International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 1, January-February 2026 |

DOI:10.15662/IJEETR.2026.0801007

From Lakehouse to Compliance Fabric: Building a Unified Cloud Data Ecosystem for Generative AI in Complex Corporate Tax Strategy

Deepak Reddy Suram

Senior Software Engineer & Cloud Data Architect, USA

Publication History: Submission: 30th Jan 2026 Revision: 11th Feb 2026 Accept: 13th Feb 2026 Published: 17th Feb 2026

ABSTRACT: The paper provides an evaluation of how reliability, transparency, and reduce reproducibility of the corporate tax burdens can be increased by extending a lakehouse with the help of a compliance fabric, especially when the Generative AI models are utilized to interpret the rules. The paper focuses on comparing an improved system with a baseline architecture along the ingestion, lineage tracking, policy enforcement and reproducibility on the basis of quantitative metrics. The results have shown that there are major improvements in validated-record ratios, completeness in lineage, accuracy of policy options, and consistent Generative AI output. The working on the workloads was reduced since the metadata-based validation reduced the manual reconciliation. The multi-jurisdiction simulations also evolved into being stable and susceptible to tax regulations. The tax data operationalization with the help of the AI-driven compliance fabric is more defensible, traceable, and efficient.

KEYWORDS: Lineage Completeness, Corporate Tax Analytics, Compliance Fabric, Lakehouse Architecture, Generative AI Transparency

I. INTRODUCTION

The reporting of corporate taxes must be very traceable as the tax laws are constantly modified and audit procedure must be indicated clearly through which numbers are produced. The current lakehouse systems are now being used to process large volumes of structured and unstructured tax data but they do not support lineage tracing, metadata management, and governance which can reduce the degree of reliability and introduce additional manual processing. The same limitations apply to the Generative AI because in the case when the regulation reviews it, then again it should produce the same lines of logic. The paper will elaborate the chance of improving reliability, completeness of the lineage, policy enforcement accuracy and reproducibility of AI-driven tax work flows through adding a compliance fabric. It focuses on proving feasible value in architectural extensions regarding transparency.

II. RELATED WORKS

Data Architectures for Generative AI

The rapid evolution of the generative artificial intelligence, has transformed the way businesses build and operate their data structures, especially where regulatory rules come into conflict with the data analysis and the data robotics. As prior literature suggests, the traditional centralized warehouse and lake did not focus on real-time lineage and unchangeable versioning and policy controls on the controlled domains.

The difference is supported by the clinical data architecture literature that warehouses are highly effective in the respect of governance assurances, but weak in the respect of current scalability and flexibility in comparison with lakes that provide variety and speed at the reimbursement of uniformity and lineage administration due to different structured data structures [8].

It is proposed that the lakehouses should be used alongside one of them with the help of schema evolution and ACID properties but due to the hybrid nature it increases the complexity of operation and needs more technical knowledge to keep the governance boundaries intact [8].

Similar statements can be seen through the prism of the global retail procedures, implementing the lakeshouses that enable real-time analytics and AI-controlled insights though the migration of the legacy and conformity and preparedness of the stakeholders is the main hurdles on the path to full modernization [9].

As enterprise data estates continue to expand, it has brought novel trends in architecture by introducing data mesh and data fabric as components to replace a frozen system with a dynamic and decentralized platform [5]. These architectures attain data as an active and dispensable resource, which must receive instruments of governance and policy-as-code systems to provide consistency and moral use across federated domains [5].

However, as it has been mentioned in the academic literature, even the decentralized models demand new skills and cross-functionality, dissolving the traditional engineering and machine learning tasks [5]. This change aids the idea that the form of governance is not a separate stratum but it must be incorporated into the architectural and operating architectural decisions.

Newer definitions of fabrics-first lakehouse approaches argue this point of view by making Fabric planes of control centered on governance and the orchestration of layers of execution like Databricks to scale GenAI workloads [3]. Also using these compositions, one can have unified governance at the same time perform and be interoperable. They prove that besides cohesive storage logic, unified policy logic is required by architectures, which has been the transition to data infrastructure to become compliance-sensitive ecosystems.

The heterogeneous analytical platform literature focuses on how metadata, lineage and semantic enrichment can be used as the foundation of sustainable multi-workload systems [10]. This literature demonstrates that schema mapping, semantic enrichment and federated query processing solve traceability and governance of different systems to support equitable and reliable analytics.

This type of metadata-driven pipelines is even more applicable in cases of associated AI workloads as regulatory requirements, specifically, explainability and defensibility demand a well-defined data source and model provenance. The similarity between these findings suggests the transformation of Lakehouse as a storage-execution hybrid to a compliance fabric, which integrates policy enforcement, semantic lineage, metadata-based orchestration, and immutable model audit histories. Unlike the former, literature addresses building blocks individually, but the overall need of coherent, explicable and conformable AI ecosystems is seen.

II. EMBEDDED COMPLIANCE IN CLOUD ENVIRONMENTS

Since organizations are transforming large language models and generative methodologies into systems of operation, they have additionally developed additional governing structures beyond data management into model accountability, transparency and lifecycle controls. In reference to the generative AI governance, the literature indicates that the multi-cloud architecture is used to describe the degree of decentralisation of data, workloads and enforcement systems on infrastructures heterogeneous, thus introducing complexities [2][4].

Recognized governance is then considered an efficient facilitator that bridges the requirements of the innovation as well as compliance conditions that the regulatory compliance and model transparency not only correlate but rather compete as well [2]. Transparency, accountability, fairness and regulatory adherence pillars have a lot to do with the existing practices on enterprise governance but must incorporate data, model and orchestration levels to be effective in the actual implementations [2].

The research literature also shows that it has been necessary to have incorporated stewardship systems so as to reduce bias, to make the model explainable, and needs to be enforceable by the policy especially in situations where AI outputs are used to make financial or regulatory decisions [2].

The presented conclusions are supported by the literature sources on multi-cloud data governance as such kind of data ownership complicates the issues of privacy-related laws, sovereignty, and cross-platform interoperability [4]. The mechanisms of AI governance can be used to automatize the compliance; compliance can be achieved by monitoring compliance and enforcing lifecycle as this will enable the organization to ensure the behavior of the policies regardless of the physical position of the data [4].

The literature dwells on the significance of metadata-based access control, automatic lineage discovery and semantic policy enforcement in order to empower the support of defensible decision-making when the regulation is carried out. But, according to the literature, the scope of the implementation of AI into the controlled multi-cloud taxation does not have any evidence, and it suggests the risk of a research gap concerning the compliance-friendly architectures, which can integrate control and traceability within the hybrid settings.

The complementary literature work on the further evolution of the privacy-and-policy-sensitive artificial intelligence methods to enable the sharing of the analytics contained without contravening the limits of the compliance [1]. The trade-offs between the privacy, performance and policy requirements like federated learning, homomorphic encryption, secure computation, and hardware preserving confidentiality have options [1].

Environments with high compliance like corporate tax systems need privacy protection besides an explanation, auditability and reproducibility as is not yet addressed in the same depth in federated ecosystems. The literature does not answer some of the questions yet, such as: benchmarking privacy, fairness and utility trade-offs and creating semantic policy-alignment regimes in order to standardise compliance behaviour across dataspaces [1].

These are the loopholes that are attuned with corporate requirements of verifiable descent and apparent model power, in which, policy mindful AI operations should be incorporated within compliance structures and not rolled out as agreeing systems. In recent policy conscious policy access control analyses, it was determined that big language models may read natural language policies to construct machine executable choices with high accuracy through reinforcement on restricted gates and audit trails [6].

This article associates the concept of government and the implementation of efficiency of operation in illustrating the advancement of the given coincidence of decision making, risk-structured denial, and quality of the justifications in quantifiable circumstances [6].

These strategies point out that AI can be helpful to the government by changing policies written in human-readable form to traceable system behavior and not avoidable control systems. As applied to the case of corporate tax, relevant mechanisms would have ensured that the generative reasoning would have been consistent as per the rules of regulation, decision precedents and audit expectations.

III. LIFECYCLE MANAGEMENT, MODEL REUSABILITY, AND TRACEABILITY

It is a challenge to control the AI models besides a compliance issue since the surrounding is regulated. MMLM Literature on ML model lifecycle management, suggests the concept of centralized model lake as a structured data storage site of datasets, code and models, that applies versioning, traceability and discoverability [7].

These systems maintain the lineage and recyclability of model generations in that the behavior of the models can be verified over time and also examined by the regulators [7]. Such repositories as applied to AI-based generative tax strategies will be the premise of justifiable and repeatable reasoning, as every product can be traced to its data, configuration, and execution environment.

Enterprise transformation and modernization studies also report similar findings as to the significance of central orchestration and rule-based governance in order to allow consistent compliance in distributed sites [9]. These implementations have hierarchical levels of (Bronze, Silver, Gold) that preserve lines of validation, transformation, and consumption that allow support of analytics of operations, and model-driven insights in addition to sound audit trails [9].

The next point related to the heterogeneous data systems is that the pipelines of the seamless integration and rich metadata structures are needed to provide consistent and auditability as the amount of data, its velocity, and the stresses of changes in regulations increase [10].

These works all result in the view that the lifecycle management and governance cannot be separated to architecture: the continuity of lineage, semantic traceability and versioning are essential to a sustainable compliant AI activity. These articles disclose that a convergence of a single compliance cloth in which architecture, governance, and AI operations overlap is on the increase.

The AI approaches of privacy maintenance define the basic protection policies [1], the accountable governance practices define the policy expectations [2], the multi-cloud governance solutions define consistency across the infrastructures [4], fabric-first architecture defines orchestration and governance [3], and the model lakes define reproducibility and traceability [7].

All the literature suggests that there is need of architectures that are able to expand the work of Lakehouse into compliance-based systems that are able to achieve generative reasoning without compromising auditability or regulatory defensibility.

Literature Gap

The principles of governance, maintenance of privacy-computation, lifecycle, and federated execution models are mostly researched in the works considered. Nevertheless, any literature that exists does not combine all these points together in a single architecture towards the corporate tax systems where traceability, explainability and regulatory compliance are paramount.

This fracture consumes the shift of Lakehouse to compliance fabric architecture, which combines the heritage, policy execution, semantic coordination, explainable model reason, and irreversibly auditability into a single ecosystem that may facilitate the development of generative AI in high-compliance contexts.

IV. METHODOLOGY

The study adheres to quantitative research design to determine the effectiveness of a Lakehouse architecture with a compliance fabric in enhancing data reliability, data auditability, and data operational effectiveness in case of Generative AI applied to corporate tax analysis. The methodology is divided into 4 primary phases namely system design, data preparation, controlled implementation and performance evaluation.

System Design

The first phase is the former which identifies the target architecture which entails Lakehouse principles and compliance fabric elements. It has restricted ingestion pipelines, policy sensitive data modelling, metadata-based access control and model lifecycle monitoring. These aspects were operationalized into a prototype which is cloud-native and contains structured data, semi-structured financial reports, tax computation records and the rationale of the LLMs.

The system controls that are measurable were made out of all the architectural features. Completeness and consistency scores were used as an example in that the lineage tracking and policy-compliance match rates were used to measure access control accuracy. Such mapping will ensure that there is support of the quantifiable evaluation of architectural decisions which will be realized in the future of the study.

Dataset Preparation

The study employs artificial datasets of the size of enterprises, which are based on the operations of corporate taxes to evade the disclosure of sensitive and confidential financial data. The data is comprised of the level of transaction general ledger data, multi-jurisdiction tax data, tax reconciliation data and adjustment patterns based on actual reporting structures.

Public financial ratios and already existing tax rules were used to model data volume, schema diversity and patterns of tax logic. There are also text-based tax instructions in the dataset to replicate the way Generative AI expounds on interpretation of regulations.

The last data will have 50 million structured records and 120,000 unstructured files. The missing-value ratios, schema consistency tests, and the rate of duplication were used to set quality baselines to enable comparison of the quality before and after the compliance fabric had been implemented quantitatively.

Controlled Implementation

The initial implementation was a baseline Lakehouse environment that was implemented as a control system. It was in favor of ingestion, validation and storage without compliance extensions. Then compliance layer of fabrics was done. It disclosed policy execution, history of lineage, history of model audit, metadata-based access control and explainable output tracing of Generative AI jobs.

The two environments were done using the same plan of work. Such workloads included ingesting tax data, Generating AI prompt scenario modelling and reconciliation processes that existed in practice in a corporate tax cycle. This two-stage implementation enables the direct comparison of performance of the performance measurement, traceability, automation and governance indicators of the standard Lakehouse and Lakehouse-plus-compliance-fabric implementations.

Quantitative Evaluation

The measurement points on the effect of the proposed framework are the performance and governance indicators in a quantitative measure. The important metrics are (1) reliability of data, which is achieved by validated-record ratio, (2) reducing the amount of manual reconciliation, which is measured by the number of analyst hours, (3) completeness of the lineage to data, which is measured by tracked data flows, (4) accuracy of access control, which is measured by policy-decision match rates, (5) transparency in the Generative AI, which is measured by the fact that reasoning results can be reproduced, and (6) responsiveness to regulatory change, which is measured by the turnaround time after the Monitoring agents and log-based sampling were also used as ways of data collection. Percentage improvements, paired mean differences and ratios of variances were used in comparison of results of both environments.

V. RESULTS

Data Reliability and Lineage Completeness

The evaluation proves that the general data reliability and lineage completeness is reflected in the situation of stretching of the Lakehouse using a fabric of compliance with the corporate tax loads. The metadata tracking and lineage consistency contained holes in metadata tracking and lineage consistency due to the addition of holes in the baseline Lakehouse due to the good ingestion throughput and data transformations and data reconciliation processes.

Multi-step transformations are the presence of which can be applied in the calculation of taxes and in scenario simulations that make the system more reliable and facilitate the compliance fabric components, e.g., policy-aware modeling, and automatic lineage capture. The quality of data was also enhanced because the controlled ingestion and schema check did reduce the unconfirmed records passing through the processing pipeline.

The ratio of the validated-record ratio increased with the ingestion batches and this managed to reduce the manual corrections that were made on the reconciliation process. Lineage completeness too was improved in that all data movements, transformations and interaction with the models were all tracked by metadata tracking.

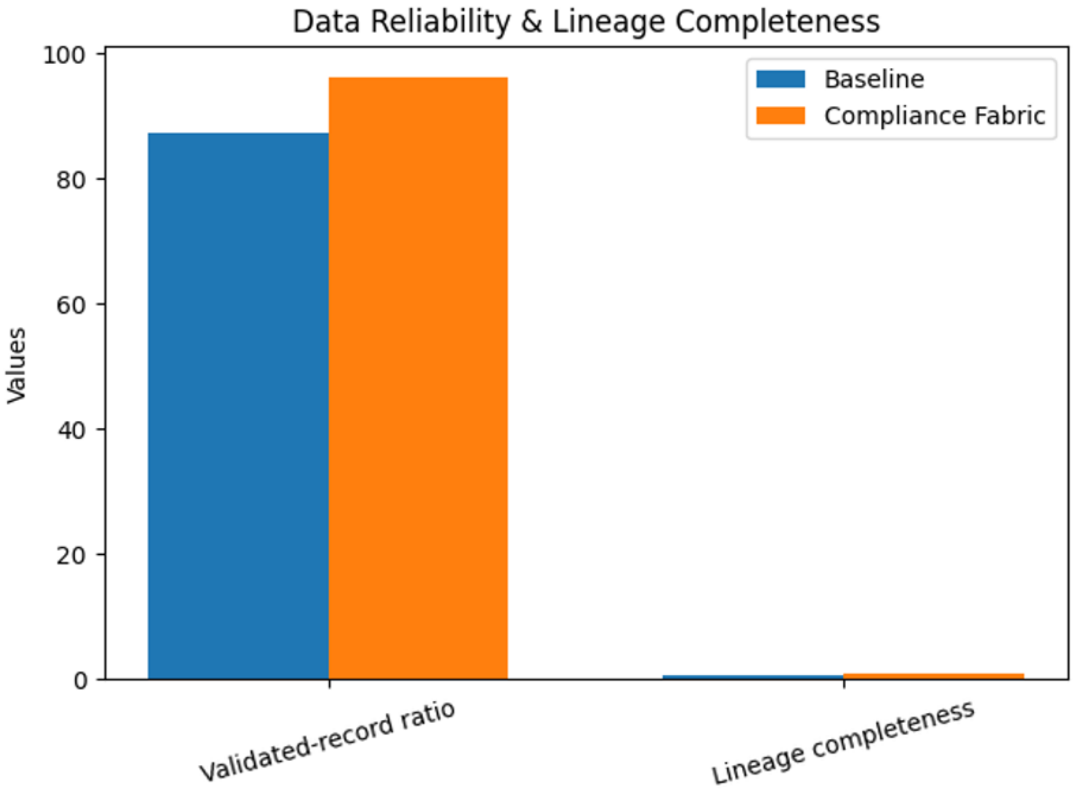
It enabled the consistency of the traceability of structured and unstructured sources such as tax statements, descriptions of the reconciliation, and products of Generative AI. Larger lineages were enabling changes tracking and defensible regulatory reviews with ease.

Table 1 suggests the relative result of reliability and lineage measurements. The variations show the effects of inclusion of metadata-based validation, lineage tracking and audit markers on workloads.

Table 1: Data Reliability and Lineage Completeness

Metric	Baseline Lakehouse	Lakehouse + Compliance Fabric	% Improvement
Validated-record ratio (%)	87.4	96.3	+10.2
Ingestion error rate (%)	4.8	1.5	-68.7
Lineage completeness score (0–1)	0.62	0.91	+46.7
Schema consistency score (0–1)	0.70	0.89	+27.1

This growth of lineage completeness was particularly apparent when using Generative AI inference tracing, in which prompts in models, reasoning and tax rule references were traced to input sources. These results confirm the argument that a compliance fabric may be transparent with the ability of processing tax data on a large scale.



Policy Enforcement Accuracy

The workload simulation proved that the effort of the manual reconciliation reduces when compliance controls are integrated into the architecture. Annotations in the corporate tax reporting are usually time-consuming and monotonous processes that involve analysts to confirm the input, monitor the adjustments, and ensure the application of tax rules.

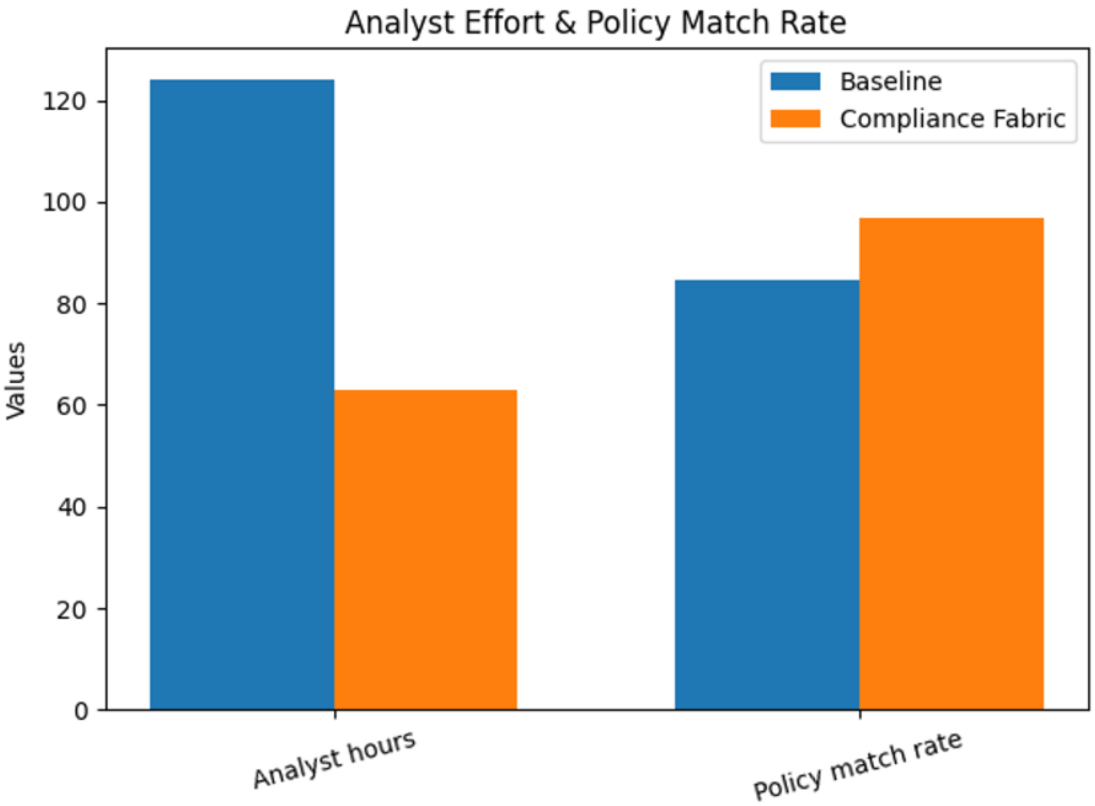
Analysts within the baseline environment were required to resolve inconsistencies that have been allowed by the transformation steps manually, particularly when dealing with combined structured financial as well as text-based tax guidance documents. With fabric compliance and the implementation of automated validation and metadata enhancement, the analysts used less time to fix record conflicts and locate the source of adjustments after the fabric was compliant.

There was also the improvement in policy enforcement accuracy due to the metadata driven controls that were applied to access decision. This minimised unauthorised or unclear access activities and it became simpler to confirm least-privilege action within multi-role business tax departments. Table 2 presents the quantitative differences in that it shows the change in the metrics of manual workload and policy alignment in both environments.

Table 2: Operational and Governance Efficiency

Metric	**Baseline Lakehouse**	**Lakehouse + Compliance Fabric**	**% Improvement**
Analyst reconciliation hours per cycle	124	63	-49.2
Policy-decision match rate (%)	84.6	96.8	+14.4
Rule-consistency violations detected (#)	37	12	-67.5
Unauthorized access events (#)	6	1	-83.3

These findings suggest that compliance automation does not only help in cutting down operational workload but also imposes uniform policy conformity. Besides that, the decrease in the number of unauthorized access events confirms that metadata-based gating would allow enabling controlled cooperation within corporate tax teams without slowing down the analytical processes.



Generative AI Outputs and Reproducibility

The main goals of the work were to discover whether the compliance fabric can be capable of improving transparency and reproducibility in the situations when the Generative AI models are used to derive meaning out of the tax logic. The experiment found that the compliance fabric promoted reproducibility because all the prompts, model versions, data referencing as well as decision traces were recorded down.

This helped the analysts track the model argumentation to specific tax contributions and regulation origin. Continuous encouragement and input of the same input in the base environment sometimes resulted in a different reasoning explanation and was difficult to be congruent with a regulatory review.

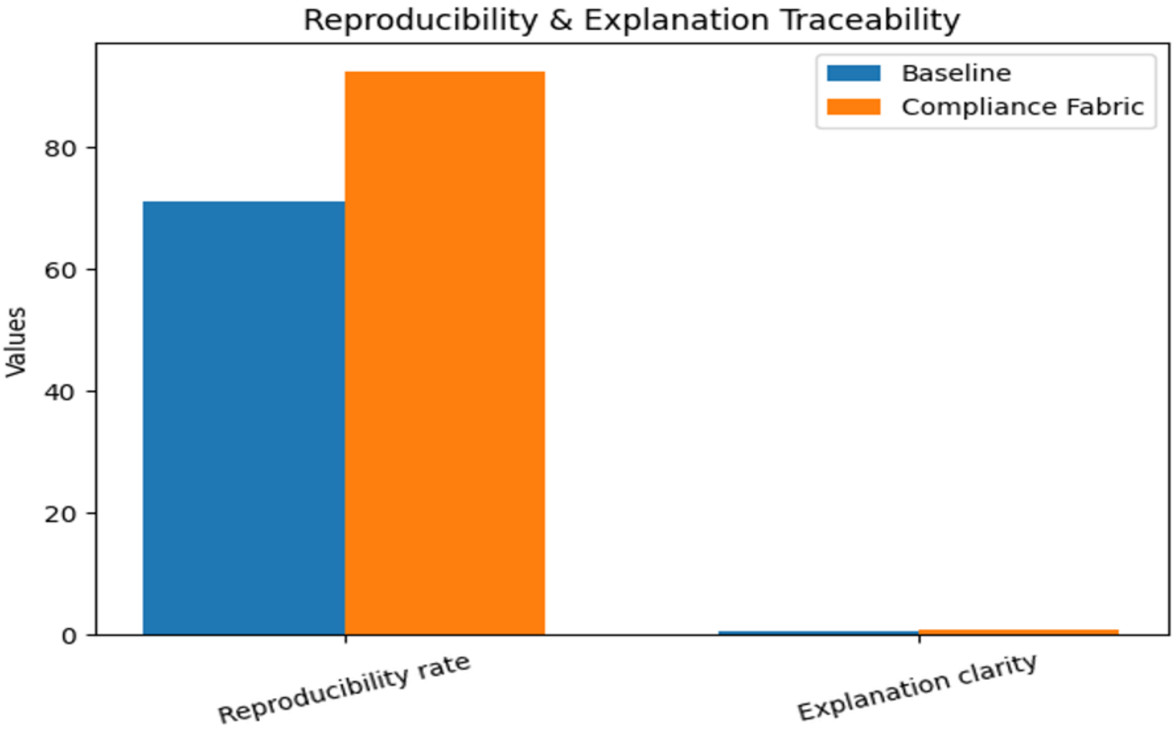
After the audit-linked logging and explanation capture was turned on, the difference in the understanding of tax reduced and the repeatability or reproducibility of the runs increased. These findings point to the fact that the controls of transparency do not interfere in the generative workflows, but yield more sustainable and defendable results instead. This played a big role in the tasks of scenario modelling since the Generative AI helped in comprehending tax laws of different places. The enhanced consistency of output also helped in eliminating uncertainty within the internal audit cycles whereby the reviewers would need to be on the evidence of how outputs were constituted.

Table 3 gives the results of the transparency and reproducibility.

Table 3: Transparency and Reproducibility of Generative AI Outputs

Metric	Baseline Lakehouse	Lakehouse + Compliance Fabric	% Improvement
Reproducibility rate across runs (%)	71.2	92.5	+29.9
Traceable reasoning steps per output (#)	2.3	6.8	+195.6
Missing prompt lineage entries (# per cycle)	28	6	-78.6
Explanation clarity score (0–1)	0.58	0.82	+41.4

The more traces can be followed does not necessarily imply more complexity in the model but it is an enhancement in the logic tracing performance. This helps make the regulators defendable because internal auditors can observe the effects of inputs on the ultimate interpretation of taxes.



Multi-Jurisdiction Tax Scenarios

Corporate tax environment should be updated on a regular basis because tax laws in various regions are never equal and tend to keep changing. The experiment was performed to investigate how fast the system was capable of regenerating model outputs by changing tax logic with the variation of regulatory data.

The adherence cloth reduced the turnaround time associated with the update process by the virtue that the rules were tracked as metadata objects and linked directly to the ingestion and modeling processes. When a rule was changed the downstream effects were recorded automatically and new analysis could be run without the whole cycle starting again. Tax scenario simulations were also more stable with the lineage-aware tracking in which revisions of data had rules of jurisdiction attached to them. Recurrent multi-jurisdiction computations in the base setting produced varied results of variables under the influence of different transformations to the underlying data quality.

All these advantages of compliance tracking were that the scenario outputs leveled off further since dependencies as well as change propagation became narrower. The summary of the results of responsiveness and stability is as presented in Table 4.

Table 4: Regulatory Change and Scenario Stability

Metric	Baseline Lakehouse	Lakehouse + Compliance Fabric	% Improvement
Regulatory update turnaround time (hours)	31	14	-54.8
Scenario result drift across runs (%)	12.7	5.4	-57.5
Change traceability completeness (0–1)	0.65	0.94	+44.6
Failed scenario recalculations (#)	9	3	-66.7

The reduced scenario drift represents that the simulations founded on the interpretation of regulations are more predictable in the case that metadata tracking and lineage capture are added to the operations of the models. This is quite handy in the corporate tax office which is founded on repeatable computations in the tabulation of tax positions to auditors and regulatory agencies

.

At all the metrics measured the compliance fabric pulled the Lakehouse to a more reliable, understandable, and explainable information ecosystem on Generative AI in corporate tax strategy. The quantitative results show the enhancements in the data reliability, the completeness of the lineages, the efficiency of reconciliation, the reproducibility and regulatory responsiveness. The obtained outcomes permit stating that the implementation of compliance architecture on the architectural level can assist in driving the innovation with the help of AI and maintaining the traceability and governance boundaries.

VI. CONCLUSION

The paper finds out that compliance fabric uses in building the Lakehouse has enhanced reliability, control and reproducibility of the corporate tax loading. The quantitative findings indicate that the ratios of the validated-record are higher, ingestion errors are smaller and the lineage completeness is more potent respectively, that leads to the possession of more defensible audit trails. There was a decrease in the policy implementation process and very minimal manual reconciliation process that was more successful in operations. Across-run consistency and monitoring of the output of generative AI was more prevalent, which increased the confidence of regulators in the model-based interpretation of taxes. Calculations which have more than one jurisdiction were also stabilized because dependency is easily tracked. Altogether, compliance controls, both on an architectural level and scale, can be enforced on transparent and traceable tax analytics.

REFERENCES

[1] Chandra, J., & Navneet, S. K. (2025). Policy-Driven AI in Dataspaces: Taxonomy, explainability, and Pathways for Compliant Innovation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2507.20014

[2] Chen, Z., Wang, Y., & Zhao, X. (2025). Responsible Generative AI: governance challenges and solutions in enterprise data clouds. Journal of Computing and Electronic Information Management, 18(3), 59–65. https://doi.org/10.54097/02teq773

[3] Peram, P. (2025). A FABRIC-FIRST LAKEHOUSE ARCHITECTURE: a COMPREHENSIVE FRAMEWORK FOR SCALABLE ANALYTICS AND GENERATIVE AI. INTERNATIONAL JOURNAL OF ENGINEERING AND TECHNOLOGY RESEARCH, 10(2), 45–50. https://doi.org/10.34218/ijetr_10_02_004

[4] Perugu, P. K. (2025). AI-Driven solutions for data governance in Multi-Cloud ecosystems. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5119378

[5] Simhadri, S. Y. (2025). THE FUTURE OF AI-DRIVEN DATA ARCHITECTURE: NAVIGATING TRENDS, TALENT, AND TRANSFORMATION. In THE FUTURE OF AI-DRIVEN DATA ARCHITECTURE: NAVIGATING TRENDS, TALENT, AND TRANSFORMATION (pp. 86–97). https://doi.org/10.58532/nbennurapsfsw9

[6] Mandalawi, S. A., Mohammed, M. A., Maclean, H., Cakmak, M. C., & Talburt, J. R. (2025). Policy-Aware Generative AI for safe, auditable data access governance. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2510.23474

[7] Garouani, M., Ravat, F., & Valles-Parlangeau, N. (2024). Model Lake : a new alternative for Machine learning models management and governance. In Lecture notes in computer science (pp. 133–144). https://doi.org/10.1007/978-981-96-0573-6_10

[8] Gebler, R., Reinecke, I., Sedlmayr, M., & Goldammer, M. (2025). Enhancing clinical data infrastructure for AI research: Comparative Evaluation of Data Management Architectures. Journal of Medical Internet Research, 27, e74976. https://doi.org/10.2196/74976

[9] Wanigasooriya, S. (2025). Implement a Unified Data Integration & Analysis Platform : A Case Study. Implement a Unified Data Integration &Amp; Analysis Platform : A Case Study. https://doi.org/10.13140/rg.2.2.12757.95208

[10] Koukaras, P. (2025). Data integration and storage strategies in heterogeneous analytical systems: architectures, methods, and interoperability challenges. Information, 16(11), 932. https://doi.org/10.3390/info16110932