WHEN SEMANTICS STAY THE SAME BUT PREDICTIONS CHANGE: AUDITING CHECKPOINT PORTABILITY FOR CONVERSATION-LOCAL SPEAKER INTERFACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Aggregate task performance can remain nearly unchanged even when a frozen checkpoint changes individual predictions under semantics-preserving preprocessing. We study this reliability gap for conversation-local nominal identifiers, where different serializers assign different integer names to the same local partition, and define *semantics-preserving checkpoint portability* as behavioral invariance across admissible re-serializations. We introduce a semantics-gated audit that first verifies preservation of the task-semantic object and then compares paired frozen executions at the decision, confidence, and unit-risk levels. The audit also separates three interface regimes: absolute-ID lookup, empirical mitigation through augmentation or JS consistency, and exact-contract canonical or relation-only representations. On held-out ECF Pair/Query evaluation, absolute-ID lookup yields unit-level risks of 0.6407 and 0.6103 while ΔF1 is only -0.0038 and $+0.0001$. Derived-ID MRDA and Entity Relation Classification reproduce the same portability pattern across task and identifier settings. Empirical mitigation improves selected portability metrics without yielding exact portability. Canonical and relation-only interfaces achieve exact consistency, zero drift, and zero portability risk under their implementation contracts. These results show that aggregate task scores do not certify checkpoint portability and motivate interface-aware reliability evaluation.

1 INTRODUCTION

Conversation-local speaker identifiers are nominal symbols: changing their integer names should not change a model's decision as long as the underlying conversation remains the same. Yet this seemingly obvious invariance is not guaranteed by common neural interfaces. Consider a frozen dialogue checkpoint evaluated through two preprocessing pipelines that assign different integers to the same speakers while preserving the utterances, order, candidate pairs, speaker-equality structure, and gold labels. With an absolute embedding lookup, $E[s_i]$, the same local participant can be routed through different embedding rows, causing the checkpoint to change individual predictions under a semantics-preserving re-serialization. Crucially, such failures may remain nearly invisible in aggregate metrics because prediction changes can cancel out at the dataset level. This exposes the central reliability gap studied in this paper: *semantic equivalence does not necessarily imply behavioral equivalence*.

We formalize the missing property as *semantics-preserving checkpoint portability*: a frozen checkpoint should preserve its functional behavior across admissible interface transformations that leave the task-semantic object unchanged. Figure 1 illustrates why aggregate evaluation is insufficient for certifying this property. Two serializers may encode exactly the same local nominal partition while assigning different integer IDs. Absolute-ID lookup can therefore alter the checkpoint's prediction profile even when the resulting change in aggregate F1 is negligible. In contrast, interfaces whose representations are invariant by construction—such as canonical or relation-only encodings—map admissible serializer views to the same representation under their stated contracts.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

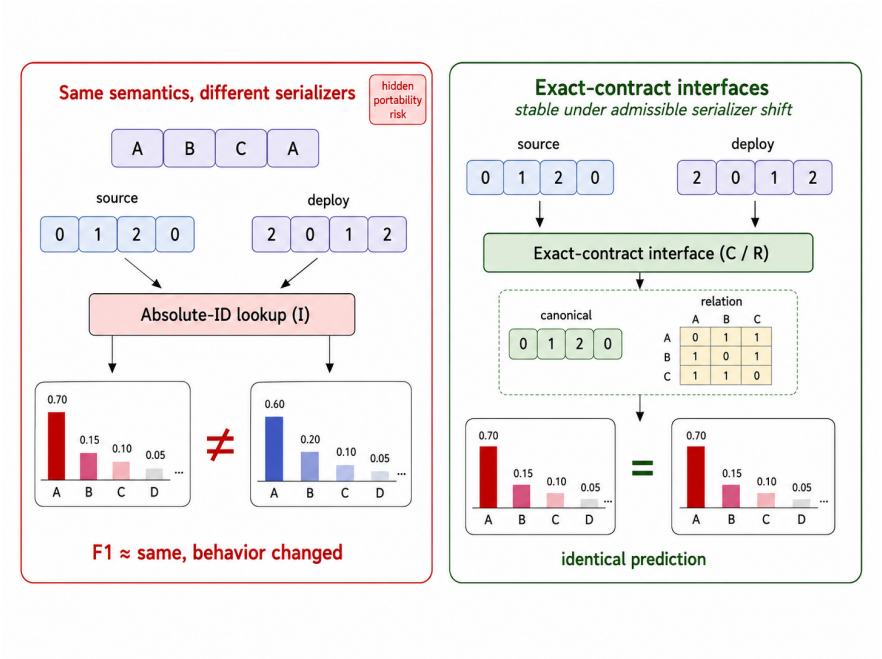

Figure 1: Motivating checkpoint-portability gap. Two serializers encode the same local nominal partition but assign different integer IDs. Absolute-ID lookup may therefore change individual predictions while leaving aggregate F1 nearly unchanged. Canonical and relation-only interfaces instead map admissible serializer views to identical representations under their respective contracts.

This issue is particularly relevant whenever implementation-level symbols are allowed to mediate a model's representation. Speaker-aware dialogue models exploit participant-specific states, typed graphs, discourse structure, and speaker-turn transitions (Majumder et al., 2019; Ghosal et al., 2019; He et al., 2021; Qamar et al., 2023; Ishiwatari et al., 2020; Shen et al., 2021; Sun et al., 2021; Liang et al., 2022), while ECPEC formulations similarly rely on structured conversational context (Wang et al., 2023; 2024b;a; Ma et al., 2026). These structures are useful, but they create a reliability hazard when an interface treats arbitrary local names as learnable identities even though the task semantics are defined by the underlying participant relations. A model may then become sensitive not only to *who is the same speaker*, but also to *which arbitrary integer happened to denote that speaker*.

Behavioral and metamorphic testing provide a natural starting point for detecting such failures: controlled transformations can expose invariance violations that ordinary aggregate accuracy does not reveal (Xie et al., 2011; Chen et al., 2018; Ribeiro et al., 2020; Gardner et al., 2020; Ribeiro & Lundberg, 2022; Elazar et al., 2021; Jang et al., 2022). However, nominal-identifier transformations introduce an additional difficulty: not every transformation is semantically valid. A bijective renaming can preserve the conversation-local participant partition, whereas speaker merges, semantic-role substitutions, or changes to persistent identity can modify the task object itself. Consequently, observing prediction differences is meaningful only after establishing that the compared inputs are genuinely semantically equivalent. Moreover, empirical techniques such as augmentation or consistency regularization may improve robustness on observed transformations (Kaushik et al., 2020; 2021; Huang et al., 2020), but improved robustness is conceptually different from an interface contract that guarantees invariance by construction.

To separate these questions, we introduce a *semantics-gated checkpoint portability audit*. The semantic gate first verifies that a transformation preserves the text, utterance order, labels, candidate keys, and conversation-local equality relation. Only admissible serializer views are then passed through the *same frozen checkpoint* and compared using exact decision consistency, structured-output agreement, confidence drift, the correct-to-incorrect (C→I) rate, unit-level portability risk, and ΔF1. This design separates three distinct interface regimes: *absolute-ID lookup*, *empirical mitigation* through augmentation or JS consistency, and *exact-contract interfaces* based on canonical or relation-only representations.

The resulting picture is substantially different from what aggregate F1 alone suggests. On held-out ECF Pair and Query evaluation, absolute-ID lookup reaches unit-level portability risks of 0.6407 and 0.6103, respectively, while ΔF1 is only -0.0038 and $+0.0001$. Thus, almost unchanged aggregate performance can coexist with substantial prediction-level instability. Derived-ID MRDA and Entity Relation Classification reproduce the same qualitative pattern beyond the original ECF setting. Empirical mitigation improves selected portability metrics but retains the underlying absolute-ID interface and does not provide exact portability. By contrast, canonical and relation-only interfaces achieve exact decision consistency, zero confidence drift, and zero portability risk under their explicit implementation contracts.

Contributions. We make three contributions:

- **Problem formulation.** We define *semantics-preserving checkpoint portability* and introduce a semantic gate that restricts cross-serializer behavioral comparisons to transformations that preserve the task-semantic object.
- **Prediction-level portability analysis.** Across held-out ECF Pair/Query evaluation, derived-ID MRDA, and Entity Relation Classification, we show that negligible changes in aggregate task scores can coexist with substantial unit-level behavioral divergence.
- **Interface-level reliability comparison.** We distinguish vulnerable absolute-ID lookup, empirical mitigation, and exact-contract interface design. Our results show that canonical and relation-only representations provide exact portability under explicit implementation contracts, whereas empirical mitigation improves robustness without establishing the same guarantee.

2 Related Work

Our work sits at the intersection of behavioral evaluation, speaker-aware dialogue modeling, and symmetry-aware representation design. These lines of work provide the ingredients for testing behavioral invariance, modeling local participant structure, and removing arbitrary naming degrees of freedom. We combine them around a checkpoint-level question: whether a fixed model preserves its decisions when a local nominal interface is re-serialized.

Behavioral and metamorphic evaluation. Metamorphic testing specifies relations that should hold across related executions (Xie et al., 2011) and has been adapted to machine-learning validation and neural-network debugging (Tian et al., 2018; Odena et al., 2019; Xie et al., 2020). NLP stress tests, contrast sets, CheckList, adaptive evaluation, and consistency tests likewise show that aggregate scores can hide capability-specific failures and prediction-level inconsistency (Naik et al., 2018; McCoy et al., 2019; Gardner et al., 2020; Ribeiro et al., 2020; Kiela et al., 2021; Ribeiro & Lundberg, 2022; Elazar et al., 2021; Jang et al., 2022). Our audit specializes this idea to local nominal re-serialization and adds an explicit semantic gate before any behavioral comparison.

Speaker-aware dialogue and ECPEC. Dialogue models exploit speaker-specific states, typed graphs, discourse relations, and turn transitions (Majumder et al., 2019; Ghosal et al., 2019; Ishiwatari et al., 2020; Shen et al., 2021; Sun et al., 2021; Liang et al., 2022; Shang et al., 2020; He et al., 2021; Qamar et al., 2023). ECPEC methods similarly use structured conversational context, graph reasoning, and emotion–cause alignment (Wang et al., 2023; 2024b;a; Ma et al., 2026). These methods establish the value of participant structure. Our evaluation holds that structure fixed and varies the integer names assigned to its local classes, isolating interface serialization from task semantics.

Invariant and empirical interface design. Canonicalization maps equivalent combinatorial objects to a common representative (McKay & Piperno, 2014); invariant and equivariant architectures encode known symmetries in sets, matrices, and graphs (Zaheer et al., 2017; Lee et al., 2019; Ravanbakhsh et al., 2017; Maron et al., 2019; Keriven & Peyré, 2019). Counterfactual augmentation and consistency objectives offer a complementary empirical route by exposing models to multiple equivalent views (Kaushik et al., 2020; 2021; Huang et al., 2020; Wang & Culotta, 2021). We use these two design families to separate learned robustness from representation-level invariance: augmentation and JS consistency provide empirical mitigation, while canonical and relation-only interfaces provide exact portability under explicit contracts.

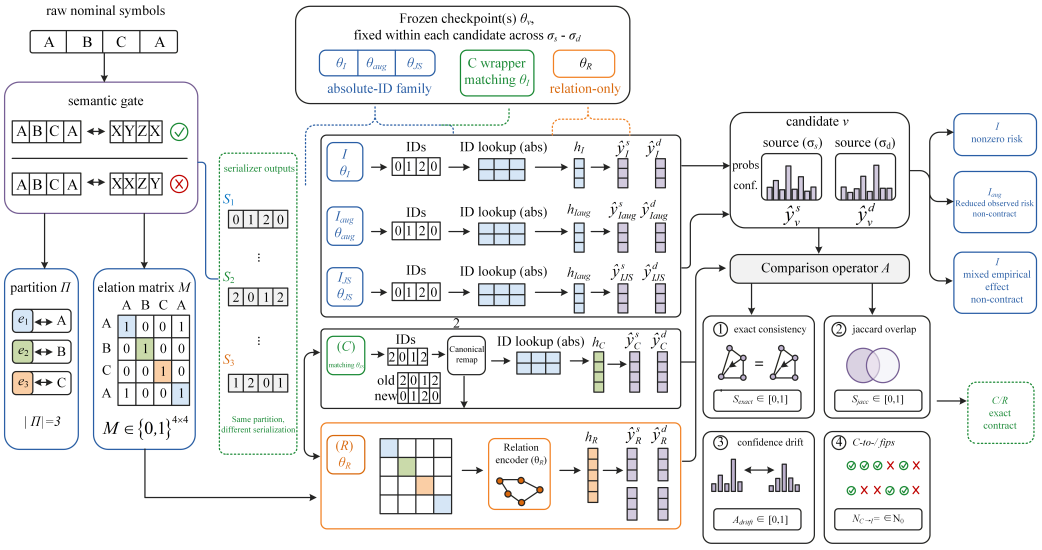

Figure 2: Semantics-gated checkpoint portability audit. Raw local symbols first pass a semantic gate that preserves the local partition and rejects semantics-changing transformations. Deterministic serializers then create equivalent identifier views for the same frozen candidate interfaces. A shared comparison operator measures exact consistency, structured overlap, confidence drift, and correct-to-incorrect transfer.

3 Semantics-Gated Checkpoint Portability Audit

Our audit has four stages: establish semantic equivalence, construct admissible serializer views, evaluate a fixed checkpoint through a declared interface, and compare paired behavior. Figure 2 summarizes this pipeline and the three interface families used in the experiments.

Figure 2 makes the audit boundary explicit. The left side determines whether two views represent the same task object before any behavioral comparison is made. The center isolates the interface choice: absolute-ID lookup retains serializer-dependent names, empirical repairs keep that lookup while training across views, and the canonical and relation-only paths remove the naming degree of freedom from the evaluated representation. The right side applies the same comparison operator to paired frozen outputs, linking each interface design directly to the portability metrics used in Section 4.

3.1 Problem Formulation and Semantic Gate

For evaluation unit d, let $x_d = (u_1, \ldots, u_n)$ denote the semantic input, $a = (a_1, \ldots, a_n)$ the raw local identifier strings, and $\mathcal{P}$ the induced local equivalence partition:

$$i \sim_{\mathcal{P}} j \quad \Longleftrightarrow \quad a_i = a_j. \tag{1}$$

A deterministic serializer σ assigns contiguous integers to the classes in $\mathcal{P}$,

$$\mathbf{s}_\sigma = \sigma(\mathcal{P}) = (s_{\sigma,1}, \ldots, s_{\sigma,n}). \tag{2}$$

The integers serve as local class names; their magnitude and ordering carry no task semantics in the setting studied here.

The semantic gate admits a source/deployment pair (σ_s, σ_d) when both serializers preserve the same local partition,

$$s_{\sigma_s,i} = s_{\sigma_s,j} \quad \Longleftrightarrow \quad s_{\sigma_d,i} = s_{\sigma_d,j}, \qquad \forall i, j, \tag{3}$$

and text, utterance order, raw identifier strings, candidate keys, and gold labels are unchanged. Local nominal bijections satisfy this gate. Speaker merges or splits, semantic role substitutions, persistent-identity changes, and fabricated identifier fields fail it and receive no portability score.

Let Σ_{adm} be the admitted serializers and let ϕ_v be interface v with frozen checkpoint θ_v. We define *semantics-preserving checkpoint portability* by

$$f_{\theta_v}\left(x, \phi_v(\mathbf{s}_{\sigma_1})\right) = f_{\theta_v}\left(x, \phi_v(\mathbf{s}_{\sigma_2})\right), \quad \forall \sigma_1, \sigma_2 \in \Sigma_{\text{adm}}. \tag{4}$$

The property concerns re-serialization of a fixed local nominal interface.

3.2 Admissible Serializer Views

For K local classes, each serializer is a bijection onto $\{0, \ldots, K-1\}$. We instantiate three deterministic views: S_1 assigns IDs by first appearance, S_2 orders distinct raw strings lexicographically, and S_3 uses stable-hash (SHA-256) order with the raw string as a deterministic tiebreaker. For example,

$$S_1 : (0, 1, 2, 0), \qquad S_2 : (2, 0, 1, 2), \qquad S_3 : (1, 2, 0, 1). \tag{5}$$

The sequence of integers changes while the equality partition and all non-identifier task inputs remain fixed. A source-trained checkpoint is executed under its source view and an admissible deployment view with the same checkpoint, decoder, candidate ordering, and labels.

3.3 Interface Designs

We evaluate five interfaces, $\{I, I_{\text{aug}}, I_{\text{JS}}, C, R\}$, spanning direct lookup, empirical mitigation, and exact-contract representations.

Absolute-ID lookup. The baseline embeds the serialized integer directly:

$$\phi_I(\mathbf{s}_\sigma)_i = E[s_{\sigma,i}]. \tag{6}$$

An admissible serializer can route the same local class to a different learned embedding row, making the interface serializer-sensitive by construction.

Empirical mitigation. I_{aug} keeps the absolute-ID interface and trains on deterministic nonidentity re-serializations. I_{JS} adds agreement between two serializer views:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{task}} + \lambda_{\text{JS}} \mathcal{L}_{\text{JS}}, \\ \mathcal{L}_{\text{JS}} &= \frac{1}{m} \sum_{k=1}^{m} \text{JS}(\text{Ber}(\widehat{p}_k^{\sigma_a}) \parallel \text{Ber}(\widehat{p}_k^{\sigma_b})) \,. \end{aligned} \tag{7}$$

These interfaces retain absolute-ID lookup, so their portability is measured empirically across the declared serializer views.

Canonical interface. C canonicalizes local classes by first appearance:

$$c(\mathbf{s}_{\sigma_1}) = c(\mathbf{s}_{\sigma_2}) = c(\mathcal{P}), \tag{8}$$

with

$$\phi_C(\mathbf{s}_\sigma)_i = E[c(\mathbf{s}_\sigma)_i]. \tag{9}$$

Because S_1 already uses first appearance, C reuses the matching frozen I model state through a canonicalization wrapper.

Relation-only interface. R removes integer names and represents the equality relation directly:

$$M_{ij}(\mathbf{s}_\sigma) = \mathbb{I}[s_{\sigma,i} = s_{\sigma,j}]. \tag{10}$$

For any admitted pair,

$$M(\mathbf{s}_{\sigma_1}) = M(\mathbf{s}_{\sigma_2}). \tag{11}$$

Model R consumes M or deterministic features derived from M.

Canonical and relation-only interfaces satisfy

$$\begin{aligned} \phi_C(\mathbf{s}_{\sigma_1}) &= \phi_C(\mathbf{s}_{\sigma_2}), \\ \phi_R(\mathbf{s}_{\sigma_1}) &= \phi_R(\mathbf{s}_{\sigma_2}), \end{aligned} \qquad \sigma_1, \sigma_2 \in \Sigma_{\text{adm}}. \tag{12}$$

With a fixed checkpoint and deterministic inference, identical interface inputs produce identical outputs. The exact contract covers every identifier-dependent path in the model input pipeline.

3.4 Frozen Behavioral Comparison

For eligible unit d, interface v, and serializer view $q \in \{s, d\}$,

$$\begin{aligned} \widehat{\mathbf{p}}^{q}_{d,v} &= f_{\theta_v}\big(x_d, \phi_v(\mathbf{s}_{d,\sigma_q})\big)\,, \\ \widehat{\mathbf{y}}^{q}_{d,v} &= g_{\mathcal{T}}(\widehat{\mathbf{p}}^{q}_{d,v}), \end{aligned} \tag{13}$$

where $g_{\mathcal{T}}$ is the frozen task decision rule. For ECF it applies the fixed threshold τ to aligned candidate-pair probabilities; the classification settings use their frozen decoders.

Let m_d be the number of aligned prediction instances and $\mathbf{y}_d$ the aligned gold labels. We report exact decision consistency,

$$S_{\text{dec}} = \frac{\sum_{d\in\mathcal{D}} \sum_{k=1}^{m_d} \mathbb{I}[\widehat{y}^{s}_{d,v,k} = \widehat{y}^{d}_{d,v,k}]}{\sum_{d\in\mathcal{D}} m_d}, \tag{14}$$

structured-output Jaccard agreement where applicable,

$$S_{\text{jacc}} = \frac{1}{|\mathcal{D}|} \sum_{d\in\mathcal{D}} \frac{|\mathcal{E}^{s}_{d,v} \cap \mathcal{E}^{d}_{d,v}|}{|\mathcal{E}^{s}_{d,v} \cup \mathcal{E}^{d}_{d,v}|}, \tag{15}$$

and mean absolute confidence drift,

$$\Delta_{\text{drift}} = \frac{\sum_{d\in\mathcal{D}} \sum_{k=1}^{m_d} |c^{s}_{d,v,k} - c^{d}_{d,v,k}|}{\sum_{d\in\mathcal{D}} m_d}. \tag{16}$$

We also count source-correct decisions that become incorrect,

$$N_{\text{C}\to\text{I}} = \sum_{d\in\mathcal{D}} \sum_{k=1}^{m_d} \mathbb{I}[\widehat{y}^{s}_{d,v,k} = y_{d,k}]\mathbb{I}[\widehat{y}^{d}_{d,v,k} \neq y_{d,k}], \tag{17}$$

with normalized rate

$$r_{\text{C}\to\text{I}} = \frac{N_{\text{C}\to\text{I}}}{\sum_{d\in\mathcal{D}} m_d}. \tag{18}$$

Unit-level portability risk is

$$R_{\text{unit}} = \frac{1}{|\mathcal{D}|} \sum_{d\in\mathcal{D}} \mathbb{I}[\widehat{\mathbf{y}}^{s}_{d,v} \neq \widehat{\mathbf{y}}^{d}_{d,v}], \tag{19}$$

where the unit is a conversation for ECF, a dialogue for MRDA, and an entity-relation group for Entity Relation Classification. Finally,

$$\Delta\text{F1} = \text{F1}_{\sigma_d} - \text{F1}_{\sigma_s}. \tag{20}$$

Together, these metrics separate decision agreement, structured overlap, confidence movement, harmful flips, affected-unit risk, and aggregate task performance. The main audit uses deterministic $S_1/S_2/S_3$ transfers.

4 Experiments: Checkpoint Portability Under Equivalent Serializers

4.1 Experimental Setup

We evaluate ECPEC Pair/Query, held-out MRDA dialogue-act classification with derived per-dialogue IDs, and Entity Index Relation Classification. ECF uses licensed train/development data for protocol design and a separately frozen held-out set (Wang et al., 2023). Matched I, C, R interfaces form the main comparison; I_{aug} and I_{JS} enter the ECF interface-design ablation. Every admitted transfer preserves text, order, candidate keys, labels, and the equality partition. We report exact consistency, portability risk, confidence drift, C→I transfer, and ΔF1.

Checkpoint, threshold, serializer, and evaluation subset are frozen before the held-out evaluation. ECF development uses 937 training and 109 development conversations with seeds 6201–6202. The held-out evaluation uses 250 of 261 conversations, 20 checkpoints, $\tau = 0.50$, and 2,000 paired bootstraps. The V++ development analysis uses seeds 7301–7303 after the held-out protocol is closed. MRDA uses 12 held-out dialogues with derived local IDs; Entity uses fixed relation groups with a post-freeze summary.

Table 1: Main frozen-checkpoint portability results. ECF and MRDA intervals use paired conversation- and dialogue-level bootstrap, respectively; Entity is a fixed post-freeze summary. C/R denotes the canonical and relation-only interfaces, which share the same exact-portability outcomes under their contracts.

Setting	Intf.	Exact	Unit risk [95% CI]	Drift	C→I rate	ΔF1 [95% CI]
ECF Pair	I	0.9490	0.6407 [0.5982, 0.6831]	0.0360	0.0270	-0.0038 [-0.0125, 0.0049]
	C/R	1.0000	0 [0, 0]	0	0	0 [0, 0]
ECF Query	I	0.9744	0.6103 [0.5668, 0.6539]	0.0214	0.0128	$+0.0001$ [-0.0051, 0.0053]
	C/R	1.0000	0 [0, 0]	0	0	0 [0, 0]
MRDA lexical	I	0.8942	0.9167 [0.7500, 1.0000]	0.0281	0.0490	-0.0049 [-0.0162, 0.0058]
	C/R	1.0000	0 [0, 0]	0	0	0 [0, 0]
MRDA stable-hash	I	0.8853	1.0000 [1.0000, 1.0000]	0.0309	0.0526	-0.0024 [-0.0135, 0.0071]
	C/R	1.0000	0 [0, 0]	0	0	0 [0, 0]
Entity relation	I	0.9125	0.5833	0.0325	0.0326	-0.0012
	C/R	1.0000	0	0	0	0

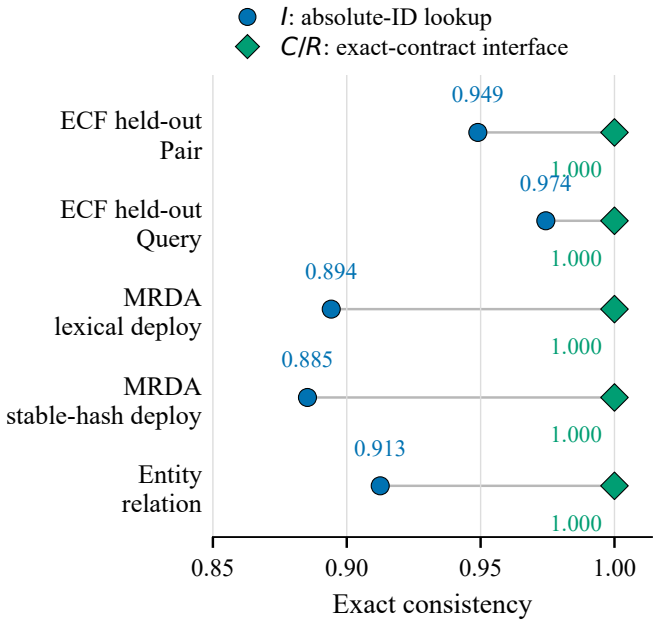

Figure 3: Exact decision consistency across the main evaluation settings. Absolute-ID lookup remains below 1.0 in ECF Pair/Query, both MRDA deployment serializers, and Entity Relation Classification. Canonical and relation-only interfaces remain at 1.0 under their contracts.

4.2 Main Result: Aggregate Stability Hides Behavioral Divergence

Table 1 shows the central mismatch. Absolute-ID lookup is prediction-level unstable in every setting: unit-level risk spans 0.5833–1.0000 while all reported ΔF1 values remain small. On held-out ECF Pair and Query, risk reaches 0.6407 and 0.6103 with ΔF1 of -0.0038 and $+0.0001$. Canonical and relation-only interfaces yield exact consistency, zero drift, zero C→I transfer, and zero risk under their implementation contracts. The MRDA stable-hash interval $[1, 1]$ is a concrete held-out outcome: all 12 dialogues change in every bootstrap resample.

Figure 3 highlights the recurring interface-level separation across ECF, MRDA, and Entity: absolute-ID lookup is non-exact in every evaluated setting, whereas C/R remain exact under their contracts.

Robustness across serializer directions. The pooled held-out effect is not driven by a single transfer path. For absolute-ID lookup, all six nonidentity source–deployment directions are non-exact. Pair exact consistency ranges from 0.9400 to 0.9620 with unit risk from 0.598 to 0.718; Query exact consistency ranges from 0.9699 to 0.9788 with unit risk from 0.546 to 0.700. The pooled result therefore reflects a recurring serializer effect rather than one outlying migration.

The decision-level effect is substantial even when aggregate change is small. Across the ECF development transfers, literal I has mean ΔF1 of -0.001135 for Pair and $+0.001374$ for Query, while 3,873 and 2,175 source-correct decisions become incorrect. Relative to S_1, S_2 changes 674 training and 76 development ID sequences; S_3 changes 669 and 79. Every changed view preserves the local partition and task-semantic fields.

4.3 Replication Across Serializers, Tasks, and Identifier Types

The result persists across three controlled preprocessing migrations. Encounter-order ($A \to B$), regenerated-vocabulary ($A \to C$), and SHA-256 cache ($B \to C$) migrations produce Pair exactness of $0.9377/0.9302/0.9279$ and Query exactness of $0.9694/0.9659/0.9648$. Pair drift is $0.0442/0.0466/0.0481$ and Query drift is $0.0217/0.0233/0.0248$. All three paths preserve text, labels, order, candidate keys, and equality for the 109 eligible conversations; canonical mapping recovers the first-appearance representation exactly.

MRDA changes the task while retaining a derived local-ID interface: raw participant strings are excluded from the model input and only the per-dialogue nominal-ID tensor is re-serialized. Entity Relation Classification changes the identifier type from speakers to local entity indices. Across the two MRDA re-serializations and Entity Relation Classification, absolute-ID unit risk ranges from 0.5833 to 1.0000 while the largest absolute F1 change is only 0.0049. The same mismatch therefore persists after changing both the prediction task and the type of local nominal identifier. Their rows in Table 1 and Figure 3 show the same interface-level signature: absolute lookup changes decisions under admissible re-serialization, whereas exact-contract interfaces remain invariant.

4.4 Interface Design: Empirical Mitigation vs. Exact Contracts

A separate V++ development analysis compares direct lookup, two empirical repairs, and two exact-contract interfaces after the held-out protocol is closed.

Table 2: Interface-design ablation over seeds 7301–7303 and two frozen nonidentity deployments. I_{aug} and I_{JS} are empirical mitigation interfaces; C and R implement exact contracts.

Fam.	Intf.	F1	Exact	Risk	Drift	C→I
Pair	I	0.3229	0.9339	0.6529	0.0454	513.7
Pair	I_{aug}	0.3808	0.9712	0.5826	0.0227	187.8
Pair	I_{JS}	0.3424	0.9399	0.6544	0.0445	450.8
Pair	C	0.3229	1.0000	0.0000	0.0000	0.0
Pair	R	0.3710	1.0000	0.0000	0.0000	0.0
Query	I	0.3942	0.9677	0.6498	0.0225	208.3
Query	I_{aug}	0.4226	0.9793	0.5765	0.0189	144.5
Query	I_{JS}	0.4299	0.9789	0.6208	0.0167	132.8
Query	C	0.3942	1.0000	0.0000	0.0000	0.0
Query	R	0.4104	1.0000	0.0000	0.0000	0.0

The ablation separates empirical mitigation from exact portability. I_{aug} raises Pair exact consistency from 0.9339 to 0.9712 and reduces risk from 0.6529 to 0.5826. For Query, I_{JS} reaches exact consistency 0.9789 with risk 0.6208. Pair I_{JS} does not uniformly improve portability risk over I (0.6544 versus 0.6529), so the empirical repairs improve selected portability metrics but remain non-exact.

Task performance and portability are also distinct within the same ablation. On Query, I_{JS} attains source F1 0.4299 yet retains unit risk 0.6208, whereas R attains F1 0.4104 with zero portability risk. Source task performance therefore does not by itself identify the interface with stronger portability.

Contract verification across frozen transfers. The exact-contract behavior is verified across the complete held-out transfer grid. For each of Pair–C, Pair–R, Query–C, and Query–R, all 12 seed-specific nonidentity transfers attain exact consistency 1.0 with zero unit risk, zero confidence drift, zero C→I transfer, and zero $|\Delta$F1—. $Across the 48 audited$ C/R $transfer rows, the empirical consequence of the representation-level contract is therefore observed on every frozen held-out serializer direction.$

4.5 Robustness and Audit-Boundary Checks

Decision-threshold robustness. Recomputing the same frozen held-out Pair probabilities over $\tau \in \{0.30, 0.40, 0.50, 0.60, 0.70\}$ preserves the diagnosis. Absolute-ID unit risk remains nonzero throughout the sweep (0.5210–0.7105), whereas C/R retain exact decision consistency and zero risk. Confidence drift is computed before thresholding and is therefore unchanged by this sweep. The portability gap is thus not an artifact of a single decision threshold.

Semantic-gate controls. We separately test transformations at the audit boundary. Local nominal bijections are admitted because they preserve the equality partition and all task-semantic fields. Speaker merges or splits are rejected because they alter the partition; semantic role swaps and global-identity renamings are rejected because their symbols are not interchangeable local names; fabricated identifier fields are rejected because no serializer-source interface exists. Rejected transformations receive no portability score. These controls ensure that the reported behavioral differences are measured only between semantically admissible views.

5 Conclusion

Semantics-preserving checkpoint portability captures a reliability property that aggregate task scores can miss. Our audit first verifies semantic admissibility and then compares the same frozen checkpoint across equivalent serializer views, exposing prediction changes that remain hidden behind small changes in F1.

Across ECF, derived-ID MRDA, and Entity Relation Classification, absolute-ID lookup shows substantial prediction-level divergence under admissible re-serialization. Empirical augmentation and JS consistency improve selected portability metrics but remain non-exact. Canonical and relation-only interfaces achieve exact consistency, zero drift, and zero portability risk under their implementation contracts. The resulting distinction—vulnerable lookup, empirical mitigation, and exact-contract design—provides the central reliability lens for local nominal interfaces.

AI use statement

We used ChatGPT (OpenAI) to assist with manuscript editing, structural revision, reviewer-style critique, and feedback on the presentation of the methodology and experimental protocol. All AI-assisted suggestions and text were reviewed and revised by the authors. Numerical results, experimental records, citations, mathematical statements, and final scientific claims were checked by the authors against the underlying code and frozen evaluation artifacts. Generative AI was not used to generate experimental data or to select reported experimental results. The authors take full responsibility for the final content of the submission.

Ethics statement

This work audits frozen checkpoints on benchmark-derived data using controlled, semantics-preserving interface transformations. It collects no new human-subject or personal data and performs no deployment intervention. The ethical considerations center on faithful reporting of the declared datasets, transformations, and implementation contracts.

Reproducibility statement

The main paper specifies the semantic gate, admissible serializer views, interface designs, frozen evaluation protocol, and behavioral metrics. The supplementary material provides complete development transfers, frozen held-out transfer details, evaluation protocols, threshold robustness, repair comparisons, preprocessing-migration controls, and semantic-gate negative controls. The anonymous companion files retain the seed-specific held-out serializer-portability rows and their machine-readable JSON record.

REFERENCES

Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, T. H. Tse, and Zhi Quan Zhou. Metamorphic testing: A review of challenges and opportunities. *ACM Computing Surveys*, 51(1), January 2018. doi: 10.1145/3143561.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410. URL https://aclanthology.org/2021.tacl-1.60/.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of EMNLP-IJCNLP*, pp. 154–164, 2019. doi: 10.18653/v1/d19-1015.

Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani. Speaker turn modeling for dialogue act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2150–2157, 2021. doi: 10.18653/v1/2021.findings-emnlp.185.

William Huang, Haokun Liu, and Samuel R. Bowman. Counterfactually-augmented SNLI training data does not yield better generalization than unaugmented data. *arXiv preprint arXiv:2010.04762*, 2020. URL https://arxiv.org/abs/2010.04762.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7360–7370, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.597. URL https://aclanthology.org/2020.emnlp-main.597/.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. BECEL: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3680–3696, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics. doi: 10.18653/v1/2022.coling-1.324. URL https://aclanthology.org/2022.coling-1.324/.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C. Lipton. Explaining the efficacy of counterfactually augmented data. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=HHiiQKWsOcV.

Nicolas Keriven and Gabriel Peyré. Universal invariant and equivariant graph neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 15868–15876. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/ea9268cb43f55d1d12380fb6ea5bf572-Abstract.html.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Chen Liang, Jing Xu, Yangkun Lin, Chong Yang, and Yongliang Wang. S+PAGE: A speaker and position-aware graph neural network model for emotion recognition in conversation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 148–157, Online, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-main.12. URL https://aclanthology.org/2022.aacl-main.12/.

Tianxiang Ma, Weijie Feng, Xinyu Wang, and Zhiyong Cheng. Emotion–cause pair extraction in conversations via semantic decoupling and alignment. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12697–12711, 2026. doi: 10.18653/v1/2026.findings-acl.617.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6818–6825, 2019. doi: 10.1609/aaai.v33i01.33016818.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Syx72jC9tm.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Brendan D. McKay and Adolfo Piperno. Practical graph isomorphism, ii. *Journal of Symbolic Computation*, 60:94–112, 2014. doi: 10.1016/j.jsc.2013.09.003.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2340–2353, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1198/.

Augustus Odena, Catherine Olsson, David Andersen, and Ian Goodfellow. TensorFuzz: Debugging neural networks with coverage-guided fuzzing. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4901–4911. PMLR, 2019. URL https://proceedings.mlr.press/v97/odena19a.html.

Ayesha Qamar, Adarsh Pyarelal, and Ruihong Huang. Who is speaking? speaker-aware multiparty dialogue act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10122–10135, 2023. doi: 10.18653/v1/2023.findings-emnlp.678.

Siamak Ravanbakhsh, Jeff Schneider, and Barnabás Póczos. Equivariance through parameter-sharing. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2892–2901. PMLR, 2017. URL https://proceedings.mlr.press/v70/ravanbakhsh17a.html.

Marco Tulio Ribeiro and Scott Lundberg. Adaptive testing and debugging of NLP models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3253–3267, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.230. URL https://aclanthology.org/2022.acl-long.230/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442.

Guokan Shang, Antoine Tixier, Michalis Vazirgiannis, and Jean-Pierre Lorré. Speaker-change aware CRF for dialogue act classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 450–464, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.40. URL https://aclanthology.org/2020.coling-main.40/.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1551–1560, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.123. URL https://aclanthology.org/2021.acl-long.123/.

Yang Sun, Nan Yu, and Guohong Fu. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics:*

EMNLP 2021, pp. 2949–2958, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.252. URL https://aclanthology.org/2021.findings-emnlp.252/.

Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. DeepTest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th International Conference on Software Engineering*, ICSE '18, pp. 303–314, New York, NY, USA, 2018. Association for Computing Machinery. doi: 10.1145/3180155.3180220.

Botao Wang, Keke Tang, and Peican Zhu. Enhancing emotion-cause pair extraction in conversations via center event detection and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10773–10783, Miami, Florida, USA, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.632. URL https://aclanthology.org/2024.findings-emnlp.632/.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844, 2023. doi: 10.1109/TAFFC.2022.3226559.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, pp. 2022–2033, 2024b. URL https://aclanthology.org/2024.semeval2024-1.273.

Zhao Wang and Aron Culotta. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14024–14031, 2021. doi: 10.1609/aaai.v35i16.17651.

Xiaoyuan Xie, Joshua W. K. Ho, Christian Murphy, Gail Kaiser, Baowen Xu, and Tsong Yueh Chen. Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software*, 84(4):544–558, 2011. doi: 10.1016/j.jss.2010.11.920.

Xiaoyuan Xie, Zhiyi Zhang, Tsong Yueh Chen, Yang Liu, Pak-Lok Poon, and Baowen Xu. METTLE: A METamorphic testing approach to assessing and validating unsupervised machine LEarning systems. *IEEE Transactions on Reliability*, 69(4):1293–1322, 2020. doi: 10.1109/TR.2020.2972266.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper/6931-deep-sets.

A COMPLETE DEVELOPMENT SERIALIZER TRANSFERS

Tables S1–S2 contain every frozen, non-diagonal development transfer used by the pooled development values in the main paper. Section B separately reports the frozen held-out transfer summaries; the anonymous Markdown and JSON companion files retain the seed-specific held-out rows. S1, S2, and S3 denote first appearance, lexical raw-speaker-string order, and stable-hash (SHA-256) raw-speaker-string order. A row changes the deployment serializer while keeping the checkpoint and task-semantic inputs fixed. Dec. Cons. is aligned decision agreement, Jaccard is averaged per conversation, drift is mean absolute candidate-confidence change, C-to-I counts decisions that move from correct to incorrect, and F1 delta is target minus source.

Table S1: Complete pair-family non-diagonal serializer transfers. Each of C and R uses one S_1-trained model state per seed across serializer views, because canonicalization or equality projection removes the serializer naming degree of freedom.

Variant	Seed	Source	Deploy	Dec. Cons.	Jaccard	Drift	C-to-I count	F1 delta
I	6201	S1	S2	0.935888	0.780963	0.044399	495	-0.008067
I	6201	S1	S3	0.928698	0.774470	0.049356	551	-0.012674
I	6201	S2	S1	0.963089	0.851352	0.031158	228	+0.007879
I	6201	S2	S3	0.957086	0.840975	0.035262	293	+0.000119
I	6201	S3	S1	0.962793	0.873633	0.026384	251	+0.000758
I	6201	S3	S2	0.959013	0.855812	0.030070	271	+0.000959
I	6202	S1	S2	0.962719	0.857832	0.035651	250	-0.001745
I	6202	S1	S3	0.966425	0.854051	0.034462	246	-0.009583
I	6202	S2	S1	0.963905	0.849104	0.032705	251	+0.002621
I	6202	S2	S3	0.961977	0.823751	0.034513	265	+0.001603
I	6202	S3	S1	0.938556	0.785041	0.041970	398	+0.005422
I	6202	S3	S2	0.941076	0.768270	0.042041	374	-0.000909
I_aug	6201	S1	S2	0.973540	0.918791	0.022696	156	+0.001947
I_aug	6201	S1	S3	0.969612	0.903091	0.024375	179	+0.003861
I_aug	6201	S2	S1	0.979766	0.934934	0.018300	139	-0.000479
I_aug	6201	S2	S3	0.977246	0.925815	0.019814	155	-0.001881
I_aug	6201	S3	S1	0.974059	0.915655	0.020318	183	-0.002912
I_aug	6201	S3	S2	0.974133	0.909969	0.021640	191	-0.003333
I_aug	6202	S1	S2	0.973836	0.913392	0.021460	182	-0.000881
I_aug	6202	S1	S3	0.970353	0.896944	0.022977	202	+0.000230
I_aug	6202	S2	S1	0.979247	0.920139	0.016524	138	+0.000454
I_aug	6202	S2	S3	0.979099	0.915690	0.018578	142	-0.001249
I_aug	6202	S3	S1	0.971168	0.892239	0.025315	209	-0.002007
I_aug	6202	S3	S2	0.970649	0.881945	0.026962	225	-0.009045
C	6201	S1	S2	1.000000	1.000000	0.000000	0	+0.000000
C	6201	S1	S3	1.000000	1.000000	0.000000	0	+0.000000
C	6202	S1	S2	1.000000	1.000000	0.000000	0	+0.000000
C	6202	S1	S3	1.000000	1.000000	0.000000	0	+0.000000
R	6201	S1	S2	1.000000	1.000000	0.000000	0	+0.000000
R	6201	S1	S3	1.000000	1.000000	0.000000	0	+0.000000
R	6202	S1	S2	1.000000	1.000000	0.000000	0	+0.000000
R	6202	S1	S3	1.000000	1.000000	0.000000	0	+0.000000

B FROZEN HELD-OUT SERIALIZER-TRANSFER DETAILS

The main paper reports the held-out means pooled over all nonidentity serializer transfers. The frozen companion artifact contains 72 aggregate rows spanning Pair/Query, variants $I/C/R$, seeds 6201–6202, and the six nonidentity source–deployment serializer pairs. Table S3 shows the arithmetic mean over the two frozen seeds for each absolute-ID transfer path. These rows are derived directly from `heldout_serializer_portability_rows.json`; the companion Markdown file retains the two seed-specific rows and their 2,000-resample percentile bootstrap intervals.

Averaging the 12 seed-specific absolute-ID rows within each task family (equivalently, the six two-seed transfer means shown in Table S3) gives Pair exact consistency 0.948958, risk 0.640667, drift 0.036001, C-to-I rate 0.026877, and ΔF1 -0.003797; Query exact consistency 0.974356, risk 0.610333, drift 0.021377, C-to-I rate 0.012754, and ΔF1 $+0.000084$. The main paper displays the corresponding entries as 0.9490/0.6407/0.0360/0.0270/-0.0038 and 0.9744/0.6103/0.0214/0.0128/$+0.0001$.

Table S2: Complete query-family non-diagonal serializer transfers. Each of C and R uses one S_1-trained model state per seed across serializer views, because canonicalization or equality projection removes the serializer naming degree of freedom.

Variant	Seed	Source	Deploy	Dec. Cons.	Jaccard	Drift	C-to-I count	F1 delta
I	6201	S1	S2	0.971613	0.894730	0.022894	167	+0.003360
I	6201	S1	S3	0.970798	0.896466	0.023928	167	+0.006757
I	6201	S2	S1	0.974133	0.909412	0.020183	205	-0.002037
I	6201	S2	S3	0.971020	0.895030	0.022296	188	+0.003858
I	6201	S3	S1	0.969019	0.887279	0.025329	229	-0.005452
I	6201	S3	S2	0.965906	0.876321	0.027390	241	-0.004809
I	6202	S1	S2	0.971390	0.896973	0.021851	183	+0.001768
I	6202	S1	S3	0.971020	0.895753	0.022898	186	+0.000034
I	6202	S2	S1	0.973614	0.901778	0.021814	170	+0.006110
I	6202	S2	S3	0.968055	0.883057	0.025419	202	+0.007013
I	6202	S3	S1	0.984732	0.943313	0.014100	117	-0.000773
I	6202	S3	S2	0.981989	0.932044	0.015850	120	+0.000661
I_aug	6201	S1	S2	0.979618	0.915633	0.019135	139	-0.000017
I_aug	6201	S1	S3	0.977765	0.910683	0.020315	155	-0.001862
I_aug	6201	S2	S1	0.984287	0.930908	0.014651	111	-0.000225
I_aug	6201	S2	S3	0.982286	0.924874	0.017222	118	+0.000730
I_aug	6201	S3	S1	0.980878	0.922912	0.018883	122	+0.001682
I_aug	6201	S3	S2	0.979543	0.915322	0.020625	131	+0.002026
I_aug	6202	S1	S2	0.988512	0.949218	0.011597	77	-0.000571
I_aug	6202	S1	S3	0.985473	0.940743	0.012958	107	-0.003641
I_aug	6202	S2	S1	0.989549	0.958100	0.010537	71	-0.000119
I_aug	6202	S2	S3	0.987771	0.949658	0.012088	75	+0.002133
I_aug	6202	S3	S1	0.989994	0.959107	0.009658	67	+0.000123
I_aug	6202	S3	S2	0.990216	0.959777	0.010135	60	+0.002853
C	6201	S1	S2	1.000000	1.000000	0.000000	0	+0.000000
C	6201	S1	S3	1.000000	1.000000	0.000000	0	+0.000000
C	6202	S1	S2	1.000000	1.000000	0.000000	0	+0.000000
C	6202	S1	S3	1.000000	1.000000	0.000000	0	+0.000000
R	6201	S1	S2	1.000000	1.000000	0.000000	0	+0.000000
R	6201	S1	S3	1.000000	1.000000	0.000000	0	+0.000000
R	6202	S1	S2	1.000000	1.000000	0.000000	0	+0.000000
R	6202	S1	S3	1.000000	1.000000	0.000000	0	+0.000000

Table S3: Held-out absolute-ID (I) portability by serializer path. Each row is the arithmetic mean over seeds 6201–6202 for the indicated frozen source–deployment transfer. S_1, S_2, and S_3 are first appearance, lexical order, and stable-hash order. No model parameter, threshold, candidate set, or label changes between the paired executions.

Family	Transfer	Exact	Risk	Drift	C-to-I rate	ΔF1
Pair	$S_1 \to S_2$	0.946022	0.616000	0.038504	0.030246	-0.009580
Pair	$S_1 \to S_3$	0.942263	0.598000	0.040733	0.031647	-0.009688
Pair	$S_2 \to S_1$	0.962021	0.608000	0.029871	0.019636	-0.001863
Pair	$S_2 \to S_3$	0.952598	0.694000	0.036131	0.023838	-0.003168
Pair	$S_3 \to S_1$	0.950878	0.610000	0.032876	0.025512	+0.001006
Pair	$S_3 \to S_2$	0.939964	0.718000	0.037892	0.030383	+0.000513
Query	$S_1 \to S_2$	0.973666	0.608000	0.021013	0.012513	+0.000329
Query	$S_1 \to S_3$	0.972311	0.606000	0.022284	0.012710	+0.001455
Query	$S_2 \to S_1$	0.975843	0.576000	0.019817	0.012924	-0.001466
Query	$S_2 \to S_3$	0.969921	0.700000	0.024844	0.013959	+0.004458
Query	$S_3 \to S_1$	0.978841	0.546000	0.018600	0.011797	-0.003200
Query	$S_3 \to S_2$	0.975553	0.626000	0.021703	0.012619	-0.001075

Table S4: Held-out verification of the exact-contract interfaces in the frozen artifact. Each family–interface pair contains all 12 nonidentity transfer rows (two seeds × six source–deployment pairs).

Family	Intf.	Rows	Exact min/max	Risk max	Drift max	$\lvert\Delta$F1$\rvert$ max
Pair	C	12	1.000/1.000	0	0	0
Pair	R	12	1.000/1.000	0	0	0
Query	C	12	1.000/1.000	0	0	0
Query	R	12	1.000/1.000	0	0	0

Table S5: Controlled datasets and frozen evaluation protocols. Each comparison changes the declared local nominal-interface realization while preserving the task-semantic inputs and labels.

Setting	Eligible units	Views / interfaces	Frozen protocol
ECF dev.	937 train; 109 dev.	$S_1/S_2/S_3$; $I, I_{\mathrm{aug}}, C, R$	Seeds 6201–6202; protocol development
ECF held-out	250 of 261 conversations	$S_1/S_2/S_3$; I, C, R	20 checkpoints; $\tau = 0.50$; 2,000 paired bootstraps
ECF V++	Frozen dev. candidate set	S_1/S_2; $I, I_{\mathrm{aug}}, I_{\mathrm{JS}}, C, R$	Seeds 7301–7303; held-out protocol closed
MRDA	12 held-out dialogues	source / lexical / stable-hash; I, C, R	Derived local IDs; dialogue bootstrap
Entity relation	Fixed relation groups	admissible re-serializations; I, C, R	Post-freeze fixed summary

Artifact traceability. The held-out JSON identifies the evaluation scope as a one-time, inference-only serializer-portability audit and records 261 source conversations, of which 250 are eligible after the single-speaker exclusion used by the frozen evaluator. It also stores the frozen protocol and preflight hashes, checkpoint hashes for every row, and all row-level bootstrap intervals. Across the 72 transfer records, these hashes identify 20 unique checkpoint artifacts, matching the held-out protocol count in the main paper. Raw held-out text and per-example predictions are excluded from the anonymous artifact.

C Evaluation Protocols and Robustness Controls

C.1 Frozen evaluation protocols

Checkpoint, threshold, serializer, and evaluation subset are fixed before held-out evaluation.

Task-specific metric instantiation. For ECF, prediction instances are candidate pairs and confidence drift is the mean absolute change in candidate probability. For MRDA, prediction instances are utterance-level dialogue-act decisions and confidence is the frozen evaluator's aligned scalar score for the selected decision. For Entity Relation Classification, prediction instances are aligned relation-class decisions and confidence is the frozen evaluator's aligned scalar score for the selected classification. Decision consistency and unit-level risk use aligned predictions from the same frozen checkpoint in every setting.

C.2 Threshold robustness

Recomputing the same frozen held-out Pair probabilities over $\tau \in \{0.30, 0.40, 0.50, 0.60, 0.70\}$ preserves the diagnosis. Absolute-ID lookup has nonzero risk throughout (0.5210–0.7105); C/R retain exact consistency and zero risk. Confidence drift is threshold-independent.

D Post-held-out controlled repair comparison

Table S6 reports a separately frozen post-held-out ECF train/internal-validation plus development comparison. The held-out source remains closed. C wraps the matching I state, and training wall time is recorded as unobserved in the frozen artifact.

Table S6: V++ repair comparison, mean over seeds 7301–7303 and the two frozen nonidentity deployments. Exact decision consistency and zero drift for C/R are guaranteed under their stated implementation contracts; augmentation and JS are empirical repairs.

Family	Condition	S1 F1	Dec. Cons.	Risk	Drift	C-to-I count
Pair	I	0.322937	0.933911	0.652905	0.045388	513.7
Pair	C	0.322937	1.000000	0.000000	0.000000	0.0
Pair	R	0.371006	1.000000	0.000000	0.000000	0.0
Pair	I_serializer_aug	0.380786	0.971205	0.582569	0.022737	187.8
Pair	I_JS	0.342420	0.939878	0.654434	0.044518	450.8
Query	I	0.394195	0.967660	0.649847	0.022475	208.3
Query	C	0.394195	1.000000	0.000000	0.000000	0.0
Query	R	0.410360	1.000000	0.000000	0.000000	0.0
Query	I_serializer_aug	0.422560	0.979284	0.576453	0.018886	144.5
Query	I_JS	0.429903	0.978852	0.620795	0.016739	132.8

Table S7: Three-path preprocessing failure taxonomy for unchanged I checkpoints. Each path preserves text, utterance order, gold edges, candidate keys, and equality partitions for all eligible development conversations. The three paths instantiate controlled preprocessing-compatibility migrations.

Path	Pair dec. cons.	Query dec. cons.	Pair drift	Query drift	Pair C-to-I count	Query C-to-I count
A→B	0.937667	0.969389	0.044164	0.021664	1474	602
A→C	0.930156	0.965930	0.046611	0.023285	1608	648
B→C	0.927908	0.964769	0.048086	0.024788	1441	689

E SEMANTIC GATE NEGATIVE CONTROLS

Table S8 reports synthetic schema-fixture controls for the semantic gate. These checks operate on the transformation schema before model execution. Rejection marks an invalid invariance premise and produces no portability score.

Table S8: Semantic-gate negative controls. The gate permits portability measurement only for a local nominal bijection. All rejected transformations change or invent the semantic source interface and therefore receive no risk or prediction-consistency score.

Transformation	Admitted?	Gate outcome	Portability metric permitted?
Local nominal bijection	Yes	`LOCAL_NOMINAL_BIJECTION_ADMITTED`	Yes
Speaker merge	No	`SAME_SPEAKER_PARTITION_CHANGED`	No
Speaker split	No	`SAME_SPEAKER_PARTITION_CHANGED`	No
Role-label swap	No	`ROLE_SEMANTICS_NOT_NOMINAL`	No
Global-identity rename	No	`GLOBAL_IDENTITY_SEMANTICS_NOT_NOMINAL`	No
Fabricated speaker field	No	`NO_SERIALIZER_SOURCE_INTERFACE`	No