# EVALUATING EVIDENCE UNDER CHANGE: A CATEGORICAL SEMANTICS FOR DECISION-RELATIVE REUSE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Evaluation results often outlive the procedures that produced them, and when tasks, verifiers, environments or reporting rules change, what an old archive still warrants depends on the question asked of it. We give a finite categorical semantics in which evaluation settings and changes form a category, typed compatibility relations carry retained evidence to possible target records and each downstream use observes them through a decision profile. Tracking record multiplicity and use dependence separates exact reuse from decision-sufficient multiplicity, unresolved disagreement, empty compatibility and invalid comparison. Since changes compose, alternative routes can be compared through the profile sets they induce, and any discrepancy yields a finite witness of decision-relevant non-transport. For a fixed family of future observations we characterize the coarsest quotient preserving them and when compressed interfaces preserve propagation exactly. In exploratory case studies, a coarse CI archive still settles release but loses triage detail, eICU report contracts agree at current thresholds yet separate at every prespecified split-decile, and a corrected DeepSWE v1-to-v1.1 re-evaluation finds no qualifying same-state use-specific sufficiency witness. The semantics shows what historical evidence still supports for a declared use, where it falls short and which distinctions an archive must keep.

## 1 INTRODUCTION

Evaluation results routinely outlive the settings that gave them meaning. Once the harness, evaluator, software environment, reporting interface or deployment question changes, a retained archive may stay consistent with several complete target records that agree for one downstream use while disagreeing for another, and two independently justified routes to the new setting may leave different possibilities open. Reusing such evidence thus means asking *what the retained information still determines, for which use and under which declared change*.

We answer this categorically, taking evaluation settings as the objects of a finite category $\mathcal{C}$ and declared changes as its morphisms. Application-supplied compatibility relations extend to a functor $R : \mathcal{C} \to \mathbf{Rel}$, which existential direct image lifts to a functor $E : \mathcal{C} \to \mathbf{Set}$ propagating retained possibility sets, and each downstream use observes the propagated state through its decision profiles, so one construction handles sequential change, incomplete retention, use-relative sufficiency and alternative paths while keeping their assumptions apart. Exploratory case studies under contracts fixed before any target calculation test these distinctions on retained artifacts, namely a CI evidence request, two archive contracts on the public eICU Demo cohort (Pollard et al., 2018), a corrected analysis of a versioned software benchmark and two workstreams on dependency updates and on $\tau^3$ (Section 6).

Our contributions are as follows.

- A finite categorical semantics for evidence reuse, with a *reuse spectrum* of exact, decision-sufficient, diagnostic, empty and invalid grades (Section 3), in which Bayesian, psychometric and agentic evaluation methods become records, relations and rules (Section 6).
- A compositional account in which possibility sets compose although local grade labels need not, with a use-specific direct/staged coherence that yields a finite witness whenever parallel routes disagree (Sections 3.5 and 4).



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.

- Two retention results, a continuation quotient that keeps only what a frozen family of future observations can distinguish and a descent criterion for when chosen source and target interfaces preserve propagation exactly (Section 5).
- A probabilistic outer bound, by which every analysis charging only relation-compatible chains, shared latent parameters, path-dependent kernels and conditioning on target data included, predicts profiles within the relational profile set, which functorial refinements into subprobability kernels fill (Section 3.6).

The powerset, minimization, quotient and descent constructions and the relational translation of evidence each have close precedents (Section 2), and combined in one semantics for evaluation they yield use-relative grades, witnesses of route disagreement and retention criteria.

## 2 Related work and positioning

**Categorical and observational structure.** Categories and functors are standard tools for compositional data transformation, as in functorial data migration, and relational semantics handles nondeterministic structure (Spivak & Wisnesky, 2015; Chamoun & Mimram, 2026). Our quotient and interface constructions (Section 5) have close precedents in determinization and minimization, observational completeness and database view preservation (Silva et al., 2013; Colcombet & Petrişan, 2020; Amato & Scozzari, 2011; Fan et al., 2012; Benedikt et al., 2021).

**Evidence theory and comparison of experiments.** At the level of supports our propagation is the multivalued mapping of Dempster (1967) and the translation of evidence along paths of compatibility relations (Lowrance et al., 1986; Shafer, 1976; Kohlas & Monney, 1995), with $R_p[U]$ translating the categorical evidence $U$. We extend it in three directions, keeping empty and invalid outcomes as reportable grades where evidence theory normalizes empty images away as conflict (Lowrance et al. require every element to have a compatible partner), auditing parallel arrows, distinct unless a path equation identifies them, through coherence and finite witnesses where Dempster's rule would fuse their translations into one intersection, and designing retention for declared future uses (Proposition 3.1, Theorems 5.1 and 5.2). Our summary order $\preceq$ is the deterministic case of the garbling order of Blackwell (1951; 1953), but where Blackwell compares experiments across every decision problem, we ask, after the fact and without priors or losses, which profiles one realized archive state leaves possible for a declared family of uses, so the coarsest summary of Theorem 5.1 is relative to that family, as a minimal sufficient statistic is relative to its model.

**Decision-relative evaluation evidence.** Evaluation methodology and reporting frameworks stress intended uses, provenance and the distinctions that aggregate scores hide (Jacobs & Wallach, 2021; Wallach et al., 2025; Ghosh et al., 2026; Jiang et al., 2026), while agent evaluations call for records holding environments, tools, trajectories, resources and outcomes (Liu et al., 2024; Zhou et al., 2024; Mialon et al., 2024; Jimenez et al., 2024). Bayesian and psychometric methods quantify uncertainty within a fixed setting (Longjohn et al., 2025; Luettgau et al., 2025; Pilditch, 2026; Fonseca Rivera et al., 2026); we treat their posteriors as fields and rules of records and a cross-setting use of them that respects the declared relations as a compatible analysis (Section 3.6). Like recent work that values data and representations by the decisions they support (Bennouna et al., 2025; Ye et al., 2026), we judge retained evidence by the decisions it can still settle after an explicit change of setting, so our notion of *transport* (Section 3) is possibilistic and decision-relative. Causal transportability and dataset-shift correction rest on further assumptions, which may justify the supplied relations but which we do not try to infer (Pearl & Bareinboim, 2014; Bareinboim & Pearl, 2016; Quiñonero-Candela et al., 2009).

## 3 A finite semantics for evidence reuse

### 3.1 Settings, changes and compatibility relations

Let $\mathcal{C}$ be a finite category whose objects are evaluation settings and whose morphisms are declared changes. A convenient default is the free category on a finite acyclic directed graph, so that morphisms are paths (identities included); applications needing path equations may instead supply an explicitly

finite presented category, since a finite presentation alone does not guarantee finitely many morphisms. Each setting $M$ has a finite space $X_M$ of complete records, and each generating arrow $f : M \to N$ carries an application-supplied typed relation $R_f \subseteq X_M \times X_N$. When these assignments respect every imposed path equation, they extend to a functor $R : \mathcal{C} \to \mathbf{Rel}$ with

$$R_{g \circ f} = R_g \circ R_f, \qquad R_{\mathrm{id}_M} = \mathrm{id}_{X_M},$$

where $(x, z) \in R_g \circ R_f$ when some $y$ satisfies $(x, y) \in R_f$ and $(y, z) \in R_g$. For $R \subseteq X \times Y$ we write $R(x) = \{y : (x, y) \in R\}$ and $R[U] = \bigcup_{x \in U} R(x)$.

The structure earns its keep on parallel routes. A presentation can contain a staged route $M_0 \xrightarrow{f} M_1 \xrightarrow{g} M_2$ alongside a separately justified direct arrow $h : M_0 \to M_2$, kept as a distinct parallel morphism because an independently declared $R_h$ need not equal $R_g \circ R_f$; a proposed path equation that the relations violate must be repaired or dropped. As a running example, take a leaderboard that evaluates a system on a fixed prompt set, labelling several stochastic generations per prompt with an automated binary judge and summarizing the labels by Bayesian posteriors as in Longjohn et al. (2025) and Luettgau et al. (2025), whose single-setting analyses leave the changes to our illustration, the staged route replacing the judge and then extending the prompt set while the direct route justifies both in one joint contract. After either change, several complete target records may remain compatible with the archive even when all of them support the same aggregate decision.

### 3.2 Records, evidence and declared uses

A complete record contains every field that the in-scope compatibility relations and declared uses need, describing an item, a bounded trajectory or an evaluation batch with fields ranging from scores, uncertainty summaries and subgroup counts to versions, provenance and resource use. Posterior uncertainty about per-prompt success rates, for instance, lives inside a record as a function of its labels and declared priors, while $R_f$, $R_g$ and $R_h$ encode only compatibility across settings and place no probability on alternative target records unless a separate probabilistic contract supplies the weights (Section 3.6).

Each setting $M$ carries a finite family $D_M$ of declared uses, and for each $d \in D_M$ a finite family $\Pi_{M,d}$ of total deterministic rules $\pi : X_M \to A_{M,d}$ defines the profile map

$$q_{M,d} : X_M \to A_{M,d}^{\Pi_{M,d}}, \qquad q_{M,d}(x) = (\pi(x))_{\pi \in \Pi_{M,d}}.$$

The format is more expressive than it looks, since a record can store whatever statistical or operational object a use needs, encoding a performance threshold, group constraint, impact assessment, procurement disposition or resource budget (Appendix A).

### 3.3 Propagation, observation and transport

Archived source evidence is a nonempty possibility set $U_i \subseteq X_i$ containing every complete source record consistent with what was retained. A singleton means one record was retained in full; partial retention can leave several completions, with no implied distribution. Such sets propagate along paths by existential direct image, the standard functor

$$\mathcal{P}_\exists : \mathbf{Rel} \to \mathbf{Set}, \qquad \mathcal{P}_\exists(R)(U) = R[U],$$

which sends a record space to its powerset and preserves identities and relational composition. Composing it with the change interpretation gives $E = \mathcal{P}_\exists \circ R : \mathcal{C} \to \mathbf{Set}$, and a declared use $d \in D_M$ observes a propagated set through

$$O_{M,d} : \mathcal{P}(X_M) \to \mathcal{P}\left(A_{M,d}^{\Pi_{M,d}}\right), \qquad O_{M,d}(V) = q_{M,d}[V].$$

Given source evidence $U_i \subseteq X_i$, a path $p : M_i \to M_j$ and a target use $d \in D_j$, we compute the reachable records and the observed profiles

$$T_j(p; U_i) = E(p)(U_i) = R_p[U_i],$$
$$B_{j,d}(p; U_i) = O_{j,d}(T_j(p; U_i)) = q_{j,d}[T_j(p; U_i)],$$

writing $T_j(R; U_i) = R[U_i]$ and $B_{j,d}(R; U_i) = q_{j,d}[R[U_i]]$ for a bare relation $R$. We assume no naturality among the $O_{M,d}$, since uses and profile spaces need not be identified across settings (Figure 1).

$$\underbrace{M_0 \xrightarrow{f} M_1 \xrightarrow{g} M_2}_{\text{staged } g\circ f} \qquad \underbrace{M_0 \xrightarrow{h} M_2}_{\text{direct } h} \qquad U_0 \xrightarrow{E(p)} T_2(p;U_0) \xrightarrow{O_{2,d}} B_{2,d}(p;U_0).$$

Figure 1: A path propagates retained evidence to reachable records, and a declared use exposes the distinctions that matter for its decision.

Evidence *transports for use* $d$ when $B_{j,d}(p;U_i)$ is a nonempty singleton, and the same reachable set can transport for one use but not another, a pattern the worked example of Section 4 makes concrete.

### 3.4 Validity, grades and the reuse spectrum

A change contract names the source and target schemas and versions, the compatibility predicate and its justification, and a comparison is structurally valid when its types, relation membership and target-use rules are all evaluable. Missing retained fields leave several completions without invalidating anything, whereas an unspecified mapping or unevaluable use rule blocks any calculation needing it. Scientific legitimacy, whether statistical, causal, psychometric or about security, provenance and governance, is a separate matter for the application.

For a structurally valid comparison, abbreviate $T = T_j(p;U_i)$ and $B_d = B_{j,d}(p;U_i)$. The transport value

$$\tau_{j,d}(p;U_i) = \begin{cases} a, & B_d = \{a\}, \\ \bot_{\text{empty}}, & T = \varnothing, \\ \bot_{\text{disagreement}}, & |B_d| > 1, \end{cases}$$

together with $\bot_{\text{invalid}}$ for an invalid comparison, keeps ordinary decision profiles apart from empty compatibility, disagreement and invalid contracts. For reporting we grade each comparison in $G = \{\mathsf{exact}, \mathsf{sufficient}, \mathsf{diagnostic}, \mathsf{empty}, \mathsf{invalid}\}$; an evaluable one receives

$$g_{j,d}(p;U_i) = \begin{cases} \mathsf{exact}, & |T| = 1, \\ \mathsf{sufficient}, & |T| > 1,\ |B_d| = 1, \\ \mathsf{diagnostic}, & |B_d| > 1, \\ \mathsf{empty}, & T = \varnothing, \end{cases}$$

and invalid is reserved for an unevaluable contract. An exact grade (one reachable target record) or a sufficient one (several whose profiles agree) licenses reuse for that use as it stands, a diagnostic grade calls for a fresh run or a finer retained interface (Section 5), and an empty or invalid outcome is reported as such, never as a null finding. Collecting grades across target uses gives the *reuse spectrum*

$$\mathbf{g}_j(p;U_i) = \big(g_{j,d}(p;U_i)\big)_{d\in D_j},$$

which we never aggregate unless a separate stakeholder contract says how. It tracks two kinds of multiplicity: all uses share the reachable set $T$ for fixed $U_i$ and $R_p$, so exactness and emptiness agree across every evaluable use, while sufficient and diagnostic grades can still vary with $q_{j,d}$. An unevaluable use invalidates only its own coordinate, whereas an unevaluable shared relation invalidates all of them.

### 3.5 Sequential propagation and local grades

Full possibility sets propagate along composite paths without loss, since for composable $f$ and $g$

$$T_2(g\circ f;U_0) = E(g\circ f)(U_0) = E(g)(E(f)(U_0)) = R_g[R_f[U_0]],$$

even when the intermediate image is empty. Local grades lack this property: a grade records current multiplicities without saying which records produce them, and a later continuation may observe distinctions the grade already discarded. In a minimal example, let $X_0 = \{s\}$, $X_1 = \{a,b,c\}$, $X_2 = \{p,f\}$ and $U_0 = \{s\}$, with an intermediate use constant on $a,b,c$ and a target use separating $p$ from $f$, and define

$$R_{12} = \{(a,p),(c,p),(b,f)\}, \quad R_{01}^A = \{(s,a),(s,c)\}, \quad R_{01}^B = \{(s,a),(s,b)\}.$$

Both first steps reach two records sharing one profile and so receive grade sufficient, yet the common continuation yields

$$g_2(R_{12} \circ R_{01}^A; U_0) = \mathsf{exact}, \qquad g_2(R_{12} \circ R_{01}^B; U_0) = \mathsf{diagnostic},$$

where we suppress the use index because each target has a single relevant use.

**Proposition 3.1** (Local grades are not continuation-complete)**.** *For the fixed $U_0$ and $R_{12}$ above, no deterministic map $F : G \to G$ can recover the endpoint grade from the first-step grade over any class containing both $R_{01}^A$ and $R_{01}^B$.*

Any such $F$ would send the two equal first-step grades to one endpoint grade, contradicting the display above (Appendix C.1). Retaining the reachable sets $\{a, c\}$ and $\{a, b\}$ determines the endpoint correctly, so the loss comes from the grade alone, and a direct/staged audit must use full path semantics or a summary proved adequate for future observations, the coarsest of which Section 5 constructs.

### 3.6 Probabilistic refinements

Probability enters reuse either within a setting, where a posterior computed from a record is a field of it or a rule applied to it (Section 6), or across settings, where a probabilistic analysis weights compatible records. Let $\mathbf{SubStoch}$ be the category of finite sets and subprobability kernels, the Kleisli category of the finite subdistribution monad and a subprobabilistic variant of the kernel categories of Giry (1982) and Fritz (2020), in which mass lost by a kernel records that a source record has no compatible continuation. Sending a kernel $k$ to $\{(x, y) : k(y \mid x) > 0\}$ defines an identity-on-objects functor $\mathrm{Supp} : \mathbf{SubStoch} \to \mathbf{Rel}$, because a composite kernel is positive at $(x, z)$ exactly when some intermediate $y$ carries positive weight in both factors.

**Definition 3.2** (Probabilistic refinement)**.** *A* probabilistic refinement *of $(R, U_i)$ is a functor $\widetilde{R} : \mathcal{C} \to \mathbf{SubStoch}$ with $\mathrm{Supp} \circ \widetilde{R} = R$, together with a distribution $\mu$ on $X_i$ whose support is $U_i$, so that the weights cannot silently discard a retained possibility. For a path $p : M_i \to M_j$ and use $d \in D_j$, its predictive profile measure is $\lambda_d(p; \mu) = (q_{j,d})_*\big(\widetilde{R}(p)_*\mu\big)$.*

When $\mathcal{C}$ is free, any choice of generator kernels with the prescribed supports extends uniquely to a refinement, whereas a presented category's path equations constrain the weights along with their supports and may leave no probabilistic refinement at all, so the statements below concern refinements that exist. More generally, writing $p = f_k \circ \cdots \circ f_1$ as a path of generators, call a finite measure $P$ on chains $(x_0, \ldots, x_k)$ of records along it *compatible* with $(R, U_i)$ when it charges only chains with $x_0 \in U_i$ and $(x_{t-1}, x_t) \in R_{f_t}$ for every $t$, with predictive profile measure $\lambda_d^P$ its image under $(x_0, \ldots, x_k) \mapsto q_{j,d}(x_k)$. Compatibility involves no Markov or functorial structure, mixing over latent parameters shared across settings or conditioning and reweighting on target data preserves it, and a refinement induces the compatible measure $\mu(x_0) \prod_t \widetilde{R}(f_t)(x_t \mid x_{t-1})$, whose predictive profile measure is $\lambda_d(p; \mu)$.

**Proposition 3.3** (Support shadow)**.** *Every compatible measure $P$ satisfies $\operatorname{supp} \lambda_d^P \subseteq B_{j,d}(p; U_i)$, with equality for the uniform measure on compatible chains, so evidence transports for $d$ exactly when nonzero compatible measures exist and all of them concentrate on the same profile. Every probabilistic refinement satisfies $\operatorname{supp} \lambda_d(p; \mu) = B_{j,d}(p; U_i)$; hence $\lambda_d(p; \mu)$ is zero exactly when the grade is* empty*, and it is concentrated on a single profile, whatever the refinement, exactly when evidence transports for $d$. If $\mathcal{C}$ is free on a finite acyclic graph and $|B_{j,d}(p; U_i)| > 1$, then for every profile $a$ in that set and every $\varepsilon > 0$, the normalized measure gives $a$ probability above $1 - \varepsilon$ under some refinement and below $\varepsilon$ under another.*

The relational semantics is thus an outer bound on every compatible analysis, attained by every refinement, and since every grade, witness, quotient and descent statement here depends only on supports, all of them hold unchanged for each refinement, the weights deciding only how mass spreads over the profiles of a diagnostic coordinate (Appendix C.4 proves the proposition, and Appendix A.1 works it out for Beta-Binomial and item-response evaluations). In Dempster–Shafer terms the categorical evidence $U_i$ gives each profile on such a coordinate plausibility one and belief zero, and in the free acyclic case refinements push its predictive probability arbitrarily close to either value, so a posterior carried across a change reports the declared kernel as much as the archive. Equal predictive measures on parallel routes imply decision coherence (Definition 4.1) but not conversely, so a refinement with independently justified kernels can still disagree in the weights.

## 4 PARALLEL PATHS AND DIRECT/STAGED DECISION COHERENCE

Because a category can hold several morphisms between the same two settings, it can represent independently justified routes whose relational interpretations may disagree, as when $R_g \circ R_f$ and $R_h$ in the running example encode different assumptions. In general we let $p, q : M_i \to M_j$ be parallel paths, fix source evidence $U \subseteq X_i$ and take a target use $d \in D_j$.

**Definition 4.1** (Use-specific decision coherence)**.** *The parallel paths $p$ and $q$ are* decision-coherent for $(U, d)$ *when*

$$B_{j,d}(p; U) = B_{j,d}(q; U),$$

*and decision-incoherent otherwise.*

Coherence compares routes only after propagation and target-use observation (Figure 2), so it is weaker than equality of the underlying relations or reachable record sets while being stronger than equality of grade labels, which cannot establish it. Coherence alone is not conclusive either, since equal multi-profile sets are coherent but diagnostic and equal empty sets support nothing.

$$\text{direct: } U \xrightarrow{R_h[\cdot]} T_2(h; U) \xrightarrow{q_{2,d}[\cdot]} B_{2,d}(h; U),$$
$$\text{staged: } U \xrightarrow{R_f[\cdot]} T_1(f; U) \xrightarrow{R_g[\cdot]} T_2(g \circ f; U) \xrightarrow{q_{2,d}[\cdot]} B_{2,d}(g \circ f; U).$$

Figure 2: The direct/staged audit compares the target-use profile sets that each route reaches from the same archive.

### 4.1 A CONSTRUCTED MULTI-USE EXAMPLE

Let $X_0 = \{s\}, U = \{s\}, X_1 = \{a, b\}$ and $X_2 = \{u, v, w\}$, with

$$R_f = \{(s, a), (s, b)\}, \quad R_g = \{(a, u), (b, v)\}, \quad R_h = \{(s, u), (s, v), (s, w)\}.$$

Each target record is a complete evaluation batch of the kind described in Section 3, with $u$ a final reanalysis in which the performance and group rules both pass within budget, $v$ a compatible reanalysis with the same conclusions at excess cost and $w$ one in which aggregate performance still passes but the group rule fails. Table 1 records only these downstream decisions, preferring neither route statistically.

| Record | Performance | Group rule | Cost |
|---|---|---|---|
| $u$ | met | met | within |
| $v$ | met | met | exceeds |
| $w$ | met | unmet | within |

Table 1: Both routes reach $u$ and $v$; only the direct route reaches $w$.

The direct route reaches $\{u, v, w\}$ and the staged route $\{u, v\}$, and the three uses respond differently. Performance is decision-sufficient and coherent, observing $\{\text{met}\}$ on both routes, and cost is coherent but diagnostic with $\{\text{within}, \text{exceeds}\}$ on both, whereas the group use is diagnostic on the direct route and decision-sufficient on the staged one, with the profile $\text{unmet}$ (realized by $w$) witnessing the incoherence. Every compatible record thus supports the same aggregate decision while the group use still depends on which record remains possible, an ambiguity about cross-setting compatibility that no average over the three records resolves.

### 4.2 INCLUSION AND FINITE WITNESSES

**Lemma 4.2** (Direct/staged monotonicity)**.** *If parallel paths $p, q : M_i \to M_j$ satisfy $R_p \subseteq R_q$, as when $p = g \circ f$ and $q = h$ with $R_g \circ R_f \subseteq R_h$, then for every source evidence set $U$ and target use $d$,*

$$B_{j,d}(p; U) \subseteq B_{j,d}(q; U).$$

*Proof.* Relational inclusion gives $R_p[U] \subseteq R_q[U]$, and direct image under $q_{j,d}$ preserves inclusion. □

Whichever route is broader, a profile it determines is also determined by a nonempty image of the narrower one, while disagreement on the narrower route persists on the broader, so one change contract can safely bound another without any total ordering of reuse grades.

**Proposition 4.3** (Finite decision-incoherence witness)**.** *If the two observed sets differ, some profile $a$ belongs to their symmetric difference. On the route whose observed set contains $a$, a reachable record $y \in X_2$ satisfies $q_{2,d}(y) = a$, while no record reachable on the other route has profile $a$.*

*Proof.* Choose $a$ in either one-sided set difference. Membership in a direct image supplies $y$, and nonmembership on the other route excludes every reachable record with that profile. □

A discrepancy record stores the route pair, the use, the profile $a$, its witness $y$ and the predecessors of $y$, and checking it means showing that the other route reaches no record with profile $a$, which asks more than that it misses $y$. The observed sets and the witness then show whether the choice of route is immaterial for the declared use or whether some decision-relevant possibility survives under one change contract alone.

## 5 Evidence interfaces for future reuse

The audits so far run on full possibility sets, although an archive need not keep complete records when a smaller interface preserves every observation that a frozen family of declared future paths and uses will make. Two retention questions follow, which record distinctions such a family can observe, answered by the quotient below, and whether a chosen source interface keeps enough for a given change and target interface, which descent decides. *Complete* here refers only to the declared profile-set observations, which the quotient keeps with their emptiness, and not to target-record cardinality, so it may blur the exact/sufficient line.

### 5.1 Continuation-complete quotient

Let $X$ be a finite intermediate record space and $\Gamma$ a finite, prospectively declared family of evaluable continuation/use pairs, where each $\gamma \in \Gamma$ specifies a path relation $R_\gamma \subseteq X \times X_{t(\gamma)}$ and a target use $d_\gamma$. With $q_\gamma = q_{t(\gamma),d_\gamma}$, define

$$\Phi_\gamma(x) = q_\gamma[R_\gamma(x)], \qquad \sigma(x) = \big(\Phi_\gamma(x)\big)_{\gamma\in\Gamma},$$

keeping the empty set as a legitimate observation value, and call two records equivalent when no frozen observation tells them apart:

$$x \equiv_\Gamma x' \quad \Longleftrightarrow \quad \sigma(x) = \sigma(x').$$

Let $p : X \twoheadrightarrow Q = X/\!\equiv_\Gamma$ be the quotient map. A deterministic summary $s : X \to S$ is $\Gamma$-complete when every $\Phi_\gamma$ factors through its image corestriction $s : X \to \mathrm{im}(s)$; when comparing summaries, we discard unused codomain values and write $s_1 \preceq s_2$ if $s_1 = h \circ s_2$ for some map $h : \mathrm{im}(s_2) \to \mathrm{im}(s_1)$.

**Theorem 5.1** (Coarsest continuation-complete quotient)**.** *The quotient $p$ is $\Gamma$-complete. For every deterministic $\Gamma$-complete summary $s$, there is a unique map $h_s : \mathrm{im}(s) \to Q$ such that $p = h_s \circ s$. Consequently $p \preceq s$.*

Each future profile set for a retained possibility set $U \subseteq X$ is the union of the observations attached to the quotient classes that $U$ meets, so the quotient keeps every declared future observation (emptiness included) while forgetting only what $\Gamma$ cannot see, its optimality being relative to the frozen family and preorder above and silent on encoded size, acquisition cost or runtime. The construction is familiar from work on observational completeness, view preservation, contextual quotients and minimization (Amato & Scozzari, 2011; Fan et al., 2012; Benedikt et al., 2021; Jacobs & Wißmann, 2023), and Appendix C.2 gives the proof.

### 5.2 CROSS-SETTING DESCENT

Across a change, a source interface suffices for propagation to a target interface when every source record with the same retained value reaches the same set of retained target values. For surjective deterministic interfaces $p_i : X_i \twoheadrightarrow Q_i$ and $p_j : X_j \twoheadrightarrow Q_j$ and an evaluable relation $R \subseteq X_i \times X_j$, write $D_R(x) = p_j[R(x)]$; the relation *descends* when some $\overline{R} \subseteq Q_i \times Q_j$ satisfies $\overline{R}(p_i(x)) = D_R(x)$ for every $x$.

**Theorem 5.2** (Cross-setting interface descent)**.** *The relation $R$ descends from $p_i$ to $p_j$ exactly when*

$$p_i(x) = p_i(x') \implies p_j[R(x)] = p_j[R(x')]$$

*for all $x, x' \in X_i$. In that case $\overline{R}$ is unique and, for every $U_i \subseteq X_i$,*

$$p_j[T_j(R; U_i)] = \overline{R}[p_i[U_i]].$$

*Otherwise, two records collapsed by $p_i$ have a target-interface state reachable from exactly one of them.*

When descent fails, the witness names a distinction the source interface discarded but the change still exposes, pointing to something specific to retain. The criterion is pairwise, testing chosen interfaces against one relation at a time, and coherent preservation across a whole network of declared changes needs further assumptions (Appendix C.3 proves the criterion, and Appendix A reads it as a commuting square in $\mathbf{Rel}$).

## 6 STRUCTURED INSTANTIATIONS AND CASE STUDIES

Domain structure enters the semantics only through complete records, compatibility relations and declared uses, so statistical, interactive and governance-heavy evaluations share one reuse layer while keeping their own scientific assumptions.

**Statistical evaluation.** A statistical method fixes what a complete record holds and which deterministic rules act on it. In the Beta-Binomial model of Longjohn et al. (2025), say, a record holds $n$ judged generations for each of $K$ prompts, and the success counts $r_m(x)$, the Beta posteriors of the per-prompt rates $\theta_m$ and a summary such as the posterior mode of $W_{>\nu} = \sum_m \mathbf{1}[\theta_m > \nu]$ are all functions of $x$, the last being a rule $\pi : X_M \to A$. Once the model, priors, sampler and seed are declared, HiBayES (Luettgau et al., 2025), optimal stopping once a credible interval is narrower than $\delta$ (Pilditch, 2026) and item response theory (Fonseca Rivera et al., 2026) become rules on records in the same way (Appendix A.1). Retention and change enter only through $U$ and $R$, so an archive keeping the counts but not the generations has $U = \{x : r(x) = r^{\mathrm{obs}}\}$, which fixes every such rule in $M$, whereas after a judge revision $b \mapsto b'$ the relabelled counts are constrained only by what the change contract asserts.

**Interactive and agentic evaluation.** For a bounded interactive generator with finite trajectory set $\mathrm{Exec}_M$, $X_M$ contains the trajectories together with the environment, tool, resource and outcome fields the declared changes and uses require, as in a SWE-bench-style repair task whose record carries the task, repository state, trajectory, final patch and verifier receipts, with resolution rules $\pi_F(x) = \mathbf{1}[\text{all fail-to-pass tests pass}]$ and $\pi_P(x) = \mathbf{1}[\text{all pass-to-pass tests pass}]$ as in the software study below. Interactive structure affects reuse only through the record space, the admissible relations and the uses, so forgetting the generator leaves every reuse, quotient, descent and direct/staged calculation unchanged (Proposition A.1). Appendix B.1 lists the fields such records needed in our studies, and the case studies below ask whether these distinctions arise in retained evaluation artifacts. We apply the finite support/profile semantics to several bounded artifacts (Table 2), treating every study as exploratory because the CI receipts, the public eICU Demo cohort, the benchmark records and the public trace were known before the final analyses. Contracts written after that point but before the target-use calculations fixed the retained interfaces, uses, support semantics and reference calculations, with every failure, exclusion, incompatibility and setup stop kept in the accounting (full contracts and provenance in Appendix B).

| Study | Instantiation | Outcome |
|---|---|---|
| CI report *evidence request* | $X = \{0,1\}^{16}$; suite counts $e_i$, total $E$; $U = \{x : E(x) = 10\}$, $p =$ id; $\pi_A =$ HOLD iff $E > 0$, $\pi_i =$ GO iff $e_i = 0$. | $\lvert U\rvert = 8{,}008$ over 60 count vectors; $B_A = \{\text{HOLD}\}$ but $\lvert B_B\rvert = 6$; retaining $e_4 = 0$ makes $U$ a singleton. |
| eICU marginals *use witness* | $X = \mathbb{N}^{2\times 100}$ age-by-risk tables on 1,683 stays; $U$ fixes the risk histogram and age margins; uses: mean risk and $\mathbf{1}[\bar{r}_{65+} > \bar{r}]$. | Mean is a singleton, while the age comparison has $B = \{0,1\}$; the source age–risk aggregate yields $\{1\}$. |
| eICU routes *route obstruction* | Same $U$; direct $R_p$ keeps histogram and age margins, staged $R_q$ decile totals and age margins, so $R_p \subseteq R_q$; uses: flagged counts $C_t$ and $G_t$ (age 65+) at threshold $t$. | Coherent at the nine current deciles; $B(p) \subsetneq B(q)$ at all ten prespecified split-deciles in both contexts; 21 scalars preserve the 19-threshold support in 102 fields. |
| Software revision *corrected zero* | $R_f$: v1.1 re-verification of each archived v1 patch; one record per task, so $\lvert U\rvert = 1$; uses: all F2P pass, all P2P pass, reward $= 1$. | 104 of 109 pairs continue and five have $R_f(s) = \varnothing$; with $\lvert T\rvert \leq 1$ every use is exact, so no same-state witness exists despite 51 with $\pi_F \neq \pi_P$. |
| Setup stops *typed refusal* | Dependency pair changes two coordinates, so no declared arrow types it; $\tau^3$ tool-contract use needs absent tool definitions. | Both return invalid; $\tau^3$ retains 96 tasks/384 simulations but stops before scoring because the evaluator cache is missing. |

Table 2: Each case study instantiates the semantics with explicit records, evidence states, relations and use rules; Appendix B gives the full reductions.

A single semantics covers CI gating, clinical risk reporting, software re-verification and agentic benchmarks, and the audit settles what a rule keeping the fields each query reads cannot. The eICU contracts agree at every current decile, yet the decile report admits strictly more profiles for both uses at every split-decile, each with a finite witness, so agreement at current thresholds can mask loss at future ones, while 21 scalars keep exactly the profile sets the fine report supports, where a query-input rule would ask for the age-by-risk table behind $G_t$ that neither contract carries. The software archive held one patch record per task, so the 51 cases in which F2P and P2P disagree are different answers from fully determined records and by themselves say nothing about needing more retained data, which the same-state criterion confirms by reducing them to zero. A comparison that is not well typed, for want of tool definitions or a scorer cache or because a change touches more coordinates than any declared arrow, is itself a reportable outcome, never a skipped case or an inconclusive null.

## 7 Limitations

The theory is finite with deterministic use rules and offers no general tractability guarantee. Every conclusion depends on compatibility relations and use contracts the application must justify scientifically, with probabilistic analyses bounded by these relations when their chains respect them and causal transport needing further assumptions. Because the CI and eICU analyses rest on constructed report completions (with both eICU contexts sharing one cohort), the software correction is retrospective with zero qualifying stronger witnesses, and two workstreams stop before any target result, none of them establishes prevalence, independent replication, re-evaluation savings or prospective field validity. Although the mathematical ingredients have mature analogues, the displayed proofs still need independent checking.

## 8 Conclusion

Evaluation results outlive the settings that produced them, and we have tried to say what such a result still supports for a declared change and use. Most of the leverage lies in declaring what practice tends to leave implicit, and justifying those declarations, which our results presuppose but cannot supply, remains the harder part.

## AI USE STATEMENT

Generative AI tools were used to comprehensively search relevant literature and identify prior art, which informed research directions. They were also used to assist in empirical data analysis and to help with checking mathematical expositions and general writing. We take responsibility for the final content of this work.

## REFERENCES


Gianluca Amato and Francesca Scozzari. Observational completeness on abstract interpretation. *Fundamenta Informaticae*, 106(2–4):149–173, 2011. doi: 10.3233/FI-2011-381. URL https://doi.org/10.3233/FI-2011-381.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. doi: 10.1073/pnas.1510507113. URL https://doi.org/10.1073/pnas.1510507113.

Michael Benedikt, Pierre Bourhis, Louis Jachiet, and Efthymia Tsamoura. Balancing expressiveness and inexpressiveness in view design. *ACM Transactions on Database Systems*, 46(4):1–40, 2021. doi: 10.1145/3488370. URL https://doi.org/10.1145/3488370.

Omar Bennouna, Amine Bennouna, Saurabh Amin, and Asuman E. Ozdaglar. What data enables optimal decisions? an exact characterization for linear optimization. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2505.21692. URL https://openreview.net/forum?id=NXkUWAxSEH.

David Blackwell. Comparison of experiments. In Jerzy Neyman (ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 93–102. University of California Press, Berkeley, CA, 1951. doi: 10.1525/9780520411586-009. URL https://doi.org/10.1525/9780520411586-009.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

Yorgo Chamoun and Samuel Mimram. Realization of relational presheaves. In *Foundations of Software Science and Computation Structures*, volume 16503 of *Lecture Notes in Computer Science*, pp. 178–197. Springer, 2026. doi: 10.1007/978-3-032-22730-0_9. URL https://doi.org/10.1007/978-3-032-22730-0_9.

Thomas Colcombet and Daniela Petrişan. Automata minimization: A functorial approach. *Logical Methods in Computer Science*, 16(1:32):1–28, 2020. doi: 10.23638/LMCS-16(1:32)2020. URL https://doi.org/10.23638/LMCS-16(1:32)2020.

Arthur P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967. doi: 10.1214/aoms/1177698950. URL https://doi.org/10.1214/aoms/1177698950.

Wenfei Fan, Floris Geerts, and Lixiao Zheng. View determinacy for preserving selected information in data transformations. *Information Systems*, 37(1):1–12, 2012. doi: 10.1016/j.is.2011.09.001. URL https://doi.org/10.1016/j.is.2011.09.001.

Joshua Fonseca Rivera, Neil Shah, David Demitri Africa, and Konstantinos Voudouris. Item response theory for AI safety. *arXiv preprint arXiv:2608.05086*, 2026. doi: 10.48550/arXiv.2608.05086. URL https://arxiv.org/abs/2608.05086.

Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020. doi: 10.1016/j.aim.2020.107239.

Avijit Ghosh, Anka Reuel, Jenny Chim, Wm. Matthew Kennedy, Srishti Yadav, Jennifer Mickel, Yanan Long, Andrew Tran, Anastassia Kornilova, Damian Stachura, Kevin Klyman, Felix Friedrich, Jeba Sania, Jan Batzner, Anoop Mishra, Eliya Habba, Yixiong Hao, Nathan Heath, Shalaleh

Rismani, Usman Gohar, Andrea Loehr, David Manheim, Ruchira Dhar, Sree Harsha Nelaturu, Aarush Sinha, Leshem Choshen, Drishti Sharma, Ishan Khire, Amit Saha, Subramanyam Sahoo, Michael Hardy, Michael Alexander Riegler, Kabir Manghnani, Michelle Lin, Yanan Jiang, Yilin Huang, Asaf Yehudai, Jessica Ji, Aris Hofmann, Mubashara Akhtar, Max Lamparth, Nuno Moniz, Yacine Jernite, Stella Biderman, Zeerak Talat, Sanmi Koyejo, Mykel Kochenderfer, and Irene Solaiman. Evaluation cards: An interpretive layer for AI evaluation reporting. *arXiv preprint arXiv:2606.09809*, 2026. doi: 10.48550/arXiv.2606.09809. URL https://arxiv.org/abs/2606.09809.

Michèle Giry. A categorical approach to probability theory. In *Categorical Aspects of Topology and Analysis*, volume 915 of *Lecture Notes in Mathematics*, pp. 68–85. Springer, 1982. doi: 10.1007/BFb0092872. URL https://doi.org/10.1007/BFb0092872.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Jules Jacobs and Thorsten Wißmann. Fast coalgebraic bisimilarity minimization. *Proceedings of the ACM on Programming Languages*, 7(POPL):1514–1541, 2023. doi: 10.1145/3571245. URL https://doi.org/10.1145/3571245.

Han Jiang, Susu Zhang, Dongyao Zhu, Yuzhuo Bai, Sang T. Truong, Xiaoyuan Yi, Sanmi Koyejo, Xing Xie, and Ziang Xiao. AI evaluation should require standardized item-level data releases. *arXiv preprint arXiv:2604.03244*, 2026. doi: 10.48550/arXiv.2604.03244. URL https://arxiv.org/abs/2604.03244.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Jürg Kohlas and Paul-André Monney. *A Mathematical Theory of Hints: An Approach to the Dempster–Shafer Theory of Evidence*, volume 425 of *Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin, Heidelberg, 1995. doi: 10.1007/978-3-662-01674-9. URL https://doi.org/10.1007/978-3-662-01674-9.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ. arXiv:2308.03688.

Rachel Longjohn, Shang Wu, Saatvik Kher, Catarina Belém, and Padhraic Smyth. Bayesian evaluation of large language model behavior. *arXiv preprint arXiv:2511.10661*, 2025. doi: 10.48550/arXiv.2511.10661. URL https://arxiv.org/abs/2511.10661.

John D. Lowrance, Thomas D. Garvey, and Thomas M. Strat. A framework for evidential-reasoning systems. In *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI-86)*, pp. 896–903, 1986. URL https://cdn.aaai.org/AAAI/1986/AAAI86-149.pdf.

Lennart Luettgau, Harry Coppock, Magda Dubois, Christopher Summerfield, and Cozmin Ududec. HiBayES: A hierarchical bayesian modeling framework for AI evaluation statistics. *arXiv preprint arXiv:2505.05602*, 2025. doi: 10.48550/arXiv.2505.05602. URL https://arxiv.org/abs/2505.05602.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fibxvahvs3. arXiv:2311.12983.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014. doi: 10.1214/14-STS486. URL https://doi.org/10.1214/14-STS486.

Toby D. Pilditch. Knowing when to stop: Bayesian optimal stopping for LLM evaluations. *arXiv preprint arXiv:2608.14425*, 2026. doi: 10.48550/arXiv.2608.14425. URL https://arxiv.org/abs/2608.14425.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo Anthony Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018. doi: 10.1038/sdata.2018.178. URL https://doi.org/10.1038/sdata.2018.178.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (eds.). *Dataset Shift in Machine Learning*. Neural Information Processing. MIT Press, Cambridge, MA, 2009. ISBN 9780262170055. doi: 10.7551/mitpress/9780262170055.001.0001. URL https://mitpress.mit.edu/9780262170055/dataset-shift-in-machine-learning/.

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976. doi: 10.1515/9780691214696. URL https://doi.org/10.1515/9780691214696.

Alexandra Silva, Filippo Bonchi, Marcello Bonsangue, and Jan Rutten. Generalizing determinization from automata to coalgebras. *Logical Methods in Computer Science*, 9(1:9):1–27, 2013. doi: 10.2168/LMCS-9(1:9)2013. URL https://doi.org/10.2168/LMCS-9(1:9)2013.

David I. Spivak and Ryan Wisnesky. Relational foundations for functorial data migration. In *Proceedings of the 15th Symposium on Database Programming Languages*, pp. 21–28. Association for Computing Machinery, 2015. doi: 10.1145/2815072.2815075. URL https://doi.org/10.1145/2815072.2815075. Extended version available as arXiv:1212.5303.

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J. Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. Position: Evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 82232–82251. PMLR, 2025. URL https://proceedings.mlr.press/v267/wallach25a.html.

Yuhan Ye, Saurabh Amin, and Asuman Özdağlar. Learning decision-sufficient representations for linear optimization. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 6924–6975. PMLR, 2026. URL https://proceedings.mlr.press/v336/ye26a.html.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx. arXiv:2307.13854.

## A SUPPLEMENTARY CATEGORICAL DETAILS

By default the change category is the free category on a finite acyclic directed graph, whose morphisms are the finite paths in the graph, and assigning a finite record space to each vertex and a correctly typed relation to each generating arrow then extends uniquely, through relational composition and identity relations, to a functor $R : \mathcal{C} \to \mathbf{Rel}$. We ask for acyclicity because a finite graph with cycles need not generate a finite category, so an application that needs cycles or path equations must supply an explicitly finite category for its scope, and any interpretation of a presentation has to respect the presentation's equations. Two morphisms with the same endpoints are not thereby equal. In

the central audit the direct generator $h$ and the staged path $g \circ f$ stay distinct, and whether their interpretations coincide, or one contains the other, has to be checked on the relations themselves, since the presentation does not settle it.

Writing $\mathcal{P}(X)$ for the powerset of a finite set $X$, existential direct image $\mathcal{P}_\exists(R)(U) = R[U]$ is a functor $\mathbf{Rel} \to \mathbf{Set}$, since direct image preserves identity relations and relational composition, and the propagation functor of the main text is the composite $E = \mathcal{P}_\exists \circ R$. A target use $d$ observes an evidence state through its decision profiles, $O_{M,d}(V) = q_{M,d}[V]$, so that every audit quantity takes the form $O_{N,d}(E(p)(U))$, and as nothing identifies uses or profile spaces across settings, the observations at different settings need not assemble into a natural transformation. For parallel paths $p, q : M \to N$, use-specific decision coherence is the pointwise equality

$$O_{N,d}(E(p)(U)) = O_{N,d}(E(q)(U))$$

after propagation and target observation, which neither identifies $p$ and $q$ in $\mathcal{C}$ nor requires $R_p = R_q$, comparing only what the declared use can see from the fixed source evidence state, and finite enumeration of the two images makes the discrepancy certificate of Proposition 4.3 explicit without added complexity, the predecessors of its witness being a source record and, for a staged witness, an intermediate one, while invalid inputs and empty path images remain separate outcomes.

The two retention results have similarly direct categorical readings. The continuation signature is a product of set-valued observations whose kernel equivalence identifies the records that every frozen path/use observation treats alike, and the quotient by that equivalence has the universal factorization property of Theorem 5.1, with the factor map determined on its whole codomain once each summary is corestricted to the values it attains. For interface descent, reading functions as their graph relations turns the defining condition into the commuting square

$$p_j \circ R = \overline{R} \circ p_i$$

in $\mathbf{Rel}$, which says that propagating a record and then reading off its target interface agrees with reading off its source interface and then applying the descended relation; Theorem 5.2 characterizes when such an $\overline{R}$ exists. The rest of the manuscript is built from these finite constructions, although results on network-wide coherence or minimization would need further assumptions and do not follow from them.

**Operational uses.** Cost, group and impact questions become separate observation contracts, each with its own accounting boundary, population, construct and decision rule, and an evaluable cost rule needs a declared accounting boundary and denominator, with missing expenditure left unresolved, not imputed as zero, and expected cost until acceptance needing a retry model.

### A.1 Statistical evaluation methods as instantiations

Each method below analyzes a single evaluation setting in its own terms, and although we rewrite its records and decision rules in our notation, the retention choices and changes we attach to it are our own illustrations; none of the cited works speaks of possibility sets or compatibility relations. A well-specified posterior can still fall short for a later use if the retained counts, generations or provenance cannot separate compatible records.

**Beta-Binomial behavioral evaluation.** Longjohn et al. (2025) fix a prompt set $c_1, \dots, c_K$, a system with stochastic decoding, $n$ generations per prompt, a deterministic binary judge $b$ and independent priors $\theta_m \sim \mathrm{Beta}(\alpha_m, \beta_m)$ on the per-prompt probability of the judged behavior. A complete record $x \in X_M$ holds the generations $y_{m,i}$ with their labels, so that $r_m(x) = \sum_i b(y_{m,i})$ and $\theta_m \mid x \sim \mathrm{Beta}(\alpha_m + r_m(x), \beta_m + n - r_m(x))$. The aggregate $W_{>\nu} = \sum_m \mathbf{1}[\theta_m > \nu]$ has a Poisson-binomial posterior with success probabilities $1 - F_{\mathrm{Beta}}(\nu; \alpha_m + r_m(x), \beta_m + n - r_m(x))$, which makes its posterior mode an exact function of $x$, whereas their Monte Carlo interval for $K^{-1} \sum_m \theta_m$ becomes a function of $x$ only once the seed is fixed. Their contrast between greedy and stochastic decoding is a change of setting of the kind our semantics tracks. Under greedy decoding a record holds one output per prompt, and since, as they note, such an evaluation cannot reveal the per-prompt rates under stochastic decoding, a relation from the greedy to the stochastic setting leaves $B_{W_{>\nu}}$ multivalued unless further assumptions are made. Any probabilistic model of the stochastic records given the greedy one, a prior on the per-prompt rates for instance, is then a compatible

analysis when it charges only records the relation admits, so by Proposition 3.3 its predictive law for the posterior mode of $W_{>\nu}$ stays within $B_{W_{>\nu}}$, and it fills that set when its support is the whole relation, which makes it a refinement of this one-arrow change. The gap is sizeable in their case studies, where greedy decoding prefers one model on 41 of 80 prompts while the posterior mode of $W_{>0.75}$ at $n = 50$ is 23, and refuses 98 of 100 harmful prompts while the posterior mode of $W_{>0.95}$ is 94.

**Hierarchical GLMs.** HiBayES (Luettgau et al., 2025) fits Binomial and Beta-Binomial generalized linear models with a logit link to success counts aggregated by cell (in its GAIA case study, by model, task, reasoning effort and difficulty), and a record holds these cell counts. Once the model, priors, sampler and seed are fixed, the fitted posterior and its 95% HPDIs are functions of the record, and the practice of calling two models different when their HPDIs are disjoint, equivalent when one is nested in the other and inconclusive otherwise becomes a rule $\pi(x) \in \{\mathsf{different}, \mathsf{equivalent}, \mathsf{inconclusive}\}$. Its extension using earlier posteriors as priors for new evaluations carries a posterior across a change, so on a diagnostic coordinate the result reports the declared kernel as much as the archive, and the reuse spectrum flags where this happens. The paper also restricts two of its use cases to the first repeat of each evaluation, a step it does not frame this way but which is a retention choice of the kind our $U$ describes.

**Precision-based optimal stopping.** The optstop framework (Pilditch, 2026) addresses a question that comes before HiBayES's analysis, namely when data collection may stop, and does so by sampling epochs of items within a grouping. In its binary pathway a record holds the per-item counts $(s_i, n_i)$, a logit-normal hierarchical model yields a posterior for $\Theta = N^{-1} \sum_i \theta_i$, and the primary rule stops once the width $W(x)$ of its 97% credible interval satisfies $W(x) < \delta$; since the paper describes this pathway as deterministic and monotonic, with outputs reproducible for a fixed seed, $\pi_\delta(x) = \mathbf{1}[W(x) < \delta]$ is a rule in our sense. In its threshold study the TruthfulQA run configured with $\delta = 0.2$ stops after 63 trials with width 0.163, meeting $\delta = 0.2$ but not $\delta = 0.05$ (a separate run at $\delta = 0.05$ needed 238 trials). Both rules are exact on that run's retained counts, so the two thresholds differ in profile values while evidence sufficiency is never in question, and the corrected software zero of Section 6 rests on the same distinction. Sufficiency can come apart once retention is coarser. The ordinal pathway stops on the modal category while evaluators may report the mean, which the paper notes can still be computed from the stopped data, yet an archive that kept only the modal estimate would leave the mean multivalued. Setting the grouping prior mean from earlier results when re-evaluating the same model and benchmark, as the framework allows, is a different matter again, changing the declared rule while leaving the retained evidence untouched.

**Item response theory.** Fonseca Rivera et al. (2026) score one zero-temperature response per model and item under the two-parameter logistic model $P(x_j = 1) = \sigma(a_j(\theta - b_j))$, calibrating $(a_j, b_j)$ by regularized marginal maximum likelihood on a set of calibration models and then holding them fixed. A model's ability is the posterior mode $\hat{\theta}(x)$ under a standard-normal prior, which is a rule on the record $x \in \{0, 1\}^J$. A reduced test on items $K \subset J$ (25 static items, or about ten adaptive ones) retains $x_K$, giving $U = \{x : x_K = x_K^{\mathrm{obs}}\}$, and while the ability estimate from $x_K$ is exact on this interface, the full-benchmark score $|J|^{-1} \sum_j x_j$ ranges over an interval on $U$. The paper recovers that score by a spline regression on the reduced score and estimated ability, fitted to calibration models, and our semantics admits such a statistical prediction through a declared probabilistic contract, a distribution over the completions in $U$, whose predicted scores Proposition 3.3 then confines to that interval. Its caveat that a reduced test does not show the same items will remain informative for future models translates, in our terms, into the requirement that the relation for a new model be declared rather than inherited. The substitution audit in the paper has the shape of an interface refinement. A 40-item ability test separates 99.5% of the 8,911 model pairs in its 134-model cohort on at least one benchmark, while the 41 remaining ability-matched pairs still disagree on a median of 1,052 individual items (and no pair in the cohort on fewer than 284), so the item-level responses separate records that the ability summary does not.

### A.2 Interactive generators

**Proposition A.1** (Conservative structured instantiation)**.** *For a bounded interactive setting whose generator induces $X_M$, forgetting the generator while retaining $X_M$, the declared uses, source evidence*

*and typed compatibility relations leaves every current reuse, quotient, descent and direct/staged calculation unchanged. Conversely, a finite record-only setting can be represented by a one-step trivial interactive generator.*

*Proof.* Reachable sets, profile sets, transport values and grades, decision coherence and its witnesses, the continuation signature with its quotient and the descent condition are all defined from the category $\mathcal{C}$, the record spaces, the relations $R_f$, the evidence sets $U_i$ and the profile maps $q_{M,d}$. The generator enters only by inducing $X_M$, so once $X_M$ and these data are retained each of those quantities is the same function of the same arguments with or without it. Conversely, for a finite record space $X_M$, take the generator that moves in one step from a single initial state to any record of $X_M$ and then stops; identifying each one-step trajectory with its endpoint gives $\mathrm{Exec}_M = X_M$, so it induces the given record space and leaves the declared uses and relations as they were. □

# B Case-study materials and provenance

Each interpretation in Section 6 depends on keeping observed records, constructed compatible completions, target-side measurements and unsupported uses apart, so we give the publication-facing evidence behind that section here in more detail than the main text allows. Private identifiers, raw reports, source trees, detailed witness tables and restricted artifacts stay in the companion research package, which remains authoritative for them, while the manuscript repository carries only publication-safe aggregate derivatives.

## B.1 Shared empirical contract

In every study a retained interface is a deterministic projection of a finite source record space, each interface value standing for all frozen source records consistent with what was retained, and a target relation classifies each source completion as commensurable, typed noncommensurable, unresolved or structurally invalid. For each declared use the audit reports the set of target profiles reached from the entire supported fibre, never a single favorable completion, and keeps failed runs, unsupported subgroups, missing fields, timeouts and noncontinuations in the accounting. The evidential bar is equally conservative. A qualifying use-dependent evidence witness requires the *same* retained state and supported relation to yield a singleton profile set for one use and a multivalued set for another, so that different singleton profile values across uses do not count, and a source-side evidence request resolves an ambiguity only when a pre-target retained field refines the same support without silently discarding members. Claims about direct and staged routes likewise need two independently specified operational routes; one route defined to equal the other by construction does not qualify.

**Retention checklist.** The software and agentic studies below identify what an evaluation record needs to keep for their comparisons to be evaluable. On the source side it holds the task with its repository and base commit, the model patch or saved trace with its event stream and task linkage, the source reward with its basis, and the environment and provenance; on the target side it holds the patch-application status, the fail-to-pass and pass-to-pass outcomes, the official reward and the failure metadata. Some artifacts must never be null or missing, among them tool definitions and scorer or evaluator caches, whose absence stopped the $\tau^3$ study, and the pinned environment images or manifests and verifier version that fix a target contract. For test reports, per-suite counts kept beside the total are what settled the CI state.

## B.2 CI report-completion study

We treat two previously completed 16-test receipts as separate exploratory instances, each fixing one setting $M$ given by an artifact, configuration, image and test universe of 16 tests, partitioned into four suites $S_1, \dots, S_4$ with capacities $n_i = |S_i|$, $(n_1, n_2, n_3, n_4) = (3, 3, 4, 6)$, and no skipped tests. A complete record is a binary assignment $x \in X_M = \{0,1\}^{16}$, with $x_k = 1$ when test $k$ has a failure or error, and it determines the suite counts $e_i(x) = \sum_{k \in S_i} x_k$ and the total $E(x) = \sum_i e_i(x)$. The coarse report is the interface $I_0(x) = E(x)$, so the source evidence is the fibre $U = I_0^{-1}(E^{\mathrm{obs}})$, and since no change of setting is involved we have $p = \mathrm{id}_M$ and $T = U$. Two uses are declared, namely

the all-tests gate $d_A$, whose single rule is $\pi_A(x) = \text{HOLD}$ if $E(x) > 0$ and GO otherwise, and the triage vector $d_B$, with rules $\pi_i(x) = \text{GO}$ iff $e_i(x) = 0$ for $i = 1, \ldots, 4$.

The all-pass receipt has $E^{\text{obs}} = 0$, so $U = \{\mathbf{0}\}$ is already exact, the coarse report admitting one suite-count vector, one binary test assignment and one all-suite gate profile, and the total error count fixes both the all-tests gate and every suite gate. The updated receipt records ten tests with a failure or error, giving $E^{\text{obs}} = 10$ and $|U| = \binom{16}{10} = 8{,}008$, and under the fixed suite capacities these assignments realize 60 count vectors $(e_1, \ldots, e_4)$ and six distinct suite-level gate profiles, all of which share the all-tests action HOLD. Thus $B_{d_A} = \{\text{HOLD}\}$ while $|B_{d_B}| = 6$, and the same evidence state is graded sufficient for the gate but diagnostic for triage. The updated receipt has suite counts $e^{\text{obs}} = (3, 3, 4, 0)$, its ten failures filling the three smaller suites. We evaluated all $2^4 = 16$ subsets of the four suite-count fields as evidence requests, and the single field $e_4$ settles the state: $e_4 = 0$ forces the other three suites to capacity, so $U \cap \{x : e_4(x) = 0\}$ is a singleton with one vector, one assignment and one triage profile and both uses become exact, whereas retaining $e_1$ or $e_2$ alone leaves 1,716 assignments and four triage profiles and retaining $e_3$ alone leaves 924 assignments and five profiles. In general one retained suite count settles the state exactly when that suite is empty or full and the remaining failures either vanish or fill every other suite, a condition the updated receipt meets through $e_4$ and through no other single count, while for each of the 60 count vectors compatible with $E = 10$ at most two suite counts settle the triage profile.

A count-vector enumerator and an independently written binary-assignment enumerator cross-check this arithmetic, returning matching multiplicities and profile sets for both receipts. Only the assignment in the source report was actually observed, and every other one is a constructed completion of the retained aggregate, carrying no probability or frequency and not claimed to be executable for the fixed build. Nor are the old and updated receipts independent replications or a causal dependency comparison, and two setup-only attempts remain unsupported instead of being folded into test-outcome denominators.

### B.3 eICU REPORT-CONTRACT STUDIES

The public Demo source contains 2,520 patient-table IDs and 3,676 APACHE rows, 1,838 each for IV and IVa. The frozen pairing rule yields 1,838 IDs with one row from each version, of which 1,683 stays remain once valid predicted-risk values are required; of the 837 excluded IDs, 682 lack the required APACHE pair and 155 have invalid or missing prediction cells in both version rows. The retained cohort has 801 stays under age 65 and 882 aged 65 or older, the frozen parser leaving no age unknown. Both studies below work with fixed one-percentage-point bin representatives in place of raw predicted-risk values, and neither indexes nor uses actual mortality.

**Marginal audit.** Fix a model context $c \in \{\text{IV}, \text{IVa}\}$ and the cohort of $N = 1{,}683$ stays. A complete record is an age-by-risk table $y \in X_M = \mathbb{N}^{2\times 100}$, whose row $a = 1$ holds stays aged 65 or older and row $a = 0$ the rest, and whose column $b$ is the predicted-risk bin $[b/100, (b+1)/100)$. The coarse interface retains the 100-bin predicted-risk histogram $h_b = \sum_a y_{ab}$ and the broad age margins $m_a = \sum_b y_{ab}$, with $(m_0, m_1) = (801, 882)$, while omitting their joint table, and the declared completion set consists of every nonnegative integer table consistent with these marginals,

$$U = \Big\{ y : \textstyle\sum_a y_{ab} = h_b^{\text{obs}} \; \forall b,\; \sum_b y_{ab} = m_a \; \forall a \Big\},$$

with $p = \mathrm{id}_M$ as before. With bin representatives $r_b$, the two predeclared uses are the overall binned mean $\pi_{\text{mean}}(y) = N^{-1} \sum_b r_b h_b$ and the age-65-plus comparison $\pi_{65}(y) = \mathbf{1}\big[m_1^{-1} \sum_b r_b y_{1b} > \pi_{\text{mean}}(y)\big]$. The mean depends only on the retained column sums, so $B_{\text{mean}}$ is a singleton, whereas $B_{65} = \{0, 1\}$; retaining the source-derived age-by-risk aggregate $y^{\text{obs}}$ refines the same support to $\{y^{\text{obs}}\}$ and yields $B_{65} = \{1\}$ in both IV and IVa.

**Threshold-route audit.** The route study imposes two information contracts on the same pooled cohort, from a source setting $M_0$ to a common target setting $M_2$ whose uses are threshold queries, both with record space $\mathbb{N}^{2\times 100}$, and we write $U_0$ for the source state $U$ above. The direct contract $p : M_0 \to M_2$ transmits the fine histogram together with the age margins, from which the recipient reconstructs every consistent table, so its relation is the kernel relation of $(h, m)$,

$$R_p = \{(y, y') : h(y') = h(y),\; m(y') = m(y)\}.$$

Since $U_0$ is a single class of this equivalence, $T(p; U_0) = U_0$, which is also the image under the identity relation, so every profile set and check reported below is the same under either reading, the ambiguity about the joint table lying entirely in the source state; the kernel relation is the one the contract warrants, because from a more refined source the identity would pass on a joint table that the contract never transmits. The staged route $q = g \circ f$ passes through an intermediate setting $M_1$ whose records are the pairs $(D, m)$ of decile totals $D_k(y) = \sum_{b=10k}^{10k+9} \sum_a y_{ab}$ and age margins, with $R_f$ the graph of $y \mapsto (D(y), m(y))$ and $R_g$ relating each pair $(D, m)$ to every table with those decile totals and margins, the recipient again reconstructing every consistent table, so that

$$R_q = R_g \circ R_f = \{(y, y') : D(y') = D(y),\ m(y') = m(y)\},$$
$$T(q; U_0) = \{y' : D(y') = D(y^{\mathrm{obs}}),\ m(y') = (801, 882)\}.$$

At threshold $t$ (in percent) the uses are the total flagged count $C_t(y) = \sum_{b \geq t} \sum_a y_{ab}$ and the age-65-plus flagged count $G_t(y) = \sum_{b \geq t} y_{1b}$, and since decile totals are sums of histogram bins, $R_p \subseteq R_q$ and Lemma 4.2 gives $B_t(p; U_0) \subseteq B_t(q; U_0)$ for both uses. At the nine current decile thresholds $t \in \{10, 20, \ldots, 90\}$, each lying on a decile boundary, the two contracts agree for both uses and the routes are decision-coherent. The primary future family, however, consists of the ten split-decile thresholds $t \in \{5, 15, \ldots, 95\}$, and at every one of them, in IV as in IVa, the inclusion is strict for $C_t$ and for $G_t$, so any staged-compatible table whose profile lies outside the direct set is a finite witness in the sense of Proposition 4.3. Coherence at the boundaries and strictness for $C_t$ follow from the contracts, since at a decile boundary both routes fix $C_t$ and leave $G_t$ the integer interval $[\max(0, C_t - m_0), \min(m_1, C_t)]$, whereas inside a decile $k$ with total $D_k > 0$ the staged route leaves $C_t$ free over $D_k + 1$ values while the direct route fixes it; the checks below confirm the analytic bounds. We did not obtain the stronger differentiated-use result, since no primary threshold identifies total workload while leaving age-65-plus workload ambiguous in the staged state.

Each context has 29 threshold records, at each of which we check the analytic profile bounds three ways, once by an expanded-slot calculation of the total interval and twice by explicit subgroup enumeration, on the direct and on the staged route. All 87 checks per context agree with the formulas, and the source membership/refinement checks pass for all 38 declared threshold/use observations per context.

**Family-specific interface.** Let $\Gamma$ be the declared family of the 38 continuation/use pairs $(p, C_t)$ and $(p, G_t)$ with $t \in \{5, 10, \ldots, 95\}$, each continuing along the direct contract $p$. The summary keeping the 21 integer scalars $(C_5, C_{10}, \ldots, C_{95}, m_0, m_1)$ together with one source-multiplicity tag reconstructs every joint $\Gamma$-profile set, and with it the full joint profile support for all 19 declared thresholds, whereas the histogram and age margins occupy 102 integer fields. We claim semantic sufficiency for the declared family only, with no byte-cost measurement and no claim of global minimality, and off-family tests mark where that sufficiency ends, a five-point interface remaining non-singleton on all ten challenged thresholds in IV and on nine of ten in IVa.

### B.4 Corrected versioned software-evaluation study

The source setting $M_0$ is the historical DeepSWE v1 evaluation, whose complete records $s \in X_0$ carry the task with its repository and base commit, together with the model patch, source reward, environment and provenance. The target setting $M_1$ is one fully bound v1.1 verifier contract, whose records $t \in X_1$ carry the patch-application status, the fail-to-pass (F2P) and pass-to-pass (P2P) test outcomes, the official reward and the failure metadata. We interpret the declared change $f : M_0 \to M_1$ by the relation pairing each source record with the receipt of verifying its patch under the v1.1 contract, so that $R_f(s) = \varnothing$ exactly for the typed patch/base incompatibilities, and the three declared uses have one rule each, namely $\pi_F(t) = \mathbf{1}[\text{all F2P obligations pass}]$, $\pi_P(t) = \mathbf{1}[\text{all P2P obligations pass}]$ and $\pi_O(t) = \mathbf{1}[\text{official reward} = 1]$.

The corrected historical analysis covers 109 primary source/target pairs, of which five are typed patch/base incompatibilities while the other 104 continue under the bound target execution contract, splitting into 37 official-reward successes, 47 functional-only failures, four regression-only failures and 16 failures on both the functional and regression dimensions. The repeat subset contains 27 tasks whose repeated records comprise 23 continuations and four repeated typed incompatibilities, and every repeated outcome is stable under the contract.

The original analysis conflated different observed profile *values* with different evidence *requirements*. The 51 source cases with $\pi_F(t) \neq \pi_P(t)$ (47 functional-only and four regression-only failures, spread across 43 repository clusters) do show different profile values across uses, but under the field-realistic historical interface $I_{\text{hist}}$ the closed corpus holds one source record per task, so every evidence state $U = I_{\text{hist}}^{-1}(z)$ is a singleton and $|T_1(f; U)| \leq 1$. Every coordinate of the reuse spectrum is then exact, or empty for an incompatibility, and since no state can have $|B_d| = 1$ for one use and $|B_{d'}| > 1$ for another, the corrected count of states meeting the frozen same-state criterion is zero, well short of the frozen advancement threshold of four qualifying cases across two repositories; a separately implemented finite enumerator agrees with the corrected classification. The corrected package also records 312 source-side partition-refinement comparisons and 102 of 104 equal binary source-to-target official-reward pairs, but both are descriptive views of the same cohort and not independent observations, and neither shows that the corresponding target outcomes were retained in the historical source archive. We report the negative result because it rules out inferring different evidence-sufficiency requirements from cross-use outcome variation alone.

### B.5 Stopped studies

Both stopped studies end in the invalid outcome before any grade is computed over records.

**Dependency update.** The intended generating arrow $f : M \to N$ is a single declared dependency change, and the follow-up begins with an outcome-blind, deterministic selection of 48 previously uninspected metadata records spanning 36 projects. The first repeated dependency stream chosen by the frozen rule has matching inventories of 43 source/test blobs across the pair, yet its build descriptors differ in two direct dependency coordinates, the nominated dependency and an ancillary test dependency. Since the frozen design requires exactly one changed dependency, no declared arrow of $\mathcal{C}$ types such a pair, which is therefore ineligible before target execution; no relation $R_f$ is ever evaluated, and no alternative stream is substituted after the obstruction is observed. A resource gate independently blocks the intended image subset, because the relevant pinned manifests exceed the frozen 1-GB cap on compressed layers and even a three-configuration subset remains above it; no image is pulled and no build or outcome is inspected. What we obtain is a typed comparability and resource obstruction, which is neither a scientific null nor an estimate of how often reuse succeeds.

**$\tau^3$ execution feasibility.** Here $M_0$ is the pre-v1.0.1 banking release and $M_1$ is v1.0.1, and records are saved traces carrying their task linkage, reward basis, event stream and tool definitions. The audit declares a direct regrading arrow $h$ and a staged route $g \circ f$ through the legacy evaluator's saved results, and on $X_1$ a reward use $d_R$ and a tool-contract use $d_T$ with profiles in $\{\text{all_calls_valid}, \text{one_or_more_invalid}, \text{unverifiable}\}$. The public banking trace contains 97 tasks and 388 simulations, of which the frozen no-model-call filter keeps 96 tasks and 384 simulations, excluding before scoring one task with four NL-assertion simulations. Two separate obstructions then stop the study. First, the rule of $d_T$ needs the source tool definitions, but every record stores `tool_defs` as null, which leaves the tool-contract use unsupported under the frozen definition and its coordinate invalid. Second, the first staged step $R_f$ halts before computing any reward, since the first official reward route exits before processing any trajectory: the pinned scorer requires a precomputed dense-retrieval embedding cache that the clean offline environment lacks, and a cache miss would require an OpenAI embedding call. No credential is mounted, network access is disconnected and no paid or model API call is made. As the direct-current and staged-current reward regrades were never reached, no coordinate of either route was evaluated, and the publication-facing conclusion for this case is limited to artifact structure, filter accounting, missing scorer and cache provenance, and missing source-use evidence, with no reward profile, no use-sufficiency witness and no direct/staged comparison.

### B.6 Evidence class and interpretation

Every study result in this manuscript is exploratory, since the CI receipts and eICU cohort were already known, the corrected software analysis is explicitly retrospective and the stopped studies end before producing the target results that broader claims would need. The studies still bear on the mechanism, their contracts making positive, negative and refusal outcomes distinguishable and reproducible at finite scale. They do not establish prevalence or population transport, nor clinical

utility, deployment safety or operational adoption, and they stand in for neither independent replication nor the prospective multi-use field result that a complete claim of this kind would require.

# C Proofs of the main results

## C.1 Construction and proof for Proposition 3.1

It suffices to exhibit two first steps with the same grade whose composites with a common continuation have different endpoint grades. Let $X_0 = \{s\}$, $X_1 = \{a, b, c\}$, $X_2 = \{p, f\}$ and $U_0 = \{s\}$, give the intermediate use a constant profile, $q_1(a) = q_1(b) = q_1(c)$, and let the target use separate the two endpoints, $q_2(p) \neq q_2(f)$. The relations

$$R_{12} = \{(a, p), (c, p), (b, f)\},$$
$$R_{01}^{A} = \{(s, a), (s, c)\},$$
$$R_{01}^{B} = \{(s, a), (s, b)\}$$

are all structurally valid, and each first step reaches two intermediate records with a common profile, so both receive grade sufficient. Their composites with $R_{12}$, however, differ:

$$R_{12} \circ R_{01}^{A} = \{(s, p)\},$$
$$R_{12} \circ R_{01}^{B} = \{(s, p), (s, f)\}.$$

The first composite reaches only $p$ and so has endpoint grade exact, while the second reaches $p$ and $f$, which the target use separates, and has endpoint grade diagnostic. A deterministic map $F : G \to G$ recovering the endpoint grade from the first-step grade would therefore have to take both values at sufficient, which is impossible.

Since the continuation and the source archive are the same in both cases, the information the grade discards concerns the intermediate possibilities alone, and retaining the set $R_{01}[U_0]$ itself and propagating it relationally recovers the endpoint correctly.

## C.2 Proof of Theorem 5.1

We show first that $p$ is $\Gamma$-complete and then that $p$ factors uniquely through every $\Gamma$-complete summary. For each continuation/use observation $\gamma \in \Gamma$, let $Y_\gamma = A_{t(\gamma), d_\gamma}^{\Pi_{t(\gamma), d_\gamma}}$ and define $\overline{\Phi}_\gamma : Q \to \mathcal{P}(Y_\gamma)$ by $\overline{\Phi}_\gamma([x]) = \Phi_\gamma(x)$, which is well defined because equivalent records agree on every frozen observation. By construction $\Phi_\gamma = \overline{\Phi}_\gamma \circ p$, which is $\Gamma$-completeness, and the factorization is unique because $p$ is surjective.

Now let $s : X \to S$ be a deterministic $\Gamma$-complete summary, corestricted to its image, with factorizations $\Phi_\gamma = \widehat{\Phi}_\gamma \circ s$. If $s(x) = s(x')$, then for every $\gamma$

$$\begin{aligned}\Phi_\gamma(x) &= \widehat{\Phi}_\gamma(s(x)) \\ &= \widehat{\Phi}_\gamma(s(x')) = \Phi_\gamma(x'),\end{aligned}$$

so $\sigma(x) = \sigma(x')$ and $p(x) = p(x')$. The assignment

$$h_s : \mathrm{im}(s) \to Q, \qquad h_s(s(x)) = p(x)$$

is therefore well defined and satisfies $p = h_s \circ s$, and it is unique because every element of $\mathrm{im}(s)$ has the form $s(x)$; in particular $p \preceq s$. Working with the image spares us any choice of how to extend $h_s$ to unused summary values, and the argument goes through unchanged when $X$ is empty.

Finally, applying $\Phi_\gamma = \overline{\Phi}_\gamma \circ p$ pointwise shows that the quotient image $H(U) = p[U]$ of any $U \subseteq X$ reconstructs the declared observations:

$$\bigcup_{x \in U} \Phi_\gamma(x) = \bigcup_{z \in H(U)} \overline{\Phi}_\gamma(z).$$

These profile sets are preserved, emptiness included, although the quotient does not count how many complete records produce each profile. A source record reaching one target record and another

reaching two, say, can yield identical singleton observations when the target use does not separate the two target records, in which case the quotient need not distinguish the corresponding exact and sufficient grades. The factorization property proved here says nothing about minimal encoded size, runtime, storage, provenance overhead or practical compression.

### C.3 Proof of Theorem 5.2

We prove the two directions of the fibre criterion, then the propagation identity and finally the failure witness. If $R$ descends and $p_i(x) = p_i(x')$, then

$$p_j[R(x)] = \overline{R}(p_i(x)) = \overline{R}(p_i(x')) = p_j[R(x')].$$

Conversely, suppose that target-interface possibilities are constant on source fibres. Since $p_i$ is surjective, every $z \in Q_i$ has a preimage $x$, and we may set $\overline{R}(z) = p_j[R(x)]$, the choice of $x$ being immaterial by fibre constancy. The relation so defined has the required pointwise property, and it is the only one that does, since every source-interface state is attained.

Taking unions of $\overline{R}(p_i(x)) = p_j[R(x)]$ over $x \in U_i$ gives the propagation identity

$$\begin{aligned}\overline{R}[p_i[U_i]] &= \bigcup_{x \in U_i} p_j[R(x)] \\ &= p_j[R[U_i]] \\ &= p_j[T_j(R; U_i)].\end{aligned}$$

When constancy fails, some $x$ and $x'$ with the same source-interface value have distinct finite sets $p_j[R(x)]$ and $p_j[R(x')]$, and any element of their symmetric difference is a target-interface state reachable from exactly one of the two records, which is what the failure clause asserts.

### C.4 Proof of Proposition 3.3

For finite measures we have $\operatorname{supp}(f_*\nu) = f[\operatorname{supp}\nu]$ for any map $f$, since the mass at each point is a sum of nonnegative terms and is positive exactly when one of them is. Let $\mathrm{Ch}$ be the set of compatible chains along $p = f_k \circ \cdots \circ f_1$, those with $x_0 \in U_i$ and $(x_{t-1}, x_t) \in R_{f_t}$ for every $t$. By the definition of relational composition their final records form $(R_{f_k} \circ \cdots \circ R_{f_1})[U_i]$, which is $R_p[U_i]$ because $R$ is a functor (for the identity path, $k = 0$, both are $U_i$). A compatible measure $P$ has $\operatorname{supp} P \subseteq \mathrm{Ch}$, so writing $\ell(x_0, \ldots, x_k) = q_{j,d}(x_k)$,

$$\operatorname{supp}\lambda_d^P = \ell[\operatorname{supp} P] \subseteq \ell[\mathrm{Ch}] = q_{j,d}[R_p[U_i]] = B_{j,d}(p; U_i),$$

with equality whenever $\operatorname{supp} P = \mathrm{Ch}$, as for the uniform measure on $\mathrm{Ch}$. If $B_{j,d}(p; U_i) = \{a\}$, then $\mathrm{Ch}$ is nonempty, so that uniform measure is nonzero, and every nonzero compatible measure, whose predictive measure has the same total mass, concentrates on $a$; if the set is empty, so is $\mathrm{Ch}$, and every compatible measure is zero; and if it contains $a \neq b$, the point masses on two compatible chains ending at records with profiles $a$ and $b$ are compatible and concentrate on different profiles. This proves the first statement. For a refinement, the chain measure $P_\mu(x_0, \ldots, x_k) = \mu(x_0) \prod_{t=1}^k \widetilde{R}(f_t)(x_t \mid x_{t-1})$ is positive exactly on $\mathrm{Ch}$, because $\operatorname{supp}\mu = U_i$ and $\operatorname{Supp}\widetilde{R}(f_t) = R_{f_t}$, and its marginal on final records is $\widetilde{R}(p)_*\mu$, since composing kernels sums over the intermediate records and $\widetilde{R}$ is a functor. Hence $\lambda_d(p; \mu) = \lambda_d^{P_\mu}$ has support exactly $B_{j,d}(p; U_i)$, so it is zero exactly when $T_j(p; U_i) = \varnothing$ and concentrated on one profile exactly when $B_{j,d}(p; U_i)$ is a singleton.

For the extremal statement, write $p = f_k \circ \cdots \circ f_1$ with distinct generators, as $\mathcal{C}$ is free on an acyclic graph; for the identity path ($k = 0$) only $\mu$ is chosen, and the bound $(1-\eta)^{k+1}$ below still applies. Given $a \in B_{j,d}(p; U_i)$, choose $x_0 \in U_i$ and $x_1, \ldots, x_k$ with $(x_{t-1}, x_t) \in R_{f_t}$ and $q_{j,d}(x_k) = a$. For $\eta \in (0, 1)$, let $\mu = (1-\eta)\delta_{x_0} + \eta\,\mathrm{Unif}(U_i)$, let the row of $\widetilde{R}(f_t)$ at $x_{t-1}$ be $(1-\eta)\delta_{x_t} + \eta\,\mathrm{Unif}(R_{f_t}(x_{t-1}))$, and let every other row of every generator be uniform on its relational image, or zero when that image is empty. These choices have supports exactly $U_i$ and the $R_f$, the generator kernels extend uniquely to a functor on the free category, and the chain carries mass at least $(1-\eta)^{k+1}$ to $a$ out of a total mass of at most one, so the normalized probability of $a$ exceeds $1 - \varepsilon$ once $\eta$ is small. Running the same construction for some $b \neq a$ in the set instead yields a refinement under which the chain carries mass at least $(1-\eta)^{k+1}$ to $b$, so that $a$ receives at most $1 - (1-\eta)^{k+1}$ of the total mass and normalized probability at most $(1 - (1-\eta)^{k+1})/(1-\eta)^{k+1}$, which is below $\varepsilon$ for $\eta$ small enough.