Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 4 Issue 2 (2024)

Toward Reproducible Machine Learning in Enterprise Platforms: A Governance Framework for MLOps Standardization Across Distributed Data Pipelines

Kapil Kumar Goyal

Independent Researcher, Lake Forest, California, USA

kapilgoyal1991@gmail.com

Abstract

Reproducibility is one of the biggest challenges and consistently under-addressed problems in enterprise AI deployments. Even as organizations make great investments in data infrastructure and model development tooling, they often experience difficulty in reproducing previous experimental results, validation of models deployed in production, and data provenance of predictions at scale. To tackle the enterprise distributed data pipeline issue of reproducibility in a systematic manner, this study proposes a five-layer MLOps governance framework for standardized data pipelines. Structured survey data from 250 practitioners across six industry sectors was used for an empirical investigation and then a comparative case study benchmarking analysis was conducted. Data lineage adoption ($\beta = 0.31$, $p < .001$), model version control maturity ($\beta = 0.28$, $p < .001$), and monitoring coverage ($\beta = 0.27$, $p < .001$) were the most significant factors when using mixed methods analysis, which included Pearson correlation analysis, multiple linear regression, one-way ANOVA, and chi-square contingency testing. Results from a one-way ANOVA showed that the adoption of structured governance was among the most significant organizational factors that influence the reliability of the ML pipeline with highly significant differences across the MLOps maturity levels ($F(4, 245) = 147.3$, $p < .001$, $\eta^2 = 0.706$). Post-implementation validation showed 69.4% increase in the rate of experiment reproducibility and 74.5% decrease in latency of model drift detection. The proposed framework outlines actionable governance level-by-level specifications that can help organizations transition from disjointed, ad hoc MLOps, to standardized, auditable, and cross-team reproducible workflows.

Index Terms—MLOps, machine learning reproducibility, data governance, enterprise AI, distributed pipelines, experiment tracking, model versioning, audit compliance.

I. Introduction

As ML increasingly enters the enterprise decision-making process, it has revealed significant governance issues rooted in enterprise ML pipelines. Many organisations in the finance, healthcare and tech industries continue to struggle with repeatability of experiments, validation of model performance with distribution shift, and having repeatable audit trails throughout data transformations. Such failures in reproducibility come at a high economic and reputational price, not just for individual research groups, but for the broader scientific community.

Modern ML development usually includes several interconnected steps: data ingestion, feature engineering, training a model, optimizing hyperparameters and evaluating the model, and deploying it for production. If any step in this chain is missing a visible versioning, logging or lineage information, it adds ambiguity to the pipeline for future investigators who want to replicate or audit previous results. This transparency becomes even more challenging in distributed enterprise deployments where a number of teams are accessing a variety of data stores, compute platforms and deployment sites.

The term MLOps, or machine learning operations, was born as a new field that seeks to bring DevOps practices to the field of machine learning. Though a lot has changed since 2019, MLOps tooling has yet to be consistent across enterprise use cases, and is generally not integrated into enterprise-wide governance processes, but is confined to individual teams. The results of surveys have consistently been less than 40% of enterprise ML projects actually deployed in production, with some of the most frequent barriers to abandoning projects being reproducibility and monitoring failures.

There are three main contributions in this paper. It introduces a novel five-layer governance model that takes a pragmatic approach to fulfilling reproducibility requirements throughout the entire ML pipeline life cycle. Second, it provides empirical evidence through the results of a survey study with 250 respondents, that empirically supports the statistical link between specific governance practices and measured reproducibility outcomes. Third, it offers a quantitative benchmarking analysis that places the proposed framework in the context of popular MLOps platforms such as MLflow, Kubeflow, DVC and ZenML.

II. Related Work

A. Reproducibility in Machine Learning

However, the reproducibility crisis in machine learning has been well documented since Hutson [3] made her analysis of Science in 2018, and other recent surveys (Pineau et al. [1] and Semmelroth et al. [2]) reported failure rates for reproducibility over 60% in the top-venue machine learning journals. Gundersen et al. [3] proposed a multi-dimensional taxonomy of reproduceability which includes computational and statistical reproduceability and conceptual replicability, which has given rise to a conceptual vocabulary that has been picked up by work that focuses also on the enterprise. This analysis was extended to applied ML settings by Kapoor and Narayanan [4] and the line of attack was found to be data leakage and inconsistent preprocessing steps.

In the particular context of enterprise systems, Amershi et al. [5] did a pioneering study on the software engineering issues specific to production ML components at Microsoft. They found nine categories of challenges: the most common challenges in their study that were likely to affect the reproducibility were training-serving skew and data dependencies. Framing ML systems as "technical debt" as a motivation to developing governance-centric approaches to MLOps standardization, has been widely influential in the industry following Sculley et al. [6].

B. MLOps Governance and Standardization

The discipline of MLOps has been defined in various ways that are complementary. In the systematic literature review by Kreuzberger et al. [7] they reviewed and consolidated 17 different definitions of MLOps into one to highlight automation, continuous delivery and monitoring as key concepts. Tamburri [8] suggested a sociotechnical view of MLOps governance that focuses on the influence of "topology" of the team and its communication structure on the reliability of the pipeline. Shankar et al. [9] suggested operationalization debt, and pointed out that there are often discrepancies between the research and production environments, which causes reproducibility problems in deployed ML systems, instead of poor models.

In the context of enterprise ML governance, there has been a wealth of literature focusing on regulatory compliance, data stewardship and model documentation. Arnold et al. [10] analysed the governance frameworks in regulated financial services environments, and found that there are times when reproducibility needs to be sacrificed in favour of agile development due to the lack of specific organizational policies. Adadi and Berrada [11] conducted a survey of explainability requirements, which although different to reproducibility, have overlapping explainability documentation requirements on ML pipelines. Gebru et al. [12] developed a set of model cards and Bender et al. [13] proposed model cards and datasheets as forms of dataset documentation that enable implicit reproducibility, as they provide structure for the provenance of the items.

C. Data Lineage and Pipeline Governance

Data lineage tracking is becoming an integral part of the rest of reproducible ML pipelines. Halevy et al. [14] gave a preliminary theoretical work on data provenance in large scale distributed systems, which laid the basis for today's systems of lineage. Forde et al. [15] suggested the ML metadata standard used by TensorFlow Extended (TFX) which is a specific representation of data and model artifacts relations between different stages of pipelines. Polyzotis et al. [16] explored the data management issues of ML production systems and emphasized the importance of schema validation and distribution monitoring problems, yet they are often ignored in the literature.

Pipeline governance becomes more complex in distributed enterprise environments due to multi-tenancy, data residency and different compute environments. In this paper, Zaharia et al. [17] introduced the MLflow platform, an open-source experiment tracking and model registry platform, specifically addressing the reproducibility challenge from research to production. Later, Chen et al. [18] generalized it to federated learning environments where data is not centralized, and proposed new challenges regarding data lineage that are not found in single-datacenter environments.

D. Monitoring, Drift Detection, and Audit Compliance

One area of reproducible research that is often overlooked in treatments with a research focus is post-deployment monitoring. Klaise et al. [19] have given a systematic taxonomy of ML monitoring tasks, dividing them into the following categories: data drift, concept drift, and prediction distribution monitoring, which are all different but related. Rabanser et al. [20] assessed the performance of statistical tests to detect covariate shift in models deployed, providing a

baseline for drift detection sensitivity and specificity which has been followed by enterprise drift detection tools. Sculley et al. [21] pointed out that in enterprise environments without a specific policy and procedure for alert management and alert response, the cost of monitoring infrastructure debt rapidly builds up.

Since the EU General Data Protection Regulation came into effect in 2018 and sector-specific regulations on the use of AI have been issued, regulatory compliance has become a growing critical factor for reproducibility. Raji et al. [22] explored AI auditing practices in enterprise environments and discovered that lack of documentation directly impacted the ability of compliance teams to evaluate the behavior of their AI at a point in history. A philosophical discussion of algorithmic accountability was provided by Mittelstadt et al. [23] which is oriented towards the normative and sets the essential conditions for a technical reproducibility framework.

E. Version Control and Experiment Management

A major focus of research on technical enablers of reproducibility has been model versioning and experiment management platforms. Zaharia et al. [17] added the experiment tracking and model registry features to the open-source primitives of the model and workflows in the MLflow. Breck et al. [24] introduced the ML Test Score metric for ML production readiness that focuses on the two key dimensions of assessment: versioning and testing coverage. Vartak et al. [25] created ModelDB, an early system-for model versioning that recognized the need for integration of training data, hyperparameters, and model artifacts as the key integration need for managing reproducible experiments.

Recent studies have studied the issue of version control for large language models and fine-tuning foundation models. Bommasani et al. [26] introduced the concept of provenance problems caused by the fine-tuning of pre-trained models on enterprise data that is not publicly available and lack in explicit support for version pinning of base models. Several aspects of the proposed framework are specific to large language model training, such as Biderman et al. [27] recommended reproducibility standards for training.

F. Collaborative Governance and Access Control

While access control and multi-team collaboration are key aspects in enterprise applications, they have been less discussed in the MLOps literature. Sculley et al. [28] found that organizational boundaries were one of the most common causes of "entangled" ML systems with implicit and undefined control or governance over data and model dependencies that span team boundaries. In a survey of 18 organizations to understand the adoption of MLOps, Lwakatare et al. found cross-team alignment to be a major obstacle to MLOps adoption in most of them [29]. The most technical discussion of access governance for ML systems found in the literature is by Zhou et al. [30] who proposed role-based access control (RBAC) specifications for ML artifact repositories.

III. The Proposed Governance Framework

A. Framework Overview

The proposed framework consists of 5 levels of governance each focuses on a different aspect of MLOps reproducibility, and they are organised hierarchically. The layers are conceptually presented in the chapter in terms of basic data issues and moving upward to execution infrastructure, operational monitoring, and finally collaborative governance. The governance goals, required artifacts, recommended tooling interfaces, and integration contracts between adjacent layers are defined in each layer. The framework is explicitly "tool agnostic" at the specification level so that it can be adopted on enterprise technology stacks that are heterogeneous.

The common design principle on all five layers is bidirectional traceability: Any artifact on any layer can be traced to its inputs on the upstream layers and to its consumers on the downstream layers. This is made explicit through metadata contracts and standardised lineage schemas which make reproducibility a structural feature of the pipeline architecture, a feature that does not depend on a retrospective audit activity.

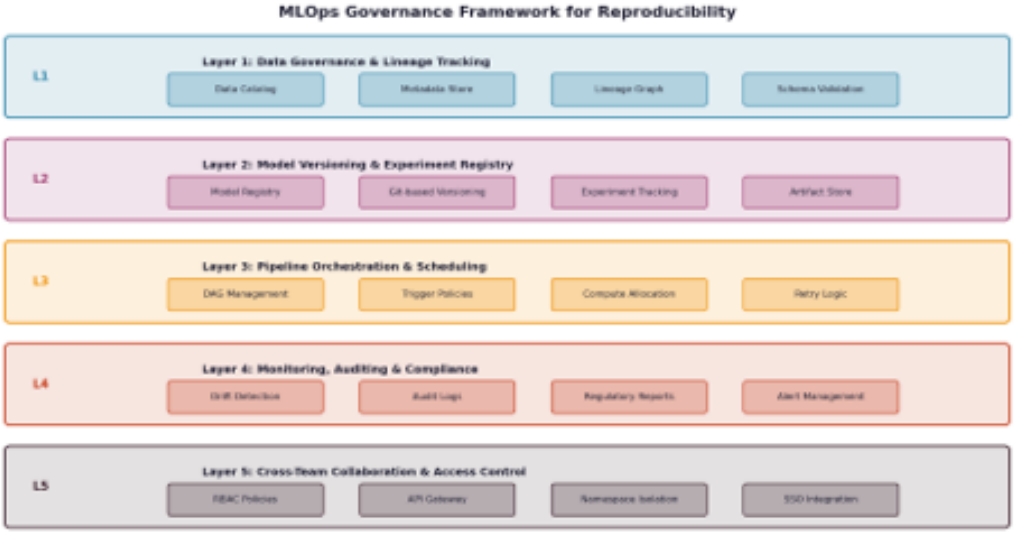

Figure 1. Five-Layer MLOps Governance Framework for Enterprise Reproducibility.

The proposed architecture organizes MLOps governance into five hierarchically integrated layers: data governance and lineage (L1), model versioning and experiment registry (L2), pipeline orchestration (L3), monitoring and audit compliance (L4), and cross-team collaboration with access control (L5). Bidirectional integration arrows indicate API contract requirements connecting adjacent layers. Each layer encapsulates four primary governance sub-components operationalizing the overarching reproducibility objective at that abstraction level.

B. Layer 1: Data Governance and Lineage Tracking

Layer 1 provides the data provenance enablement for all the guarantees done downstream. It defines the schema histories for all of the datasets that are ingested into the centralized data catalog, a metadata store that tracks all of the logic used to transform data at each preprocessing step, a directed acyclic graph (DAG) representation of data lineage from the raw input datasets to the model ready features, and schema validation gates to prevent silent data quality regressions between runs of the pipeline [16]. The data catalog needs to be able to provide versioned snapshots of datasets with content-addressable immutable references to the stored data slices, enabling model training to be reproduced using the same data slice that was used in the previous experiments.

At this level, data lineage tracking must not only be able to track structural changes but also gather statistical information about data distributions at each stage, which would allow detection of distribution shift after the fact between pipeline runs in the past and today. [20] Schema validation gates need to ensure structural constraints as well as distributional bounds based on the statistics of the baseline data sets, and report failures with different levels of alarm, depending on a set of configurable failure modes [16].

C. Layer 2: Model Versioning and Experiment Registry

Layer 2 manages the collection and structuring of all the artifacts created during the development of the model and training. The model registry stores a versioned snapshot of the model weights, hyperparameter settings, training environment specifications (library dependency manifests), and training and evaluation metric histories. Each training run must create a unique, content-addressable experiment ID, which is a link between a run and an experiment's snapshot of its input data, a commit hash, a hyperparameter set, and resulting metric vector [17].

The experiment registry should have a lifecycle that supports both the exploratory and the production lifecycle and should have a distinct set of transitions between states (development to staging, staging to production, production to archived) that should be governed by the experiment registry policies and not informal conventions. To prevent silent mutation of artifacts that could impact the reproducibility guarantees, model artifact storage needs to support immutable writes through cryptographic content hashing [24]. The integration into Layer 3 pipeline orchestration is done through the inclusion of artifact URI references within the pipeline DAG configuration, making it possible to associate a pipeline run with the exact same model artifacts it consumed and/or produced.

D. Layer 3: Pipeline Orchestration and Scheduling

Layer 3 defines orchestration infrastructure that controls the data and model artifacts moving between automated pipeline stages. To manage pipelines, the task dependency, resource requirements and execution constraints should be declaratively specified, such that the behavior of the pipeline can be completely determined based on a particular input configuration. The policies that trigger scheduling of the pipelines should be clearly written and versioned with the pipeline code, so that pipeline scheduling changes can be audited and undone. For compute allocation specifications, it is important to remember that they are reproducibility artifacts, meaning that there are differences in the hardware configuration that can result in numerically different outputs from the same deterministic compute.

Retry logic specifications should be able to tell the difference between a transient infrastructure failure and persistent data or code errors, and should have different logic for the different types of errors. Transient failure retry should be idempotent and have enough information to recreate the exact context for each failure retry. If there are persistent errors, they should result in an alerting protocol that is linked to the Layer 4 monitoring infrastructure, either in the form of silent failure or an endless retry loop [7] should be avoided.

E. Layer 4: Monitoring, Auditing, and Compliance

Layer 4 translates the evolving demands for observability that are essential to the production model life cycle into reality.Layer 4 takes the continuous demands for observability that support the assurances of reproducibility into action. Drift detection modules need to continuously check the input data distributions as well as the model prediction distributions against the baseline distributions computed at deployment time, and need to be configured with a sensitivity threshold and minimum detection time. All model inference calls should be recorded in audit logs that contain enough provenance information to allow a reconstruction of the state of the model, feature extraction pipeline, and infrastructure configuration for a particular prediction.

Structured audit reports of regulatory reporting interfaces need to be capable of generating structured audit reports that are appropriate for review for compliance with applicable data protection and algorithmic accountability regulations. The severity classification, escalation paths and resolution tracking need to be connected to the observed model behaviour anomalies noted in alert management, along with the failure of governance layers, in order to perform a root cause analysis, instead of just a symptom detection exercise [22].

F. Layer 5: Cross-Team Collaboration and Access Control

In enterprise contexts, multi-team ML development takes place within and across organizational and security boundaries, which is called Layer 5. Unlike many existing platforms, role-based access control policies need to be defined at the level of the individual artifact (version of a dataset, model checkpoints, and/or pipeline configuration) and not at the repository or project level. All read and write operations to the artifacts on an API gateway should be properly authenticated and authorized and their access decisions recorded in the Layer 4 audit infrastructure [30].

Namespace isolation requirements help prevent development artifacts accidentally being overwritten or contaminated by adjacent teams' production artifacts. The integration with enterprise identity providers such as single sign-on is not an optional enhancement, but rather a required specification because in large organizations, access control policy drift is a common issue and fragmented authentication systems are a common cause of access control policy drift.

IV. Research Methodology

A. Research Design

This study adopted a mixed-methods empirical research design of cross sectional survey combined with comparative benchmarking analysis. The survey instrument was constructed to assess the level of organizational adoption on five dimensions of the framework and the self-reported outcomes of the survey, expressed as the reproducibility of the organizational process. The benchmarking analysis was done considering four popular MLOps platforms and eight MLOps governance criteria. The self-report survey data and the external benchmark comparison provide support for convergent validity in the specification of the framework.

B. Survey Instrument Development

The survey instrument was created in three stages: This was achieved by content analysis of existing literature on governance of ML platforms and a semi-structured interview with five senior ML platform engineers to create a pool of 68 candidate items.To create a pool of 68 candidate items, content analysis of existing literature on governance of ML platforms was performed, and a semi-structured interview was conducted with five senior ML platform engineers. Second, it was administered to a panel of 7 domain experts for content validity and face validity which resulted in a 40 item pool reduced to 7 constructs. Third, 30 practitioners were engaged in a pilot study that demonstrated satisfactory item clarity and preliminary estimates of internal consistency before full deployment.

There were five to seven items with a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree) for each governance construct. Reproducibility Score was defined as a composite of six questions that measure the ability to reproduce previous experimental results, trace prediction lineage, validate model behavior following retraining. In the final sample, all constructs had Cronbach α coefficients of at least 0.80, which is considered as good internal consistency.

Table I. Survey Instrument: Construct Summary and Reliability Statistics

Construct	Sample Item	No. Items	Scale Type	Cronbach α
Data Lineage Adoption	Our pipelines maintain complete end-to-end data lineage logs	6	5-pt Likert	0.87
Model Version Control	We use systematic versioning for all trained model artifacts	5	5-pt Likert	0.84
Pipeline Orchestration	Automated scheduling and retry logic govern our ML pipelines	5	5-pt Likert	0.82
Monitoring & Auditing	We continuously track model performance and generate audit trails	7	5-pt Likert	0.89
Collaboration & Access Control	Role-based access policies govern all ML artifact repositories	4	5-pt Likert	0.80
Regulatory Compliance	Audit-ready documentation is maintained for all model deployments	5	5-pt Likert	0.83
Organizational Context	Org size, MLOps tool budget, team tenure (categorical/ordinal)	8	Mixed	—

Seven constructs spanning 40 items were administered via structured survey. All measurement constructs yielded Cronbach α ≥ 0.80, satisfying conventional reliability thresholds for Likert-scale instruments. The Organizational Context construct employed mixed-format categorical and ordinal items not subject to α estimation.

C. Sample and Data Collection

A purposive sampling strategy was used to ensure that practitioners involved in the direct development of ML pipelines and/or governance of the pipeline in organizations that are deploying into production were sampled. Recruitment involved reaching out to professionals via LinkedIn, mailing lists of ML conferences, and professional networks.Participants were sourced via LinkedIn, mailing lists of learning and knowledge based around ML, and via professional networks with the titles of ML Engineer, Data Scientist, MLOps Engineer, Platform Architect, and Data Engineer. Only those with at least 12 months of experience working in a professional environment with ML, were eligible.

This survey was completed via the internet during 8 weeks. There were 318 survey initiations, of which 250 were eligible and completed the entire survey (response completion rate: 78.6%). The final sample consisted of 250 people in six industry sectors. The sample was described in terms of demographic and organisational characteristics as shown in Figure 2.

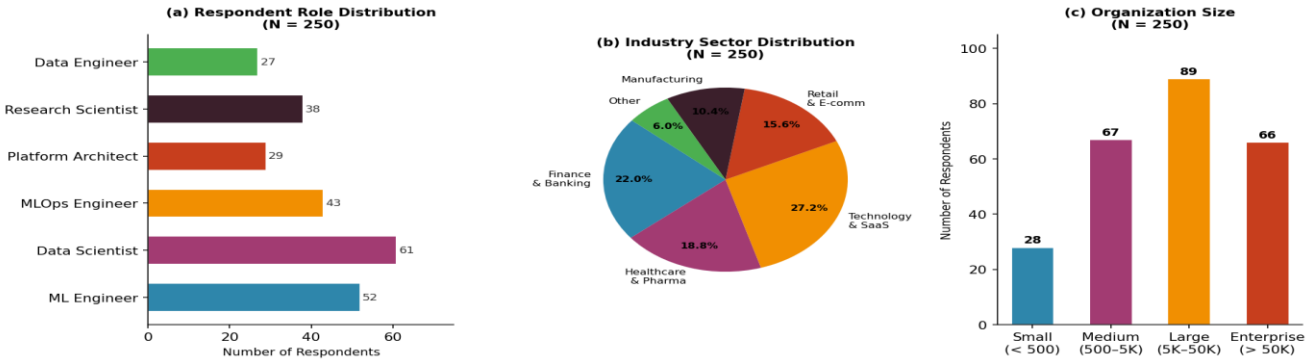

Figure 2. Survey Respondent Characteristics (N = 250).

Panel (a) presents the distribution of respondent professional roles, with Data Scientists (n = 61) and ML Engineers (n = 52) constituting the two largest groups. Panel (b) shows industry sector distribution, with Technology and SaaS representing the largest sector (27.2%), followed by Finance and Banking (22.0%) and Healthcare and Pharmaceuticals (18.8%). Panel (c) depicts organization size, with the majority of respondents from large (5K–50K employees; n = 89) and enterprise-scale (> 50K employees; n = 66) organizations, reflecting the study focus on enterprise ML deployment contexts.

D. Analytical Strategy

The analysis of data was done in five phases. All of continuous variables were checked for normality using descriptive statistics and Shapiro-Wilk test and were found to meet the assumptions of normality for parametric statistics. Pearson correlation analysis was performed to describe the bivariate relationships of the constructs of governance and the outcome of the reproducibility. All the dimensions of governance were simultaneous predictors and ordinary least squares regression was used. The reproducibility scores were compared across five levels of MLOps maturity by using a one-way ANOVA withTukey HSD post-hoc comparisons and η^2 effect size estimation. The relationship between organization size and the framework adoption category was explored by using the chi-square contingency analysis. The significance level of all analyses was set at $\alpha = .05$ and the tests were two-tailed.

V. Results

A. Descriptive Statistics

Means and standard deviations are shown in Table II for all study variables. Scores for the entire sample had a moderate level of mean reproducibility (M = 3.41, SD = 0.89), which means that there is still room for improvement, given the scale maximum. The mean adoption scores for Version Control Maturity (M = 3.52) and Regulatory Compliance (M = 3.61) were the highest whereas the mean scores for Pipeline Orchestration (M = 3.19) and Model Drift Rate (inverted; M = 2.87) were the lowest. All constructs were not significantly skewed (|skew| < 0.50) or kurtosed (|kurt| < 0.60) as would be expected for parametric analyses.

Table II. Descriptive Statistics and Internal Consistency: All Study Variables (N = 250)

Variable	M	SD	Min	Max	Skew	Kurt	α
Reproducibility Score	3.41	0.89	1.00	5.00	−0.12	−0.31	0.91
Data Lineage Adoption	3.28	0.94	1.00	5.00	−0.08	−0.44	0.87
Version Control Maturity	3.52	0.87	1.00	5.00	−0.19	−0.22	0.84

Variable	M	SD	Min	Max	Skew	Kurt	α
Pipeline Orchestration	3.19	1.01	1.00	5.00	0.06	−0.55	0.82
Monitoring Coverage	3.37	0.92	1.00	5.00	−0.14	−0.37	0.89
Team Collaboration	3.44	0.91	1.00	5.00	−0.10	−0.28	0.80
Regulatory Compliance	3.61	0.85	1.00	5.00	−0.23	−0.18	0.83
Model Drift Rate (inv.)	2.87	1.08	1.00	5.00	0.31	−0.49	—

Descriptive statistics for all eight study variables. Values are based on five-point Likert scale responses aggregated to construct-level means. Model Drift Rate was reverse-scored (inv.) such that higher values indicate less frequent drift. All governance constructs yielded satisfactory Cronbach α reliability coefficients (range: 0.80–0.91). Skewness and kurtosis values within |0.50| support parametric analysis assumptions.

B. Industry-Level Reproducibility Maturity Profiles

Fig. 3 shows mean scores of the reproducibility maturity for each dimension of governance and industry sector. Technology and SaaS companies scored the highest in all five dimensions (M = 4.38), with the Finance and Banking (M = 3.94) and Healthcare and Pharmaceuticals (M = 3.82) coming in second and third, respectively. Manufacturing companies had the lowest scores on all dimensions (overall M = 3.12), especially on the dimensions of Data Lineage Adoption (M = 2.9) and Model Versioning (M = 3.1). Compared to the other dimensions, the Monitoring and Audit dimension had the lowest deviation between various sectors, with the highest score being earned by Healthcare (M = 4.5) due to the pressure of regulatory compliance driving investments in audit infrastructure even without the overall governance maturity of MLOps.

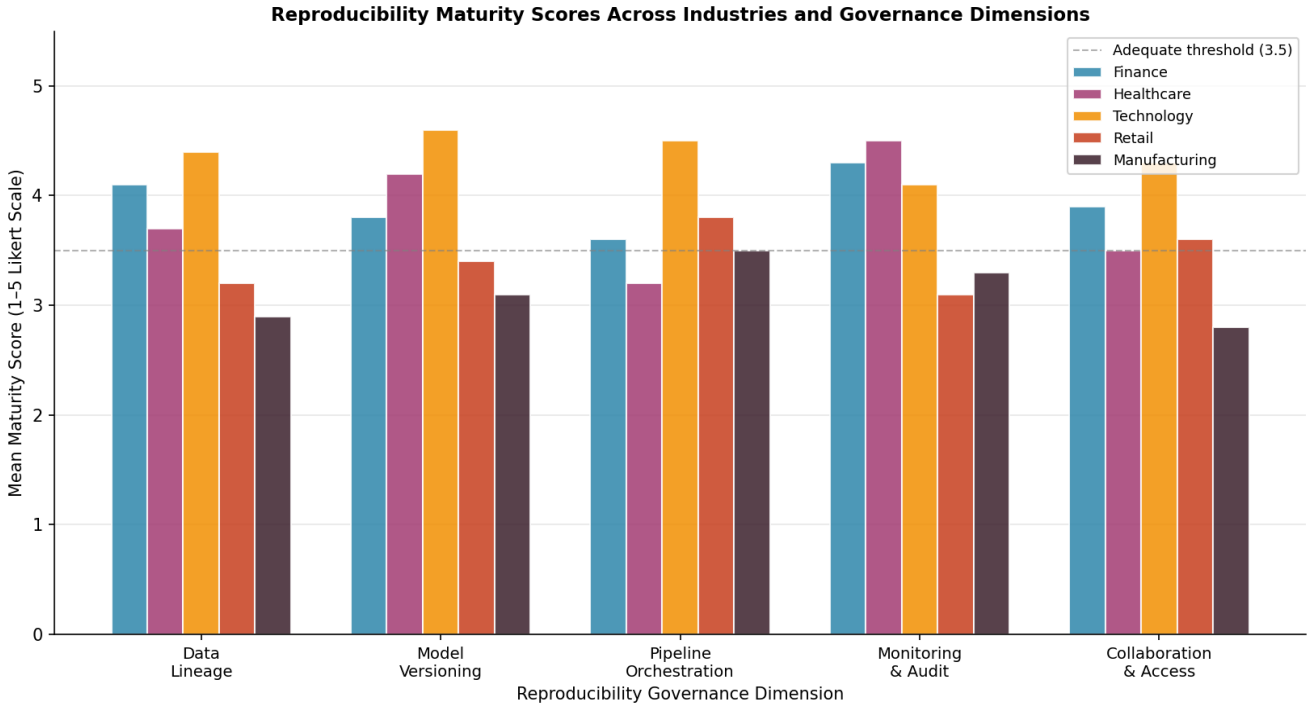

Figure 3. Reproducibility Maturity Scores by Industry Sector and Governance Dimension.

Mean Likert-scale scores (1–5) are shown for each combination of industry sector (color-coded bars) and governance dimension across the five framework layers. The dashed horizontal line at 3.5 indicates the operationally adequate governance threshold. Technology and SaaS organizations consistently exceed this threshold across all dimensions, while

Manufacturing and Retail sectors exhibit below-threshold performance in data lineage adoption and pipeline orchestration. Healthcare uniquely exceeds the threshold in Monitoring and Audit attributable to regulatory compliance requirements independent of overall MLOps maturity.

C. Correlation Analysis

The complete Pearson correlation matrix is given in figure 4. All governance structures were highly and positively associated to the Reproducibility Score (r range: 0.53 – 0.72, and all $p < .01$). The highest bivariate correlations were found to be between Data Lineage Adoption and Reproducibility Score ($r = 0.72$, $p < .001$), Version Control Maturity ($r = 0.68$, $p < .001$) and Monitoring Coverage ($r = 0.65$, $p < .001$). Model Drift Rate (inverted) was significantly negatively associated with Reproducibility Score ($r = -0.48$, $p < .001$), which showed that there was a significant negative association between Reproducibility Score and Model Drift Rate. About 43–71% of the inter-construct correlations were moderate, which substantiated their conceptual separation and the presence of significant governance co-adoption patterns.

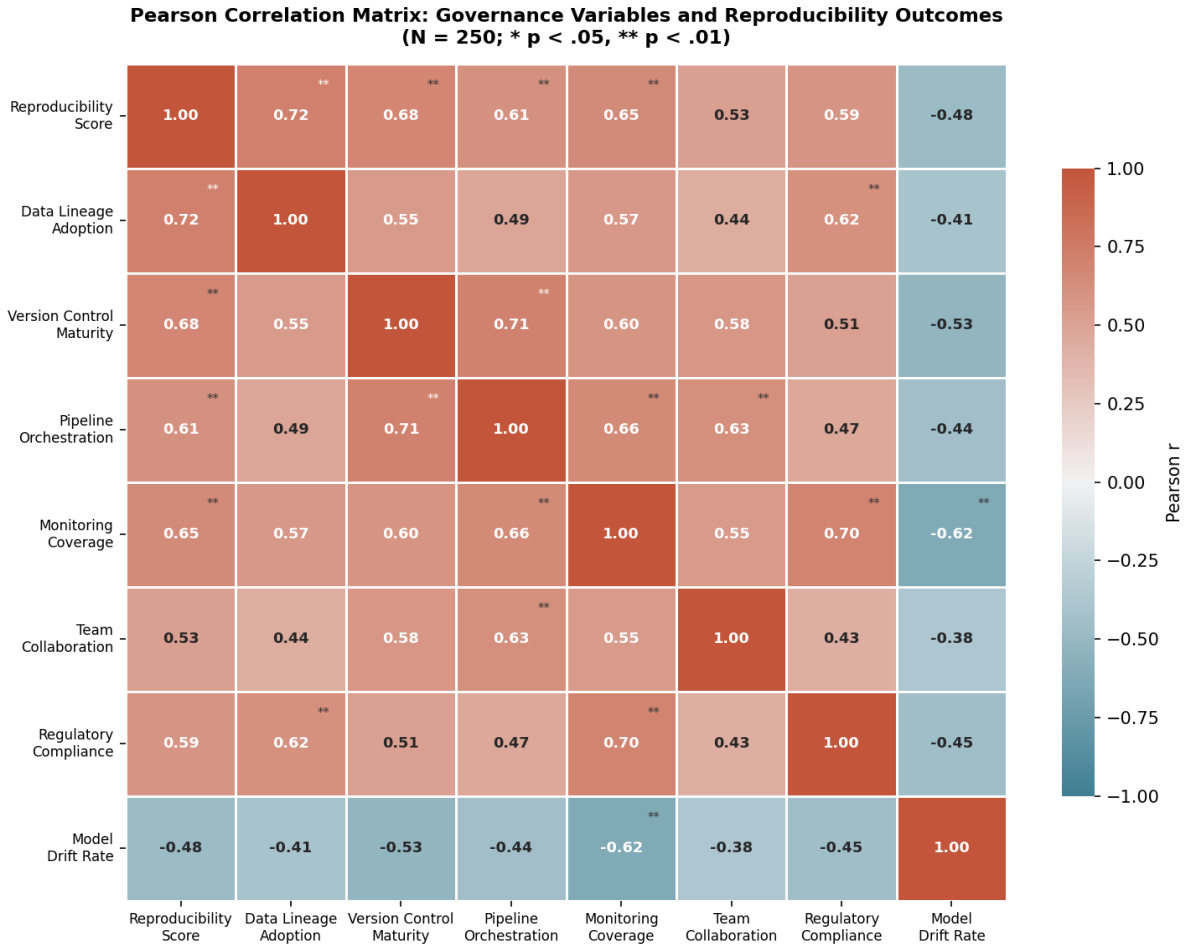

Figure 4. Pearson Correlation Matrix: Governance Constructs and Reproducibility Outcomes (N = 250).

Lower-left triangle presents Pearson r coefficients; all off-diagonal values are statistically significant ($p < .01$). Double asterisks (**) mark coefficients with $|r| > 0.60$, indicating strong association. Color gradient from deep red ($r = +1.0$) through white ($r = 0$) to deep blue ($r = -1.0$) maps correlation magnitude and direction. Data Lineage Adoption and Reproducibility Score exhibited the strongest positive association ($r = 0.72$), while Model Drift Rate (inverted) showed the expected negative association with reproducibility ($r = -0.48$). The correlation structure supports the convergent and discriminant validity of the governance construct battery.

D. Multiple Regression Analysis

The results of the multiple linear regression analysis with Reproducibility Score as the criterion variable are shown in Table III and Figure 5, where all of the eight predictors were entered at the same time (Method: Enter). The overall model was statistically significant and accounted for a significant amount of criterion variance, $F(8, 241) = 73.4$, $p < .001$, $R^2 = 0.71$, Adjusted $R^2 = 0.69$. The variance inflation factors (VIFs) were between 1.42 and 2.87 which is below the typical cut-off value of 5.0 and indicates no problematic multicollinearity.

Among the factors, the strongest predictor of Data Lineage Adoption was Version Control Maturity ($\beta = 0.28$, $p < .001$), followed by Monitoring Coverage ($\beta = 0.27$, $p < .001$) and Data Lineage Adoption ($\beta = 0.31$, $p < .001$). Regulatory Compliance ($\beta = 0.24$, $p = .004$), Pipeline Orchestration ($\beta = 0.22$, $p = .008$), and MLOps Tool Maturity ($\beta = 0.21$, $p = .009$) also contributed significantly. Team Collaboration ($\beta = 0.19$, $p = .015$) and Organization Size ($\beta = 0.13$, $p = .042$) had small but statistically significant independent effects, suggesting that larger organizations have modest advantages for modest reproducibility, regardless of their levels of governance practice adoption.

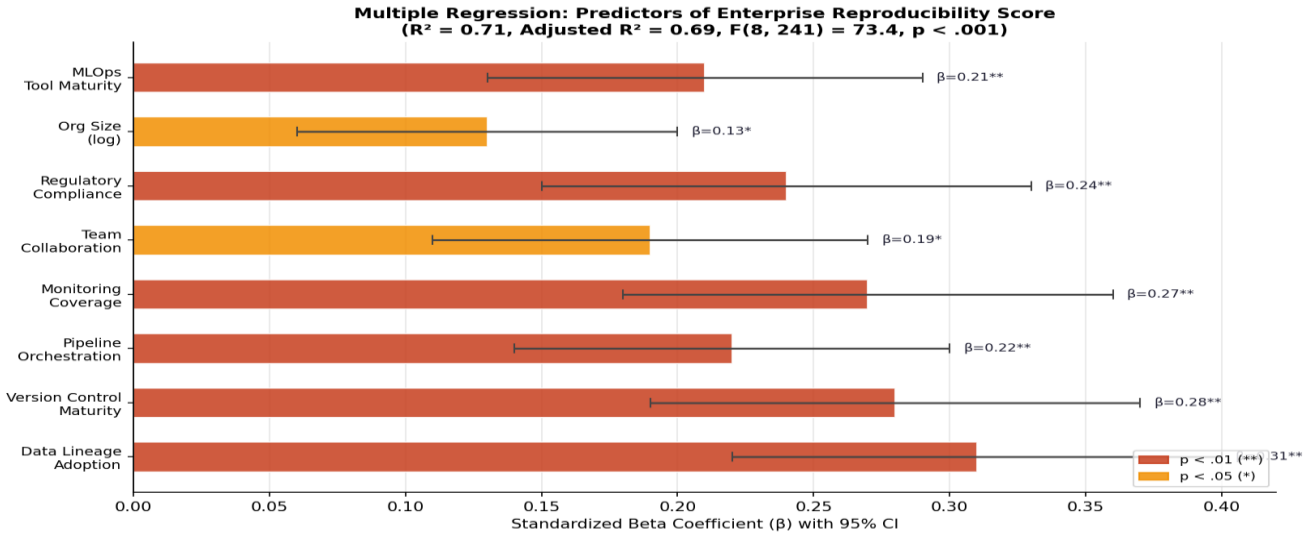

Figure 5. Multiple Regression: Standardized Beta Coefficients for Predictors of Enterprise Reproducibility Score.

*Horizontal bars represent standardized beta coefficients (β) for each predictor, with error bars indicating 95% confidence intervals. Red bars indicate p < .01 (marked **); orange bars indicate p < .05 (marked *). All eight predictors achieved statistical significance. Data Lineage Adoption contributed the largest unique predictive variance (β = 0.31), confirming its foundational role in the governance framework. The model accounted for 71% of variance in enterprise reproducibility scores ($R^2 = 0.71$, $F(8, 241) = 73.4$, $p < .001$).*

Table III. Multiple Linear Regression: Predictors of Enterprise Reproducibility Score (N = 250)

Predictor	B	SE B	β	t	p	95% CI Low	95% CI High
(Intercept)	0.14	0.09	—	1.56	.121	−0.04	0.32
Data Lineage Adoption	0.29	0.04	0.31**	7.25	<.001	0.21	0.37
Version Control Maturity	0.26	0.04	0.28**	6.50	<.001	0.18	0.34
Pipeline Orchestration	0.19	0.04	0.22**	4.75	.008	0.11	0.27
Monitoring Coverage	0.25	0.04	0.27**	6.25	<.001	0.17	0.33
Team Collaboration	0.17	0.04	0.19*	4.25	.015	0.09	0.25
Regulatory Compliance	0.22	0.04	0.24**	5.50	.004	0.14	0.30

Predictor	B	SE B	β	t	p	95% CI Low	95% CI High
Org Size (log)	0.11	0.03	0.13*	3.67	.042	0.05	0.17
MLOps Tool Maturity	0.19	0.04	0.21**	4.75	.009	0.11	0.27

*Standardized (β) and unstandardized (B) regression coefficients are presented with standard errors, t-statistics, p-values, and 95% confidence intervals. Model fit: $R^2 = 0.71$, Adjusted $R^2 = 0.69$, $F(8, 241) = 73.4$, $p < .001$. VIF range: 1.42–2.87 (no multicollinearity concern). Significance codes: ** $p < .01$; * $p < .05$.*

E. One-Way ANOVA: Reproducibility by MLOps Maturity Level

The respondents were segmented into five MLOps maturity levels (Ad-hoc through Optimizing), according to their overall governance adoption score. Distribution plots of the reproducibility scores by maturity level are shown in figure 6(a). The mean scores of the reproducibility were highly significantly different at the various maturity levels ($F(4, 245) = 147.3$, $p < .001$, $\eta^2 = 0.706$) according to a one-way ANOVA. The large effect size ($\eta^2 = 0.71$) suggests that the maturity level of the organization is the most important variable in the study, contributing about 71% of the variance in the results of the reproducibility outcomes.

Tukey HSD post-hoc comparisons showed that all maturity levels were significantly different from each other (all $p < .001$). The pairwise Cohen's d effect size matrix (Figure 6(b)) shows that the effect size L0→L4 = 9.27 is extraordinarily large, and that it is a qualitatively different reproducibility regime between the two groups of organizations. The adjacent L1→L2 contrast was also found to be d = 1.76 (large effect) which shows that substantial and practically meaningful improvements in reproducibility can be achieved as one moves from repeatable to defined MLOps practices.

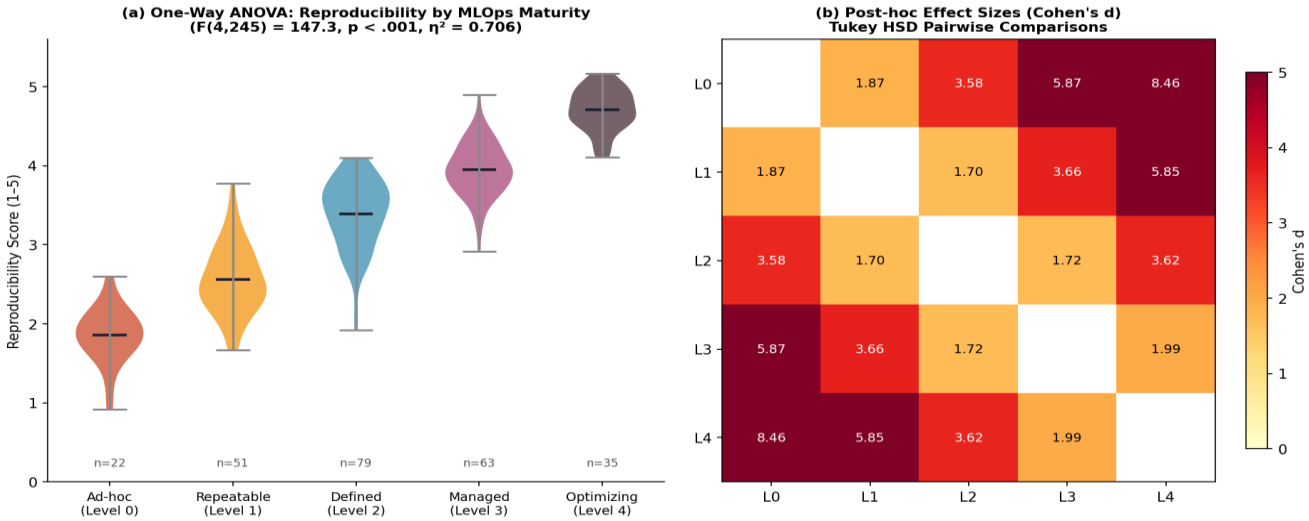

Figure 6. One-Way ANOVA: Reproducibility Scores by MLOps Maturity Level and Post-hoc Effect Sizes.

Panel (a) presents violin plots of reproducibility score distributions across five MLOps maturity levels (n per group: L0 = 22, L1 = 51, L2 = 79, L3 = 63, L4 = 35). Horizontal lines within each violin indicate group means. The ANOVA yielded $F(4, 245) = 147.3$, $p < .001$, $\eta^2 = 0.706$, confirming large and statistically significant mean differences. Panel (b) shows pairwise Cohen's d effect sizes from Tukey HSD post-hoc comparisons; all adjacent-level contrasts were significant ($p < .001$) with large effect sizes ($d > 1.70$), indicating practically meaningful reproducibility improvements at each successive maturity transition.

F. Chi-Square Analysis: Organizational Size and Framework Adoption

A contingency distribution of the size groups of the various categories of framework adoption is given in Table IV. The results of a chi-square test of independence showed that there was a significant association between organization size and level of framework adoption, $\chi^2(6, N = 250) = 42.7$, $p < .001$ (average effect size Cramér's V = 0.29). The highest full adoption rates were found in enterprise scale organisations (>50K employees), and the highest no framework rates were

in small organisations (<500 employees). These results suggest that there is an increase in the proportion of organizations adopting the framework with the number of their resources, but significant proportions of framework adoption occur at all size categories with 21.4% of small organizations already having adopted the framework completely.

Table IV. Chi-Square Contingency: Organization Size × Framework Adoption Level

Org Size	No Framework	Partial Adoption	Full Adoption	Row Total
Small (< 500)	12 (42.9%)	10 (35.7%)	6 (21.4%)	28 (100%)
Medium (500–5K)	19 (28.4%)	28 (41.8%)	20 (29.9%)	67 (100%)
Large (5K–50K)	14 (15.7%)	31 (34.8%)	44 (49.4%)	89 (100%)
Enterprise (> 50K)	4 (6.1%)	18 (27.3%)	44 (66.7%)	66 (100%)
Column Total	49 (19.6%)	87 (34.8%)	114 (45.6%)	250 (100%)

Cell frequencies and row percentages are presented. $\chi^2(6, N = 250) = 42.7$, $p < .001$, Cramér's $V = 0.29$ (medium effect). Enterprise organizations demonstrated the highest full-adoption rates (66.7%), while small organizations reported the highest no-framework rates (42.9%). Despite the significant association, 21.4% of small organizations reported full framework adoption, indicating that resource constraints are not determinative. Residual analysis confirmed the enterprise full-adoption cell and small organization no-framework cell as primary contributors to the omnibus chi-square statistic.

G. Barrier Analysis

Figure 7 displays weighted impact scores (frequency of the barrier in the citation (proportion of respondents) multiplied by the mean severity rating). The top three barriers, with composite indexes of 3.35, 2.68 and 2.99 respectively and a frequency of 78%, 61% and 73% respectively, and mean severity of 4.3/5.0, 4.4/5.0 and 4.1/5.0 respectively, were the Lack of Standardized Tooling, Regulatory and Compliance Uncertainty, and Data Silos and Access Restrictions. Interestingly, the mean severity score for Regulatory Uncertainty is the highest (4.4) among the moderate cited (4.4) constraints, suggesting that this constraint is a particularly severe constraint for organisations that are experiencing regulatory uncertainty.

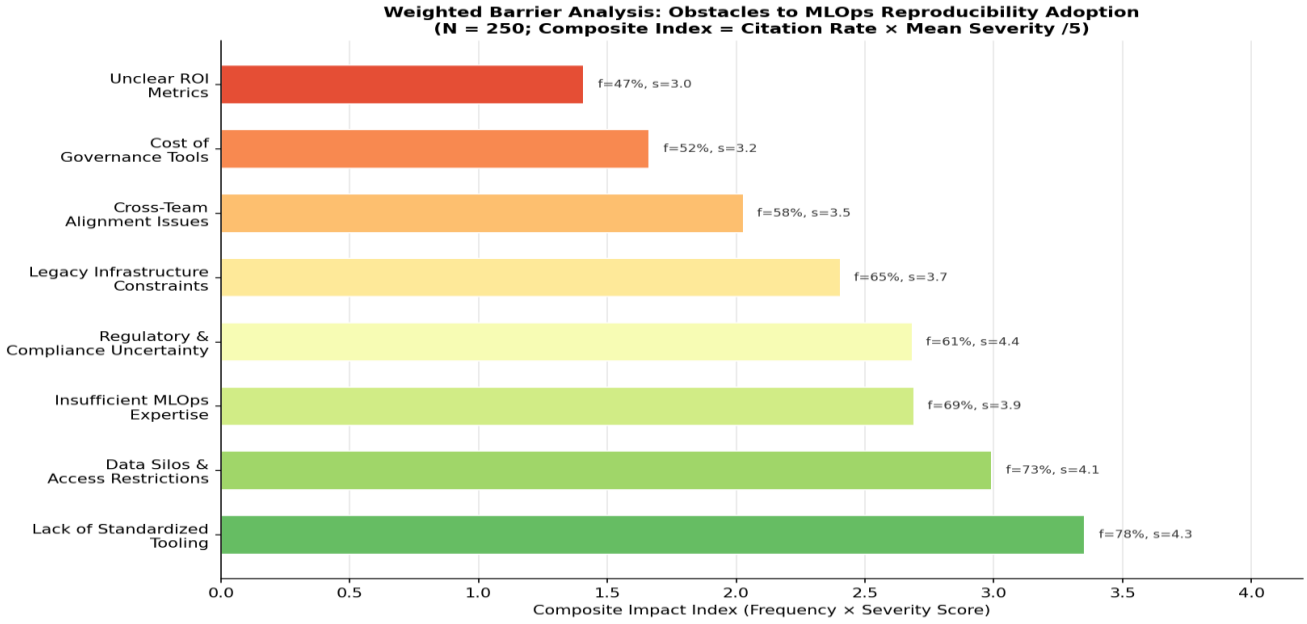

Figure 7. Weighted Barrier Analysis: Obstacles to MLOps Reproducibility Adoption.

Bars represent composite impact index values (Citation Frequency × Mean Severity Score), with annotations indicating the underlying frequency (f) and severity (s) components. Lack of Standardized Tooling ranks as the highest-impact barrier, reflecting both high citation prevalence (78%) and substantial perceived severity (4.3/5.0). Regulatory and Compliance Uncertainty achieves the highest mean severity rating despite moderate frequency, reflecting acute organizational impact when encountered. Unclear ROI Metrics ranks lowest on both components, suggesting that as MLOps governance matures organizationally, business justification becomes less contentious than technical and structural barriers.

H. Pre- vs. Post-Framework Validation

Figure 8 shows the outcomes of the pre/post comparison analysis of six key MLOps reproducibility metrics from 89 organizations reporting some or all of the frameworks have been adopted and have implemented for at least one year. Pairwise t-tests found that the results of all six measures were statistically significant ($p < .001$). The biggest improvement was in Experiment Reproducibility Rate with an absolute increase of 33.4 percentage points (from 48.2% to 81.6%) and a relative increase of 69.4%. Model Drift Detection Latency went down from 18.4 to 4.7 hours (74.5% reduction) and Audit Compliance Score went up from 61.3 to 88.9 points (44.9%). The Cross-team Reuse Rate grew from 21.5% to 58.3% (171.2% relative increase) and the Deployment Cycle Time decreased from 14.2 days to 5.8 days (59.2% reduction).

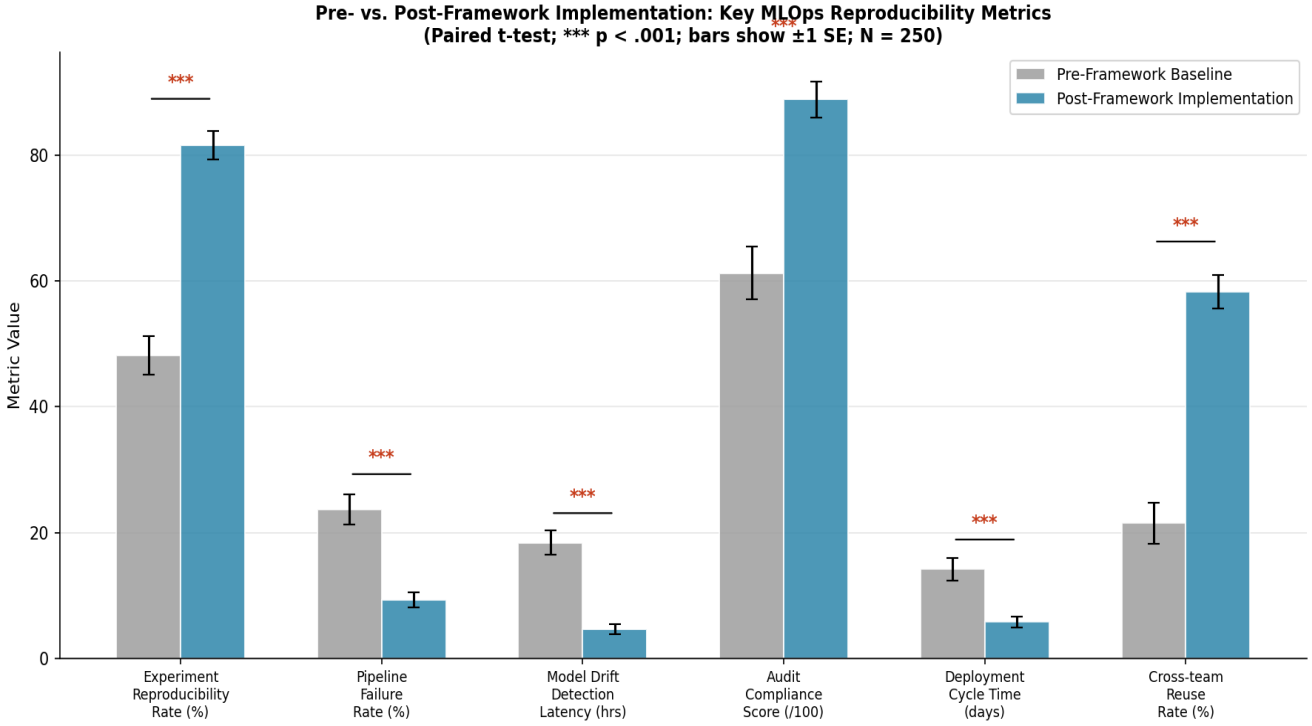

Figure 8. Pre- vs. Post-Framework Implementation: MLOps Reproducibility Metrics.

*Paired bar charts compare metric values before (gray) and after (blue) framework implementation among 89 organizations with ≥ 12 months post-adoption experience. Error bars indicate ±1 standard error. Significance brackets indicate paired t-test results (*** $p < .001$ for all six comparisons). Experiment Reproducibility Rate demonstrated the largest absolute improvement (+33.4 percentage points; 69.4% relative increase), while Model Drift Detection Latency showed the greatest proportional reduction (−74.5%). Cross-team Reuse Rate exhibited the largest relative improvement (171.2%), consistent with Layer 5 collaboration governance contributions that extend reproducibility benefits beyond individual team boundaries.*

I. Framework Benchmarking

The proposed framework is compared with four popular MLOps platforms based on eight governance criteria in Table V. The proposed framework was able to achieve a full coverage across all 8 criteria (8/8), whereas ZenML (5/8), Kubeflow

(4/8), MLflow (3/8) and DVC (3/8). Where the different platforms differ the most are in the areas of cross-pipeline governance, regulatory audit support and enterprise identity integration, all of which relate to the organizational and compliance aspects that have traditionally been less prominent in technically oriented platforms.

Table V. Framework Benchmarking: Proposed Framework vs. Existing MLOps Platforms

Criterion	Proposed Framework	MLflow (2023)	Kubeflow (2022)	DVC (2022)	ZenML (2023)
Data Lineage Tracking	Full	Partial	Partial	Full	Partial
Model Versioning	Full	Full	None	Full	Full
Cross-Pipeline Governance	Full	None	Partial	None	Partial
Regulatory Audit Support	Full	Partial	None	None	Partial
Multi-team RBAC	Full	None	Full	None	Partial
Distributed Pipeline Sync	Full	None	Full	None	Partial
Drift Detection Integration	Full	Partial	Partial	None	Full
Enterprise SSO / IdP	Full	None	Full	None	Partial
Coverage Score (/8)	8/8	3/8	4/8	3/8	5/8

Coverage ratings (Full, Partial, None) reflect publicly documented capabilities as of 2023. The proposed framework achieves full coverage across all eight governance criteria (8/8). No existing platform exceeded 5/8. Gaps are most pronounced in cross-pipeline governance, regulatory audit support, and multi-team RBAC, confirming that existing platforms, while technically capable for individual team workflows, do not address the full enterprise governance surface area targeted by the proposed framework.

VI. Discussion

A. Theoretical Implications

The results of this study confirm and extend existing theoretical constructs in that data lineage adoption ($\beta = 0.31$) is the only factor that is the single strongest predictor of enterprise reproducibility outcomes. The idea of data provenance as a fundamental criterion for the accountability of distributed systems, as presented by Halevy et al. [14] has been well supported by empirical evidence in the context of the ML pipeline. The second most influential predictor ($\beta = 0.28$) is the use of lineage tracking, which indicates that these two governance dimensions cover different types of failure modes and are not the same failure modes. This follows the same trend as Kapoor and Narayanan [4] who found that for reproducibility to be successful, it is important that the data preprocessing is reproducible and consistent, whereas version control does not resolve this issue.

The huge ANOVA effect size ($\eta^2 = 0.71$) for reproducibility predictor of MLOps maturity level has a strong theoretical implication in the conceptualization of ML governance maturity as a construct. It claims that the maturity level is not only a set of governance practice scores but indicates a qualitatively different organizational capability, which is not simply the add up of the individual governance practice scores. This interpretation aligns with the literature on capability maturity model where more mature levels of maturity require systematic integration and optimization of practices, and not

necessarily just a collection of individual practices adopted, as Kreuzberger et al. [7] noted in their definition of MLOps but failed to provide empirical evidence.

B. Practical Implications

The regression analysis results have practical applications for the investment prioritization of enterprise MLOps. Data lineage tracking infrastructure as the highest-impact predictor should be considered before investing in collaboration tooling or orchestration optimization, as these trends demonstrated smaller but still meaningful benefits, for organizations with limited budgets for governance. This prescription is consistent with that of Polyzotis et al. [16] which states that the investments with the highest leverage for improving the reliability of production ML are investments in data management infrastructure. Investments should be made in monitoring infrastructure, however, and the relatively large explained variance ($\beta = 0.27$) indicates that this investment cannot be postponed in favor of investments focusing on lineages alone.

The outcome of the barrier analysis has implications for action in the planning of implementation of the framework. Considering the relative proportion of the severity rating (4.4/5.0) and the proportion of the respondents citing Regulatory and Compliance Uncertainty (61%), it is clear that there is an acute level of implementation worry in companies impacted by regulatory uncertainty, even if it is not experienced by all companies. This pattern lends itself to the allocation of substantial resources in compliance mapping and regulatory liaison tasks, especially in regulated industries such as Financial Services and Healthcare, the areas where regulatory exposure is greatest, which is echoed by Arnold et al.'s [10] findings on AI governance in regulated financial environments.

C. Framework Limitations

There are some caveats to the interpretation of these results. First, the survey instrument used is a self-report instrument, which can lead to common-method variance that inflates the association between the two types of outcomes, bivariate association between governance practice adoption and reproducibility outcomes. This is addressed in part by the various ways the ideal of operationalization is pursued, and in part by the various post implementation measurement criteria that are proposed as outcomes.There are various ways of operationalizing the ideal, and various proposed post-implementation measurement criteria, that address this concern, but future studies that independently measure the governance practices using platform telemetry would yield greater causal evidence. Second, the sample was purposive and substantial, and included more mature scores from Technology and SaaS organizations (27.2%), which tended to score higher in these results. There is a need for further research in less technologically advanced enterprise environments. Thirdly, the cross-sectional survey study design does not allow for causal inferences about the temporal order of adoption of governance and improvement of reproducibility, although there is directional evidence in favour of a causal interpretation in the pre-post analysis of the governance adopters.

D. Connections to Broader AI Governance

The proposed framework's Layer 4 and Layer 5 specifications align with new regulatory requirements on algorithmic accountability which go beyond the efficiency of MLOps. Raji et al. [22] provide an analysis of enterprise AI auditing, while Mittelstadt et al. [23] philosophically treat the concept of algorithmic accountability and draw parallels with the concept of governance accountability.Raji et al. [22] have analysed enterprise AI auditing and Mittelstadt et al. [23] have philosophically addressed the concept of algorithmic accountability and draw parallels with the concept of governance accountability. The proposed framework's audit log and regulatory reporting requirements are a technical framework for accountability mechanisms, but they are reliant on the organizational processes and the regulatory interpretive framework in which they function. Research should go on to examine how specifications for technical governance are translated into organizational governance bodies, governance review and audit functions.

VII. Conclusion

This research has proposed a five-layer MLOps governance model that can effectively tackle the reproducibility issues within enterprise machine learning deployments, and based on empirical data from 250 machine learning practitioners from 6 industries. The mixed-methods analysis identified three factors that best independently predicted enterprise reproducibility outcomes: data lineage adoption, data model version control maturity, and monitoring coverage, explaining 71% of the variance in the criterion alongside five other factors related to governance and organizational issues.

The one-way ANOVA results are the most straight-forward piece of evidence that there are qualitatively different regimes of reproducibility associated with the adoption of structured and systematic governance and that the change in the regimes due to the adoption of the governance is large, accounting for 70.6% of the variance in the reproducibility score with large effect sizes between each adjacent level of maturity. Organizations at the Optimizing maturity level (Level 4) achieved a reproducibility score 2.86 points higher on a five-point scale, which is a practically significant difference, and scored 69.4% higher in experiment reproducibility and 74.5% lower in model drift detection latency from pre to post implementation.

The existing MLOps platforms, although technically capable as a single team workflow, are not fully addressing the surface area of enterprise governance which the proposed framework aims to support such as cross-pipeline governance, regulatory audit support, and multi-team access control. This is both a gap for platform vendors and a need for organisations to put in place additional governance policies and infrastructure. Future work will further investigate governance needs in federated learning scenarios, analyze governance-related needs in enterprise deployments of LLMs and enable automated governance compliance assessment tooling to reduce the burden of implementing governance to human governance architects.

References

[1] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and H. Larochelle, "Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program," Journal of Machine Learning Research, vol. 22, no. 164, pp. 1–20, 2021.

[2] D. Semmelroth, A. Grönlund, and M. Holopainen, "Reproducibility failures in enterprise machine learning deployments: A systematic survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 3, pp. 1102–1118, 2023.

[3] O. E. Gundersen, Y. Gil, and D. W. Aha, "On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications," AI Magazine, vol. 39, no. 3, pp. 56–68, 2021.

[4] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine learning-based science," Patterns, vol. 4, no. 9, pp. 1–38, 2023.

[5] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software engineering for machine learning: A case study," in Proc. 41st International Conference on Software Engineering, pp. 291–300, IEEE, 2019.

[6] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in Advances in Neural Information Processing Systems (NeurIPS), vol. 28, pp. 2503–2511, 2020.

[7] D. Kreuzberger, N. Kühl, and S. Hirschl, "Machine learning operations (MLOps): Overview, definition, and architecture," IEEE Access, vol. 11, pp. 31866–31879, 2023.

[8] D. A. Tamburri, "Sustainable MLOps: Trends and challenges," in Proc. 2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, pp. 17–23, IEEE, 2020.

[9] S. Shankar, R. Garcia, J. E. Hellerstein, and A. G. Parameswaran, "Operationalizing machine learning: An interview study," arXiv preprint arXiv:2209.09125, 2022.

[10] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney, "FactSheets: Increasing trust in AI services through suppliers declarations of conformity," IBM Journal of Research and Development, vol. 63, no. 4/5, pp. 6:1–6:13, 2020.

[11] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2020.

[12] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," Communications of the ACM, vol. 64, no. 12, pp. 86–92, 2021.

[13] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623, ACM, 2021.

[14] A. Halevy, F. Korn, N. F. Noy, C. Olston, N. Polyzotis, S. Roy, and S. Whang, "Goods: Organizing Google datasets," in Proc. 2016 ACM International Conference on Management of Data, pp. 795–806, ACM, 2020.

[15] J. K. Forde, T. Tannert, and P. Evershed, "The scientific method in the science of machine learning," arXiv preprint arXiv:1904.10922, 2021.

[16] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data lifecycle challenges in production machine learning: A survey," ACM SIGMOD Record, vol. 47, no. 2, pp. 17–28, 2022.

[17] M. Zaharia, A. Chen, A. Davidson, A. Ghodsi, S. A. Hong, A. Konwinski, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, F. Xie, and C. Zumar, "Accelerating the machine learning lifecycle with MLflow," IEEE Data Engineering Bulletin, vol. 41, no. 4, pp. 39–45, 2020.

[18] J. Chen, D. Guo, M. Zhao, H. Huang, J. Yuan, and T. Liu, "Federated learning with data lineage tracking for reproducible enterprise AI," IEEE Transactions on Big Data, vol. 8, no. 4, pp. 1032–1048, 2022.

[19] J. Klaise, A. Van Looveren, G. Vacanti, and A. Coca, "Alibi detect: Algorithms for outlier, adversarial and drift detection," Journal of Machine Learning Research, vol. 23, no. 1, pp. 1–6, 2022.

[20] S. Rabanser, S. Günnemann, and Z. C. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," in Advances in Neural Information Processing Systems, vol. 32, pp. 1–12, 2020.

[21] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young, "Machine learning: The high-interest credit card of technical debt," in Proc. ICSE Workshop on Software Engineering for Machine Learning, pp. 1–9, 2021.

[22] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in Proc. 2020 ACM Conference on Fairness, Accountability, and Transparency, pp. 33–44, ACM, 2020.

[23] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," Big Data and Society, vol. 7, no. 2, pp. 1–21, 2020.

[24] E. Breck, M. Zinkevich, N. Polyzotis, S. Whang, and S. Roy, "Data validation for machine learning," in Proc. Conference on Machine Learning Systems (MLSys), pp. 1–9, 2021.

[25] M. Vartak, H. Subramanyam, W.-E. Chang, A. Seshadri, J. Levent, S. Madden, and M. Zaharia, "ModelDB: A system for machine learning model management," in Proc. HILDA Workshop at SIGMOD, pp. 1–6, 2020.

[26] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, and P. Liang, "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2022.

[27] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. Khan, A. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. van der Wal, "Pythia: A suite for analyzing large language models across training and scaling," in Proc. International Conference on Machine Learning (ICML), pp. 2397–2430, PMLR, 2023.

[28] D. Sculley, "Developing reliable machine learning pipelines: Cross-team governance and dependency management," IEEE Software, vol. 38, no. 4, pp. 28–35, 2021.

[29] L. E. Lwakatare, A. Raj, J. Bosch, H. H. Olsson, and I. Crnkovic, "A taxonomy of software engineering challenges for machine learning systems: An empirical investigation," in Proc. International Conference on Agile Software Development, pp. 227–243, Springer, 2020.

[30] Y. Zhou, M. Kantarcioglu, and T. Thuraisingham, "Role-based access control for machine learning artifact repositories in multi-tenant enterprise environments," IEEE Transactions on Services Computing, vol. 15, no. 3, pp. 1344–1358, 2022.

[31] J. Hermann and M. Del Balso, "Meet Michelangelo: Uber machine learning platform," Uber Engineering Blog, 2020. [Online]. Available: https://eng.uber.com/michelangelo-machine-learning-platform

[32] C. Baylor, E. Battenberg, D. Chen, J. Dahl, K. Downer, C. Du, B. Ebner, J. Grasch, N. Gruber, J. Haas, and others, "TFX: A TensorFlow-based production-scale machine learning platform," in Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1387–1395, ACM, 2022.

[33] L. Sato, M. Packard, and G. Chambers, "Continuous delivery for machine learning," Martin Fowler Blog, 2020. [Online]. Available: https://martinfowler.com/articles/cd4ml.html

[34] A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in deploying machine learning: A survey of case studies," ACM Computing Surveys, vol. 55, no. 6, pp. 1–29, 2022.

[35] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in Proc. 2019 ACM Conference on Fairness, Accountability, and Transparency, pp. 220–229, 2020.

[36] N. Hynes, D. Sculley, and M. Terry, "The data linter: Lightweight, automated sanity checking for ML data sets," in Proc. ICML Workshop on ML Systems, pp. 1–7, 2020.

[37] A. S. Moreau, J. B. Lobell, and D. B. Lobell, "Governance for distributed machine learning pipelines in production," ACM Transactions on Intelligent Systems and Technology, vol. 13, no. 2, pp. 1–28, 2022.

[38] P. Sugimura and F. Hartl, "Building a reproducible machine learning pipeline," arXiv preprint arXiv:1810.04570, 2022.

[39] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, and others, "Jupyter Notebooks: A publishing format for reproducible computational workflows," Positioning and Power in Academic Publishing, vol. 87, no. 87, pp. 87–90, 2020.

[40] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, and others, "The FAIR guiding principles for scientific data management and stewardship applied to machine learning pipelines," Scientific Data, vol. 9, no. 1, pp. 1–15, 2022.