RELEASE-STATE AUDITS FOR RETROSPECTIVE MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A dataset name or event-time cutoff does not uniquely specify the evidence available to a retrospective machine-learning evaluation. We introduce release-state audits: a closed reference specification linking exact source-state commitments, support accounting, estimands, and admissible claims. It separates support migration from retained-unit change and specifies typed non-results when identity, label, or missing-support requirements fail. The current v0.2 one-input checker validates only closed aggregate types and listed algebra over caller-supplied declarations and always withholds; it does not establish source replay or authorization. Three formal boundaries motivate the specification: a terminal event-time view need not identify earlier availability; finite agreeing label states do not establish permanence; and union-paired contrasts with release-exclusive support depend on registered missing-side restrictions. A separate closed synthetic adapter/evaluator fixture supports only finite implementation conformance, not real-world detection performance. We examine one exploratory pre-merge BIG-bench task-state case using three exact model revisions and a fixed census of 998 effective inputs. The complete four-cell panel exhibits descriptive grade differences between task states. Under no-appended-choice rendering, all three models have identical aggregate grades near the cell-specific uniform-choice expectation, without establishing identical predictions. This limited vignette illustrates why task state belongs to the evaluation object; it does not establish post-release benchmark drift, causal repair effects, or general model capability.

1 INTRODUCTION

A record can describe an event that precedes a forecast origin yet be absent from the source state available at that origin. Likewise, a benchmark can keep the same task name while its rendering rule, action space, or targets change. Filtering a current database by event date or naming a benchmark without its state can therefore reconstruct a different evidence object from the one that an earlier evaluation could have used.

This problem is distinct from ordinary temporal splitting. A cross-state score difference can mix three changes: the release-native population changes when records enter or exit; retained records or task items change while identity is stable; and endpoint labels evolve across later states. Moreover, an item observed in only one state has no ordinary counter-state loss. Treating that missing quantity as zero, or silently restricting to retained support, changes the claim rather than merely changing an estimator.

Real-time macroeconomic data make data vintages part of the historical information set (Croushore & Stark, 2001; 2003). Temporal benchmarks and evolving corpora address chronological distribution shift and changing knowledge (Yao et al., 2022; Huang et al., 2023; Jang et al., 2022; Zaporojets et al., 2022); longitudinal API data document changing outputs from externally supplied systems (Chen et al., 2022). We ask a narrower question: what must be bound and reported before a fixed retrospective ML claim can be compared across two named source states?

Our answer is a release-state audit: a reference contract linking exact state identity, population, estimand, and admissible claim language, plus an executable card-internal checker for closed aggregate types and registered algebra. The reference contract returns a bounded quantity or a named non-result;

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.

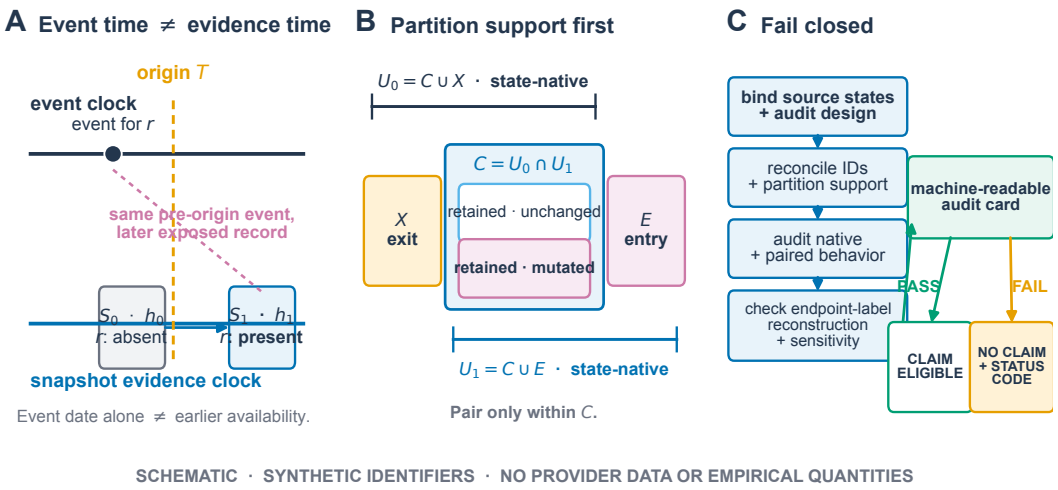


Figure 1: A release-state audit binds the evidence states, partitions eligible support into retained (C), exit (X), and entry (E) sets, assigns each quantity its population, and ends in an eligible claim or explicit non-result. This is the reference specification; the implemented one-input checker checks only declared aggregates and always withholds.

structural failure, quantity unavailability, claim eligibility, replay, and authorization remain separate layers.

We make three contributions.

1. **Formal claim boundaries.** We separate event time from evidence time; partition native support into retained, exit, and entry sets; decompose additive loss change; and characterize missing-support conclusions relative to explicit completion restrictions.
2. **A closed audit contract.** The card binds source, identity, eligibility, labels, models, metrics, seeds, and decision order. Failed gates cannot become numerical zeros or post-hoc subgroup claims.
3. **A bounded empirical instantiation.** A complete crossed panel evaluates three fixed model revisions on 998 effective inputs from two pre-merge BIG-bench task states. We report all four cells and all registered contrasts, including zero and mixed directions, separately from finite synthetic implementation-conformance evidence.

The observed-state grade differences in this case range from -1.50 to -1.10 percentage points. Under rendering without appended choices, all three models have the same aggregate grade, close to the cell-specific uniform-choice expectation. Equal means do not imply equal predictions, and neither the net differences nor their sign identify a causal repair effect. This modest evidence illustrates task-state dependence in one selected case; it does not validate the full replay specification or generalize to modern model capability. The planned production-snapshot application remains unexecuted and contributes no empirical results to this paper.

Figure 1 summarizes the chain. The paper does not introduce a new predictor, authenticate a provider from a hash alone, infer why a source changed, or interpret a revision as an error. At the card-internal level, its machine-checkable contribution couples caller-declared evidence-state commitments to a claim boundary; an explicit stronger replay specification governs source-facing claims.

2 Problem Setting

Two clocks and ordered states. Fix a retrospective task with forecast origin T. Let $[v_i^L, v_i^U]$ be the represented event-time interval for unit i, and let S_0, S_1 be ordered evidence states with availability times $a_0 < a_1$ and content identifiers h_0, h_1. Event time describes the represented event; evidence time names the exact source state that exposed its representation. A state difference alone does

not identify correction, backfill, error, or latent truth. An eligibility rule may, for example, require $v_i^U \leq T$, but the contract must freeze the endpoint convention and apply it separately to each evidence state.

Proposition 1 (event time does not recover evidence time). Suppose a terminal state reports a unit's represented event time but supplies neither an authenticated earlier state nor a trustworthy first-seen time. If delayed exposure is admissible, availability at a_0 is not identified, even when availability is monotone. Two histories can share the same terminal state: the unit is visible at a_0 in one and first appears in $(a_0, a_1]$ in the other. Appendix A gives the construction.

Release-native populations. Apply one frozen eligibility rule separately to the two states, obtaining U_0 and U_1. Define

$$C = U_0 \cap U_1, \qquad X = U_0 \setminus U_1, \qquad E = U_1 \setminus U_0. \tag{1}$$

The audit requires $|U_0| = |C| + |X|$, $|U_1| = |C| + |E|$, and $|U_0 \cup U_1| = |C| + |X| + |E|$. A failed identity invalidates downstream quantities.

Retained-unit and composition accounting. For model procedure m, let ℓ_{is}^m be an additive loss and let L_s^m be its state-native mean. Assume U_0, U_1, and C are nonempty, and write L_{Cs}^m for the common-support mean in state s. Then

$$\begin{aligned} L_1^m - L_0^m = (L_{C1}^m - L_{C0}^m) &+ \frac{1}{|U_1|}\sum_{i\in E}(\ell_{i1}^m - L_{C1}^m) \\ &- \frac{1}{|U_0|}\sum_{i\in X}(\ell_{i0}^m - L_{C0}^m). \end{aligned} \tag{2}$$

The first term is retained-unit change; the remaining terms are entry and exit composition. Equation 2 is descriptive and assigns no mechanism. Empty E or X contributes an empty sum, so no empty-set mean is fabricated. If C is empty, this common-support decomposition is unavailable; if either native population is empty, its native mean is unavailable.

Missing-support boundary. For $i \in E$, native state-0 loss is undefined; for $i \in X$, native state-1 loss is undefined. A union-population paired contrast therefore requires completion restrictions. Suppose $N = |U_0 \cup U_1| > 0$. If the registered differences $\delta_i^m = \ell_{i1}^m - \ell_{i0}^m$ lie in the side-uniform interval $[a_E, b_E]$ for entries and $[a_X, b_X]$ for exits, then

$$\left[\frac{\sum_{i\in C}\delta_i^m + |E|a_E + |X|a_X}{N}, \frac{\sum_{i\in C}\delta_i^m + |E|b_E + |X|b_X}{N}\right] \tag{3}$$

is sharp relative to the product-box restrictions. Without defensible completion semantics, the union target is unavailable rather than estimated from C. Appendix A states the item-specific version.

Accordingly, native, retained, and union-completion quantities enter the contract as distinct claim types, even when their realized values coincide.

All quantities above are descriptive functionals of the two named finite populations. Sampling uncertainty, model-seed variability, and superpopulation inference are separate targets and require separate registered procedures.

Proposition 2 (finite label witnesses do not prove permanence). Agreement across any finite, prelisted set of label states identifies agreement only on those states. A path that remains constant forever and one that changes immediately after the last witness are observationally equivalent on the registered states. A terminal release or justified absorbing rule is additional information, not a consequence of finite agreement.

3 A CLOSED RELEASE-STATE AUDIT CONTRACT

An audit card binds the ordered source identifiers, ordering basis, task and fold, stable-identity and eligibility rules, forecast origin, label state, metric orientation, model roster, tuning and seed policies, and complete population partition. The public card stores aggregates and commitments rather than raw identifiers, labels, predictions, credentials, or provider secrets.

Closed object and assurance. Let $\mathcal{B} = (\Sigma, \mathbf{z}, \mathbf{h}, D, R, W, \mathcal{K}, \alpha)$ denote a verification bundle: a recursively closed schema, an exact artifact roster and its ordered content commitments, a frozen design, an aggregate payload, finite label witnesses, an optional completion class, and an authorization declaration whose trust is not established by this object. The public card is an aggregate projection of $\mathcal{B}$. We distinguish $\mu = \mathsf{CARD_INTERNAL}$, which checks schema closure and reported cross-field arithmetic, from the stronger specification $\mu = \mathsf{BUNDLE_REPLAY}$, which also checks the complete supplied-byte roster and recomputes source-derived fields. The current v0.2 one-input aggregate checker represents either CARD_INTERNAL_ONLY or REPLAY_EVIDENCE_DECLARED_UNVERIFIED; both labels have the same declared-unverified claim ceiling and the sole terminal WITHHELD. It neither implements nor emits BUNDLE_REPLAY. The superseded v0.1 companion is retained only as an adversarial audit baseline and is not a replay, chronology, or authorization gate. Source-facing scientific types require a separate implementation that actually satisfies the replay obligations below. Neither the declared-unverified checker nor content hashing proves provider issuance, upstream completeness, or record truth.

Availability and typed claims. Every requested quantity receives exactly one tagged value $\mathsf{Value}(v)$, $\mathsf{Set}(I, \mathcal{K})$, or $\mathsf{Unavailable}(r)$; none is encoded by an unqualified null or zero. The closed claim grammar (arguments abbreviated) is

$$\begin{aligned} c ::= {} & \mathsf{Native}_\mu \mid \mathsf{Retained}_\mu \mid \mathsf{Decomp}_\mu \mid \mathsf{Union}_\mu(\mathcal{K}) \\ & \mid \mathsf{Witness}_\mu(Q) \mid \mathsf{Order}_\mu \mid \mathsf{NonResult}_\mu(r). \end{aligned} \tag{4}$$

At replay level these constructors are, respectively, descriptive on (U_0, U_1), descriptive on C, accounting on (C, E, X), set-identified on the union relative to $\mathcal{K}$, agreement only on named releases, point-descriptive on C, and a reason-qualified non-result. There is no rule that casts them to historical availability, causal effect, permanent truth, provider intent, or model superiority.

Ordered reference semantics. The only successful structural trace in the reference semantics is a prefix of

$$\mathsf{SCHEMA} \prec \mathsf{BIND} \prec \mathsf{PARTITION} \prec \mathsf{AVAILABILITY} \prec \mathsf{RECOMPUTE} \prec \mathsf{TYPE} \prec \mathsf{COMMIT}. \tag{5}$$

Unknown fields or a failed structural guard terminate in $\mathsf{REJECT}(r)$ and introduce no later premise. Quantity absence instead produces a typed $\mathsf{Unavailable}(r)$. After binding sources, code, and the frozen task, the replay branch derives U_0, U_1, C, E, X, rejects duplicate identities, checks all conservation equations, and recomputes every eligible aggregate and assumption-indexed interval. The card-internal branch checks only closed-card and reported-field identities. Both branches assign one availability tag per quantity, validate its guards, derive exactly the claims licensed by Equation 4, and atomically commit the trace and claim set. Publication is outside the one-input aggregate checker. In the normative replay specification, a distinct wrapper may emit PUBLISHABLE only after opening an external authorization envelope, authenticating its issuer under a named trust policy, and verifying that its bytes bind the complete semantic-payload digest and declared purpose. A caller-supplied authorization label or digest is insufficient and cannot strengthen the scientific type.

> **ReferenceSemantics($\mathcal{B}, \mu, \alpha$).** Closed-decode the card; bind source/design/code commitments; in replay mode verify every artifact and derive the exact partition (internally, check only reported-count identities); assign every requested quantity one availability tag; recompute permitted quantities and witnesses; derive claims by the closed typing rules; then atomically commit WITHHELD. Only a distinct, externally authenticated authorization wrapper outside this one-input checker may issue a purpose-specific PUBLISHABLE decision. Any failed structural guard returns $\mathsf{REJECT}(r)$ before later premises are introduced.

Theorem 1 (conditional type soundness of the reference semantics). Assume deterministic closed decoding and canonical serialization; bound deterministic identity, eligibility, metric, witness, and completion rules; in replay mode, opening every rostered artifact, verifying its bytes, digest, and size with the running schema and validator bound, and deriving every source-facing field rather than trusting a supplied replay receipt; a versioned, design-bound numerical policy with no claim-local tolerance; and correct execution of the reference semantics. Every committed claim then has a derivation in the closed grammar at its recorded assurance level. In replay mode, its premises recompute under its exact source, task, population, and assumption commitments; internal mode

guarantees only closed-card typing and cross-field consistency. In particular, a union interval is sharp only relative to its named product box, finite witness agreement is confined to the named releases, and any later publishable decision additionally requires the external authenticated-envelope check above. This is a specification-level software-safety statement, not a refinement theorem for the current aggregate checker, provider authentication, causal identification, permanence, external validity, or proof that a production implementation refines the reference semantics. Appendix B gives the constructor obligations, state machine, pseudocode, and proof.

Closed synthetic software QA. A separate closed synthetic adapter/evaluator fixture, not the claim-receipt companion, contains 54 author-specified designs (108 folded cases). On that finite fixture, the card-internal adapter matched a separately encoded evaluator on the registered schema, type, and algebra decisions; its unkeyed content commitments test only internal consistency, not authenticity. The v0.2 one-input claim-receipt checker has a separate 24-test synthetic regression suite and always withholds. Neither fixture establishes source replay, historical ordering, authorization, source-derived estimands, or real-world detection performance.

4 AN EXPLORATORY PRE-MERGE TASK-STATE CASE

We instantiate the stable-support part of the audit on one public BIG-bench task (BIG-bench authors, 2023), not a sample of benchmark revisions. Two immutable feature-branch states of `salient_translation_error_detection` differ in a rendering configuration and four target/action maps. The change from the nonempty string `"false"` to Boolean `false` removes appended choices under the unchanged runner. The repaired state precedes pull-request merge; these are not two formal benchmark releases.[1]

Design and estimands. The locally frozen 2×2 panel crosses rendering q_0 (append choices) and q_1 (do not append) with target/action states y_0, y_1. Only q_0y_0 and q_1y_1 are observed feature-branch states; the off-diagonal cells are diagnostic constructions, not historical releases or causal interventions. Three exact pre-case model revisions—DistilGPT2 (Hugging Face, 2021), T5-small, and T5-base (Raffel et al., 2020)—are evaluated on the same 998 effective input identities. Conditional target-token log probabilities are summed without length normalization, followed by the official deterministic tie rule and the selected target's raw multiple-choice grade. For each model, $g(q, y) = 998^{-1} \sum_i G_i(q, y)$. We report every cell and all six frozen contrasts: $\Delta_{\text{obs}} = g(1, 1) - g(0, 0)$, $\Delta_q(y) = g(1, y) - g(0, y)$ for both y, $\Delta_y(q) = g(q, 1) - g(q, 0)$ for both q, and $\Delta_{q\times y} = g(1, 1) - g(1, 0) - g(0, 1) + g(0, 0)$. These are fixed-census descriptions, not sampling estimates; no confidence interval, significance test, or equivalence claim is introduced.

Complete-panel results. Table 1 reports the complete three-model panel. Observed-state grade differences are -1.5030, -1.1022, and -1.3026 percentage points for DistilGPT2, T5-small, and T5-base, respectively. Both conditional rendering contrasts are negative for every model. Target/action contrasts at q_0 are positive, negative, and zero in that model order; at q_1 they are all zero. The interaction consequently has negative, positive, and zero signs. The fixed-panel macro-average does not replace these heterogeneous rows.

All three models attain the identical aggregate grade 16.8337% in both q_1 cells, only about 0.177 and 0.167 percentage points above the corresponding uniform-choice expected grades. This reference is an expectation under uniform selection, not a lower bound on model performance or an empirically run random model. Equal aggregate grades do not establish equal predictions, a constant label preference, or an implementation error. The panel provides limited evidence of state sensitivity in a near-reference scoring regime, not a demonstration of general task capability.

Diagnostics and interpretation. Between the two observed states, selected-choice-index disagreement is $998/998$, $898/998$, and $993/998$; grade changes occur on $351/998$, $319/998$, and $347/998$ items, respectively. Appendix D reports every count, denominator, rate, uniform-choice reference,

[1] Immutable repair commit: `https://github.com/google/BIG-bench/commit/70e9f8702c12f77b12100326c0a52adcf80d1414`; review timeline: `https://github.com/google/BIG-bench/pull/374`.

Table 1: Complete fixed BIG-bench panel: 998 items in every model–cell. Panel A reports $100g(q, y)$; panel B reports contrasts in percentage points. q_0 appends choices and q_1 does not; y indexes target/action state. Uniform choice is an expected-grade reference, not a performance lower bound. No sampling intervals or causal interpretation are implied.

A. Mean grades on the complete fixed roster

Model	q_0y_0	q_1y_0	q_0y_1	q_1y_1
DistilGPT2	18.3367	16.8337	18.4369	16.8337
T5-small	17.9359	16.8337	17.0341	16.8337
T5-base	18.1363	16.8337	18.1363	16.8337
Fixed-panel mean	18.1363	16.8337	17.8691	16.8337
Uniform-choice reference	16.6571	16.6571	16.6667	16.6667

B. Every registered contrast; no model selection

Model	Δ_{obs}	$\Delta_q(y_0)$	$\Delta_q(y_1)$	$\Delta_y(q_0)$	$\Delta_y(q_1)$	$\Delta_{q\times y}$
DistilGPT2	-1.5030	-1.5030	-1.6032	+0.1002	0.0000	-0.1002
T5-small	-1.1022	-1.1022	-0.2004	-0.9018	0.0000	+0.9018
T5-base	-1.3026	-1.3026	-1.3026	0.0000	0.0000	0.0000
Fixed-panel mean	-1.3026	-1.3026	-1.0354	-0.2672	0.0000	+0.2672

Both task files are feature-branch states; the repair precedes merge. These results do not establish post-release drift, leaderboard impact, or general model ranking. The fixed-panel mean weights the three named models equally.

and fixed-panel macro-average. Indices are runner-relative positions, not aligned semantic labels across changed action spaces: near-total index disagreement is not evidence of near-total semantic answer change. Small net mean differences can coexist with many item-level grade changes; the aggregate does not identify their direction split.

The y factor is not a pure relabelling operation. Under q_0, the official renderer advances one sequential random stream; the four choice-cardinality changes can alter later appended-choice permutations. This registered propagation is part of the y contrast and interaction, not a defect to remove after observing results. We therefore do not interpret T5-small's q_0 target/action contrast as a direct effect confined to four edited items.

The complete output underwent independent aggregate recomputation. Two implementation corrections were made after automated processing but before human or agent inspection of scientific values; the original failures were retained and no model was rerun (Appendix D.3). Exact recomputation establishes agreement on stored outputs and frozen calculations, not independent proof that every model invocation is correct. This exploratory pre-merge vignette shows why exact task state belongs to the evaluation object; it does not establish causal repair effects, post-release drift, leaderboard impact, or model superiority. The proposed DBLP production-snapshot study remains unexecuted future scope and contributes no empirical evidence here.

5 Related work

Versioned data and provenance. Real-time vintages distinguish revised histories from the information set available to an earlier forecaster (Croushore & Stark, 2001; 2003). Dynamic-data citation and versioning bind persistent identifiers, queries, and changing objects (Klump et al., 2021; Rauber et al., 2021); standardized dataset-version labels additionally encode schema, statistical drift, and timestamps (González-Cebrián et al., 2024). ML systems also use data versioning, time travel, and captured pipeline provenance for reproducibility and debugging (Ormenisan et al., 2020). Datasheets, Data Cards, and benchmark-repository studies document provenance, intended use, and governance (Gebru et al., 2021; Pushkarna et al., 2022; Longjohn et al., 2024), while a systematic review documents practical gaps in dataset provenance and version control (Wu et al., 2024). Auditable bibliometric workflows likewise preserve queries, parameters, logs, and manifests because databases evolve (El Majjaoui et al., 2026).

Table 2: Coverage among selected closest lines of work. • denotes direct treatment, ○ an adjacent ingredient, and – no direct treatment evidenced in the inspected source. This is a positioning aid, not an exhaustive priority claim.

Line of work	Version	Support	Estimand	Claim gate	Receipt
Leakage / learn–predict separation	–	–	○	○	–
Partial identification with missing data	–	○	•	–	–
ML time travel and pipeline provenance	•	–	–	–	○
Prospective temporal/infrastructure shift	○	–	•	–	–
Benchmark lifecycle and quality assurance	○	–	○	○	○
Dataset-management/version-control review	•	–	–	–	–
Release-State Audits (this contract)	•	•	•	•	•

Temporal and evolving evaluation. Leakage analysis requires an explicit modeling goal and separates information available for learning from information legitimately available at prediction time (Kaufman et al., 2011). Prospective validation further distinguishes temporal changes in populations from changes in data-access infrastructure (Otles et al., 2021). Temporal benchmarks address chronological shift, graph events, evolving entities, and changing knowledge snapshots (Yao et al., 2022; Huang et al., 2023; Jang et al., 2022; Zaporojets et al., 2022). Adjacent work measures knowledge-graph evolution and regenerates evaluation corpora when retrieval sources update (Bobed et al., 2020; Chernogorskii et al., 2026); longitudinal API datasets expose system changes (Chen et al., 2022). New CIFAR-10 and ImageNet test-set work also measures accuracy across carefully reconstructed data and intermediate dataset revisions (Recht et al., 2019). Our target is different from a temporal split or benchmark replacement: it asks whether the same nominal retrospective task is supported by the same evidence and task objects across named releases.

Support mismatch and bounded claims. Distribution-shift diagnostics already separate performance contributions from common and newly encountered regions (Cai et al., 2026). Our common support is an exact stable-identity intersection, not merely an overlap region in feature space. More generally, partial identification makes explicit what observed evidence identifies and which assumptions narrow a region when values are missing (Manski, 2005). A release-exclusive unit lacks an ordinary native counter-state loss; therefore a union-paired target requires explicit completion semantics. Benchmark-lifecycle work supplies quality-assurance and maintenance criteria (Reuel et al., 2024), but a checklist is not an evidence-to-claim type system.

Table 2 locates the proposed contract among these selected neighbors. The marks summarize only the inspected sources: a dash is not a claim that no implementation exists elsewhere. We do not introduce versioning, provenance, leakage control, partial identification, temporal evaluation, or benchmark quality assurance in isolation. We contribute their replay-level coupling in a reference specification: exact release identities determine conserved support; that partition types the estimand or bound; and validation maps evidence to a closed claim state. The present executable implements only the declared-unverified aggregate subset: it validates closed types and listed algebra over caller-supplied fields and emits a withheld receipt or registered non-result.

6 Limitations and data availability

Scope of the evidence. The formal boundaries organize what a cross-state comparison can mean; they are not a new predictor or a general theorem of data quality. The synthetic fixture covers author-defined software behavior, not detection under unknown mechanisms. The BIG-bench study is one selected pre-merge task-state pair with three small, older model revisions and a fixed census, not a random sample of tasks, releases, or models. Its near-uniform-reference grades weaken its value as an ability experiment. No general performance ranking, prevalence estimate, contamination exclusion, post-release drift, or leaderboard impact follows. The production-snapshot application planned for DBLP remains unexecuted.

Interpretation and implementation. The crossed cells are diagnostic constructions, not randomized treatment assignments. Target/action changes interact with the registered sequential renderer RNG, so a conditional contrast is not a pure label-correction effect. Index disagreement concerns positions, not aligned semantic answers. Equal aggregate grades do not establish equal predictions,

constant-choice behavior, or an implementation defect. The available aggregate cannot distinguish these explanations. Separately implemented aggregation checks arithmetic on stored outputs; it does not independently certify the model invocation or shared scoring code. No post hoc raw-output diagnostic or model rerun was used to strengthen the interpretations reported here. Frozen output can depend on tokenizers, weights, numerical libraries, and hardware.

Assurance boundary. Content hashes detect byte changes but do not authenticate provider issuance or establish upstream completeness. A self-authored replay or authorization digest proves neither replay nor permission. The one-input checker always withholds; the stronger reference semantics is not a production replay system. Record or task revision is observable state, not latent truth or evidence of misconduct. Selection of this narrow empirical scope after aggregate access is not confirmatory evidence or external preregistration.

Data and code availability. The accompanying anonymous supplement supplies the reported scientific aggregate, deterministic table-generation code, the aggregate checker, synthetic conformance fixtures, and bounded reproduction instructions. The ZIP's `paper/tables/bigbench_scientific_aggregate.json` supplies the numerical table data; `paper/tools/render_bigbench_tables.py` regenerates their displayed form, and `paper/artifacts/public_source_selectors.json` lists the upstream assets. The supplement distinguishes reproducing tables from re-running model inference and from independently recomputing the original stored outputs. This aggregate-only release does not redistribute benchmark task text, checkpoints, individual model outputs, or bibliographic records. Exact upstream states and model revisions are specified in the appendix; upstream terms govern access to those assets. Restricted Dimensions data and derivatives are outside this paper and its supplement. No separate public repository or persistent deposit is claimed.

7 Conclusion

Release-state audits couple named evidence states, conserved support, estimands, and admissible claims. The formal boundaries separate historical availability from event time, finite label agreement from permanence, and observed native quantities from assumption-dependent union completions. The current executable checks only declared aggregate types and listed algebra; it always withholds and does not implement bundle replay. A complete, limited BIG-bench vignette shows descriptive task-state sensitivity alongside near-uniform-reference grades, mixed conditional contrasts, and substantial grade changes that a net mean alone would conceal. The practical implication is to state exactly which evidence and task objects support an evaluation, report the complete bounded comparison, and preserve an explicit non-result when a stronger interpretation lacks its required evidence.

AI USE STATEMENT

Generative AI tools were used extensively in this project. They assisted with research-topic refinement; literature discovery and synthesis; development and critique of the conceptual framework, formal claims, assumptions, and proof sketches; experimental and audit design; generation of synthetic conformance fixtures; software implementation and testing; construction and review of data-processing and statistical procedures; preparation of figures, tables, artifact documentation, and reproducibility checks; interpretation and stress-testing of intermediate results; translation; and drafting and editing the manuscript.

The workflow used separate AI-assisted implementation and review roles to find inconsistencies, missing gates, unsupported claims, and software failures. These roles do not constitute independent human validation. Citations in the manuscript were checked against identifiable public sources rather than accepted from model memory. Data-derived quantities must be generated by bound analysis code and cannot be invented or selected by a language model. Restricted data and results remain subject to separate access and disclosure controls.

The human author is responsible for the research question, authorization to use each data source, final design and claim boundaries, review of the formal and empirical work, citation accuracy, and the final manuscript. The author retains responsibility for the complete submission, including text, claims, code, and artifacts produced with generative-AI assistance.

ETHICS STATEMENT

The study does not intervene on human participants and does not infer traits about individual researchers. Neither task revision nor bibliographic revision is treated as evidence of error, misconduct, retraction, or research quality. The reviewer artifact omits row-level bibliographic records, task text, model outputs, restricted data, credentials, and internal paths. The analysis is limited to the disclosed research scope; data access and publication permission are not inferred from a numerical result or from a content hash.

REPRODUCIBILITY STATEMENT

Section 2 defines the populations and estimands; Section 3 states the reference execution order and implemented checker boundary; Section 4 reports the entire fixed-model crossed panel. Appendix A supplies proofs and boundary cases, and the empirical appendix specifies source/model identities, inference and aggregation semantics, diagnostics, and execution limitations. The anonymous supplement regenerates the tables from the complete disclosed aggregate and tests the included software on synthetic fixtures. This is distinct from replicating model inference or the earlier record-level independent recomputation; no such inference rerun is claimed for the supplement check.

REFERENCES

BIG-bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Carlos Bobed, Pierre Maillot, Peggy Cellier, and Sébastien Ferré. Data-driven assessment of structural evolution of RDF graphs. *Semantic Web*, 11(5):831–853, 2020. doi: 10.3233/SW-200368. URL https://doi.org/10.3233/SW-200368.

Tiffany Tianhui Cai, Hongseok Namkoong, and Steve Yadlowsky. Diagnosing model performance under distribution shift. *Operations Research*, 74(2):898–916, March 2026. doi: 10.1287/opre.2023.0217. URL https://pubsonline.informs.org/doi/10.1287/opre.2023.0217.

Lingjiao Chen, Zhihua Jin, Evan Sabri Eyuboglu, Christopher Ré, Matei Zaharia, and James Zou. Hapi: A large-scale longitudinal dataset of commercial ml api predictions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24571–24585. Curran Associates, Inc., 2022. doi: 10.52202/068431-1784. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9bcd0bdb2777fe8c729b682f07e993f1-Abstract-Datasets_and_Benchmarks.html.

Fedor Chernogorskii, Sergei Averkiev, Liliya Kudraleeva, Zaven Martirosian, Maria Tikhonova, Valentin Malykh, and Alena Fenogenova. DRAGOn: Designing RAG on periodically updated corpus. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 622–638. Association for Computational Linguistics, March 2026. doi: 10.18653/v1/2026.eacl-srw.48. URL https://aclanthology.org/2026.eacl-srw.48/.

Dean Croushore and Tom Stark. A real-time data set for macroeconomists. *Journal of Econometrics*, 105(1):111–130, November 2001. doi: 10.1016/S0304-4076(01)00072-0. URL https://doi.org/10.1016/S0304-4076(01)00072-0.

Dean Croushore and Tom Stark. A real-time data set for macroeconomists: Does the data vintage matter? *The Review of Economics and Statistics*, 85(3):605–617, August 2003. doi: 10.1162/003465303322369759. URL https://doi.org/10.1162/003465303322369759.

Abdelilah El Majjaoui, Anass El Haddadi, Abdelkhalek Bahri, Oumaima El Haddadi, Fadwa Bouhafer, and Sara Ouald Chaib. Auditable bibliometric analysis: A generic data pipeline for research evaluation. *Journal of Informetrics*, 20(2):101808, June 2026. doi: 10.1016/j.joi.2026.101808. URL https://www.sciencedirect.com/science/article/pii/S1751157726000416.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, November 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Alba González-Cebrián, Michael Bradford, Adriana E. Chis, and Horacio González-Vélez. Standardised versioning of datasets: A FAIR-compliant proposal. *Scientific Data*, 11(1):358, April 2024. doi: 10.1038/s41597-024-03153-y. URL https://doi.org/10.1038/s41597-024-03153-y.

Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. Temporal graph benchmark for machine learning on temporal graphs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 2056–2073. Curran Associates, Inc., 2023. doi: 10.52202/075280-0099. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/066b98e63313162f6562b35962671288-Abstract-Datasets_and_Benchmarks.html.

Hugging Face. distilbert/distilgpt2: Immutable model repository revision 96699c9fd598cfb0babef4e8d217a9f9586b2bb6. Hugging Face Model Hub, May 2021. URL https://huggingface.co/distilbert/distilgpt2/tree/96699c9fd598cfb0babef4e8d217a9f9586b2bb6. Accessed 2026-09-15.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6237–6250. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.418. URL https://aclanthology.org/2022.emnlp-main.418/.

Shachar Kaufman, Saharon Rosset, and Claudia Perlich. Leakage in data mining: Formulation, detection, and avoidance. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pp. 556–563. ACM, August 2011. doi: 10.1145/2020408.2020496. URL https://doi.org/10.1145/2020408.2020496.

Jens Klump, Lesley Wyborn, Mingfang Wu, Julia Martin, Robert R. Downs, and Ari Asmi. Versioning data is about more than revisions: A conceptual framework and proposed principles. *Data Science Journal*, 20(1):12, 2021. doi: 10.5334/dsj-2021-012. URL https://doi.org/10.5334/dsj-2021-012.

Rachel Longjohn, Markelle Kelly, Sameer Singh, and Padhraic Smyth. Benchmark data repositories for better benchmarking. In *Advances in Neural Information Processing Systems*, volume 37, pp. 86435–86457. Curran Associates, Inc., 2024. doi: 10.52202/079017-2744. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9d41fddf3169f4df3025fc60bfac94fa-Abstract-Datasets_and_Benchmarks_Track.html.

Charles F. Manski. Partial identification with missing data: Concepts and findings. *International Journal of Approximate Reasoning*, 39(2–3):151–165, June 2005. doi: 10.1016/j.ijar.2004.10.006. URL https://doi.org/10.1016/j.ijar.2004.10.006.

Alexandru A. Ormenisan, Moritz Meister, Fabio Buso, Robin Andersson, Seif Haridi, and Jim Dowling. Time travel and provenance for machine learning pipelines. In *2020 USENIX Conference on Operational Machine Learning (OpML 20)*. USENIX Association, July 2020. URL https://www.usenix.org/conference/opml20/presentation/ormenisan.

Erkin Otles, Jeeheh Oh, Benjamin Li, Michelle Bochinski, Hyeon Joo, Justin Ortwine, Erica Shenoy, Laraine Washer, Vincent B. Young, Krishna Rao, and Jenna Wiens. Mind the performance gap: Examining dataset shift during prospective validation. In Ken Jung, Serena Yeung, Mark Sendak, Michael Sjoding, and Rajesh Ranganath (eds.), *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pp. 506–534. PMLR, 06–07 Aug 2021. URL https://proceedings.mlr.press/v149/otles21a.html.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1776–1826. Association for Computing Machinery, 2022. doi: 10.1145/3531146.3533231. URL https://doi.org/10.1145/3531146.3533231.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Andreas Rauber, Bernhard Goesswein, Carlo Maria Zwoelf, Chris Schubert, Florian Woerister, James Duncan, Katharina Flicker, Koji Zettsu, Kristof Meixner, Leslie D. McIntosh, Reyna Jenkyns, Stefan Proell, Tomasz Miksa, and Mark A. Parsons. Precisely and persistently identifying and citing arbitrary subsets of dynamic data. *Harvard Data Science Review*, 3(4), October 2021. doi: 10.1162/99608f92.be565013. URL https://hdsr.mitpress.mit.edu/pub/si7wzxxa.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 21763–21813. Curran Associates, Inc., 2024. doi: 10.52202/079017-0685. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/26889e8359e7ef8a7f5d77457364ca55-Paper-Datasets_and_Benchmarks_Track.pdf.

Yiwei Wu, Leah Ajmani, Shayne Longpre, and Hanlin Li. A systematic review of NeurIPS dataset management practices. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 32813–32827. Curran Associates, Inc., 2024. doi: 10.52202/079017-1032. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/398d2261db1a72f5d4dd03673d9a7c94-Paper-Datasets_and_Benchmarks_Track.pdf.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei W Koh, and Chelsea Finn. Wild-time: A benchmark of in-the-wild distribution shift over time. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10309–10324. Curran Associates, Inc., 2022. doi: 10.52202/068431-0749. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/43119db5d59f07cc08fca7ba6820179a-Abstract-Datasets_and_Benchmarks.html.

Klim Zaporojets, Lucie-Aimée Kaffee, Johannes Deleu, Thomas Demeester, Chris Develder, and Isabelle Augenstein. TempEL: Linking dynamically evolving and newly emerging entities. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc.,

2022. doi: 10.52202/068431-0135. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0c3464f16c854d395b880cf9e7bcaf2f-Abstract-Datasets_and_Benchmarks.html.

A FORMAL BOUNDARIES AND PROOFS

A.1 TERMINAL-STATE NONIDENTIFICATION

Let $A_i(a) \in \{0,1\}$ indicate whether stable unit i is exposed by the source at evidence time a. Proposition 1 assumes a terminal state at a_1, an earlier target time $a_0 < a_1$, at least one terminal unit whose represented event precedes T, and admissible delayed exposure. Construct histories h and h' with identical terminal identities, content, and event times. Set $A_i(a_0) = 1$ in h; set $A_i(a_0) = 0$ in h' and expose i between a_0 and a_1. Both histories can satisfy monotone availability. The terminal observation map is many-to-one, so any earlier-support functional that changes with i is not identified. An authenticated earlier snapshot or an availability-time field with adequate semantics adds information excluded by the proposition.

A.2 FINITE-WITNESS NONIDENTIFICATION

Let $q_0 < \cdots < q_K$ be any finite set of registered label states and suppose $Y_i^{(q_0)} = \cdots = Y_i^{(q_K)}$ on the frozen roster. Assume at least one component admits two label values. One path keeps the vector fixed after q_K; a second agrees at every observed state, changes that component immediately after q_K, and retains the new value. Unless q_K is known terminal or an externally justified absorbing rule applies, both paths generate the same finite observations but different permanence indicators. The audit can report agreement only over the named states.

A.3 SHARP UNION-COMPLETION INTERVAL

Write $D_C = \sum_{i \in C}(\ell_{i1}^m - \ell_{i0}^m)$, let $M = E \cup X$, and assume $N = |C \cup M| > 0$. If the registered completion class is exactly the product box $\prod_{i \in M}[a_i, b_i]$, with finite $a_i \le b_i$, its image under the linear map $(D_C + \sum_{i \in M} \delta_i^m)/N$ is

$$\mathcal{I}_U = \left[\frac{D_C + \sum_{i \in M} a_i}{N}, \frac{D_C + \sum_{i \in M} b_i}{N} \right]. \tag{6}$$

Endpoint assignments attain both limits. For any $t \in [0,1]$, assigning $\delta_i^m = a_i + t(b_i - a_i)$ reaches the same fractional position in the interval, proving sharpness relative to the stated box. If no completion target is defined, the target is unavailable rather than unbounded.

For finite squared losses with only nonnegativity imposed on missing-side loss, an entry difference ranges over $(-\infty, \ell_{i1}^m]$ and an exit difference over $[-\ell_{i0}^m, \infty)$. The identified set is a point with no migration, a one-sided interval with entry-only or exit-only migration, and all of $\mathbb{R}$ when both occur.

A.4 NATIVE, RETAINED, AND UNION TARGETS ARE NOT INTERCHANGEABLE

Let $C = \{c\}$, $E = \{e\}$, and $X = \varnothing$. Loss triples $(\ell_{c0}, \ell_{c1}, \ell_{e1}) = (0,0,2)$ and $(0,1,1)$ have the same native difference 1 but retained-support differences 0 and 1. Triples $(0,0,2)$ and $(0,0,0)$ have the same retained-support difference but native differences 1 and 0. Holding the first observed triple fixed while varying the nonnegative missing completion $\widetilde{\ell}_{e0}$ makes the union mean $(2 - \widetilde{\ell}_{e0})/2$ range over $(-\infty, 1]$. With no migration, the three objects coincide when labels, procedures, weights, and orientations are held fixed; with migration, equality requires realized-value coincidence or additional restrictions.

A.5 INTERFACE COLLISION WITNESSES

Let Ω contain structurally admissible audit histories and let $\Gamma_{\mathcal{Q}}(\omega)$ be the qualified claim records emitted by the closed interface. A projection R is collision-free for this interface if $R(\omega) = R(\omega')$ implies $\Gamma_{\mathcal{Q}}(\omega) = \Gamma_{\mathcal{Q}}(\omega')$. The following witnesses show only that dropping a field can create a collision for this chosen interface; they do not establish a globally minimal representation.

The proof is by paired witnesses. Hold numerical payloads and all other blocks fixed while changing: (i) the ordered digest pair; (ii) an otherwise unrecoverable cross-state identity relation; (iii) the named label state while realized labels coincide; (iv) the product-box restrictions while scalar endpoints

coincide; or (v) the reason for a withheld claim. In each pair, the allowed qualified record differs although a projection that drops the changed block is identical. Content commitments do not prove provider issuance, completeness, causal intent, or authorization.

B CLOSED CONTRACT SEMANTICS AND VALIDATOR SOUNDNESS

B.1 CLOSED SCHEMA AND ASSURANCE-INDEXED OBJECTS

For each tagged record sort t, let Req_t, Opt_t, and Inv_t denote its required keys, branch-local optional keys, and cross-field invariants. A record satisfies the closed schema exactly when, for some $O \subseteq \mathrm{Opt}_t$, its keys are $\mathrm{Req}_t \cup O$, every value has its declared sort, and all invariants hold. Tagged branches specify exact key sets; unknown or duplicate keys, unknown enum values, and non-finite numbers are invalid. Canonical serialization sorts object keys, preserves array order, and fixes finite-number rendering before hashing.

At CARD_INTERNAL assurance, a syntactically valid digest is a commitment string and source-facing claims have only an InternallyConsistent type. At BUNDLE_REPLAY assurance, the verifier additionally requires the exact artifact roster, checks every digest and size, binds the executable specification, and deterministically recomputes every source-derived field. The assurance tag is therefore derived from completed guards rather than selected by the caller. This paragraph defines the stronger reference obligation: the present aggregate companion does not open that roster and therefore cannot emit replay assurance or a publishable terminal.

Every quantity has the disjoint availability type

$$\mathsf{Avail}(A) = \mathsf{Value}(A) \mid \mathsf{Set}(A, \mathcal{K}) \mid \mathsf{Unavailable}(r).$$

Quantity reasons form a finite enum, including NO_COMMON_SUPPORT, NO_COMPLETION_TARGET, CONSTANT_PREDICTION, and INSUFFICIENT_LABEL_WITNESSES. Authorization is a separate enum and is never a numerical availability reason.

B.2 CONSTRUCTOR OBLIGATIONS

Table 3: Replay-level claim constructors and their maximum interpretations.

Constructor	Necessary replay premises	Type and hard boundary
Native	bound pair/task/model/loss; nonempty native sets; recomputed means	descriptive on (U_0, U_1); not paired change
Retained	exact partition; $C \neq \varnothing$; paired identities and losses	descriptive on C; not union behavior
Decomp	preceding premises; conservation; Eq. 2; empty-side sum rule	accounting on (C, E, X); no mechanism
Union	defined union target; exact $\mathcal{K}$; complete migrated-unit coverage; exact linear image	set-identified relative to $\mathcal{K}$; no missing loss observed
Witness	exact ordered witness roster and label binding; complete recomputation	agreement on named witnesses; no permanence
Order	$C \neq \varnothing$; same models, orientation, and tie tolerance	point description; no superiority
NonResult	exactly one valid unavailable or structural-stop reason	non-result; no numeric payload

For example, with $D_C = \sum_{i \in C} \delta_i^m$ and $\mathcal{K} = \prod_{i \in E \cup X} [a_i, b_i]$, the Union constructor is licensed only if its interval equals

$$[(D_C + \textstyle\sum_i a_i)/N, (D_C + \sum_i b_i)/N]$$

and stores the commitment to $\mathcal{K}$. With migration but no completion semantics it instead yields NonResult(NO_COMPLETION_TARGET). Equal realized scalars do not permit a cast between native, retained, and union types. A publication wrapper preserves the derived type exactly and may emit a publishable decision only after it opens an external authorization envelope, authenticates the issuer under a named trust policy, and verifies that the external bytes bind the complete semantic-payload digest and declared purpose.

B.3 State machine and reference algorithm

Let q_0 be start and let $q_1, \ldots, q_7$ denote schema closure, design binding, partition validation, availability assignment, recomputation, claim typing, and committed decision. The stored trace must be one valid prefix followed by exactly one terminal state. A failed structural guard enters $\mathsf{REJECT}(r)$; a well-formed but unavailable quantity remains on the successful structural path with an explicit non-result. If $C = \varnothing$, for example, the partition may be valid while paired metric blocks are never evaluated.

The reference algorithm first closed-decodes the card and binds the ordered sources, frozen design, artifact roster, and code. In replay mode it then verifies the complete supplied-byte roster, derives statewise eligibility and the exact $C/E/X$ partition, rejects duplicate identities, and checks conservation; internal mode checks reported-count conservation only. It assigns one availability branch to every requested quantity, evaluates only the guards and arithmetic of that branch, and validates finite witnesses on their named roster. It derives the complete claim set by the finite constructor rules. Finally, a conforming replay implementation atomically commits the assurance level, trace, availability map, claims, commitments, and a withheld scientific receipt. Only a distinct wrapper may later issue a publishable decision after the external-envelope checks above. The current card-internal companion does not evaluate, publish, or authorize an outcome-bearing scientific claim before all preceding semantic guards pass; this logical ordering is not a trusted timestamp or evidence of historical execution order.

B.4 Proof of Theorem 1 and trusted boundary

Induct on the unique trace. Schema closure gives a unique typed record. Design binding introduces only assurance-appropriate facts: exact supplied-byte facts exist only in replay mode. Partition guards establish either the replayed sets and their conservation identities or only the weaker reported-count facts. The tagged availability sum makes value, set, and non-result branches disjoint, so an unavailable quantity cannot supply a numerical premise. Recomputation establishes each constructor equality under the fixed, design-bound numerical policy with no claim-local tolerance; Equation 2 follows by expanding native means, and linearity plus endpoint attainability gives the product-box interval. Inversion on each final typing derivation recovers the constructor's exact source, population, witness, and assumption premises. Any later publication rule lies in a distinct wrapper and additionally requires an opened external authorization envelope, authenticated under the named trust policy and verified against the complete semantic-payload digest and purpose. Since the grammar contains no introduction or cast rule for the forbidden interpretations, the validator cannot derive them. This proves the stated reference-semantics claim soundness under deterministic syntax and semantics, complete replay inputs when replay is claimed, the fixed design-bound numerical policy, and correct execution of the reference runtime and atomic writer.

It does not prove that the aggregate companion refines the replay semantics. The theorem also does not establish provider issuance, upstream completeness, the substantive correctness of identity or eligibility choices, absence of unregistered artifacts, record truth, causal interpretation, future label permanence, external validity, power, model superiority, or production-code refinement. The closed synthetic differential tests cover only their finite fixture.

Table 4: Audit objects, populations, and maximum interpretations.

Object	Reports	Does not establish
Source state	named content and order	provider issuance or completeness
Native support	entry, exit, migration	retained-unit change
Common support	paired loss or grade	union behavior
Label witnesses	agreement on named states	final truth or permanence
Sensitivity	conclusions under named bounds	observed counter-state losses
Claim gate	eligible wording or non-result	causal provider intent

Table 5: Minimum logical blocks in the closed audit card and decision receipt.

Block	Bound content
Source	exact identifiers, digests, order, acquisition semantics
Task	forecast origin, unit, eligibility, folds, feature and label state
Identity	stable matching rule, duplicate policy, $C/X/E$ commitments
Procedure	complete model and metric roster, orientation, seeds, tuning policy
Sensitivity	label witnesses and missing-side completion restrictions
Availability	requested quantity, value or registered non-result reason
Decision	structural status, claim branch, access and manuscript authorization

C AUDIT-CARD INTERFACE

The representation may use signed manifests or restricted commitments instead of public row identifiers. A card-internal checker establishes schema closure and selected cross-field consistency. A conforming bundle-replay validator would additionally establish agreement between named commitments and the exact supplied bytes, subject to its hash and trusted-runtime assumptions; accepting self-reported digests is insufficient. Neither establishes provider issuance, upstream completeness, or record truth without external evidence.

D COMPLETE BIG-BENCH RESULTS AND EXECUTION PROVENANCE

This appendix accompanies the exploratory stable-support vignette in Section 4. It contains the entire frozen diagnostic panel, exact model identities, reporting semantics, and the implementation amendment history. No raw input text, task rows, target maps, row identifiers, or row-level predictions are distributed in these tables.

D.1 STATE, MODEL, AND SCORING IDENTITIES

The two task states are the unique parent `7d331b3e8b5939e90442ab07d3e8903e0de5ac78` and repair commit `70e9f8702c12f77b12100326c0a52adcf80d1414` in `google/BIG-bench`, at `bigbench/benchmark_tasks/salient_translation_error_detection/task.json`. The repair was committed on 15 June 2021, before pull request 374 merged on 20 July 2021. The source record does not establish that the parent state was formally released, used on a leaderboard, or used in a published evaluation.

Table 6: Exact frozen model revisions; the complete panel is retained.

Model	Repository	Revision
DistilGPT2	`distilbert/distilgpt2`	`96699c9fd598cfb0babef4e8d217a9f9586b2bb6`
T5-small	`google-t5/t5-small`	`255e6b7587fa94e9cce60b7051755f072601f82f`
T5-base	`google-t5/t5-base`	`5a7ff2d8f5117c194c7e32ec1ccbf04642cca99b`

The 1,000 source examples in each state reduce to 998 effective inputs under the frozen runner's exact duplicate handling. Stable input identities are shared across all model–cell blocks. The earlier target state contains 994 six-choice and four seven-choice items; the later target state contains 998 six-choice items. Target/action changes do not change the fixed input identity roster. Local content commitments bind the selected objects; they are not external preregistration or evidence of first public availability.

Each crossed cell starts from a fresh copy and a fresh `RandomState(42)`. Under q_0, choices sorted by their string keys are permuted and appended using the runner's sequential random stream. Different choice cardinalities between y_0 and y_1 can change subsequent permutations even for items whose target maps did not change. Under q_1, choices are not appended to the prompt. These semantics were frozen before the scientific outcome was inspected; a per-item reseeding analysis is not substituted for them.

The frozen adapters sum conditional target-token log probabilities without length normalization. The causal adapter identifies the token boundary in the concatenated context and candidate; the T5 adapter uses conditional encoder–decoder target scoring. Scores are normalized in NumPy float32, and the official SHA-256 rule selects among exactly tied maxima. The inference profile uses CPU, `torch_eager`, float32, batch size one, deterministic algorithms, and one computational thread; no training or outcome-driven tuning is performed. All 11,976 ordered records were required for the complete private bundle. The render oracle and aggregate verifier are distinct from the always-withheld claim-receipt checker in the methods. The bound runtime records Linux x86_64, CPython 3.8.10, PyTorch 2.4.1+cpu, Transformers 4.46.3, NumPy 1.24.4, and SciPy 1.10.1. Processor model and measured total-run elapsed time are not recorded in the retained runtime and aggregate receipts; synthetic-profile projections are not reported as measured scientific-run costs.

D.2 Complete registered quantities and numerical reporting

For each model, the four finite-roster mean grades and all six contrasts appear in Table 1. A positive contrast always means the later-indexed state minus the earlier-indexed state:

$$\begin{aligned}
\Delta_{\mathrm{obs}} &= g(1,1) - g(0,0), \\
\Delta_q(y_0) &= g(1,0) - g(0,0), & \Delta_q(y_1) &= g(1,1) - g(0,1), \\
\Delta_y(q_0) &= g(0,1) - g(0,0), & \Delta_y(q_1) &= g(1,1) - g(1,0), \\
\Delta_{q\times y} &= g(1,1) - g(1,0) - g(0,1) + g(0,0).
\end{aligned}$$

The model macro-average weights the three fixed models equally, not by architecture, size, or observed score. It is not an estimate for a population of models. Means use the frozen `math.fsum` accumulation followed by division by the exact count; contrasts use `math.fsum` of the signed stored means in the fixed cell order. Display rounding is applied only after these recorded computations, without changing the bound aggregate.

For cell (q, y), the uniform-choice reference is

$$u(q,y) = \frac{1}{998} \sum_{i=1}^{998} \frac{1}{J_i(y)} \sum_{j=1}^{J_i(y)} t_{ij}(y).$$

Here $J_i(y)$ is the number of available choices for item i in target state y, and $t_{ij}(y)$ is the target score assigned to choice j. The machine field retains the historical name `cell_uniform_choice_expected_grade_floor`, but u is an expectation, not a mathematical lower bound for a model. Its denominator is cell-specific, and it is not used to normalize the reported grades.

The observed-state diagnostics are

$$d_m = \frac{1}{998} \sum_i \mathbf{1}\{a_{mi}(0,0) \neq a_{mi}(1,1)\}, \qquad c_m = \frac{1}{998} \sum_i \mathbf{1}\{G_{mi}(0,0) \neq G_{mi}(1,1)\}.$$

Here a is the persisted selected-choice index, not a common semantic label. Grade inequality uses the exact stored finite numeric values after conversion to Python float. All diagnostic numerators and denominators are retained.

The target is the entire fixed roster, not a randomly sampled population. Bootstrap output is disabled in the final pre-outcome analysis configuration; the older conditional reporting-shell mention of a possible descriptive bootstrap does not override that decision. No post hoc confidence intervals, significance tests, equivalence margins, or subgroup analyses are added.

The anonymous machine-readable companion `tables/bigbench_scientific_aggregate.json` preserves all 92 scalar leaves of the approved scientific subtree, including exact unrounded numbers, contrast direction labels, model revisions, denominators, and the fixed-panel non-generalization flag. It excludes private filesystem paths, author identifiers, and row-level payloads. The deterministic table renderer also emits a display manifest binding the 30 main model-result slots and all generated output digests. These derivative bindings do not themselves grant publication authority or independently establish model-call correctness.

Table 7: All observed-state diagnostics, q_0y_0 versus q_1y_1. Rates are percentages. Selected indices are not aligned semantic labels. Macro rates are arithmetic means over the three fixed models.

	Index disagreement			Grade change		
Model	Count	n	Rate	Count	n	Rate
DistilGPT2	998	998	100.0000	351	998	35.1703
T5-small	898	998	89.9800	319	998	31.9639
T5-base	993	998	99.4990	347	998	34.7695
Fixed-panel mean	—	—	96.4930	—	—	33.9679

Cell	q_0y_0	q_1y_0	q_0y_1	q_1y_1
Uniform-choice expectation (%)	16.6571	16.6571	16.6667	16.6667

D.3 Chronology and software corrections

The source, model roster, scoring semantics, and reporting branches were locally frozen before inspecting scientific outcomes. The aggregate analysis contract and its first independent implementation were frozen before scientific execution or access. The work is exploratory, not externally preregistered or independently controller-authorized.

After the complete model run, both original aggregate parsers contained an invalid prerequisite: equality of choice-vector lengths across the two target states. This rejected the four documented seven-to-six-choice changes with `OBSERVED_STATE_CHOICE_DIMENSION_MISMATCH`. The v0.2 correction removed only this cross-state equality. Within-record dimensions, selected-index bounds, fixed-y target invariance, complete roster, finite values, and lineage checks remained. The frozen formulas, denominators, model roster, and choice-index-only interpretation did not change. The corrected implementation passed 62 synthetic tests locally and in the analysis runtime.

The v0.2 primary aggregate then passed, but its independent verifier stopped with `PRIMARY_AGGREGATE_RECEIPT_CONTENT_MISMATCH`. A value-free comparison localized three cross-model mean fields. Synthetic examples reproduced a rounding-stage discrepancy: the primary rounded the sum before division, as frozen, while the verifier rounded the exact rational mean only once. The verifier-only v0.3 change restored the registered sum-then-division rounding stages. It introduced no tolerance, did not change the successful primary aggregate, and passed 37 synthetic tests locally and in the analysis runtime; a separate review checked 12,100 additional invented mean vectors.

Both amendments occurred after automated processing of outcome records but before any human or agent inspected scientific values. This timing is not a pre-run amendment or external preregistration. Both original failure records and earlier implementation versions were retained. No model was rerun, no item was excluded, and no outcome-driven model selection or retuning occurred. The unchanged primary v0.2 aggregate subsequently matched the independent v0.3 verifier on all 92 scientific-subtree leaves and 217 complete-receipt leaves.

Interpretation limits. The three identical q_1 mean grades near the uniform-choice expectation do not identify whether outputs are identical, concentrated on one label, affected by label-length preference, or associated with a shared invocation defect. The approved aggregate contains no selected-index histogram, cross-model prediction concordance, probability entropy, or score-margin summary. No additional record-level diagnostic is claimed to have been run. The present evidence establishes descriptive sensitivity within this single task-state pair, not a resolved mechanism or a generally capable model evaluation. The implementation amendments and independent agreement are necessary numerical evidence, not proof of substantive validity.