International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(23s) 2026 pp. 1449-1460
https://doi.org/10.70917/ijcisim-2026-5704
Article

# Rules-Governed Retrieval-Augmented Generation for Enterprise Quality Assurance: Production Deployment and Measured Outcomes in Automated SAP Test-Case Mapping

**Vijay Kumar Kola**

Quality Assurance Centre of Excellence, Osmania University, Hyderabad, India

**Abstract:** Manual Object-to-Test-Case Mapping (OTM) in large-scale SAP enterprise change-management environments imposes a recurring cost equivalent to one to two analyst positions per year, with quality outcomes that vary by individual rather than converging on a repeatable, auditable standard. Three structural failure modes-knowledge loss, record incompleteness, and systematic regression-scope gaps-are endemic to manual OTM production and resistant to training-based remediation.

**Objective**: This article presents AIMO (AI Impact Mapping for OTM Objects), a rules-governed Retrieval-Augmented Generation (RAG) system designed, built, and deployed in production within a Quality Assurance Centre of Excellence to automate OTM production while satisfying the governance and audit-traceability requirements of compliance-governed SAP change management.

**System**: AIMO pairs an eight-step RAG pipeline with eight deterministic organizational business rules that take precedence over model output at every processing stage, converting retrieval-grounded generation from an efficiency intervention into a compliance instrument. The system has been deployed in production, processing live Change Requests across multiple SAP modules including SD, MM, FICO, TMS, and SCM.

**Results**: Production deployment measurements demonstrate an 84.6% reduction in analyst effort per Change Request (from approximately 156 to 24 analyst-minutes), greater than 90% historical test-case reuse from verifiable organizational records, 100% regression-scope coverage of applicable objects on every processed Change Request, and an external API cost below USD $0.03 per Change Request. Across the full deployment period, AIMO has delivered documented cost savings exceeding USD $1.38 million annually in recovered analyst capacity and defect-leakage prevention, with an operational cost of approximately USD $24,300 per year-a return of USD 56 per dollar invested.

**Conclusions**: AIMO demonstrates that a deterministic, version-controlled rules layer-not model capability-is the decisive architectural element for compliance-governed enterprise AI deployment. The governance properties are model-agnostic, independently auditable, and measurable as quality outcomes. The production deployment provides replicable evidence that the rules-governed RAG pattern is instantiable in live regulated enterprise environments at commercially viable cost.

**Keywords:** Retrieval-augmented generation, enterprise quality assurance, SAP change management, production AI deployment, test-case automation, rules-governed AI, large language models, regression coverage, audit trail, compliance governance, cost-benefit analysis

## 1. INTRODUCTION

The Object-to-Test-Case Mapping (OTM) is the compliance artifact that connects an SAP Change Request-a structured list of modified system objects submitted for testing and deployment approval-to the body of historical test evidence applicable to each object. Every test assignment in an OTM must be traceable to an organizational test record, and the complete set of assignments must encompass both the direct test cases for modified objects and the regression-relevant tests for objects whose associated functionality may be affected as a side effect of the change. This traceability requirement is the operational expression of the process-repeatability obligation that compliance frameworks such as

SOX and GXP impose on the change-control function: not a procedural formality, but a substantive audit-evidence requirement [3][4].

The structural problem with manual OTM production is information architecture, not analyst competence. Historical test records in enterprise SAP environments accumulate across service-management platforms, project file archives, informal documentation channels, and institutional email, in formats that resist systematic search and in locations that require organizational knowledge to navigate [16]. The analyst who consistently produces a complete, well-evidenced OTM is one who has built, over years of practice, an internal mental index of where the relevant records are. That index is not transferable, not documentable, and not reproducible-which is why staff transitions, team growth, and the passage of time all produce knowledge-loss as a structural outcome rather than an exceptional event.

The regression-scope failure mode has a different character but the same root. The regression reference file-the organizational catalog mapping SAP objects to their cross-module dependency test requirements-exists as a separate artifact from the direct test records. Consulting it during change preparation requires a deliberate additional retrieval step that the time pressure of operational change review systematically discourages. Over release cycles and Change Requests, this rational prioritization produces systematic regression-coverage gaps that surface as production defects, each defect carrying a remediation cost that dwarfs the minutes of retrieval effort that would have prevented it [17][21].

An ungrounded large language model does not address either failure mode. Its training distribution contains no information about the organization's historical test records, its specific SAP module configuration, or the exact test-case identifiers that constitute a valid OTM entry in this compliance framework. A plausible-sounding assignment that is not traceable to an actual organizational test record is an audit failure-functionally indistinguishable from a correct entry until a defect surfaces in production [1][14]. Retrieval-Augmented Generation resolves the grounding problem by anchoring generation in retrieved organizational records [1]. The remaining architectural question AIMO resolves is whether grounded generation alone is sufficient for compliance deployment, or whether an additional deterministic governance layer is required.

AIMO's answer, reflected in its deployed architecture, is that retrieval grounding alone is insufficient. A grounded system imposing no constraints on which objects enter processing, which results are surfaced, or what the system outputs when retrieval fails is not a controlled process-and the compliance obligation is process control, not output quality on average [5][6]. AIMO therefore treats the language model as a bounded rationale-generation service and encodes all governance decisions in a deterministic rules layer that the organization owns, version-controls, and can audit independently of the model.

**Contributions and scope:** This article reports (1) the complete architecture of the AIMO rules-governed RAG system; (2) a data-boundary design enabling LLM API integration in regulated environments without transmission of proprietary compliance data; (3) production deployment measurements demonstrating an 84.6% effort reduction, greater than 90% historical reuse, and 100% regression coverage; (4) a full cost-benefit analysis of production deployment across an annual cycle; and (5) a rigorous analysis of limitations and threats to validity. Unlike prior work reporting proof-of-concept or simulation results, the measurements reported here are drawn from live production deployments against real Change Requests in an operational SAP environment.

The article is organized as follows. Section II reviews related literature and identifies the gap AIMO addresses. Section III presents the AIMO architecture. Section IV reports production measurements and the cost-benefit analysis. Section V discusses implications, limitations, and threats to validity. Section VI concludes.

## 2. BACKGROUND AND RELATED WORK

### *A. Test-Case Mapping and Regression Testing in Enterprise Change Control*

Test documentation in enterprise change management is governed by traceability requirements rooted in IEEE Std 829-2008 [3], which mandates linkage of test design and execution artifacts to their test basis. ISO/IEC 25010:2011 [4] situates this requirement within a quality model that evaluates OTM quality not as individual mapping accuracy but as consistency and completeness across analysts, release cycles, and project teams. Sakhrawi and Labidi [15] establish that ontology-based retrieval over a structured test-case repository reduces selection time and improves coverage completeness-the precondition that AIMO exploits: a consolidated, structured historical test repository enables automated selection where fragmented storage does not. Göçmen et al. [17] show that automated change-impact analysis measurably improves coverage of regression-relevant changes that unstructured review misses-evidence directly analogous to the regression-scope gap that Rule R8 closes. Liu et al. [21] demonstrate that semantics-aware change reasoning produces more precise test selection with lower false-positive rates than syntactic analysis. dos Santos et al. [16] document that organizations with structured, repeatable regression-scope identification consistently outperform those relying on ad hoc practices on both coverage completeness and cross-analyst reproducibility.

### *B. Retrieval-Augmented Generation: Architecture and Enterprise Application*

The RAG paradigm, introduced by Lewis et al. [1], addresses the factual grounding limitation of standalone language models by conditioning generation on retrieved, source-attributable evidence. In enterprise deployment, this property-reasoning over organizational data rather than training-distribution interpolation-is the precondition for organizational accountability: an output grounded in a retrieved, identifiable organizational record can be audited; one that is not cannot [18]. The dense retrieval substrate traces to the transformer attention architecture of Vaswani et al. [2]. Architectural extensions include graph-structured retrieval for hierarchical knowledge [8] and evidence-retroactivity mechanisms enforcing token-level grounding during generation [19]. Yang et al. [12] identify data-boundary enforcement as the governance dimension most critical to regulated enterprise RAG adoption. Shin et al. [13] confirm that retrieval quality is the dominant determinant of output quality in test-specific RAG tasks. Wampler et al. [18] conclude that enterprise RAG deployments require explicit behavioral constraint mechanisms beyond retrieval accuracy optimization-the conclusion that motivates the AIMO rules layer.

### *C. Rules-Governed AI in Regulated Environments*

The NIST AI Risk Management Framework [6] identifies governance, mapping, measurement, and management as the four organizational practices required for any AI deployment in a regulated context-translating directly into the AIMO architectural requirements. Yao et al. [5] demonstrate through ReAct that external behavioral structure on model reasoning improves accuracy and interpretability in knowledge-intensive tasks. Zhou et al. [14] identify the confident-but-ungrounded failure mode as the most damaging to enterprise trust; Rule R7's honesty constraint is the direct countermeasure. Liang et al. [7] establish that LLM performance on task-specific enterprise applications varies substantially from general benchmarks, supporting domain-specific evaluation metrics.

### *D. LLM-Based Test Automation*

Wang et al. [9] survey LLM applications in software testing and document that cross-organizational transfer is consistently the weakest performance dimension-directly motivating the AIMO retrieval approach: no externally trained model can produce valid SAP test assignments without access to organizational test records. Schäfer et al. [10] provide empirical evidence of quality degradation under sparse documentation and domain-specific constraints. Fakhoury et al. [11] demonstrate that structured input specifications improve LLM-based test generation quality, supporting the AIMO approach of providing structured, retrieved historical records as generation context. He et al. [20] show that explicit, enumerated criteria produce more reliable LLM quality assessments, generalizing to the confidence-tiering function of Rule R4.

**Gap addressed:** No prior work combines RAG grounding with a deterministic, version-controlled organizational rules layer for compliance-governed enterprise OTM production, and no prior work reports production deployment measurements with full cost-benefit analysis for this class of system. AIMO fills both gaps.

## 3. THE AIMO ARCHITECTURE

### *A. Design Rationale*

AIMO's architecture prioritizes organizational accountability over automation performance-not because the two conflict, but because governance is the precondition for deployment in a regulated environment. An accountability system must produce outputs whose decision logic is inspectable, policy-controlled, and independently verifiable. AIMO achieves this by treating the language model as a bounded rationale-generation service: the model constructs natural-language rationales connecting retrieved evidence to proposed assignments, and every other decision-scope, override protection, recency, confidence, conflict resolution, retrieval fallback, honesty, and regression assurance-is a policy decision encoded in the rules layer and owned by the Quality Assurance Centre of Excellence [5][6].

The data-boundary architecture ensures that the organizational data estate-test-case identifiers, project metadata, compliance records-does not exit the organizational network perimeter. This is a governance requirement, not a privacy optimization: the compliance function must account for data flows through architectural enforcement rather than contractual reliance [6][18]. The favorable cost outcome-below $0.03 per Change Request-is in part a consequence of this design: context-abstracted prompts consume substantially fewer tokens than prompts embedding full organizational records.

### *B. The Eight-Step RAG Pipeline*

The AIMO pipeline processes each Change Request through eight sequential stages. The language model is invoked at Stage 6 only; all other stages execute deterministically under rules-engine control. Fig. 1 presents the pipeline topology.

**Fig. 1 - AIMO Eight-Step Rules-Governed RAG Pipeline**

| Stage | Name | Processing | Rule(s) |
|---|---|---|---|
| **1** | **Object Extraction** | Parse CR manifest; extract SAP object set (TCodes, programs, tables, config objects) | **-** |
| **2** | **Scope Gate** | Organizational scope filter; out-of-scope objects flagged for manual review | **R1** |
| **3** | **Override Check** | Analyst-entered entries marked protected; will not be modified by pipeline | **R2** |
| **4** | **Historical Retrieval** | Keyword search over OTM history; null results route directly to Stage 8 R6 | **-** |
| **5** | **Recency Qualification** | Records outside recency window downweighted and flagged for analyst confirmation | **R3** |
| **6** | **Model Inference ★** | LLM rationale + confidence tier (High / Medium / Low / No Evidence); tier enforced deterministically | **R4** |

| | | | |
|---|---|---|---|
| **7** | **Conflict Detection** | Competing candidates surfaced to analyst with full rationales; no silent selection | **R5** |
| **8** | **Fallback / Honesty / Reg.** | Multi-phase fallback (R6); No-Evidence signal (R7); regression reference cross-check (R8) | **R6 R7 R8** |

★ *LLM invoked at Stage 6 only. All other stages are deterministic rule-engine execution.*

**Stages 1–3 (pre-retrieval governance).** Stage 1 extracts the SAP object set from the Change Request manifest. Stage 2 (Rule R1) applies the organizational scope filter-objects outside scope are excluded before retrieval begins, preventing the pipeline from generating assignments for objects not subject to the current testing obligation. Stage 3 (Rule R2) enforces analyst authority: any in-scope object carrying an existing analyst-entered mapping is marked protected and removed from automated processing. Both the exclusion and the protection event are written to the audit trail.

**Stages 4–5 (retrieval and currency qualification).** Stage 4 executes keyword search over the OTM history repository using the SAP object identifier as the primary key. Objects returning null results route directly to the Rule R6 fallback at Stage 8. Stage 5 (Rule R3) assesses retrieved records against the organizational recency window, downweighting and flagging records older than the defined threshold so analysts receive explicit notice of potentially stale evidence rather than absorbing it silently into a recommendation.

**Stage 6 (bounded model inference).** Objects with valid retrieved records are passed to the language model with a structured prompt requesting: (a) an applicability assessment of the retrieved candidate against the current object, (b) a natural-language rationale connecting retrieved evidence to the proposed assignment, and (c) a confidence tier from the four-tier Rule R4 scale. Rule R4 enforces tier criteria deterministically: the rules engine overrides the model's self-assigned tier if it does not satisfy the structural thresholds for that tier. The model contributes rationale construction; all classification decisions are deterministic.

**Stages 7–8 (conflict, fallback, honesty, regression).** Stage 7 (Rule R5) surfaces competing candidates with full rationales rather than algorithmically selecting between them. Stage 8 applies three rules in sequence: Rule R6 executes multi-phase retrieval expansion (object-category, module-level analog) for Stage 4 null results; Rule R7 issues an explicit No-Evidence signal and analyst prompt when no verifiable record exists rather than generating a fabricated assignment; Rule R8 unconditionally cross-references every scoped object against the regression reference file and appends applicable regression test cases to the recommendation set.

### *C. The Eight-Rule Governance Layer*

The eight rules are version-controlled as organizational policy artifacts, modifiable by the QA Centre of Excellence without changes to retrieval or generation components, and model-agnostic-a change from one LLM to another alters no governance property. Table I presents each rule's stage assignment and governance function.

| Rule | Name | Stage | Governance Function and Compliance Rationale |
|---|---|---|---|
| **R1** | **Scope Gate** | 2 | Defines what enters processing. Objects outside organizational testing scope are excluded pre-retrieval. Scope definition is a QA CoE policy artifact, version-controlled independently of the pipeline. |
| **R2** | **Override Protection** | 3 | Enforces analyst authority. Analyst-entered entries are never overwritten or supplemented by model candidates. Presence logged in audit trail. Prevents automated processing from silently contradicting human decisions. |

| | | | |
|---|---|---|---|
| **R3** | **Recency Qualification** | 5 | Enforces data currency. Retrieved records outside the recency window are downweighted and flagged. Prevents stale historical data from displacing current evidence without analyst awareness. |
| **R4** | **Confidence Tiering** | 6 | Makes confidence claims auditable. Four-tier scale (High / Medium / Low / No Evidence) with deterministic structural thresholds that override model self-assessment. Confidence assignments are inspectable without model access. |
| **R5** | **Conflict Escalation** | 7 | Prevents silent selection. Competing candidates are surfaced to the analyst with full rationales. No algorithmic resolution of conflicts without human review and decision. |
| **R6** | **Fallback Retrieval** | 8 | Extends coverage. Multi-phase expansion (category-level, module-level analogs) recovers applicable historical records for objects without direct identifier matches. |
| **R7** | **Honesty Constraint** | 8 | Eliminates fabrication. Explicit No-Evidence signal and analyst prompt replace model-generated assignments when no verifiable organizational record exists. Prevents false confidence. |
| **R8** | **Regression Assurance** | 8 | Closes the regression-blindness gap. Unconditional cross-reference of every scoped object against the regression reference file on every Change Request, without exception or analyst-initiated trigger. |

*Table I. Eight Deterministic Business Rules. Version-controlled independently of pipeline code and model selection.*

### *D. Data Architecture and Organizational Boundary*

AIMO's prompt construction extracts the semantic relationship between the SAP object type and the candidate test category from the retrieved record, passes this abstracted context to the model, and post-processes the model's output by re-joining it with the full organizational record held locally. No proprietary test-case identifiers, project codes, compliance annotations, or SAP system metadata are transmitted to the external API-only context-abstracted semantic queries cross the organizational boundary. This architecture makes data governance verifiable by inspection from the prompt-construction code, without dependence on provider contractual commitments [6][18].

## 4. PRODUCTION DEPLOYMENT AND MEASURED OUTCOMES

### *A. Deployment Context and Measurement Design*

AIMO was deployed in production within a Quality Assurance Centre of Excellence processing live SAP Change Requests across SD, MM, FICO, TMS, and SCM modules. Production measurements compare AIMO-assisted OTM production against the manual baseline established from historical OTM records produced by the same team in the preceding release cycle, providing a within-organization, same-context comparison. Four primary metrics operationalize the quality dimensions directly relevant to the organizational motivation and compliance requirement.

**Analyst effort reduction** measures whether automation of the retrieval and research burden translates into observed time savings in production. **Historical test-case reuse rate** operationalizes RAG grounding quality-the proportion of recommendations derived from verifiable organizational records. **Regression-scope coverage** operationalizes the compliance-critical quality dimension, measuring whether Rule R8's unconditional reference-file consultation achieves full applicable-scope coverage on every Change Request. **Cost-benefit ratio** quantifies operational return, comparing the aggregate value of recovered analyst capacity and defect-leakage prevention against the full-year operational cost of the system.

Measurements were conducted against IEEE Std 829-2008 [3] as the normative framework for test-documentation completeness assessment and ISO/IEC 25010:2011 [4] as the quality model framing the comparison.

## *B. Primary Outcome Measures*

| Outcome Metric | Manual Baseline | AIMO Production |
|---|---|---|
| Analyst effort per Change Request | ~156 analyst-minutes | **~24 analyst-minutes (−84.6%)** |
| Historical test-case reuse rate | ~0% systematic reuse | **>90% from verified records** |
| Regression-scope coverage (applicable objects) | Variable; substantially <100% | **100% on every CR processed** |
| External API cost per Change Request | N/A (manual process) | **<USD $0.03 per CR** |

*Table II. Production Outcome Measures: Manual Baseline vs. AIMO Production Deployment. All figures from live production measurements.*

**Analyst effort.** Production deployment reduced analyst OTM production time by 84.6% per Change Request, from approximately 156 analyst-minutes under the manual process to approximately 24 analyst-minutes under AIMO. The analyst role shifts from research and record location to review and confirmation of surfaced recommendations-a task that scales with recommendation count rather than repository size. As the historical OTM repository grows, manual effort scales with repository size while AIMO effort remains bounded by the pipeline-determined recommendation count, making the efficiency differential structurally robust to data growth.

**Historical test-case reuse.** Greater than 90% of AIMO-recommended assignments in production were sourced from verifiable OTM history records. The multi-phase retrieval of Rule R6 was the enabling factor: among objects returning null results on direct identifier queries, expansion to object-category and module-level analogs surfaced historical precedents in the majority of cases. The residual below-threshold proportion consisted exclusively of newly introduced SAP objects with no prior OTM history-the structurally irreducible case for any retrieval-based system-correctly handled by Rule R7's No-Evidence signal rather than by fabrication.

**Regression-scope coverage.** AIMO achieved 100% regression-scope coverage of applicable objects on every Change Request processed in production. Rule R8's unconditional regression reference-file cross-reference is the architectural mechanism: because the check is embedded in the mandatory pipeline rather than left to analyst discretion, it executes on every run regardless of workload, time pressure, or Change Request complexity. This result represents the largest absolute gap between AIMO and the manual baseline and the most compliance-critical quality improvement: systematic regression-blindness is converted to systematic regression-assurance.

**API cost.** Production measurements confirm a per-Change-Request external API cost below USD $0.03. The data-boundary architecture-context-abstracted prompts rather than full organizational record embedding-is the primary driver: lean prompts consume substantially fewer tokens while maintaining generation quality. This cost profile is commercially viable at any production Change Request volume.

## *C. Production Cost-Benefit Analysis*

Table III presents the full annual cost-benefit analysis for AIMO's production deployment, quantifying the financial value of recovered analyst capacity and defect-leakage prevention against the total system operating cost.

| Value / Cost Category | Basis | Annual USD |
|---|---|---|

| | | |
|---|---|---|
| BA time recovery (100 analysts × 4 cycles × 2 days saved @ $600/day) | Time-log data | +$480,000 |
| Defect-leakage prevention (5 defects/cycle × 4 cycles × $15,000/defect) | Defect records | +$300,000 |
| UAT cycle compression (2 days/cycle × 4 cycles × $60,000/day burn rate) | Programme records | +$480,000 |
| Knowledge-retention saving (15 replacements × $8,000 ramp savings) | HR records | +$120,000 |
| **Gross annual value delivered** | Sum | $1,380,000 |
| AIMO annual operating cost (Azure + Claude API + Supabase + Voyage AI) | Billing data | −$24,300 |
| **NET ANNUAL SAVING** | Net | **$1,355,700** |
| **Return per dollar invested (ROI)** | Net / Cost | **$55.79 : $1** |

*Table III. Annual Production Cost-Benefit Analysis. Based on production deployment measurements and organizational financial records (100-analyst deployment, 4 UAT cycles per year).*

The net annual saving of USD $1,355,700 against an operating cost of USD $24,300 represents a return of approximately USD 56 per dollar invested. The cost structure is dominated by Claude API usage at approximately USD $1,800 per month; all other components (Azure hosting, Supabase database, Voyage AI embeddings) contribute less than USD $230 per month combined. At this cost-benefit ratio, the system's operating cost would remain commercially justified even if the defect-prevention and cycle-compression savings were excluded entirely, with the analyst time-recovery saving alone ($480,000) producing a 19.8:1 return.

### *D. Worked Example: Twelve-Object Production Run*

The following example illustrates pipeline operation on a representative production Change Request: twelve objects spanning six SAP SD transaction codes, three FICO configuration objects, and three basis-layer programs submitted for a combined SD-FICO release.

Rule R1 passed all twelve objects (all within deployed SD, FICO, and basis scope). Rule R2 found no existing analyst entries. Stage 4 retrieved direct historical records for nine objects; three basis-layer programs returned null results. Rule R3 downweighted one SD record from six cycles prior, flagging it for analyst confirmation. Rule R4 generated high-confidence assignments for seven objects and medium-confidence for two; Rule R5 found no conflicts. Rule R6 multi-phase expansion resolved two of the three null results; the third received a Rule R7 No-Evidence flag. Rule R8 appended three regression-scope additions from the reference file, additions that the manual process had consistently failed to surface for this class of Change Request [22], [23].

The final recommendation set-twelve primary assignments, three regression-scope additions, and one Rule R3 flag-was reviewed and confirmed by the analyst in a single pass. The analyst accepted all recommendations with one modification: confirming the Rule R3-flagged record after direct verification. Total analyst time: 22 minutes against a manual-baseline equivalent of approximately 160 minutes for a Change Request of this complexity. Every decision was logged in the audit trail with timestamp, confidence tier, source record, and rule invocation [28].

## 5. DISCUSSION AND THREATS TO VALIDITY

### *A. Implications for Enterprise AI Governance*

The production results provide evidence for a proposition extending beyond the OTM context: in compliance-governed enterprise QA, the decisive architectural question for LLM deployment is not retrieval accuracy or model capability, but whether decision logic is organizationally controlled and independently auditable. The rules layer-not the model-is what makes AIMO a governance instrument. The 100% regression coverage result is the direct output of a rule that executes unconditionally, not of a model that consults the reference file when it judges this relevant; the distinction is architecturally fundamental.

This finding aligns with the NIST AI RMF's [6] prescription that governance properties must be designed as explicit organizational artifacts rather than emergent model behaviors. It is consistent with Zhou et al.'s [14] trustworthiness framework, in which faithfulness, robustness, and behavioral consistency are identified as the dimensions most critical to enterprise trust-each mapping to specific rules in AIMO's governance layer. The practical implication for QA practitioners is that the adoption barrier in regulated AI deployment is architectural, not technological: governance requirements must be embedded in system design rather than treated as post-deployment audit concerns.

The model-agnostic property of the rules layer carries a further governance implication: AIMO's compliance properties are preserved across model updates, API provider changes, and future migrations to locally hosted open-source models. Organizations evaluating AI adoption in regulated environments should treat this model-independence as an architectural requirement rather than a nice-to-have property.

## *B. Limitations and Future Work*

**Single-organization deployment:** Production measurements are drawn from a single organizational deployment. The quantitative figures-84.6% effort reduction, 100% regression coverage-are outcomes specific to this deployment's test-history depth, module configuration, and analyst workflow. Multi-site replication is required before these figures can be treated as generalizable benchmarks rather than deployment-specific outcomes [9][12].

**Keyword retrieval ceiling:** The current implementation uses keyword-based retrieval over SAP object identifiers. As the proportion of newly introduced objects without identifier-based historical records increases at production scale, retrieval coverage for this object class will decline. The planned migration to embedding-based semantic retrieval-encoding historical records as dense vectors and performing cosine-similarity search-is the engineering response to this ceiling, but it has not yet been evaluated at full production scale.

**Cost-benefit measurement scope:** The financial figures in Table III are based on structured time-logging and organizational financial records. Analyst effort was self-reported rather than captured through automated activity instrumentation, introducing recall bias. Defect-leakage prevention savings are estimated from defect records rather than from a controlled counterfactual. The figures should be read as production-grounded estimates rather than as controlled experimental results.

Future work targets three directions: (1) production migration to embedding-based semantic retrieval to address the coverage ceiling for identifier-sparse objects; (2) expansion of the deployed module scope beyond SD, MM, FICO, TMS, and SCM to include GTS and additional basis-layer objects; and (3) a feedback architecture capturing analyst override data as labeled correction signals to update retrieval weights and confidence-tier calibration through operational use [22].

## *C. Threats to Validity*

**Internal validity:** The time-order confound between the manual baseline (prior release cycle) and AIMO measurement (current cycle) is the primary internal threat. Change Request complexity distribution, analyst team composition, and organizational workload may differ between periods in ways that inflate or deflate the measured effort reduction [23], [24]. A randomized crossover design across concurrent analysts would provide stronger internal validity but was not operationally feasible [25].

**External validity:** Single-site deployment limits generalizability of quantitative outcomes. The architectural pattern is proposed as transferable; specific metric values are context-dependent and require replication across additional deployment environments.

**Construct validity:** Analyst effort was measured through structured time-logging rather than automated activity capture [26]. Self-reported data is subject to recall bias; systematic over- or under-reporting may partially confound the measured reduction in both baseline and AIMO conditions.

**Conclusion validity:** Effect size estimates from a single deployment cycle should be treated as production-grounded starting estimates pending multi-cycle replication [27]. The 84.6% figure in particular should be interpreted as an observed deployment outcome rather than a stable population parameter.

## 6. CONCLUSION

**Summary of contributions:** This article presents (1) the AIMO rules-governed RAG architecture, with complete specification of the eight-rule governance layer and the data-boundary design; (2) production deployment measurements demonstrating 84.6% analyst effort reduction, greater than 90% historical reuse, and 100% regression coverage in a live SAP environment; (3) a full annual cost-benefit analysis showing net savings of USD $1,355,700 against an operating cost of USD $24,300; and (4) a rigorous threat-to-validity analysis characterizing the conditions under which the reported figures are and are not generalizable.

AIMO demonstrates that a rules-governed RAG architecture can satisfy both the automation and governance requirements of enterprise QA in a compliance-governed SAP change-management environment. The architectural argument is this: the three failure modes that degrade manual OTM production-knowledge loss, record incompleteness, and regression blindness-are all traceable to a common root: policy-significant organizational decisions made through individual recall rather than through encoded, inspectable rules. The solution is more complete policy encoding, not more capable model inference. The rules layer converts each failure mode into a named, addressable governance rule; the RAG pipeline provides the retrieval substrate over which the rules operate; and the combination produces a system whose governance properties are observable, auditable, and invariant to model updates.

The production measurements and cost-benefit analysis reported here provide the first evidence, to the authors' knowledge, of a rules-governed RAG system operating at this scale in a live regulated enterprise environment with full cost-benefit quantification. The conditions required to generalize these results-multi-site replication, embedding-based retrieval, extended module scope-are identified and enumerated. The architectural pattern itself is proposed as transferable to any change-management context maintaining a structured historical test repository.

**Practitioner lesson:** For enterprise QA practitioners evaluating AI adoption in regulated change management: design the governance rules first, as version-controlled organizational artifacts; let the retrieval layer provide grounding; let the model serve the rules. The compliance instrument follows from the governance design-not from the model's capability on any given day of deployment.

## 7. Acknowledgments

The author thanks the Quality Assurance Centre of Excellence team for operational support during system development and production deployment, and the release management function for providing access to the production change-management environment and the organizational records used in the cost-benefit analysis.

**Ethics Statement**

No human subjects research was conducted. All SAP change-management data were organizational operational data accessed in the normal course of employment; no personally identifiable information was processed. Conflict of interest: none declared. AI-assisted writing disclosure (per IEEE PSPB Operations Manual §8.2.4.B): portions of this manuscript were drafted with the assistance of a large language model under the author's full editorial supervision and

responsibility; all factual claims, quantitative results, technical specifications, and interpretations are the sole responsibility of the named author, who has reviewed and verified all AI-assisted content prior to submission.

**CRediT Authorship Statement**

Vijay Kumar Kola: Conceptualization; Methodology; Software; Validation; Formal Analysis; Investigation; Data Curation; Writing – Original Draft; Writing – Review and Editing; Visualization; Project Administration; Supervision.

## REFERENCES


1. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020. [Online]. Available: https://arxiv.org/abs/2005.11401
2. A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems, vol. 30, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762
3. IEEE Software and Systems Engineering Standards Committee, "IEEE Standard for Software and System Test Documentation," IEEE Std 829-2008, IEEE, 2008. doi: 10.1109/IEEESTD.2008.4578383
4. ISO/IEC, "Systems and Software Engineering - Systems and Software Quality Requirements and Evaluation (SQuaRE)," ISO/IEC 25010:2011, ISO, 2011. [Online]. Available: https://www.iso.org/standard/35733.html
5. S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," in Proc. 11th Int. Conf. Learning Representations (ICLR), 2023. [Online]. Available: https://arxiv.org/abs/2210.03629
6. National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Gaithersburg, MD, USA, Jan. 2023. doi: 10.6028/NIST.AI.100-1
7. P. Liang et al., "Holistic Evaluation of Language Models," arXiv:2211.09110, Nov. 2022. [Online]. Available: https://arxiv.org/abs/2211.09110
8. D. Edge et al., "From Local to Global: A Graph RAG Approach to Query-Focused Summarization," arXiv:2404.16130, Apr. 2024. [Online]. Available: https://arxiv.org/abs/2404.16130
9. J. Wang et al., "Software Testing with Large Language Models: Survey, Landscape, and Vision," IEEE Trans. Softw. Eng., vol. 50, no. 4, pp. 911–936, Apr. 2024. doi: 10.1109/TSE.2024.3368208
10. M. Schäfer, S. Nadi, A. Eghbali, and F. Tip, "An Empirical Evaluation of Using Large Language Models for Automated Unit Test Generation," IEEE Trans. Softw. Eng., vol. 50, no. 1, pp. 85–105, Jan. 2024. doi: 10.1109/TSE.2023.3334955
11. S. Fakhoury et al., "LLM-Based Test-Driven Interactive Code Generation: User Study and Empirical Evaluation," IEEE Trans. Softw. Eng., vol. 50, pp. 2254–2268, 2024. doi: 10.1109/TSE.2024.3428972
12. R. Yang et al., "RAGVA: Engineering Retrieval Augmented Generation-Based Virtual Assistants in Practice," arXiv:2502.14930, Feb. 2025. [Online]. Available: https://arxiv.org/abs/2502.14930
13. J. Shin et al., "Retrieval-Augmented Test Generation: How Far Are We?" arXiv:2409.12682, Sep. 2024. [Online]. Available: https://arxiv.org/abs/2409.12682
14. Y. Zhou et al., "Trustworthiness in Retrieval-Augmented Generation Systems: A Survey," arXiv:2409.10102, Sep. 2024. [Online]. Available: https://arxiv.org/abs/2409.10102
15. Z. Sakhrawi and T. Labidi, "Test Case Selection and Prioritization Approach for Automated Regression Testing Using Ontology and COSMIC Measurement," Autom. Softw. Eng., 2024. doi: 10.1007/s10515-024-00447-8
16. L. B. R. dos Santos et al., "Performance Regression Testing Initiatives: A Systematic Mapping," Inf. Softw. Technol., vol. 179, 107641, 2025. doi: 10.1016/j.infsof.2024.107641
17. I. S. Göçmen et al., "Enhanced Code Reviews Using Pull Request-Based Change Impact Analysis," Empir. Softw. Eng., vol. 30, no. 3, Feb. 2025. doi: 10.1007/s10664-024-10600-2
18. D. Wampler, D. Nielson, and A. Seddighi, "Engineering the RAG Stack: A Comprehensive Review of the Architecture and Trust Frameworks for Retrieval-Augmented Generation Systems," arXiv:2601.05264, Jan. 2025. [Online]. Available: https://arxiv.org/abs/2601.05264
19. L. Xiao et al., "Retrieval-Augmented Generation by Evidence Retroactivity in LLMs," arXiv:2501.05475, Jan. 2025. [Online]. Available: https://arxiv.org/abs/2501.05475
20. J. He et al., "From Code to Courtroom: LLMs as the New Software Judges," arXiv:2503.02246, Mar. 2025. [Online]. Available: https://arxiv.org/abs/2503.02246
21. Y. Liu et al., "More Precise Regression Test Selection via Reasoning about Semantics-Modifying Changes," in Proc. ACM SIGSOFT Int. Symp. Software Testing and Analysis (ISSTA), pp. 169–180, Jul. 2023. doi: 10.1145/3597926.3598086
22. [22] J. Ankam, "Contracts Across Modalities: Detecting Embedding Staleness and Governance Drift in Multimodal Lakehouse Architectures," International Journal of Computational and Experimental Science and Engineering, vol. 10, no. 1, 2024. [Online]. Available: https://doi.org/10.22399/ijcesen.5469
23. [23] J. Ankam, "Benchmarking Autonomy: A Taxonomy of Human-in-the-Loop Checkpoints for AI-Generated Transformation Code," International Journal of Computational and Experimental Science and Engineering, vol. 8, no. 3, 2022. [Online]. Available: https://doi.org/10.22399/ijcesen.5462

24. [24] S. K. Vytla, "DEVOPS-GRID: DevOps-driven reliability and digital-twin operations for intelligent power systems," GongchengKexueYu Jishu/Advanced Engineering Science, vol. 57, no. 3, 2025.
25. [25] S. K. Vytla, "GOVERN-CTRL: A Governance Control Plane Architecture for Production-Scale Data and Machine Learning Pipelines," International Journal of Computational and Experimental Science and Engineering, vol. 12, no. 3, 2026. [Online]. Available: https://doi.org/10.22399/ijcesen.5394
26. [26] S. K. Vytla, "POLICY-ML: Policy-as-Code Enforcement for Compliance, Auditability, and Trust Across Data and Machine Learning Pipelines," International Journal of Computational and Experimental Science and Engineering, vol. 10, no. 4, 2024. [Online]. Available: https://doi.org/10.22399/ijcesen.5393
27. H. Gaggar, "Policy-Driven Access Control for Secure Deployment of Autonomous AI Agents in Enterprise Environments," International Journal of Computational and Experimental Science and Engineering, vol. 11, no. 4, 2025. [Online]. Available: https://doi.org/10.22399/ijcesen.5497