HISTORICAL EVALUATION WITHOUT A UNIQUE PAST: IDENTIFICATION, COMPLEXITY, AND EVIDENCE FROM A REAL-TIME ARCHIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Historical evaluation often uses the archive available today to approximate the data available at a past decision. Even when the decision protocol is known, incomplete provenance can leave several historical states consistent with it. We represent this ambiguity by a set of admissible worlds. A loss, loss contrast, or winner is identified when it is invariant across that set, even if the historical state itself remains unknown. We prove that strict winner reversal is NP-complete even for one cutoff, binary availability, fixed targets, frozen linear predictors, and squared loss. Ordered revisions admit an exact cutoff quotient, while fixed-dimensional affine timing permits exact evaluation over hyperplane-arrangement faces. An audit of 38 quarterly decisions using 1,180 official ALFRED snapshots shows why the admissible set matters. When one feed age is shared across feeds and fixed across decisions, same-world comparisons certify all ten pairwise rankings on a declared four-age support; separate marginal ranges leave four unresolved. Adding intermediate daily ages exposes one reversal, and allowing feed-specific fixed ages exposes two. If ages may instead vary by decision, only four rankings remain certified. Reusing identical resolved states reduces 256 worlds with fixed feed-specific ages to 32 states and cuts model calls eightfold. These results connect historical claim identification to explicit temporal assumptions, exact computational structure, and verifiable archival evidence.

1 INTRODUCTION

Retrospective evaluation asks a deceptively simple question: how would a frozen model have performed using what was known at a past decision time? A current database may no longer preserve that historical state. Preliminary values are revised, series are rebased, and records arrive asynchronously; the latest snapshot may not reveal which version was used in a decision. Standard leakage controls ask whether a declared pipeline uses information that was available at prediction time (Kaufman et al., 2012). We ask the complementary question: *after the protocol is fixed, what can still be concluded when today's evidence admits more than one protocol-consistent past?*

Let a decision contract C and surviving evidence E induce admissible worlds $\mathcal{W}(C, E)$. A frozen evaluator computes its loss from the resolved evaluation state: $Q_f(w) = q_f(R_C^{\mathrm{eval}}(w))$. For two models, $\Delta_{AB}(w) = Q_A(w) - Q_B(w)$, with both models evaluated in the *same* world. The evidence may determine an evaluation claim even when it does not determine the historical state. Encoding the winner by the sign of the loss contrast, with zero denoting a tie, gives

$$\text{state identified} \Rightarrow \text{joint losses identified} \Rightarrow \text{contrast identified} \Rightarrow \text{winner identified}. \tag{1}$$

Every converse can fail: two different inputs can yield the same losses; loss pairs $(1, 2)$ and $(2, 3)$ have the same contrast; and contrasts -1 and -2 have the same winner. Identifying an evaluation claim can therefore require less information than reconstructing the full state.

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.

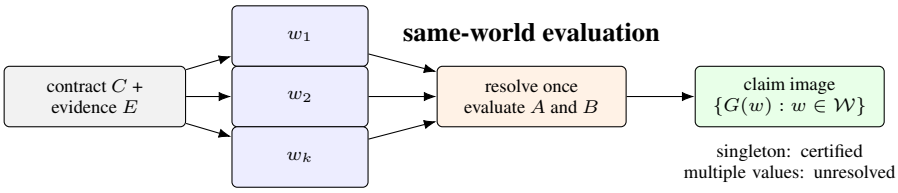

Figure 1: **Claim-specific historical identification.** Evidence can admit several decision-time worlds. Certification propagates the *same* world through the resolver and frozen evaluator, then asks whether the downstream claim is invariant.

Contributions.

- **A claim-specific evaluation target.** We distinguish protocol validity, state identification, loss identification, and winner identification, then specify which historical claims are determined by a contract and its surviving evidence.
- **A computational boundary.** Strict winner reversal is NP-complete in a restricted setting with frozen linear predictors, and exact lower-bound computation is strongly NP-hard. In contrast, ordered revisions and fixed-dimensional shared affine timing admit exact structural reductions.
- **A dependence-preserving comparison.** We optimize loss contrasts while evaluating both models in the same world and characterize when separate marginal ranges are strictly wider. The matched evaluation procedures use the same model calls.
- **An auditable empirical study.** Real ALFRED vintages support a controlled experiment with hidden schedule identity. Support expansion, feed-specific timing, decision-varying timing, and block resampling distinguish historical invariance from sampling stability. A caching experiment across all feed subsets measures exact state reuse; a later-vintage replay assesses the effect of revised inputs.

Takeaway. Historical ambiguity should be propagated to the claim being made: report the chosen point estimate, the admissible set, and whether the conclusion changes across that set.

2 Historical Claims as Identified Sets

A contract

$$C = (\mathsf{P}, \mathsf{D}, \mathsf{A}, \rho, \mathsf{L}, \mathsf{M}, \mathsf{Y}) \tag{2}$$

specifies the knowledge principal P (whose information counts), decision times and cutoffs D, allowed sources A, resolver ρ, transformation lineage L, label-maturity rule M, and target definition Y. The resolver maps each admissible history to the inputs and outcomes used by the frozen evaluator; their collection over the evaluation cases is $R_C^{\mathrm{eval}}(w)$. With surviving evidence E,

$$\mathcal{W}(C, E) = \{w : w \text{ is jointly consistent with } C \text{ and } E\}. \tag{3}$$

We assume $\mathcal{W}(C, E) \neq \varnothing$ and coverage $w^\star \in \mathcal{W}(C, E)$ for the realized history $w^\star$. An empty set indicates that the contract and evidence are inconsistent. Even if all arrivals are retained, a revision can replace a preliminary value in the resolved state.

For a real-valued frozen loss (smaller is better), assumed bounded over the admissible set,

$$\Theta_f = \{Q_f(w) : w \in \mathcal{W}(C, E)\}, \qquad \underline{Q}_f = \inf_w Q_f(w), \quad \overline{Q}_f = \sup_w Q_f(w). \tag{4}$$

The interval $[\underline{Q}_f, \overline{Q}_f]$ is its interval hull; values inside it need not be attainable. For a general claim G, its *claim image* is the set $\Theta_G(C, E) = \{G(w) : w \in \mathcal{W}(C, E)\}$ of admissible values. The claim is identified exactly when this set is a singleton. We call this a *historical certificate*: it concerns the observed evaluation cases and their admissible histories. A strict winner is certified by $\sup_w \Delta_{AB}(w) < 0$ or $\inf_w \Delta_{AB}(w) > 0$. For a finite admissible set, these conditions are also necessary for an invariant strict winner.

2.1 SAME-WORLD COMPARISONS

Separate marginal loss ranges $I_A = [\underline{Q}_A, \overline{Q}_A]$ and $I_B = [\underline{Q}_B, \overline{Q}_B]$ imply

$$I_{\text{marg}} = [\underline{Q}_A - \overline{Q}_B,\ \overline{Q}_A - \underline{Q}_B]. \tag{5}$$

The exact paired comparison instead uses

$$I_{\text{pair}} = [\inf_w \{Q_A(w) - Q_B(w)\},\ \sup_w \{Q_A(w) - Q_B(w)\}]. \tag{6}$$

Proposition 1 (Same-world containment)**.** $I_{\text{pair}} \subseteq I_{\text{marg}}$. *For either endpoint, if the corresponding paired extremum is attained—in particular, if $\mathcal{W}(C, E)$ is finite—then the paired bound is strictly tighter whenever no single world jointly attains the two extrema defining the corresponding marginal bound.*

Every attainable pair $(Q_A(w), Q_B(w))$ lies in $I_A \times I_B$, whereas the Cartesian product also contains cross-world pairs. Evidence refinement is monotone: if $\mathcal{W}(C, E') \subseteq \mathcal{W}(C, E)$, every claim image can only shrink.

3 EXACT CERTIFICATION: HARDNESS AND STRUCTURE

Certification can be computationally hard even without training: one uncertain availability vector affects multiple evaluation cases. The decision instance explicitly encodes independent binary timing choices and two rational frozen predictors; it asks whether two worlds encoded with polynomially many bits yield opposite strict winners. Fix one decision time $t = 0$. For source i, archival evidence permits $\kappa_i \in \{-1, +1\}$ and $a_i(w) = \mathbf{1}\{\kappa_i(w) \le 0\}$, with every $a \in \{0,1\}^m$ reachable. Given a nonempty Exact Cover by 3-Sets (X3C) instance with incidence matrix $A \in \{0,1\}^{N \times m}$, where $N = 3q$, create one evaluation case per universe element. Case j receives binary feature $A_{ji}a_i(w)$ from source i; a fixed vector of unit weights predicts $(Aa(w))_j$ against a target of one. Its summed squared loss is

$$L_f(w) = \|Aa(w) - \mathbf{1}\|_2^2. \tag{7}$$

Then $L_f(w) = 0$ exactly when the selected sets form an exact cover. Give the comparator g an always-available coordinate and constant prediction $1 - 1/(2N)$ on every case, hence $L_g = N(1/(2N))^2 = 1/(4N) \in (0,1)$; the all-unavailable world has $L_f = N > L_g$.

Theorem 1 (Strict historical winner reversal)**.** *Deciding whether two admissible worlds yield opposite strict winners is NP-complete, even with one cutoff, binary before/after availability, frozen scalar linear prediction, binary payloads, fixed unit targets, squared loss, and no training. Exact computation of $\min_w L_f(w)$ is strongly NP-hard.*

Membership in NP follows from two polynomial-size world certificates and exact rational loss evaluation. Strong NP-hardness follows because the minimization construction uses only binary payloads and unit weights and targets; Appendix A gives the complete reduction.

3.1 TWO EXACT REGIMES

For an ordered chain $s_0 \to \cdots \to s_q$ with transition times $\kappa_1 \le \cdots \le \kappa_q$ and decisions $t_1 < \cdots < t_n$, let b_r be the first decision index at which transition r has occurred, or $n + 1$ if it occurs after all decisions.

Theorem 2 (Cutoff quotient)**.** *Under a deterministic path-independent resolver, a frozen evaluator, and no direct use of exact transition times, equal cutoff vectors $b(\kappa) = b(\kappa')$ imply identical resolved trajectories and evaluations. One chain has at most $\binom{n+q}{q}$ cutoff classes relevant to evaluation.*

Alternatively, suppose source times depend on a shared latent parameter, $\kappa_i(\theta) = r_i + a_i^\top \theta$, with $\theta \in \Theta \subseteq \mathbb{R}^p$. Availability predicates are affine halfspaces. We call a resolver *sign-factorized* when it is path-independent and accesses θ only through the finite vector of signs of the availability predicates and the payloads it selects; the frozen evaluator likewise receives no exact transition time. The induced hyperplanes then partition Θ into relatively open faces on which every predicate, resolved state, and frozen loss is constant.

Theorem 3 (Fixed-dimensional exact evaluation)**.** *For fixed p, rational timing data and polytope Θ with B_0 inequalities, a deterministic boundary rule, a sign-factorized resolver, and frozen finitely represented losses, exact loss and paired-contrast extrema can be computed in*

$$(1 + M + B_0)^{O(p)} \operatorname{poly}(L, T_f), \tag{8}$$

where M is the number of distinct timing hyperplanes, L is the input bit length, and T_f is the cost of one frozen evaluation.

The result is polynomial for fixed p; the exponent depends on p. In contrast, the cutoff quotient above is an exact finite reduction whose size can still grow exponentially when both n and q vary.

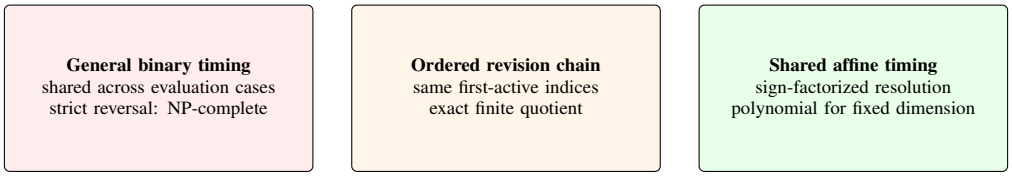

Figure 2: **Computational structure.** The general hardness construction and two exact reductions have distinct assumptions. Ordered revisions and shared affine timing are separate regimes.

4 Real-Data Audit With Incomplete Schedule Evidence

4.1 Contract, evidence, task, and frozen models

We use official ALFRED snapshots because the archive returns values as they existed on specified dates rather than retroactively replacing them with today's values (ALFRED, 2026; Croushore and Stark, 2001). The primary audit uses 154 distinct snapshots, including training and target vintages; a support-expansion audit adds every intermediate age from 0 through 30 days, for 1,180 snapshots in total. The audit covers 38 feature quarters from 2010Q1 through 2019Q2 and uses four series: real GDP (GDPC1), nonfarm payrolls (PAYEMS), CPI (CPIAUCSL), and industrial production (INDPRO).

We use these archived revisions to construct an audit with controlled uncertainty about the ingestion schedule. At each decision, 90 days after feature-quarter end, the model predicts GDP growth at feature quarter $t + 2$, approximately one quarter later. The resolver selects the newest ingested version of each feed. We define feed age as the number of days between a decision cutoff and the selected snapshot date; age zero denotes the *fresh state*. The primary contract specifies that one fixed age $d \in \{0, 7, 14, 30\}$ applies to all four feeds *and all 38 decisions*, while the value of d is unknown. Thus w_d uses the official snapshot at each decision cutoff minus d, producing four complete evaluation trajectories. The schedule restriction is imposed by the audit design; it is not inferred from ALFRED. Each candidate snapshot is dated no later than its decision cutoff. We then relax the support, cross-feed coupling, and cross-decision coupling separately. Appendix B gives the construction.

The training snapshot is fixed at 2010-02-15. We train once on 1985Q1–2008Q4, then freeze five rules: a historical mean; direct two-quarter-horizon OLS using current GDP growth alone or with three lags (AR(1) and AR(4)); ridge regression on seven macroeconomic and lagged features; and eight-nearest-neighbor regression (KNN-8) on the same features. Mature outcomes use the fixed 2026-08-21 target vintage. Replaying features from that same later vintage provides a post-hoc revision-leakage diagnostic, outside $\mathcal{W}(C, E)$. This design changes historical state while holding the model, cases, target vintage, and hyperparameters fixed. Throughout the audit, Q_f is mean squared error (MSE); RMSE is reported for predictive performance.

4.2 Predictive performance and state variation

All ten pairwise rankings are invariant across the four fixed shared-age worlds even though the inputs differ. Relative to the fresh feed, a 30-day-stale feed changes current-quarter growth features for 35/38 GDP cases, 36/38 payroll cases, and 38/38 industrial-production cases; the CPI growth

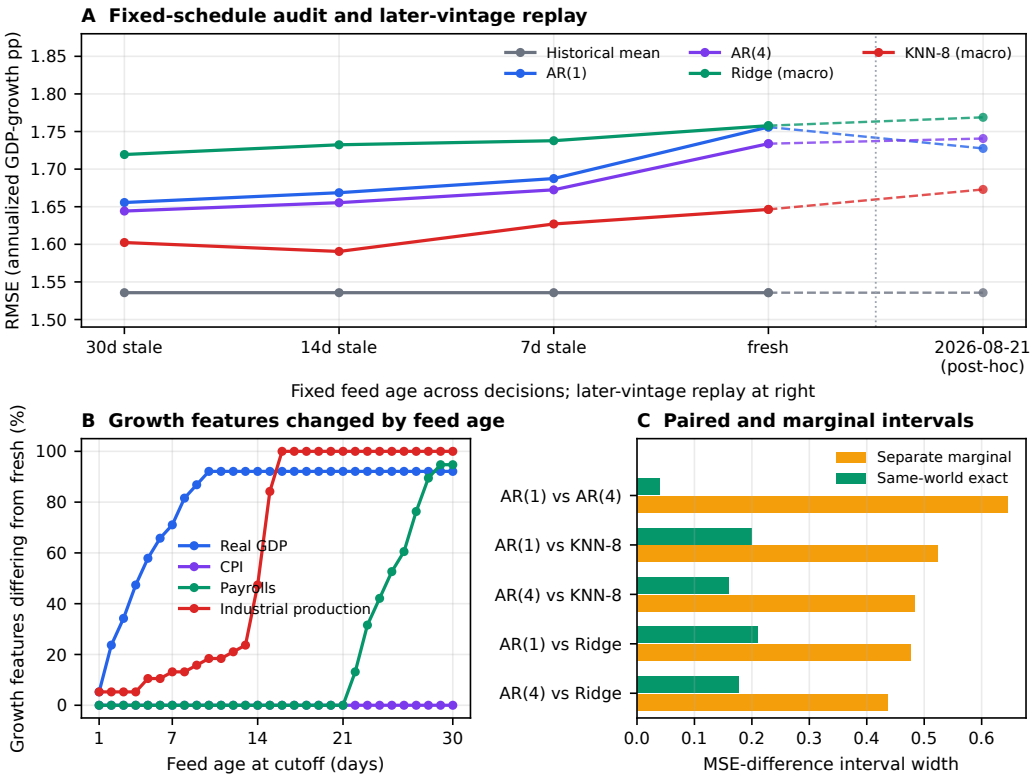

Figure 3: **A real-data audit separates state from claim uncertainty. A:** frozen-rule RMSE across four fixed feed-age schedules, followed by the separate later-vintage replay. **B:** fraction of the 38 quarterly growth values for each series that differ from the fresh state at each feed age. **C:** the five model pairs with the widest marginal intervals, comparing separate marginal and exact paired intervals over the four declared worlds.

feature is unchanged over this window. Thus, this audit identifies the rankings without identifying a unique input state. The later-vintage replay reverses AR(1) versus AR(4), illustrating the effect of substituting revised features for admissible historical inputs.

Table 1: **Predictive baselines at the fresh state and 2026-08-21 replay.** The 95% interval is a percentile interval from 2,000 circular moving-block bootstrap draws (four-quarter blocks) for the frozen model at the fresh state. It quantifies evaluation-sample uncertainty conditional on the fitted model.

Rule	Fresh RMSE	95% block-bootstrap CI	Later RMSE
Historical mean	1.536	[1.156, 1.895]	1.536
AR(1)	1.756	[1.217, 2.349]	1.728
AR(4)	1.734	[1.216, 2.300]	1.741
Ridge (macro)	1.758	[1.397, 2.145]	1.769
KNN-8 (macro)	1.646	[1.233, 2.049]	1.673

4.3 Evaluator baselines: performance and efficiency

We compare three evaluation procedures. A point reconstruction uses only the fresh state. Separate marginal evaluation combines the bounds of each model's loss range. Same-world exact evaluation computes both losses in each world and optimizes their difference over worlds. For AR(1) versus ridge, the point contrast is -0.006 MSE units and the exact admissible interval is $[-0.217, -0.006]$,

Table 2: **Evaluation-method comparison for AR(1) versus ridge.** Runtime is median prediction-plus-aggregation time from resolved states over 11 batches of 400 runs in the recorded environment. Both set-valued methods make the same model calls.

Method	Worlds	Model calls	Runtime (ms)	Contrast interval	Conclusion
Fresh-state point	1	2	1.71	-0.006	AR(1) wins
Separate marginal	4	8	6.55	$[-0.349, 0.127]$	unresolved
Same-world exact	4	8	6.55	$[-0.217, -0.006]$	AR(1) wins

Table 3: **Interaction ablation.** Feed-product worlds combine the four admissible ages separately for each feed. The table shows the two new reversals and the pair with the largest width expansion among those without a reversal; "new reversal" means the shared-age interval has one strict sign but the 256-world interval crosses zero.

Model pair	Shared width	Feed-product width	Expansion	New reversal
AR(1) vs ridge	0.210	0.253	$1.20\times$	yes
AR(4) vs KNN-8	0.160	0.394	$2.47\times$	yes
Ridge vs KNN-8	0.099	0.296	$3.00\times$	no

so AR(1) is certified better throughout $\mathcal{W}(C, E)$. Separate marginal evaluation instead reports $[-0.349, 0.127]$ and loses that conclusion by introducing impossible cross-world pairs.

Across all ten model pairs, exact same-world intervals certify every ranking, whereas separate marginal ranges conservatively leave four unresolved. The median paired-to-marginal width ratio is 0.424. For each pair, the two set-valued procedures require the same four state resolutions and eight model-world calls. Pairing preserves the dependence induced by evaluating both models in the same world, yielding tighter bounds with the same predictive workload; Proposition 1 gives the containment guarantee.

4.4 Cross-source interaction stress test

The base contract uses one age across feeds and decisions. We first relax only the cross-feed constraint, retaining one fixed age per feed across the 38 decisions. GDP, CPI, payrolls, and industrial production can each take one of the same four admissible ages. The Cartesian product contains $4^4 = 256$ worlds, each using snapshots dated no later than the corresponding decision cutoff. This feed-product support specifies admissible schedules without assigning probabilities.

Relative to the four shared-age worlds, the median contrast width expands by $1.20\times$ and the maximum by $3.00\times$. Two formerly certified comparisons become reversible: AR(1) versus ridge and AR(4) versus KNN-8. Thus cross-feed timing constraints can determine whether a ranking is identifiable.

4.5 Support sensitivity, sampling stability, and computational scaling

Removing admissible worlds cannot overturn an existing certificate because it only shrinks the claim image. To test sensitivity beyond the declared support, we add every intermediate feed age from 0 through 30 days. The declared $\{0, 7, 14, 30\}$ grid certifies all ten rankings, but the 31-age support certifies nine; AR(1) versus ridge changes from $[-0.217, -0.006]$ to $[-0.253, 0.020]$. Separate marginal evaluation certifies five rankings on the daily grid. The reversal first appears when days 1–6 are added: stability on a sparse support does not extend to intermediate archive states. Additional retained results cover the 15 subset-deletion settings and product-support sensitivity.

Relaxing cross-decision coupling. We next relax the constraint that ages remain fixed across decisions. If each decision has its own shared feed age from $\{0, 7, 14, 30\}$, the support contains 4^{38} trajectories. Because the frozen predictors act on each case separately and MSE is additive, exact bounds require no trajectory enumeration: minimize or maximize each case's paired loss difference over its four ages, then average. Only four of ten rankings remain certified. AR(1) versus ridge widens to $[-0.390, 0.165]$, and six pairs admit strict reversals. The daily shared-feed and four-age

feed-product supports also retain four rankings when timing may vary by decision (Appendix C). The contract must therefore specify whether ages can vary across decisions, as well as which ages are allowed.

Sampling stability. We assess sampling stability using 2,000 circular moving-block draws. Within each draw, we use the same sampled quarter indices for every model in all four fixed-schedule worlds. For block lengths 2, 4, and 8, the median fraction of draws retaining the full-sample sign across every world is respectively 70.6%, 70.5%, and 75.1% over the ten pairs. At the declared four-quarter length the pairwise range is 50.9%–95.5%. These bootstrap frequencies describe resampling stability, while historical certificates concern the fixed observed panel. They are not posterior probabilities that a model is better.

Finally, we measure the cost of exact evaluation across all 15 nonempty subsets of the four separately timed feeds. At one, two, and three active feeds, unique resolved-state counts range respectively over 1–4, 2–16, and 8–32; the four-feed product has 256 nominal worlds and 32 unique states. Its model calls fall from 1,280 to 160 and end-to-end median runtime from 3657.2 to 475.1 ms (7.70×). The quotient timing includes exact byte-level class discovery, feature assembly, prediction, and loss aggregation, and a per-model equality check preserves nominal-world multiplicity. This finite resolved-state quotient demonstrates reuse across identical feature matrices; the affine-face theorem instead identifies states through timing geometry.

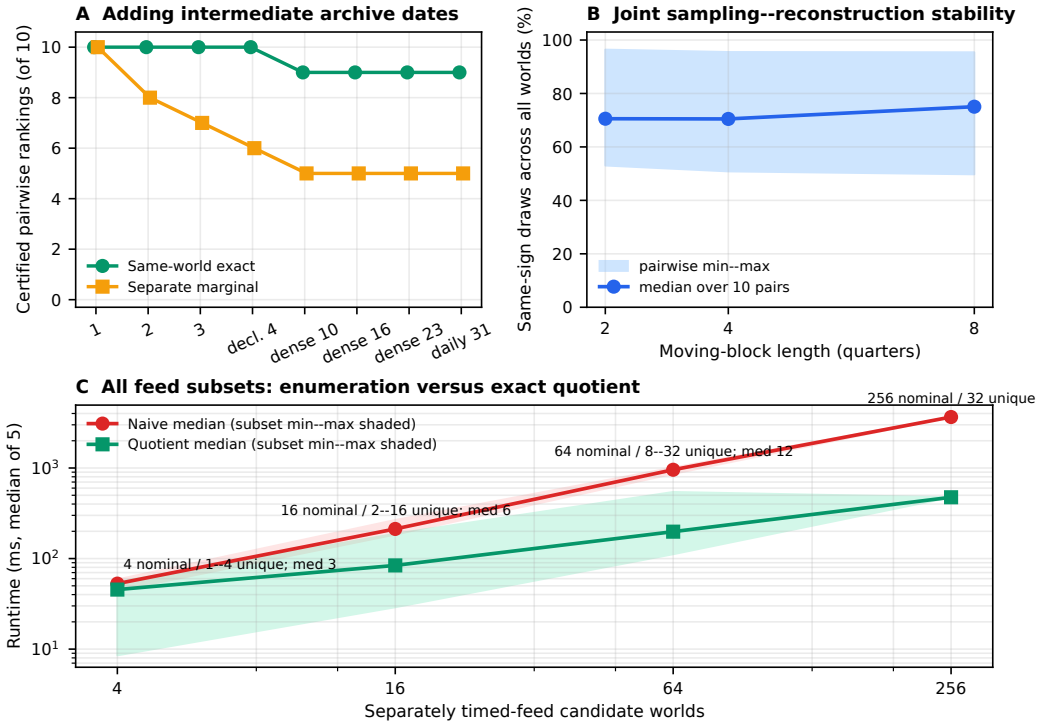

Figure 4: **Robustness across evidence, sampling, and scale. A:** nested supports from a fresh point through the declared grid and the daily 0–30-day support; the latter exposes one same-world reversal. **B:** across ten model pairs, the median and range of the fractions of moving-block draws whose paired intervals retain the full-sample sign. **C:** median runtime with subset min–max bands over all 15 nonempty feed subsets; class discovery is timed.

5 Related Work

Leakage and real-time data. Data leakage introduces information that would not legitimately be available at prediction time (Kaufman et al., 2012); leakage has also produced documented reproducibility failures across ML-based science (Kapoor and Narayanan, 2023). Timeline violations

can change model rankings in offline recommender evaluation (Ji et al., 2023), while a concurrent study of stateful data discovery shows that freezing a corpus without its interaction memory still creates hindsight leakage (Zhou et al., 2026). Real-time macroeconomic evaluation has long shown that revised data can change forecasts and model comparisons (Croushore and Stark, 2001; Stark and Croushore, 2002; Pfarrhofer, 2024). We address the uncertainty that remains before any single replay can be justified: surviving evidence may leave several protocol-consistent states admissible, so we characterize the image of a claim, prove complexity boundaries, and separate reconstruction from sampling uncertainty.

Temporal infrastructure and versioned storage. Event-time systems use watermarks and triggers to manage late, out-of-order records (Akidau et al., 2015), while versioned table systems retain transactional snapshots and time travel (Armbrust et al., 2020). These mechanisms can make a past state retrievable when the relevant logs and versions were retained. Our problem begins when the decision contract survives but the exact ingestion or reconciliation event does not. We therefore treat temporal infrastructure as a source of evidence E, not as a substitute for reasoning over the states that E still leaves possible.

Possible worlds and partial identification. Possible-world semantics provide the database foundation for incomplete states (Imieliński and Lipski, 1984); Karlaš et al. (2021) study certain predictions over incomplete training data. Classical partial identification separates an identified set from sampling uncertainty (Manski, 2003), while performance partial identification studies set-valued metrics under weak labels (Polo et al., 2024). Here the uncertain object is the historical arrival/revision state, the models and mature outcomes are frozen, and pairwise ranking is a primary target. The cutoff-specific hardness and temporal quotient results arise from that object.

Frozen neural evaluators. The contract acts at the evaluator's input interface. Transformers (Vaswani et al., 2017), diffusion Transformers (Peebles and Xie, 2023), and other models can be evaluated on each admissible resolved input, with weights and inference randomness fixed. Reusing identical inputs is architecture-independent; the cost of a forward pass determines its practical value. The present experiments establish the interface and exact reuse with small predictors.

6 Limitations

Evidence and temporal scope. Certification requires the realized history to lie in the admissible set and the resolver to match the decision process. The audit uses real archived values with controlled schedule ambiguity. Its four-age, daily-age, feed-product, and decision-varying supports expose different consequences of temporal assumptions; none covers undocumented manual overrides or sub-day events. Transactional coupling, rollback, and path-dependent inference require richer world and resolver definitions.

Empirical scope. The evidence comprises four U.S. macroeconomic feeds, 38 pre-pandemic decisions, five small frozen rules, and one target vintage. It establishes concrete identification failures and exact reuse opportunities, while leaving their prevalence across domains and larger models open. Target values may be revised after 2026-08-21. The block bootstrap conditions on the fitted models and measures evaluation-sample variability; training and distribution-shift uncertainty require separate analyses.

Computational scope. Worst-case hardness concerns shared timing choices across evaluation cases. Ordered revisions compress timestamp histories but may retain exponentially many cutoff classes, and affine evaluation is polynomial only at fixed timing dimension. Decision-varying timing is tractable in our additional audit because its product support and casewise additive loss permit separable optimization; this simplification does not hold for arbitrary path-dependent evaluators.

7 Conclusion

A valid decision protocol can leave several historical states consistent with the surviving evidence. Historical evaluation should therefore report the admissible values of its evaluation claim. Strict

winner reversal is NP-complete in a restricted setting with frozen linear predictors, while ordered revisions, fixed-dimensional timing geometry, and casewise separability provide exact reductions under their respective assumptions. In the ALFRED audit, ten rankings are certified under four fixed shared-age schedules, nine under daily fixed ages, eight under fixed feed-specific ages, and four when ages vary by decision. Reusing identical resolved states cuts model calls eightfold in the fixed feed-product audit. A point result becomes interpretable when accompanied by its paired interval and the temporal constraints on which that interval depends.

Reproducibility Statement

The code and reproducibility materials are available at `https://anonymous.4open.science/r/historical-evaluation-reproducibility-E408/`. They support a one-command cached-data rerun with 1,180 raw ALFRED payloads and their URL/byte-count/SHA-256 provenance, predictions, interaction/support/bootstrap/scaling outputs, environment metadata, and build instructions. Models and target vintage are frozen; the same seed and fitted models are used across worlds; the normalized manifest digest is recorded in `summary.json`.

AI Use Statement

AI tools were used to polish the language and improve the readability of the manuscript. The authors take full responsibility for the content of the paper.

Ethics Statement

The study uses public aggregate BEA, BLS, and Federal Reserve series and no human-subject, personal, or credentialed data. Source URLs and provenance are retained. The main risk is using an incomplete contract to overstate certification, so coverage assumptions and failure conditions are explicit.

References

Tyler Akidau, Robert Bradshaw, Craig Chambers, Slava Chernyak, Rafael J. Fernández-Moctezuma, Reuven Lax, Sam McVeety, Daniel Mills, Frances Perry, Eric Schmidt, and Sam Whittle. The dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *PVLDB*, 8(12):1792–1803, 2015.

Federal Reserve Bank of St. Louis. Archival Federal Reserve Economic Data: Download data by vintage date. `https://alfred.stlouisfed.org`, accessed 2026-09-24.

Michael Armbrust, Tathagata Das, Liwen Sun, Burak Yavuz, Shixiong Zhu, Mukul Murthy, Joseph Torres, Herman van Hovell, Adrian Ionescu, Alicja Łuszczak, Michał Świtakowski, Michał Szafrański, Xiao Li, Takuya Ueshin, Mostafa Mokhtar, Peter Boncz, Ali Ghodsi, Sameer Paranjpye, Pieter Senster, Reynold Xin, and Matei Zaharia. Delta Lake: High-performance ACID table storage over cloud object stores. *PVLDB*, 13(12):3411–3424, 2020.

Dean Croushore and Tom Stark. A real-time data set for macroeconomists. *Journal of Econometrics*, 105(1):111–130, 2001.

Herbert Edelsbrunner. *Algorithms in Combinatorial Geometry*. Springer, 1987.

Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

Tomasz Imieliński and Witold Lipski, Jr. Incomplete information in relational databases. *Journal of the ACM*, 31(4):761–791, 1984.

Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems*, 41(3):75:1–75:27, 2023.

Bojan Karlaš, Peng Li, Renzhi Wu, Nezihe Merve Gürel, Xu Chu, Wentao Wu, and Ce Zhang. Nearest neighbor classifiers over incomplete information: From certain answers to certain predictions. *PVLDB*, 14(3):255–267, 2021.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), Article 15, 2012.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

William Peebles and Saining Xie. Scalable diffusion models with Transformers. In *ICCV*, 2023.

Michael Pfarrhofer. Forecasts with Bayesian vector autoregressions under real time conditions. *Journal of Forecasting*, 43(3):771–801, 2024.

Felipe Maia Polo, Subha Maity, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Weak supervision performance evaluation via partial identification. In *NeurIPS*, 2024.

Alexander Schrijver. *Theory of Linear and Integer Programming*. Wiley, 1986.

Tom Stark and Dean Croushore. Forecasting with a real-time data set for macroeconomists. *Journal of Macroeconomics*, 24(4):507–531, 2002.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Thomas Zaslavsky. Facing up to arrangements: Face-count formulas for partitions of space by hyperplanes. *Memoirs of the American Mathematical Society*, 1(154), 1975.

Yixi Zhou, Fan Zhang, Sikun Wang, Yingfan Xu, and Haipeng Zhang. A historical corpus is not a historical system: Auditing hindsight leakage in stateful data discovery. arXiv:2609.12766, 2026.

A PROOFS

A.1 PROOF OF PROPOSITION 1

For every w, $Q_A(w) \in I_A$ and $Q_B(w) \in I_B$. Hence $Q_A(w) - Q_B(w)$ lies in the Minkowski difference $I_A - I_B = I_{\mathrm{marg}}$. Taking the infimum and supremum over same-world contrasts yields $I_{\mathrm{pair}} \subseteq I_{\mathrm{marg}}$. Suppose a world w_- attains the paired lower endpoint. Equality with the lower marginal endpoint then forces $Q_A(w_-) = \underline{Q}_A$ and $Q_B(w_-) = \overline{Q}_B$; otherwise one of the two defining inequalities is strict. The upper endpoint is analogous. Thus, if no single world jointly attains the relevant marginal extrema, the attained paired bound is strictly tighter than the corresponding marginal bound. Finite $\mathcal{W}(C, E)$ guarantees paired attainment. Without that condition, an unattained paired infimum or supremum can equal a marginal endpoint in the limit, which is why the stated strictness guarantee assumes attainment.

A.2 PROOF OF THEOREM 1

Given an X3C instance $(U, \mathcal{S})$, construct its incidence matrix A. Every $a \in \{0,1\}^m$ is reachable by the independent binary timing choices. For evaluation case j, source i contributes binary payload $A_{ji}a_i$ and the fixed predictor sums the source coordinates, giving prediction $(Aa)_j$ against a target of one. Then $L_f = \|Aa - \mathbf{1}\|_2^2 = 0$ if and only if the chosen sets cover each element exactly once. Because each chosen set contains three elements and $|U| = 3q$, an exact cover selects exactly q sets. The construction is polynomial and uses only binary payloads, unit weights, unit targets, and binary timing values.

Let g use one always-available feature with weight $1 - 1/(2N)$, so every one of its N predictions has error $-1/(2N)$ and $L_g = 1/(4N) \in (0, 1)$. In a YES instance, an exact-cover world has $L_f = 0 < L_g$. In a NO instance, $Aa - \mathbf{1}$ is a nonzero integer vector for every a, so $L_f \geq 1 > L_g$. The all-unavailable world always has $L_f = N > L_g$. Thus opposite strict winners exist if and only if X3C is YES. Two binary worlds form a polynomial certificate, and their rational losses are exactly evaluable in polynomial time, proving NP-completeness. Exact lower-bound computation would decide whether $\min_w L_f = 0$ using only the $0/1$ incidence construction, so the optimization problem is strongly NP-hard (Garey and Johnson, 1979).

A.3 PROOF OF THEOREM 2

At decision t_j, the active version index is $k_j = |\{r : b_r \leq j\}|$. Equal cutoff vectors therefore induce the same active version at every decision. Deterministic path-independent resolution and the factorization $Q_f = q_f(R_C^{\mathrm{eval}})$ then give identical resolved trajectories and losses. The weakly increasing vector $1 \leq b_1 \leq \cdots \leq b_q \leq n + 1$ has $\binom{n+q}{q}$ possible values by combinations with repetition.

A.4 PROOF OF THEOREM 3

Let $N = 1 + M + B_0$ and collect the M timing hyperplanes together with the supporting hyperplanes of the B_0 inequalities defining Θ. A standard incremental arrangement algorithm enumerates every nonempty relatively open face of these hyperplanes restricted to Θ (Edelsbrunner, 1987). An arrangement of N hyperplanes in $\mathbb{R}^p$ has $N^{O(p)}$ faces of all dimensions (Zaslavsky, 1975); for fixed p, the incidence structure and sign vector of every face can be produced in $N^{O(p)}\,\mathrm{poly}(L)$ time.

For completeness, a face sign vector partitions its defining affine forms into H_0, H_+, H_-. After clearing rational denominators, obtain a relative-interior representative by solving the rational linear program that imposes $h(\theta) = 0$ for $h \in H_0$, $h(\theta) \geq \epsilon$ for $h \in H_+$, $h(\theta) \leq -\epsilon$ for $h \in H_-$, the inequalities of Θ, and $0 \leq \epsilon \leq 1$, then maximizes ϵ. Signs incompatible with Θ are discarded. For a nonempty reported face, either both strict-sign sets H_+ and H_- are empty or the optimum has $\epsilon > 0$; standard rational linear programming returns a representative with bit length polynomial in L (Schrijver, 1986). This construction includes zero-dimensional faces and faces on $\partial\Theta$.

The sign of each availability predicate is constant on each relatively open face. Equality is retained as a third sign, so the declared deterministic boundary/tie rule fixes availability there as well. By sign

factorization, the resolver therefore returns one resolved state throughout that face, and the frozen evaluator returns one exactly comparable loss without accessing the exact value of θ. Conversely, every $\theta \in \Theta$ belongs to exactly one enumerated face. Evaluating one representative per face consequently visits every attainable loss, including losses attained only on boundaries; taking minima and maxima is exact. For paired contrasts, evaluate both frozen models at the same representative and subtract their losses. Multiplying the $N^{O(p)}$ faces by polynomial-size linear programming and evaluator cost gives $N^{O(p)}\,\mathrm{poly}(L, T_f)$. Here T_f includes exact evaluation and comparison of the finitely represented loss. This proves the claim.

B Experimental Details

Levels are converted to annualized quarterly log growth, $400\log(x_t/x_{t-1})$. Monthly series use the last observation in each quarter. Ridge minimizes summed squared error plus $4\|\beta\|_2^2$ after standardization, with an unpenalized intercept. KNN-8 averages the targets of the eight nearest training cases in standardized Euclidean distance. The AR(1) and AR(4) rules are direct two-quarter-horizon OLS regressions using current GDP growth alone or with three lags; they are not iterated one-step forecasts. The historical mean ignores features and supplies an invariant control. All preprocessing statistics are fitted on the training snapshot. For a stochastic evaluator, the random seed would also be part of the frozen evaluation configuration.

For every feature quarter t, the decision cutoff is quarter end plus 90 days and the mature target is real-GDP growth at $t + 2$ from the fixed target vintage; the horizon is measured from the feature quarter, not from the later decision date. The controlled schedule specification requires one shared feed age $d \in \{0, 7, 14, 30\}$ across all feeds and decisions; the corresponding ALFRED vintage is cutoff minus that age. The earliest selected snapshot is therefore dated 60 days after feature-quarter end. All 38 cases are complete at every daily age from 0 through 30. The later-vintage replay uses 2026-08-21 and is excluded from $\mathcal{W}(C, E)$. A 2,000-draw circular moving-block bootstrap with four-quarter blocks and fixed seed 20260924 supplies conditional evaluation-sample intervals for the frozen models in Table 1.

The support audit first retains the 15 nonempty subsets of the four declared ages, then adds nested intermediate dates until all 31 daily ages are present. It recomputes paired and marginal intervals at every step. The joint bootstrap uses 2,000 draws at block lengths 2, 4, and 8; within a draw, identical quarter indices are used for every state and model. The scaling ablation evaluates all 15 nonempty subsets of GDP, CPI, payrolls, and industrial production. Its exact quotient places two timing worlds in the same class exactly when their resolved feature matrices over all 38 quarters are byte-identical, caches the five-model loss vector, and retains nominal-world multiplicity in an aggregate checked separately for every model. Timings are medians of five alternating naive/quotient passes and include class discovery, feature assembly, prediction, and loss aggregation; they exclude data retrieval and model fitting.

Table 4: Fresh-state metrics ($n = 38$); bias is mean prediction minus target. RMSE, MAE, and bias use annualized log-growth percentage points.

Rule	RMSE	MAE	Bias
Historical mean	1.536	1.183	0.293
AR(1)	1.756	1.276	0.007
AR(4)	1.734	1.283	−0.062
Ridge (macro)	1.758	1.407	0.693
KNN-8 (macro)	1.646	1.249	0.571

Table 5: Growth-feature ablation: fraction of 38 current-quarter growth values differing from the fresh feed at the same cutoff (absolute tolerance 10^{-10}).

Series	7d stale	14d stale	30d stale
Real GDP	71.1%	92.1%	92.1%
CPI	0%	0%	0%
Payrolls	0%	0%	94.7%
Industrial production	13.2%	47.4%	100%

Table 6: Support-expansion sensitivity. After the declared four-age support, intermediate dates are added rather than removed. Entries are certified pairwise rankings out of ten.

Support	Ages	Same-world	Marginal
Fresh point	1	10	10
$\{0, 7, 14, 30\}$ declared	4	10	6
Declared + daily 1–6	10	9	5
Declared + daily 1–13	16	9	5
Declared + daily 1–21	23	9	5
All daily 0–30	31	9	5

Table 7: Joint sampling–reconstruction stability. Each percentage is the fraction of 2,000 moving-block draws for which the paired interval across all four admissible worlds retains the full-sample sign.

Pair	Full-sample interval	2-quarter	4-quarter	8-quarter
AR(1) vs AR(4)	$[0.037, 0.077]$	72.7%	71.9%	75.3%
AR(1) vs KNN-8	$[0.173, 0.373]$	56.7%	56.1%	58.9%
AR(1) vs ridge	$[-0.217, -0.006]$	53.0%	50.9%	49.8%
AR(4) vs KNN-8	$[0.136, 0.295]$	53.4%	53.2%	55.4%
AR(4) vs ridge	$[-0.261, -0.084]$	58.4%	55.1%	55.6%
Historical mean vs AR(1)	$[-0.725, -0.383]$	84.0%	82.7%	86.0%
Historical mean vs AR(4)	$[-0.648, -0.345]$	81.4%	82.2%	85.0%
Historical mean vs KNN-8	$[-0.352, -0.171]$	68.4%	69.1%	74.8%
Historical mean vs ridge	$[-0.731, -0.598]$	96.4%	95.5%	95.4%
Ridge vs KNN-8	$[0.373, 0.471]$	83.7%	84.2%	85.7%

Table 8: Product-world scaling over all feed subsets. Unique states show min/median/max across subsets of the same size; calls and runtimes are naive/quotient medians. Speedup shows median [min,max]. Class discovery is timed.

Feeds	Subsets	Nominal	Unique	Calls	Runtime (ms)	Speedup
1	4	4	1/3/4	20/15	52.9/45.4	$1.12\,[0.97, 6.13]\times$
2	6	16	2/6/16	80/30	212.4/84.0	$2.54\,[1.29, 7.17]\times$
3	4	64	8/12/32	320/60	956.7/198.2	$5.10\,[1.79, 7.72]\times$
4	1	256	32/32/32	1280/160	3657.2/475.1	$7.70\times$

C Exact Bounds When Ages Vary by Decision

Let $\mathcal{D}_t$ be the finite set of allowed feed-age configurations at decision t. When $\mathcal{W} = \prod_{t=1}^{n} \mathcal{D}_t$, both predictors act on each case separately, and the loss is additive, write

$$\Delta_{AB}(w) = \frac{1}{n} \sum_{t=1}^{n} \delta_t(d_t), \qquad d_t \in \mathcal{D}_t.$$

The exact endpoints are

$$\min_{w \in \mathcal{W}} \Delta_{AB}(w) = \frac{1}{n} \sum_{t=1}^{n} \min_{d \in \mathcal{D}_t} \delta_t(d), \qquad \max_{w \in \mathcal{W}} \Delta_{AB}(w) = \frac{1}{n} \sum_{t=1}^{n} \max_{d \in \mathcal{D}_t} \delta_t(d).$$

The pointwise minima and maxima bound every admissible trajectory. Choosing a minimizing or maximizing configuration separately at each decision attains the corresponding endpoint. This uses $\sum_t |\mathcal{D}_t|$ case–configuration loss differences instead of enumerating $\prod_t |\mathcal{D}_t|$ trajectories. Cross-decision constraints or path-dependent model state can invalidate this factorization. In particular, the hardness reduction shares one availability vector across all evaluation cases, so it does not have this product structure.

Table 9: Certified rankings under different temporal constraints. A configuration is a shared feed age in the first two rows and a four-feed age tuple in the third. Fixed configurations are held across all 38 decisions.

Local configurations	Fixed worlds	Fixed certified	Varying worlds	Varying certified
Four shared ages	4	10	4^{38}	4
Daily shared ages	31	9	31^{38}	4
Four ages per feed	256	8	256^{38}	4

We compute all 30 pair–support intervals and 60 endpoint witnesses, recording the attaining age or age tuple at every quarter. We check each witness by direct substitution and check separability against all 4^3 age assignments on the first three cases for each of the ten model pairs. Four rankings remain certified: the historical mean outperforms AR(1), AR(4), and ridge, and AR(4) outperforms AR(1). The same four rankings survive all three decision-varying supports.