Logistic and Operation Management Research, Vol. 5 No.1 (2026) https://doi.org/10.31098/lomr.v5i1.4258

Check for updates

Research Paper

Evidence-Grounded Hybrid Framework for ISO/IEC 17025 Laboratory Operations Audits: A Neuro-Symbolic Decision-Support Approach

Sadam Al Rasyid[2], Aldi Wiranata[1], Korediantо Usman[2*], Suryo Adhi Wibowo[2], Gunadi Dwi Hantoro[1], Marshaniswah Syamsul[2], Yudha Indah Prihatini[1]

[1] PT Telkom Indonesia Tbk, Bandung, Indonesia

[2] Telkom University, Bandung, Indonesia

Received: December 02, 2025	Revised : April 27, 2026	Accepted : May 18, 2026	Online : May 29, 2026

Abstract

The accreditation of testing and calibration laboratories under ISO/IEC 17025:2017 requires systematic mapping of operational evidence to regulatory clauses—a labour-intensive process dependent on subjective auditor interpretation. As critical quality-infrastructure nodes in industrial supply chains, accredited laboratories directly influence product safety, environmental monitoring, and service reliability. Although Retrieval-Augmented Generation (RAG) technologies offer partial automation, conventional RAG systems produce extrinsic hallucinations at rates of 17–33 %, posing unacceptable liability for compliance governance. This study proposes an Evidence-Grounded Hybrid Intelligence framework that enforces verifiable compliance mapping through a neuro-symbolic architecture. Adopting a Design Science Research methodology, the framework integrates dual-retrieval (TF-IDF + SBERT), a verifiable RAG pipeline (Qwen3-32B, temperature 0.1) with mandatory evidence-span extraction and character-offset verification, a deterministic rule-based validation engine, and RDF/Turtle provenance export. Benchmarked against a conventional RAG baseline across 847 query-clause pairs from 50 purposively sampled Indonesian audit reports, the framework eliminated propagated hallucinations (14 % → 0 %; McNemar $\chi^2 = 118.6$, $p < .001$), raised evidence-grounding precision to 0.94 (F1 = 0.91), reduced reviewer correction rates by 37 %, and achieved 100 % detection of predefined non-compliance patterns. The framework contributes to laboratory operations management by delivering a scalable compliance decision- support architecture that reduces audit-cycle effort, standardises verification processes, mitigates operational risk, and establishes machine-readable provenance infrastructure for longitudinal quality assurance.

Keywords: *Hybrid Intelligence, Retrieval-Augmented Generation, Verifiable AI, Compliance Auditing, Quality Management Systems, Laboratory Operations Management, Decision-Support Systems, Knowledge Graph, ISO/IEC 17025*

INTRODUCTION

The accreditation of testing and calibration laboratories under ISO/IEC 17025:2017 constitutes a fundamental pillar of global quality infrastructure, ensuring the validity and international recognition of measurement results (ISO/IEC, 2017; UNIDO, 2020). Beyond a mere regulatory formality, these accredited laboratories function as essential quality-infrastructure nodes within complex industrial supply chains. Their operational output directly governs product safety, environmental monitoring integrity, and downstream service reliability across diverse manufacturing and service sectors (van der Wiele et al., 2011). Consequently, any deficiency or delay in the accreditation process inevitably propagates throughout the supply chain, precipitating disrupted production schedules and compounding operational risks for all dependent stakeholders (Flynn et al., 1994; Zu et al., 2008).

From an operations management perspective, laboratory compliance auditing presents a

Corresponding author's email: korediantousman@telkomuniversity.ac.id

formidable process-management challenge that is highly amenable to standardization and decision-support enhancement. The conformity assessment process is notoriously labor- intensive, compelling auditors to systematically map thousands of pages of operational evidence against open-textured regulatory clauses. The operations and quality management literature increasingly emphasizes that optimizing audit cycle times, enforcing process standardization, and establishing robust quality assurance systems are critical for effective operational governance (Slack et al., 2016; Power, 2002). Addressing these operational friction points requires scalable decision-support systems that enhance procedural efficiency and mitigate operational risk without supplanting essential human professional judgment (Arnott & Pervan, 2005; Hevner et al., 2004).

Despite this operational imperative, the current landscape of Automated Compliance Checking (ACC) remains fundamentally bifurcated. Rigid, rule-based systems guarantee absolute auditability but critically falter when navigating the semantic nuances of regulatory natural language. Conversely, while recent advancements in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) exhibit remarkable proficiency in processing unstructured text, standard RAG architectures suffer from a critical deficit in deterministic reliability. These systems frequently produce extrinsic hallucinations—fabricating citations or inventing measurement values at rates of 17–33% (Magesh et al., 2025; Dahl et al., 2024)—which poses an unacceptable liability for compliance governance. This dichotomy exposes a profound operations management practice gap: the critical absence of scalable, high-assurance audit decision-support systems capable of combining semantic adaptability with uncompromising deterministic verification for regulated operations (Garcez & Lamb, 2020; Pakina & Pujari, 2024).

The Indonesian quality infrastructure ecosystem serves as a highly representative crucible for addressing this gap. As a rapidly developing economy with aggressively expanding industrial and testing sectors, Indonesia confronts formidable hurdles in scaling its quality infrastructure to satisfy the rigorous imperatives of international trade and regulatory enforcement (World Bank, 2020). The national accreditation body, *Komite Akreditasi Nasional* (KAN), currently presides over a burgeoning portfolio of laboratories, thereby precipitating an urgent operational mandate for enhanced audit efficiency. The intricate confluence of entrenched legacy documentation practices, heterogeneous laboratory domains, and acute resource constraints positions Indonesia as an exemplary proving ground for evaluating the efficacy of AI-augmented compliance systems.

To systematically resolve these intersecting operational challenges, this study introduces a novel Evidence-Grounded Hybrid Framework. This architecture deliberately enforces "architectural verifiability"—a foundational design philosophy that guarantees system reliability through rigorous structural constraints, effectively mitigating the inherent unpredictability of probabilistic model behavior (Garcez & Lamb, 2020).

Research Objectives and Questions

This research is propelled by five sequential objectives: (O1) to engineer a verifiable RAG pipeline that mandates verbatim evidence extraction anchored by precise character offsets; (O2) to operationalize a robust dual-retrieval (TF-IDF + SBERT) architecture; (O3) to architect a deterministic, rule-based validation engine; (O4) to establish a comprehensive knowledge-graph provenance infrastructure utilizing RDF/Turtle export; and (O5) to empirically benchmark the framework's performance against conventional baselines utilizing authentic audit documentation.

These objectives are strategically operationalized through the following decision-support and operations-focused research questions:

RQ1: How does the proposed evidence-grounded hybrid framework mitigate hallucination rates and diminish compliance risk when benchmarked against a conventional RAG baseline?

RQ2: To what extent does the integration of a dual-retrieval architecture optimize evidence recall

and fortify grounding accuracy?
RQ3: What quantifiable operational efficiency gains, specifically regarding audit effort reduction and cycle-time acceleration, does this framework deliver for laboratory accreditation bodies?

Contributions

This study delivers three distinct and impactful contributions to the existing body of literature and industrial practice:

1. A practical contribution to audit operations and compliance governance by providing a highly deployable decision-support architecture that demonstrably truncates audit-cycle effort, rigorously standardizes the compliance verification process, and systematically mitigates operational risk.
2. A methodological contribution achieved through the novel integration of verifiable RAG with deterministic neuro-symbolic validation, offering a replicable and robust design pattern for deploying high-assurance AI systems within strictly regulated operational environments.
3. A theoretical contribution demonstrating that architectural verifiability constraints, rather than isolated probabilistic model enhancements, are fundamentally requisite to ensure hybrid decision-support reliability in operations. This empirical insight formally extends the hybrid intelligence paradigm into the complex spheres of quality management and operational governance (Garcez & Lamb, 2020; Slack et al., 2016).

LITERATURE REVIEW

Automated Compliance Checking in Regulatory Domains

Automated Compliance Checking has evolved over two decades from hard-coded logic systems in the Architecture, Engineering, and Construction sector to applications in financial regulation and quality management (Pakina & Pujari, 2024; Chasandras, n.d.). Rule-based approaches demonstrate effectiveness for quantitative checks but achieve less than 45 % accuracy on natural-language requirements, whereas transformer-based models can process regulatory semantics but introduce interpretability deficits and confident yet factually unsupported conclusions (Pakina & Pujari, 2024; Cheng et al., 2025; Huang et al., 2025).

Synthesis.

Scholars agree that neither purely symbolic nor purely neural approaches suffice for standards combining quantitative and qualitative requirements. What remains unresolved is whether hybrid architectures can achieve both semantic flexibility and deterministic rigour without unacceptable performance trade-offs—a gap this study addresses through mandatory evidence grounding within a neuro-symbolic architecture.

Compliance Auditing as an Operations Management Problem

The operations management literature frames compliance auditing as a process-management and quality-assurance challenge amenable to standardisation and decision-support enhancement (Slack et al., 2016; van der Wiele et al., 2011). Audit operations involve systematic flows—document collection, evidence mapping, compliance determination, and corrective-action tracking—comparable to manufacturing process controls (Zu et al., 2008; Power, 2002). Laboratories function as critical quality-infrastructure nodes whose accreditation status directly affects downstream product certifications and safety evaluations; audit inefficiencies therefore propagate through supply chains as operational risks (Flynn et al., 1994). Decision-support systems in operations have evolved from rule- based tools to intelligent architectures incorporating machine learning and human-in-the- loop design (Arnott & Pervan, 2005; Hevner et al., 2004).

Synthesis

Operations management scholars agree that compliance auditing is amenable to decision-support automation, but the contested dimension is the degree to which AI systems can be trusted for high-stakes regulatory decisions. The unresolved gap is the absence of empirically validated decision-support frameworks that maintain verifiability while delivering measurable efficiency gains in compliance audit workflows.

RAG and the Hallucination Problem

RAG was designed to ground LLM outputs in external documentary evidence, reducing reliance on potentially incorrect parametric knowledge (Oche et al., 2025). Huang et al. (2025) distinguish intrinsic hallucinations (contradicting retrieved sources) from extrinsic hallucinations (fabricating facts absent from any source), finding that extrinsic hallucination remains prevalent even in RAG-augmented systems. Legal-domain evaluations by Magesh et al. (2025) documented 17–33% hallucination rates involving fabricated citations and invented reference errors constituting professional malpractice (Dahl et al., 2024). For ISO/IEC 17025 auditing, these findings imply that standard RAG may fabricate calibration certificates, invent uncertainty values, or misattribute measurement traceability (Rawte et al., 2023; Tonmoy et al., 2024).

Synthesis

While RAG reduces hallucination relative to standalone LLMs, it does not eliminate the problem. The unresolved challenge is how to architecturally guarantee output verifiability rather than merely reduce hallucination probability. This study addresses this gap through mandatory evidence grounding with character offsets and deterministic validation.

Neuro-Symbolic AI and Hybrid Intelligence

Garcez and Lamb (2020) characterised neuro-symbolic integration as the "third wave" of AI, combining neural pattern recognition with symbolic logical rigour. Belle and Papantonis (2021) argued that explanation capabilities in high-stakes domains must be architecturally integral, not post-hoc additions. The Hybrid Intelligence paradigm positions AI as a decision-support agent within human-supervised workflows, acknowledging that regulatory compliance ultimately requires professional judgement while leveraging AI for evidence retrieval and consistency checking (Pakina & Pujari, 2024; Garcez & Lamb, 2020).

Synthesis

High-stakes domains require architectures combining neural flexibility with symbolic rigour and human oversight. What remains unresolved is implementing this principle in compliance auditing with measurable verifiability guarantees, which this study operationalises through verifiable RAG with deterministic validation.

Knowledge Graphs and Provenance in Audit Contexts

Knowledge graphs encode entities, relationships, and attributes in formats supporting both human interpretation and automated reasoning (Yu et al., 2025). RDF/Turtle serialisation provides standardised provenance vocabularies capturing what, when, by whom, and based on what evidence compliance determinations were made (Qi et al., 2021). GraphRAG approaches extend these concepts to multi-hop reasoning across connected entities (Tonmoy et al., 2024). For

laboratory accreditation, where compliance often depends on cross- document relationships, knowledge graphs offer infrastructure for longitudinal audit analytics and "Smart Accreditation" (Qi et al., 2021; Taylor, n.d.).

Synthesis

Knowledge graph technologies offer established provenance infrastructure, but existing implementations focus on knowledge representation rather than integration with AI-driven compliance pipelines. This study bridges the gap by embedding RDF/Turtle export directly within the automated audit workflow.

Theoretical Benchmark and Design Propositions

This study adopts hybrid intelligence and verifiability-by-design as the theoretical benchmark for decision-support systems in regulated operations, grounded in three converging principles: (1) *reliability*—deterministic, reproducible system behaviour (Slack et al., 2016; van der Wiele et al., 2011); (2) *traceability*—the ability to link every output to its specific evidence source (Qi et al., 2021; Flynn et al., 1994); and (3) *operational control*—constraining system behaviour through structural mechanisms rather than solely probabilistic performance (Garcez & Lamb, 2020; Hevner et al., 2004).

Following Design Science Research methodology (Peffers et al., 2007), the study presents three testable design propositions: *DP1*: Mandatory evidence grounding with character offsets eliminates propagated hallucinations. *DP2*: Dual-retrieval architecture improves evidence recall compared to single-path retrieval. *DP3*: Deterministic rule-based validation catches compliance errors that neural components miss, reducing overall correction rates.

RESEARCH METHOD

Research Design

This study adopts a Design Science Research (DSR) methodology (Peffers et al., 2007) with quantitative experimental evaluation, following the build-and-evaluate cycle established in information systems research (Gregor & Hevner, 2013). The research proceeds through five phases: (1) problem identification from the operations management practice gap; (2) system design based on hybrid intelligence principles; (3) artifact implementation; (4) controlled experimental comparison against a baseline; and (5) evaluation using quantitative performance metrics aligned with both information systems and operations management criteria. The quantitative approach is appropriate because audit operations require measurable improvements in accuracy, efficiency, and reliability.

System Architecture

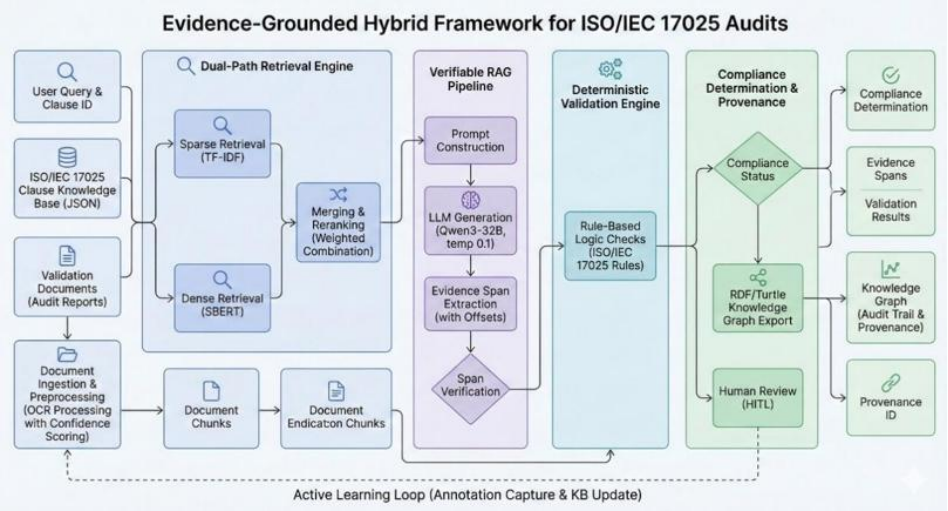

Figure 1. Evidence-Grounded Hybrid Framework for ISO/IEC 17025 Audits. Overview of the end-to-end system architecture combining dual-path retrieval, verifiable RAG reasoning, deterministic rule checks, and provenance-based compliance determination.

The framework integrates five modular components: Document Ingestion and Preprocessing, Dual-Path Retrieval Engine, Verifiable RAG Pipeline, Deterministic Validation Engine, and Knowledge Graph Exporter (Figure 1). Verifiability is enforced through structural constraints at multiple stages: mandatory checkpoints reject outputs failing evidence-grounding requirements, ensuring unverifiable claims never propagate to downstream audit processes. This "fail-safe" philosophy prioritises safety over coverage— a deliberate design trade-off aligned with risk-management principles underlying ISO/IEC 17025:2017 itself.

The knowledge base is derived from the Indonesian translation of ISO/IEC 17025:2017 and is segmented into 3,097 discrete clauses, following the standard's hierarchical structure to enable granular retrieval. Each entry contains clause_id, full Indonesian text, mandatory_evidence_types, parent_clause references, normative cross-references, and domain keywords. The validation corpus comprises 50 internal audit reports selected through stratified purposive sampling across three laboratory sectors: chemical testing (n = 18), mechanical calibration (n = 17), and electrical measurement (n = 15). This strategy ensures domain-specific coverage of documentation practices, terminology variations, and compliance patterns (Patton, 2015). Three criteria determined sample size: (1) coverage across primary sectors; (2) sufficient query-clause pairs (847 total) for statistical evaluation; and (3) representation of digital-native and legacy documentation (46 % OCR-processed).

Table 1. Dataset Summary Statistics

Component	Count	Description
ISO/IEC 17025 Clauses	3,097	Discrete requirements segmented from standard
Validation Documents	50	Purposively sampled internal audit reports

Evidence Types Defined	47	Mandatory evidence categories mapped to clauses
Average Clause Length	78 words	Mean text length per clause entry
OCR-Processed Documents	23 (46%)	Legacy scanned documents requiring OCR
Mean OCR Quality Score	0.87	Character-level accuracy
Query-Clause Pairs	847	Total evaluation instances

Dual-Path Retrieval Architecture

The framework combines sparse and dense retrieval to leverage complementary strengths. The sparse path (TF-IDF, scikit-learn; min_df = 2, max_df = 0.95, ngram_range = (1,2), sublinear_tf = True) captures exact keyword matches, clause numbers, and technical acronyms. The dense path (SBERT, paraphrase-multilingual-mpnet-base-v2, 768- dimensional embeddings) enables semantic matching without lexical overlap (Cheng et al., 2025; Oche et al., 2025). Both paths operate at top-k = 5 with a cosine similarity threshold of 0.7. Candidates are merged and re-ranked: *combined_score = α × dense_score + (1 − α) × sparse_score* (α = 0.6).

Verifiable RAG Pipeline

The generative component employs Qwen3-32B via Ollama local deployment at temperature 0.1, prioritising output determinism (Qwen et al., 2025). This model selection balances generative capability with practical deployment considerations for on-premises processing of potentially confidential audit documentation. The pipeline enforces a "No- Hallucination Constraint": every compliance claim must include (1) a verbatim text span from source documents, (2) character offsets locating the span, and (3) a source-document identifier. Outputs lacking these elements are automatically rejected with "Evidence not found" rather than propagating ungrounded claims.

Deterministic Validation Engine

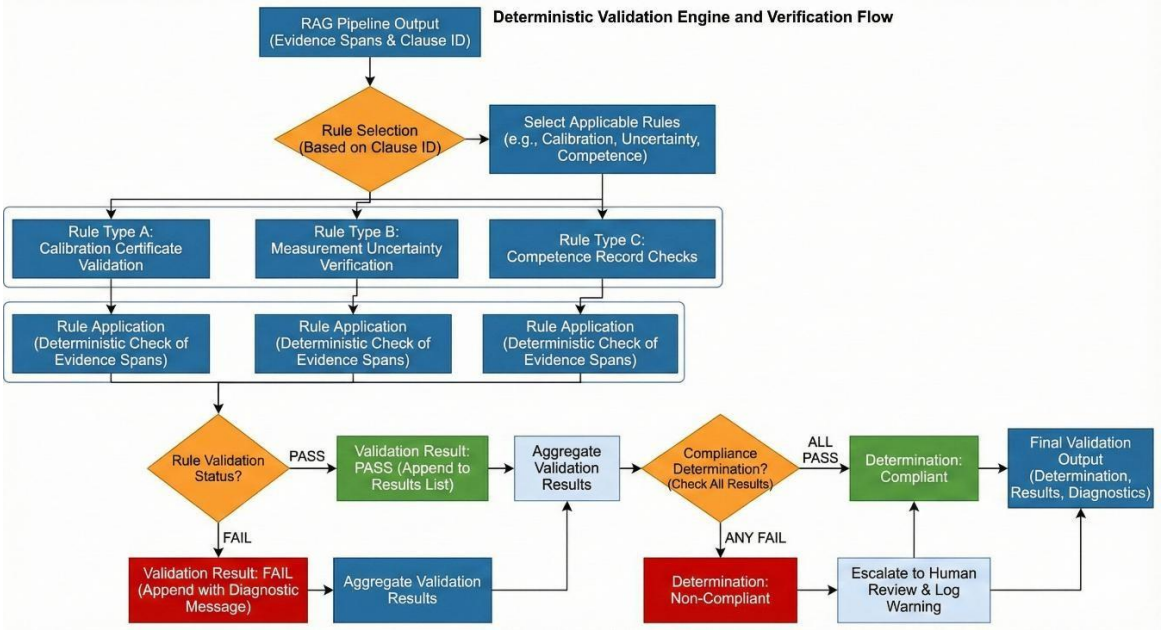

Figure 2. Deterministic Validation Engine and Verification Flow. Rule-based verification process that applies clause-specific checks to RAG outputs and generates final compliance determinations.

The symbolic reasoning layer subjects RAG outputs to deterministic logic checks derived from ISO/IEC 17025:2017 requirements, operating independently of the generative model (Pakina & Pujari, 2024; Garcez & Lamb, 2020). Rules cover calibration certificate validation (traceability statements, uncertainty declarations, certificate-number formats, validity periods), measurement uncertainty verification (Type A/B budget structure, coverage factor, combined uncertainty), and competence record checks (authorization scope, training records). Each rule produces a binary pass/fail outcome, achieving 100 % detection of predefined non-compliance patterns (Figure 2).

Knowledge Graph Export and Active Learning

The framework generates RDF/Turtle provenance triples extending the PROV-O ontology—capturing entity nodes, evidence-determination linkages, timestamps, system version, and human-review annotations (Qi et al., 2021; Yu et al., 2025). An active learning mechanism incorporates high-confidence (≥ 0.9) human reviewer annotations to update clause-evidence mappings and retrieval weights over time without model retraining.

Validity and Reliability

Internal validity is established through controlled baseline comparison: both systems share identical infrastructure (Qwen3-32B, SBERT, same corpus), isolating evidence-grounding constraints and deterministic validation as experimental variables. Reliability is ensured through deterministic rule reproducibility (intraclass correlation = 1.0 on repeated execution, n = 10) and low-temperature RAG generation (0.1). Construct validity is addressed through formal operationalisations: "hallucination" follows Huang et al. (2025), annotated by two certified lead auditors with inter-rater agreement of Cohen's κ = 0.91; "evidence grounding" requires verified verbatim spans with character offsets; "correction rate" is the proportion of outputs requiring human modification. External validity is bounded by the Indonesian context across three laboratory sectors; the modular architecture facilitates adaptation through knowledge-base replacement and

rule extension.

Evaluation Metrics

Performance metrics include: hallucination rate (%), evidence grounding rate (%), precision/recall/F1 against expert annotations, reviewer correction rate (%), rule detection accuracy (%), and query rejection rate (%). Statistical comparison uses McNemar's test for paired binary outcomes and paired t-tests for continuous metrics ($\alpha = .05$). From an operations management perspective, these map to: operational risk mitigation (hallucination reduction), audit cycle-time efficiency (correction rate reduction), compliance process standardisation (grounding rate), and governance reliability (rule accuracy) (Slack et al., 2016; van der Wiele et al., 2011).

FINDINGS AND DISCUSSION

The framework was evaluated across 847 query-clause pairs extracted from 50 audit reports spanning chemical testing, mechanical calibration, and electrical measurement laboratories. The baseline system used identical infrastructure but permitted free-form generation without span verification or deterministic validation. Two independent KAN-certified lead auditors (Expert A: 12 years, >200 audits; Expert B: 9 years, >150 audits) conducted blind reviews of all outputs (Table 2).

Table 2. Expert Reviewer Profile

Reviewer	Certification	Experience	Protocol
Expert A	Lead Auditor (KAN)	12 yrs; >200 lab audits	Independent blind review, 847 outputs
Expert B	Lead Auditor (KAN)	9 yrs; >150 lab audits	Independent blind review, 847 outputs

Note. Inter-rater agreement: Cohen's $\kappa = 0.91$. Disagreements resolved by consensus.

Hallucination Mitigation and Evidence Grounding (RQ1)

Table 3 presents comparative results. The framework eliminated all propagated hallucinations (14.0 % → 0.0 %; McNemar $\chi^2 = 118.6$, $p < .001$), confirming DP1. Common baseline hallucination types included fabricated certificate numbers (38 %), invented uncertainty values (29 %), misattributed calibration dates (21 %), and non-existent traceability references (12 %). The proposed framework's 8.7 % rejection rate represents an intentional safety–coverage trade-off: the system refuses to generate claims without sufficient evidence, prioritising audit integrity over throughput.

Table 3. Experimental Results—Baseline vs. Proposed Framework

Metric	Baseline RAG	Proposed	Improvement
Extrinsic Hallucination Rate	14.0%	0.0%*	100% reduction ($p < .001$)

Evidence Grounding Rate	62.3%	100.0%	+37.7 pp
Reviewer Correction Rate	41.2%	26.0%	37% reduction
Query Rejection Rate	3.1%	8.7%	+5.6 pp
Rule Detection Accuracy	N/A	100.0%	—
Precision	0.78	0.94	+0.16
Recall	0.83	0.89	+0.06

Metric	**Baseline RAG**	**Proposed**	**Improvement**
F1 Score	0.80	0.91	+0.11
Mean Processing Time (s)	2.3	3.8	+65%

**Ungrounded outputs rejected rather than propagated.*

Managerial interpretation

The elimination of hallucinated outputs transforms audit risk management from reactive error-detection to proactive evidence-verification. In the baseline system, each of the 14 % hallucinated outputs represented a potential accreditation liability requiring costly downstream correction. The proposed framework converts this probabilistic risk into a deterministic safety guarantee: no unverifiable claim reaches the human reviewer. For accreditation bodies managing large audit portfolios, this architectural guarantee reduces auditor cognitive burden and enables confident delegation of routine evidence extraction to the AI system (Slack et al., 2016; van der Wiele et al., 2011).

Dual-Retrieval Performance (RQ2)

The dual-retrieval architecture confirms DP2: queries drew from sparse retrieval alone in 23 % of cases, dense retrieval alone in 31 %, and both paths in 46 %. Precision of 0.94 indicates that 94 % of system-generated determinations were confirmed accurate; the 6 % error rate involved partial compliance situations. Recall of 0.89 indicates that 11 % of expert-identified compliant items were missed, primarily due to OCR quality issues (41 % of false negatives), implicit evidence requiring inferential reasoning (35 %), and terminology variations (24 %).

Managerial interpretation

The dual-retrieval design addresses a fundamental operational challenge: the diversity of terminology across laboratories. By combining lexical precision with semantic flexibility, the system achieves robust evidence identification regardless of whether documentation uses

standardised ISO terminology or laboratory-specific jargon. The 46 % of queries benefiting from both paths validates the dual-path architecture as a compliance process-standardisation mechanism (Zu et al., 2008).

Deterministic Validation and Neuro-Symbolic Advantage (RQ3)

The rule-based engine processed 1,247 rule checks, identifying 89 non-compliance instances—missing uncertainty declarations (31), incomplete traceability chains (24), certificate format violations (19), and authorization-scope misalignments (15)—all confirmed by expert reviewers (zero false positives), confirming DP3. Critically, in 23 instances (2.7 % of corpus), the RAG pipeline generated apparently compliant assessments subsequently flagged by deterministic rules detecting missing required elements. Without the symbolic layer, these would have propagated as false positives.

Integration of the hybrid framework reduced reviewer correction rates by 37 % (41.2 % → 26.0 %). The baseline required corrections for hallucination removal (14 %), evidence citation addition (24 %), and determination reversal (3 %). The proposed framework eliminated the first two categories through architectural constraints. Residual corrections involved partial compliance refinements (18 %), additional context supplementation (5 %), and edge-case adjustments (3 %). Active learning further reduced correction rates from 29 % to 23 % after the first 25 reports.

Managerial interpretation

The deterministic validation layer functions as a quality-control checkpoint analogous to automated inspection in manufacturing. The 23 instances caught by rules but missed by the neural component represent a 2.7 % "defect escape rate" that would compromise audit reliability without neuro-symbolic redundancy. For laboratory managers and regulators, this assures that system reliability does not depend solely on LLM performance—a critical consideration for technology adoption in regulated operations (Flynn et al., 1994; Hevner et al., 2004).

Trade-offs and Unexpected Findings

The framework's 8.7 % rejection rate (vs. 3.1 % baseline) and 65 % processing-time increase (2.3 s → 3.8 s per query; ≈54 min vs. 32 min for full audit corpus) represent deliberate trade-offs. The elevated rejection rate reflects verifiability constraints analogous to how quality management systems accept modest throughput reductions to prevent defective outputs. The processing overhead is justified by the substantial reduction in downstream review effort and the elimination of hallucination-related liability. These trade-offs are consistent with risk-management principles underlying ISO/IEC 17025:2017 itself prioritising measurement integrity over processing speed (ISO/IEC, 2017).

Comparison with Existing Approaches

The framework's results distinguish it from existing audit automation approaches along three dimensions. First, compared to commercial RAG-based legal AI tools exhibiting 17–33 % hallucination (Magesh et al., 2025), the architectural verifiability approach—mandatory evidence grounding with character offsets—provides a fundamentally more reliable mechanism than the confidence-threshold filtering and prompt-engineering strategies documented in the hallucination mitigation literature (Huang et al., 2025; Rawte et al., 2023). Where those approaches reduce hallucination probability, this framework eliminates propagation by architectural design, shifting the reliability guarantee from the model layer to the system layer.

Second, compared to traditional rule-based ACC systems achieving less than 45 % accuracy on natural-language requirements (Pakina & Pujari, 2024), the hybrid framework maintains complete rule detection accuracy while adding semantic processing capabilities through neural

retrieval and generation. This result corroborates the neuro-symbolic literature's prediction that hybrid architectures can exceed the performance of either component approach alone (Garcez & Lamb, 2020). The specific contribution here is demonstrating this principle in the compliance auditing domain, where the costs of both false positives and false negatives carry direct regulatory consequences.

Third, situated within the operations management literature on digital quality infrastructure (Slack et al., 2016; Power, 2002), the framework demonstrates that AI-assisted compliance systems can deliver measurable operational efficiency gains (37 % correction rate reduction) while preserving the traceability and audit integrity standards emphasised by quality management scholars (van der Wiele et al., 2011; Flynn et al., 1994). The Human-in-the- Loop design—positioning AI as a pre-screening decision-support agent rather than an autonomous decision-maker—aligns with established principles for decision-support implementation in high-stakes contexts (Arnott & Pervan, 2005; Hevner et al., 2004).

OCR Quality Impact and Knowledge Graph Utility

Documents with OCR quality below 0.85 exhibited elevated false-negative rates (14.3–31.6 %), as character-level errors disrupted pattern matching for structured data elements. This finding underscores a practical prerequisite for AI-assisted auditing: laboratories should establish minimum OCR quality thresholds (≥ 0.90) and prioritise digitisation of legacy documentation. The RDF/Turtle export generated 847 provenance triples enabling SPARQL queries for systemic issue identification (Clause 7.2 as most frequently non-compliant at 17 %), longitudinal trend tracking, and cross-laboratory assessment consistency verification capabilities addressing both ISO/IEC 17025:2017 audit trail requirements and the vision of "Smart Accreditation" (Qi et al., 2021).

CONCLUSION

This study developed and validated an Evidence-Grounded Hybrid Framework for ISO/IEC 17025 compliance audits, achieving all five research objectives. O1 was accomplished through mandatory evidence grounding with character offsets, eliminating propagated hallucinations (14 % → 0 %; $p < .001$). O2 was confirmed by the dual-retrieval architecture, with 46 % of queries benefiting from both sparse and dense retrieval paths. O3 was validated through 100 % detection accuracy of predefined non-compliance patterns. O4 was realised through RDF/Turtle provenance export, generating 847 queryable triples. O5 was established through systematic benchmarking demonstrating statistically significant improvements across all evaluation metrics (F1 = 0.91; correction rate reduction = 37 %). Addressing the research questions: RQ1 is answered by the complete elimination of hallucination propagation; RQ2 by the demonstrated complementarity of dual retrieval; and RQ3 by the 37 % reduction in reviewer correction burden.

Managerial and Policy Implications

The findings generate four actionable recommendations. First, accreditation bodies should integrate AI-assisted compliance systems as pre-screening mechanisms within staged audit workflows, assigning routine evidence extraction tasks to automated systems while reserving professional judgment for complex compliance assessments. Second, laboratory managers should establish OCR quality thresholds of at least 0.90 and prioritize document digitization as a fundamental prerequisite for effective AI-augmented operations. Third, regulators should require deterministic validation fail-safes for any AI system deployed in high-risk compliance decision-making, since large language model (LLM)-based systems necessitate independent rule-based verification. Fourth, accreditation ecosystems should adopt knowledge graph-based provenance frameworks to support longitudinal governance, enabling programmatic trend analysis and continuous monitoring capabilities beyond the limitations of conventional static audit reports.

Contributions in Operations Terms

The framework delivers: (1) audit cycle-time reduction through automated evidence extraction, reducing outputs requiring human correction by 37 %; (2) compliance process standardisation through deterministic grounding and rule-based validation, ensuring consistent outputs regardless of auditor interpretation; (3) operational risk mitigation through architectural elimination of hallucinated claims, converting a probabilistic 14 % error risk to a deterministic zero-propagation guarantee; and (4) digital quality infrastructure modernisation through machine-readable provenance supporting longitudinal analytics and smart accreditation (Slack et al., 2016; van der Wiele et al., 2011). The architectural approach—enforcing trustworthiness through structural constraints rather than depending on probabilistic model behaviour—provides a generalisable template for evidence-critical domains including legal compliance, medical documentation, financial auditing, and supply chain certification.

LIMITATIONS AND FUTURE RESEARCH

Several limitations shape applicability: (1) reliance on Indonesian translations constrains multilingual generalisability; (2) deterministic rules impose maintenance burden as standards evolve and currently cover approximately 60 % of ISO/IEC 17025:2017 requirements; (3) OCR sensitivity degrades performance below 0.85 quality; and (4) the static, document-bound pipeline prevents multi-hop reasoning across interconnected evidence. Future research should explore GraphRAG for multi-document reasoning, multi- agent adversarial verification, automated rule learning from annotated audit decisions, extension to related standards (ISO 15189, ISO/IEC 17020), and integration with laboratory information management systems for continuous auditing and temporal evidence reasoning.

REFERENCES

Arnott, D., & Pervan, G. (2005). A critical analysis of decision support systems research. *Journal of Information Technology, 20*(2), 67–87. https://doi.org/10.1057/palgrave.jit.2000035

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data, 4*, Article 688969. https://doi.org/10.3389/fdata.2021.688969

Chasandras, I. (n.d.). *Retrieval augmented generation on regulatory documents* [Unpublished manuscript].

Cheng, M., Luo, Y., Ouyang, J., et al. (2025). *A survey on knowledge-oriented retrieval-augmented generation* (arXiv:2503.10677). arXiv. https://doi.org/10.48550/arXiv.2503.10677

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis, 16*(1), 64–93.

https://doi.org/10.1093/jla/laae003

Flynn, B. B., Schroeder, R. G., & Sakakibara, S. (1994). A framework for quality management research and an associated measurement instrument. *Journal of Operations Management, 11*(4), 339–366. https://doi.org/10.1016/S0272-6963(97)90004-8

Garcez, A. d'A., & Lamb, L. C. (2020). *Neurosymbolic AI: The 3rd wave* (arXiv:2012.05876). arXiv. https://doi.org/10.48550/arXiv.2012.05876

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly, 37*(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75–105. https://doi.org/10.2307/25148625

Huang, L., Yu, W., Ma, W., et al. (2025). A survey on hallucination in large language models. *ACM Transactions on Information Systems, 43*(2), 1–55. https://doi.org/10.1145/3703155

ISO/IEC. (2017). *ISO/IEC 17025:2017 general requirements for the competence of testing and calibration laboratories*. International Organization for Standardization.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies, 22*(2), 216–242. https://doi.org/10.1111/jels.12413

Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). *A systematic review of key retrieval-augmented generation (RAG) systems* (arXiv:2507.18910). arXiv. https://doi.org/10.48550/arXiv.2507.18910

Pakina, A. K., & Pujari, M. (2024). Neuro-symbolic compliance architectures. *International Journal of Science and Technology, 4*(1), 56–66. https://doi.org/10.56127/ijst.v4i1.1961

Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). Sage.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems artifacts. *Journal of Management Information Systems, 24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Power, D. J. (2002). *Decision support systems: Concepts and resources for managers*. Quorum Books.

Qi, Q., Tao, F., Hu, T., et al. (2021). Enabling technologies and tools for digital twin. *Journal of Manufacturing Systems, 58*, 3–21. https://doi.org/10.1016/j.jmsy.2019.10.001

Qwen, A. Y., Yang, B., et al. (2025). *Qwen2.5 technical report* (arXiv:2412.15115). arXiv. https://doi.org/10.48550/arXiv.2412.15115

Rawte, V., Sheth, A., & Das, A. (2023). *A survey of hallucination in large foundation models* (arXiv:2309.05922). arXiv. https://doi.org/10.48550/arXiv.2309.05922

Slack, N., Brandon-Jones, A., & Johnston, R. (2016). *Operations management* (8th ed.). Pearson Education.

Taylor, B. N. (n.d.). *The International System of Units (SI)* (NIST Special Publication 330). National Institute of Standards and Technology.

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., et al. (2024). *A comprehensive survey of hallucination mitigation techniques in large language models* (arXiv:2401.01313). arXiv. https://doi.org/10.48550/arXiv.2401.01313

UNIDO. (2020). *Complying with ISO 17025: A practical guidebook*. United Nations Industrial Development Organization.

van der Wiele, T., van Iwaarden, J., Williams, R., & Eldridge, S. (2011). A new foundation for quality management in the business environment of the twenty-first century. *Total Quality Management & Business Excellence, 22*(5), 587–598. https://doi.org/10.1080/14783363.2011.568262

World Bank. (2020). *Quality infrastructure in Indonesia: Assessment and recommendations*. World Bank Group.

Yu, H., Gan, A., Zhang, K., et al. (2025). Evaluation of retrieval-augmented generation: A survey. In *Lecture Notes in Computer Science* (Vol. 2301). Springer. https://doi.org/10.1007/978-981-96-1024-2_8

Zu, X., Fredendall, L. D., & Douglas, T. J. (2008). The evolving theory of quality management: The role of Six Sigma. *Journal of Operations Management, 26*(5), 630–650. https://doi.org/10.1016/j.jom.2007.12.002