HAKEMLİ Denetişim
Ortak Aklın Harmanı

Kamu İç Denetçileri Derneği Meşrutiyet Caddesi Konur Sokak No: 36/6 Kızılay - ANKARA
www.kidder.org.tr/denetisim/ • denetisim@kidder.org.tr
ISSN 1308-8335
Yıl: 17, Sayı: 34, 174-184, 2026

Research Article *(Araştırma Makalesi)*

CAN LARGE LANGUAGE MODELS ACT AS "CO-AUDITORS"?

(BÜYÜK DİL MODELLERİ "ORTAK DENETÇİ" GİBİ HAREKET EDEBİLİR Mİ?)

Hakan EMEKCİ[1]

ABSTRACT

This study explores the integration of large language models (LLMs) into audit workflows as "co-auditors," emphasizing the necessity of embedding them within frameworks that ensure evidence traceability, governance, and human accountability. Despite growing interest in AI-augmented auditing, prior work has not systematically bridged LLM technical capabilities with audit standards and regulatory compliance requirements. Through a narrative literature review synthesizing audit doctrine, AI governance frameworks, and natural language processing research, the study examines how such integration can be achieved.

Rather than substituting professional judgment, LLMs offer auditable support that enhances assurance processes. By incorporating hybrid retrieval, policy-constrained generation, and cryptographic provenance, the proposed architecture addresses both factual reliability and regulatory compliance. The findings underscore that effective LLM deployment requires strict alignment with standards. Ultimately, the research confirms that trustworthy AI in auditing depends on robust technical safeguards, governance structures, and sustained human oversight.

Keywords: Large Language Models (LLMs), Co-Auditor Architecture, Evidence Traceability, Artificial Intelligence (AI) Governance, Regulatory Compliance

JEL Classification: M42, C45, O33

ÖZ

Bu çalışma, büyük dil modellerinin (BDM) denetim iş akışlarına "eş denetçiler" olarak entegrasyonunu inceleyerek, kanıt izlenebilirliğini, yönetişimi ve insan hesap verebilirliğini sağlayan çerçevelere entegre edilmesinin gerekliliğini vurgulamaktadır. Yapay zeka destekli denetime artan ilgiye rağmen, önceki çalışmalar BDM'nin teknik yeteneklerini denetim standartları ve düzenleyici uyumluluk gereklilikleriyle sistematik olarak birleştirmemiştir. Denetim doktrini, yapay zeka yönetişim çerçeveleri ve doğal dil işleme araştırmalarını sentezleyen bir anlatı literatür taraması yoluyla, çalışma böyle bir entegrasyonun nasıl sağlanabileceğini incelemektedir.

BDM'ler, mesleki yargının yerini almak yerine, güvence süreçlerini geliştiren denetlenebilir destek sunmaktadır. Hibrit erişim, politika kısıtlamalı üretim ve kriptografik köken gibi özellikleri bir araya getirerek, önerilen mimari hem olgusal güvenilirliği hem de düzenleyici uyumluluğu ele almaktadır. Bulgular, etkili bir BDM uygulamasının standartlarla sıkı bir uyum gerektirdiğinin altını çizmektedir. Sonuç olarak araştırma, denetimde güvenilir yapay zekânın sağlam teknik güvenlik önlemlerine, yönetişim yapılarına ve sürekli insan gözetimine bağlı olduğunu doğrulamaktadır.

Anahtar Kelimeler: *Büyük Dil Modelleri (BDM), Ortak Denetçi Mimarisi, Kanıt İzlenebilirliği, Yapay Zeka Yönetişimi, Düzenleyici Uyum*

JEL Kodları: M42, C45, O33

[1] Assist. Prof., TED University, Graduate School, Applied Data Science, Ankara, Orcid Id: 0000-0002-4074-5600, hakan.emekci@tedu.edu.tr

1. INTRODUCTION

Auditing is based on a simple but stringent assumption: opinions are founded on sufficient appropriate evidence properly recorded so a qualified auditor can understand what was done, by whom, when, and why. For external audit, such requirements are embedded within Public Company Accounting Oversight Board (PCAOB) standards on audit evidence and documentation; within international practice, ISA (International Standard on Auditing) 230 similarly grounds documentation on professional judgment and reviewability. They define both possibilities and limitations of any technology that makes audit work possible (PCAOB, 2025).

LLMs deliver real-world advantages like faster policy searches, cleaner workpapers, and more unstructured evidence coverage. However new risks of reliability, security, and governance that are now within standard-setters' and regulators' sights. PCAOB’s 2024 “Generative AI Spotlight” identifies new uses and red flags, while International Auditing and Assurance Standards Board (IAASB) has made technology a priority at-board and is actively revamping ISA 500 (Audit Evidence). Internal-audit functions are also in motion: the Institute of Internal Auditors (IIA) issued an AI Auditing Framework consistent with its 2024 Global Internal Audit Standards. Nevertheless, supervisory reviews continue to cite capability-measurement gaps. For example, the Financial Reporting Council (FRC) of the United Kingdom (UK) recently found that big four firms are not always tracking AI impact on audit quality. Individually and collectively, these developments map out a relevant research question: are LLMs able to be “co-auditors” within existing audit expectations? (IAASB, 2010; IIA, 2025; PCAOB, 2024). By "co-auditor," the paper does not mean an independent decision-maker. It rather denotes an augmentative system utilized under human accountability that allows evidence compilation and analysis and documentation and upholds respect for objectivity and professional independence. The 2024 Standards of the IIA affirm independence as the state of assurance and work supported by AI remains attributable, reviewable, and owned by the engagement team. That is, LLMs can aid in audit work but never substitute human accountability, sign-off, or governance (IIA, 2025).

The literature points out why these weaknesses are significant. LLMs are powerful but fallible: recent surveys map hallucinations and brittle reasoning; communities of interest document prompt-injection, insecure output handling, and data-poisoning weaknesses. There is a unifying thread of work that asserts retrieval-augmented generation (RAG), combining dense/sparse retrieval and re-ranking with citation, can improve factuality and produce provenance that auditors can analyze. These findings instigate designs where each AI-aided assertion has citable sources and each transformation is auditable (Huang et al., 2023; Ji et al., 2023; Karpukhin et al., 2020; Lewis et al., 2021; Nogueira & Cho, 2019; OWASP Foundation, 2025).

Similarly, adoption of audits must co-exist within agreed governance processes. National Institute of Standards and Technology (NIST) Artificial Intelligence (AI) Risk Management Framework (AI RMF 1.0) outlines a risk-based cycle (Govern–Map–Measure–Manage); International Organization for Standardization - ISO/IEC 42001:2023 specifies an AI Management System standard; ISO/IEC 23894:2023 specifies guidance on AI risk management; and the European Union (EU) AI Act (Regulation [EU] 2024/1689) specifies requirements when bringing in general-purpose and high-risk applications. For companies, these co-exist with The Committee of Sponsoring Organizations of the Treadway Commission’s (COSO) Internal Control–Integrated Framework, still used as the basis of designing and monitoring controls within new tools. An effective co-auditor therefore must be designed and operated as an auditable control system and not a smart assistant (COSO, 2013; [ISO], 2023b, 2023a; NIST, 2024).

Despite these advances, existing literature exhibits a critical gap: while technical research demonstrates LLM capabilities in retrieval and generation, and governance frameworks articulate principles for AI accountability, these streams have not been systematically integrated within the specific context of audit practice. Technical studies rarely address audit evidence standards or professional responsibility requirements, while audit and governance literature has not yet comprehensively engaged with the architectural patterns such as retrieval-augmented generation and provenance tracking. That makes LLM assistance auditable and trustworthy. This study addresses that gap by bridging audit doctrine with contemporary natural language processing research. The primary contribution is a conceptual framework synthesizing LLM technical capabilities with audit standards and AI governance principles to define conditions under which LLMs can function as co-auditors.

This study aims to examine whether LLMs can function as “co-auditors” within the boundaries of current audit standards and professional accountability frameworks. It argues that such a role is only feasible if LLM outputs are evidence-based, traceable, secure, and embedded within robust governance structures. By bridging audit doctrine with contemporary natural language processing (NLP) technologies, the study seeks to ensure that AI-enabled efficiencies do not compromise the core principles of assurance, transparency, and responsibility that underpin the auditing profession.

This study addresses two research questions:

- Can large language models function as "co-auditors" within existing audit standards while maintaining evidence traceability and professional accountability?
- What technical and governance mechanisms are necessary to ensure LLM outputs meet audit documentation requirements and regulatory compliance?

2. RELATED WORKS

The application of LLMs to evidence-seeking and document composition has crystallized around a few core pillars: retrieval-augmented generation (RAG), document and passage ranking, and mechanisms for citation and attribution. Alongside these, complementary strands of research address hallucination mitigation, security vulnerabilities, and structure-constrained generation. At a conceptual level, RAG integrates generative models with non-parametric memory, enabling claims to be substantiated by external sources rather than relying solely on the model’s internal priors (Lewis et al., 2021). Dense retrievers (such as DPR [Dense Passage Retrieval]) and state-of-the-art rankers substantially outperform traditional sparse methods in terms of recall and precision, which is particularly vital in domains characterized by long-tail or specialized content (Karpukhin et al., 2020; Nogueira & Cho, 2019). On the generative side, architectures like Fusion-in-Decoder (FiD) aggregate multiple passages to support long-form, evidence-grounded writing and remain foundational in this subdomain (Izacard & Grave, 2021).

2.1 Retrieval and Ranking

The contemporary retrieval stack encompasses sparse retrieval (e.g., BM25), dense bi-encoder models, and late-interaction approaches such as ColBERT and ColBERTv2. These are often combined into hybrid pipelines that balance wide coverage (via sparse methods) with semantic depth (via dense models), typically followed by cross-encoder re-ranking for precision refinement (Khattab & Zaharia, 2020; Nogueira & Cho, 2019; Santhanam et al., 2022). Recent developments have emphasized optimizing FiD and listwise re-ranking for throughput, a critical consideration in audit-like scenarios where a large number of candidate passages must be evaluated (Formal et al., 2021). Empirical findings consistently affirm that this architecture constitutes the state of the art in knowledge-intensive generation contexts demanding verifiable evidence.

2.2 Hallucination and Factuality

Extensive literature documents the tendency of LLMs to hallucinate both facts and rationales, with unique error patterns emerging particularly in open-ended and long-form settings (Huang et al., 2023; Ji et al., 2023). While retrieval reduces hallucination frequency, it is insufficient on its own. Techniques like Self-Reflective Retrieval-Augmented Generation (Self-RAG) train models to dynamically decide when retrieval is necessary and to evaluate their own outputs critically. Post-generation tools such as Retrieve, Attribute, Rewrite, Repeat (RARR) go further by identifying and amending unsupported segments in a draft through targeted research (Asai et al., 2023; Gao et al., 2023a). Dialogue-based RAG frameworks have shown marked improvements in reducing hallucination relative to purely parametric models (Li et al., 2021).

2.3 Attribution and Citation-Aware Writing

In domains requiring high assurance, such as auditing or legal reasoning, not only must generated outputs be factually correct, but they must also be attributable to cited sources. Emerging benchmarks and evaluation metrics now explicitly assess these dimensions: AIS (Attributable to Identified Sources) evaluates whether outputs are supported by cited evidence; Automatic LLM Citation Evaluation (ALCE) measures citation correctness in long-form responses; and Factual Accuracy Score (FActScore) decomposes text into discrete factual claims to quantify precision (Gao et al., 2023b; Min et al., 2023; Rashkin et al., 2023). Recent work has explored fine-grained citation alignment and large-scale multi-source attribution, highlighting that the problem of “correct answer, wrong citation” remains prevalent and warrants direct evaluation (Gao et al., 2023b; Min et al., 2023).

2.4 Security and Provenance Risks in Evidence-Seeking

The integration of retrieval functions introduces novel security vulnerabilities. When LLMs follow or fetch links, they are susceptible to prompt injection, data leakage, and contamination of retrieved evidence. These threats cataloged within the Open Web Application Security Project (OWASP) Top-10 risks for LLM systems and demonstrated through indirect prompt-injection attacks (Greshake et al., 2023; OWASP Foundation, 2025). Additionally, extraction studies reveal that LLMs can memorize and inadvertently regurgitate sensitive training data, complicating claims around provenance and retrieval when such content is mislabeled as externally sourced (Carlini et al., 2021). These findings underscore the need for architectural safeguards such as retrieval allow-lists, sanitization of retrieved content, enforced citation practices, and human-in-the-loop review mechanisms.

2.5 Structure- and Policy-Constrained Generation

In audit-style documentation, outputs must adhere to predefined formats (e.g., structured JSON with fields for evidence, source, and timestamp) and avoid off-policy generation. Techniques like grammar-constrained decoding and programmatic prompting especially with tools like Language Model Query Language (LMQL). It enforces structural and policy constraints during inference by combining grammar-aware decoding with multi-step control flows (Beurer-Kellner et al., 2023; Geng et al., 2023). These methods enhance the auditability and reliability of generated artifacts.

2.6 Domain-Specific Retrieval for Law and Policy.

Because audit-relevant evidence frequently resides within legal, regulatory, or policy corpora, domain-adapted benchmarks are essential. LegalBench evaluates legal reasoning, LexGLUE covers legal natural language understanding, and LegalBench-RAG benchmarks the retrieval step in legal contexts. Across these datasets, results consistently show that accurate snippet retrieval and robust ranking are foundational for faithful long-form answers with correct citations, an exacting requirement in audit settings where traceability and verifiability are paramount (Chalkidis et al., 2022; Guha et al., 2023; Pipitone & Alami, 2024).

2.7 Synthesis.

The prevailing technical consensus is pragmatic and converges on four requirements for trustworthy LLM-assisted auditing workflows:

(i) robust, often hybrid retrieval and re-ranking pipelines.

(ii) constrained, citation-aware generation.

(iii) explicit evaluation of factuality and attribution.

(iv) security safeguards embedded in retrieval workflows.

Collectively, these capabilities align with core assurance values like traceability of claims, rigor in documentation, and preserved human accountability (Lewis et al., 2021; Min et al., 2023; Nogueira & Cho, 2019; OWASP Foundation, 2025). RAG has been established as a principled approach for grounding sequence models in external corpora, thereby improving factuality and controllability on knowledge-intensive tasks (Lewis et al., 2021). Sparse lexical retrieval based on the probabilistic relevance framework (e.g., BM25) remains a strong and interpretable first-stage baseline and continues to play a central role in hybrid systems (Robertson & Zaragoza, 2009). Dense Passage Retrieval (DPR) demonstrated that dual-encoder embeddings trained over question–passage pairs can surpass strong BM25 pipelines on open-domain QA recall, reshaping first-stage retrieval practice (Karpukhin et al., 2020). Late-interaction multi-vector retrieval (ColBERT) preserves token-level matching while maintaining scalable indexing and has been adopted for high-recall passage search in large corpora (Khattab & Zaharia, 2020). Learned sparse expansion methods (SPLADE) recover inverted-index efficiency and interpretability with competitive effectiveness, making them attractive in compliance-sensitive domains (Formal et al., 2021). Cross-encoder re-rankers based on BERT further improve precision by jointly scoring query–document pairs and remain standard in RAG pipelines for ordering candidate contexts (Nogueira & Cho, 2019).

3. METHODOLOGY AND CONCEPTUAL FRAMEWORK

This study employs a narrative literature review to explore the integration of LLMs into audit workflows. The review synthesizes three domains: professional audit standards and regulatory frameworks (ISA 230, PCAOB, COSO, NIST AI Risk Management Framework, ISO/IEC AI standards, EU AI Act); natural language processing research on retrieval-augmented generation, factuality, and attribution; and AI governance literature addressing accountability and risk management.

Literature was identified through academic databases (ACM Digital Library, arXiv, IEEE Xplore) using keywords: "large language models," "retrieval-augmented generation," "audit evidence," and "AI governance." Standards repositories (IAASB, PCAOB, NIST, ISO) and regulatory sources were consulted for authoritative guidance.

Inclusion criteria:

1- studies on LLM retrieval, generation, and factuality (2020–2025);

2- audit standards applicable to AI deployment;

3- technical work on attribution and security.

Exclusion criteria:

1- purely theoretical AI work without audit context;

2- non-English sources.

The synthesis prioritized foundational architectures (RAG, DPR), attribution benchmarks (ALCE, FActScore), and regulatory frameworks (NIST, AI RMF, EU AI Act).

Based on this review, the study develops a conceptual "co-auditor" as a design proposal that illustrates how LLM capabilities can be embedded within audit-compliant controls. This framework serves as a blueprint for future implementation and empirical validation.

The design prioritizes audit documentation and evidence requirements: work must be attributable, reviewable, and reproducible by a competent auditor (IAASB, 2010; PCAOB, 2025). The system therefore emphasizes traceability, provenance, security, and governance over raw model power.

"Co-auditor" consists of five cooperating planes, namely Governance and Risk Plane, Evidence and Provenance Plane, Retrieval and Analysis Plane, Constrained Generation and Attribution Plane, Safety and Security Plane.

Governance and Risk Plane establish the oversight structure necessary for accountable AI deployment in audit contexts. It ensures that all AI-supported audit activities operate within defined policies, documented controls, and assigned human accountability. The plane provides role-based access control, change management procedures, and risk registers tailored to AI tools used in audit engagements, ensuring that LLM assistance remains traceable, reviewable, and subject to professional oversight. The governance design follows a risk-based cycle (Govern–Map–Measure–Manage) aligned with the NIST AI Risk Management Framework (AI RMF 1.0) and its Generative AI Profile, while satisfying the requirements of ISO/IEC 42001 AI Management System standards and ISO/IEC 23894 guidance on AI risk management(ISO, 2023a, 2023b; NIST, 2024).

Evidence and Provenance Plane maintain a comprehensive audit trail of all LLM-supported activities, enabling reviewers to verify what was done, by whom, when, and based on what sources. It records content hashes, timestamps, user identities, model versions, prompts, retrieval sets, re-ranking scores, and generated excerpts in an append-only evidence ledger. Such traceability is essential for satisfying audit documentation standards that require work to be attributable and reproducible by a competent auditor. Provenance metadata is represented using the The World Wide Web Consortium (W3C) Provenance (PROV) data model, while file attachments carry cryptographic content credentials via the Coalition for Content Provenance and Authenticity (C2PA) specification. The ledger employs Merkle-tree techniques based on certificate-transparency log designs to ensure tamper-evidence and long-term verifiability (C2PA, 2024; W3C, 2013).

Retrieval and Analysis Plane implement a policy-bounded retrieval-augmented generation stack that queries only approved corpora such as client-provided documents, organizational policies, contracts, and regulatory filings within defined scope and time boundaries. By restricting retrieval to authorized sources, the plane ensures that generated outputs are grounded in verifiable evidence rather than the model's internal parameters. The retrieval pipeline combines dense and sparse retrieval methods for high recall, applies cross-encoder re-ranking for precision, and uses multi-document reasoning to synthesize evidence from multiple sources. Technical components include dense passage retrieval encoders, late-interaction models for token-level matching, fusion-in-decoder architectures for aggregating multiple passages, and reflective generation techniques that assess retrieval necessity and output quality (Asai et al., 2023; Izacard & Grave, 2021; Karpukhin et al., 2020; Lewis et al., 2021; Nogueira & Cho, 2019; Santhanam et al., 2022).

Constrained Generation and Attribution Plane enforces structured, citation-backed outputs to prevent unsupported claims and ensure that every material assertion is traceable to a cited source. Outputs are constrained to predefined schemas (e.g., JSON structures specifying purpose, evidence, conclusions, and reviewer comments), and generation follows a "no source, no claim" policy requiring line-level citations for all substantive content. Schema enforcement is implemented through grammar-constrained decoding and policy-aware prompt design, techniques shown to improve compliance with minimal performance overhead. This structured approach reduces free-form drift, enhances reviewability, and aligns with audit requirements for documented bases of conclusions (Beurer-Kellner et al., 2023; Geng et al., 2023).

Safety and Security Plane protect against adversarial inputs, data leakage, and model vulnerabilities that could compromise audit integrity. It implements defenses against prompt injection attacks, insecure output handling, and training-data poisoning through strict allow-listing of retrieval sources and tools, input sanitization, egress filtering, and reduction of personally identifiable information in processing pipelines. Regular red-teaming exercises test resilience to known attack vectors. The design is informed by the OWASP Top-10 risks for LLM applications and NIST AI 600-1 guidance on managing generative AI risks, while acknowledging residual risks from memorization and adversarial poisoning that require ongoing monitoring (Carlini et al., 2021; NIST, 2024; OWASP Foundation, 2025; Xu et al., 2023).

3.1 Retrieval Policies and Controls

Policy-bounded retrieval enforces scope (approved repositories), time-boxing (e.g., period under audit), and versioning (e.g., file hashes). The pipeline uses DPR or SPLADE/ColBERT for recall, cross-encoder BERT re-rankers for precision, and FiD/Self-RAG for multi-document reasoning. Each claim is trace-linked to snippets (doc ID, page/line) captured in the ledger (Asai et al., 2023; Izacard & Grave, 2021; Karpukhin et al., 2020; Nogueira & Cho, 2019; Santhanam et al., 2022).

3.2 Constrained Review

Structure-first generation: the system generates a structured workpaper object (e.g., purpose, processes followed, evidence referenced, conclusions made, comment of reviewer) verified and checked in-flight by grammar/JSON-schema constraints. This reduces free-form drift and requires citation density. Experiments show that grammar- or automata-constrained decoding materially improves format adherence and efficiency for structured output; LMQL describes programmable constraints on top of LLM decoding (Beurer-Kellner et al., 2023; Geng et al., 2023; Jain et al., 2023).

3.3 Evidence Ledger and Provenance

Each transformation (ingest → retrieval set → ranked list → excerpt → generated text → human edits) is recorded using cryptographic digests. Provenance complies with W3C PROV model (agents, activities, entities) and carries content credentials of ancillary artifacts using Coalition for Content Provenance and Authenticity C2PA; the append-only log can itself be implemented using transparency-log strategies (Merkle trees) so that at a later time it can be examined and conflict resolved (Coalition for Content Provenance and Authenticity [C2PA], 2024; The World Wide Web Consortium (W3C), 2013).

3.4. Security

It assumes untrusted web inputs and adversarial corpora. It:

- Prevents prompt-injection/indirect injection via content isolation, allow-listed tools, and de-tainting of retrieved text before execution (Greshake et al., 2023; OWASP Foundation, 2025).
- Prevents memorization leakage by disabling training on client data and scanning outputs for substrings of high entropy or Personally Identifiable Information (PII); risks of memorization are logged empirically (Carlini et al., 2021).
- Hardens against poisoning/backdoors when fine-tuning or training data using dataset governance, differential sampling, and red-team seeding grounded on recent work on poisoning practicability and tenacity (Carlini et al., 2023; Xu et al., 2023).
- Aligns with NIST AI 600-1 controls related to Generative AI risk management such as supply-chain, content provenance, incident taxonomy (NIST, 2024)

3.5. Human Responsibility and Control Charting

Operationalization is equivalent to COSO (control environment, risk assessment, control activities, information & communication, monitoring) and to audit standards (IAASB, 2010):

- Control environment: policies, role definition, segregation of duties regarding AI tools.
- Control activities: pre-issuance review gates; required evidence-to-assertion mappings; exception workflows if the evidence is not adequate.
- Monitoring: data/model cards, measures of drift and quality (e.g., attributable-to-source rates), red-team results monitoring.
- Documentation: Immutable logs enable a mature auditor to understand what was undertaken, by whom, when, and why (IAASB, 2010; COSO, 2013).

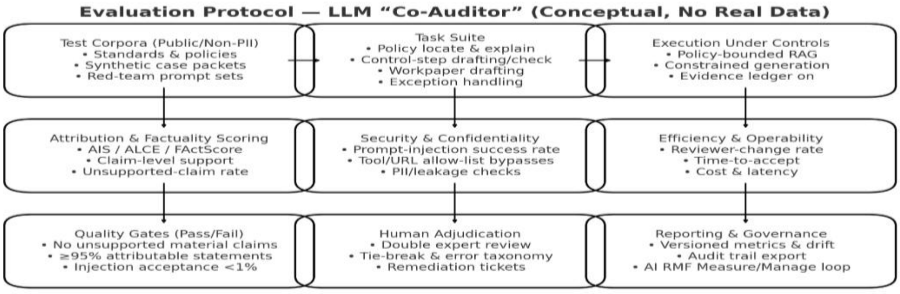
Compliance overlays are ISO/IEC 42001 (requirements of AI management system) and EU AI Act requirements (logging, transparency, risk management, post-market surveillance) for high-risk or general-purpose deployments. (European Union, 2024; ISO, 2023a, 2023b).

3.6 Evaluation & Quality Gates

Quality is measured by attribution-focused metrics and task accuracy: FActScore, ALCE, and AIS measures receive groundedness and identifiability of sources; internal Key Performance Indicators (KPIs) are “claims per citation,” “review-change rate,” and “unverifiable claim rate.” They are forwarded to quality management and offer constant improvement within NIST AI. It is co-auditor assistance, not automation: it raises evidence, drafts workpapers, and highlights gaps but never makes a professional judgment, independence, or sign-off. It must fail closed (retain conclusions withheld) when retrieved evidence is incomplete, contradictory, or outside of policy scope. This stance is aligned with ISA 500 series of changes development and general standard-setter recommendations about technology use in audit. The target environment presumes a versioned corpus of policies, procedures, audit programs, and templates managed at paragraph-level granularity. Each paragraph carries metadata such as document and paragraph identifiers, version and effective dates, section labels, and access tags that bind content to roles or attributes. Paragraph-sized chunks (approximately 300–800 tokens with overlap) are favored because they are specific enough for retrieval yet long enough to preserve context, in line with established findings in RAG (Karpukhin et al., 2020; Lewis et al., 2021; Nogueira & Cho, 2019; Robertson & Zaragoza, 2009). Access controls and version discipline follow enterprise governance frameworks (ISO, 2023b; NIST, 2024; COSO, 2013).

Three task families are in scope: (a) policy locate-and-explain, identifying and explaining the applicable rule, including scope, exceptions, and effective dates; (b) control-step verification, checking that planned or executed test steps align with the authoritative audit program; and (c) working-paper drafting, producing structured documentation

Figure 1: Evaluation Protocol – LLM “Co-Auditor”



Source: (Authors’s own)

The figure describes end-to-end protocol of a “co-auditor” judgment absent client data: public non-PII corpora and synthesized case packs are fed into a task suite (policy locate-and-explain, control-step drafting/checking, workpaper drafting and exception handling). Execution is under strict controls with policy-constrained RAG, constrained "no source, no claim" prod which is produced and an always-running evidence ledger that captures documents, hashes and timestamps and prompts and versions of models, is such that all output is attributable and reproducible. Performance is then rated on three axes:

(v) attribution/factuality (AIS/ALCE/FActScore, citation accuracy/recall, unsupported-claim rate),
(vi) security/confidentiality (acceptance of prompt injection, allow-list evasions, PII/leakage)
(vii) operability (change-in-reviewer rate and time-to-accept and latency/cost).

Quality gates apply pass/fail thresholds (e.g., no unsupported material claims; ≥95% attributable statements; injection acceptance <1%) and fail-closed responses absent evidence. Double expert judgment makes final decisions on edge cases and Error Type Classification (attribution and retrieval and reasoning and policy and security) and raises remediation tickets. And versioned metrics and drift reports are fed into governance and continuous improvement pursuant to the NIST AI RMF "Measure/Manage" loop such that the system is simultaneously auditable and fit-for-use over time.

4. LIMITATIONS AND FURTHER RESARCH

This study is purely conceptual and lacks empirical validation. The proposed architecture has not been implemented or tested in real audit engagements. The narrative literature review may be subject to selection bias despite specified inclusion criteria. The evaluation metrics (AIS, ALCE, FActScore) have not been validated in audit contexts. Organizational, economic, and behavioral dimensions of LLM adoption are beyond the scope of this work. Finally, rapid LLM evolution means this framework reflects knowledge as of early 2025 and requires ongoing revision.

Future research should pursue empirical validation of the proposed co-auditor architecture through pilot implementations in real audit settings with a case study. Investigating cost-benefit trade-offs and productivity gains across different audit task types and developing audit-specific benchmarks and measures with metrics should be researched.

5. CONCLUSION

This paper addresses whether large language models can function as "co-auditors" within existing audit standards and accountability frameworks.

Regarding the first research question on whether LLMs can function as co-auditors within existing audit standards, the literature confirms that this is feasible only when LLMs are embedded within systems ensuring evidence traceability, governance, and human accountability. Technical capabilities alone are insufficient; they must align with audit documentation standards (ISA 230, PCAOB AS 1215) and professional responsibility frameworks.

Regarding the second research question on necessary technical and governance mechanisms, the study identifies the following requirements: hybrid retrieval-augmented generation with policy-bounded sources; citation-constrained generation enforcing "no source, no claim" policies, cryptographically verifiable provenance (W3C PROV, C2PA), security safeguards addressing OWASP LLM risks; and governance aligned with NIST AI RMF and ISO/IEC 42001.

The proposed "co-auditor" architecture is a conceptual design framework, not an implemented system. It defines conditions under which LLMs can augment but not replace professional judgment

This conforming renders human auditors the final assessors of assurance with indelible logs and formatted output permitting the necessary transparency for oversight and post-review.

Overall, the paper confirms that the route to auditable and trustworthy AI is neither by uncontrolled use nor by uncontrolled deployment but rather by stringent incorporation of technical controls, governance layers, and human judgment. On these terms, LLMs can take the profession forward: scaling up evidence coverage, standardizing workpaper quality, and facilitating successive improvement. All while holding fast to the core principles of transparency, accountability, and independence that underlie the profession of audit.

References

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2310.11511

Beurer-Kellner, L., Fischer, M., & Vechev, M. (2023). Prompting Is Programming: A Query Language for Large Language Models. Proceedings of the ACM on Programming Languages, 7 (PLDI), 1946–1969. https://doi.org/10.1145/3591300

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., & Tramèr, F. (2023). Poisoning Web-Scale Training Datasets is Practical (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2302.10149

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting Training Data from Large Language Models. 30th USENIX Security Symposium (USENIX Security 21), 2633–2650. Retrieved December 10, 2025, from https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D., & Aletras, N. (2022). LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 4310–4330. https://doi.org/10.18653/v1/2022.acl-long.297

Coalition for Content Provenance and Authenticity [C2PA]. (2024). Content Credentials: C2PA Technical Specification. Retrieved December15,2025,from https://spec.c2pa.org/specifications/specifications/2.1/specs/C2PA_Specification.html

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules (Artificial Intelligence Act). Retrieved November 18, 2025, from http://data.europa.eu/eli/reg/2024/1689/oj

Formal, T., Lassance, C., Piwowarski, B., & Clinchant, S. (2021). SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2109.10086

Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V., Lao, N., Lee, H., Juan, D.-C., & Guu, K. (2023a). RARR: Researching and Revising What Language Models Say, Using Language Models. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 16477–16508. https://doi.org/10.18653/v1/2023.acl-long.910

Gao, T., Yen, H., Yu, J., & Chen, D. (2023b). Enabling Large Language Models to Generate Text with Citations. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 6465–6488. https://doi.org/10.18653/v1/2023.emnlp-main.398

Geng, S., Josifoski, M., Peyrard, M., & West, R. (2023). Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 10932–10952. https://doi.org/10.18653/v1/2023.emnlp-main.674

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, 79–90. https://doi.org/10.1145/3605764.3623985

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., … Li, Z. (2023). LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2308.11462

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. https://doi.org/10.48550/ARXIV.2311.05232

Institute of Internal Auditors [IIA]. (2025). AI in Internal Audit | Knowledge center to empower internal audit teams with the latest info and practical guidance in AI. Retrieved November 7, 2025, from https://www.theiia.org/en/resources/knowledge-centers/artificial-intelligence/

International Auditing and Assurance Standards Board [IAASB]. (2010). IFAC Releases 2010 Handbooks Containing All IAASB Pronouncements and the Code of Ethics for Professional Accountants. Retrieved November 1, 2025, from https://www.iaasb.org/news-events/2010-04/ifac-releases-2010-handbooks-containing-all-iaasb-pronouncements-and-code-ethics-professional

International Organization for Standardization [ISO]. (2023a). ISO/IEC 23894:2023. Retrieved November 1, 2025, from https://www.iso.org/standard/77304.html

International Organization for Standardization [ISO]. (2023b). ISO/IEC 42001:2023. Retrieved November 1, 2025, from https://www.iso.org/standard/42001

Izacard, G., & Grave, E. (2021). Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 874–880. https://doi.org/10.18653/v1/2021.eacl-main.74

Jain, N., Schwarzschild, A., Wen, Y., Somepalli, G., Kirchenbauer, J., Chiang, P. Y., & Goldstein, T. (2023). Baseline defenses for adversarial attacks against aligned language models. arXiv Preprint arXiv:2309.00614. https://doi.org/10.48550/arXiv.2309.00614

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 39–48. https://doi.org/10.1145/3397271.3401075

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (No. arXiv:2005.11401). arXiv. https://doi.org/10.48550/arXiv.2005.11401

Li, Z., Qu, L., & Haffari, G. (2021). Total Recall: A Customized Continual Learning Method for Neural Semantic Parsers. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 3816–3831. https://doi.org/10.18653/v1/2021.emnlp-main.310

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation (No. arXiv:2305.14251). arXiv. https://doi.org/10.48550/arXiv.2305.14251

National Institute of Standards and Technology [NIST]. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (Nos. 600–1). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.AI.600-1

Nogueira, R., & Cho, K. (2019). Passage Re-ranking with BERT (Version 5). arXiv. https://doi.org/10.48550/ARXIV.1901.04085

OWASP Foundation. (2025). OWASP Top 10 for Large Language Model Applications. OWASP Top 10 for Large Language Model Applications. Retrieved November 1, 2025, from https://owasp.org/www-project-top-10-for-large-language-model-applications/

Pipitone, N., & Alami, G. H. (2024). LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2408.10343

Public Company Accounting Oversight Board [PCAOB]. (2024). Generative-AI-Spotlight. Retrieved November 1, 2025, from https://pcaobus.org/documents/generative-ai-spotlight.pdf

Public Company Accounting Oversight Board [PCAOB]. (2025). AS 1215: Audit Documentation. Retrieved November 1, 2025, from https://pcaobus.org/oversight/standards/auditing-standards/details/AS1215

Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G. S., Turc, I., & Reitter, D. (2023). Measuring Attribution in Natural Language Generation Models. Computational Linguistics, 49(4), 777–840. https://doi.org/10.1162/coli_a_00486

Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends® in Information Retrieval, 3(4), 333–389. https://doi.org/10.1561/1500000019

Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., & Zaharia, M. (2022). ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3715–3734. https://doi.org/10.18653/v1/2022.naacl-main.272

The Committee of Sponsoring Organizations of the Treadway Commission [COSO]. (2013). Internal Control. Retrieved November 1, 2025, from https://www.coso.org/internal-control

The World Wide Web Consortium [W3C]. (2013). PROV-DM: The PROV Data Model. PROV-DM: The PROV Data Model. Retrieved November 1, 2025, from https://www.w3.org/TR/prov-dm/

Xu, J., Ma, M. D., Wang, F., Xiao, C., & Chen, M. (2023). Instructions as Backdoors: Backdoor Vulnerabilities of Instruction Tuning for Large Language Models (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2305.14710