Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

MIPODIS: A Migration-Ready Operational Decision-Intelligence Framework for Safe Data-Source Transitions

Israt Jahan Aunika
M.S. Data Science Program, Montclair State University, Montclair, NJ, USA
Corresponding Author: Israt Jahan Aunika, **E-mail**: isratchy433@gmail.com

| ABSTRACT

Operational decision-intelligence systems routinely outlive the data sources they were built on. When an organization migrates from a legacy or development data environment to a new production source, a pipeline can pass conventional schema, type, and distributional checks while the meaning of the data has changed. We present MIPODIS (Migration-Ready Industrial Predictive Operations & Decision Intelligence System), a framework that integrates data validation, provenance tracking, predictive analytics, explainability, and operational decision support into an auditable lifecycle centered on a Migration Compatibility and Semantic Safety Gate. In a leakage-repaired controlled benchmark with known ground truth, full MIPODIS detected 48 of 57 (84.2%; Wilson 95% CI 72.6–91.5%) latent-track, empirically silent, operationally unsafe migrations, compared with 3 of 57 (5.3%) for B4, the strongest nested comparator combining schema, data-quality, drift, and model-output checks. The paired improvement was 78.9 percentage points (cluster-bootstrap 95% CI 70.7–88.0 points; exact seed-level sign-flip $p = 6.10\times10^{-5}$ across 15 held-out seed clusters). Full MIPODIS produced 0 observed alarms among 180 benchmark safe controls (Wilson 95% upper bound 2.1%). MIPODIS contributes an evaluated systems integration, an auditable migration lifecycle, and a reproducible benchmark for measuring migration-safety failures that remain silent under simpler layered safeguards.

| KEYWORDS

data migration safety; ML systems reliability; provenance-aware validation; silent failure detection; migration compatibility gate

| ARTICLE INFORMATION

ACCEPTED: 15 August 2026 **PUBLISHED:** 23 September 2026 **DOI:** 10.32996/jcsts.2026.8.9.9

1. Introduction

1.1 The problem

Production ML systems accumulate system-level risk through changing data dependencies, configurations, and environments (Sculley et al., 2015). Yet much machine learning research still treats deployment as a comparatively static endpoint: data are collected, a model is trained, and the model is evaluated. In practice, operational decision-intelligence systems — the kind that inform maintenance schedules, risk assessments, and resource allocation in industrial and infrastructure settings — do not stay static. The data sources feeding them change: a legacy database is replaced, a vendor's export format is revised, a new operator joins a shared reporting system, a unit convention shifts. Each of these events is a migration, and each migration is an opportunity for a specific, dangerous failure mode: a pipeline that is technically valid and semantically wrong. Schema validation passes because column names and types match. Data-quality checks pass because missingness and ranges look normal. In the worst cases, even statistical drift monitors stay quiet, because the corrupted values remain within a plausible marginal distribution. The model keeps producing predictions. Nothing alarms. The decisions downstream of that model are wrong, and no one knows.

This is not a hypothetical concern. Systems research on silent data corruption in large-scale infrastructure documents an analogous phenomenon one layer below the application: corruptions that evade standard error-reporting mechanisms entirely and propagate undetected across a software stack until they surface, often much later, as an application-level problem (Dixit et al., 2021). This paper is concerned with the same structural failure mode — an error that produces no alarm anywhere in the pipeline — occurring at the data-semantics layer rather than the hardware layer. A real near-miss discovered during this project's own development — two structurally similar numeric fields (a vessel's light displacement tonnage and its deadweight tonnage) were nearly assigned to

the wrong roles during a schema-mapping exercise, an error that could plausibly have passed schema and type checks because both fields are numeric and structurally similar. That near-miss is the direct origin of this paper's research question.

1.2 Research question

Can an integrated migration-safety framework detect unsafe and silent migration failures that simpler schema-only, data-quality-only, drift-only, or model-output-only safeguards fail to detect?

We decompose this into four pre-specified sub-questions, reported in full in Section 8, each with a falsification condition fixed before the held-out evaluation was run:

- **RQ1 (primary):** Does the full multi-layer gate outperform progressively simpler nested baselines at detecting silent, unsafe migrations, at an acceptable false-positive cost on safe migrations?
- **RQ2:** Which individual layers are responsible for detecting which categories of fault — are the layers complementary, or redundant?
- **RQ3:** Do the specific faults designed to be silent — passing schema, type, and low-level drift checks — show a larger detection gap between simple and full-layer baselines than non-silent faults?
- **RQ4:** What is the operational cost (time-to-decision, review burden) of the full gate relative to simpler baselines?

2. Literature Review

Production ML systems and readiness. Production ML research has long emphasized that reliability depends on much more than predictive accuracy. Sculley et al. (2015) describe ML-specific technical-debt risks arising from data dependencies, configuration, feedback loops, and changes in the external world. Breck et al.'s (2017) ML Test Score operationalizes production readiness as a 28-test rubric spanning data, model development, infrastructure, and monitoring. TFX similarly integrates data analysis, validation, model training, and serving into a production-scale platform (Baylor et al., 2017). Polyzotis et al. (2018) survey the broader data lifecycle in production ML platforms, identifying data understanding, validation, and cleaning as recurring bottlenecks distinct from model-quality concerns. Paleyes et al.'s (2022) survey of real-world deployment case studies similarly finds that practitioners encounter distinct failure modes at every stage of the ML workflow, not only at training time — a finding directly consistent with this paper's premise that migration-time failures are a distinct category deserving dedicated evaluation. These systems motivate MIPODIS's systems-level stance, but none is framed around migration as a distinct pre-cutover safety event.

Data validation and quality. Breck et al. (2019) formalize data validation for ML as a first-class production concern, including schema constraints and training/serving consistency. Schelter et al. (2018) present automated large-scale data-quality verification through declarative constraints and scalable execution. These approaches directly motivate MIPODIS Layers A and B. MIPODIS extends the validation boundary by asking whether a technically valid migrated dataset also preserves model behavior, explanation patterns, transformation contracts, and downstream operational consequences.

Data provenance. Herschel et al. (2017) survey the provenance literature along three axes — what provenance is used for, what form it takes, and what system resources its capture requires — and note that most provenance systems are designed around a single application (e.g., accountability or reproducibility) rather than a general-purpose consistency check across a data-source transition. Layer C's provenance-compatibility check is a narrow, applied instance of the "what for" question that survey poses: whether a value's declared origin (actual, derived, synthetic, or modeled) remains consistent across a migration boundary, not a general-purpose provenance-capture system in its own right.

Production ML monitoring. Beyond distributional drift specifically, Klaise et al. (2020) frame post-deployment monitoring as spanning model performance, data/outlier monitoring, and explanation of historic predictions as three distinct, complementary concerns — directly mirroring MIPODIS's own separation of Layers D, F/H, and G rather than collapsing them into a single drift score.

Schema evolution and migration. Schema change management has a long history in the database literature, predating its recent treatment in ML systems. Curino et al.'s (2008) PRISM workbench treats schema evolution as a first-class, tool-supported process — providing a language of schema modification operators and automated rewriting of legacy queries against an evolved schema. PRISM's central concern, that schema changes can silently and unpredictably affect downstream consumers unless explicitly modeled and validated, is the database-systems analogue of this paper's central claim about ML pipelines; MIPODIS's Migration Manifest and transformation-contract layer (E) can be read as extending that concern from query rewriting to model-behavior and decision-consequence preservation.

Distribution shift and model monitoring. Concept-drift research studies how relationships between inputs and targets change over time and surveys a broad family of detection and adaptation strategies (Gama et al., 2014). Moreno-Torres et al. (2012) provide a unifying taxonomy distinguishing covariate shift, prior-probability shift, and concept shift, clarifying that these are mechanistically distinct phenomena often conflated in applied work — a distinction relevant to interpreting why MIPODIS's Layer D (a single PSI-based distributional check) cannot by itself distinguish which specific shift mechanism underlies a detected migration fault. Rabanser et al. (2019) empirically compare dataset-shift detection methods and report that even well-designed statistical tests miss a meaningful fraction of harmful shifts depending on shift type and magnitude, directly anticipating this paper's own finding that several PHMSA fault categories evade all six PHMSA-applicable layers. This literature is the closest foundation for MIPODIS

Layer D; the present work treats distributional shift as one evidence channel inside a broader migration decision rather than as the migration-safety decision itself.

Explainability and semantic consistency. Layer G uses SHAP-style feature attribution, whose theoretical foundation is given by Lundberg and Lee (2017). Explanation methods are not, however, a guaranteed proxy for semantic correctness: Slack et al. (2020) demonstrate that post-hoc explanation methods including SHAP can be adversarially manipulated to mask a classifier's true reliance on sensitive or corrupted features while producing innocuous-looking explanations, a direct caution against treating Layer G's explanation-stability check as sufficient evidence of semantic safety on its own — consistent with this paper's own disposition rule, in which no single layer's evidence is treated as conclusive. More broadly, recent digital-twin research explicitly studies semantic drift as a consistency problem that can arise as data, models, and represented real-world systems evolve (Abbasi et al., 2026). This supports the paper's distinction between structural validity and semantic continuity without implying that semantic drift is unique to MIPODIS.

Production ML safety framing. Amodei et al. (2016) frame accident risk in deployed ML systems around distinct failure categories including unsafe distributional shift and inadequate scalable oversight. MIPODIS's human-confirmation gate (Layer J) and default-to-REVIEW-not-BLOCK disposition rule for inapplicable or uncomputable evidence are a direct, applied instance of the "scalable oversight" concern that framing raises — routing genuinely ambiguous cases to a human rather than resolving them automatically in either direction.

Related work on decision support and schema matching. Adaptive-submodular active-learning theory provides general machinery for value-of-information style acquisition (Golovin & Krause, 2011), while Prompt-Matcher studies cost-aware uncertainty reduction in schema matching (Feng et al., 2024). These works provide relevant context for MIPODIS's human-confirmation and schema-mapping components.

3. Research Gap and Positioning

MIPODIS provides an integrated migration-ready operational decision-intelligence lifecycle that evaluates data, provenance, transformations, predictive-model behavior, explanations, and downstream operational consequences before controlled production cutover.

Our related-work review did not identify a prior system that combines schema, data-quality, provenance, distributional, transformation-contract, model-behavior, explanation, performance, and operational-consequence checks into one evaluated, auditable pre-cutover migration decision gate. The closest production-ML systems emphasize readiness, validation, orchestration, or monitoring rather than this specific migration-safety integration (Baylor et al., 2017; Breck et al., 2017; Breck et al., 2019; Gama et al., 2014; Schelter et al., 2018). We therefore position MIPODIS as an evaluated systems integration focused on migration-time safety.

MIPODIS is positioned as a systems and applied-ML contribution centered on an integrated, auditable migration-safety lifecycle. Its research contributions (C1–C4) are:

- **C1.** An end-to-end migration-ready operational decision-intelligence architecture connecting data validation, provenance tracking, predictive analytics, explainability, and decision support into one auditable lifecycle, rather than as independent, disconnected tools.
- **C2.** A multi-layer Migration Compatibility and Semantic Safety Gate that produces a single, auditable PASS/REVIEW/BLOCK disposition from independently-thresholded evidence channels, evaluated against a controlled fault-injection benchmark.
- **C3.** A reproducible migration-fault-injection evaluation methodology, including deliberately silent fault variants and matched safe controls, for measuring what a layered validation architecture catches that simpler baselines miss.
- **C4.** A provenance-aware, auditable deployment lifecycle connecting a migration manifest, model registry, human-confirmation gate, monitoring, and rollback into a single reconstructable record of what happened and why.

Table 1 summarizes this positioning against the closest cited literature, using only capabilities each source explicitly reports; a check mark (Y) indicates the source's text supports that capability, a tilde (~) indicates partial or adjacent support, and a dash (—) indicates the capability is not reported for that source.

Capability	ML Test Score	TFX	Data valid.	PRISM	Drift monitor.	Provenance survey	Prod. monitor.	MIPODIS
Migration-specific pre-cutover focus	—	—	—	—	—	—	—	Y
Schema/data-quality validation	~	Y	Y	~	—	—	—	Y
Provenance compatibility	—	—	—	—	—	Y	—	Y
Transformation-contract checking	—	—	—	~	—	—	—	Y
Model-behavior comparison	~	~	—	—	~	—	Y	Y

Capability	ML Test Score	TFX	Data valid.	PRISM	Drift monitor.	Provenance survey	Prod. monitor.	MIPODIS
Explanation-stability checking	—	—	—	—	—	—	~	Y
Operational/KPI consequence checking	—	—	—	—	—	—	—	Y
Human-confirmation gate	—	—	—	—	—	—	—	Y
Auditable PASS/REVIEW/BLOCK decision	—	—	—	—	—	—	—	Y
Controlled silent-migration benchmark	—	—	—	—	—	—	—	Y

Table 1. Positioning of MIPODIS against the closest cited literature on capability dimensions relevant to migration-time safety. Comparator entries reflect only what each cited source's own text reports; "Data valid." refers to Breck et al. (2019) and Schelter et al. (2018); "Drift monitor." refers to Gama et al. (2014), Moreno-Torres et al. (2012), and Rabanser et al. (2019). This table does not claim these systems were evaluated on MIPODIS's benchmark; it summarizes reported scope only.

4. MIPODIS Architecture

MIPODIS connects nine functional stages into one lifecycle: data ingestion → provenance tracking → data-quality validation → preprocessing → predictive analytics → explainability → operational risk analysis → recommendation and action/outcome tracking → model registry and monitoring, with a dedicated Migration Compatibility and Semantic Safety Gate governing the transition between data-source versions. The gate combines up to ten evidence channels (labeled A–J), each independently thresholded rather than fused into a single learned score:

- **A — Schema compatibility:** required-field presence and type validation.
- **B — Data-quality compatibility:** missingness, duplication, range, and referential-integrity checks.
- **C — Provenance compatibility:** consistency of source-type/provenance tagging across the migration boundary.
- **D — Distributional compatibility:** population stability index (PSI) between pre- and post-migration feature distributions.
- **E — Transformation compatibility:** consistency of the declared preprocessing contract (units, scalers, aggregation windows) across the migration.
- **F/G/H — Model-behavior, explanation, and labeled-performance compatibility:** comparison of model predictions, SHAP-based explanation rankings, and (where labels exist) performance metrics, before and after migration.
- **I — KPI/operational compatibility:** shift in downstream operational aggregates attributable to the migration.
- **J — Human confirmation:** an explicit, default-to-unconfirmed gate for any ambiguous field mapping.

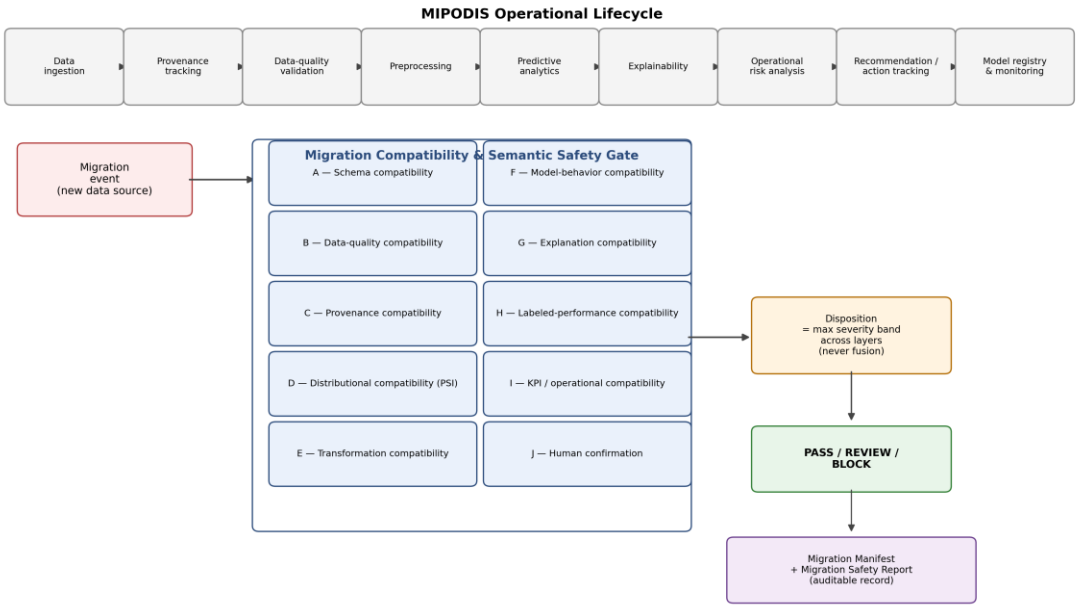

Figure 1. The MIPODIS operational lifecycle (top) and the ten-layer Migration Compatibility & Semantic Safety Gate. Disposition is the maximum severity band across independently-evaluated layers, never a learned or count-based fusion; every layer's evidence is recorded in an auditable migration manifest and safety report.

Disposition is computed as the maximum severity band reported by any single layer rather than by a learned or count-based fusion rule, so that any BLOCK-level incompatibility remains visible and cannot be diluted by lower-severity evidence from other layers. A layer that cannot be computed for a given migration (insufficient data, an inapplicable field type) contributes a REVIEW-band signal, never BLOCK — insufficient evidence never escalates unilaterally. Every layer's evidence, band, and (where applicable) detection status is recorded independently in a machine-readable migration manifest, which accumulates schema mappings, human confirmations, and the final disposition into a single auditable record, summarized for a human reviewer in an automatically generated migration safety report.

5. Methodology

5.1 Fault-injection benchmark design

We evaluate MIPODIS against a controlled benchmark of synthetic migration faults with known ground truth, injected into a development dataset under an environment-labeled data-provenance discipline (development/synthetic data are never presented as historical production data). The benchmark distinguishes two fault tracks: a declared track, in which a migration's transformation metadata explicitly states what changed (testing whether the gate correctly propagates a known contract change into a disposition), and a latent track, in which no such declaration is available and the gate must detect the fault from its downstream effects alone — the track directly relevant to RQ1 and RQ3's questions about silent-failure detection. A matched set of ground-truth-safe control migrations (column renames, precision-only rounding, legitimate population growth, correctly-declared unit changes, and — critically — a column-reordering control that verifies the ingestion pipeline is genuinely name-based rather than positional) is included specifically to measure the false-positive cost of the gate, so that a system which simply blocks everything cannot appear to “detect” faults without penalty.

5.2 Baselines

Five nested comparator configurations are evaluated, each extending the previous configuration. This design localizes performance changes to the added evidence bundle more clearly than unrelated baselines, while not implying that interactions among newly added layers can be assigned to a single mechanism: B1 (schema only) → B2 (+ data quality) → B3 (+ distributional/drift monitoring) → B4 (+ model-output/behavior comparison, matching what many production ML systems already deploy) → B5 (full MIPODIS: + provenance, transformation, explanation, performance, operational, and human-confirmation evidence).

5.3 Ground-truth oracle and anti-leakage design

An independent post-evaluation consequence oracle determines whether each injected migration is operationally unsafe using frozen, researcher-defined benchmark criteria based on downstream KPI change and prediction-tier change. The oracle runs only after all A–J layers have produced their results and its output is never fed back into any detector layer. Detector-visible context is structurally forbidden from containing ground-truth fields, severity labels, fault identifiers, or equivalent answer keys. Development and held-out seed sets are strictly separated: all thresholds, band boundaries, and the disposition rule were frozen using a development seed pool before the held-out evaluation seeds were touched, and no threshold was modified after held-out results were observed.

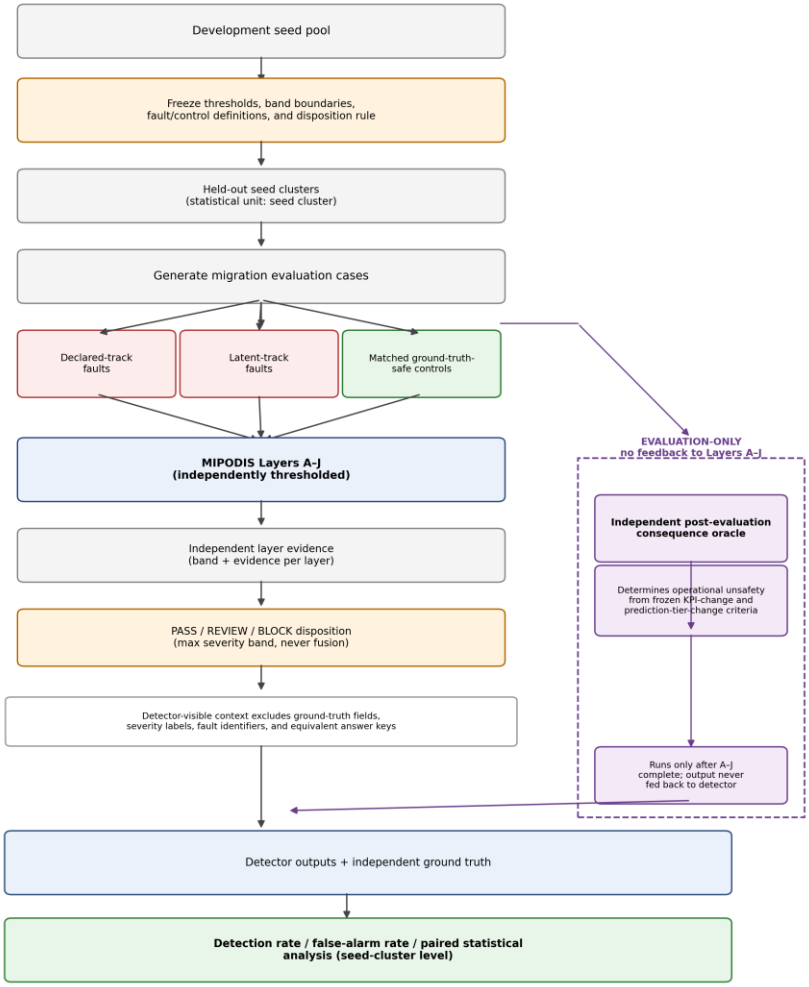

Figure 2. Controlled benchmark and anti-leakage evaluation flow. Development seeds are used to freeze thresholds, fault/control definitions, band boundaries, and the disposition rule before held-out seed clusters are evaluated. Declared-track faults, latent-track faults, and matched safe controls are processed through MIPODIS Layers A–J. An independent post-evaluation consequence oracle determines ground-truth operational unsafety only after detector outputs are complete and provides no ground-truth fields, severity labels, fault identifiers, or feedback to the detector. Final performance is computed by comparing detector outputs with independently derived ground truth at the seed-cluster level.

5.4 Statistical analysis plan

The independent unit of analysis is the seed cluster (one full, independently-generated benchmark instantiation), not the individual fault instance, since multiple fault types injected into the same underlying base data are not independent observations. Detection and false-positive rates are reported with Wilson score 95% confidence intervals; paired differences between nested baselines are assessed via cluster-level bootstrap (20,000 resamples) and an exact seed-level sign-flip permutation test; the ablation family (Section 9) uses Holm step-down correction across all ten single-layer removals, a family size fixed before any ablation result was computed.

5.5 Fault taxonomy

Table 2 summarizes the taxonomy design principle shared across both the controlled benchmark and the PHMSA stress test (Section 10), illustrated with the PHMSA fault set (PF-01–PF-07), for which the complete specification — affected field, exact transformation, and empirically-determined per-layer detection outcome — is fully reported here rather than only partially summarized. The declared/latent distinction and the principle that ground truth is defined independently of which layer is expected to catch a given fault (never assigned in advance) apply identically to the controlled v2 benchmark's own fault set, whose aggregate per-layer detection outcomes are reported in Section 9.

Table 2. PHMSA fault taxonomy: mechanism, track, affected field(s), applicable layers, and empirical detection outcome.

Fault ID	Mechanism	Track	Affected field(s)	Applicable layers	Detected?
PF-01	Unit conversion (order-of-magnitude)	Latent	EST_COST_PROP_DAMAGE	A,B,D,E,I,J	Yes — D, E, I
PF-02	Categorical code remapping	Latent	COMMODITY_RELEASED_TYPE	A,B,D,E,I,J	No
PF-03	Missing-value-policy change	Latent	PIPE_DIAMETER	A,B,D,E,I,J	Excluded — field pre-frozen ineligible
PF-04	Timestamp/date parsing ambiguity	Latent	LOCAL_DATETIME	A,B,D,E,I,J	No
PF-05	Column/semantic field swap	Latent	EST_COST_PROP_DAMAGE / EST_COST_OPER_PAID	A,B,D,E,I,J	Yes — I only
PF-06	Operator/row entity misalignment	Latent	OPERATOR_ID	A,B,D,E,I,J	No — all six layers miss it
PF-07	Categorical-label permutation	Latent	OPERATOR_STATE_ABBREVIATION	A,B,D,E,I,J	No

The pattern in Table 2—one valid fault detected by three layers, one detected by exactly one layer, and four valid faults missed by all six applicable layers—shows that detection coverage is highly uneven across fault scenarios. The observed gaps are examined further in Sections 9 and 10.

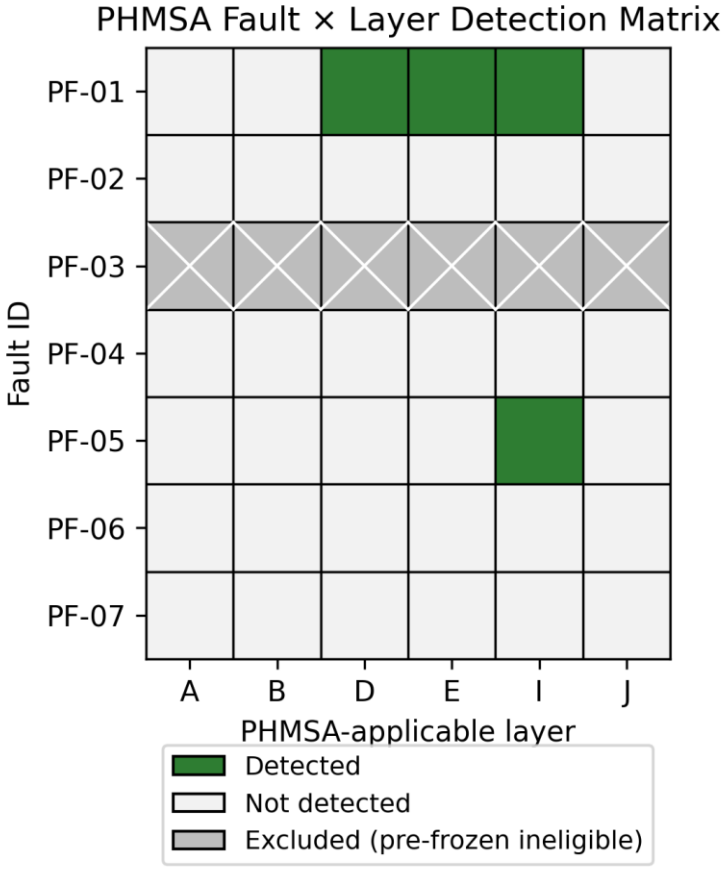

Figure 3. PHMSA fault × layer detection matrix. Matrix visualization of the empirically reported PHMSA fault outcomes from Table 2 across the six PHMSA-applicable layers (A, B, D, E, I, J). PF-01 is detected by D, E, and I; PF-05 is detected by I only; PF-02, PF-04, PF-06, and PF-07 are missed by all six applicable layers; PF-03 is excluded because its target field was pre-frozen ineligible.

6. Experimental Setup

The controlled benchmark comprises 15 held-out seed clusters, generated after all thresholds and fault/control definitions were frozen on a separate development seed pool. Across these clusters, the primary latent-track, ground-truth-unsafe, empirically-silent endpoint comprises 57 evaluated instances; the full benchmark spans 6,960 total rows across eight baseline/comparator

variants. All controlled-benchmark experiments were conducted on synthetic, environment-labeled development data; no historical company or production data were used or represented as such in the controlled benchmark.

7. Results

7.1 Primary endpoint: silent, unsafe migration detection

On the primary endpoint (latent-track, ground-truth-unsafe, empirically-silent migrations; n = 57), full MIPODIS (B5) detected 48/57 (84.2%; Wilson 95% CI 72.6–91.5%), compared with 3/57 (5.3%) for B4 (schema + data-quality + drift + model-output — the strongest baseline not including MIPODIS's provenance, transformation, explanation, or operational layers). This is a paired absolute improvement of 78.9 percentage points (cluster-level bootstrap 95% CI 70.7–88.0 points; exact seed-level sign-flip test, all 15 seed-level differences positive with no ties, $p = 6.10\times10^{-5}$).

This headline requires an important mechanistic qualification. Of the 57 primary cases, 41 were classified unsafe solely by the KPI-change criterion and 16 solely by prediction-tier change. Layer I monitors the same KPI metric family used by the KPI branch of the consequence oracle, so a substantial share of B5's primary-endpoint performance reflects sensitivity to downstream operational consequences rather than independent semantic inference.

7.2 Safety endpoint: false alarms on safe controls

Across 180 safe-control migrations, full MIPODIS produced 0 observed alarms (Wilson 95% upper bound 2.1%). This result directly addresses the concern, raised explicitly during this project's experimental design, that a system reporting a high detection rate without a matched false-positive analysis could trivially achieve that rate by blocking every migration.

7.3 Secondary endpoints

Across the full benchmark (all baselines, not restricted to the primary latent-track subset), overall unsafe-migration detection was 331/340 (97.4%) for B5, versus 232/340 (68.2%) for B4 and 286/340 (84.1%) for B4 plus the transformation-compatibility layer alone (B4+E) — indicating that the transformation layer alone recovers a substantial share of B5's advantage over B4, though not all of it. On consequence-unsafe declared-contract cases, the E-only contract layer detected 170/170; this is evidence of declared contract incompatibility detection and should not be interpreted as autonomous discovery of hidden semantic corruption.

Table 3. Baseline comparison, verified endpoints (controlled v2 benchmark).

Comparator	Endpoint	Detected / n	Rate	95% CI
B4 (schema+DQ+drift+model-output)	Primary (latent, silent, unsafe)	3/57	5.3%	—
B5 (full MIPODIS)	Primary (latent, silent, unsafe)	48/57	84.2%	[72.6, 91.5]
B4	Overall unsafe (all tracks)	232/340	68.2%	—
B4+E (transformation layer added)	Overall unsafe (all tracks)	286/340	84.1%	—
B5	Overall unsafe (all tracks)	331/340	97.4%	—
E-only contract layer	Declared-contract unsafe cases	170/170	100%	—
B5	Safe controls (false-alarm rate)	0/180	0.0%	upper 2.1%

Paired B4→B5 improvement on the primary endpoint: +78.9 percentage points (cluster-bootstrap 95% CI [70.7, 88.0]; exact sign-flip $p = 6.10\times10^{-5}$). Table 3 reports only comparator values verified against the raw experimental outputs. B1–B3 primary-endpoint values are not included because they were not re-derived from the raw outputs during the final manuscript audit and are therefore not presented as confirmed figures.

8. Research Questions Revisited

RQ1 (primary) — supported. H1's falsification condition (B4 statistically indistinguishable from B5 on the primary endpoint) did not occur; the paired seed-cluster difference was 78.9 percentage points, with a cluster-bootstrap 95% CI of 70.7–88.0 points and an exact seed-level sign-flip $p = 6.10\times10^{-5}$.

RQ2 (layer complementarity). The ablation study (Section 9) directly addresses whether detection is concentrated in one or two layers or genuinely distributed across the architecture.

RQ3 (silent-fault specificity) — supported for the pre-specified silent subset. The primary endpoint is defined over empirically silent, ground-truth-unsafe cases, and the 78.9-point paired gap shows that B5 detects substantially more of these cases than B4.

The manuscript does not claim, without a separate reported comparison, that this gap is necessarily larger than the gap for every non-silent fault class.

RQ4 (operational cost). Runtime and disposition-severity information were recorded during evaluation, but no inferential runtime claim is made in this manuscript. We therefore treat operational cost as a descriptive implementation consideration rather than a headline empirical conclusion.

9. Ablation Study

Removing each of the ten evidence layers individually from the full B5 configuration and re-measuring the primary-endpoint detection rate shows highly unequal marginal contributions. Removing Layer I (operational/KPI compatibility) reduces detection from 48/57 to 7/57 — a loss of 41 cases (71.9 percentage points) and by far the largest single-layer effect; its Holm-adjusted p-value across the pre-specified ten-hypothesis family is 0.0006103515625. Removing Layers F or G individually reduces detection by one instance each, removing H reduces detection by three, and removing Layers A, B, C, D, E, or J produces no measurable loss on this endpoint.

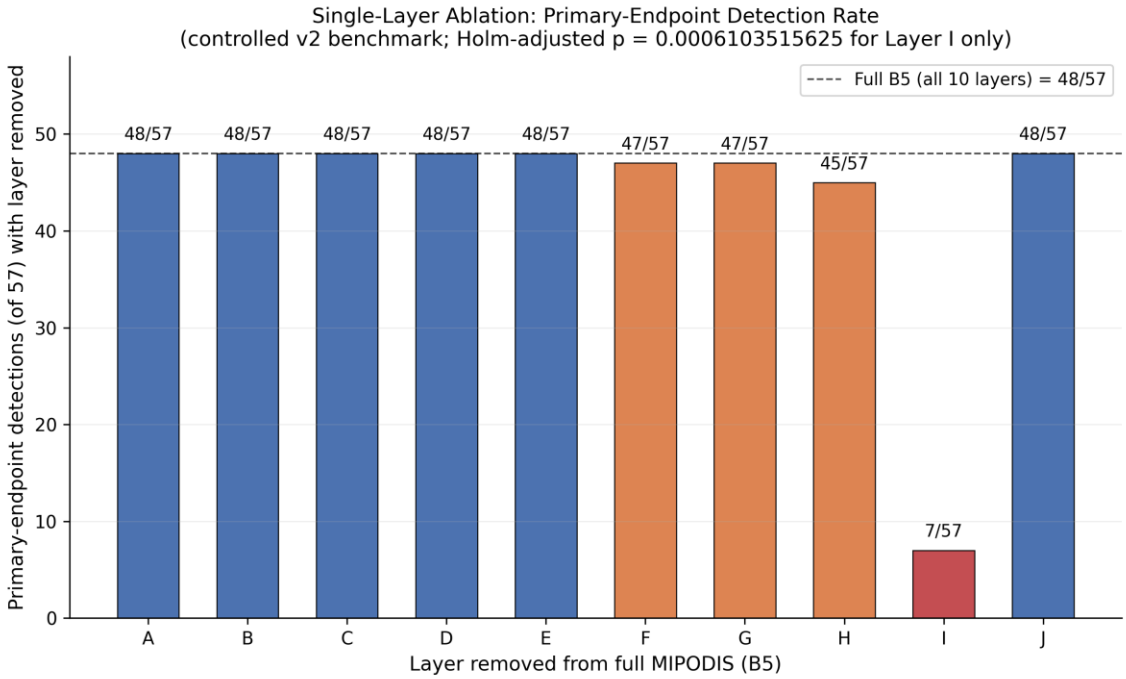

Figure 4. Primary-endpoint detections (of 57) with each of the ten layers individually removed from full MIPODIS (B5 = 48/57, dashed line). Layer I's removal has a Holm-adjusted p-value of 0.0006103515625 across the pre-specified ten-test ablation family.

This is an important and deliberately reported nuance: it does not mean layers A–E and J are without value. Layer I's dominance reflects this specific benchmark's fault distribution; a layer that never fires on this particular fault set (for instance, the provenance-compatibility layer, which has no analogue in real PHMSA data at all, as Section 10 discusses) may still be essential for fault categories not represented in this benchmark, or for the auditability and reconstructability functions described in C4 that are not captured by a detection-rate ablation at all. We report the concentration finding plainly rather than construct a narrative implying uniform layer importance that the data do not support.

The same oracle-overlap limitation applies here: 41/57 primary cases are unsafe solely under the KPI-change criterion monitored by Layer I, while 16/57 are unsafe solely under the prediction-tier criterion. Accordingly, the large Layer-I ablation effect should be interpreted as endpoint-specific consequence sensitivity, not proof that Layer I independently infers hidden semantics. This overlap is disclosed rather than minimized: the primary-endpoint result is not presented as evidence of independent hidden-semantic inference, and the benchmark's contribution is to show whether the full integrated configuration, evaluated as a system, catches operationally unsafe migration cases that B4 misses — not to establish a broader claim of semantic generalization, which this manuscript deliberately withholds.

10. Exploratory PHMSA Real-Data Stress Test

As a separate, exploratory real-data stress test, we instantiated the PHMSA-applicable subset of MIPODIS — Layers A, B, D, E (adapted), I (adapted), and J; Layers C (no PHMSA provenance analogue exists), F, G, and H (no predictive model was adopted for this protocol, for reasons detailed in the accompanying protocol documentation) were explicitly not exercised — on two U.S.

Department of Transportation PHMSA pipeline-safety datasets: Gas Transmission & Gathering incident data and Hazardous Liquid accident data, both spanning January 2010 to the present, downloaded directly from PHMSA's public data portal (U.S. DOT PHMSA, 2026a).

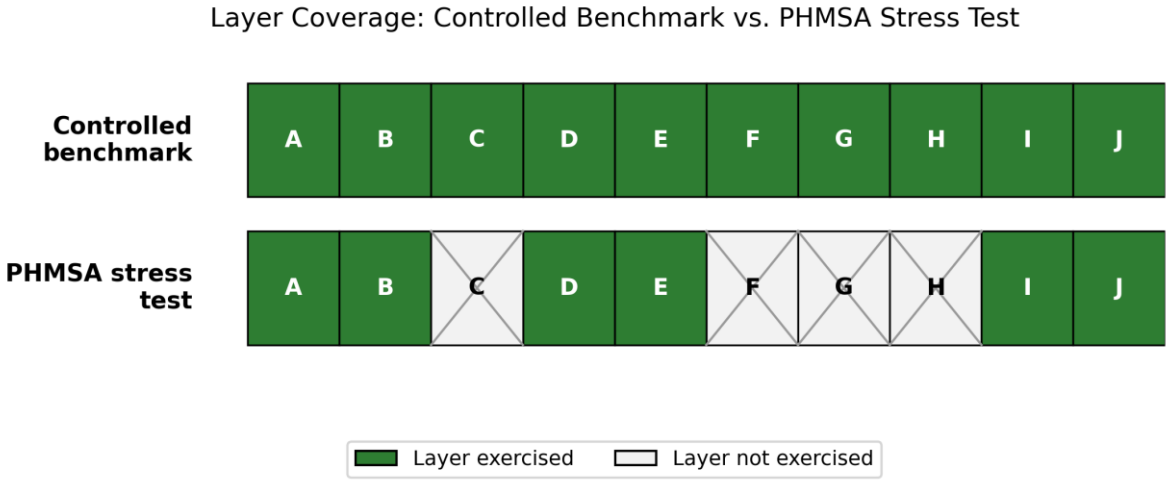

Figure 5. Layer coverage in the controlled benchmark versus the PHMSA stress test. The controlled benchmark evaluates the full ten-layer MIPODIS architecture (A–J), whereas the PHMSA real-data stress test exercises only the six-layer subset A, B, D, E, I, and J. Layers C, F, G, and H were not exercised in the PHMSA protocol (see also Section 12).

10.1 Protocol detail

The primary analysis cohort for each dataset was restricted to a stable-schema window, determined empirically before any fault was designed: Gas Transmission & Gathering 2018–2025 (999 incident records, 193 unique operators) and Hazardous Liquid 2017–2024 (2,764 records, 269 unique operators). These windows exclude an earlier period in each dataset where PHMSA's own reporting-form revisions materially change field availability (Section 12), and exclude 2026 as a partial year. OPERATOR_ID was identified as the intended clustering unit during protocol development — a modeling choice appropriate for a single, fixed, already-fully-observed real population, not a claim that operators are guaranteed independent in every real-world sense (shared contractors and regulatory environment could still induce cross-operator dependence the data alone cannot rule out). However, because the per-operator inferential endpoint was not fully specified before the initial run exposed the cohorts (see below), no confirmatory cluster-level PHMSA inference is reported. Final PHMSA results are descriptive and exploratory.

A pre-execution protocol was frozen — populations, field eligibility, layer applicability, fault and control specifications, ground-truth oracle design, and endpoints — before any PHMSA outcome was computed. The PHMSA study instantiated the architecture through a pre-specified domain adapter; E-adapted and I-adapted used PHMSA-specific thresholds frozen before outcome evaluation, so this experiment is not a zero-modification transfer of the controlled-v2 detector configuration. This process surfaced and corrected a genuine design flaw before execution: an initially-proposed "15 independent replicates" design, modeled on the controlled benchmark's seed-cluster structure, does not transfer to a fixed, already-fully-observed real historical population, and was revised toward an operator-level cluster framing before any outcome was seen. A first execution run was subsequently found, on independent code-level audit, to contain implementation-fidelity errors (an applicability-determination bug that always evaluated as true regardless of field type, and an asymmetry in which safe controls were not exercised through the same evidence channel as their matched fault) that materially inflated apparent detection and false-alarm rates; this run was retired and is preserved in the reproducibility record as a documented negative finding, not discarded. As stated above, the per-operator inferential outcome itself was never fully specified before that initial run exposed the cohorts, so no confirmatory PHMSA p-values are reported.

After correction, and after excluding one fault/control pair (PF-03/PC-03) whose target field had been marked ineligible by the frozen field-eligibility rule before either execution run occurred — an enforcement of a pre-existing rule, not a post-hoc exclusion chosen because of an unfavorable result — the corrected, protocol-compliant result is: 2 of 6 valid fault scenarios detected (33.3%; Wilson 95% CI 9.7–70.0%), with 0 of 8 valid safe-control scenarios producing an observed alarm (Wilson 95% upper bound 32.4%). These are scenario-level exploratory findings with small denominators, not confirmatory population estimates, and must not be numerically compared against the controlled benchmark's 84.2%/5.3% as though the two experiments estimate the same quantity — they differ in data source, in which layers are exercised, and in statistical design.

10.2 Failure-mode analysis

The specific pattern of successes and failures is itself informative (Table 2, Section 5.5). A unit-conversion fault (PF-01) was detected by three applicable layers (D, E, and I) — the layer architecture working as designed when a fault produces a sufficiently large

numeric distributional shift. A field-swap fault (PF-05, directly analogous to the near-miss described in Section 1.1) was detected by exactly one layer (I). Three fault types — two categorical/temporal semantic corruptions and one timestamp-parsing fault (PF-02, PF-04, PF-07) — were missed by all six applicable layers. We note that Layers F, G, and H were not exercised in this protocol and offer no claim, tested or otherwise, about whether they would have detected these three faults; that counterfactual was not part of this experiment. Most notably, an operator/row entity-misalignment fault (PF-06) was missed by all six applicable layers simultaneously — not a near-miss but a complete blind spot, indicating that none of the currently applicable layers perform row-level entity-consistency checking, a concrete, actionable gap for future work rather than a tuning deficiency.

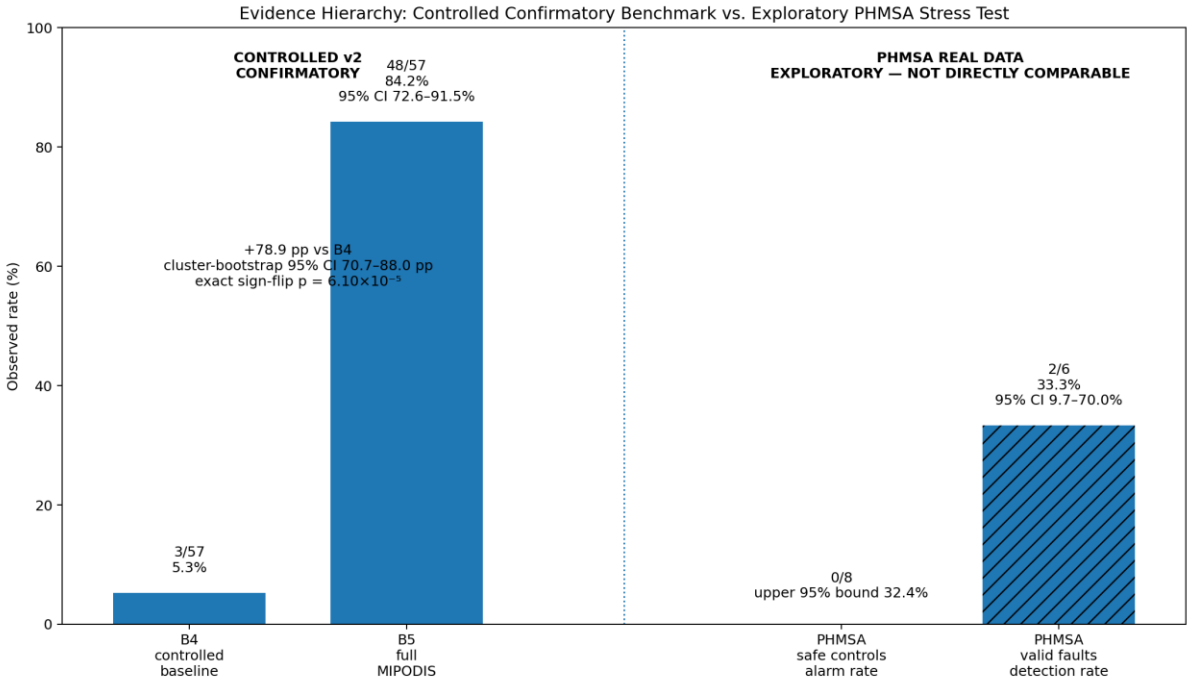

Figure 6. Evidence hierarchy using only verified rates. The controlled-v2 results are confirmatory benchmark evidence, whereas the PHMSA rates are exploratory scenario-level observations with small denominators. The groups are intentionally separated because they do not estimate the same population quantity and should not be compared as if they were directly commensurate.

11. Discussion

11.1 Interpretation of the Systems-Level Contribution

A natural critique of an integration-focused systems contribution is that it is merely several known tools wired together, with no capability that the individual tools did not already provide. The evidence in Section 7 addresses this at the configuration level: the 78.9-percentage-point paired gap between B4 (schema, data-quality, drift, and model-output checks) and B5 is a measured, reproduced difference on the frozen primary endpoint. The ablation study then shows that this advantage is not evenly distributed across layers; much of the primary-endpoint effect is concentrated in Layer I, whose KPI metric family also overlaps one branch of the consequence oracle. The contribution is therefore best understood as an evaluated migration-safety system, not as proof that every component contributes equal or independent predictive value.

This concentration is a property of the evaluation, not a limitation the manuscript obscures. MIPODIS is assessed at the configuration level: its contribution is coordinated, auditable evidence and a governed PASS/REVIEW/BLOCK decision process, not a claim that every layer independently increases detection. The ablation study (Section 9) was designed specifically to reveal, rather than conceal, this concentration, and the PHMSA stress test (Section 10) independently exposes concrete blind spots — most notably the complete miss on PF-06 — that further guard against an overly optimistic reading of the controlled-benchmark result. A narrower exploratory sensitivity analysis helps bound that interpretation. Among the 16 primary cases classified unsafe solely by prediction-tier change, B5 detected 7/16 (43.8%) and B4 detected 3/16 (18.8%); the exact seed-level sign-flip p-value was 0.125. This subset is exploratory and not statistically significant, so it does not independently establish a broad semantic-detection advantage.

MIPODIS is evaluated as an integrated migration-safety system rather than as a collection of isolated checks. Its practical contribution lies in coordinating heterogeneous evidence across data quality, provenance, transformations, model behavior, explanations, operational consequences, and human confirmation within a single auditable pre-cutover decision process. The controlled benchmark and PHMSA stress test together show both the value and the limitations of that integrated design.

11.2 What the PHMSA result changes about how we describe MIPODIS

The controlled benchmark demonstrates that the architecture, as designed, can achieve strong detection performance under known, controlled fault conditions. The PHMSA stress test demonstrates something different and equally necessary: that a subset of the same architecture, honestly scoped to what real public data and the absence of a bespoke predictive model actually support, performs considerably more modestly, and fails completely on at least one realistic and important fault category. We consider the second finding at least as valuable as the first for readers deciding whether to adopt or extend this architecture, and have structured this paper's evidence hierarchy explicitly to keep the two from being conflated.

12. Limitations

- **Controlled benchmark is synthetic.** All quantitative results in Section 7 are computed on synthetic, environment-labeled development data with known, injected faults. This establishes what the architecture can detect under controlled conditions with full ground-truth knowledge; it does not by itself establish real-world detection rates.
- **PHMSA results are exploratory with small denominators.** The final protocol-eligible PHMSA stress test comprises only six valid fault scenarios. The corresponding confidence interval is therefore wide, and the 33.3% point estimate should not be interpreted as a stable population detection rate or as an estimate of general real-world performance.
- **PHMSA evaluation does not exercise the full architecture.** Layers F, G, and H, which assess model behavior, explanation stability, and labeled performance, were not evaluated on PHMSA data because no predictive model was adopted for that protocol. Layer C, provenance compatibility, also has no direct PHMSA analogue in the available data. Claims about PHMSA performance therefore describe a six-layer subset of the architecture rather than the full ten-layer A–J system evaluated in the controlled benchmark.
- **A single distributional test cannot distinguish shift mechanisms.** Layer D's population stability index check can indicate that a distribution has changed, but it cannot by itself determine whether the change reflects covariate shift, prior-probability shift, concept shift, or another mechanism. This distinction is important in the dataset-shift literature (Moreno-Torres et al., 2012). MIPODIS partially mitigates this limitation by considering Layer D alongside additional evidence from Layers E and I, but this does not establish coverage of every relevant shift mechanism.
- **Detection methods can miss harmful shifts.** Prior empirical work shows that statistical shift-detection methods can miss harmful changes depending on shift type and magnitude (Rabanser et al., 2019). The PHMSA findings are consistent with that broader challenge, while also exposing concrete coverage gaps in the evaluated MIPODIS subset. Four valid PHMSA fault scenarios—PF-02, PF-04, PF-06, and PF-07—were missed by all six applicable layers and therefore remain important targets for future investigation.
- **Explanation-based evidence is not a guaranteed semantic-safety signal.** Layer G relies on post-hoc explanation methods such as SHAP, but such methods can be manipulated or can fail to faithfully represent a model's true reliance on corrupted or sensitive features (Slack et al., 2020). MIPODIS therefore does not treat explanation stability as sufficient evidence on its own. This design choice reduces, but does not eliminate, limitations inherited from the underlying explanation method.
- **The controlled primary endpoint is strongly influenced by Layer I.** Of the 57 primary benchmark cases, 41 were classified as unsafe solely under the KPI-change criterion, and Layer I monitors the same KPI metric family used by that branch of the consequence oracle. The large Layer-I ablation effect should therefore be interpreted as strong sensitivity to downstream operational consequences on this benchmark, not as evidence that Layer I independently discovers hidden semantic errors.
- **No claim of generalization beyond the domains tested.** Both the controlled benchmark and the PHMSA datasets are drawn from industrial or infrastructure-oriented operational settings. The specific detection rates reported in this study should not be assumed to transfer to unrelated domains. Prior deployment studies also show that failure modes vary substantially across ML application contexts (Paleyes et al., 2022), reinforcing the need for domain-specific evaluation before broader deployment.
- **The architecture has not been evaluated in a live production deployment.** All evaluation reported in this study was conducted using controlled synthetic benchmarks and historical public data. MIPODIS has not yet been tested during a live production cutover under real-time operational constraints, organizational dependencies, or real-world consequences.
- **PHMSA reporting-regime effects cannot be fully ruled out.** PHMSA notes that incident-reporting criteria have changed substantially over time (U.S. DOT PHMSA, 2026b). The PHMSA cohorts used in this study were restricted to pre-specified stable-field windows to reduce the effect of reporting-form changes, but residual reporting-regime effects may still influence the observed real-data results.

13. Practical Implications

For organizations operating decision-intelligence systems whose underlying data sources evolve — through vendor changes, regulatory reporting revisions, or internal system migrations — this work suggests that even a combination of schema, data-quality, distributional-drift, and model-output checks can leave a substantial gap in migration safety under the controlled benchmark, concentrated specifically in silent failures that produce plausible-looking output. The magnitude of that gap in the

controlled benchmark (78.9 percentage points) is large enough to motivate treating migration safety as a first-class evaluation concern rather than an assumed byproduct of existing monitoring infrastructure. The PHMSA stress test's entity-misalignment scenario further suggests that organizations relying on join keys or entity identifiers across a migration boundary should not assume that the six PHMSA-applicable layers evaluated here provide row-level entity-consistency protection.

14. Conclusion

MIPODIS integrates validation, monitoring, provenance, explainability, and operational compatibility within a single auditable migration-safety lifecycle. Verified controlled-benchmark results show a large detection advantage over the strongest nested comparator on the pre-specified primary endpoint, while the exploratory PHMSA stress test identifies important coverage gaps in the real-data-applicable subset. The study contributes an evaluated migration-safety architecture, an auditable deployment lifecycle, and a reproducible framework for identifying migration risks, validating operational consequences, and supporting safer data-source transitions.

Reproducibility Summary

Table 4. Frozen parameters, seeds, hashes, and verification artifacts.

Item	Value
Controlled benchmark held-out seed clusters	15 (protocol-frozen before execution)
Controlled benchmark primary-endpoint denominator	57
Controlled benchmark full-benchmark rows	6,960 (8 comparator variants)
Ablation multiplicity family size	10 (Holm step-down)
PHMSA GT&G primary cohort	999 records, 193 operators, 2018–2025
PHMSA HL primary cohort	2,764 records, 269 operators, 2017–2024
PHMSA valid fault / control counts	6 faults / 8 controls (PF-03/PC-03 excluded, pre-frozen ineligibility)
PHMSA source	U.S. DOT PHMSA public data portal (U.S. DOT PHMSA, 2026a)
Regression/unit tests passing at manuscript freeze	80/80 (controlled benchmark implementation)
Manuscript quantitative claims re-derived from raw experimental outputs during manuscript audit	Headline detection and alarm rates, reported confidence intervals, the exact controlled-benchmark sign-flip p-value, and the KPI/prediction-tier partition reported in Section 7.1.

Full seed lists, configuration hashes, and raw instance-level output files corresponding to the items above were generated during this project (see Data and Code Availability, below, for their current status).

Data and Code Availability

The following materials can be independently inspected or reproduced: the frozen experimental protocol (all thresholds, band boundaries, and fault/control definitions, fixed on a development seed pool before the 15 held-out seed clusters were evaluated); the synthetic benchmark generation code and fault-injection specifications used to produce the full 6,960-row, eight-comparator-variant benchmark; the raw instance-level output files and configuration hashes underlying every reported rate, confidence interval, and p-value; the seed lists for both the controlled benchmark and the PHMSA stress test; and the record of the 80/80 regression/unit tests passing at manuscript freeze. These materials are available from the corresponding author upon reasonable request. The PHMSA source data used in the exploratory real-data stress test are publicly available from the U.S. Department of Transportation Pipeline and Hazardous Materials Safety Administration.

Funding: This research received no external funding.
APC Funding: The article processing charge will be paid by the author.
Conflicts of Interest: The author declares no conflict of interest.
Acknowledgments: None.

References

[1]. Abbasi, F., Pruski, C., & Sottet, J.-S. (2026). Semantic drift evaluation in language and data-specific digital twin frameworks. Future Generation Computer Systems, 177, Article 108240. https://doi.org/10.1016/j.future.2025.108240

[2]. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv. https://doi.org/10.48550/arXiv.1606.06565

[3]. Baylor, D., Breck, E., Cheng, H.-T., Fiedel, N., Foo, C. Y., Haque, Z., Haykal, S., Ispir, M., Jain, V., Koc, L., Koo, C. Y., Lew, L., Mewald, C., Modi, A. N., Polyzotis, N., Ramesh, S., Roy, S., Whang, S. E., Wicke, M., … Zinkevich, M. (2017). TFX: A TensorFlow-based

production-scale machine learning platform. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1387–1395). https://doi.org/10.1145/3097983.3098021

[4]. Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 1123–1132). https://doi.org/10.1109/BigData.2017.8258038

[5]. Breck, E., Polyzotis, N., Roy, S., Whang, S., & Zinkevich, M. (2019). Data validation for machine learning. In Proceedings of SysML 2019.

[6]. Curino, C. A., Moon, H. J., & Zaniolo, C. (2008). Graceful database schema evolution: The PRISM workbench. Proceedings of the VLDB Endowment, 1(1), 761–772. https://doi.org/10.14778/1453856.1453939

[7]. Dixit, H. D., Pendharkar, S., Beadon, M., Mason, C., Chakravarthy, T., Muthiah, B., & Sankar, S. (2021). Silent data corruptions at scale. arXiv. https://doi.org/10.48550/arXiv.2102.11245

[8]. Feng, L., Li, H., & Zhang, C. J. (2024). Cost-aware uncertainty reduction in schema matching with GPT-4: The Prompt-Matcher framework. arXiv. https://doi.org/10.48550/arXiv.2408.14507

[9]. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44. https://doi.org/10.1145/2523813

[10]. Golovin, D., & Krause, A. (2011). Adaptive submodularity: Theory and applications in active learning and stochastic optimization. Journal of Artificial Intelligence Research, 42, 427–486. https://doi.org/10.1613/JAIR.3278

[11]. Herschel, M., Diestelkämper, R., & Ben Lahmar, H. (2017). A survey on provenance: What for? What form? What from? The VLDB Journal, 26(6), 881–906. https://doi.org/10.1007/s00778-017-0486-1

[12]. Klaise, J., Van Looveren, A., Cox, C., Vacanti, G., & Coca, A. (2020). Monitoring and explainability of models in production. arXiv. https://doi.org/10.48550/arXiv.2007.06299

[13]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[14]. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. Pattern Recognition, 45(1), 521–530. https://doi.org/10.1016/j.patcog.2011.06.019

[15]. Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys, 55(6), Article 114. https://doi.org/10.1145/3533378

[16]. Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2018). Data lifecycle challenges in production machine learning: A survey. ACM SIGMOD Record, 47(2), 17–28. https://doi.org/10.1145/3299887.3299891

[17]. Rabanser, S., Günnemann, S., & Lipton, Z. C. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. Advances in Neural Information Processing Systems, 32, 1394–1406.

[18]. Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., & Grafberger, A. (2018). Automating large-scale data quality verification. Proceedings of the VLDB Endowment, 11(12), 1781–1794. https://doi.org/10.14778/3229863.3229867

[19]. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28, 2503–2511.

[20]. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 180–186). https://doi.org/10.1145/3375627.3375830

[21]. U.S. Department of Transportation, Pipeline and Hazardous Materials Safety Administration. (2026a). Distribution, transmission & gathering, LNG, and liquid accident and incident data. Retrieved September 2026, from https://www.phmsa.dot.gov/data-and-statistics/pipeline/distribution-transmission-gathering-lng-and-liquid-accident-and-incident-data

[22]. U.S. Department of Transportation, Pipeline and Hazardous Materials Safety Administration. (2026b). Pipeline incident flagged files; History of PHMSA incident reporting criteria. Retrieved September 2026, from https://www.phmsa.dot.gov/data-and-statistics/pipeline/pipeline-incident-flagged-files