



Scripta Technica: Journal of Engineering and Applied Technology

Vol 2 No 2 December 2026, Hal. 16-39
ISSN:3110-0775(Print) ISSN: 3109-9696(Electronic)
Open Access: <https://scriptainteletektual.com/scripta-technica>

A Modular Retrieval-Augmented Generation Architecture for Grounded Document Analysis

Stanley Nwakamma^{1*}, Serif Oyindamola Oyesiji², Harouna Wendpanga Yann Christian Sankara³, Emmanuel Eniola Ajala⁴

^{1,3-4} Independent Researcher, United States of America

² Harrisburg University of Science and Technology, United States of America

email: nwakammastanley@gmail.com^{1*}

Article Info :

Received:
01-09-2026
Revised:
07-10-2026
Accepted:
14-10-2026

Abstract

Organizations increasingly employ language models for the analysis of contracts, regulatory filings, technical manuals, and clinical guidance, where unsupported or misattributed statements create material reliability risks. Retrieval-augmented generation (RAG) mitigates some limitations of parametric models by conditioning generation on external evidence, but existing architectures vary substantially in retrieval, control, attribution, and verification mechanisms. Many implementations do not preserve sentence-level provenance or provide stage-level diagnostics for unsupported outputs. This study develops a conceptual RAG architecture using a design-oriented research method and a structured narrative synthesis of literature on retrieval, failure modes, mitigation, measurement, and governance. The synthesis yields eight requirements for grounded document analysis and informs a modular architecture comprising structure-aware segmentation, hybrid indexing, query planning, retrieval, reranking, provenance-preserving evidence assembly, citation-mandatory generation, and claim-level verification. The architecture specifies evidence-sufficiency and claim-verification gates, bounded repair, and single-pass, iterative, and abstaining execution policies. Its principal contribution is an explicit system specification that integrates span-level attribution, structural preservation, conflict transparency, calibrated abstention, and stage-level telemetry. Because no implementation or empirical benchmark is reported, the proposed mechanisms are treated as design propositions. Seven testable hypotheses and a four-phase evaluation protocol are specified for subsequent empirical assessment.

Keywords: retrieval-augmented generation; grounded document analysis; claim-level attribution; hallucination verification; modular architecture; conceptual framework.



©2026 Authors.. This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.
(<https://creativecommons.org/licenses/by-nc/4.0/>)

INTRODUCTION

Retrieval-augmented generation (RAG) conditions language-model inference on externally retrieved evidence and thereby mitigates documented limitations of purely parametric knowledge, including temporal staleness, incomplete coverage of long-tail facts, and limited provenance (Kandpal et al., 2023; Mallen et al., 2023). Its technical foundations include end-to-end trained retriever-generator formulations (Lewis et al., 2020; Guu et al., 2020; Izacard & Grave, 2021) and in-context retrieval over frozen models (Ram et al., 2023). Empirical evidence further indicates that retrieval augmentation can reduce hallucination relative to closed-book generation (Shuster et al., 2021). Collectively, these developments establish RAG as a mechanism for extending model access to externally verifiable information.

RAG encompasses several architectural families rather than a single implementation pattern. The retrieve-then-read formulation (D. Chen et al., 2017) has developed into jointly trained retriever-generator models (Lewis et al., 2020; Guu et al., 2020; Izacard & Grave, 2021; Borgeaud et al., 2022), in-context retrieval over frozen models (Ram et al., 2023; W. Shi et al., 2023), iterative and active retrieval interleaved with reasoning (Trivedi et al., 2023; Jiang et al., 2023), agentic systems in which a planning model determines retrieval actions (Singh et al., 2025), and graph-based approaches that retrieve over entity and community structures (Edge et al., 2024). Contemporary surveys distinguish naive, advanced, and modular configurations (Y. Gao et al., 2024a), with modular RAG extending this classification through reconfigurable operator graphs (Y. Gao et al., 2024b). The underlying retrieval technologies include lexical ranking (Robertson & Zaragoza, 2009), dense passage retrieval (Karpukhin et al., 2020), late interaction (Khattab & Zaharia, 2020), cross-encoder reranking (Nogueira & Cho,

2019), and large-scale vector indexing (Johnson et al., 2021; Malkov & Yashunin, 2020). Cross-domain benchmarks nevertheless show that retrieval quality remains domain-dependent (Thakur et al., 2021). The present study therefore focuses on a modular architecture compatible with both in-context and iterative execution.

Retrieval access alone does not establish evidentiary fidelity. Engineering studies of deployed systems document failures involving missing content, ranking errors, context consolidation, incorrect specificity, and specification drift between pipeline stages (Barnett et al., 2024). Behavioral benchmarks likewise show reduced reliability in the presence of noisy, absent, or conflicting evidence (J. Chen et al., 2024; Longpre et al., 2021; Xie et al., 2024), while long-context studies identify positional underuse of relevant evidence and degradation caused by irrelevant material (Liu et al., 2024; F. Shi et al., 2023). Attribution studies further demonstrate that generated citations may fail to support the associated claims (T. Gao et al., 2023; Rashkin et al., 2023). These observations are consistent with broader analyses of hallucination (Ji et al., 2023; Huang et al., 2025) and temporal degradation of parametric knowledge (Lazaridou et al., 2021). In legal, regulatory, technical, and clinical settings, such failures directly affect reliability, traceability, and auditability. Empirical studies in law and medicine illustrate the resulting risk profile (Dahl et al., 2024; Singhal et al., 2023), while regulated deployment additionally requires reconstruction of the evidence and system configuration underlying released outputs (Ominyi et al., 2026; Nwakamma et al., 2024a).

Mitigation methods are distributed across distinct stages of the RAG pipeline. Query decomposition, planning, and rewriting improve evidence acquisition (Press et al., 2023; Trivedi et al., 2023; Ma et al., 2023), as do active retrieval (Jiang et al., 2023) and methods designed to reduce sensitivity to irrelevant context (Yoran et al., 2024). Self-reflective and corrective approaches address grounding during generation (Asai et al., 2024; Yan et al., 2024), whereas post-hoc verification and revision target unsupported generated content (Dhuliawala et al., 2024; Gou et al., 2024; Peng et al., 2023; Madaan et al., 2023; L. Gao et al., 2023). Uncertainty estimation provides additional signals for gating and abstention (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024; Kadavath et al., 2022), subject to calibration constraints (Guo et al., 2017). Measurement methods have developed in parallel through claim-level factuality scoring (Min et al., 2023), attribution and citation-quality metrics (Rashkin et al., 2023; T. Gao et al., 2023), factual consistency classifiers (Kryscinski et al., 2020; Laban et al., 2022; Honovich et al., 2022), and composed-system evaluation suites (Es et al., 2024; Saad-Falcon et al., 2024). These methods provide substantial stage-specific coverage but do not, individually, define system-level release conditions.

A complementary literature addresses governance requirements for AI systems operating in regulated or high-accountability environments. Research on responsible AI governance, organizational readiness for the European Union's Artificial Intelligence Act, and enterprise AI governance emphasizes traceable evidence, reviewability, and institutional accountability (Ominyi & Anichukwueze, 2024; Ominyi et al., 2024b; Ominyi et al., 2026). Related studies address algorithmic accountability and model risk management in financial services (Ominyi & Anichukwueze, 2023), privilege and confidentiality in enterprise language-model deployment (Ominyi et al., 2025c), data integrity in regulated life sciences (Aneke et al., 2023), evidence chains in cross-border financial auditing (Agu et al., 2023), and accountability structures for AI-driven health management (Afrihyia et al., 2024). Data-governance research contributes sensitivity classification, regulatory traceability, compliance-oriented lakehouse governance, and legal and ethical risk modeling (Aliliele et al., 2024; Aliliele et al., 2025b; Mbonu et al., 2018). Assurance research contributes continuous auditing, compliance analytics, analytics-supported audit quality, and control automation (Aliliele et al., 2025a; Akomolafe et al., 2025; Ozowara et al., 2025; Mbonu, Aliliele, et al., 2022). Research on model-based judging and human-in-the-loop evaluation further identifies calibration bias and expert disagreement as measurable governance concerns (Nwakamma et al., 2024a; Ladapo et al., 2022; Ladapo et al., 2026). These findings support explicit requirements for traceability, controlled release, and post hoc review in a grounded RAG architecture.

Existing integrative frameworks address these requirements only partially. Modular RAG represents pipelines as reconfigurable operators and flows (Y. Gao et al., 2024b), but does not define attribution, verification, or abstention as mandatory release conditions. Self-reflective generation internalizes support judgments within the generator (Asai et al., 2024), which limits their independence as externally auditable controls. Corrective retrieval evaluates and repairs retrieved evidence (Yan et

al., 2024) without requiring claim-level verification of the final output. Agentic designs provide flexible control flow (Singh et al., 2025), although their grounding behavior depends on agent decisions and may therefore be more difficult to audit consistently. The resulting gap concerns the absence of an integrated specification that combines provenance, verification, conflict handling, abstention, and stage-level observability as enforceable architectural properties.

The unresolved research problem is the absence of a system-level RAG architecture that composes retrieval, grounding, and verification mechanisms under explicit release constraints. Such an architecture requires (a) binding each released claim to an immutable and auditable document span, (b) conditioning release on evidence-sufficiency and claim-verification criteria, and (c) attributing failures to specific pipeline stages through standardized telemetry. Explicit module contracts and release gates are therefore required to operationalize grounding as a verifiable system property rather than an implicit behavior of the generative model.

Accordingly, the study addresses four research questions. RQ1: Which requirements must a RAG architecture satisfy to support grounded analysis of high-stakes documents, and which documented failure evidence motivates each requirement? RQ2: To what extent do representative existing RAG architectures satisfy these requirements as enforceable architectural properties rather than optional behaviors? RQ3: What modular architecture, expressed through module contracts, typed interfaces, release gates, and execution policies, can operationalize these requirements? RQ4: Which hypotheses and evaluation procedures are required to make the architecture's central claims empirically falsifiable?

The study contributes an architecture-level synthesis in four respects. It derives eight requirements for grounded document analysis from documented failure modes and deployment constraints; comparatively evaluates seven representative RAG architectures against those requirements; specifies a modular reference architecture with typed interfaces, immutable span identity, explicit release gates, a verification-verdict taxonomy, and bounded repair; and formulates seven testable hypotheses with a four-phase empirical evaluation protocol. The contribution is intentionally limited to architectural specification and analytical evaluation rather than demonstrated performance improvement.

The scope of inference is therefore restricted to the internal coherence, traceability, and testability of the proposed design. Claims concerning realized improvements in attribution accuracy, hallucination reduction, latency, cost, or operational robustness require implementation and empirical validation.

RESEARCH METHOD

Research Design

A conceptual, design-oriented research design was employed. The study follows the model-oriented form of conceptual research, which develops a framework that specifies relationships among constructs rather than reporting primary observations (Jaakkola, 2020), and adopts the design-science view that an architecture specification constitutes a research artifact whose utility requires subsequent evaluation (Hevner et al., 2004; Peffers et al., 2007). Consistent with the distinction between *ex ante* and *ex post* evaluation in design science (Venable et al., 2016), the present analysis evaluates the internal coherence and requirement coverage of the architecture *ex ante*, while empirical effectiveness is reserved for a separately specified validation protocol. This design is appropriate because the research questions concern architectural requirements and their formal composition rather than observed system performance.

Literature Identification

Literature was identified through structured searches of the ACL Anthology, arXiv (cs.CL and cs.IR categories), the ACM Digital Library, IEEE Xplore, and Google Scholar. Search strings combined a RAG term ("retrieval-augmented generation", "retrieval augmented", "open-domain question answering") with one of four strand terms: failure ("hallucination", "faithfulness", "knowledge conflict", "failure"), mitigation ("verification", "self-reflection", "query decomposition", "reranking", "abstention", "uncertainty"), measurement ("attribution", "citation", "factual consistency", "evaluation"), or architecture ("modular", "agentic", "pipeline", "framework"). The core search window was 2017 to 2025, beginning with the retrieve-then-read formulation (D. Chen et al., 2017). Seminal earlier works, such as probabilistic relevance ranking and calibration studies, entered through backward and forward snowballing from included papers (Wohlin, 2014). The governance strand was identified

by targeted searches combining AI governance, audit, traceability, and regulatory compliance terms with the document domains the framework targets.

Selection Criteria

Machine-learning sources were included if they (a) proposed, analyzed, or evaluated a RAG stage, a failure mode of evidence-conditioned generation, a mitigation, or a measurement instrument for attribution or faithfulness; (b) were published in a peer-reviewed venue or were widely cited preprints; and (c) described their mechanism in sufficient technical detail to determine the stage at which it operates. Governance sources were included if they articulated requirements for traceability, auditability, review, or release decisions in regulated or high-accountability settings. Sources were excluded if they addressed open-ended conversational generation without an evidence component, provided no mechanism-level detail, or duplicated a more complete version of the same work.

Synthesis Procedure

The literature was synthesized using the narrative-synthesis procedure described by Popay et al. (2006). The first operation coded each source according to pipeline stage, documented or addressed failure mode, mechanism, evidence type (benchmark, field measurement, or design argument), and emitted signal. The second mapped failure modes to their stage of origin and linked mitigation methods to the failures they address, thereby identifying stage-specific concentrations of mitigation techniques. The third translated this mapping into architectural requirements through a traceability rule under which a requirement was retained only when supported by at least one documented failure mode and one deployment constraint. The fourth examined the robustness of the resulting requirements against counter-evidence and through comparative assessment of representative existing frameworks.

From Requirements to Architecture

Translation of the requirements into architecture followed three design rules. Stage alignment places a module boundary at each point where the failure literature identifies a distinct failure mode, thereby assigning detection responsibility to a defined component. Contract allocation maps every requirement to at least one module contract and one telemetry signal. Falsifiability links every requirement to an evaluation phase and a testable hypothesis. Table 1 summarizes the resulting traceability from motivating evidence to architectural mechanism and empirical test.

Methodological Limitations

The synthesis is narrative rather than systematic. It did not employ a pre-registered protocol, duplicate independent screening, or PRISMA-style study counts; consequently, the literature coverage is indicative rather than exhaustive. Coding and requirement derivation also involved author judgment. The governance literature includes work from the authors' research network, which may introduce selection bias and should be considered when interpreting the synthesis. Ex ante analytical evaluation can establish internal consistency and specification completeness but cannot establish empirical effectiveness. The latter remains dependent on implementation and controlled evaluation.

RESULT AND DISCUSSION

The design-oriented synthesis produced four categories of analytical output: requirements for grounded document analysis, a comparative assessment of representative RAG frameworks, a modular reference architecture, and a set of falsifiable hypotheses with an associated evaluation protocol. The results are therefore conceptual and analytical rather than measurements of implemented system performance. Interpretation focuses on whether the proposed mechanisms form a coherent architecture for preserving provenance, constraining release, localizing failure, and supporting subsequent empirical validation.

Requirements for Grounded Document Analysis

Eight requirements were derived for grounded document analysis. R1, evidentiary fidelity, requires every released claim to be supported by cited corpus spans at claim granularity, with support verified rather than assumed (Rashkin et al., 2023; Min et al., 2023). R2, attribution auditability, requires citations to resolve to exact document spans with stable identifiers, enabling direct review of

the evidentiary basis for each claim; this function is analogous to chain-of-custody and audit-trail mechanisms in regulated distribution and configuration management (Aneke & Adesemoye, 2019; Oshoba et al., 2020). R3, structural preservation, requires segmentation and retrieval to retain document structures such as clauses, sections, tables, and definitions, because structure contributes materially to meaning in analytical documents (Barnett et al., 2024). R4, conflict transparency, requires contradictory evidence to be represented explicitly with attribution to the relevant sources rather than resolved implicitly by the generator (Longpre et al., 2021; Xie et al., 2024; J. Chen et al., 2024).

R5, calibrated abstention, requires absent or insufficient evidence to produce a controlled non-answer or partial answer supported by the retrieval record and calibrated uncertainty signals (Farquhar et al., 2024; Kadavath et al., 2022). R6, stage observability, requires each module to emit scores and telemetry sufficient to localize failures to pipeline stages (Barnett et al., 2024; Saad-Falcon et al., 2024; Mbonu et al., 2019). R7, modularity, requires components to remain replaceable behind typed interfaces without cross-stage rewiring (Y. Gao et al., 2024a). R8, configuration traceability, requires the model versions, index snapshots, gate thresholds, and other configuration elements associated with a released answer to be recorded and reviewable (Ominyi et al., 2026; Ominyi et al., 2024b).

Table 1 records, for each requirement, the documented failure evidence that motivates it, the modules and mechanisms that realize it, and the evaluation phase and hypothesis that would test it.

Table 1. Requirement Traceability: From Failure Evidence To Mechanism And Test

Requirement	Motivating failure evidence	Realizing modules and mechanisms	Evaluation phase and hypothesis
R1 Evidentiary fidelity	Citations that do not support their claims (T. Gao et al., 2023; Rashkin et al., 2023)	Citation-mandatory generation (M7); claim-verification gate (M8); repair policy	Phases 2 and 3; H2, H5
R2 Attribution auditability	Absent provenance in parametric and monolithic systems (Kandpal et al., 2023; Barnett et al., 2024)	Span identity and hashing (M1); span-typed interfaces; provenance formatting (M6)	Phases 1 and 2; H1
R3 Structural preservation	Context-severing chunking (Barnett et al., 2024)	Canonical document model, structure-aware spans, reference edges (M1); structural metadata (M2)	Phase 1; H1
R4 Conflict transparency	Knowledge-conflict behavior (Longpre et al., 2021; Xie et al., 2024)	Conflict detection (M6); attributed disagreement (M7)	Phase 2
R5 Calibrated abstention	Fluent answers from absent evidence (J. Chen et al., 2024)	Sufficiency gate (M6); verdict taxonomy (M8); abstention schema	Phase 2; H3
R6 Stage observability	Stage-specific failure points (Barnett et al., 2024)	Required telemetry for every module (Table 3)	Phase 4; H6
R7 Modularity	Rapid component turnover (Y. Gao et al., 2024a)	Typed interfaces; format adapters; module registry	Phase 4; H7
R8 Configuration traceability	Defensibility requirements in regulated settings (Ominyi et al., 2026)	Module registry; configuration versioning; gate-decision logs	Phase 4; H4

Note. M1 to M8 denote the modules summarized in Table 3; H1 to H7 denote the hypotheses specified in the evaluation protocol.

Comparison with Existing Modular and Agentic Frameworks

Seven representative architectures were assessed against the derived requirements. The selected systems represent major architectural families and mitigation strategies identified in the literature. Ratings were assigned from each framework's published specification rather than reported empirical performance. A rating of "Yes" indicates an explicitly specified and externally checkable mechanism;

"Partial" indicates a mechanism that is internal to a model, optional, or not coupled to release; "No" indicates that no relevant mechanism is specified; and "Varies" indicates implementation-dependent behavior.

Table 2. Comparison of Representative RAG Architectures Against The Requirements For Grounded Document Analysis

Framework (family)	Claim-level verification at release	Span-level attribution	Structural spans	Conflict surfacing	Calibrated abstention	Telemetry and provenance	Typed modularity
Naive and advanced RAG (Y. Gao et al., 2024a)	No	No	Partial	No	No	No	No
Modular RAG (Y. Gao et al., 2024b)	No	No	Partial	No	No	No	Yes
Self-RAG (Asai et al., 2024)	Partial	Partial	No	No	No	Partial	No
Corrective RAG (Yan et al., 2024)	No	No	Partial	No	No	Partial	Partial
Iterative and active retrieval (Trivedi et al., 2023; Jiang et al., 2023)	No	No	No	No	No	Partial	No
Post-hoc verification and revision (L. Gao et al., 2023; Dhuliawala et al., 2024)	Partial	Partial	No	No	No	Partial	Partial
Agentic RAG (Singh et al., 2025)	Varies	Varies	Varies	Varies	Varies	Varies	Partial
Proposed framework (this paper)	Specified	Specified	Specified	Specified	Specified	Specified	Specified

Note. Ratings reflect each framework's published specification, not empirical performance. "Specified" for the proposed framework denotes an architectural specification whose effectiveness has not yet been tested. "Telemetry and provenance" covers stage observability (R6) and configuration provenance (R8).

The comparative assessment in Table 2 indicates that none of the reviewed frameworks jointly specifies claim-level verification at release, span-level attribution, explicit conflict handling, and calibrated abstention. Self-reflective generation includes verification-related behavior but locates the support judgment within the generator. Post-hoc revision evaluates generated claims against evidence snippets but does not require stable span identity or a structural document model. Modular RAG provides a flexible composition mechanism while leaving grounding semantics to the implementation.

The comparative result therefore supports the need for an architecture in which these mechanisms are defined as interoperable and enforceable contracts.

Architecture Overview

The framework decomposes the pipeline into eight modules on two planes, coordinated by an orchestration layer. The document plane runs ahead of queries: ingestion and segmentation convert source files into a structure-preserving span store, and indexing builds lexical and dense representations over the spans with metadata. The query plane runs per request: query planning interprets and decomposes the request; retrieval executes hybrid search per planned sub-query; reranking refines candidate sets; evidence assembly constructs the provenance-preserving context; grounded generation produces claim-cited output; and verification checks each claim against its citations before release.

The orchestration layer sequences modules under declared execution policies: single-pass processing for simple lookups, iterative processing when planning or verification requires additional evidence (Trivedi et al., 2023; Jiang et al., 2023), and abstention when release criteria are not satisfied. Interfaces are typed by span identity; after segmentation, each index entry, candidate, context element, citation, and verification verdict references an immutable span identifier. This interface discipline is intended to preserve attribution auditability across the pipeline. Related platform-engineering practices demonstrate how API governance, infrastructure-as-code, and portable component models can support reviewable configuration and component substitution (Aliliele et al., 2023; Mbonu et al., 2020; Ojukwu et al., 2025). Table 3 summarizes the eight modules, their contracts, and required telemetry.

Table 3. Module Specification: Contract and Required Telemetry

Module	Contract	Required telemetry
1. Ingestion and segmentation	Emit spans aligned to document structure, each with source, location, version, and hash	Parse coverage, structure-preservation flags, span size distribution
2. Indexing	Build lexical and dense representations over spans with structural and governance metadata	Index snapshot version, coverage gaps, embedding model identifier
3. Query planning	Convert a request into a logged, inspectable retrieval program of scoped sub-queries	Plan shape, request class, decomposition depth, applied scopes
4. Retrieval	Execute hybrid search per sub-query with score fusion under plan scopes	Hit distributions, fused relevance mass, calibrated scores
5. Reranking	Refine fused candidates by cross-encoder scoring	Rerank deltas, score calibration, latency per candidate set
6. Evidence assembly	Deduplicate, detect conflict, order by position findings, format with visible provenance	Conflict flags, context composition, sufficiency-gate decision
7. Grounded generation	Emit claims bound to span identifiers; render conflict as attributed disagreement	Claim count, citation density, uncited-claim rate
8. Verification	Adjudicate every claim-citation pair before release; repair or remove failures	Per-claim verdicts, repair outcomes, abstention rate, judge calibration

Document Plane: Segmentation and Indexing

Canonical document model and supported formats

Because PDF documents, scanned images, HTML, word-processing files, spreadsheets, and regulatory databases encode structure differently, the architecture does not assume a single parser. Instead, format-specific parsing adapters, each behind the ingestion interface, map sources into one canonical document model: a tree of typed nodes (section, clause, list item, definition, table, table cell, footnote, appendix or schedule, figure caption) in which every node retains a pointer to its source object. Table 4 specifies the parsing route, structural signals, and principal risks for each supported format.

Downstream modules see only the canonical model, so supporting a new format means adding an adapter rather than changing any other module.

Table 4. Supported Source Formats and Parsing Adapters

Format	Parsing route	Structural signals used	Principal risks and handling
Born-digital PDF	Layout analysis: reading order, block and heading detection, table detection	Font size and weight, numbering patterns, page and bounding-box coordinates	Multi-column reading-order errors, running headers and footers; source pointer is page plus bounding box
Scanned PDF and images	OCR combined with a layout model	As for born-digital PDF, plus token-level OCR confidence	Recognition errors; low-legibility flags, stricter numeric thresholds, and retention of the page image for verification
Word-processing files (DOCX)	Native XML parsing	Paragraph styles, list numbering, native tables, footnote parts	Manual formatting used instead of styles; fallback to layout heuristics
HTML and XML	DOM parsing; schema-aware parsing of legislative or regulatory markup where available	Heading and sectioning elements, semantic markup, anchors and links	Navigation and boilerplate content removed; source pointer is element path
Spreadsheets	Cell-grid parsing	Sheets, header rows, named ranges, merged cells, number formats and units	Implicit headers and formulas; values stored typed, formulas recorded; pointer is sheet and cell range
Regulatory databases and structured feeds	Native record schema through an API or export	Record fields, consolidation status, effective and repeal dates	Drift between consolidated and original texts; addressed through effective-date metadata and explicit versioning

Note. All adapters emit the same canonical document model, so downstream modules are format-independent.

Structure-aware segmentation

Segmentation emits spans aligned to the canonical model: clause-level units for contractual and regulatory text, row-level units with headers for tables, and definition units linked to their defined terms. The path of headings above each span is stored as metadata rather than flattened into the text. Span size adapts to structure rather than to a fixed token count; units longer than the embedding model's effective input are subdivided at sentence boundaries, and overlapping windows are used only within structural units, never across their boundaries.

Restricting overlap to structural units is a design proposition rather than an established empirical result. Structural and semantic boundaries may diverge when clauses continue across page breaks, sections contain multiple topics, or a provision depends on an exception stated elsewhere. The architecture therefore combines structure-aware segmentation with explicit reference edges and parent-context expansion, allowing evidence assembly to include an enclosing section when several related spans are retrieved. H1 evaluates whether this combination improves evidence recall and attribution precision relative to fixed-size and semantic chunking.

OCR Errors, Tables, Cross-References, Footnotes, and Appendices

OCR errors. For scanned sources, the OCR adapter records token-level confidence, and each span stores its minimum and mean confidence together with the page image region from which it was read.

Spans below a calibrated confidence threshold are flagged as low-legibility. They remain retrievable, but the verifier treats claims that rest solely on them as uncheckable unless corroborated by another span, and numeric tokens, where a single misread digit changes meaning, are held to a stricter threshold. Low-legibility rates are reported per document as ingestion telemetry.

Tables. A table is represented as a table node with its header hierarchy, merged cells expanded, units, and footnote markers. Segmentation emits one span per row, serialized with its full header path (for example, "Fee schedule > Tier 2 > Annual fee: 4,500"), plus a whole-table span for small tables. Numeric cell values are also stored in typed form, so that the verifier can check numeric claims deterministically rather than by textual entailment.

Cross-references. Internal references to related clauses, definitions, schedules, annexes, or other document components are resolved during ingestion as directed edges between the corresponding spans. Defined terms are linked to the spans containing their formal definitions, while unresolved references are recorded as parsing exceptions. During evidence assembly, directly referenced spans may be incorporated within a one-hop expansion limit to preserve the dependency structure of the source document and reduce the risk of interpreting a provision without its associated qualification, exception, or definition.

Footnotes and appendices. Footnotes are emitted as separate spans linked to their anchor span, so that they can be retrieved independently and are also visible as context of the anchor. Appendices, schedules, and annexes are top-level nodes with a type label, and documents incorporated by reference are linked by document-level edges when they are present in the corpus.

Span Identity, Versioning, and Hashing

Each span has a logical identifier composed of a document identifier, a structural path, and an ordinal, plus a version identifier. Two hashes are computed using a standard cryptographic hash function such as SHA-256. A source-object hash is computed over the raw bytes of the ingested file and supports chain-of-custody verification. A span content hash is computed over normalized span text (Unicode normalization, whitespace collapse, and removal of line-break hyphenation, with case, punctuation, and numerals preserved) and supports identity and change detection. Hashing raw text alone would make identity brittle to re-parsing, and hashing only the source object could not localize changes within a document.

When a new version of a document is ingested, spans are aligned to the previous version by structural path and content similarity. A span whose content hash is unchanged keeps its logical identifier and receives the new version pointer; a changed span receives a new version linked to its predecessor by a "supersedes" edge; a removed span is retained as a tombstone. Earlier versions are never overwritten, so a citation issued against a prior version continues to resolve to the exact text that was cited, and effective-date metadata allows queries to be scoped to the version in force on a given date.

Curation and Indexing

Corpus curation is enforced during ingestion because the span store defines the universe of evidence that may subsequently be cited. Context-aware curation governs admission to the corpus, while sensitivity classification attaches handling constraints that remain associated with each span (Oyesiji et al., 2024; Aliliele et al., 2024).

Indexing combines lexical and dense representations with structural and governance metadata. Lexical indexing supports exact matching for defined terms, section numbers, and names (Robertson & Zaragoza, 2009), whereas dense indexing supports semantic recall across paraphrases (Karpukhin et al., 2020). Late-interaction retrieval can be used when token-level matching justifies its computational cost (Khattab & Zaharia, 2020), using established vector-search infrastructure (Johnson et al., 2021; Malkov & Yashunin, 2020). Structural metadata (document type, section path, table flags, and effective dates) and governance metadata (access constraints, retention rules, and sensitivity labels) are indexed with content so that retrieval can be scoped and handling constraints can be enforced (Aliliele et al., 2025b; Badmus et al., 2024). Embedding models remain replaceable under the indexing interface, subject to in-domain evaluation because retrieval quality varies across domains (Thakur et al., 2021). Parameter-efficient adaptation provides an additional option for multilingual or domain-distant corpora without altering the index contract (Oyesiji et al., 2023).

Query Plane: Planning, Retrieval, and Assembly

Query planning converts each request into an explicit retrieval program. The planner classifies the request as lookup, comparison, aggregation, or cross-document analysis; decomposes compound questions into dependent sub-queries (Press et al., 2023); rewrites sub-queries into corpus-effective forms (Ma et al., 2023); and attaches structural scopes such as document sets, sections, and date ranges. The resulting plan is stored as an inspectable intermediate artifact that can be revised during subsequent iterations, supporting diagnosis of multi-hop retrieval behavior (Trivedi et al., 2023). It also separates analytical intent from the retrieval operations used to satisfy that intent (Adelanwa et al., 2024) and enables compliance-oriented constraints, including restriction to instruments effective on a specified date (Ominyi et al., 2024a).

Retrieval executes each sub-query through hybrid search with score fusion under the structural and governance scopes defined by the plan. Reranking then applies cross-encoder refinement to the fused candidate set (Nogueira & Cho, 2019), with both stages emitting calibrated relevance scores as telemetry. Because reranking can contribute substantially to inference latency and computational cost, model compression and efficient deployment techniques provide relevant mechanisms for meeting operational latency constraints (Oyesiji et al., 2025; Nwakamma et al., 2025).

Evidence assembly constructs the generation context under explicit policy. The procedure deduplicates candidates by span identity; detects disagreement among spans addressing the same query point so that conflicting evidence is represented explicitly (J. Chen et al., 2024; Xie et al., 2024); preserves provenance by retaining span identifiers and structural paths; expands relevant reference edges by one hop; and orders evidence with attention to positional effects while excluding low-relevance content shown to impair generation (Liu et al., 2024; F. Shi et al., 2023). Evidence assembly also applies the first release gate. When fused relevance mass falls below the calibrated threshold for the request class, control is returned for additional retrieval or the request is terminated through abstention rather than passed directly to generation.

Generation and Verification

Citation-Mandatory Generation

Grounded generation operates under a citation-mandatory contract in which each generated claim is bound to one or more span identifiers from the assembled evidence. Decoding is configured to prioritize evidentiary fidelity, and conflicting evidence is represented as attributed disagreement. The interface accommodates different grounding mechanisms, including instructed frozen models and self-reflective generators (Asai et al., 2024). Multi-agent designs that distribute drafting and verification across specialized roles can be integrated through the same generation interface (Ominyi et al., 2025a).

Claim decomposition

Because verification operates at claim granularity, the architecture specifies how output is atomized. Generation emits a claim-first structure in which each claim is intended to express one proposition with its citations. A decomposition step then checks atomicity, following the atomic-fact approach of Min et al. (2023): conjunctions of independent facts, embedded relative clauses that assert facts, and comparisons are split into separate propositions. For example, "The agreement is governed by New York law and expires on 31 December 2027" becomes two claims, each verified against its own citation. Qualifiers, however, are not split away from the proposition they qualify. Conditions, exceptions, temporal scope, and modality (for example, "shall" versus "may") remain attached, because a legal or technical statement stripped of its qualifier can be verifiable and yet misleading. Each atomic claim inherits its parent's citations and must be supported by at least one of them. Decomposition is logged, and because it is itself an error source on discursive output, its accuracy is measured separately in Phase 1 of the evaluation.

Verification Verdicts and Verifier Limits

Verification is the second release gate. Each claim-citation pair receives one of four verdicts: supported, unsupported, contradicted, or uncheckable. The architecture does not assume that a single entailment model can adjudicate every pair reliably. Existing factual-consistency and natural language inference (NLI) systems are known to degrade on technical, legal, numerical, and temporal content

(Honovich et al., 2022; Laban et al., 2022). The verifier is therefore routed by claim type. Quotation claims are checked by exact match against normalized span text; numeric and quantity claims by extracting, unit-normalizing, and comparing values against typed span or table values; temporal claims additionally against effective-date metadata; and paraphrase claims by an entailment verifier from the faithfulness lineage (Kryscinski et al., 2020; Laban et al., 2022; Honovich et al., 2022) with per-claim-type thresholds calibrated on held-out in-domain data.

The verdict "uncheckable" is retained as an explicit verification outcome. It applies when calibrated verifier confidence lies between acceptance and rejection thresholds, when a claim depends exclusively on low-legibility spans, or when no validated verifier is available for the relevant claim type in the deployment domain. Depending on the policy for the request class, such claims are removed, released with an explicit unverified designation, or referred for human review. Reporting verifier accuracy separately by claim type permits the reliability of the verification gate to be measured rather than assumed.

Additional consistency and uncertainty signals are used when verifier confidence is low, including sampling-based stability (Manakul et al., 2023), semantic-entropy screening (Kuhn et al., 2023; Farquhar et al., 2024), and calibrated self-assessment as a low-cost prior (Kadavath et al., 2022; Guo et al., 2017). When the verifier is itself a language model, known biases in model-based judging, including position and verbosity effects, require fixed rubrics, held-out calibration sets, and periodic human adjudication (Nwakamma et al., 2024a; Ladapo, Jooda, et al., 2024). Disagreement among adjudicators is retained as a diagnostic measure of claim ambiguity and potential verifier drift rather than being eliminated through forced consensus (Ladapo et al., 2026).

Repair Policy

Unsupported, contradicted, and miscited claims trigger differentiated repair actions because regeneration, re-retrieval, citation reassignment, and removal address different failure causes. The repair policy therefore selects an intervention from the diagnosed condition rather than applying undifferentiated resampling. This approach is consistent with corrective-action principles in quality management (Aneke et al., 2022) and with revision methods developed for language-model outputs (Peng et al., 2023; Dhuliawala et al., 2024; Gou et al., 2024; Madaan et al., 2023; L. Gao et al., 2023). Table 5 specifies the resulting decision policy.

Table 5. Repair Decision Policy By Diagnosed Cause

Diagnosed condition	Repair action	Bound	Terminal outcome if repair fails
Unsupported: cited span relevant, but the claim overstates it or adds detail	Constrained regeneration of the claim against the same cited spans, biased toward extractive restatement	One attempt	Claim removed; gap stated
Unsupported: cited spans irrelevant or insufficient	Targeted re-retrieval using the claim as query within the plan scopes, then one constrained regeneration	One retrieval round	Claim removed; gap stated
Miscited: claim supported by a different span in the assembled context	Citation reassignment followed by re-verification	One attempt	Claim removed
Contradicted by the cited span	No regeneration of the same assertion; replaced by extractive restatement of the span, or rendered as attributed conflict where spans disagree	One attempt	Claim removed; contradiction reported
Uncheckable	No automatic repair	None	Removed, marked "unverified", or routed to

Share of removed claims above policy threshold	Response-level conversion	None	human review, per request-class policy Partial answer with stated gaps, or abstention
--	---------------------------	------	--

Note. Repaired claims re-enter verification; a claim that fails verification a second time is removed.

Six constraints limit repair and reduce the risk of circular or cascading errors. Each response has a global repair budget, and each claim may undergo no more than one repair cycle; a repaired claim that fails a second verification is removed. Repair is monotone, meaning that it may remove content or replace it only with content whose support has been verified, and it may not introduce propositions outside the scope of the claim being repaired. Re-retrieval remains within the scopes defined by the original query plan. The orchestrator records a fingerprint of each claim and cited span set and terminates repair when a state recurs. When the proportion of removed claims exceeds a policy threshold, the response is converted to a partial answer with explicit gaps or to abstention rather than being released after substantial evidentiary loss. Every repair action is logged with its diagnosed cause and outcome, and each released claim retains its citations and verification verdict.

Orchestration and Observability

The orchestration layer executes declared policies over the module graph. Policies define iteration budgets, conditions under which verification failures or evidence-sufficiency failures trigger re-planning or abstention, and request-class-specific gate thresholds; active retrieval remains an optional policy mechanism rather than a mandatory loop (Jiang et al., 2023). Policy design therefore formalizes the trade-off among latency, cost, and evidentiary sufficiency, consistent with risk-based approval models and security-orchestration practices (Eze et al., 2024; Dosunmu & Ogundele, 2025b; Oteri & Edviri, 2026). Abstention is represented as an explicit terminal outcome with a defined response schema that records the requested information, retrieved evidence, and reason the evidence failed the release criterion (J. Chen et al., 2024).

Observability is implemented through structured telemetry aligned with documented pipeline failure modes. Each module records plan characteristics, hit distributions, fused relevance scores, reranking deltas, conflict flags, context composition, claim-level verdicts, repair outcomes, and gate decisions keyed by request and span identifiers. These signals permit missing-content failures, ranking errors, and grounding failures to be associated with the stages at which they occur (Barnett et al., 2024). The telemetry design is consistent with enterprise log analytics and security-operations practice (Mbonu et al., 2019; Ogbole et al., 2025; Ladapo, Dosunmu, et al., 2024). The same records support longitudinal evaluation of faithfulness, relevance, and attribution (Es et al., 2024; Saad-Falcon et al., 2024) and aggregate reporting through governed indicators with assigned owners and thresholds (Aliliele et al., 2025a; Dosunmu & Ogundele, 2025a; Edviri & Oteri, 2024; Tonoyan et al., 2024). Versioned module registries record model checkpoints, indexes, and thresholds so that released outputs can be reconstructed under the configuration that produced them (Mbonu et al., 2020; Oshoba et al., 2020; Ladapo et al., 2023).

Positioning and Novelty

The proposed architecture combines established components in a configuration defined by five system-level properties not jointly specified by the comparative frameworks in Table 2. First, immutable span identity is carried through typed interfaces from segmentation to released claims and across document versions. Second, release is conditioned on two explicit gates: evidence sufficiency before generation and claim verification after generation. Third, verification uses a claim-type-specific verdict taxonomy that includes an explicit uncheckable outcome rather than assuming a universal entailment verifier. Fourth, repair is bounded, monotone, and selected according to diagnosed failure cause. Fifth, telemetry is defined as a structured schema linked to documented failure modes, while configuration versioning supports reproducibility of released outputs. These properties constitute the principal architectural contribution; their operational effectiveness remains an empirical question.

Testable Hypotheses

The architecture gives rise to seven hypotheses for empirical evaluation. H1 (segmentation): on clause- and table-heavy corpora, structure-aware segmentation with reference edges yields higher evidence recall and attribution precision than fixed-size and semantic chunking at comparable index size. H2 (verification efficacy): the claim-verification gate reduces the rate of unsupported and contradicted claims in released output relative to the same pipeline without the gate. H3 (abstention trade-off): gate thresholds can be identified at which the reduction specified in H2 is achieved while over-abstention, defined as refusal or claim removal for questions answerable from the corpus, remains below a pre-registered tolerance. H4 (latency and cost): the gated pipeline's 95th-percentile latency and per-request cost remain within pre-registered budgets for each request class. H5 (repair): bounded repair recovers a meaningful proportion of initially failed claims without increasing the contradicted-claim rate. H6 (failure localization): stage telemetry attributes seeded faults to the correct module at a rate substantially above a naive baseline. H7 (modularity): replacement of a module under a fixed interface requires no changes outside that module and preserves contract-level test results. H2, H3, and H4 jointly evaluate whether improved release fidelity can be achieved without unacceptable abstention, latency, or cost.

Evaluation Protocol

Evaluation proceeds in four phases. Phase 1 evaluates modules independently against their specified contracts: segmentation against structure-preservation criteria on document suites containing tables, clause hierarchies, cross-references, footnotes, and scanned pages (H1); retrieval and reranking on in-domain relevance sets, with the hybrid configuration compared against lexical-only and dense-only baselines (Thakur et al., 2021; Robertson & Zaragoza, 2009; Karpukhin et al., 2020); claim decomposition against human atomization; and verifier accuracy by claim type on labeled claim-evidence pairs, including adversarial near-support cases (Honovich et al., 2022; Laban et al., 2022). Annotation procedures report inter-annotator disagreement because measured verifier performance is constrained by human agreement (Ladapo, Jooda, et al., 2024; Ladapo et al., 2026). Module test suites are retained as regression assets to detect contract violations after component replacement (Mbonu et al., 2021; Borah et al., 2022).

Phase 2 evaluates the composed system on grounded document tasks, including question answering and analytical summarization over document collections with reference answers and reference evidence, using task structures derived from knowledge-intensive evaluation (Petroni et al., 2021; Kwiatkowski et al., 2019; Thorne et al., 2018). Reported measures include answer quality, claim-level attribution precision and recall (T. Gao et al., 2023; Rashkin et al., 2023; Min et al., 2023), behavior under noisy, absent, and conflicting evidence (J. Chen et al., 2024), latency and cost by request class (H4), and abstention operating characteristics across gate thresholds (H3). Evaluation tasks are drawn from the targeted document classes, including regulated-document review and operational decision support (Ominyi et al., 2025a; Adelanwa et al., 2024).

Phase 3 evaluates verification and repair efficacy (H2 and H5). Unsupported claims, miscitations, and subtle contradictions are introduced at controlled rates, after which detection, repair success, and residual leakage are measured. This design permits direct estimation of verification performance under deliberately adverse conditions and is consistent with adversarial validation approaches used in security assurance (Dosunmu & Ogundele, 2024a; Dosunmu & Ogundele, 2024b).

Phase 4 evaluates modularity and observability (H6 and H7). Module-replacement studies using alternative embedding models, rerankers, generators, and verifiers measure integration effort and quality change under fixed interfaces. Failure-localization experiments introduce controlled faults, including index gaps, degraded rerankers, and incorrectly configured gates, and assess whether telemetry identifies the responsible module (Barnett et al., 2024; Saad-Falcon et al., 2024). Each injected fault is tracked from detection through remediation verification (Adegbite et al., 2024; Aneke et al., 2022).

Anticipated Applications

Potential applications are concentrated in settings where evidentiary fidelity, provenance, and reviewability are primary requirements. Legal and contract analysis can use clause-level attribution and

explicit conflict representation in a domain where hallucination has been empirically documented as consequential (Dahl et al., 2024), subject to established constraints concerning privilege and confidentiality, algorithmic accountability and trade secrets, authorship of machine-generated text, cross-border privacy, and platform governance (Ominyi et al., 2025c; Annan, 2024; Annan, 2023; Annan, 2022; Ominyi & Anichukwueze, 2022a; Ominyi et al., 2025b). Regulatory and compliance analysis can use structural scoping, effective-date metadata, and cross-document conflict detection for obligations mapping, extending AI-assisted compliance monitoring with span-level attribution (Ominyi et al., 2024a). Relevant regulatory settings include readiness assessment for the European Union's Artificial Intelligence Act, anti-money-laundering and anti-bribery regimes, environmental disclosure obligations, and trade compliance (Ominyi et al., 2024b; Ominyi & Anichukwueze, 2021; Ominyi & Anichukwueze, 2022b; Jones & Ominyi, 2023; Jones & Ominyi, 2024; Jones & Ominyi, 2025).

Audit and financial reporting constitute another relevant application domain because their evidentiary requirements depend on traceable support for assertions (Akomolafe & Agu, 2018; Akomolafe et al., 2025; Ozowara et al., 2025; Agu et al., 2023; Mbonu, Aliliele, et al., 2022; Oyeleye et al., 2025), with corrective action planning as a related downstream process (Ebhojie et al., 2023). Life-science and clinical applications provide a further rationale for calibrated abstention and verification (Singhal et al., 2023), particularly in relation to data integrity, regulatory harmonization, risk-based approval, pharmaceutical traceability, digital health governance, privacy compliance, and operational planning (Aneke et al., 2023; Aneke et al., 2024; Eze et al., 2024; Aneke & Adesemoye, 2019; Afrihyia et al., 2024; Afrihyia et al., 2025; Ekechi et al., 2026; Orise & Niniola, 2025; Akintola et al., 2025; Adelanwa et al., 2024). Technical documentation and governed enterprise assistance are additional candidate domains (Badmus et al., 2025; Shah et al., 2025). Uniform typed interfaces and telemetry also permit component-level research to be conducted within a fixed end-to-end system, supporting more precise attribution of performance changes to individual modules (Y. Gao et al., 2024a; Es et al., 2024).

Limitations

The proposed guarantees are conditional on verifier accuracy, corpus quality, parsing reliability, and calibrated release thresholds. For paraphrase claims, verification depends on model-based judgment; verifier errors may therefore admit unsupported claims or reject valid ones, and adversarial evaluation can estimate but not eliminate residual risk (Honovich et al., 2022; Saad-Falcon et al., 2024). Correlated errors remain possible when one language model evaluates another, so periodic human adjudication is a mitigation rather than a complete solution (Nwakamma et al., 2024a; Ladapo et al., 2026). Attribution establishes consistency with the corpus but does not establish the factual correctness of the source documents, making corpus curation a continuing governance requirement (Oyesiji et al., 2024). Structural parsing also remains dependent on source format and document quality. Hybrid retrieval, reranking, claim-level verification, and repair increase inference cost and latency (Oyesiji et al., 2025; Nwakamma et al., 2025), and conservative release thresholds may increase over-abstention; these trade-offs are explicitly evaluated by H3 and H4. Claim decomposition also remains an independent source of error in discursive output (Min et al., 2023).

The architecture constrains the evidence-handling pipeline but does not directly eliminate vulnerabilities in the underlying models or surrounding execution environment. Prompt-borne adversarial content, tool poisoning, memory manipulation, and excessive agency remain outside the scope of the verification gate (Nwakamma et al., 2024b); complementary controls include adversarial machine learning, secure software engineering, generative-AI pipeline security, and zero-trust architectures (Adebayo et al., 2022; Odejobi et al., 2025; Adegbite et al., 2026; Ahmed et al., 2021). Data-protection obligations apply to the corpus and telemetry as well as to generated outputs (Mbonu, Iwuanyanwu, et al., 2022; Mbonu et al., 2018), and the proposed architecture does not itself discharge legal duties concerning automated decision-making, privilege, or institutional governance (Annan, 2025; Ominyi et al., 2025c; Ominyi et al., 2026). Regulated deployments also retain domain-specific human-review obligations (Dahl et al., 2024; Huang et al., 2025; Ladapo et al., 2022). The study is conceptual, its evidence base is a narrative rather than systematic synthesis, and the proposed mechanisms have not yet been implemented or empirically evaluated.

CONCLUSION

The study specifies a conceptual RAG architecture in which grounding is defined through system-level controls rather than prompt-dependent behavior. The resulting contribution comprises eight requirements derived from documented failure evidence, a comparative assessment of representative modular, self-reflective, corrective, iterative, and agentic RAG approaches, and a reference architecture based on immutable span identity, evidence-sufficiency gating, claim-level verification, claim-type-specific verdicts, bounded repair, and stage-level telemetry. The architecture also defines seven hypotheses and a four-phase evaluation protocol, thereby converting its principal design claims into empirically testable propositions.

The conclusions are limited to architectural specification and analytical coherence. No implementation or empirical performance results are reported, so the effectiveness of span-level attribution, conflict transparency, calibrated abstention, repair, and stage-level telemetry remains unverified. The evidentiary base is a narrative synthesis whose coverage is indicative rather than exhaustive, and the comparative assessment reflects the mechanisms documented in the selected framework specifications.

Subsequent research should implement the reference architecture and evaluate it on legal, regulatory, audit, clinical, and technical-document corpora. Priority tests concern whether verification gates reduce unsupported and contradicted claims (H2) without excessive over-abstention (H3) or unacceptable latency and cost (H4), whether structure-aware segmentation improves attribution relative to conventional chunking (H1), whether bounded repair recovers failed claims without increasing contradiction rates (H5), and whether stage telemetry correctly localizes controlled faults (H6). Empirical results from these tests are required before claims of production-level grounding reliability can be sustained.

REFERENCES

- Adebayo, A., Adegbite, M. P., & Ahmed, M. O. (2022). Adversarial machine learning in critical infrastructure: A conceptual framework for threat modeling AI enabled OT systems. *World Journal of Innovation and Modern Technology*, 6(1), 184–234. <https://doi.org/10.56201/wjimt.v6.no1.2022.pg184.234>
- Adegbite, M. P., Adebayo, A., & Ahmed, M. O. (2024). A vulnerability governance architecture for power and utilities corporations: From exposure mapping to remediation verification. *World Journal of Innovation and Modern Technology*, 8(6), 185–246.
- Adegbite, M. P., Ahmed, M. O., & Adebayo, A. (2026). Securing generative AI pipelines in critical infrastructure environments: A conceptual architecture for identity, data integrity, and adversarial risk. *Advances in IT and Digital Security*, 1(2), 1–46.
- Adelanwa, A., Basnet, A., & Anene, U. N. (2024). Advanced AI based decision support systems for healthcare operations and resource planning. *Gyanshauryam, International Scientific Refereed Research Journal*, 7(1), 244–276.
- Afrihyia, E., Akinse, S. G., & Ojukwu, P. U. (2024). Comparative governance of AI-driven healthcare management: Executive oversight, regulatory structures, and accountability in the United States and developing countries. *International Journal of Advanced Multidisciplinary Research and Studies*, 4(6), 3087–3102. <https://doi.org/10.62225/2583049X.2024.4.6.5920>
- Afrihyia, E., Akinse, S. G., & Ojukwu, P. U. (2025). Organizational readiness for generative AI integration in healthcare operations: Comparative management capabilities between the U.S. and low- and middle-income countries. *Iconic Research and Engineering Journals*, 8(10), 1673–1697. <https://doi.org/10.64388/IREV8I10-1714670>
- Agu, M. U., Akomolafe, O., & Bello, A. (2023). A comparative review of SOX compliance frameworks in cross-border financial auditing. *International Journal of Advanced Multidisciplinary Research and Studies*, 3(6), 2297–2306. <https://doi.org/10.62225/2583049X.2023.3.6.5362>
- Ahmed, M. O., Adegbite, M. P., & Adebayo, A. (2021). Zero trust architecture for operational technology in North American critical infrastructure: A framework for implementation and resilience optimization. *International Journal of Engineering and Modern Technology*, 7(1), 66–113.
- Akintola, A. S., Fawehinmi, Y. O., Aiyenitaju, O., Chinemerem, B., & Emmanuella, O. (2025). Digital transformation in telehealth: A systematic review of user trust, privacy, and regulatory governance

- in AI-powered remote monitoring systems. *Journal of Scientific Research and Reports*, 31(11), 163–182. <https://doi.org/10.9734/jsrr/2025/v31i113658>
- Akomolafe, O., & Agu, M. U. (2018). A conceptual model for enhancing internal audit quality through technology-enabled risk assessment frameworks. *Iconic Research and Engineering Journals*, 1(9), 458–475.
- Akomolafe, O., Agu, M. U., & Bello, A. (2025). A conceptual model for advancing risk governance through data-driven compliance analytics in financial institutions. *Journal of Accounting and Financial Management*, 11(11), 211–228. <https://doi.org/10.56201/jafm.vol.11.no11.2025.pg211.228>
- Aliliele, C., Mbonu, I. S., & Iwuanyanwu, U. (2023). A review of API governance and risk prioritization frameworks in modern financial institutions. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 9(10), 395–433. <https://doi.org/10.32628/CSEIT2361072>
- Aliliele, C., Mbonu, I. S., & Iwuanyanwu, U. (2024). A conceptual framework for enterprise data sensitivity classification and regulatory traceability mechanisms. *International Journal of Advanced Multidisciplinary Research and Studies*, 4(6), 3103–3124. <https://doi.org/10.62225/2583049X.2024.4.6.5991>
- Aliliele, C., Mbonu, I. S., Uzoka, E., & Iwuanyanwu, U. (2025a). A review of AI assisted continuous auditing systems in technology risk and cybersecurity oversight. *Gyanshauryam, International Scientific Refereed Research Journal*, 8(4), 210–250. <https://doi.org/10.32628/GISRRJ258369>
- Aliliele, C., Mbonu, I. S., Uzoka, E., & Iwuanyanwu, U. (2025b). Advances in data lakehouse governance architectures for enterprise data loss prevention and compliance assurance. *Shodhshauryam, International Scientific Refereed Research Journal*, 8(4), 171–213. <https://doi.org/10.32628/SHISRRJ258474>
- Aneke, O. B., & Adesemoye, A. C. (2019). Toward an integrated conceptual framework for traceability and chain-of-custody integrity in pharmaceutical distribution. *Iconic Research and Engineering Journals*, 3(4), 655–672.
- Aneke, O. B., Adesemoye, A. C., & Sydney, E. (2022). Corrective and preventive action effectiveness in pharmaceutical quality management systems: A systematic review of root-cause methodologies. *International Journal of Health and Pharmaceutical Research*, 7(2), 78–111.
- Aneke, O. B., Adesemoye, A. C., & Sydney, E. (2024). Regulatory harmonization and convergence in global medicine quality assurance: A review of frameworks, pathways, and implementation barriers. *International Journal of Medical Evaluation and Physical Report*, 8(6), 213–252.
- Aneke, O. B., Sydney, E., & Adesemoye, A. C. (2023). Data integrity and governance in regulated life sciences operations: A review of failures, controls, and remediation strategies. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 9(10), 538–568.
- Annan, A. O. (2022). Data privacy governance models for cross-border digital platforms. *International Journal of Multidisciplinary Research and Growth Evaluation*, 3(6), 949–964.
- Annan, A. O. (2023). Intellectual property ownership and authorship in AI-generated works. *Gyanshauryam, International Scientific Refereed Research Journal*, 6(3), 425–450.
- Annan, A. O. (2024). Algorithmic accountability and trade secret protection in artificial intelligence. *Shodhshauryam, International Scientific Refereed Research Journal*, 7(5), 315–347.
- Annan, A. O. (2025). Automated decision-making and anti-discrimination compliance under U.S. law. *International Journal of Advanced Multidisciplinary Research and Studies*, 5(2), 2522–2540.
- Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *Proceedings of the International Conference on Learning Representations*.
- Badmus, O., Dosunmu, A. A., & Anunagba, C. O. (2025). A governance framework for AI-assisted CRM workflows: Agentforce deployment, risk management, and organizational readiness in enterprise Salesforce environments. *International Journal of Scientific Research in Humanities and Social Sciences*, 2(1), 59–80.
- Badmus, O., Jooda, D., & Anunagba, C. O. (2024). A privacy-by-design framework for role-based security architecture in Salesforce environments handling sensitive personal data. *International*

- Journal of Advanced Multidisciplinary Research and Studies, 4(6), 3244–3254. <https://doi.org/10.62225/2583049X.2024.4.6.6243>
- Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven failure points when engineering a retrieval augmented generation system. Proceedings of the IEEE/ACM International Conference on AI Engineering, 194–199. <https://doi.org/10.1145/3644815.3644945>
- Borah, S., Aliliele, K. C., Rakshit, S., & Vajjhala, N. R. (2022). Applications of artificial intelligence in software testing. In P. K. Mallick, A. K. Bhoi, P. Barsocchi, & V. H. C. de Albuquerque (Eds.), Cognitive informatics and soft computing (Lecture Notes in Networks and Systems, Vol. 375, pp. 727–736). Springer. https://doi.org/10.1007/978-981-16-8763-1_60
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., ... Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. Proceedings of the International Conference on Machine Learning, 2206–2240.
- Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1870–1879. <https://doi.org/10.18653/v1/P17-1171>
- Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. Proceedings of the AAAI Conference on Artificial Intelligence, 38(16), 17754–17762. <https://doi.org/10.1609/aaai.v38i16.29728>
- 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>
- Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-verification reduces hallucination in large language models. Findings of the Association for Computational Linguistics: ACL 2024, 3563–3578. <https://doi.org/10.18653/v1/2024.findings-acl.212>
- Dosunmu, A. A., & Ogundele, P. O. (2024a). Breach and attack simulation frameworks for continuous validation of enterprise security controls. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(3), 1100–1119.
- Dosunmu, A. A., & Ogundele, P. O. (2024b). Threat informed defence engineering models for measuring security control effectiveness at scale. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 2847–2858.
- Dosunmu, A. A., & Ogundele, P. O. (2025a). Cyber defense performance measurement frameworks for executive and board level governance. Computer Science and IT Research Journal, 6(11), 895–913. <https://doi.org/10.51594/csitrj.v6i11.2164>
- Dosunmu, A. A., & Ogundele, P. O. (2025b). Security orchestration and automation models for accelerating incident detection and response. Computer Science and IT Research Journal, 6(11), 878–894. <https://doi.org/10.51594/csitrj.v6i11.2163>
- Ebhojie, O., Dogbatsey, E. A., & Oyeleye, A. O. (2023). Audit liaison, corrective action planning, and control compliance in multinational organisations: A systematic literature review. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2839–2850. <https://doi.org/10.62225/2583049X.2023.3.6.6200>
- Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2024). From local to global: A graph RAG approach to query-focused summarization. arXiv. <https://doi.org/10.48550/arXiv.2404.16130>
- Edivri, J., & Oteri, O. (2024). A conceptual KPI-driven decision and optimization framework for IT service delivery, portfolio performance, and adoption. Gyanshauryam, International Scientific Refereed Research Journal, 7(1), 167–187.
- Ekechi, N. V., Ozowara, D. E., & Anunagba, C. O. (2026). Conceptual framework for AI governance, data privacy compliance, and financial sustainability in digital health. Computer Science and IT Research Journal, 7(4), 275–299.
- Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. Proceedings of the Conference of the European Chapter of the

- Association for Computational Linguistics: System Demonstrations, 150–158. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- Eze, F. I., Akinleye, O. K., & Anene, U. N. (2024). Development of a risk-based accelerated approval decision model balancing speed, safety, and evidence quality in pharmaceutical regulatory science. *Research Journal of Pure Science and Technology*, 7(6), 120–161. <https://doi.org/10.56201/rjpst.v7.no6.2024.pg120.161>
- Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017), 625–630. <https://doi.org/10.1038/s41586-024-07421-0>
- Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V. Y., Lao, N., Lee, H., Juan, D.-C., & Guu, K. (2023). RARR: Researching and revising what language models say, using language models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 16477–16508. <https://doi.org/10.18653/v1/2023.acl-long.910>
- Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 6465–6488. <https://doi.org/10.18653/v1/2023.emnlp-main.398>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024a). Retrieval-augmented generation for large language models: A survey. *arXiv*. <https://doi.org/10.48550/arXiv.2312.10997>
- Gao, Y., Xiong, Y., Wang, M., & Wang, H. (2024b). Modular RAG: Transforming RAG systems into LEGO-like reconfigurable frameworks. *arXiv*. <https://doi.org/10.48550/arXiv.2407.21059>
- Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., & Chen, W. (2024). CRITIC: Large language models can self-correct with tool-interactive critiquing. *Proceedings of the International Conference on Learning Representations*.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the International Conference on Machine Learning*, 1321–1330.
- Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). Retrieval augmented language model pre-training. *Proceedings of the International Conference on Machine Learning*, 3929–3938.
- 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>
- Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A., & Matias, Y. (2022). TRUE: Re-evaluating factual consistency evaluation. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 3905–3920. <https://doi.org/10.18653/v1/2022.naacl-main.287>
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), Article 42. <https://doi.org/10.1145/3703155>
- Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 874–880. <https://doi.org/10.18653/v1/2021.eacl-main.74>
- Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, 10(1–2), 18–26. <https://doi.org/10.1007/s13162-020-00161-0>
- 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), Article 248. <https://doi.org/10.1145/3571730>
- Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). Active retrieval augmented generation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 7969–7992. <https://doi.org/10.18653/v1/2023.emnlp-main.495>
- Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547. <https://doi.org/10.1109/TBDATA.2019.2921572>
- Jones, B. C., & Ominyi, M. (2023). Reviewing compliance frameworks for technology transfers between developed and emerging markets in an era of export controls and sanctions. *World Journal of Innovation and Modern Technology*, 7(2), 121–149.

- Jones, B. C., & Ominyi, M. (2024). Applications of artificial intelligence in customs administration and trade facilitation: A review. *World Journal of Innovation and Modern Technology*, 8(6), 181–217. <https://doi.org/10.56201/wjimt.v8.no6.2024.pg181.217>
- Jones, B. C., & Ominyi, M. (2025). Conceptualizing AI-driven advisory services for expanding trade compliance access to small and medium enterprises. *IIARD International Journal of Economics and Business Management*, 11(12), 441–474. <https://doi.org/10.56201/ijebm.vol.11.no12.2025.pg441.474>
- Kadavath, S., Conerly, T., Askeel, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., ... Kaplan, J. (2022). Language models (mostly) know what they know. arXiv. <https://doi.org/10.48550/arXiv.2207.05221>
- Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. *Proceedings of the International Conference on Machine Learning*, 15696–15707.
- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 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 International ACM SIGIR Conference on Research and Development in Information Retrieval*, 39–48. <https://doi.org/10.1145/3397271.3401075>
- Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 9332–9346. <https://doi.org/10.18653/v1/2020.emnlp-main.750>
- Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Proceedings of the International Conference on Learning Representations*.
- Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7, 452–466. https://doi.org/10.1162/tacl_a_00276
- Laban, P., Schnabel, T., Bennett, P. N., & Hearst, M. A. (2022). SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10, 163–177. https://doi.org/10.1162/tacl_a_00453
- Ladapo, O. O., Dosunmu, A. A., Jooda, D., & Abolaji, T. O. (2022). Human-in-the-loop machine learning: A state of the art. *Journal of Frontiers in Multidisciplinary Research*, 3(1), 656–669. <https://doi.org/10.54660/JFMR.2022.3.1.656-669>
- Ladapo, O. O., Dosunmu, A. A., Jooda, D., & Abolaji, T. O. (2023). An IT service management literature review: Challenges, benefits, opportunities, and implementation practices. *International Journal of Advanced Multidisciplinary Research and Studies*, 3(6), 2875–2889. <https://doi.org/10.62225/2583049X.2023.3.6.6205>
- Ladapo, O. O., Dosunmu, A. A., Jooda, D., & Abolaji, T. O. (2024). Integrated network and security operation center: A systematic analysis. *International Journal of Multidisciplinary Futuristic Development*, 5(1), 65–80. <https://doi.org/10.54660/IJMFD.2024.5.1.65-80>
- Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2024). Keeping humans in the loop: Human-centered automated annotation with generative AI. *International Journal of Multidisciplinary Futuristic Development*, 5(1), 81–95. <https://doi.org/10.54660/IJMFD.2024.5.1.81-95>
- Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2026). On the disagreement problem in human-in-the-loop federated machine learning. *International Journal of Multidisciplinary Research and Growth Evaluation*, 7(3), 178–192. <https://doi.org/10.54660/IJMRGE.2026.7.3.178-192>
- Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kocisky, T., Ruder, S., Yogatama, D., Cao, K., Young, S., & Blunsom, P.

- (2021). Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34, 29348–29363.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173. https://doi.org/10.1162/tacl_a_00638
- Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., & Singh, S. (2021). Entity-based knowledge conflicts in question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 7052–7063. <https://doi.org/10.18653/v1/2021.emnlp-main.565>
- Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). Query rewriting in retrieval-augmented large language models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 5303–5315. <https://doi.org/10.18653/v1/2023.emnlp-main.322>
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-Refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 46534–46594.
- Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4), 824–836. <https://doi.org/10.1109/TPAMI.2018.2889473>
- Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 9802–9822. <https://doi.org/10.18653/v1/2023.acl-long.546>
- Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 9004–9017. <https://doi.org/10.18653/v1/2023.emnlp-main.557>
- Mbonu, I. S., Aliliele, C., Iwuanyanwu, U., & Oluoha, O. M. (2018). A conceptual framework for legal and ethical risk modeling in enterprise data protection governance systems. *Iconic Research and Engineering Journals*, 2(2), 207–226. <https://doi.org/10.64388/IREV2I2-1714911>
- Mbonu, I. S., Aliliele, C., Iwuanyanwu, U., & Uzoka, E. (2021). Advances in artificial intelligence techniques for secure software testing and automated regression control mechanisms. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 7(5), 468–496. <https://doi.org/10.32628/CSEIT217565>
- Mbonu, I. S., Aliliele, C., Iwuanyanwu, U., & Uzoka, E. (2022). A conceptual framework for AI enabled IT general controls and SOX audit automation processes. *Gyanshauryam, International Scientific Refereed Research Journal*, 5(5), 384–414. <https://doi.org/10.32628/GISRRJ2256239>
- Mbonu, I. S., Iwuanyanwu, U., Aliliele, C., & Uzoka, E. (2020). Advances in infrastructure as code governance for secure terraform based enterprise cloud deployments. *International Journal of Multidisciplinary Research and Growth Evaluation*, 1(5), 811–828. <https://doi.org/10.54660/IJMRGE.2020.1.5.811-828>
- Mbonu, I. S., Iwuanyanwu, U., Aliliele, C., & Uzoka, E. (2022). A review of data protection impact assessment models in multi cloud financial infrastructure systems. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 8(1), 589–623. <https://doi.org/10.32628/CSEIT25442>
- Mbonu, I. S., Iwuanyanwu, U., Uzoka, E., & Oluoha, O. M. (2019). Advances in enterprise log analytics and automated incident response architectures using Python and SIEM platforms. *Iconic Research and Engineering Journals*, 3(2), 1000–1019. <https://doi.org/10.64388/IREV3I2-1714915>
- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 12076–12100. <https://doi.org/10.18653/v1/2023.emnlp-main.741>

- Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. arXiv. <https://doi.org/10.48550/arXiv.1901.04085>
- Nwakamma, S., Ojukwu, J., & Oyesiji, S. O. (2024a). LLM-as-a-judge for automated model governance: A review of methods, biases, and release-gating practices. *World Journal of Innovation and Modern Technology*, 8(6), 185–216. <https://doi.org/10.56201/wjimt.v8.no6.2024.pg185.216>
- Nwakamma, S., Ojukwu, J., & Oyesiji, S. O. (2024b). Securing agentic AI enterprise workflows against prompt injection, tool poisoning, memory manipulation, and excessive agency threats. *World Journal of Innovation and Modern Technology*, 8(6), 217–248. <https://doi.org/10.56201/wjimt.v8.no6.2024.pg217.248>
- Nwakamma, S., Ojukwu, J., & Oyesiji, S. O. (2025). On-device multimodal small language models for privacy-preserving edge intelligence through quantization, pruning, distillation, and energy-efficient inference. *World Journal of Innovation and Modern Technology*, 9(12), 271–302. <https://doi.org/10.56201/wjimt.v9.no12.2025.pg271.302>
- Odejobi, O. D., Okonkwo, C. S., Ahiaekwe Patrick, M. C., Okeke, O. T., & Mayo, W. (2025). AI-augmented secure software engineering: Leveraging deep learning for autonomous threat detection and mitigation. *International Journal of Engineering and Modern Technology*, 11(12), 101–121.
- Ogbole, J. I., Okoruwa, P. O., Fadayomi, O., Akeju, B., Edivri, J., & Abolaji, T. O. (2025). Security analytics and digital forensics for enterprise risk management, advances and practical implications. *International Journal of Advanced Multidisciplinary Research and Studies*, 5(6), 2017–2028.
- Ojukwu, J., Oyesiji, S. O., & Nwakamma, S. (2025). WebAssembly component model and WASI for portable cloud-to-edge software: Language interoperability, security, performance, and container comparisons. *International Journal of Computer Science and Mathematical Theory*, 11(12), 232–263. <https://doi.org/10.56201/ijcsmt.vol.11.no12.2025.pg232.263>
- Ominyi, M., & Anichukwueze, C. C. (2021). A systematic review of anti-money laundering and anti-bribery compliance regimes across emerging and developed markets. *International Journal of Economics and Financial Management*, 6(2), 89–118. <https://doi.org/10.56201/ijefm.v6.no2.2021.pg89.118>
- Ominyi, M., & Anichukwueze, C. C. (2022a). A conceptual framework for harmonizing data protection compliance across GDPR, NDPR, and United States privacy regimes. *International Journal of Social Sciences and Management Research*, 8(5), 148–174. <https://doi.org/10.56201/ijssmr.v8.no5.2022.pg148.174>
- Ominyi, M., & Anichukwueze, C. C. (2022b). Reviewing greenhouse gas disclosure and ESG compliance obligations for energy and infrastructure enterprises under evolving regulatory regimes. *International Journal of Economics and Financial Management*, 7(6), 94–124. <https://doi.org/10.56201/ijefm.v7.no6.2022.pg94.124>
- Ominyi, M., & Anichukwueze, C. C. (2023). A review of algorithmic accountability and model risk management approaches in financial services. *Journal of Accounting and Financial Management*, 9(12), 218–238. <https://doi.org/10.56201/jafm.v9.no12.2023.pg218.238>
- Ominyi, M., & Anichukwueze, C. C. (2024). Conceptualizing responsible AI governance structures for automated decision systems in regulated industries. *International Journal of Social Sciences and Management Research*, 10(11), 403–430. <https://doi.org/10.56201/ijssmr.v10.no11.2024.pg.403.430>
- Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2024a). Developing a conceptual framework for AI-assisted compliance monitoring and anomaly detection in high-volume transaction environments. *World Journal of Innovation and Modern Technology*, 8(6), 204–228. <https://doi.org/10.56201/wjimt.v8.no6.2024.pg204.228>
- Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2024b). A systematic review of organizational readiness for the EU Artificial Intelligence Act in multinational enterprises. *World Journal of Innovation and Modern Technology*, 8(6), 185–203. <https://doi.org/10.56201/wjimt.v8.no6.2024.pg185.203>
- Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2025a). A conceptual framework for multi-agent AI quality control in the review of regulated documents. *International Journal of Social Sciences and Management Research*, 11(8), 534–557. <https://doi.org/10.56201/ijssmr.vol.11no8.2025.pg534.557>

- Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2025b). A review of intermediary liability and content governance obligations for digital platforms under the DSA, DMCA, and CDA regimes. *International Journal of Social Sciences and Management Research*, 11(8), 558–582. <https://doi.org/10.56201/ijssmr.vol.11no8.2025.pg558.582>
- Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2025c). Reviewing legal privilege and confidentiality challenges in enterprise deployment of large language models. *Journal of Law and Global Policy*, 10(3), 150–176. <https://doi.org/10.56201/jlgp.vol.10.no3.2025.pg150.176>
- Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2026). Conceptualizing enterprise AI governance functions as institutional standards for auditable and defensible risk management. *Journal of Accounting and Financial Management*, 12(7), 326–356. <https://doi.org/10.56201/jafm.vol.12.no7.2026.p326.356>
- Orise, C., & Niniola, S. (2025). Data governance in digital health learning systems: A systematic review of implementation frameworks and best practices. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 11(5), 520–538.
- Oshoba, T. O., Hammed, N. I., & Odejobi, O. D. (2020). Blockchain-enabled compliance and audit trail model for cloud configuration management. *Journal of Frontiers in Multidisciplinary Research*, 1(1), 193–201.
- Oteri, O., & Edivri, J. (2026). An operational reliability and service assurance framework for enterprise IT systems supporting large user populations. *Gulf Journal of Engineering & Technology*, 2(1), 1–19. <https://doi.org/10.51594/gjet.v2i1.205>
- Oyeleye, A. O., Dogbatsey, E. A., & Ebhojie, O. (2025). Embedding automated close controls in public sector ERP systems: A conceptual model for audit readiness and financial reporting quality. Zenodo. <https://doi.org/10.5281/zenodo.20097950>
- Oyesiji, S. O., Nwakamma, S., & Ojukwu, J. (2023). Parameter-efficient multilingual adaptation of on-device language models: A survey. *International Journal of Engineering and Modern Technology*, 9(3), 287–320. <https://doi.org/10.56201/ijcsmt.v9.no3.2023.pg287.320>
- Oyesiji, S. O., Nwakamma, S., & Ojukwu, J. (2024). A conceptual framework for context-aware data curation in fine-tuning pipelines. *International Journal of Engineering and Modern Technology*, 10(11), 197–228. <https://doi.org/10.56201/ijemt.v10.no11.2024.pg197.228>
- Oyesiji, S. O., Nwakamma, S., & Ojukwu, J. (2025). Compressing recommender and ranking models for edge deployment: A survey. *International Journal of Engineering and Modern Technology*, 11(12), 205–237. <https://doi.org/10.56201/ijemt.vol.11.no12.2025.pg205.237>
- Ozowara, D. E., Adebayo, A., & Anunagba, C. O. (2025). Advanced conceptual model for strengthening audit quality using data analytics across financial institutions. *International Journal of Advanced Multidisciplinary Research and Studies*, 5(6), 2246–2268. <https://doi.org/10.62225/2583049X.2025.5.6.6050>
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., & Gao, J. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. arXiv. <https://doi.org/10.48550/arXiv.2302.12813>
- Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., & Riedel, S. (2021). KILT: A benchmark for knowledge intensive language tasks. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2523–2544. <https://doi.org/10.18653/v1/2021.naacl-main.200>
- Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC Methods Programme. Lancaster University.
- Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A., & Lewis, M. (2023). Measuring and narrowing the compositionality gap in language models. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5687–5711. <https://doi.org/10.18653/v1/2023.findings-emnlp.378>

- Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11, 1316–1331. https://doi.org/10.1162/tacl_a_00605
- 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/10.1561/15000000019>
- Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 338–354. <https://doi.org/10.18653/v1/2024.naacl-long.20>
- Shah, R., Oziri, S. T., & Seyi-Lande, O. B. (2025). A framework for leveraging artificial intelligence in strategic business decision-making. *Gulf Journal of Advance Business Research*, 3(11), 1517–1558. <https://doi.org/10.51594/gjabr.v3i11.171>
- Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., Schärli, N., & Zhou, D. (2023). Large language models can be easily distracted by irrelevant context. *Proceedings of the International Conference on Machine Learning*, 31210–31227.
- Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., & Yih, W.-t. (2023). REPLUG: Retrieval-augmented black-box language models. *arXiv*. <https://doi.org/10.48550/arXiv.2301.12652>
- Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803. <https://doi.org/10.18653/v1/2021.findings-emnlp.320>
- Singh, A., Ehtesham, A., Kumar, S., & Talaei Khoei, T. (2025). Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv*. <https://doi.org/10.48550/arXiv.2501.09136>
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., ... Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180. <https://doi.org/10.1038/s41586-023-06291-2>
- Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *Proceedings of the NeurIPS Datasets and Benchmarks Track*.
- Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 809–819. <https://doi.org/10.18653/v1/N18-1074>
- Tonoyan, A., Dada, O., & Ayivi-Donkor, S. S. (2024). Real-time KPI tracking systems: A review of automated performance monitoring and data-driven decision making. *World Journal of Innovation and Modern Technology*, 8(6), 247–281.
- Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2023). Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 10014–10037. <https://doi.org/10.18653/v1/2023.acl-long.557>
- Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems*, 25(1), 77–89. <https://doi.org/10.1057/ejis.2014.36>
- Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering*, Article 38. <https://doi.org/10.1145/2601248.2601268>
- Xie, J., Zhang, K., Chen, J., Lou, R., & Su, Y. (2024). Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *Proceedings of the International Conference on Learning Representations*.

- Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024). Corrective retrieval augmented generation. arXiv. <https://doi.org/10.48550/arXiv.2401.15884>
- Yoran, O., Wolfson, T., Ram, O., & Berant, J. (2024). Making retrieval-augmented language models robust to irrelevant context. Proceedings of the International Conference on Learning Representations.