.*, *Personalization of Large Language Models: A Survey*, May 2025. doi: 10.48550/arXiv.2411.00027. arXiv: 2411.00027 [cs].

[15] H. Abu-Rasheed, C. Weber, and M. Fathi, "Knowledge Graphs as Context Sources for LLM-Based Explanations of Learning Recommendations," in *2024 IEEE Global Engineering Education Conference (EDUCON)*, May 2024, pp. 1–5. doi: 10.1109/EDUCON60312.2024.10578654. arXiv: 2403.03008 [cs].

[16] J. Liu, J. Lin, and Y. Liu, *How Much Can RAG Help the Reasoning of LLM?* Oct. 2024. doi: 10.48550/arXiv.2410.02338. arXiv: 2410.02338 [cs].

[17] H. N. Patel, A. Surti, P. Goel, and B. Patel, "A Comparative Analysis of Large Language Models with Retrieval-Augmented Generation based Question Answering System," in *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Kirtipur, Nepal: IEEE, Oct. 2024, pp. 792–798, isbn: 9798350376425. doi: 10.1109/I-SMAC61858.2024.10714814.

[18] X. Han *et al.*, "OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 169–174. doi: 10.18653/v1/D19-3029.

[19] A. Gupte *et al.*, *REBEL: Rule-based and Experience-enhanced Learning with LLMs for Initial Task Allocation in Multi-Human Multi-Robot Teams*, Sep. 2024. doi: 10.48550/arXiv.2409.16266. arXiv: 2409.16266 [cs].

[20] V. A. Batista, D. S. M. Gomes, and A. G. Evsukoff, *SESAME - Self-supervised framework for Extractive queStion Answering over docuMent collEctions*, Mar. 2024. doi: 10.21203/rs.3.rs-4018202/v1.

[21] S. Setty, H. Thakkar, A. Lee, E. Chung, and N. Vidra, *Improving Retrieval for RAG based Question Answering Models on Financial Documents*, Aug. 2024. doi: 10.48550/arXiv.2404.07221. arXiv: 2404.07221 [cs].

[22] X. Zhou, X. Zhao, and G. Li, *LLM-Enhanced Data Management*, Feb. 2024. doi: 10.48550/arXiv.2402.02643. arXiv: 2402.02643 [cs].

[23] B. P. Allen, L. Stork, and P. Groth, *Knowledge Engineering using Large Language Models*, Oct. 2023. doi: 10.48550/arXiv.2310.00637. arXiv: 2310.00637 [cs].

[24] D. Dua, E. Strubell, S. Singh, and P. Verga, "To Adapt or to Annotate: Challenges and Interventions for Domain Adaptation in Open-Domain Question Answering," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 14 429–14 446. doi: 10.18653/v1/2023.acl-long.807.

[25] S. Siriwardhana *et al.*, "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, Jan. 2023, issn: 2307-387X. doi: 10.1162/tacl_a_00530.

[26] M. E. J. Newman, *Modularity and community structure in networks*, https://www.pnas.org/doi/epdf/10.1073/pnas.0601602103, 2006. doi: 10.1073/pnas.0601602103.

[27] V. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: Guaranteeing well-connected communities," *Scientific Reports*, vol. 9, no. 1, p. 5233, Mar. 2019, issn: 2045-2322. doi: 10.1038/s41598-019-41695-z. arXiv: 1810.08473 [cs].

[28] *Ollama*, https://ollama.com.

[29] D. Edge *et al.*, *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*, Apr. 2024. doi: 10.48550/arXiv.2404.16130. arXiv: 2404.16130 [cs].

[30] *Code Implementation*, https://tinyurl.com/43zb4duk.

[31] *Opik LLM development Platform | Observability, Evaluation & Security*, https://www.comet.com/docs/opik.