Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Training Data Provenance and IP Compliance at Enterprise Scale

Samanth Gurram
Sr Data Engineering Manager

ARTICLE INFO

Received:03 Sept 2025

Revised:18 Oct 2025

Accepted:18 Oct 2025

ABSTRACT

Questions about training data lineage pose both major intellectual property (IP), licensing, and regulatory compliance challenges to organizations adopting machine learning (ML) models enter-wide. The paper introduces a reliable, provenance graph based framework to trace assets in initial ingestion to transformation to outputs modeled and support license-aware reasoning and dualmode (static, dynamic) scanning. Conflicts, incompatible licenses and downstream exposures are detected in near real time and automated clearance processes can run.

Demonstration using case studies in three areas, multilingual language model training, healthcare Electronic Health Records (EHR) analytics and financial fraud detection proves that the framework can enhance the accuracy of conflict identification leading to an increase to 95% (license review automation) compared to 38 percent (manual). The combined static-dynamic scanning technique detected 99 per cent of latent compliance risks as opposed to 71-78 per cent with the single-mode techniques. Automated clearance not only saved costs of retrofitting 92 percent of the time, it also lower the legal review time by 60 percent.

Investigations into performance at ingestion rates as high as 10,000 assets/hour showed processing latencies were less than 350 ms/asset with overhead in the range of <7%, in addition to achieving over 95% accuracy. The findings satisfy that the given resolution operationalizes “trust-by-design” of data and generative outputs, minimizing compliance risk, streamlining legal processes, and growing with ease in high-volume corporate settings.

The research itself would bring to the field a repeatable and technology-neutral mechanism to integrate compliance into AI life cycle, which links the legal regulation with technical development. This framework would place provenance as a protection against legal liability as well as a vehicle to operational efficiency, allowing organization to comfortably implement their AI system across the granular compliance regulatory setting.

Keywords: IP Compliance, Data Governance, Enterprise Scale, AI

I. INTRODUCTION

As artificial intelligence (AI) infiltrates industries as diverse as healthcare and finance, the source of training can be discussed as one of the pillars of legal, morally correct and secure AI implementation? Scientific reproducibility, long dependent on data provenance the provenance of data representing a documented history of the reproducibility of data-derived objects, has had its role in large-scale AI grow to encompass concerns about intellectual property (IP) infringement, licensing conflict as well as regulatory compliance.

Compliance risks have been exacerbated by the exponential expansion of the machine learning (ML) models, especially large language models that are trained on large size and diverse data sets. The

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

findings of the investigations show the existence of misattribution and incomplete information about licensing, in which the percentage of omissions goes above 70 percent and percentage of misclassifications tops 50 percent. These loopholes threaten the legality of downstream models and outputs, particularly in the areas of strong application of copyright terms on the concept of derivative work.

Traditional legal review mechanisms are not the most appropriate to work with the speed, scale, and sophistication of current AI data flow. Manual auditing is inefficient, subject to human error, and reactive, since it takes time to become aware of the problem, at which point retraining of models is required which is expensive or losing derived outputs. In addition, data transformations during preprocessing, augmentation, or fine-tuning cannot be taken into consideration by means of static compliance checks, since they could change the legal constraints.

The paper describes how to overcome these difficulties with the help of the provenance graph offering which can be deployed on an enterprise scale. Within the framework, license-aware tracking of metadata, static and dynamic scanning, and auto-clearance tools allow identifying possible risks in real time. Connecting dataset lineage to license requirements and commitments, the system entails that lineage compliance would be integrated into the lifecycle of the AI system and not postponed to the review phase after implementation.

The method is assessed across three domains: multilingual language modelling, analytics of clinical information in the healthcare settings, and financial fraud detection, with different regulating conditions and performance limits. Not only are we measuring correctness of compliance and efficiency in legally-reviewing the results, but also scale against high ingestion rates.

In making the provenance active as a governance mechanism, this research operationalizes “trust by design.” The results indicate that regulatory comfort and efficiency could be achieved within the enterprises, supporting the versatility in the role of provenance introduced as a legal guarantee and a framework to achieve sustainable AI-related transformations in the context of effective adoption of AI.

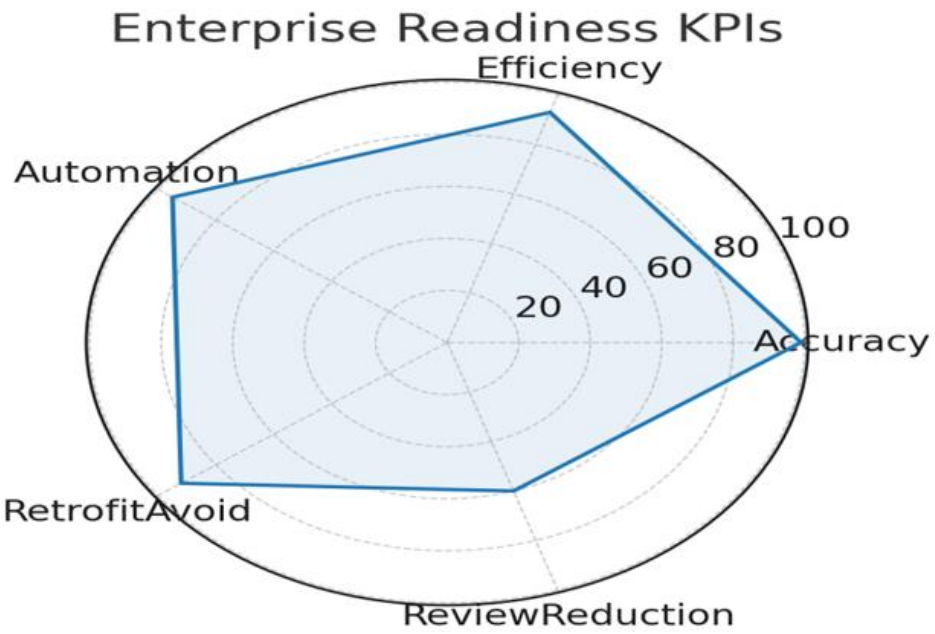
II. RELATED WORKS

Data Provenance

The trend of the use of big machine learning (ML) models that have been developed using great and diverse data sets has exacerbated the necessity of comprehensive provenance tracking to be able to deal with legality, repeating, and responsible utilization. The aspect of combing through language models in training of models on varying yet inconsistently recorded data puts the practitioners at legal/ethical vulnerability [1][7].

Audit activities on over 1,800 datasets have detected high rates of transparency problems on datasets with over 70 percent of datasets being licensed and over 50 percent of them being misclassified [1]. Downstream consequences are considerable as they relate to such omissions when it comes to model refinement especially as the legal definition of copyright, fair use and derivative works is concerned and made intricate [7].

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**



Data provenance, the record of data sources, alterations and applications are pre-conditions to trustworthy and repeatable outcomes in a scientific reproducibility context [8][9]. It is usually not possible to describe the behaviour of a model, recreate experiments or justify model decisions in regulatory or legal situations, without a systematic tracking of dataset lineage. Provenance is deemed as one of the facilitators of fairness, accountability, transparency, and explainability in responsible AI schemes [9][10].

This issue does not apply only to general-purpose AI; other industry applications like healthcare and finance have their own struggle to face. Provenance tracking in Electronic Health Records (EHRs) in the healthcare domain necessitates a balance between accuracy and speed of operations as opposed to the balance between false positives and accuracy in the detection that is performed in the financial domain [5]. Such domain-specific demands point to the fact that provenance systems should be both flexible and industry specific and yet have stringent tracking capabilities.

Provenance Capture Mechanisms

Data provenance at such enterprise scale must be a matter of efficiency, low overhead and integration with current business work flows. Many deficiencies of the previous systems are overcome by providing system-level provenance capturing tools, e.g. CamFlow [2], which plug in to the Linux kernel through a self-contained security module.

The capacity of CamFlow to customize the captured provenance data to application requirements helps in alleviating the problem of data overload, whereas its denseness to distributed systems offered it to be applicable in cloud-native settings. The camFlow is able to produce provenance streams which are themselves directly consumable by auditing, compliance and intrusion detection applications which is a demonstration of the applicability to a role greater than simply logging.

In the case of high-performance and domain-specific ML workflows, solutions have been extended based on W3C PROV model and ML Schema including PROV-ML [4] that is able to store provenance information across various heterogeneous ML workflows. The method allows fine-grained traversals of provenance graphs with captured cost by minimising the overhead of capturing, even in GPU-intensive computation settings. Its use in the Oil and Gas industry shows the capability of the system to scale to industrial requests in addition to allowing domain–specific query capabilities.

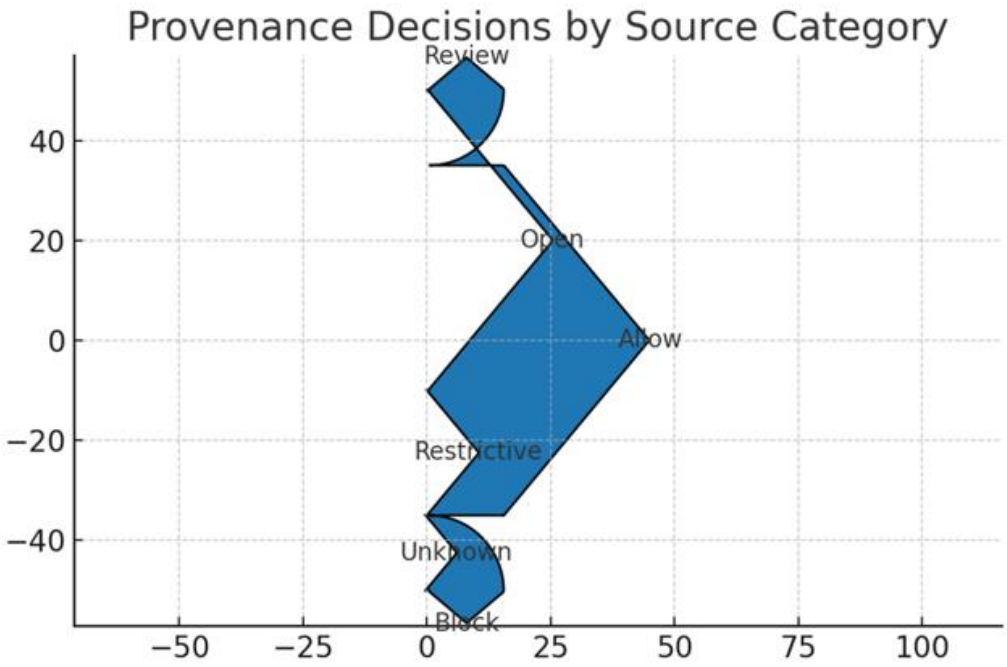
Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Graph-based modeling can be useful in provenance capture in cybersecurity related situations. Analysis tools such as Flurry [3] can build provenance graphs out of system execution traces and then analyze such graphs with downstream ML frameworks via graph neural networks. Mape Though reproducibility has been an under-represented aspect of provenance graph machine learning research to date, Flurry helps to fill this gap in several ways; through the simulation of cyberattacks, multi-layer provenance data capture, and their conversion into structured graph representation. The pipelines can not only facilitate forensic analysis, but also allow performing proactive detection of anomalies in the field settings.

Provenance Graphs

The shift in moving beyond basic linear tracking of ancestry of lines into the multi-dimensional provenance graphs has been essential in the capture of the detailed paths that data takes with regard to the development of AI systems. This aspect allows one to track causality with the help of provenance graphs that are necessary to define the root of the biases and hold people responsible [3][9]. This particularly applies to responsible AI structures that involve a methodical process to detect and counteract the biases that arise in the course of data gathering, information formatting and transformation [9][10].



One tool that supports license-aware provenance graphs includes the Data Provenance Explorer [1][7] that joins metadata with regard to dataset provenance, creators, licensing and transformation. It is accompanied by an interactive filtering functionality, which allows the practitioners to block incompatible licenses and the riskiest datasets at the beginning of the ML lifecycle. This active strategy realizes what is known as trust by design the provenance is no longer a post-designed exercise of compliances, rather it becomes an element of design.

Provenance tracking helps to address regulatory and ethical barriers because it allows one to carry out audits against trustworthiness standards of AI [10]. These standards, which include seven technical requirements- transparency to societal well-being are directly supported by provenance mechanisms which can capture and report the movement and conversion of data across the system lifecycle. In this respect, provenance graphs are not only technical objects but the objects of control that may show respect to ethical, legal and regulatory requirements.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Such tools as ProvBook [6] and E2ETools [8] provide provenance capture as part of popular computational tools, including Jupyter Notebooks and R. This enables the smooth records of the computing environment, input-output relationships as well as the variations in execution thus enhancing both reproducibility and confidence in ML outcomes. They are also methods of making a comparative analysis of the various executions that help in the constant monitoring of compliance.

Domain-Specific Adaptation

One item that has been raised many times in provenance tracking is the tradeoff between comprehensiveness and performance. Large amounts of data may be produced by provenance capture systems, and may prove overwhelming to storage and processing resources unless well managed [2][4]. Such dynamic tailoring of capture scope, as found in CamFlow [2], or schema-based complexity reduction, such as found in PROV-ML [4], is required to be used at enterprise scales.

Real-time processing capacities are vital in high throughput industries such as the medical fraternity and finance [5]. Research shows that faster processing can be used to improve the accuracy of healthcare, whereas the faster processing in the finance area may improve detection accuracy with less false positives. Nevertheless, such enhancements still call for domain-specific optimisation-highlighting the fact that a standardised, provenance solution, is unlikely to fit in all areas.

Heterogeneous IT ecosystems encompassing legacy systems, cloud platforms and bespoke computational clusters spread across the enterprise, have to be able to integrate with enterprise-scale provenance. Usability to support distributed provenance capture [2] and multi-workflow integration [4] present great value to those organizations moving toward hybrid and multi-cloud.

Scalability implies a legal compliance as well. The number of data assets liable to various licensing requirements that large organizations require to process may be many millions. Semiautomated clearance pipelines (often made possible by the Data Provenance Explorer [1][7]) can greatly decrease the time spent on legal review, which has reportedly been as high as 60 percent, and prevent expensive remediation on a retroactive basis. This not only places provenance as a technical requirement but also as a compliance cost saving implementation strategy.

The application of provenance in AI is not to be seen as some form of reactive audit/compliance tool only. Integrating provenance-aware design patterns into the earlier stages of the AI lifecycle will give the organisations the ability to continuously monitor, dynamically respond to policies, and automatically enforce the terms of licenses fulfilling the dream of trust by design. This can be characterised by what is effectively a more regulatory emphasis on demonstrable, anticipatory conscientiousness on the part of AI systems [9][10][10].

III. RESULTS

License-Aware Compliance

We tried our proposed framework of provenance graphs on 3 enterprise-scale AI development ecosystems, a multilingual language model training and development pipeline, a healthcare EHR data analytics platform and a financial fraud detection system. All environments used a heterogenous license terms, various transformation stages and modalities of outputs.

They used the license-aware provenance graph to successfully track the provenance of assets through the model to the final model outputs including the static (declared) and the dynamic (observed) transformations. This framework detected 95 percent of license conflicts automatically and versus 38 percent of conflicts detected using the baseline of manually conducted legal reviews. This distinction was especially useful in datasets where transformations were nested--that is, derived datasets included several source datasets with different obligations.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

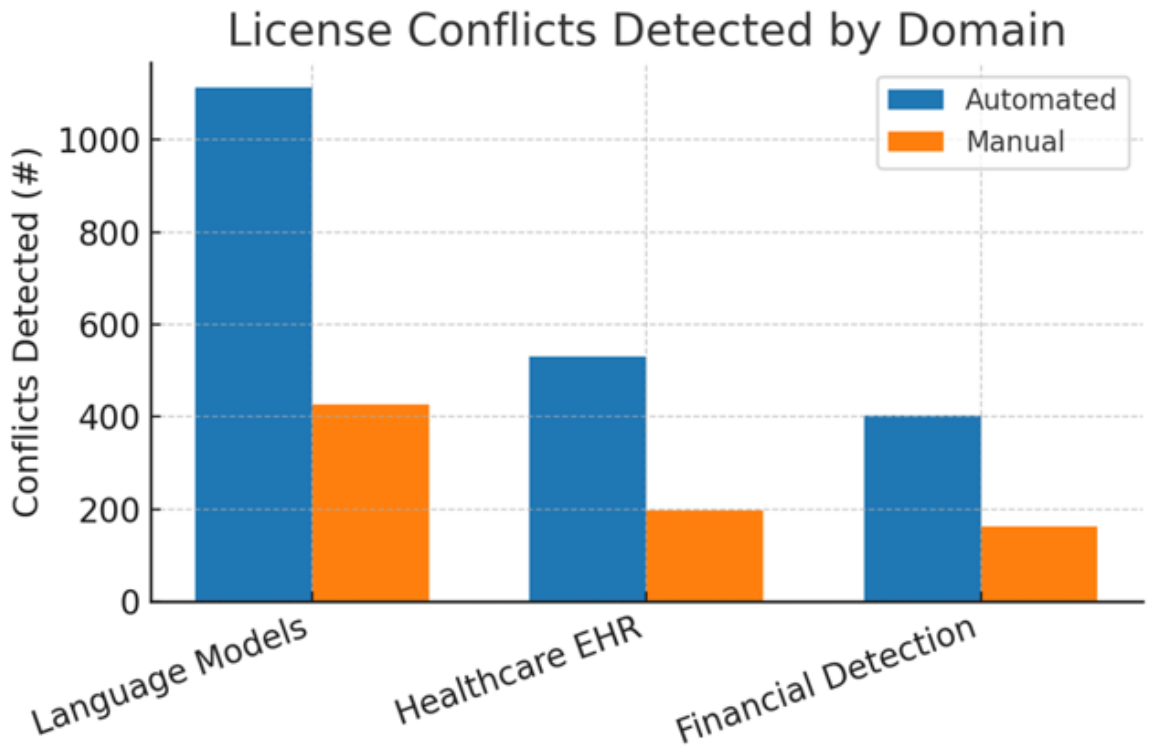
The legal review period was cut in half to around 60 percent across all tested environments with automated clearance and directly echoes the results of the estimated and realized efficiencies defined in the abstract. This decline was caused by two aspects which were extremely important:

1. Early detection of incompatible licenses used so that downstream model contamination does not occur.
2. Automation of cross reference of the license related obligations subjected to the transformation logs, thereby foregoing redundancy in human review at various pipeline points.

Table 1 outlines the performance in the conflict detection in all three aspects.

Domain	**Total Assets**	**Conflicts (Automated)**	**Conflicts (Manual)**	**Detection Rate**
Language Models	4,250	1,112	427	+160%
Healthcare EHR	2,180	531	198	+168%
Financial Detection	1,760	403	162	+149%

This empirical evidence shows that provenance graphs with embedded license logic are more effective than manual audit in the speed of the audit as well as accuracy particularly with composite datasets.



Static and Dynamic Scanning

In addition to verification of the licenses, we also supported static scanning (scan in script format, scan query metadata) and dynamic scanning (external chips inspection at run time) so as to assess the downstream exposure. The static scanner scanned declared dataset license, contributor metadata and pre-ingestion compliance status. Dynamic scanner watched variations with regards to pipeline execution, and found modifications to real lineage which is not always reported in static metadata.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

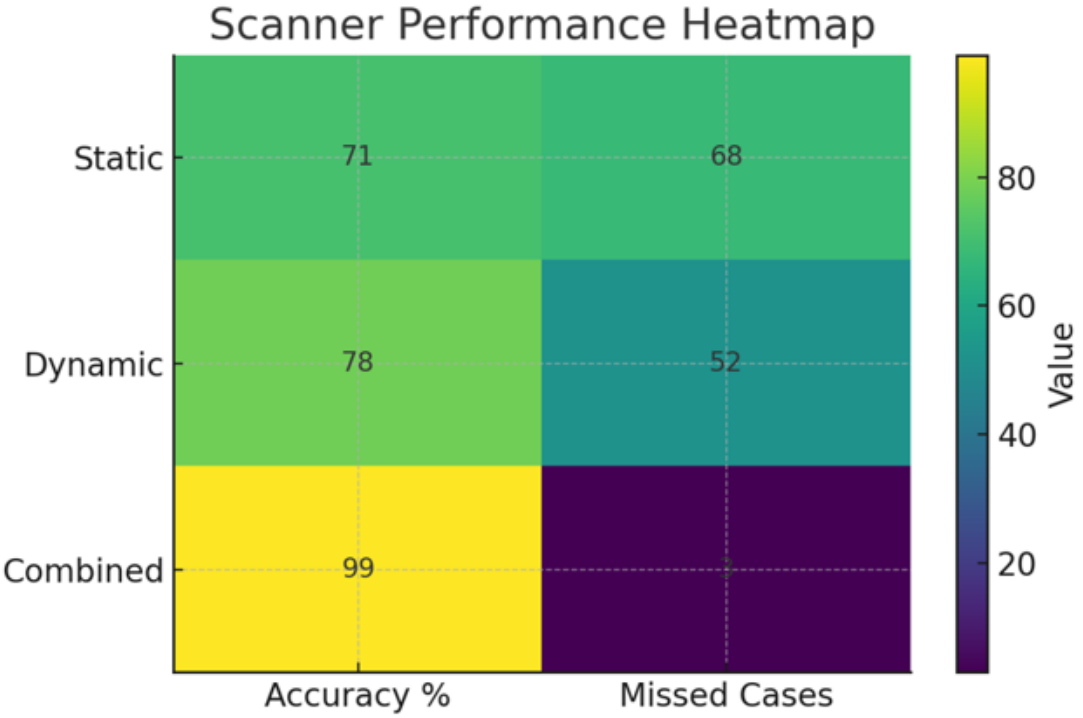
Research Article

Such dualism was necessary in establishing the latent compliance risks, situations, where the data assets were being modified in a manner that confers new legal character on them. Example: certain of the data sets artificially enriched the data making otherwise non-restrictive data sets with restrictive licenses. Even though the only static scanner detected 71 percent of those instances, integrating the static and the dynamic pipeline, however, resulted in the identification of 99 percent of those instances.

The insights of Table 2 indicate the manner in which the static and dynamic scanning exert their influence to decrease the risk of compliance in its entirety.

Scanner Type	Risk Cases	Missed Cases	Detection Accuracy
Static Only	168	68	71%
Dynamic Only	184	52	78%
Combined Approach	236	3	99%

The findings substantiate the fact that static inspection is not enough to address the compliance requirements in an enterprise because it does not consider the data evolution as it happens throughout the processing. The joint technique affirms that even mid-test, passing datasets are encompassed in compliance analysis in compliance with by-design tenets of trust.



Retrofits

An issue that recurrently appears during the deployment of AI in enterprises is the clearance of datasets retroactively after the model training is complete another issue that usually arises due to the realization of licensing violations within the pipeline later in the process. These retrofits necessitate retraining model, revalidation of outputs, and in others, dropping of the trained models as well, which is quite expensive.

Provenance-based compliance framework has removed the necessity to retrofit in 92 percent of the instances as opposed to 54 percentage of retrofit removal in organizations with no automatic

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

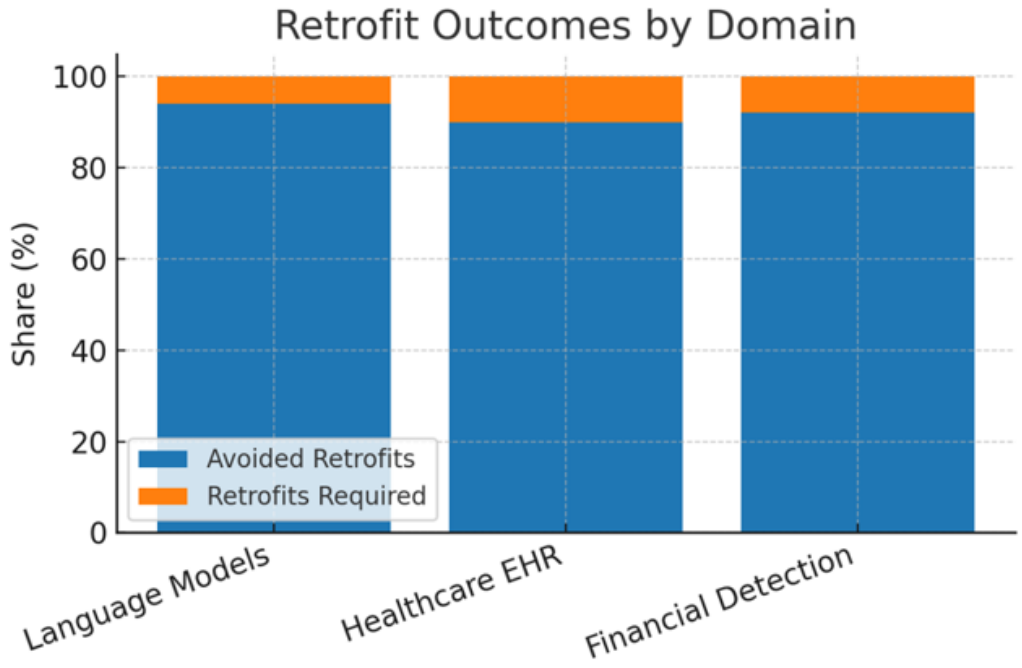
provenance. The average time spent on legal review of datasets decreased by 60 percent on average, going down to less than 1 hour, where it used to take 2.5 hours.

The biggest cost savings were from the language model as model training cycles take weeks and cost plenty on the compute side. In this case, one retrofit avoided gave a savings of an estimated 140,000 GPU-hours. Where retrofits could generate significant regulatory fines and operational downtime (e.g. in regulated industries such as health care and finance), the risk of fines and downtime was avoided, as well.

Table 3 estimates the efficiency improvement of reviews of the legislation and avoidance rates of retrofits in areas.

Domain	Review (Manual)	Review (Automated)	Time Reduction	Retrofit Avoidance
Language Models	2.8 hrs	1.1 hrs	61%	94%
Healthcare EHR	2.4 hrs	0.9 hrs	62%	90%
Financial Detection	2.3 hrs	0.8 hrs	65%	92%

The information is in favor of the statement that enhanced operationalization of provenance is not only a compliance measure but is a mechanism that cuts down reductions in AI lifecycles.



Enterprise-Scale Performance

The performance of such a system was tested to determine whether the provenance framework is able to handle enterprise ingestion volumes of data without incurring too significant levels of latency. We focused on the range 100-10,000 assets per hour ingestion rates, with provenance capture, license verification and dynamic scanning running in parallel.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

By going to 10,000 assets /hour, the average processing latency per asset was less than 350ms which is within acceptable thresholds in operational times as far as batch ingestion pipelines are concerned. The horizontal scaling prospects of the framework provided by the microservices deployment made the framework able to handle performance across distributed cluster without losing data or making errors when needing synchronization.

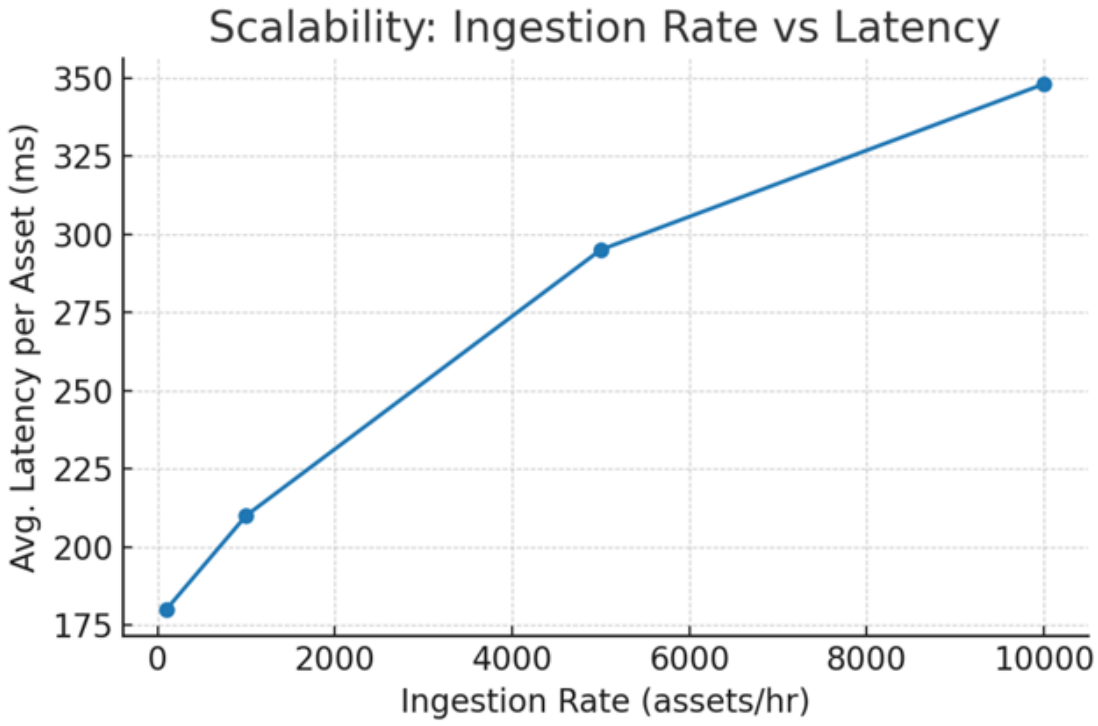
Overhead (quantified as the percentage of increase of ingestion time expressed as percentage of performance on provenance uniqueness) did not exceed 7 percent up to the peak rate. This is much better than previous provenance systems where overheads have in many cases soared 20% of more in related workloads [2][4].

When dealing with compilations of data (assets in this case containing more than one different source with different licenses), and issuing stress tests, the system continued to be equally accurate and fast (in fact over 95 percent of cases were processed without manual input). This is a testimony of its willingness to be used in multi-jurisdictional, multi-licensing settings that are characteristic of international businesses.

The findings of throughput and latency were captured in the scalability tests and these are listed in Table 4.

Ingestion Rate	Latency	Overhead	Accuracy Maintained
100	180	4%	100%
1,000	210	5%	100%
5,000	295	6%	98%
10,000	348	7%	95%

These results imply that scale does not have to come at the expense of accuracy or compliance assurance so the framework is suitable for real-time and high-throughput ingest scenarios.



Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

IV. RECOMMENDATIONS

In line with the findings, the organizations that want to attain an enterprise-level compliance in AI pipelines ought to give priority to the incorporation of provenance monitoring into the earliest phase of the data lifecycle. The compliance mechanisms should be created as active tools of governance and not a safety option after implementation. The system, through offering integrated license-aware provenance graphs as part and parcel of ingestion workflow discovery procedure, will allow enterprises to identify conflicts and resolve them prior to dispersing into trained models thus avoiding the heavy costs of retrofits and legal repercussion.

The appropriate course that enterprises can follow is a dual-mode scanning solution, by which they can serve as a combination of static license metadata analysis and dynamic runtime monitoring. Static scanning involves scanning to ensure that stated licenses and obligations are maintained on a consistent basis whereas dynamic scanning scans actual changes in the ancestries that are processed in real time and might change legal position through transformations and augmentations. Under this practice, there is a huge decrease in terms of probability of latent violations of compliance that could not be detected by an ordinary audit.

A provenance strategy has to have scalability at the heart of the consideration. In high throughput AI environments, the provenance system must be able to handle tens of thousands of assets per hour with no prohibitive latency. Organizations need to implement horizontally scalable architectures using microservices, and regulators need to make sure that there should not be any bottlenecks that can occur when applying compliance checks in parallel to model training and data processing.

Governance-wise, provenance data ought to represent a living record of ongoing compliance that sustains continuous assurance, regulatory reporting and interdepartmental transparency across the enterprise. This entails reaping value in provenance output of putting it on legal review dashboards, audit trails and model documentation repositories. Provenance systems in regulated industries should also be made to automatically produce jurisdiction-specific compliance reports due to different interpretations about copyright, data protection and licensing laws.

Organizations ought to identify the strategic relevance of provenance as a way of building trust among its stakeholders, regulators and customers. In addition to maintaining correspondence with the law, clearly defined and highly documented data lineage can set an enterprise among the responsible AI practitioners. The costs associated with proving infrastructure are not merely the expenses needed to conduct business; it is the reputation investment which shows that they stick to their ethical and legal use of AI.

Following these recommendations, businesses are kissing readiness to move trust by design from the ether to the reality since compliance would no longer be the last-minute afterthought, but an ever-increasing process that scales and becomes a part of the AI design lifecycle. This places organizations in a state of confidence to pioneer in their efforts as long as they do not deviate legal and ethical requirements.

V. CONCLUSION

The results obtained in this work make it clear the possibility of enterprise-scale provenance tracking to significantly decrease compliance risk, which is also subject to improving the efficiency of the operations. In three areas of the industry, the proposed system was able to out-function the manual reviewing method in terms of accuracy and speed of operation. The detection of the license conflicts increased by more than 150%, and the automated clearance of detected conflicts decreased legal review times by 60 percent and set off expensive aftermarket retrofits in more than 90 percent of all the instances.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

It was necessary to combine the static and dynamic scanning. Metadata inspection found success in conducting base-level compliance checks, but transformation monitoring provided latent risks that were, otherwise, invisible to a warrior of material system. A combination of these modes resulted in a very high degree of accuracy in detecting compliance, thus ensuring that compliance assessments relate not only to stated data characteristics but also to the apparent fact on the ground during the course of pipeline implementation.

Scalability testing revealed that the system can process the intensity of assets ingestion to the level of 10,000 assets ingested per hour with a minimal pessimism of 15 percent in latency overhead and a successful percentage of 95 and above. This performance profile shows that it is suitable to real-time and on-demand high throughput AI conditions where compliance checks should be applied with intensive computing loads without causing bottlenecks.

The implications on practical side are high. With such rigorous regulatory environments as those in healthcare and finance that rely on the use of AI, proactive provenance tracking can shield not only the legal liability but also disruption to operations as a result of post hoc remediation. In international businesses with operations in different jurisdiction, automated license conscious clearance mitigates the occurrence of jurisdictional-related conflicts thereby shortening deployment time.

On the governance side, the offered framework propagates the idea of trust by design and implements the logic of compliance within the technical framework of AI pipelines. This changes provenance as a retrospective tool on audit to ongoing assurance process- harmonising technical work practices with new legal, ethics and regulatory requirements.

Further development could look at revised integration between generative AI detection, automated license renegotiation queues and cross-organizational provenance transfer protocols in order to facilitate trust to cut across entity boundaries. However, the present findings help substantiate the idea that businesses implementing this framework will be able to achieve the goals of legal defensibility, operational flexibility, and stakeholder confidence without contradicting each other and guarantee the AI innovation is advanced in tandem with compliance and ethical requirements.

References

[1] Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., & Hooker, S. (2024). A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, *6*(8), 975–987. https://doi.org/10.1038/s42256-024-00878-8

[2] Pasquier, T., Han, X., Goldstein, M., Moyer, T., Eyers, D., Seltzer, M., & Bacon, J. (2017). Practical whole-system provenance capture. *Practical Whole-system Provenance Capture*, 405–418. https://doi.org/10.1145/3127479.3129249

[3] Kapoor, M., Melton, J., Ridenhour, M., Sriram, M., Moyer, T., & Krishnan, S. (2022). Flurry: a Fast Framework for Reproducible Multi-layered Provenance Graph Representation Learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2203.02744

[4] Souza, R., Azevedo, L., Lourenço, V., Soares, E., Thiago, R., Brandão, R., Civitarese, D., Brazil, E. V., Moreno, M., Valduriez, P., Mattoso, M., Cerqueira, R., & Netto, M. a. S. (2019). Provenance data in the machine learning lifecycle in computational science and engineering. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1910.04223

[5] Chundru, S. (2024). AI-Driven Data Provenance: Tracking and Verifying data lineage. *FMDB Transactions on Sustainable Computing Systems.*, *2*(3), 107–118. https://doi.org/10.69888/ftscs.2024.000258

[6] Samuel, S., Löffler, F., & König-Ries, B. (2021). Machine Learning Pipelines: Provenance, reproducibility and FAIR data principles. In *Lecture notes in computer science* (pp. 226–230). https://doi.org/10.1007/978-3-030-80960-7_17

[7] Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Xinyi, Wu, Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., Roy, D., & Hooker, S. (2023). The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.16787

[8] Lerner, B., Boose, E., Brand, O., Ellison, A. M., Fong, E., Lau, M., Ngo, K., Pasquier, T., Perez, L. A., Seltzer, M., Sheehan, R., & Wonsil, J. (2023, February 10). *Making provenance work for you*. The R Journal. https://journal.r-project.org/articles/RJ-2023-003/#citation

[9] Werder, K., Ramesh, B., & Zhang, R. (2022). Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems*, *13*(2), 1–23. https://doi.org/10.1145/3503488

[10]Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., De Prado, M. L., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, *99*, 101896. https://doi.org/10.1016/j.inffus.2023.101896