An Evidence-Aware and Risk-Sensitive Retrieval-Augmented Generation Framework for Internal Auditing

Tareq Fahad Aljabri, Mariam Abdulaziz Alnajim
Saudi Telecom Company, Riyadh, Saudi Arabia

***Abstract*—Large Language Models (LLMs) that are enhanced with Retrieval-Augmented Generation(RAG) can aid in internal auditing, particularly in search and analysis of documents. However, in general, most RAG-based audit tools focus more on quick document access and being easy to use, then about deeper auditing reasons. They don't do much to help with significant audit procedures, such as maintenance of clear evidence, calculating risk, and making intelligent decisions. Because of this, they are yet to find a place in ongoing internal auditing, which needs serious evidence and must adhere to recommended auditing standards. This study introduces an *Evidence-Aware and Risk-Sensitive Retrieval-Augmented Generation (ER^2-RAG)* to help in internal auditing. The framework doesn't just see how well documents can be retrieved, but also manages audit evidence and considers risk. It connects audit conclusions to supporting documents with confidence levels, modifies the information retrieval based on the audit risk and materiality, and restricts the process of generation to the standard audit reasoning practices. These design choices make AI assistance more transparent, reliable, and defensible in audit judgments. ER^2-RAG has been developed and evaluated using the normal audit situations. These situations are related to the analysis of exceptions, evaluation of control efficiency, and control over procedural adherence. The research uses design science methodology. Compared to the older methods of RAG, ER^2-RAG is efficient and presents a higher scope of evidence, references the sources much better, and the argument is clearer. The results indicate that risk sensitivity must be taken into account, and evidence should be used when adopting AI systems to perform continuous internal audits. The given research transforms RAG into not only aiding in information retrieval but also as a powerful reasoning foundation of professional assurance. It strives to enhance audit reliability and guide future development of evidence-aware AI systems.**

***Keywords*—LLM; RAG; Audit; digitalization; automation**

I. Introduction

Transformer-based Large Language Models (LLMs) have begun a new period in AI applications in many professional fields [1], [2]. This has changed the way we comprehend natural language [3]. More researchers and practitioners are using models like GPT3 and GPT4 to perform financial and audit procedures such as the analysis of documents, authoring planning, and doing analytical processes [4]. Due to this, there is an increasing interest to solve problem in internal auditing using LLMs. This includes handling more and more information, complex regulations, and necessity of constant assurance [5].

In spite of the current developments, not all of the LLM application is suitable for internal auditing [6]. The majority of general conversation tools are only able to maintain limited context when chatting and require users to post documents individually. They are not very clear regarding the data that was provided to them with the answers. This is a significant weakness of internal audits, since the conclusions reached in the audit should be backed by traceable evidence, and should conform to specified standards [7]. The previous research has attempted to address this issue through Retrieval-Augmented Generation (RAG) with AI agents. This enabled them to gain access to confidential sources of audit knowledge and generate more accurate responses [8].

The latest studies have revealed that RAG-based AI agents are more successful than standalone LLMs at tasks such as searching documents, table processing data, and automation of audit workflows [9]. Nonetheless, such systems are more concerned with the data that they access rather than the way that auditors make inferences based on that data [10]. Internal auditors need to collect the appropriate data, evaluate its reliability, take audit risks into account, and report the findings to further review and regulation [11]. Current RAG frameworks do not clearly explain the reasoning and evidence needed [12].

A continuous auditing model, in which controls, transactions, and risks are continuously assessed rather than through recurring engagements, is becoming increasingly popular in internal auditing [13]. This change alters the expectations for supporting technologies. AI systems must be able to gather evidence over time, reevaluate risk when circumstances change, and support expert decisions that can be convincingly defended in such circumstances [14]. Because they lack explicit mechanisms for risk-aware reasoning and evidence attribution, current RAG-based agents fall short in meeting these needs. Their suitability for this new internal auditing method is therefore still limited [15].

This study presents an Evidence-Aware and Risk-Sensitive RAG (ER^2-RAG) framework to address these problems and improve intelligent internal auditing. ER^2-RAG integrates audit-specific reasoning constraints, explicit evidence traceability, and risk-weighted knowledge retrieval within an AI agent architecture, in contrast to earlier methods that prioritize retrieval efficiency. From basic information assistants to sophisticated audit reasoning systems that can support continuous internal auditing procedures, the framework seeks to develop RAG-enhanced agents.

Three major contributions are made by this study. It first presents a system that links verifiable source documents and confidence indicators to audit conclusions produced by artifi-

cial intelligence. Second, it proposes a risk-sensitive retrieval approach that aligns information access with materiality and audit risk considerations. Third, it demonstrates how AI-assisted audit decisions can be made more transparent and consistent by using reasoning-limited generation. When taken as a whole, these contributions improve RAG-based auditing systems and set the stage for further studies on reliable AI in assurance services.

The rest of the study is organized as follows: Section II discusses the related work. Section III discusses the proposed methodology. Section IV discusses the detailed experiment results, and Section V concludes the findings of the proposed work.

II. Related Work

According to recent research, the procedural, multi-step, and document-intensive nature of auditing makes it a good fit for LLM-based agents [16]. Accordingly, auditors gain more from a structured agent workflow that can: 1) direct a query to the appropriate subtask, 2) apply standardized audit methods via fixed prompts, and 3) connect to private audit repositories through retrieval mechanisms than from a single chatbot interaction. In order to minimize manual document handling while maintaining methodological consistency, Xiong et al. [17] describe an audit-agent design that revolves around these concepts and combines retrieval-augmented generation, customizable workflows (such as sequential/parallel agents), and prompt orchestration.

Retrieval-Augmented Generation is commonly used to address two persistent limitations of foundation LLMs in professional applications, namely knowledge cutoff and hallucinated outputs. In enterprise settings, RAG research has focused on building complete pipelines that span document ingestion, text chunking, embedding generation, vector database indexing, reranking strategies, and security and privacy controls. This work is typically accompanied by evaluation methods that assess retrieval quality and the faithfulness of generated answers [18], [19], [20], [21], [22]. In auditing contexts, the requirement for grounded responses becomes more critical, since audit conclusions must be directly traceable to authoritative evidence. Within audit-agent designs, RAG is therefore used as the primary mechanism for linking auditors to internal procedure manuals, workpapers, and regulatory guidance, without relying on repeated manual document uploads [17], [23], [24].

Audit procedures routinely require auditors to work with evidence that appears in different formats. Structured tables, including trial balances, accounts payable aging reports, and revenue listings, must often be analyzed together with narrative policy documents and prior audit workpapers [25], [26], [27]. This combination creates practical challenges for automated systems. Prior research shows that general-purpose LLMs perform poorly in such contexts, particularly when numerical consistency must be preserved, when information is spread across multiple documents, or when tabular evidence needs to be accurately reflected in generated responses [24], [28], [29], [30], [31]. These shortcomings have increased interest in RAG-enhanced agents for tabular processing and have driven the development of architectures that enable retrieval from both unstructured text and table-aware representations, ensuring that responses are grounded in retrieved evidence and adhere to a predefined audit response schema [32], [33].

Baseline vector RAG can work well for local, pinpoint questions, but audit repositories contain dense cross-document dependencies (policies $\rightarrow$ procedures $\rightarrow$ templates $\rightarrow$ exceptions), where the "best" evidence may be distributed across multiple files. Graph-based RAG approaches address this by building an entity-relationship index and generating community summaries that support corpus-level (global) questions and sensemaking [34]. In parallel, blended/hybrid retrieval methods combine dense and sparse retrievers, or integrate knowledge graphs with vector retrieval to improve recall and robustness under terminology variation [35], [36]. Audit-agent work explicitly discusses evaluating baseline RAG against GraphRAG to better handle interconnected audit documentation [17].

"Helpful" responses are insufficient for audit use unless they are consistent with the engagement context, verifiable, and permission-aware. Evaluation beyond fluency, such as retrieval metrics (e.g., precision/recall, MRR) and generation metrics related to faithfulness and correctness, is becoming more and more important in the RAG literature [22], [19], [21]. Layered controls and validation gates are frequently motivated by governance concerns (confidentiality, scoped access, and evidence-only claims) in applied internal audit contexts. The need for careful design and monitoring is reinforced by related work in internal-auditing education, which also highlights privacy, ethical use, and reliability issues when implementing generative systems [37], [38], [39].

III. Methods

The ER^2-RAG framework, shown in Fig. 1, accepts a straightforward natural language question from an auditor, which can be supplemented with basic audit context like the business process being examined and the time frame of the audit. The query can be stated in simple terms and does not need to employ technical auditing terminology or established query formats.

The framework creates a structured, evidence-based audit response that includes a clear conclusion, clear references to the audit evidence that supports it, an assessment of the evidence's strength, and a confidence level that is suitable for professional evaluation and documentation.

This framework aims to provide structure and clarity to an investigation that is usually informal. There are several steps involved in the process. These involve the breaking down of tasks, collecting evidence while also minding the potential risks, plainly displaying the source of evidence, and employing normal audit procedures rather than simply answering the question. Each step adds context and organization. It gradually introduces risk factors, audit-related issues, and supporting evidence. As a result, a simple question such as "Does the company check invoices before paying them?", is converted into a structured audit evaluation that links findings with evidence-based data and is displayed to the auditor in a form suitable for professional review and documentation.

A. Layer-Wise Explanation of the ER^2-RAG Framework

This subsection explains how an input query is progressively transformed as it passes through each layer of the pro-

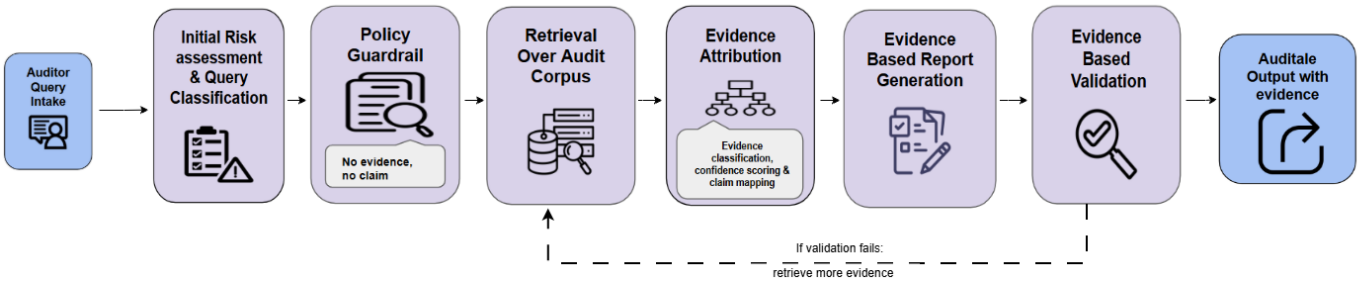

Fig. 1. Flowchart representation of the ER^2-RAG methodology for internal auditing.

posed ER^2-RAG framework. For demonstration lets consider a simple, layman-level query: *"Are invoices checked properly before payments are made?"*

1) Layer 1: Intake and task routing: The auditor submits a raw natural-language query to the layer. The query is currently unstructured and might not include specific audit-related terms. Understanding the purpose of the query and classifying it into the proper audit task type are the goals of the intake and task routing layer. The system identifies that the query is about checking how effective the controls are for verifying invoices in the accounts payable process. It uses intent recognition and contextual clues to do this. The system also assesses an initial risk level based on the features of the process and any potential financial impacts. This step creates a clear task outline that directs all further actions by specifying the audit goal, the type of task, and the response format needed.

2) Layer 2: Policy guardrails: The structured task definition of the last layer is the input to the policy guardrails layer. This layer is not altering the meaning of the query. Rather, it establishes explicit guidelines on how one will access it. The system maintains access control as well as confidentiality, which ensures that only authorized audit documents are utilized. An important purpose of this layer is to abide by the principle that if there is no evidence, there should be no claim. This is a rule that prevents the system from making conclusions without evidence to support that conclusion. This layer provides a safe layer that enables the query to be processed safely whilst adhering to the audit regulations.

3) Layer 3: Risk-Sensitive Retrieval (ER layer): The retrieval layer is risk-sensitive and utilizes initial query and task parameters to locate the relevant audit evidence in the private knowledge base. First, it broadens the query to fit certain terms of the process, and auditing of it is useful in improving retrieval. The system then does a hybrid search, which involves searching by semantic similarity and by keyword matching. The retrieved documents are then ranked with the help of a scoring system that concentrates on risk. Such a system places emphasis on the most recent versions, documents about the high-risk areas, and documents by trusted sources. The result of this layer is a ranked list of evidence items in decreasing order of their significance. The information in each piece comprises the type of the source, version, and relevance score.

4) Layer 4: Evidence Attribution Engine (E^2 layer): The evidence attribution layer processes the ranked evidence chunks that have been retrieved previously. In this layer documents are retrieved and assessed according to the contribution each has had to a particular auditing task rather than being considered as an unstructured collection. The query is divided into the explicit audit claims, i.e. whether invoices are inspected and whether the inspections take place regularly. The documents are then linked to every claim and classified as either direct, indirect, or inferred evidence. The confidence levels are based on the reliability and the consistency of evidence available. Propositions which are not supported enough are removed out of reasoning. Consequently, this layer generates organized claim-evidence mapping, which can be revised and traced to references.

5) Layer 5: Reasoning-constrained generation: This is where the emphasis is placed on putting together an audit response that will be based on the previous claims. This response is grounded on fundamental auditing principles and is dependent upon the already gathered evidence. In the example of verifying invoices, an answer is provided on whether there is a review process, where it might not be complete enough, and what risks are associated with that. The result of this is an early audit report that adheres to regular audit reasoning and can be traced to supporting documentation.

6) Layer 6: Validation and quality gates: Here, the draft audit response is verified, after which it is finalized. This is to ensure that nothing significant is left out and that the response is sensible. Simple checks are made to determine whether the format intended is correct, whether evidence supports the claims, and whether it brings any contradiction. In case an issue is discovered, additional information may be entered, or the confidence level may be re-established. Once this check is made, the audit reply will be available to be checked.

7) Layer 7: Audit trail and continuous monitoring: In the final stage, final audit response and supporting information are stored. This involves the original question, the documents involved, the findings that have been made, and the confidence level. Maintaining these records simplifies the reviewing of the work in the future and allows the audits to be tracked over time. In case the same questions crop up, these records can be reexamined to determine whether there has been any change of risk or controls. A brief summary of the layers and their functions is provided in Table I.

IV. Experimental Results

A. Audit Query Set

In order to test the ER^2-RAG model, we applied a series of real-life audit questions. The questions were provided in

TABLE I. STEP-BY-STEP PROCESSING OF AN AUDIT QUERY IN THE PROPOSED ER^2-RAG FRAMEWORK

Stage	Component	Input	Processing Function	Output	Role of RAG
0	Auditor / User	Natural-language query + audit context	User submits a question in everyday language (e.g., *"Are invoices checked properly before payments are made?"*)	Raw audit query	–
1	Intake & Task Router	Raw query + context	Classifies the query type (e.g., control evaluation), detects risk and materiality, and selects the appropriate audit response template	Structured audit task definition	RAG not yet applied
2	Policy Guardrails	Task definition	Enforces audit scope, access permissions, confidentiality constraints, and the *no-evidence-no-claim* rule	Constrained execution environment	RAG constrained by governance
3	Risk-Sensitive Retrieval (ER Layer)	Structured task + expanded query	Performs query expansion and hybrid retrieval (vector, keyword, and graph-based search). Ranks documents based on risk relevance, authority, and recency	Ranked evidence chunks with metadata	**Primary RAG retrieval stage**
4	Evidence Attribution Engine (E^2 Layer)	Retrieved evidence chunks	Maps retrieved documents to explicit audit claims, classifies evidence strength (direct, indirect, inferred), and assigns confidence scores	Evidence-backed claim structure	**Transforms RAG output into grounded evidence**
5	Reasoning-Constrained Generation	Evidence-backed claims + templates	Generates audit responses using structured audit reasoning (criteria, condition, cause, effect, and recommendation)	Draft audit conclusion with citations	**RAG constrains generation to evidence only**
6	Validation & Quality Gates	Draft response + evidence	Verifies citation coverage, detects contradictions, validates numerical consistency, and enforces output schema compliance	Validated audit response	**RAG feedback loop (retrieve or regenerate if needed)**
7	Audit Trail & Continuous Monitoring	Final response + metadata	Logs the query, supporting evidence, confidence scores, and version changes to support reviewability and continuous auditing	Versioned audit record	RAG supports future comparisons
–	Final Output	Validated response	Presents the final answer with explicit citations, confidence level, and evidence summary	Review-ready audit response	Fully RAG-grounded output

the same format that auditors typically pose when they are doing their planning, field work, and review. Technical or other fixed formats were not used. We kept the language simple and focused on common audit concerns, such as risk, compliance, and whether controls are actually working.

Twenty five de-identified audit questions covering multiple internal audit domains, such as accounts payable, payroll, access control, procurement, and policy compliance, made up the final query set. Instead of just requiring factual retrieval, each query was intended to require evaluative reasoning backed by documentary evidence. This design decision made sure that the evaluation assessed the framework's capacity to convert unstructured inquiries into organized, fact-based audit conclusions.

Common audit task categories, including control operation assessment, compliance verification, exception identification, and monitoring activities, were used to classify the queries. A subset of representative queries used in the experiments is shown in Table II, along with the audit focus and task type that correspond to each query.

For both the proposed ER^2-RAG framework and the baseline RAG system, each query was run independently under the same conditions. This made sure that variations in performance could be linked to the design of the framework rather than the creation of the queries. The evaluation was also able to determine how well the framework enriches minimal user input with audit-specific context, risk awareness, and evidence grounding thanks to the use of straightforward, intuitive queries.

All things considered, the audit query set provides a useful standard for evaluating the framework's capacity to close the gap between unofficial auditor questions and audit outputs that can be professionally defended.

B. Evaluation Metrics

In order to measure the system applied in internal auditing, our measures need to be more than the soundness of the responses. When evaluating an automated response in professional audit practice, we evaluate its relevance or clarity, the effectiveness with which it is supported by data, its potential fit to risk factors, and its relevance to audit working papers. As a result, both technical performance and audit-specific quality dimensions are evaluated using the evaluation metrics employed in this study.

1) Evidence coverage rate: The percentage of material statements in the generated response that are specifically backed up by cited audit evidence is known as the evidence coverage rate. For an audit to be defensible, conclusions must have a stronger foundation in verifiable documents, which is indicated by a higher coverage rate.

2) Citation correctness: Citation correctness assesses whether the referenced materials actually bolster the assertions they are linked to. In order to make sure that citations are not only pertinent but also significantly support the corresponding audit statements, this metric was evaluated manually.

3) Unsupported claim rate: The percentage of statements in the output that are not supported by direct or indirect evidence is known as the "unsupported claim rate". Since unsubstantiated claims erode professional skepticism and raise audit risk, lower values are preferred.

4) Response consistency: The variability of results when the same query is run repeatedly under the same circumstances is measured by response consistency. In audit contexts where reproducibility is crucial, high variability may be a sign of unstable reasoning, which is undesirable.

5) Audit readiness: The qualitative indicator of whether a generated response can be added to audit documentation with little change is called audit readiness. The answers received the

TABLE II. EXAMPLES FROM THE AUDIT QUERY SET USED IN EXPERIMENTAL EVALUATION

ID	Audit Query (Example)	Audit Area	Task Type
Q1	Are invoices checked properly before payments are made?	Accounts Payable	Control effectiveness
Q2	Are purchase orders approved before goods are received?	Procurement	Compliance verification
Q3	Is access to the payroll system reviewed regularly?	Payroll	Control monitoring
Q4	Are terminated employees removed from systems on time?	Access Control	Risk identification
Q5	Are changes to vendor master data properly authorized?	Vendor Management	Fraud risk assessment
Q6	Are manual overrides of system controls reviewed and approved?	IT Controls	Exception analysis
Q7	Are financial policies updated and communicated to staff?	Policy Governance	Compliance monitoring

TABLE III. COMPARISON OF BASELINE RAG AND ER^2-RAG USING AUDIT-ORIENTED EVALUATION METRICS

Metric	Baseline RAG	ER^2-RAG
Evidence Coverage Rate (%)	68.4	**92.1**
Citation Correctness (%)	71.2	**94.5**
Unsupported Claim Rate (%)	18.7	**4.3**
Response Consistency (Std. Dev.)	0.21	**0.08**
Audit Readiness	Medium	**High**

TABLE IV. ABLATION STUDY RESULTS FOR THE ER^2-RAG FRAMEWORK

Configuration	Evidence Coverage (%)	Citation Correctness (%)	Unsupported Claims (%)	Audit Readiness
Full ER^2-RAG	**92.1**	**94.5**	**4.3**	High
Without Risk-Sensitive Retrieval	84.7	88.9	8.6	Medium–High
Without Evidence Attribution	76.2	79.5	15.8	Medium
Without Validation Gates	81.4	83.1	11.2	Medium

rating Low, Medium, or High depending on the professional tone, clarity, structure, and evidence utilization.

All in all, these ratings give a clear picture of the extent to which the framework is capable of supporting practical internal auditing assignments.

As indicated by the outcome in Table III, ER^2-RAG has always demonstrated better performance than the base RAG strategy in all the considered dimensions. The high performance of the evidence attribution and Risk-sensitive retrieval layers is indicated by the marked increase in the accuracy of citation and evidence coverage. Reasoning constraints and validation gates also help produce audit output that is more stable and defensible as demonstrated by the reduction of unsupported claims and the rise in the consistency of responses. All these findings indicate that ER2-RAG is more compliant with professional internal auditing demands, compared to the traditional systems that rely on RAG.

C. Ablation Study

To test the contribution of each of the components of the proposed ER^2-RAG framework, an ablation study was conducted. This analysis was conducted to know the influence of key elements on the overall performance of the system with special attention to audit critical areas such as support of evidence, reasonableness, and audit preparation.

We started with the entire ER^2-RAG system and switched off each of the elements one by one to observe the difference. We did not use the risk-based retrieval in certain tests. In other cases, we omitted the stage of evidence attribution or omitted the last checks of validation. All other factors were retained in order to make the comparison fair.

All versions were tested with the same audit questions as well as evaluation measures. By doing this, any variation in the outcome is a result of the part that was eliminated and not the variations in the data or the system configuration.

Table IV demonstrates the impact of the different elements of the ER^2-RAG framework on system performance. When the risk-sensitive retrieval mechanism is eliminated, we observe a vast decrease in the coverage of evidence and accuracy of citations. This demonstrates that the ranking of sources according to risk is essential in obtaining reliable audit sources. The unsupported claims are significantly higher with the evidence attribution layer turned off, and the performance decreases the most. This highlights the significance of definite claim-evidence mapping towards auditing defensibility. Equally, the absence of quality and validation gates lets statements that have minimal or no support, which lessens reliability.

All things considered, the ablation study confirms that the recommended components function together rather than separately. The full ER^2-RAG configuration consistently outperforms all ablated variations, validating the architectural

design choices and confirming the framework's suitability for professional internal auditing applications.

V. Conclusion

This research study has proposed ER2-RAG, an evidence-conscious and risk-sensitive RAG model of internal audit. The primary theoretical contribution is the design of audit-oriented RAG as a multi-layered architecture, which combines retrieval, evidence attribution, structured reasoning, validation, and audit trail generation into one structure. In this respect, the study adds new knowledge, demonstrating how generative AI may be customized to address the traceability, reliability and compliance requirements of internal audit.

The findings indicate that ER2-RAG has a significant positive impact on evidence coverage, citation correctness, response consistency, and audit readiness and fewer unsupported claims than usual RAG. The ablation experiment also supports the fact that risk-sensitive retrieval, evidence attribution, and validation gates have all effects on the overall performance. These results indicate the significance of evidence-based and audit-focused controls of credible AI-assisted auditing.

There are certain limitations in this study. It was tested on a small sample of de-identified audit queries and was not aimed at benchmarking at scale or more diverse corpus or security hardening. Future directions can build on the framework by assessing the framework on larger sets of auditors, more comparison to baseline, and more defense against an adversarial or misleading retrieved content. All in all, this study offers a conceptual and practical basis of reliable and audit compliant internal auditing generative AI systems.

References

[1] M. A. Bakr, A. J. Khan, S. D. Khan, M. H. Zafar, M. Ullah, and H. Ullah, "Evaluation of learning-based models for crop recommendation in smart agriculture," *Information*, vol. 16, no. 8, p. 632, 2025.

[2] H. Ullah, S. D. Khan, M. Ullah, M. Mahmud, and F. A. Cheikh, "Generative adversarial networks: A short review," *Electronic Imaging*, vol. 32, pp. 1–7, 2020.

[3] M. Raza, Z. Jahangir, M. B. Riaz, M. J. Saeed, and M. A. Sattar, "Industrial applications of large language models," *Scientific Reports*, vol. 15, no. 1, p. 13755, 2025.

[4] M. M. Dong, T. C. Stratopoulos, and V. X. Wang, "A scoping review of chatgpt research in accounting and finance," *International Journal of Accounting Information Systems*, vol. 55, p. 100715, 2024.

[5] H. Li, M. M. de Freitas, H. Lee, and M. Vasarhelyi, "Enhancing continuous auditing with large language models: A framework for cross-verification using exogenous textual data," *Available at SSRN*, vol. 4692960, 2024.

[6] F. Xiong, Q. Han, and C. Zhang, "Design ai agent for auditing: Applying large language models (llms) and retrieval augmented generations (rag) to audit workflows," *Journal of Emerging Technologies in Accounting*, pp. 1–10, 2025.

[7] S. Wang, F. Zhao, D. Bu, Y. Lu, M. Gong, H. Liu, Z. Yang, X. Zeng, Z. Yuan, B. Wan *et al.*, "Lins: A general medical q&a framework for enhancing the quality and credibility of llm-generated responses," *Nature communications*, vol. 16, no. 1, p. 9076, 2025.

[8] E. Karakurt and A. Akbulut, "Retrieval-augmented generation (rag) and large language models (llms) for enterprise knowledge management and document automation: A systematic literature review," *Applied Sciences*, vol. 16, no. 1, p. 368, 2025.

[9] T. Nguyen, P. Chin, and Y.-W. Tai, "Ma-rag: Multi-agent retrieval-augmented generation via collaborative chain-of-thought reasoning," *arXiv preprint arXiv:2505.20096*, 2025.

[10] H. Lu and Z. Wu, "Revisiting intelligent audit from a data science perspective," *Neurocomputing*, p. 129431, 2025.

[11] S. Selishchev, "Structural aspect of checking the continuity of the enterprise during internal audit," *Statistics of Ukraine*, vol. 8990, no. 2-3, pp. 155–162, 2020.

[12] K. Xu, K. Zhang, J. Li, W. Huang, and Y. Wang, "Crp-rag: A retrieval-augmented generation framework for supporting complex logical reasoning and knowledge planning," *Electronics*, vol. 14, no. 1, p. 47, 2024.

[13] H. B. Hazar, "New paradigm in auditing: Continuous auditing," in *Ethics and Sustainability in Accounting and Finance, Volume II*. Springer, 2020, pp. 253–268.

[14] A. Sabuncuoglu, C. Burr, and C. Maple, "Justified evidence collection for argument-based ai fairness assurance," in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2025, pp. 18–28.

[15] B. Ni, Z. Liu, L. Wang, Y. Lei, Y. Zhao, X. Cheng, Q. Zeng, L. Dong, Y. Xia, K. Kenthapadi *et al.*, "Towards trustworthy retrieval augmented generation for large language models: A survey," *arXiv preprint arXiv:2502.06872*, 2025.

[16] M. Valadao, N. Freire, M. de Paula, L. Almeida, and L. Marques, "Using large language models to support the audit process in the accountability of interim managers in notary offices."

[17] F. Xiong, "Design ai agent for auditing: Applying large language models (llms) and retrieval augmented generations (rag) to audit workflows," *Journal of Emerging Technologies in Accounting*, 2025.

[18] C. Joshua, A. Banerjee, M. A. Kaplan, A. Willie, P. R. Kalluri, W. Agnew, and V. Chouhan, "Building retrieval-augmented generation (rag) for internal knowledge bases," Preprint/technical report, 2025.

[19] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[20] S. Wu *et al.*, "A survey on retrieval-augmented generation," *arXiv preprint arXiv:2310.11166*, 2023.

[21] A. Waqas, S. D. Khan, Z. Ullah, M. Ullah, and H. Ullah, "Comparative analysis of deep learning models for intrusion detection in iot networks," *Computers*, vol. 14, no. 7, p. 283, 2025.

[22] J. Lin *et al.*, "A framework for evaluating retrieval-augmented generation models," *arXiv preprint arXiv:2308.01899*, 2023.

[23] M. A. Rahim, M. Mushafiq, S. D. Khan, R. Ullah, S. Khan, and M. Ishaque, "Technical analysis-based unsupervised intraday trading djia index stocks: is it profitable in long term?" *Appl. Intell.*, vol. 55, no. 2, p. 199, 2025.

[24] F. Xiong, "Retrieval-augmented generation ai agent for tabular data processing in auditing procedures," SSRN Working Paper, 2024, sSRN abstract id: 5100205.

[25] M. El-Haj, P. Alves, P. Rayson, M. Walker, and S. Young, "Retrieving, classifying and analysing narrative commentary in unstructured (glossy) annual reports published as pdf files," *Accounting and Business Research*, vol. 50, no. 1, pp. 6–34, 2020.

[26] M. A. Rahim, S. D. Khan, S. Khan, M. Rashid, R. Ullah, H. Tariq, and S. Czapp, "A novel spatio–temporal deep learning vehicle turns detection scheme using gps-only data," *IEEE Access*, vol. 11, pp. 8727–8733, 2023.

[27] D. Appelbaum, "Securing big data provenance for auditors: The big data provenance black box as reliable evidence," *Journal of emerging technologies in accounting*, vol. 13, no. 1, pp. 17–36, 2016.

[28] A. L. Silva, A. M. Lima, G. Valença, and G. G. Cabral, "Ad-hoc vs llm based system for information retrieval in large tabular data: A comparative study in public medicine procurement audits," in *Simpósio Brasileiro de Sistemas de Informação (SBSI)*. SBC, 2025, pp. 751–758.

[29] R. Ullah, M. H. bin Hasan, S. D. Khan, and M. A. Rahim, "Secure transmission of compressed medical image sequences on communication networks using motion vector watermarking." *Computers, Materials & Continua*, vol. 78, no. 3, 2024.

[30] M. Munsif, H. Afridi, M. Ullah, S. D. Khan, F. A. Cheikh, and M. Sajjad, "A lightweight convolution neural network for automatic disasters recognition," in *2022 10th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2022, pp. 1–6.

[31] M. Ahmad, "Toward a unified framework for information retrieval in large language model applications: Balancing textual and graph-based knowledge sources," 2025.

[32] S. S. Murtaza, Y. Nie, E. Avan, U. Soni, W. Liao, A. Carnegie, C. J. Mathias, J. Jiang, and E. Wen, "Implementing retrieval augmented generation technique on unstructured and structured data sources in a call center of a large financial institution," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, 2025, pp. 598–606.

[33] S. Adhikari, "Large language models in modern data engineering: A systematic review of architectures, use cases, and limitations," *International Journal of Business & Computational Science*, vol. 2, no. 1, 2025.

[34] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.

[35] K. Sawarkar, A. Mangal, and S. R. Solanki, "Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers," *arXiv preprint arXiv:2404.07220*, 2024.

[36] B. Sarmah *et al.*, "Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction," *arXiv preprint arXiv:2408.04948*, 2024.

[37] O. Stumke and F. Ndlovu, "Transforming internal auditing: Harnessing retrieval-augmented generation technology," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 4, 2025.

[38] S. K. Parimi and R. Yallavula, "Generative ai for enterprise trust: A governance-aligned framework for safe and transparent automation at global scale," *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, vol. 6, no. 1, pp. 218–225, 2025.

[39] A. O. Adebayo, O. F. Makinde, O. A. Olasehan, N. A. Akande, and U. J. Eziokwu, "Harnessing the digital prometheus: A strategic framework for generative ai governance, risk, and control." *Educational Research (IJMCER)*, vol. 7, no. 6, pp. 204–214, 2025.