CORRECT SCORES, UNSUPPORTED CONCLUSIONS: BENCHMARKING EVALUATION UNDER INCOMPLETE SEMANTIC REFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

An evaluation score can be computed exactly against the available reference and still fail to support the scientific conclusion drawn from it. We formulate model comparison under incomplete semantic reference as an evidence-qualified evaluation problem: when observational proxies, current snapshots, or incomplete records substitute for the required semantic reference, what evidence is sufficient to certify which model is better? Holding predictions fixed across structured domains reveals that proxy distortion alone does not determine conclusion validity: selected-only reference drops chess forest set F_1 from 0.934 to 0.076 while leaving the model winner unchanged, yet the identical substitution reverses model selection in Planning Domain Definition Language (PDDL) tasks. This divergence establishes that reference fidelity and conclusion fidelity are distinct: exact reference identification is sufficient but not necessary if every compatible reference induces the same model comparison. We introduce a typed benchmark across public and executable sources spanning stable, reversed, recoverable, certifiable, and uncertifiable regimes, and operationalise an executable certification protocol that determines whether available evidence already certifies the comparison, target-matched recovery suffices, additional evidence must be acquired, or the comparison must remain UNKNOWN. In process panels, target-matched recovery resolves ranking conclusions despite imperfect reference reconstruction; in controlled and natural-history panels, exact query certificates determine whether evidence certifies the query, requires witnesses or exhaustion, or yields UNKNOWN. Holding outputs fixed, stronger references can preserve rankings, alter pairwise comparisons, or reverse top-system selection, with the latter observed in both formal planning and a contemporary speech-recognition panel evaluated against independently corrected transcripts. Evaluators should therefore determine whether available evidence is sufficient to certify the declared model comparison, rather than forcing an ungrounded point reference and mistaking predictive confidence for semantic warrant.

1 INTRODUCTION

A process log records that action A occurred. A frozen model accepts A and rejects candidate B. Scoring against the recorded event treats B as a true negative. A qualified rule or process model may instead state that both A and B were enabled, turning the same rejection into a false negative. Nothing about the model changed. The available record and the semantic reference required by the evaluation were different objects (Figure 1).

This paper studies the missing evaluation layer that connects source authority to compatible semantic references, an evaluation metric, and a declared scientific query or model comparison. The gap appears whenever recorded evidence is narrower than the claim-required reference: a chess move is one selected choice among legal actions; a planning step is one operator among applicable operators; and a current snapshot omits revisions needed for a historical query. Standard benchmarks often flatten these objects into one ground-truth column. The resulting score may be computed exactly against that column while the available evidence remains insufficient to warrant the conclusion drawn from it.

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.

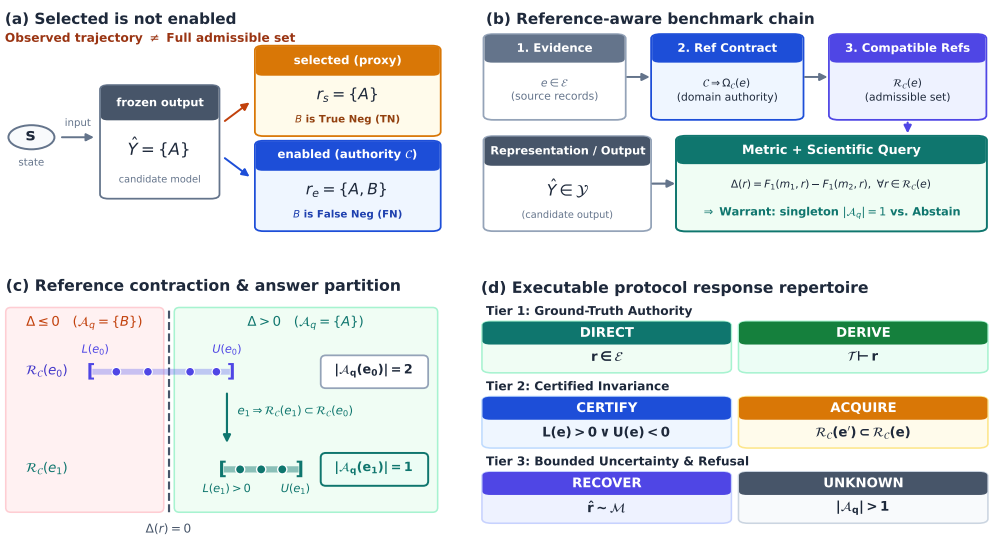

Figure 1: **From observed record to warranted conclusion.** The overview joins three stages: a frozen output scored against selected versus enabled actions; the typed chain from evidence through semantic reference and metric to a declared query; and the compatible-answer geometry in which several references can already occupy one answer region. The actions DIRECT, DERIVE, RECOVER, CERTIFY, ACQUIRE, and UNKNOWN are consequences of that diagnosis, not interchangeable processing modules.

The fundamental scientific challenge arises because evaluation intuition assumes a monotonic relationship between reference accuracy and decision validity: namely, that large proxy distortion must corrupt scientific conclusions, while high-fidelity proxies guarantee sound findings. Our empirical findings directly contradict this assumption. Holding model predictions strictly fixed, replacing the full enabled reference under the declared contract with an observational selected-only proxy produces two diametrically opposed regimes. In chess, scientific workflows, and generated Petri nets, selected-only reference causes severe absolute score distortion (e.g., set-F_1 falls by more than 0.85), yet the top winner remains unchanged and there are no material pair-order changes in these stable-control regimes. Conversely, in Planning Domain Definition Language (PDDL) tasks, the identical class of proxy substitution induces a material reversal of the scientific conclusion: a model that dominates under the full evaluator reference is rejected under the proxy. Fixed-output evaluations on contemporary machine-learning benchmarks reproduce these divergent behaviors: under strengthened references, SWE-bench Verified preserves all tested model rankings, EvalPlus alters system-pair orders while leaving the top system unchanged, and rescoring 89 speech-recognition systems against independently human-corrected transcripts reverses the top-performing system.

This divergence establishes the central scientific insight of this work:

$$\text{reference fidelity} \neq \text{conclusion fidelity}.$$

Reference fidelity measures the closeness of an available proxy to the full semantic reference admitted by the contract. Conclusion fidelity measures whether the decision induced by the proxy matches the decision under that full reference in the declared query space. Because decision boundaries in model evaluation partition the metric space into discrete answer regions, large reference errors that leave candidate models on the same side of a decision threshold are entirely conclusion-irrelevant. Conversely, even minor reference incompleteness that shifts a model comparison across the threshold $\Delta = 0$ inverts the scientific outcome.

Therefore, the critical question for structured evaluation is not *how closely does the available proxy approximate the full reference?* Rather, the central organizing question is:

> *When the evaluation reference is incomplete, what evidence is sufficient to certify which model is better?*

For a declared query, exact reference identification is sufficient but not necessary if every compatible reference induces the same answer. This query-relative perspective changes the evaluation target: rather than attempting exhaustive, costly, and often impossible reconstruction of the full reference, evaluators need only establish that the admissible reference set is contained within a single decision region. The principle reference fidelity $\neq$ conclusion fidelity explains the unified empirical story: an incomplete reference may still be sufficient for model comparison (as in chess); it may instead reverse the comparison (as in PDDL); target-matched recovery can restore the tested ranking conclusion despite imperfect reference reconstruction; additional evidence can certify comparisons before full reference identity; and some comparisons remain uncertifiable (UNKNOWN) under the admitted evidence contract.

We make three connected contributions:

1. **Evaluation task and benchmark**: We formulate model comparison under incomplete semantic reference as an evidence-qualified evaluation problem and introduce a typed benchmark over public and executable sources for incomplete semantic reference. The benchmark provides controlled regimes spanning stable comparison, comparison reversal, recoverable comparison, certifiable comparison, and uncertifiable (UNKNOWN) outcomes to study whether available evidence is sufficient to support a model comparison.
2. **Executable certification protocol**: We develop an executable protocol over compatible semantic references that determines whether available evidence already certifies the model comparison, deterministic derivation or target-matched recovery is sufficient, additional qualified evidence must be acquired, or the comparison must remain UNKNOWN under an explicit source-authority contract.
3. **Empirical characterization**: We provide an empirical characterization across structured domains and fixed-output contemporary ML evaluations, covering stable ranking, pairwise comparison changes, and top-system reversal. We demonstrate that large score distortion does not imply model-selection error (where chess, workflows, and software maintenance preserve the winning model despite score drops) and that unresolved reference variation can invert top-system selection (both in formal planning and in contemporary speech recognition), while showing that strengthened references can alter only intermediate pairwise conclusions, target-matched recovery restores ranking conclusions despite imperfect reference reconstruction, and partial evidence can certify model comparisons before the full reference is identified.

2 Related Work and Positioning

Claims, constructs, and incomplete information. Benchmark validity depends on the inference made from a score (Salaudeen et al., 2025; Freiesleben & Zezulka, 2025; Bean et al., 2025). Certain-answer and partial-identification frameworks provide the logical basis for determining a query across multiple compatible worlds (Libkin, 2016; Manski, 2003); database completeness, provenance, and temporal formalisms provide machinery for reasoning about which facts and histories support a query (Corman et al., 2024; Green et al., 2007; Kulkarni & Michels, 2012). We adapt this perspective to an executable ML-evaluation interface that connects source authority, semantic reference, metric, query, recovery, and reason-coded unsupported answers.

Missing labels, proxy evaluation, and selective risk. Selective labels, missing evaluation labels, proxy judges, and active testing can alter scores and model rankings (Lakkaraju et al., 2017; Dervovic & Cashmore, 2025; Dorner et al., 2025; Kossen et al., 2021). Missing-label statistical recovery remains appropriate when its sampling and target contract holds (Dervovic & Cashmore, 2025); selective abstention and calibrated risk control address different risk–coverage objectives (El-Yaniv & Wiener, 2010; Angelopoulos et al., 2024). Dominguez-Olmedo et al. (2025) isolate a training-side confound: unequal exposure to the test task can distort relative model evaluation. Our fixed-output design isolates the reference side instead: outputs remain unchanged while the semantic reference admitted by the evaluation contract changes. While Sevetlidis (2026) studies representation sufficiency for Bayes-optimal prediction under a specified decision problem, this paper studies whether source-qualified evidence identifies the semantic reference, or at least the evaluation conclusion, required by a scientific query. Our benchmark targets a different interface: the missing object may be a set

of enabled actions, a version history, a relationship lifecycle, or a completeness assertion whose authority cannot be replaced by predictive confidence.

Process references and imperfection benchmarks. Translucent logs explicitly distinguish executed from enabled activities (Beyel & van der Aalst, 2024; Beyel et al., 2025). MIRA separates normative, system, and recording models; FLAWD and object-centric imperfection patterns provide controlled data-quality mechanisms (Sommers et al., 2025; Comuzzi et al., 2025; Sadeghianasl et al., 2026). Inquiry-validation work records queries and results to support process-analysis validation (Zerbato et al., 2024). These works supply important benchmark families and components. Our object of evaluation is the propagation from a declared source authority to a declared conclusion, including when ordinary recovery is sufficient and when no admitted source can support the claim.

Acquisition and certification. Decision-region testing, adaptive submodular cover, feature acquisition, and implicit hitting sets provide established acquisition machinery (Golovin et al., 2010; Javdani et al., 2014; Chen et al., 2017; Cui & Nagarajan, 2023; Covert et al., 2023; Ignatiev et al., 2016). These foundations supply planning components when their assumptions hold; the exact evidence certificate remains the scientific endpoint. Our benchmark evaluates the typed reference contract and the resulting end-to-end model-comparison decision.

3 BENCHMARK

3.1 UNIT, AUTHORITY, AND DEGRADATION

A benchmark unit is $(e, \mathcal{C}, z, r, m, q)$: admitted evidence e; a versioned semantic contract $\mathcal{C}$; frozen analysis output z; the claim-required reference r; metric m; and declared scientific query q. Reference-rich panels hide evaluator truth from the tested proxy or recovery component. Natural panels use source-native limitations such as current snapshots or retained histories. Each unit records whether reference information is direct, derivable, statistically recoverable, acquirable, or unavailable under the contract.

Table 1 specifies the benchmark. To study evidence sufficiency across complementary regimes without treating every domain as an independent headline claim, the families serve defined experimental roles: Chess is the primary stable reference case, testing whether massive proxy distortion alters the model winner; workflows and generated nets provide additional stable-regime evidence; PDDL isolates model-selection reversal; process panels examine recovery before complete reference reconstruction; natural-history and certification panels evaluate evidence depth and stopping conditions; and context-support, MIRA, and GS1 isolate representation, uncertainty, and source-authority boundaries. Controlled degradation preserves evaluator-only truth; natural histories preserve source-record truth without asserting physical or normative completeness.

3.2 LAYERED REFERENCE DEPENDENCE

Reference dependence is layered rather than binary. A method may execute on recorded data while a semantic reference is still required to define its training target, evaluate its output, or license the scientific claim. The complete method-by-method matrix, including reproduction status and the admissible response to missing authority, is in Appendix D; the main benchmark therefore reports the layer and query affected instead of treating successful execution as evidence of semantic validity.

The selected-versus-enabled panels use January 2013 Lichess games and independently executed legal rules (Lichess.org, 2026), WfInstances execution records and directed acyclic graph (DAG) dependencies (Coleman et al., 2022), public PDDL domains and grounded operator schemas (McDermott et al., 1998; Plaans contributors, 2026), and PM4Py-generated Petri nets (Berti et al., 2019). Additional panels use Hospital Billing, Road Traffic Fines, and Sepsis-derived enabled sets (Mannhardt, 2017; de Leoni & Mannhardt, 2015; Mannhardt, 2016); MediaWiki revisions and Wikidata licensing (MediaWiki contributors, 2026b;a; Wikidata contributors, 2026); GitHub lifecycle events (GitHub, 2026); and MIRA controls (Sommers et al., 2025). Source cards, pins, hashes, masks, licenses, and failed runs appear in the Appendix.

Table 1: **Benchmark specification.** Each family declares its authority, truth level, evaluation unit, scale, reference substitution or degradation, and scientific query. Scale is the fixed evaluation unit, not a claim of independent population sampling.

Family	Authority	Truth level	Unit	N / scale	Proxy or degradation	Declared query
Chess	Executed rules	Rule legality	State × action	480 / 46,069	Played for legal	Model winner
Workflow	DAG dependencies	Source readiness	State × task	245 / 13,131	Started for ready	Model winner
Generated nets	Executable net	Generated model	Marking × transition	3,659 / 64,882	Fired for enabled	Model winner
PDDL	Formal operators	Model applicability	State × operator	620 / 168,310	Selected; 4 masks	Model winner
Process recovery	Published annotations	Generated enabled set	Model pair	3 families × 75	Selected; recovered	Ranking sign
Context / support	Frozen process	Generated exact	History prefix	32,000 prefixes	Context / support	Answerability
Wikidata	Revision API	Source record	Revision / query	434 / 44	Snapshot for history	Witness / invariant
GitHub	Timeline API	Source lifecycle	Issue / query	4 / 36	Snapshot for lifecycle	Membership
MIRA	Author model	Model relative	Held-out row	8,470 rows	Target / coarsening	Error / uncertainty
GS1	Standard examples	Documented events	Query cell	630 cells	Missing initial relation	Relation / UNKNOWN

4 WARRANTED EVALUATION UNDER COMPATIBLE REFERENCES

Let $\Omega_{\mathcal{C}}(e)$ be the nonempty set of worlds compatible with evidence e under contract $\mathcal{C}$. The reference required by the claim is $r_{\mathcal{C}}(x)$, so

$$\mathcal{R}_{\mathcal{C}}(e) = \{r_{\mathcal{C}}(x) : x \in \Omega_{\mathcal{C}}(e)\}. \tag{1}$$

For frozen output z, metric m, and query q, define

$$\mathcal{A}_q(e, z) = \{q(m(z, r)) : r \in \mathcal{R}_{\mathcal{C}}(e)\}. \tag{2}$$

The conclusion is warranted exactly when $|\mathcal{A}_q(e, z)| = 1$. Exact reference identification, $|\mathcal{R}_{\mathcal{C}}(e)| = 1$, is sufficient but unnecessary. This is the established certain-answer and partial-identification principle (Libkin, 2016; Manski, 2003), operationalised here as a typed evaluation interface.

For pairwise model selection, let $\Delta(r) = m_A(r) - m_B(r)$ denote the score difference between candidate models A and B evaluated under reference r. The induced score-gap interval is defined by the extremal projections over all compatible references:

$$L(e) = \inf_{r \in \mathcal{R}_{\mathcal{C}}(e)} \Delta(r), \qquad U(e) = \sup_{r \in \mathcal{R}_{\mathcal{C}}(e)} \Delta(r). \tag{3}$$

The pairwise ranking is certified if $L(e) > 0$ (Model A strictly dominates) or $U(e) < 0$ (Model B strictly dominates). Figure 1 illustrates the geometry: qualified evidence contracts the set of compatible references into a unified decision region before selecting one exact reference. The interval $[L(e), U(e)]$ is neither an arbitrary uncertainty score nor a subjective Bayesian heuristic. It is the compatible score-gap interval induced by all references still compatible with qualified evidence under contract $\mathcal{C}$; the scientific answer is obtained from its position relative to the declared decision boundary, e.g., $\Delta = 0$.

This formulation explains why proxy/reference distortion magnitude alone does not determine conclusion stability. Standard error analysis implicitly presumes that smaller distortion guarantees conclusion fidelity and large distortion invalidates it. However, scientific conclusions in comparative evaluation depend on topological separation from the decision boundary $\Delta = 0$. The interval width $U(e) - L(e)$ quantifies the extent of unresolved semantic ambiguity, but its position relative to zero determines whether that ambiguity propagates into the scientific finding. If the entire interval lies strictly on one side of zero ($L(e) > 0$ or $U(e) < 0$), unresolved reference variation is completely absorbed by the query geometry: every admissible world agrees on the superior model, rendering large semantic distortions conclusion-irrelevant. If it contains zero, the ranking is not certified, because the qualified evidence permits equally valid semantic references that produce opposite model rankings. In that situation, any point proxy that confidently selects a winner is an ungrounded artifact of arbitrary reference completion.

Representation remains a separate layer. For a fixed weighted table, representation ϕ partitions examples into cells g. For target y and total weight W, define

$$A(\phi, y) = \frac{1}{W} \sum_{g} \left[\sum_{i \in g} w_i - \max_{c} \sum_{i \in g : y_i = c} w_i \right]. \tag{4}$$

Every deterministic predictor using only ϕ incurs at least this weighted table error, and a refinement cannot increase the quantity. This expresses empirical conditional-risk accounting for target-relevant distinction collapse. A singular linear map $Z = RX$ with $\ker R \neq \{0\}$ provides an intuitive analog: non-invertibility matters when the lost direction changes y or q.

Executable protocol. The evaluator declares $(\mathcal{C}, r, m, q)$ before inspecting the hidden reference. It uses direct source facts or deterministic derivation when available, reports assumption-labelled recovery as an estimate, and checks Equation 2. If answers still differ and qualified evidence is reachable, it acquires evidence and rechecks the certificate. Predictions may order requests but never enter the hard certificate set. If no admitted evidence can separate opposite answers, the protocol returns a reason-coded UNKNOWN. For panels that include abstention or acquisition, we report warrant coverage, certified-query error, and source-native acquisition cost together. These quantities form a panel-specific reporting tuple rather than a pooled score because requests, bytes, revisions, and expert checks are not commensurate cost units.

Across the reported panels, certification uses finite enumeration when compatible references are explicit, analytic or structural bounds where available, and exact-state solving when tractable. Sound outer bounds can certify a query when every bounded answer lies in one decision region, without enumerating or reconstructing every exact reference. The criterion itself does not supply a universally efficient solver for arbitrary implicit reference spaces; scalable certification there is a separate algorithmic problem.

5 RESULTS: WHEN DOES MISSING REFERENCE MATTER?

5.1 SCORE DISTORTION DOES NOT DETERMINE CONCLUSION DISTORTION

Table 2 gives the exact negative and positive regimes for the same frozen outputs. We first ask: *can severe proxy distortion leave model comparison unchanged?* The naive expectation that proxy distortion scales monotonically into conclusion error is refuted by the stable controls. As our primary stable reference case, Chess forest F_1 falls from 0.933792 under legal-action reference to 0.076419 under the selected-only proxy, yet forest remains the winner. Workflow and generated-net scores also shift sharply without changing the winner (macro F_1 drops to 0.5135 and 0.8217, with expressive models tied at rank 1), generating zero material pairwise ranking changes across all three stable domains. Consequently, severe score distortion alone is insufficient to invalidate model comparison, prompting the second question: *when does incomplete reference reverse model comparison?* The complementary regime appears in PDDL: at 25% hiding, forest full/proxy F_1 is $0.703642/0.130485$, while prior full/proxy F_1 is $0.071003/0.196774$. Full reference selects forest; the proxy selects prior. Here, proxy substitution reverses three out of six model pairs, converting a dominant win for the Random Forest into an ungrounded victory for the naive frequency prior. This contrast forces the sharper scientific question: *what causes unresolved reference variation to cross the declared answer boundary?*

For the primary PDDL winner, the reversal occurs in all frozen problem folds at every nonzero mask and also in one unmasked fold; partial observability is therefore not necessary for the phenomenon. Alternative metrics and an exact local analysis further rule out raw missingness and omitted-action prediction disagreement as governing explanations (Appendix E). Once these explanations are separated, the compact object is the compatible score-gap interval: stable controls remain within one answer region, whereas the PDDL proxy and full reference select opposite sides of the declared ranking boundary.

Table 3 extends this progression to contemporary fixed-output evaluation. SWE-bench supplies a stable software-maintenance case (OpenAI, 2024); EvalPlus changes system-pair orders without changing the winner (Liu et al., 2023); and VoxPopuli ASR provides an independently governed

Table 2: **Fixed outputs under full and selected-only reference.** Each model pair reports full/proxy macro state-level set-F_1; bold marks the winner under that reference. Strict and material columns count changed pair orders out of six, where material changes exceed the prospective 0.01 margin threshold. "F/L/T" denotes a three-way tie among forest, linear, and tree. "Invariant" records local pair-order reversals persisting under identical omitted-action prediction sets (Appendix E), demonstrating that ranking inversion stems from reference incompleteness rather than unobserved action errors. Exact finite panels, not independent population replicates.

	Forest		Linear		Tree		Prior		Winner		Strict	Material	Invariant
Family / mask	**Full**	**Proxy**	**Full**	**Proxy**	**Full**	**Proxy**	**Full**	**Proxy**	**Full**	**Proxy**	**/6**	**/6**	**pairs**
Chess	**.9338**	**.0764**	.7872	.0439	.7801	.0451	.4754	.0355	F	F	1	0	12
Workflow	**1.0000**	**.5135**	**1.0000**	**.5135**	**1.0000**	**.5135**	.3111	.0766	F/L/T	F/L/T	0	0	0
Generated	**1.0000**	**.8217**	**1.0000**	**.8217**	**1.0000**	**.8217**	.0000	.0000	F/L/T	F/L/T	0	0	0
PDDL / 0%	**1.0000**	**.1972**	**1.0000**	**.1972**	**1.0000**	**.1972**	.0710	.1968	F/L/T	F/L/T	0	0	0
PDDL / 25%	**.7036**	.1305	.6352	.0795	.6191	.0745	.0710	**.1968**	F	P	3	3	76
PDDL / 50%	**.4980**	.0864	.4712	.0470	.4544	.0435	.0710	**.1968**	F	P	3	3	120
PDDL / 75%	.3409	.0604	**.3519**	.0337	.3462	.0306	.0710	**.1968**	L	P	5	4	165

Table 3: **Contemporary fixed-output external validation.** The same released outputs are scored under the weaker and qualified references. Mean Δ is qualified minus proxy score; pass@1 are higher-is-better, whereas word error rate (WER) is lower-is-better. Changed orders report strict sign reversals, with tie-boundary changes in parentheses. Full statistics, top-system scores, contracts, and robustness checks appear in Appendix Tables 8–10.

Panel: proxy → qualified	**Scale**	**Mean Δ**	**Pair changes**	**Top changed**	**Regime**
HumanEval → HumanEval+	89 systems; 3,916 pairs	-5.42 pp	76 (54 ties)	No	Pairwise-sensitive
MBPP → MBPP+	89 systems; 3,916 pairs	-11.29 pp	125 (32 ties)	No	Pairwise-sensitive
SWE-bench repair-only → full	5 systems; 2,500 evaluations	-3.24 pp	0/10	No	Stable
VoxPopuli original → corrected	89 systems; 628 clips	-2.99 WER pp	1,268 (17 ties)	**Yes**	**Top reversal**

top-system reversal (Wang et al., 2021; Srivastav et al., 2025; Artificial Analysis, 2026). Under human-corrected transcripts, the lowest-WER system changes from `nvidia/parakeet-rnnt-1.1b` (4.68% WER) to `mistralai/Voxtral-Small-24B-2507` (2.00% WER) across 89 public models. The reversal persists in all 10 contiguous leave-one-block checks and under plain word-level WER. Together, these outcomes show that the consequence of reference change is query-dependent rather than monotonic in score distortion, leading directly to the third question: *must the full semantic reference be completely reconstructed to resolve a comparative ranking?*

5.2 Target-matched recovery need not reconstruct the reference

If conclusion reversals arise from reference incompleteness, one might assume sound evaluation requires exhaustive semantic reconstruction. In three held-out process panels, selected-only scoring changes 51/75, 52/75, and 49/75 pairwise signs for history-count representations. A target-matched random forest trained on separate cases leaves 7/75, 4/75, and 0/75 strict disagreements; every remaining disagreement has a full-reference margin below 0.1 percentage point, while reference fidelity reaches 0.9912, 0.9755, and 0.9988. Exact semantic reconstruction is therefore stronger than this pairwise-ranking query requires (strict sign is the prospectively specified recovery endpoint; 0.1 percentage point provides descriptive sensitivity). Thus reference fidelity and conclusion fidelity are distinct. Unlike Chess, where proxy distortion leaves the winner intact without recovery, process and planning domains require addressing reference ambiguity; yet process recovery shows that target-matched recovery restores the tested ranking conclusion despite imperfect reference reconstruction.

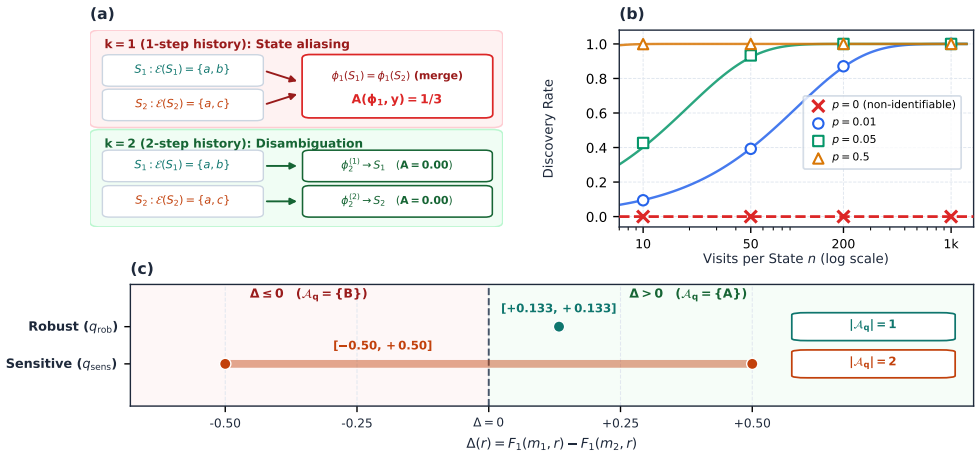

Figure 2: **Mechanisms governing semantic reference ambiguity and certification.** **(a)** State aliasing: latent states S_1, S_2 with enabled sets $\{a, b\}$ and $\{a, c\}$ alias at memory depth $k = 1$, yielding exact representation ambiguity $A(\phi_1, y) = 1/3$; extending history to $k = 2$ separates representations and resolves state identity ($A(\phi_2, y) = 0.00$). **(b)** Selection exposure boundary: empirical discovery rates across 2,000 generated units match $1-(1-p)^n$; zero support ($p = 0$) forms a strict non-identifiability boundary where unseen actions persist indefinitely. **(c)** Query-relative certification geometry: at depth $k = 1$, the robust query comparison interval $[+0.133, +0.133]$ lies entirely in the positive half-space ($|\mathcal{A}_q| = 1$); conversely, the sensitive query interval $[-0.50, +0.50]$ straddles ranking boundary $\Delta = 0$ ($|\mathcal{A}_q| = 2$), demanding context refinement.

This poses the fourth question: *when is partial evidence already sufficient to certify the comparison, and what governs the persistence of reference ambiguity?*

5.3 Context, Support, and Query Geometry Determine Ambiguity

Controlled sequential dynamics isolate how compatible reference sets $\mathcal{R}_C(e)$ arise through three structural mechanisms: representational aliasing, behavioural selection support, and query sensitivity.

Let $A_{t+1} \sim \pi(\cdot \mid S_t)$ and $Z_k = \phi_k(H_t)$ denote the policy and length-k history representation. First, representational coarsening merges environmental states with distinct enabled sets $\{a, b\}$ and $\{a, c\}$ (Figure 2a): under one-step history ($k = 1$), state aliasing leaves irreducible reference ambiguity, bounding exact representation ambiguity from below at $1/3$, whereas extending memory to $k = 2$ resolves latent histories and reduces ambiguity to zero.

Second, selection support forms an empirical observability boundary (Figure 2b): an enabled alternative chosen with probability p is observed after n visits with probability $1 - (1 - p)^n$. Across 2,000 generated units per cell, observed discovery matches this finite-exposure relation within frozen 0.035 tolerance. At $p = 0$, the unchosen alternative remains unseen through 1,000 visits, creating an observational non-identifiability boundary where predictive entropy is zero while reference entropy is one bit (certainty about selected actions provides zero information about enabled actions; a reverse control confirms predictive and reference uncertainties are decoupled). Third, while Chess illustrates that severe score distortion need not change the winner, the controlled interval experiment establishes the decision geometry enabling partial evidence to certify model comparison (Figure 2c). Under depth $k = 1$, the robust comparison interval $[0.133, 0.133]$ lies strictly in the positive half-space, yielding certified singleton answer set $|\mathcal{A}_q| = 1$ despite incomplete reference identity. Conversely, the sensitive comparison spans $[-0.5, 0.5]$ and straddles ranking threshold $\Delta = 0$, producing $|\mathcal{A}_q| = 2$ and requiring context refinement: evaluation warrant is thus query-relative rather than reference-exhaustive.

5.4 PREDICTIVE CONFIDENCE CANNOT SUBSTITUTE FOR EVIDENCE

The MIRA control crosses training target with representation coarsening. Under selected-only training, 1,037 of 6,570 low-entropy held-out rows are wrong for the semantic target, showing in this finite panel that predictive confidence does not establish reference validity. Under full coarsening, predictive entropy increases and detects 18.5714% of target loss, rejecting the claim that predictive uncertainty is always blind. Equation 4 diagnoses distinctions the representation cannot express, Equation 1 diagnoses possible semantic targets, and Equation 2 tests whether either uncertainty alters the conclusion.

Abstention must also be evaluated with coverage: on fully masked MIRA and Sepsis controls, forced maximum a posteriori answers match evaluator truth on 15/15 queries each, yet all 30 assertions are unsupported at decision time. In GS1, 225/630 relation queries are source-eligible and 405 are source-unknown; removing UNKNOWN creates 405 unsupported assertions. Correctness after disclosure is not evidence-scoped warrant, raising the fifth question: *when must additional evidence be acquired, and when must the comparison remain* UNKNOWN*?*

5.5 AUTHORITY AND QUERY FORM DETERMINE THE ADMISSIBLE ACTION

Natural revision history supplies an independent certificate geometry: across 434 reachable Wikidata revisions and 44 frozen queries, a current-snapshot proxy is wrong 5 times. Each positive existential claim admits a qualified witness; all 39 negative interval invariants require complete admitted coverage. Evidence can stop early for one query form while another requires source exhaustion. The complementary boundary is structural absence of authority: GitHub lifecycle replay confirms current snapshots can misstate historical membership, while inaccessible history remains outside the contract (Appendix L). Together, these results implement the action boundary: direct or derived references support exact use, recoverable references can be query-sufficient without exact reconstruction, acquisition shrinks compatible references when qualified evidence exists, and UNKNOWN is warranted when competing answers cannot be separated, yielding the exact query certificate as the scientific endpoint.

6 DISCUSSION AND LIMITATIONS

The benchmark connects three mechanisms. **Source authority** defines compatible worlds and distinguishes unobserved from impossible behavior. **Context, support, and representation** determine which distinctions evidence can expose. **Query geometry** projects compatible references onto $[L(e), U(e)]$ and certifies superiority when the interval lies on one side of $\Delta = 0$. The contemporary panels show that the resulting stable, pairwise-sensitive, and top-reversal regimes also occur in modern fixed-output pipelines. Evaluators can therefore declare the contract, characterize compatible references, and choose direct use, derivation, recovery, certification, acquisition, or UNKNOWN without requiring exhaustive reconstruction.

Limitations. Warrant is conditional on domain-specific evidence contracts. Wikidata and GitHub expose source records rather than unrecorded physical reality; controlled masks do not model deployed missingness prevalence; and process results use case-hash splits within finite logs rather than population draws. VoxPopuli establishes a contemporary top-system reversal under independently corrected references, not its prevalence across modern ML evaluation.

7 CONCLUSION

Incomplete semantic reference creates distinct evaluation regimes. Across structured and contemporary fixed-output evaluations, changing the reference can preserve the selected system, alter pairwise comparisons, or reverse the top-system choice. Scientific warrant depends on the answers induced across references compatible with qualified evidence. If they agree, certify; if qualified evidence can separate them, acquire it; if the contract lacks authority, return UNKNOWN.

Reproducibility Statement

All reported experimental results and benchmark metrics are derived from deterministic, fully specified pipelines with fixed random seeds, pinned dependencies, and versioned data sources. The Appendix documents the complete experimental protocol, formal source contracts, preprocessing steps, degradation masks, sample sizes, and reproduction commands for all public benchmarks. Anonymous reproduction code, evaluation scripts, and machine-readable result tables are included with this submission; public release will follow publication.

Artificial Intelligence (AI) Use Statement

OpenAI Codex was used to assist with experimental code implementation, debugging, and refactoring. Google Gemini was used for drafting portions of the manuscript, language editing, and text revision. Elicit was used to identify and retrieve relevant literature. The authors reviewed all AI-assisted outputs and remain fully responsible for the research design, empirical analyses, conclusions, and final manuscript.

References

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Artificial Analysis. VoxPopuli-Cleaned-AA: Cleaned ground truth transcripts for VoxPopuli english test set. Public dataset, version 1.0, https://huggingface.co/datasets/ArtificialAnalysis/VoxPopuli-Cleaned-AA, 2026. Apache-2.0 license.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025. doi: 10.52202/085713-0590. URL https://openreview.net/forum?id=mdA5lVvNcU.

Alessandro Berti, Sebastiaan J. van Zelst, and Wil M. P. van der Aalst. Process mining for python (PM4Py): Bridging the gap between process- and data science. In *International Conference on Process Mining Demo Track*, volume 2374 of *CEUR Workshop Proceedings*, pp. 13–16, 2019. URL https://ceur-ws.org/Vol-2374/paper4.pdf.

Harry H. Beyel and Wil M. P. van der Aalst. Improving process discovery using translucent activity relationships. In *Business Process Management (BPM)*, volume 14940 of *Lecture Notes in Computer Science*, pp. 146–163. Springer, 2024. doi: 10.1007/978-3-031-70396-6_9.

Harry H. Beyel, Christopher T. Schwanen, and Wil M. P. van der Aalst. Translucent alignments. In *Conference on Advanced Information Systems Engineering (CAiSE)*, volume 15702 of *Lecture Notes in Computer Science*, pp. 167–184. Springer, 2025. doi: 10.1007/978-3-031-94571-7_10.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7331–7348. PMLR, 2023. URL https://proceedings.mlr.press/v206/bickfordsmith23a.html.

Bluesky Project. Bluesky event model. Official documentation, https://blueskyproject.io/event-model/, 2026. Accessed September 2026.

Yuxin Chen, Hamed Hassani, and Andreas Krause. Near-optimal bayesian active learning with correlated and noisy tests. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pp. 223–231. PMLR, 2017. URL https://proceedings.mlr.press/v54/chen17b.html.

T. Coleman, H. Casanova, L. Pottier, M. Kaushik, E. Deelman, and R. Ferreira da Silva. WfCommons: A framework for enabling scientific workflow research and development. *Future Generation Computer Systems*, 128:16–27, 2022. doi: 10.1016/j.future.2021.09.043.

Marco Comuzzi, Jonghyeon Ko, and Fabrizio Maria Maggi. A language to model and simulate data quality issues in process mining. *ACM Journal of Data and Information Quality*, 17(2):6:1–6:36, 2025. doi: 10.1145/3743144.

Julien Corman, Werner Nutt, and Ognjen Savković. Complete approximations of incomplete queries. *arXiv preprint arXiv:2407.20932*, 2024. URL https://arxiv.org/abs/2407.20932.

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J. White, and Su-In Lee. Learning to maximize mutual information for dynamic feature selection. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6424–6447. PMLR, 2023. URL https://proceedings.mlr.press/v202/covert23a.html.

Yubing Cui and Viswanath Nagarajan. Minimum cost adaptive submodular cover. In *Symposium on Simplicity in Algorithms (SOSA)*, pp. 12–27, 2023. doi: 10.1137/1.9781611977585.ch2.

Massimiliano de Leoni and Felix Mannhardt. Road traffic fine management process. https://data.4tu.nl/articles/dataset/Road_Traffic_Fine_Management_Process/12683249, 2015.

DeepMind. PySC2: Starcraft ii learning environment. Official source repository, https://github.com/google-deepmind/pysc2, 2026. Pinned interface source inspected September 2026.

Danial Dervovic and Michael Cashmore. Model evaluation in the dark: Robust classifier metrics with missing labels. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1909–1917. PMLR, 2025. URL https://proceedings.mlr.press/v258/dervovic25a.html.

Ricardo Dominguez-Olmedo, Florian E. Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ab8c971c2ccd12bac0ab249f75e2c16d-Abstract-Conference.html.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=NO6Tv6QcDs.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Timo Freiesleben and Sebastian Zezulka. The benchmarking epistemology: Construct validity for evaluating machine learning models. *arXiv preprint arXiv:2510.23191*, 2025. URL https://arxiv.org/abs/2510.23191.

Alessandro Gianola, Marco Montali, and Sarah Winkler. Object-centric processes with structured data and exact synchronization. In *Conference on Advanced Information Systems Engineering (CAiSE)*, volume 15702 of *Lecture Notes in Computer Science*, pp. 185–202. Springer, 2025. doi: 10.1007/978-3-031-94571-7_11.

GitHub. Rest api endpoints for timeline events. https://docs.github.com/en/rest/issues/timeline, 2026. Official endpoint documentation; accessed September 2026.

Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 23, pp. 748–756, 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/1e6e0a04d20f50967c64dac2d639a577-Paper.pdf.

Alexandre Goossens, Johannes De Smedt, and Jan Vanthienen. Extracting process-aware decision models from object-centric process data. *Information Sciences*, 682:121263, 2024. doi: 10.1016/j.ins.2024.121263.

Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pp. 31–40, 2007. doi: 10.1145/1265530.1265535.

GS1. EPCIS standard, release 2.0. GS1 standard, https://ref.gs1.org/standards/epcis/2.0.0/, 2022.

Hugging Face Audio contributors. Open ASR leaderboard reference-error scoring. Official source repository, commit 48e6f0d75ad7dad68c962112356fec44776f3bbb, https://github.com/huggingface/open_asr_leaderboard, 2026. Apache-2.0 license.

Alexey Ignatiev, Antonio Morgado, and Joao Marques-Silva. Propositional abduction with implicit hitting sets. In *European Conference on Artificial Intelligence (ECAI)*, pp. 1327–1335, 2016. doi: 10.3233/978-1-61499-672-9-1327.

Shervin Javdani, Yuxin Chen, Amin Karbasi, Andreas Krause, J. Andrew Bagnell, and Siddhartha Srinivasa. Near optimal bayesian active learning for decision making. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 33 of *Proceedings of Machine Learning Research*, pp. 430–438. PMLR, 2014. URL https://proceedings.mlr.press/v33/javdani14.html.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/edac78c3e300629acfe6cbe9ca88fb84-Abstract-Conference.html.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763. PMLR, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Dina Kretzschmann, Alessandro Berti, and Wil M. P. van der Aalst. State-aware object-centric process mining: Enhancing OCEL 2.0 with explicit state transitions. In *Enterprise Design, Operations, and Computing (EDOC)*, volume 16213 of *Lecture Notes in Computer Science*, pp. 103–118. Springer, 2026. doi: 10.1007/978-3-032-15140-7_6.

Krishna Kulkarni and Jan-Eike Michels. Temporal features in SQL:2011. *ACM SIGMOD Record*, 41 (3):34–43, 2012. doi: 10.1145/2380776.2380786.

LabOP Contributors. Laboratory open protocol language (LabOP). Official source repository, https://github.com/Bioprotocols/labop, 2026. Pinned source inspected September 2026.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 275–284, 2017. doi: 10.1145/3097983.3098066.

Sander J. J. Leemans, Dirk Fahland, and Wil M. P. van der Aalst. Discovering block-structured process models from event logs: A constructive approach. In *Application and Theory of Petri Nets and Concurrency*, volume 7927 of *Lecture Notes in Computer Science*, pp. 311–329. Springer, 2013. doi: 10.1007/978-3-642-38697-8_17.

Leonid Libkin. Certain answers as objects and knowledge. *Artificial Intelligence*, 232:1–19, 2016. doi: 10.1016/j.artint.2015.11.004.

Lichess.org. Lichess open database. https://database.lichess.org/, 2026. Official public game exports; accessed September 2026.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. doi: 10.52202/075280-0943. URL https://papers.nips.cc/paper_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract-Conference.html.

Felix Mannhardt. Sepsis cases – event log. https://data.4tu.nl/articles/dataset/Sepsis_Cases_-_Event_Log/12707639/1, 2016.

Felix Mannhardt. Hospital billing – event log. https://data.4tu.nl/articles/dataset/Hospital_Billing_-_Event_Log/12705113/1, 2017.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003. doi: 10.1007/b97478.

Drew McDermott, Malik Ghallab, Adele Howe, Craig Knoblock, Ashwin Ram, Manuela Veloso, Daniel Weld, and David Wilkins. PDDL: The planning domain definition language, version 1.2. Technical Report CVC TR-98-003/DCS TR-1165, Yale Center for Computational Vision and Control, 1998. URL https://ipc08.icaps-conference.org/deterministic/PddlResources.html.

MediaWiki contributors. API:Continue. https://www.mediawiki.org/wiki/API:Continue, 2026a. Official continuation documentation; accessed September 2026.

MediaWiki contributors. API:Revisions. https://www.mediawiki.org/wiki/API:Revisions, 2026b. Official documentation; accessed September 2026.

National Electrical Manufacturers Association. DICOM PS3.1 2026d: Introduction and overview. Official standard, https://dicom.nema.org/medical/dicom/current/output/chtml/part01/PS3.1.html, 2026. Current NEMA edition resolved as 2026d; accessed September 2026.

OpenAI. Introducing SWE-bench Verified. Official benchmark documentation, https://openai.com/index/introducing-swe-bench-verified/, 2024. Published August 13, 2024; updated February 24, 2025.

Plaans contributors. tyr-ipc-domains: PDDL and HDDL benchmark instances. Official source repository, commit 6034088ada70c51de2a462950e00523ab7b1a7a0, https://github.com/plaans/tyr-ipc-domains, 2026. MIT-licensed repository of International Planning Competition instances.

Anja Rebmann. OCEL to DOCEL. Author source repository, https://github.com/a-rebmann/ocel_to_docel, 2026. Pinned source inspected September 2026.

Sareh Sadeghianasl, Moe Thandar Wynn, Robert Andrews, Wil M. P. van der Aalst, and Jonghyeon Ko. Object-centric event-data imperfection patterns. *Information Systems*, 141:102766, 2026. doi: 10.1016/j.is.2026.102766.

O. Salaudeen, A. Reuel, A. Ahmed, S. Bedi, Z. Robertson, S. Sundar, B. Domingue, A. Wang, and S. Koyejo. Measurement to meaning: A validity-centered framework for AI evaluation. *arXiv preprint arXiv:2505.10573*, 2025. URL https://arxiv.org/abs/2505.10573.

Vasileios Sevetlidis. Bayes-sufficient representations in supervised learning. *arXiv preprint arXiv:2606.04045*, 2026. URL https://arxiv.org/abs/2606.04045.

Dominique Sommers, Natalia Sidorova, and Boudewijn van Dongen. A ground truth approach for assessing process mining techniques. *Process Science*, 2:1, 2025. doi: 10.1007/s44311-025-00006-8.

Vaibhav Srivastav, Steven Zheng, Eric Bezzam, Eustache Le Bihan, Nithin Rao Koluguri, Piotr Żelasko, Somshubra Majumdar, Adel Moumen, and Sanchit Gandhi. Open ASR leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation, 2025. URL https://arxiv.org/abs/2510.06961.

Jan Niklas van Detten, Pol Schumacher, Benedikt Knopp, and Sander J. J. Leemans. Noise-resistant discovery of identifier-sound object-centric process models with dynamic identity relations. *Process Science*, 3:17, 2026. doi: 10.1007/s44311-026-00052-w.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 993–1003. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.80. URL https://aclanthology.org/2021.acl-long.80/.

Wikidata contributors. Wikidata:Licensing. https://www.wikidata.org/wiki/Wikidata:Licensing, 2026. Official licensing overview, including CC0 for structured data; accessed September 2026.

Francesca Zerbato, Marco Franceschetti, and Barbara Weber. A framework to support the validation of process mining inquiries. In *BPM Forum*, volume 526 of *Lecture Notes in Business Information Processing*, pp. 249–266. Springer, 2024. doi: 10.1007/978-3-031-70418-5_15.

A Source Authority and Truth Levels

This Appendix provides formal source contracts, complete experimental results, robustness analyses, implementation details, and reproducibility protocols. Four truth levels are separated throughout:

1. **Rule or specification truth**: chess legality, PDDL applicability, and DAG readiness under pinned formal specifications.
2. **Executable model truth**: generated or author-supplied process states and enabled transitions.
3. **Source-record truth**: revisions, lifecycle events, and stream assertions reachable through a versioned source interface.
4. **World truth**: physical, clinical, normative, or complete historical reality. No public experiment upgrades the first three levels to this level.

Controlled degradation preserves evaluator-only rule, model, or source truth while hiding it from the tested method. Natural panels use the reachable source record as authority and retain deleted, inaccessible, or unrecorded history as a limitation. Repeated states, masks, queries, case splits, and portfolios are not counted as independent systems.

Source-card abbreviations are data-aware object-centric event log (DOCEL), StarCraft II (SC2), Laboratory Open Protocol language (LabOP), and Digital Imaging and Communications in Medicine (DICOM).

Table 4: Complete source authority cards. "Qualification" means that the source defines a boundary but does not supply evaluator truth for a deep quantitative claim.

Source	Admitted authority	Natural / controlled	Paper role	Not established
Lichess Jan. 2013	Positions plus independently executed chess rules	Natural selection; proxy substitution	Selected/legal control	Player intent, optimality, prevalence
WfInstances	Execution record plus DAG dependencies	Natural execution; derived readiness	Scientific workflow panel	Physical completion, normative scheduling
International Planning Competition / PDDL	Domain/problem files and grounded preconditions	Formal truth; predicate masks	Positive reversal and bounds	Real planning prevalence
PM4Py nets	Generated markings and executable enablement	Generated exact truth	Multiplicity control	Natural operational prevalence
Hospital, Road, Sepsis	Published generated enabled sets over public-log records	Controlled masks and splits	Recovery and portfolio evaluation	Clinical or normative permission
MIRA	Author model and transition/binding interface	Model truth; masks/coarsening	Representation and uncertainty	Physical package-delivery truth
TrialVersion	Retained generated versions and visibility	Controlled version masking	Early answerability	Natural version-system prevalence
Wikidata	Reachable revisions under frozen intervals	Natural source history	Witness/exhaustion and sharing	Valid-time/world history; suppressed revisions
GitHub	Reachable labelled/unlabelled issue events	Natural source lifecycle	Relationship-history validation	Deleted history; immutable label identity
Bluesky	Valid public run documents and stored streams	Natural records; controlled deletion	Completion boundary	Physical experiment completion
DOCEL / GS1	Generated operation effects or standard examples	Mixed	Lifecycle and UNKNOWN boundary	Unknown initial/world relations
SC2, LabOP, DICOM	Public interfaces, standard, or metadata	Qualification	Unsupported target boundary	Paired longitudinal evaluator truth

Primary source contracts are cited at first use: public process logs (Mannhardt, 2017; de Leoni & Mannhardt, 2015; Mannhardt, 2016), workflow and planning sources (Coleman et al., 2022; McDermott et al., 1998; Plaans contributors, 2026; Berti et al., 2019), source-history interfaces (MediaWiki contributors, 2026b;a; GitHub, 2026), and qualification standards or repositories (Bluesky Project, 2026; GS1, 2022; National Electrical Manufacturers Association, 2026; LabOP Contributors, 2026; Rebmann, 2026; DeepMind, 2026).

B Formal Definitions and Defensible Properties

Let $\Omega_{\mathcal{C}}(e)$ be the worlds compatible with hard evidence e under contract $\mathcal{C}$, $r_{\mathcal{C}}(x)$ the required reference, and $a_q(z, r) = q(m(z, r))$ the declared answer for frozen output z. Then

$$\mathcal{R}_{\mathcal{C}}(e) = \{r_{\mathcal{C}}(x) : x \in \Omega_{\mathcal{C}}(e)\}, \qquad \mathcal{A}_q(e, z) = \{a_q(z, r) : r \in \mathcal{R}_{\mathcal{C}}(e)\}. \tag{5}$$

Proposition 1 (Observational non-identifiability)**.** *If two admissible worlds produce the same admitted evidence but different required references, no deterministic function of that evidence uniquely recovers the reference in both worlds. If their query answers differ, no deterministic evidence-only answer is valid in both.*

Proof. The deterministic function receives the same input in both worlds and returns one value. That value cannot equal two distinct references or answers. Additional authority, a set-valued result, or UNKNOWN is required. This is an elementary application of established identifiability and certain-answer reasoning (Manski, 2003; Libkin, 2016). □

Proposition 2 (Answerability without reference identification)**.** *There exist* e, z, q *such that* $|\mathcal{R}_{\mathcal{C}}(e)| > 1$ *but* $|\mathcal{A}_q(e, z)| = 1$.

Proof. Take compatible references $r_1 = \{A, B\}$ and $r_2 = \{A, B, C\}$ and ask whether A is enabled. The references differ, but both answers equal one. Metric and ranking queries use the same geometry. □

For a fixed weighted table, let representation cells be g and let $W = \sum_i w_i$. The target-relative empirical ambiguity used below is

$$A(\phi, y) = \frac{1}{W} \sum_g \left[\sum_{i \in g} w_i - \max_c \sum_{i \in g : y_i = c} w_i \right].$$

Proposition 3 (Finite-table representation floor)**.** *For a deterministic predictor measurable with respect to representation cells* g*, its weighted 0–1 error on the fixed table is at least* $A(\phi, y)$ *as defined above. If* ϕ_2 *refines* ϕ_1*, then* $A(\phi_2, y) \leq A(\phi_1, y)$.

Proof. Within each cell the best deterministic label selects the largest class mass; the remaining mass is unavoidable. Refinement partitions a cell and cannot increase the sum of within-cell maxima. □

Under a fixed, consistent, hard-evidence contract, adding qualified facts contracts $\Omega_{\mathcal{C}}(e)$ and therefore cannot enlarge $\mathcal{R}_{\mathcal{C}}(e)$ or $\mathcal{A}_q(e, z)$. No such statement is made for soft predictions treated as beliefs.

C METHOD REPRODUCTION SUMMARY

Method abbreviations used below are Integrated Object-centric Decision Discovery Algorithm (IODDA), data-aware object-centric Petri nets with identifiers (DOPID), Expected Predictive Information Gain (EPIG), Bayesian Active Learning by Disagreement (BALD), equivalence-class edge-cutting (EC2), equivalence-class edge-discounting (ECED), hyperedge cutting (HEC), active feature acquisition (AFA), implicit hitting set (IHS), and Query-Certifying Reference Acquisition (QCRA).

Table 5: Native, reference-level, and aligned implementation statuses. Status is limited to the named component; it never implies paper-wide parity.

Method/component	Status	Executed or checked	Unreproduced boundary / use
Ordinary Inductive Miner (Leemans et al., 2013)	Native reproduced	PM4Py discovery, conversion, export/reimport, replay	Bounded construction only; no physical validity
Translucent discovery/precision (Beyel & van der Aalst, 2024; Beyel et al., 2025)	Partial native	Author enabled-prefix and precision components; PM4Py fit	No full alignment campaign or publication-table parity
FLOWVAULT (Kretzschmann et al., 2026)	Native, bounded	52/52 golden/source queries and bounded assignments	Supplied mappings/formulas; no paper-wide or clinical-state parity
IODDA (Goossens et al., 2024)	Partial native	Native Publication path and aggregate task	Publication parity unproven; Shipping not reproduced
DOPID / CoCoMoT (Gianola et al., 2025)	Reference-level	Native examples and finite definition checks	No full CoCoMoT campaign or world-permission reference
Dynamic identity discovery (van Detten et al., 2026)	Partial native	Hinge normalized author-tree parity	Orders non-parity retained; no external lifecycle truth
MIRA	Partial native	Published model, enabled firings, successor API	No author alignment table or full simulation campaign
FLAWD / OCIP	Not reproduced	Literature and repository inspection	No implementation failure inferred
Random forest, linear, tree, prior	Aligned reimplementation	Frozen benchmark learners and 128-tree recovery	Existing learners; no new algorithm claim
PDDL exact / three-valued	Reference-level / aligned	Grounded applicability and sound lower/complete upper sets	Formal truth only
EPIG / BALD	Scoring-kernel parity only	36+36 author-kernel score checks	No native training or dataset-result reproduction
EC2 / ECED	Reference-level	Exhaustive finite score checks	No author experiment campaign
HEC and modern AFA	Not reproduced / literature only	Formulation and assumption comparison	No performance absence or superiority claim
Random, fixed, count, influence	Aligned reimplementation	Matched acquisition controls	Baselines only
Entropy/query information/VoI	Aligned reimplementation	Finite-world acquisition scores	Soft planning does not certify hard evidence
Counterexample / IHS	Aligned / reference-level	Exact counterexample and bounded basic IHS	No native Hyper parity; exponential boundary retained
QCRA / QCRA-Cover	Aligned reimplementation	Query certificate interface and cover heuristic	No new cover theorem or universal selector claim
Exact Bellman planning	Reference-level	Finite explicit oracle	Exponential state space
Cache/portfolio controls	Aligned reimplementation	Exact D, U, K, G cost decomposition	Attribution control, not acquisition novelty

Ordinary and translucent discovery use PM4Py and author components (Berti et al., 2019; Beyel & van der Aalst, 2024; Beyel et al., 2025). MIRA is cited through its primary publication (Sommers et al., 2025). Acquisition foundations are attributed to their primary work (Golovin et al., 2010; Javdani et al., 2014; Chen et al., 2017; Bickford Smith et al., 2023; Ignatiev et al., 2016). The accompanying source record contains author pins, licences, inputs, deviations, and parity criteria.

D ALGORITHM–REFERENCE DEPENDENCY MATRIX

The matrix supports a bounded claim: the studied methods expose distinct reference dependencies at execution, training, evaluation, and interpretation. It does not estimate prevalence over all process-mining or machine-learning algorithms.

E BENCHMARK LEARNERS AND FULL SELECTED–ENABLED RESULTS

The four benchmark outputs are: a random forest with 128 trees; logistic/linear multilabel scores; a depth-limited decision tree; and an empirical prior. Hyperparameters, seeds, candidate vocabulary, and split assignments are fixed in the source configurations. Candidate outputs are fixed before swapping full and selected-only reference.

Table 6: Complete layer-specific dependency matrix. "Run" means execution input, not scientific validity.

Method/task	Execution	Training	Evaluation	Claim	Failure and warranted response
Ordinary discovery	Ordered traces	Recorded log	Replay/model	Semantic mapping	Stronger claim unsupported; direct recorded result or UNKNOWN
Translucent discovery	Enabled sets	Selected+enabled	Enabled/model	Permission interpretation	Input failure or proxy distortion; direct/derive
FLOWVAULT	Mapping or formula	Optional abstraction	Formula/source query	Named state	Acquire mapping or report source-relative result
IODDA	DOCEL ownership/target	Observed target	Observed decisions	Allowed decisions	Input-contract failure or descriptive-only claim
DOPID	Typed net/log	Given model	Model conformance	World permission	Derive model-relative result; UNKNOWNbeyond mapping
Dynamic relations	Typed event/object data	Relation learner	Model/source relation	Lifecycle truth	Recover/acquire lifecycle; UNKNOWNif history absent
MIRA	Supplied models/tags	Generated data	Generator tags	External realism	Direct model truth; no world-frequency claim
Benchmark learner	Features	Declared target	Full/proxy target	Ranking	Recover, certify, acquire, or UNKNOWN
PDDL exact	Formal state/model	None	Applicability	Formal ranking	Deterministic derive
PDDL partial	Partial state/model	None	Lower/upper sets	Ranking	Certify if interval one-sided; else acquire/UNKNOWN
Temporal query	Retained history	None	Version records	Source/world history	Snapshot distortion; acquire or UNKNOWN
Representation diagnostic	Fixed cells	None	Target table	Target/table risk	Unavailable without declared target
Predictive uncertainty	Fitted model	Posterior/data	Calibration labels	Semantic warrant	Soft planning only; does not certify authority
EC2/HEC/ECED	World/test model	Prior/likelihood	Decision region	Contract-relative answer	Input-contract failure if model missing
EPIG/BALD	Predictive samples	Posterior or particles	Target inputs	Semantic warrant	Inner predictive score only
QCRA/QCRA-Cover	Evidence menu/checker	Optional ordering model	Exact certificate	Warranted answer	Acquire qualified facts; UNKNOWNif unresolvable
Bellman oracle	Explicit states/costs	Prior/cost	Exact terminal check	Contract-relative optimality	Computational boundary, not semantic proof
Cache/portfolio	Stable source identity	None	Shared source-page inventory	Cost attribution	Reuse only under version/scope agreement

Table 7: Macro state-level set F_1 for every model and admitted cross-domain regime. The final columns give strict and prospectively declared material pair-order changes among six model pairs.

	Forest		Linear		Tree		Prior		Strict	Material
Family/regime	Full	Proxy	Full	Proxy	Full	Proxy	Full	Proxy	/6	/6
Chess	.933792	.076419	.787217	.043907	.780139	.045078	.475433	.035451	1	0
Workflow	1.000000	.513464	1.000000	.513464	1.000000	.513464	.311103	.076550	0	0
Generated nets	1.000000	.821671	1.000000	.821671	1.000000	.821671	.000000	.000000	0	0
PDDL 0%	1.000000	.197233	1.000000	.197233	1.000000	.197233	.071003	.196774	0	0
PDDL 25%	.703642	.130485	.635216	.079531	.619070	.074538	.071003	.196774	3	3
PDDL 50%	.497964	.086351	.471198	.047021	.454417	.043455	.071003	.196774	3	3
PDDL 75%	.340853	.060423	.351909	.033711	.346199	.030585	.071003	.196774	5	4

Primary units are states; candidate actions within a state are dependent. Chess uses 480 test states and 46,069 candidate rows; workflow 245 and 13,131; generated nets 3,659 and 64,882; PDDL 620 and 168,310. Local pair-order changes with identical omitted-action prediction sets are 12 in Chess and 76, 120, and 165 PDDL state-pairs at 25%, 50%, and 75% hiding, respectively. The exact local algebra is in the machine table `data/mechanism_summary.csv`.

F CONTEMPORARY FIXED-OUTPUT EXTERNAL VALIDATION

We prospectively qualified public fixed-output panels outside the benchmark's original structured process families. The inclusion rule admitted only identical released outputs scored under an independently defined weaker and stronger obligation. Tables 8–10 provide the numerical evidence behind the compact main-text summary, while the attempted-family qualification registry preserves ineligible and infeasible candidates. These panels serve as external corroboration rather than discovery claims. In particular, EvalPlus already shows that strengthened executable tests expose errors and model mis-ranking in code generation (Liu et al., 2023).

Table 8: **Detailed contemporary fixed-output statistics.** Proxy and qualified columns are panel means over admitted systems. Mean Δ is qualified minus proxy score in percentage points. HumanEval and MBPP use pass@1; SWE-bench uses the fraction of 500 official instances resolved; VoxPopuli uses compound-aware corpus WER, for which lower is better. Strict changes are sign reversals among all system pairs. These are exact descriptive finite-panel quantities, not independent population replicates.

Panel	Sys.	Units	Proxy	Qualified	Δ	Better	Metric
HumanEval → HumanEval+	89	Aggregate[a]	60.83	55.40	−5.42	Higher	pass@1
MBPP → MBPP+	89	Aggregate[a]	68.51	57.21	−11.29	Higher	pass@1
SWE-bench repair-only → full	5	2,500	39.28	36.04	−3.24	Higher	resolution (%)
VoxPopuli original → corrected	89	628 clips	7.74	4.75	−2.99	Lower	WER (%)

[a]The pinned EvalPlus leaderboard publishes paired aggregate pass@1 values but does not bundle the generations required to reconstruct task-level transitions. "Aggregate" therefore identifies the available evidence level rather than a missing system-level denominator.

Table 9: **Pair-order and rank evidence for the same panels.** Strict changes are sign reversals among all admitted system pairs; ties are additional tie-boundary changes. Kendall τ_b and Spearman ρ compare complete system rankings under the proxy and qualified references.

Panel	Pairs	Strict	Ties	Strict (%)	τ_b	ρ	Top changed
HumanEval → HumanEval+	3,916	76	54	1.94	.9539	.9949	No
MBPP → MBPP+	3,916	125	32	3.19	.9318	.9903	No
SWE-bench repair-only → full	10	0	0	0.00	1.0000	1.0000	No
VoxPopuli original → corrected	3,916	1,268	17	32.38	.3488	.4838	**Yes**

EvalPlus: HumanEval and MBPP. All 89 entries with complete finite pass@1 values across HumanEval, HumanEval+, MBPP, and MBPP+ on the pinned official leaderboard snapshot (commit `8741845`) were evaluated. On HumanEval → HumanEval+, 76 of 3,916 strict system-pair orders change, with 54 additional tie-boundary changes; rank correlations remain high (Kendall $\tau_b = 0.953914$, Spearman $\rho = 0.994921$), mean score changes by -5.421 percentage points, and the top-system set remains O1 Mini and O1 Preview. On MBPP → MBPP+, 125 of 3,916 strict pair orders change, with 32 tie-boundary changes; rank correlations are Kendall $\tau_b = 0.931778$ and Spearman $\rho = 0.990282$, mean score changes by -11.291 percentage points, and O1 Preview remains the top system. Thus strengthened executable references produce pairwise comparison sensitivity without altering the declared top-system query. Raw generation artifacts are not bundled with the pinned leaderboard snapshot, so this analysis uses official paired aggregate scores and cannot report program-level transitions.

SWE-bench Verified: repair-only versus repair-and-maintenance. SWE-bench Verified (OpenAI, 2024), built from SWE-bench (Jimenez et al., 2024), evaluates the same submitted patch under two nested contracts. The proxy obligation requires only issue-repair tests (`FAIL_TO_PASS`); the official qualified obligation additionally requires maintenance tests (`PASS_TO_PASS`) to prevent breaking existing functionality. Across five systems with complete public structured report artifacts and 2,500 system–instance evaluations, 81 evaluations pass the repair-only view but fail the full maintenance obligation. System scores decrease by 1.8–5.8 percentage points (mean -3.24 pp), yet all 10 system-pair orders remain unchanged (Kendall $\tau_b = 1$, Spearman $\rho = 1$), with Engine Labs remaining the top system. Six older prospective submissions lacked structured report files and are retained as explicit exclusions. This represents official test-contract truth for the submitted patch, rather than a proof of complete real-world software correctness.

Automatic speech recognition: VoxPopuli original versus human-corrected transcripts. For speech recognition, the Open ASR Leaderboard publishes fixed hypotheses evaluated against both the original parliamentary transcripts of the VoxPopuli English test set and an independently human-

Table 10: **Top-system conclusions and robustness.** Scores use each panel's primary metric; up-arrows denote higher-is-better and down-arrows lower-is-better. The robustness column reports the check that most directly supports the headline conclusion, rather than an exhaustive experiment inventory.

Panel / metric	Proxy top	Qualified top	Top change	Robustness / boundary
HumanEval / pass@1 ↑	O1 Mini and O1 Preview (96.3)	O1 Mini and O1 Preview (89.0)	No	Top equivalence set unchanged; task-level transitions unavailable in the pinned aggregate release.
MBPP / pass@1 ↑	O1 Preview (95.5)	O1 Preview (80.2)	No	Top system unchanged despite 125 strict pair-order changes.
SWE-bench / resolution ↑	Engine Labs (54.2)	Engine Labs (51.8)	No	81/2,500 evaluations pass repair-only tests but fail the full maintenance obligation; all 10 pair orders remain stable.
VoxPopuli / WER ↓	Parakeet RNNT 1.1B (4.68)	Voxtral Small 24B (2.00)	**Yes**	Disjoint top sets in 10/10 leave-one-block checks; plain WER retains the reversal with 1,267 strict changes.

corrected 628-clip reference (`VoxPopuli-Cleaned-AA`) (Wang et al., 2021; Srivastav et al., 2025; Artificial Analysis, 2026; Hugging Face Audio contributors, 2026). Across all 89 public models with usable hypothesis manifests evaluated on the identical 628 audio clips, mean compound-aware corpus WER decreases by 2.99 percentage points (from a panel mean of 7.74% to 4.75%), while 1,268 of 3,916 strict system-pair orders change (32.4%) along with 17 tie-boundary changes (Kendall $\tau_b = 0.348816$, Spearman $\rho = 0.483817$). Crucially, the lowest-WER top system reverses: under the original proxy transcript, `nvidia/parakeet-rnnt-1.1b` is selected at 4.68% WER (with `mistralai/Voxtral-Small-24B-2507` ranking 39.5 at 6.14%); under the human-corrected reference, `mistralai/Voxtral-Small-24B-2507` wins at 2.00% WER (margin 2.68 pp relative to the proxy winner, and outperforming Parakeet's corrected score of 3.64% by 1.64 pp, with Parakeet falling to rank 31). Independent verification with a separate loader reproduces all 178 source scores exactly; plain word-level Levenshtein edit distance preserves the identical disjoint top systems and 1,267 pair changes; and all ten contiguous leave-one-block tests preserve top-system reversal. This finding constitutes an empirical existence result under declared benchmark transcript authority rather than an estimate of reversal prevalence across all foundation-model evaluation; dataset cleaning and leaderboard governance belong to Artificial Analysis and the Open ASR team.

Attempted-family qualification registry. Other prospective candidate families were evaluated against the fixed-output evidence contract but excluded for principled reasons: Spider/SParC/CoSQL (no multi-system fixed-prediction panel retrievable under the cutoff), BFCL (AST versus execution evaluators apply to different categories rather than nested obligations on identical items), WebArena-Verified and OSWorld-Verified (lacked matched original and verified contracts on identical multi-system trajectory panels), ImageNet-ReaL (no multi-system per-image predictions in pinned releases; corrected-label ranking effects belong to prior work), corrected-label vision and Cleanlab regression (lacked multi-system prediction tensors, and generation would require prohibited model training), and WeatherReal (observation benchmark without precomputed multi-system forecasts). The complete specification is included in the anonymous reviewer package (`data/contemporary_reversal_candidate_registry.csv`).

Together, these contemporary panels reproduce three distinct regimes outside the original benchmark: stable system ordering under a stronger software-maintenance obligation (SWE-bench), pairwise comparison sensitivity under stronger executable code tests (EvalPlus), and top-system reversal under independently human-corrected speech transcripts (VoxPopuli ASR).

G PDDL Robustness and Partial Information

The admitted inventory has nine training and six test problem units from Blocks, Gripper, and Logistics. Predicate hiding is deterministic at 0, 25, 50, and 75%. Full applicability is computed from the complete formal state; the proxy contains only the selected plan action. Four alternative problem folds show the partial-mask winner reversal in 4/4 cases at every nonzero mask. It also occurs in one unmasked problem fold. Leave-one-domain-out analyses retain reversal in Blocks and Logistics and expose Gripper as the main macro-F_1 boundary.

Table 11: Three-valued PDDL reference bounds. Lower sets are sound; upper sets are complete.

Hidden	Lower P	Lower R	Upper P	Upper R	Unknown atoms
0%	1.000	1.000	1.000	1.000	0
25%	1.000	.476949	.344733	1.000	27,707
50%	1.000	.183886	.172327	1.000	64,231
75%	1.000	.037005	.103132	1.000	110,416

Jaccard and micro-F_1 agree with the primary winner change across partial folds. Exact-set accuracy, component metrics, precision, and recall do not uniformly agree. The formal reference is therefore stable; the downstream scientific query changes with the metric. Full fold-level rows are in `data/pddl_stability.csv`.

H Recovery and Query-Relative Sufficiency

Public event records are split by case-hash into train/dev/test partitions; the five test splits overlap and are treated descriptively. Enabled-set targets are author-generated model-relative reference. The random forest uses 128 trees and only separate training cases. It never sees the evaluator-hidden target cells at test.

Table 12: Full recovery table. "Proxy changes > 1 pp" is retrospective sensitivity; strict sign is the prospectively specified endpoint.

Family	Representation	Cells	Proxy errors	Recovery errors	Fidelity	Proxy signs	Recovery signs	Max residual pp
Hospital	Last activity	88,438	17,599	970	.989032	48/75	0/75	0
Hospital	History counts	88,438	17,599	782	.991158	51/75	7/75	.036412
Road	Last activity	37,008	3,769	930	.974870	51/75	11/75	.119554
Road	History counts	37,008	3,769	906	.975519	52/75	4/75	.074111
Sepsis	Last activity	10,413	2,971	472	.954672	50/75	0/75	0
Sepsis	History counts	10,413	2,971	13	.998752	49/75	0/75	0

For the main history-count obligation, all residual strict disagreements have full-reference margin below 0.1 percentage point. The Road last-activity negative control reaches 0.119554 point and is not folded into that claim. Recovery solves the tested material-ranking obligation while leaving atom-level reference error.

I Controlled Context and Selection Support

The exact process has states s_0, s_1 with enabled sets $\{a, b\}$ and $\{a, c\}$. One-step history aliases the states; two-step history separates them. The reverse control uses one shared reference $\{a, b\}$ and two different policies.

Table 13: Exact context, uncertainty, and query results.

Quantity	**Depth 1**	**Depth 2 / control**
Representation ambiguity	$1/3$	0
Reference identity	unresolved	resolved
Robust F_1 difference	$[.1333, .1333]$	same answer
Sensitive F_1 difference	$[-.5, .5]$	$+.5$ or $-.5$
$p = 0$ predictive entropy	0 bits	n/a
$p = 0$ reference entropy	1 bit	n/a
Reverse-control reference entropy	0 bits	n/a
Reverse-control predictive entropy	1 bit	n/a

Table 14: Generated units observing the state-specific alternative at least once; 2,000 units per cell with nested prefixes.

p	$n = 10$	$n = 50$	$n = 200$	$n = 1000$
0	0	0	0	0
.01	189	785	1,741	2,000
.05	853	1,867	2,000	2,000
.5	1,998	2,000	2,000	2,000

The maximum absolute deviation from $1 - (1 - p)^n$ is 0.025237, below the prospectively frozen 0.035 tolerance. These 32,000 prefix units are generated under one controlled mechanism and are not 32,000 independent systems. Exact enumeration, deterministic replay, and an independent recomputation agree.

J Representation and Uncertainty

Reference uncertainty is multiplicity in $\mathcal{R}_{\mathcal{C}}(e)$; conclusion uncertainty is multiplicity in $\mathcal{A}_q(e, z)$; representation ambiguity is the finite-table quantity defined in Section B; model uncertainty concerns the fitted predictive distribution. They are reported separately.

The MIRA held-out repeated-row denominator is 8,470 across five overlapping state splits, coarsenings, and two targets. For selected-only training, 6,570 rows lie below predictive entropy 0.1 and 1,037 are wrong for the semantic target. Under full coarsening and true-target training, predictive entropy detects 18.5714% target loss. The paired finding rejects both "confidence certifies semantics" and "predictive uncertainty is always blind."

The no-abstention ablation on fully masked MIRA and Sepsis answers 15/15 queries in each panel and happens to match hidden evaluator truth, yet all 15 assertions per panel lack an evidence-scoped certificate. GS1 has 225/630 eligible queries and 405 source-unknown queries; forcing answers creates 405 unsupported assertions. Selective risk must always be paired with answered coverage and warrant.

EPIG and BALD comparisons reproduce only the author scoring kernels at the pinned commit. Their finite model particles are inner approximations and cannot certify every admissible source world.

K Acquisition and QCRA

QCRA is the following task. Given a compatible-world contract, exact query checker, source-authorised evidence menu, costs, and optional soft ordering model, select evidence until the compatible answer set is a singleton or the menu is exhausted. A terminal certificate always comes from hard qualified evidence. The certificate is the invariant scientific endpoint; the selector is a replaceable planning layer that changes efficiency under its own assumptions.

1. Declare contract $\mathcal{C}$, evidence e, query q, frozen output z, and menu M.

2. While $|\mathcal{A}_q(e, z)| > 1$ and a qualified request remains, score requests using a declared policy and current evidence, acquire one request, append source identity/version/cost/outcome to e, and recompute the exact compatible answer set.

3. Return the singleton answer with its evidence certificate, or a reason-coded UNKNOWN.

The evaluated selectors are random, source-native fixed orders, unresolved-count/incidence, influence, reference entropy, query information, EC2-style edge cutting, one-step value of information, fresh counterexample, certificate cover, bounded implicit hitting set, and exact Bellman planning where the explicit state space is tractable. QCRA-Cover solves a predicted certificate covering/knapsack subproblem for additive grouped queries. It adapts established cover and abduction ideas and has no new approximation theorem.

For acquisition and abstention panels we report the tuple of warrant coverage, certified-query error, and source-native acquisition cost. Warrant coverage is the number of queries whose compatible answer set is a singleton divided by all declared queries. Certified-query error is computed only on that covered set and is never shown without coverage. Cost remains in the source's native unit, so the tuple supports within-panel trade-off analysis rather than cross-source scalar aggregation.

Across 150 matched portfolio conditions, let D be unchanged independent-query traces, U exact duplicate elimination, K cached sequential rechecking, and G aggregate ordering. The comparison $G - K$ gives 13 wins, 3 losses, and 134 ties; MIRA ties in all 30 matched conditions. Large $D - G$ savings therefore cannot be attributed to selector order without cache controls.

At 25% preview noise, treating soft previews as authority answers all 1,140 temporal and 300 relational cells but is wrong on 262 and 41. Keeping previews soft answers only when exact acquired evidence certifies and has zero wrong certified answers in the matched panels. No learned or reinforcement-learning selector survived as a paper method: a structural dispatcher lost to the best fixed policy on mixed regimes, and small exact tasks were often solved by simple baselines.

L NATURAL HISTORY AND RELATIONSHIP LIFECYCLE

Wikidata. Eight entities and 22 properties define 44 binary source-history queries over 434 reachable revisions. The current-snapshot proxy is wrong on 5/44 queries. Five positive existential queries admit a qualified witness; 39 negative interval invariants require complete admitted coverage. Both newest-first and oldest-first retrieval are retained. The shared source-page inventory uses 46 source pages; unchanged independent-query traces use 227 to 242 requests. Raw shared response burden is 81,080,963 bytes. These are API costs, not human or monetary costs.

GitHub. The original prospective run fixed four public closed issues, three exact label names per issue, three historical cuts, and a resource cap. It stopped after two complete timelines. Its 2/4 completion and 8/18 current-snapshot errors document the failure mode under tight resource limits. A prospective continuation retained the identical panel and query contract while increasing only the resource envelope. It completed 4/4 issues using 57 total pages, of which 30 were newly fetched, and 2,189,743 newly retrieved raw bytes. The snapshot is wrong on 12/36 queries. Forward replay and reverse-from-current reconstruction agree 36/36. Timeline page order is not a licensed chronological prefix certificate, and the API does not exclude deleted or inaccessible history.

Completion and unsupported history. TrialVersion's final export is wrong on 1,176/1,692 visibility queries, while 428/24,300 masked cells already determine the declared payload query before complete reference identity. In Bluesky, the unchanged-record proxy is correct on 119/119 streams, while deleting the tail makes it wrong on 119/119 and interior deletion on 48/48. These are stored-stream completeness results, not physical experiment completion. GS1's missing initial relation state yields 405/630 source-unknown cells.

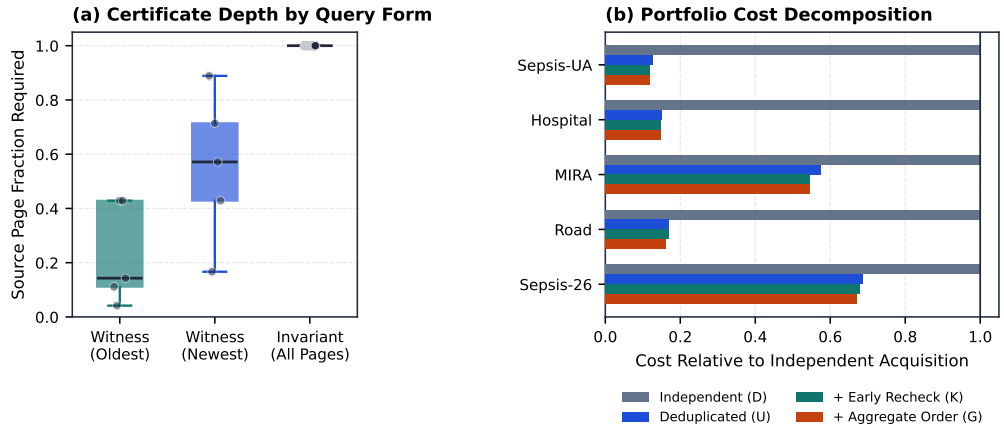

Figure 3: **Natural evidence burden depends on query form and shared authority.** Positive Wikidata witnesses can certify before full retrieval; negative interval invariants require admitted exhaustion. Portfolio cost decomposition separates ordinary duplicate/cache savings from the much smaller incremental effect of aggregate ordering.

M Actual Ablations and Controls

Table 15: Executed ablations and controls. Conceptual alternatives that were not run are excluded.

Ablation/control	Observed result	Scientific consequence
Proxy substitution	Cross-domain score distortion and PDDL reversal	Endpoint score alone does not identify the missing semantic layer
No recovery	49 to 52/75 process pair signs change	Proxy substitution can materially alter evaluation
Target-matched recovery	0 to 7/75 strict residuals; no material history-count residual	Existing recovery solves the tested material query
Full identity requirement	Robust context query would wait for depth two	Exact reconstruction can over-acquire
No acquisition	Fully masked MIRA/Sepsis certified coverage is 0/15 per panel	Coverage collapses honestly when nothing certifies
No abstention	30 unsupported masked assertions; 405 unsupported GS1 assertions	Correct hidden answers can still be unwarranted
Soft treated as hard	262/1,140 temporal and 41/300 relational errors	Predictions may guide order but cannot certify authority
Independent vs shared cache	D, U, K, G decomposition; G vs K: 13/134/3	Most portfolio saving is ordinary reuse
Reference vs representation loss	MIRA target/coarsening controls differ	The diagnostics identify different layers
Selected singleton vs three-valued PDDL	Sound lower/complete upper bounds preserve uncertainty	Existing formal reasoning narrows without always deciding
Forward vs reverse lifecycle replay	GitHub 36/36 agreement	Source-history reconstruction is internally verified
Independent certificate encoding	47,145/47,145 agreement	Encoding check supports exact certificate implementation

N Design Safeguards and Falsified Hypotheses

Design safeguards are not empirical ablations. They include freezing outputs before reference replacement; withholding evaluator truth from recovery; preserving source versions and hashes; separating soft planning signals from hard evidence; recording unsupported targets; and reporting coverage with selective risk.

Falsified or weakened hypotheses remain part of the evidence:

- Ranking changes do not require models to disagree on omitted actions; exact local counterexamples exist in all four fixed-output families.
- Predictive uncertainty is not universally blind to representation loss; the full-coarsening MIRA control raises entropy.

- Aggregate portfolio planning is not the source of most savings; matched caching explains most of the gain.
- A structural learned selector is not a robust cross-regime improvement; the best fixed policy wins on the mixed held-out comparison.
- The original GitHub run did not complete. Its continuation is a separate prospective addendum, not a retroactive success.
- An early PDDL mask that leaked hidden information was invalidated and excluded before the admitted run.
- SC2, LabOP, and some DICOM targets fail the source contract; they are unsupported targets, not algorithm failures.

O Statistics and Experimental Units

Exact finite panels receive exact descriptive counts. No independent-and-identically-distributed interval is attached to candidate actions within a state, overlapping case splits, dependent queries from one source history, or exhaustive generated panels. PDDL fold and exclusion analyses are sensitivity checks over a fixed problem inventory, not prevalence samples. Context-support Monte Carlo rates use 2,000 generated units per cell and are compared with a predeclared exact law and tolerance. Natural GitHub and Wikidata panels are purposive qualified examples.

The PDDL one-percentage-point material category was prospectively declared. In recovery, strict sign was the prospectively specified endpoint; > 1 percentage point and < 0.1 point are retrospective descriptive sensitivity categories. The predictive entropy threshold 0.1 is a frozen diagnostic threshold. No retrospective category is presented as a prospective success criterion.

Table 16: Threshold provenance.

Panel	Threshold	Status
PDDL	> 1 pp material reversal	Prospective endpoint
Recovery	strict pair sign	Prospective endpoint
Recovery	> 1 pp and < 0.1 pp	Retrospective sensitivity
MIRA	predictive entropy $\leq .1$	Frozen diagnostic
Context	max deviation $\leq .035$	Prospective tolerance

P Reproducibility Details

The anonymized package accompanying this submission reproduces the paper's result tables, figures, semantic-direction checks, and deterministic unit tests from the included evidence tables and public experiment code. Raw public-source acquisition is documented where redistribution or network access prevents a self-contained rerun. The package is organized as follows:

1. **Evidence tables and configurations**: Contains the exact reported result tables, benchmark configurations, thresholds, source qualifications, and hashes used by the manuscript.
2. **Evaluation pipelines**: Implements the reference-aware evaluation, recovery, acquisition, source-history, and context-support components used in the reported panels.
3. **Verification and rendering**: Provides deterministic tests, a semantic table-to-manuscript verifier, and figure regeneration directly from the included evidence tables.

To verify the reported results and regenerate the figures, the primary commands are:

```
python3 scripts/analysis/verify_reported_results.py
python3 -m pytest -q tests/
python3 scripts/visualization/render_reference_paper_figures.py -input-dir
data -output-dir manuscript/figures -atlas-output
data/cross_domain_reference_atlas.csv -manifest-output
data/figure_manifest.json
```

The reviewer-facing README maps each reported panel to its code, input, output, and validation level. Public-source panels retain source citations and acquisition contracts; the package does not assume that every external source can be redistributed or rerun without network access.