IDENTIFIABLE EVALUATION OF LEARNED SYSTEMS: FROM BEHAVIORAL COLLISIONS TO MEASUREMENT DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluation scores can be precise even when the protocol does not identify the behavioral property they are used to measure. We formulate protocol identifiability relative to a declared finite behavior class, an estimand, and a candidate measurement family. A protocol fails when two behaviors with different estimand values induce the same observable signature; such cross-estimand collisions are executable certificates of under-identification, and the collision structure itself defines a measurement-design problem: synthesize a minimum support that separates every declared alternative, or repair a failed protocol by minimum augmentation. The result is a constructive audit-to-repair workflow—expose the witness, add the cheapest distinguishing measurements, re-audit—assembled from classical hitting-set and concept-identification machinery. In a controlled language-model testbed with solver-verified labels, exhaustive audits show that base-only measurements collapse seven frozen policies, whereas a two-cell synthesized support separates them; on a fifty-policy stress class, certified synthesis identifies where 50,000 same-budget random supports all fail. Applied to classes derived from the models' own predictions, a frozen 15-cell synthesized support identifies all 86 derivation signatures, while held-out re-audit exposes one residual collision and its one-cell repair; same-budget entropy and random supports leave substantially more unresolved alternatives. A confirmatory rerun with surface-matched target and sham edits then measures a positive base-conditioned target–sham asymmetry across four model families after controlling the prespecified edit-surface features. The framework turns a validity concern into an auditable procedure for measurement design; all empirical results are controlled-testbed consequences, not external validation.

1 INTRODUCTION

A reported score is an estimate of what a protocol observed, not automatically of the behavioral property that motivated it (Raji et al., 2021; Salaudeen et al., 2025; Bowman & Dahl, 2021). This matters whenever a learned system can answer correctly for different reasons. For example, base-only questions cannot distinguish an ideal semantic updater from a policy that simply repeats its base answer; adding more base questions leaves that ambiguity intact. A target-world intervention and a matched sham provide the missing behavioral contrast.

We call this failure *under-identification*. Before estimating a behavioral quantity, one should ask whether the protocol identifies it over the plausible alternatives. We make this executable: declare a behavior class, an estimand, and a candidate measurement family; project each behavior onto the selected support; and search for cross-estimand collisions. If one exists, no estimator using only that support can recover the estimand on the declared class, and the distinguishing measurements directly define a support-synthesis problem; Figure 1 summarizes the workflow.

The paper makes three claims. **Structural (S):** finite-class identifiability can be audited mechanically by excluding cross-estimand observational collisions. **Design (D):** the same collision family induces measurement-selection and repair procedures. **Empirical (E):** in a controlled language-

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.

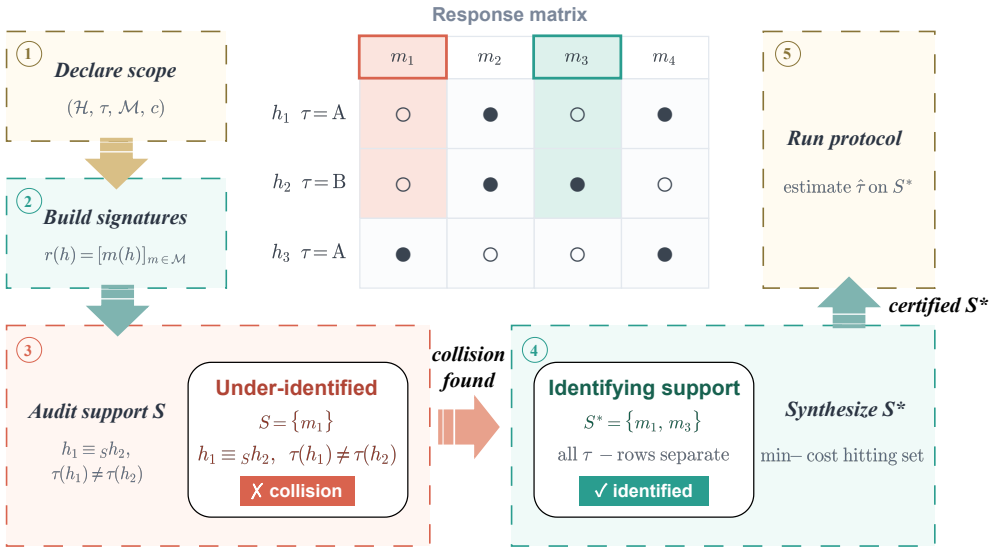


Figure 1: Generic protocol-identifiability workflow. Declare the behavior class, estimand, candidate measurement family, and costs; construct protocol-visible signatures $r(h)$; audit the selected support S for cross-estimand collisions (here $h_1 \equiv_S h_2$ with $\tau(h_1) \neq \tau(h_2)$); synthesize a minimum or minimum-cost distinguishing support; and then run the fixed protocol for estimation. In the response matrix, filled and hollow circles denote distinct observed outcomes; a shared filled/hollow pattern under m_1 is the collision, and m_3 separates it. Repetition cannot repair structural collisions. Identification remains relative to the declared $(\mathcal{H}, \tau, \mathcal{M}, S)$.

model instantiation, base correctness and intervention-response fidelity are distinct observed quantities.

Contributions. We provide an auditable workflow that separates design validity from statistical precision, with formal audit, synthesis, and minimum-augmentation repair (Proposition 1; Algorithms 1 and 2). We then show that design, not raw budget, controls identifiability: exact synthesis requires 2–15 cells across $\mathcal{H}_7$–$\mathcal{H}_{50}$, while random same-budget supports fail on the frozen stress class. Finally, a matched-control analysis finds positive target–sham asymmetry across four model families, and a model-derived bridge identifies 86 signatures before exposing and repairing a held-out collision.

2 Related Work

2.1 Evaluation Validity and Behavioral Testing

Evaluation-validity work shows scores inherit meaning from measurement design; benchmark critiques document failures from data construction and surface changes (Jacobs & Wallach, 2021; Wallach et al., 2025; Bean et al., 2025; Siska et al., 2024; Raji et al., 2021; Bowman & Dahl, 2021; Schaeffer et al., 2023; Golchin & Surdeanu, 2024; Sclar et al., 2024). Behavioral tests, contrast sets, and counterfactual probes operationalize these concerns (Ribeiro et al., 2020; Gardner et al., 2020; Wu et al., 2024; Yang et al., 2026). We instead ask whether a fixed protocol separates every declared alternative with a different estimand value, returning an unresolved pair when it does not.

2.2 Faithfulness and Causal Use of Reasoning Traces

Reasoning-trace studies distinguish plausible from faithful explanations and report reasoning–behavior dissociations under corrupted or altered traces (Wei et al., 2022; Turpin et al., 2023; Lanham et al., 2023); related analyses examine decodability, local causal use, and computational structure (Yee et al., 2024; Wu et al., 2024; Yang et al., 2026). These estimands concern traces or

particular perturbations, not identifiability of a complete behavioral property over competing strategies; we keep them distinct from our protocol-level estimand.

2.3 Causal and Model-Discriminating Measurement Design

Classical identification determines which observations identify a target, model-discriminating design separates rival models, and Q-matrix methods identify attributes under explicit diagnostic assumptions (Koopmans & Reiersøl, 1950; Hunter & Reiner, 1965; Atkinson & Fedorov, 1975; Tatsuoka, 1983; Rupp et al., 2010). Query-based identification and test-suite minimization address concept learning and coverage-preserving measurement selection (Angluin, 1988; Harrold et al., 1993); our support synthesis has the hitting-set formulation (Chvátal, 1979). Unlike these relatives, we separate relative to an estimand and protocol-visible outcomes, expose cross-estimand collisions, and use them for minimum augmentation repair (Appendix B).

2.4 Positioning of Our Framework

Our central unit is an estimand-relative collision in a declared behavior class, not an unconstrained mechanism or task-level score: the workflow audits, exposes a witness, synthesizes a separating support, or repairs one. Identification remains relative to $(\mathcal{H}, \tau, \mathcal{M})$ and does not claim to identify all natural-model mechanisms.

3 Protocol Identifiability as Measurement Design

Let $\mathcal{H}$ be a finite class of deterministic behaviors and $\mathcal{M}$ a candidate measurement family; each $m \in \mathcal{M}$ maps a behavior to a protocol-visible outcome $m(h)$, and the selected support is $S \subseteq \mathcal{M}$. The target estimand is a finite-valued function $\tau : \mathcal{H} \to \mathcal{T}$; the binary strict contract of our controlled instantiation is one special case.

Two behaviors are observationally equivalent on S, written $h \equiv_S h'$, when $m(h) = m(h')$ for every $m \in S$. The protocol identifies τ on $\mathcal{H}$ if every pair with different estimand values is separated by at least one selected measurement, following Koopmans & Reiersøl (1950).

Proposition 1 (Finite-class identifiability by collision exclusion)**.** *A finite-valued estimand τ is point-identified on S if and only if, for every $h, h' \in \mathcal{H}$, $\tau(h) \neq \tau(h')$ implies $h \not\equiv_S h'$.*

The characterization is elementary but turns an informal validity concern into an executable audit and downstream design problem. It is independent of sample size: more repeated observations cannot repair a structural collision under the selected support. Statistical precision presupposes this design condition.

A concrete collision. The failure is visible without formalism. Take two declared policies from our testbed class: an *ideal semantic updater*, which re-derives its answer after a state edit, and *target inertia*, which carries its base answer forward under the target edit. On a protocol observing only base items, both emit the correct base prediction everywhere, so their signatures coincide; the strict estimand assigns them different values (τ_{full}=1 for the updater, 0 for inertia), so no base-only estimator can recover the distinction, and more repetitions do not help. The audit returns the pair as a certificate; a single target-world measurement, whose oracle differs from the base world, separates them immediately. This witness is not hypothetical: exhaustive enumeration places all seven H7 policies in one equivalence class under base-only support, and all six cross-estimand collisions pit the ideal updater against a failing policy.

Declaring the behavior class. The class is an audit scope, not an ontology of all possible behavior. A defensible workflow: (1) declare domain-driven failure archetypes; (2) enrich them with held-out or historical signatures; (3) search adversarially for alternatives that collide under the current support; (4) freeze the class before evaluating target systems; and (5) report sensitivity as the class is enriched. Identification is always relative to $(\mathcal{H}, \tau, S)$.

Stochastic behavior. The formalization takes $m(h)$ as the outcome of measurement m under behavior h and is exact when behavior and decoding are deterministic, as in our experiments (greedy

decoding with constrained candidate emission). For stochastic systems, $m(h)$ is the cell-outcome distribution, repeated observations estimate it, and identification is relativized accordingly: a declared mixture class is separable only up to its declared granularity, since two mixtures with identical support-projected outcome distributions are observationally equivalent by construction (the stress-test family includes such a mixture; Appendix A). Finite samples add a limitation the exact audit does not model: estimated distributions can be near-identical without being equal, so exact guarantees transfer only at the declared granularity.

Two design spaces. We distinguish the canonical component family $\mathcal{P}_{\text{component}}$ (whole components included or removed) from the unrestricted cell family $\mathcal{P}_{\text{cell}}$ (arbitrary measurements from the full candidate tensor). A component can be necessary in the first family while a small cross-component cell set is sufficient in the second; these are different optimization questions, not contradictory conclusions.

4 Collision Certificates and Support Synthesis

For each cross-estimand pair (h, h'), define its distinguishing set

$$D_{h,h'} = \{m \in \mathcal{M} : m(h) \neq m(h')\}.$$

A selected support identifies τ exactly when it intersects every $D_{h,h'}$. The pipeline separates three tasks that evaluation practice usually conflates: auditing a given protocol, synthesizing a support from scratch, and repairing a failed one.

Audit. Given a support S, the audit computes the unresolved pairs

$$U(S) = \{(h, h') : \tau(h) \neq \tau(h'),\ S \cap D_{h,h'} = \varnothing\}, \tag{1}$$

and returns an audit record $\mathcal{R}(S) = (U(S), w, D_w)$, where $w = \varnothing$ and $D_w = \varnothing$ when $U(S) = \varnothing$; otherwise w is the lexicographically first pair in $U(S)$ and D_w is its distinguishing set. Thus S identifies τ exactly when the first component of $\mathcal{R}(S)$ is empty. The witness is inspectable: it names two declared behaviors that the protocol cannot tell apart, and no estimator on S can separate them.

De novo synthesis. Designing from scratch selects a minimum or minimum-cost hitting set over $\{D_{h,h'}\}$ with nonnegative costs $c(m)$, inheriting the logic of model-discriminating design (Hunter & Reiner, 1965; Atkinson & Fedorov, 1975); exact enumeration or branch-and-bound suits the small finite classes used here and can enumerate all minima; greedy variants add the measurement separating the most unresolved pairs, or the most per unit cost.

Repair by minimum augmentation. When a support S has already been collected and fails the audit, discarding it is wasteful. Repair adds the cheapest measurements that resolve every remaining collision:

$$A^* = \arg\min_{A \subseteq \mathcal{M} \setminus S} \textstyle\sum_{m \in A} c(m) \quad \text{s.t.} \quad A \cap D_{h,h'} \neq \varnothing\ \ \forall (h, h') \in U(S), \tag{2}$$

yielding the repaired support $S \cup A^*$. Repair preserves prior measurement investment; de novo synthesis does not. Throughout, an *identifying support* is any S with $U(S) = \varnothing$; *minimum*/*minimum-cost* qualify optimality; *repair* always means augmentation, never silent replacement.

Algorithms and limits. General minimum hitting set is NP-hard, so the framework does not promise scalable exact synthesis for arbitrary classes; greedy selection carries the classical set-cover guarantee. Report outputs with $|\mathcal{H}|$, $|\mathcal{M}|$, collision count, optimizer, tie-breaking, costs, and runtime.

5 Controlled Instantiation

We instantiate the framework with three-valued reasoning problems. Each cluster contains a base world, a target world whose oracle changes, and a sham world whose oracle is preserved (Figure 2). Label mappings permute TRUE/FALSE/UNKNOWN into A/B/C, with generated-choice

Algorithm 1 Audit protocol

Require: Frozen class $\mathcal{H}$, estimand τ, family $\mathcal{M}$, support S
Compute $D_{h,h'}$ for all pairs with $\tau(h) \neq \tau(h')$
$U(S) \leftarrow$ unresolved pairs by equation 1
if $U(S) = \varnothing$ **then**
    return $\mathcal{R}(S) = (U(S), \varnothing, \varnothing)$
else
    $w \leftarrow$ lexicographically first pair in $U(S)$
    return $\mathcal{R}(S) = (U(S), w, D_w)$
end if

Algorithm 2 Repair by minimum augmentation

Require: $\mathcal{H}$, τ, $\mathcal{M}$, costs c, failing support S
$(U, w, D_w) \leftarrow$ AUDITPROTOCOL$(\mathcal{H}, \tau, \mathcal{M}, S)$
if $U = \varnothing$ **then**
    return $A^* = \varnothing$, S, and (U, w, D_w)
end if
Build $\{D_{h,h'} \cap (\mathcal{M} \setminus S)\}_{(h,h') \in U}$
Solve equation 2 for A^* with declared optimizer and tie-break
$(U', w', D_{w'}) \leftarrow$ AUDITPROTOCOL$(\mathcal{H}, \tau, \mathcal{M}, S \cup A^*)$
return A^*, repaired support $S \cup A^*$, and $(U', w', D_{w'})$

and candidate-scoring readouts. All experiments use instruction-tuned models with greedy decoding (no sampling): Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct (Qwen Team, 2024; Llama Team, AI @ Meta, 2024) throughout, joined by Gemma-2-9B-IT and Mistral-7B-Instruct-v0.3 (Gemma Team, 2024; Jiang et al., 2023; Mistral AI Team, 2024) in the balanced-protocol extension; these results characterize the declared testbed and do not establish coverage of frontier-scale models; the confirmatory experiments enforce token-trie constrained candidate emission (at most eight tokens), while the pilot's generated-choice readout was free-form. Frozen revision identifiers accompany the released artifacts, and an independent solver fixes all oracle labels before inference.

The candidate family has 36 cells per model–cluster: three worlds, six mappings, and two readouts. The matched-control rerun adds direct semantic generation, giving 54 measurements per model–cluster; S_{req} denotes the complete required panel (36 cells for the structural audit, 54 for the rerun). The H7 class contains seven deterministic policies (ideal semantic updater, target inertia, any-edit reaction, fixed-label response, mapping-conditional shortcut, generated-only failure, scoring-only success), deliberately finite and diagnostic rather than a realistic enumeration of language-model behavior. A nested $\mathcal{H}_7$–$\mathcal{H}_{50}$ family is used only as a frozen synthetic stress test.

The policy-level strict estimand is $\tau_{\mathrm{full}}(h) = \mathbf{1}\{B_m(h) = T_m(h) = Q_m(h) = 1 \; \forall m \in S_{\mathrm{req}}\}$. Here B, T, Q are cell-level indicators and the strict full-support rate is the empirical fraction of units with $\tau_{\mathrm{full}} = 1$. Separately, $P(T \mid B)$ is target fidelity, $P(Q \mid B)$ sham preservation, and $P(T, Q \mid B)$ the base-conditioned joint selective rate; these are distinct from pooled descriptive selective rates.

6 Structural Audit Results

Exhaustive enumeration validates the audit without model calls: base-only support places all seven deterministic policies in one equivalence class (six cross-estimand collisions), full support yields seven classes and zero collisions (Appendix C), and removing any one of target, matched sham, paired readout, or full mapping support leaves a constructive collision witness.

In the unrestricted cell family, exact synthesis finds minimum support size two for $\mathcal{H}_7$, while only 26 of the 630 unordered two-cell subsets identify (Appendix F). Under the stated cost model, the exact two-cell support has cost four; lexicographic greedy uses three cells at cost seven and cost-aware greedy uses three cells at cost five. These figures are instance results, not universal measurement budgets.

On the frozen nested stress-test family, certified minimum support increases from 2 cells at $\mathcal{H}_7$ to 15 at $\mathcal{H}_{50}$ (Appendix E, Figure 7); uniform random same-budget selection has zero successes

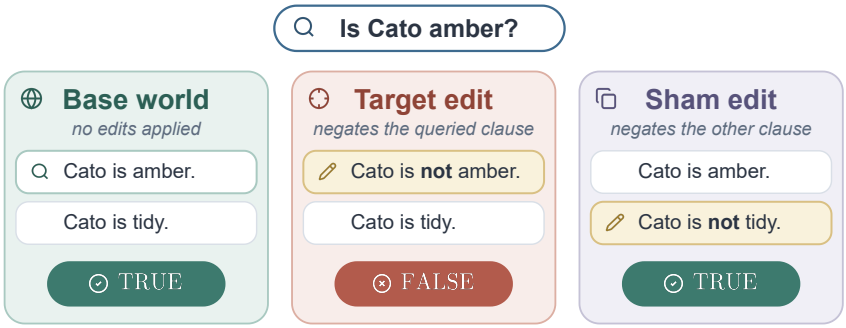


Figure 2: One frozen balanced cluster and its two surface-matched edits. The target edit negates the queried clause and flips the oracle label (TRUE→FALSE); the sham edit negates the other clause and the oracle holds. Highlighted clauses carry the edit; the badge at each card bottom is the solver-fixed oracle label.

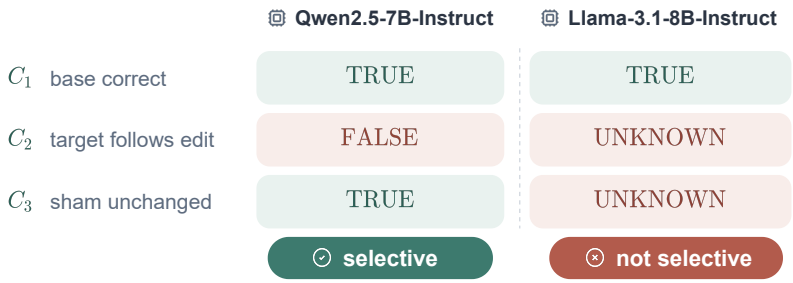


Figure 3: The strict selective contract on one measurement cell (direct-semantic readout, mapping 0, the cluster of Figure 2). Cell colour marks whether the condition is satisfied. Qwen2.5-7B-Instruct satisfies all three conditions; Llama-3.1-8B-Instruct answers the base world correctly but collapses to UNKNOWN under both edits, violating conditions 2 and 3.

in 50,000 draws at $\mathcal{H}_{50}$ (Appendix D, Figure 4). Measurement design, not raw budget, controls success on these declared classes; this does not establish that the synthetic class matches natural model behavior.

7 Empirical Consequences

The original frozen pilot contains two instruction-tuned models (Qwen Team, 2024; Llama Team, AI @ Meta, 2024), 24 diagnostic and 48 balanced clusters, six mappings, two readouts, and three worlds. Its base and selective rates, along with pilot parseability and the second-source check, are reported in Appendix I; all confirmatory panels use constrained emission.

We replace the confounded edit-type separator with target and sham edits matched on edit type, distance, length change, and lexical-cue count. The confirmatory rerun has 7,776 responses (5,184 balanced and 2,592 second-source; 1,728 and 864 cells times three world responses; each cell is model $\times$ cluster $\times$ mapping $\times$ readout). On base-correct cells, target fidelity is $P(T \mid B)$, sham preservation is $P(Q \mid B)$, and the primary asymmetry is $P(Q \mid B) - P(T \mid B)$; positive values mean greater sham preservation than target fidelity.

Table 1 shows the pattern. Conditioned on answering the base world correctly, sham preservation exceeds target fidelity in both sources, with intervals excluding zero; the direction also holds across all model–source combinations (Appendix I). The matched comparison removes the prespecified edit-type, distance, and lexical-cue separators; unmeasured cues remain outside this contrast.

The joint selective rate on base-correct cells is close to its independence null in both sources; the permutation analysis gives the same conclusion (Appendix I). Joint failure therefore carries no ex-

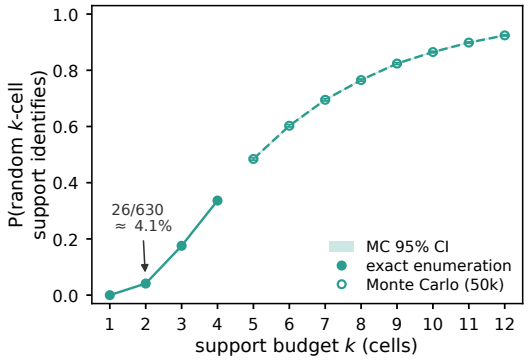


Figure 4: Random same-budget baseline on the frozen $\mathcal{H}_7$ class. Only 26 of 630 two-cell supports identify the estimand. For Monte Carlo points ($k = 5$–12), capped 95% intervals are shown; their 0.23–0.44 percentage-point half-widths are narrower than the markers.

Table 1: Base-conditioned paired intervention fidelity on the matched-control rerun; all conditional quantities share the base-correct denominator (942 and 444 cells). Asymmetry is sham preservation minus target fidelity; intervals are 2,000-replicate cluster bootstraps; the independence null is $P(T \mid B)\,P(S \mid B)$. The unconditional accuracy–selective gap (0.326 $[0.313, 0.340]$; 0.321 $[0.303, 0.340]$) mixes marginal and joint aggregation and is retained only as a diagnostic. Human audit: both sources complete (48 and 96 items).

Source	Cells	Base	Target|Base	Sham|Base	Asymmetry [95% CI]	Joint|Base	Indep. null
Balanced	1,728	0.545	0.507	0.834	0.327 [0.262, 0.393]	0.415	0.423
Second source	864	0.514	0.500	0.959	0.460 [0.368, 0.549]	0.487	0.480

cess coupling beyond error accumulation: the dissociation is carried by the target marginal, which is exactly what the asymmetry measures. We therefore treat the base-conditioned paired analysis as primary rather than headline the aggregation-sensitive unconditional accuracy–selective gap.

All three readouts have validity 1.0 in the rerun (Appendix I), and the four prespecified measured surface features are balanced between target and sham. Readout decomposition is exploratory: the direct semantic readout roughly halves the asymmetry for one model and leaves it unchanged for the other. The 48 clusters are generated deterministically by the solver at a frozen seed, so point estimates are exact for this audit population and the bootstrap quantifies variability over solver-generable clusters, not finite-population sampling.

A balanced-protocol extension to two further instruction-tuned families shows the asymmetry is not specific to one model pair. Gemma-2-9B-IT and Mistral-7B-Instruct-v0.3 run the identical frozen design with validity 1.0 on every readout. The asymmetry is positive for all four families, with model-wise estimates and intervals reported in Appendix I. Its magnitude is family-dependent, so the same frozen protocol separates families by intervention response rather than base accuracy alone.

The secondary diagnostics answer distinct questions: in the original pilot panels, conditional selective response among base-correct units is 0.138, local paired response is 0.056, and strict full-support realizability is $0/48$. The increment over counterfactual-probe studies is threefold: the sham control is surface-matched on measured edit features, the estimand is a base-conditioned paired contrast rather than a difference of unconditional rates, and the protocol is audited before the dissociation is interpreted.

7.1 From Audit to Repair on Real Model Signatures

The bridge applies the same machinery to classes derived from the models' own frozen predictions rather than to hand-designed H7. Each of the 96 model–cluster derivation units is encoded by validity and oracle-relative correctness over the 54-cell matched measurement family. The estimand here is defined directly on the realized signature (world-level correct counts plus the invalid count), so identification asks whether a sub-support separates these finite-valued summary counts—

Table 2: Same-budget support comparison on the model-derived class. Derivation: unresolved cross-estimand pairs among 3,626 (86 frozen policies). Holdout: ambiguous and separated units among 48 under a full-panel re-audit. Random row: median [IQR] over 50,000 uniform 15-of-54 draws (seed 20260817); no draw identifies. Worst ambiguity: largest undecidable set among held-out units.

15-cell support	Deriv. unres.	Hold. ambiguous	Hold. separated	Worst ambig.
Exact synthesis	0	2	95.8%	2
Coverage-greedy	1	3	93.8%	2
Entropy top-15	72	31	35.4%	4
Random (50k)	26 [19,35]	22 [17,27]	–	3 [2,3]

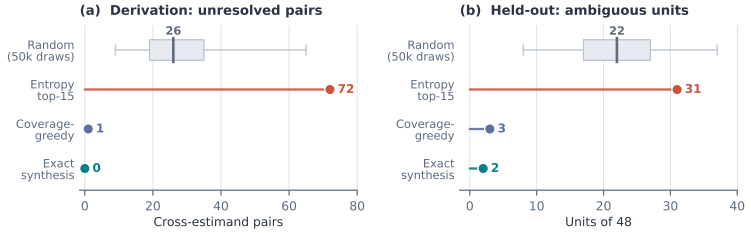


Figure 5: Same-budget support re-audit. Left: unresolved derivation pairs; right: ambiguous held-out units under full-panel re-audit. Gray band: 50,000 random 15-of-54 supports (IQR, whiskers 2.5–97.5th percentiles, tick median); lollipops: exact synthesis, coverage-greedy, entropy top-15 at 15 cells.

measurement sufficiency, not recovery of a latent mechanism—and the unit of analysis is the model–cluster pair. A two-policy seed class built from the modal signatures has exact minimum support size one; freezing all 86 distinct derivation signatures before opening the second controlled source raises the exact minimum to 15 cells and separates all 3,626 cross-estimand pairs. The held-out re-audit exposes one collision among 1,125 pairs, whose witness identifies a one-cell augmentation resolving the omission. This is a train-design–freeze–re-audit result: the held-out full panel tests the frozen certificate and informs a post-holdout repair, rather than serving as a partial-observation deployment decoder.

Is synthesis necessary at this budget? Fifteen of 54 cells is a substantial budget, so we compare collision-aware synthesis with coverage-greedy, entropy, and uniform random supports at the same budget (Table 2; Figure 5). Only the exact support identifies the derivation class; it also leaves fewer held-out ambiguities than the alternatives. The entropy baseline illustrates the failure mode: high single-cell uncertainty does not imply pair-separating power. The exact support is therefore the only design with the audit guarantee and an actionable residual witness.

The cycle—audit, enrich, freeze, hold out, redesign—consumes no new model calls. It partially answers the standard objection that a finite declared class is a strawman: the audit scope need not be invented by hand, but can be derived from observed behavior, frozen, and audited against held-out behavior. What derivation from observations alone cannot do is anticipate alternatives that were never observed; Appendix H therefore stress-tests the frozen class against adversarially constructed mixture policies that no source unit realizes, finding that identifiability is preserved at every stage while the frozen 15-cell certificate develops collisions that re-audit and re-synthesis repair at mildly growing cost (exact minimum 15 to 19 cells; greedy bound 29 cells at 5,086 policies).

8 Discussion

Three layers of validity. Evaluation validity decomposes into layers that are often conflated: (1) *structural identification*—whether the observation map separates the declared behavioral alternatives; (2) *statistical distinguishability*—whether, given identification, finite-sample estimates separate estimand values; and (3) *construct validity*—whether the estimand captures the property the community cares about. This paper targets the first layer, a precondition for the others: bootstrap in-

tervals quantify uncertainty conditional on an audited protocol, and layer 3 is addressed only locally via surface-matched sham edits and prespecified estimands.

The central practical implication is a change in the order of operations: state the target property and plausible alternatives, ask which measurements separate them, and only then optimize prompts, sample size, or estimator variance. This prevents treating more observations from a non-identifying support as evidence about an unobserved distinction: at $\mathcal{H}_{50}$, same-budget random selection succeeded zero times in 50,000 draws, while the certified 15-cell design identifies the class.

A collision is not evidence of a particular hidden strategy; it is evidence that the protocol cannot distinguish two declared strategies. The response is constructive: expose the witness, add a distinguishing measurement, and re-audit (Section 7.1 leaves one collision in 1,125 held-out pairs under a 15-cell support).

This perspective separates robustness to sampling noise, model revision, and behavioral alternatives: bootstrap intervals address the first; class enrichment the second and third. A protocol can be statistically stable and still fail the third test—the accuracy–selective gap (0.326 and 0.321) is stable yet silent about mechanism.

AI Use Statement

Generative AI was used only to translate the authors' non-native English and to draft figure visuals, not for the research ideas, theory, experimental design, implementation, analysis, or interpretation. The authors reviewed all AI-assisted work and take full responsibility for the content.

Ethics Statement

The work involves no human subjects or personal data. The main ethical risk is over-interpreting controlled measurements as claims about internal mechanisms; the manuscript scopes its claims and reports known protocol confounds.

Reproducibility Statement

The anonymous supplement bundles the frozen behavioral policies and executable generator, the audit and support-synthesis code, the matched response tables, and a provenance manifest with all analysis seeds; the definitions and procedures behind each reported quantity are specified in the main text and appendices.

References

Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1988. doi: 10.1007/BF00116828. URL https://doi.org/10.1007/BF00116828.

Anthony C. Atkinson and Valerii V. Fedorov. The design of experiments for discriminating between two rival models. *Biometrika*, 62(1):57–70, 1975. doi: 10.1093/biomet/62.1.57. URL https://doi.org/10.1093/biomet/62.1.57.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Sophia Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrø m, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott Hale, Deborah Raji, Christopher Summerfield, Philip Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 19868–19949. Curran Associates, Inc., 2025. doi: 10.52202/085713-0590.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385.

Vasek Chvátal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979. doi: 10.1287/moor.4.3.233.

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

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2308.08493.

Mary Jean Harrold, Rajiv Gupta, and Mary Lou Soffa. A methodology for controlling the size of a test suite. *ACM Transactions on Software Engineering and Methodology*, 2(3):270–285, 1993. doi: 10.1145/152388.152392. URL https://doi.org/10.1145/152388.152392.

William G. Hunter and Alfred M. Reiner. Designs for discriminating between two rival models. *Technometrics*, 7(3):307–323, 1965. doi: 10.1080/00401706.1965.10490265. URL https://doi.org/10.1080/00401706.1965.10490265.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Tjalling C. Koopmans and Olav Reiersøl. The identification of structural characteristics. *The Annals of Mathematical Statistics*, 21(2):165–181, 1950. doi: 10.1214/aoms/1177729837. URL https://doi.org/10.1214/aoms/1177729837.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Llama Team, AI @ Meta. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Mistral AI Team. Mistral 7b instruct v0.3. Hugging Face model card, 2024. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. AI and the everything in the whole wide world benchmark. In *Proceedings of the NeurIPS 2021 Datasets and Benchmarks Track (Round 2)*. Curran Associates, Inc., 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html. OpenReview: https://openreview.net/forum?id=zhahjq12BH.

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

André A. Rupp, Jonathan Templin, and Robert A. Henson. *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press, New York, 2010.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for AI evaluation. *arXiv preprint arXiv:2505.10573*, 2025. URL https://arxiv.org/abs/2505.10573.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. URL https://arxiv.org/abs/2304.15004.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11324.

Charlotte Siska, Katerina Marazopoulou, Melissa Ailem, and James Bono. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10406–10421, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.560. URL https://aclanthology.org/2024.acl-long.560/.

Kikumi K. Tatsuoka. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4):345–354, 1983. doi: 10.1111/j.1745-3984.1983.tb00212.x. URL https://doi.org/10.1111/j.1745-3984.1983.tb00212.x.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. URL https://arxiv.org/abs/2305.04388.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. URL https://arxiv.org/abs/2201.11903.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1819–1862, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.102. URL https://aclanthology.org/2024.naacl-long.102/.

Zihao Yang, Mosh Levy, Yoav Goldberg, and Byron C. Wallace. Compared to what? baselines and metrics for counterfactual prompting. *arXiv preprint arXiv:2605.01048*, 2026. URL https://arxiv.org/abs/2605.01048. Accepted at COLM 2026.

Evelyn Yee, Alice Li, Chenyu Tang, Yeon Ho Jung, Ramamohan Paturi, and Leon Bergen. Dissociation of faithful and unfaithful reasoning in LLMs. *arXiv preprint arXiv:2405.15092*, 2024. URL https://arxiv.org/abs/2405.15092.

A Frozen Policy Definitions

Table 3 lists the seven deterministic policies of the $\mathcal{H}_7$ class together with the separately analyzed stochastic mixture. The class is an audit scope: it supplies concrete cross-estimand collisions and leave-one-out witnesses, not a claim about how language models actually fail.

Table 3: Frozen behavioral policies used in the structural audits. All entries are deterministic except the stochastic mixture, which is recovered only at its declared mixture granularity: two mixtures with identical support-projected behavior are observationally equivalent by construction.

Policy	τ_{full}	Behavior on the full 36-cell support
Ideal semantic updater	1	Correct on base; follows the target oracle; keeps sham identical to base; consistent across readouts and all six mappings.
Target inertia	0	Correct on base; repeats the base answer on target and sham regardless of oracle change.
Any-edit reactor	0	Changes answer whenever the input is edited, on both target and matched sham.
Fixed-label responder	0	Always returns the same surface label; separable only once all six mappings are observed.
Mapping-conditional shortcut	0	Updates only on a subset of mappings; appears semantic on the observed mapping.
Generated-only failure	0	Fails semantically under constrained generation while answering correctly under candidate scoring.
Scoring-only success	0	Answers correctly only under candidate scoring; the symmetric counterpart of the previous policy.
Stochastic mixture	0–1	Mixes ideal-updater behavior with target inertia at a declared rate.

B Positioning Among Method Families

Table 4 contrasts our framework with the closest method families. Q-matrix methods identify latent attributes under diagnostic-model assumptions (Tatsuoka, 1983; Rupp et al., 2010); model-discriminating design selects experiments that separate rival models (Hunter & Reiner, 1965; Atkinson & Fedorov, 1975); query-based concept identification learns a target concept from membership queries (Angluin, 1988); test-suite minimization finds a small suite preserving a declared coverage (Harrold et al., 1993); and active selection optimizes the statistical efficiency of a score rather than separation of behavioral alternatives. The synthesis step is a set-cover problem with the classical greedy guarantee (Chvátal, 1979); our workflow uses collision certificates to connect identifiability diagnosis directly to protocol redesign.

How this differs from ordinary benchmark expansion. Adding more examples is useful only when the new examples refine the observation map in a way that separates unresolved cross-estimand pairs. Our synthesis objective makes this criterion explicit: an added cell has value because it hits collision witnesses, not merely because it increases the dataset size. This gives a measurable stopping rule for protocol expansion and a principled explanation for why random same-budget additions can fail.

Table 4: Positioning among method families. Our guarantee is exact identification relative to the declared $(\mathcal{H}, \tau, \mathcal{M})$; Q-matrix methods identify latent attributes under diagnostic-model assumptions; active selection optimizes statistical efficiency of a score, not separation of behavioral alternatives.

Family	Target	Alternatives	Observable object	Failure certificate	Design output
Construct-validity critiques	construct meaning of scores	–	scores, internal structure	conceptual or statistical critique	checklists, recommendations
Q-matrix / diagnostic classification	latent attribute profiles	attribute profiles	item responses	non-identifying Q (condition)	evaluate a given design
Model-discriminating design	rival models	declared rival class	experimental outcomes	–	discriminating design
Query-based concept identification	target concept	concept class	membership queries	–	query strategy
Test-suite minimization	declared coverage	fault classes	executions	coverage gap	minimized suite
Active benchmark selection	score at lower cost	–	item responses	–	adaptive item subset
Ours	estimand identification	declared finite behavior class	protocol-visible signatures	cross-estimand collision pair	minimum distinguishing support or augmentation

C Complete 16-Subset Component Lattice

Let T, S, R, and M denote target-world, matched-sham, paired-readout, and full-mapping support. Table 5 enumerates the complete power-set lattice over the canonical component family; Figure 6 visualizes it. Only full support identifies the estimand (seven classes, zero collisions); each leave-one-out support retains at least one collision, which is realized by the witness pairs listed in the figure. Removing further components only merges classes and adds collisions.

Table 5: Complete 16-subset support lattice. "Coll." counts cross-estimand policy pairs. The four three-component rows are the leave-one-out supports.

Support	$\lvert O\rvert$	Classes	Coll.	Support	$\lvert O\rvert$	Classes	Coll.
$\varnothing$	1	1	6	TS	3	3	2
T	2	2	3	TR	4	4	2
S	2	2	5	TM	12	4	2
R	2	2	5	SR	4	3	4
M	6	2	5	SM	12	3	4
RM	12	3	4	TSR	6	5	1
TSM	18	5	1	TRM	24	6	1
SRM	24	4	3	TSRM	36	7	**0**

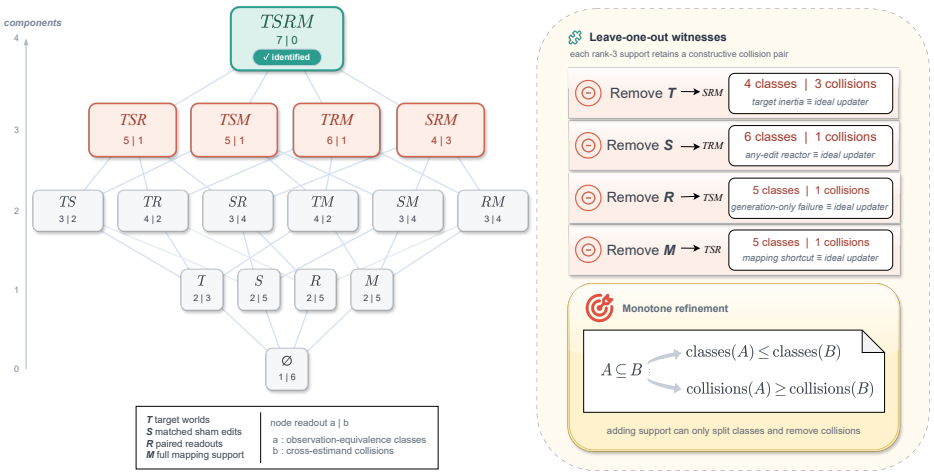


Figure 6: Support-component essentiality over the frozen seven-policy class. Each node reports its observation-equivalence class count and cross-estimand collision count $(a \mid b)$; only full support $TSRM$ is identified. The four rank-3 nodes are the leave-one-out supports; the right panel lists a constructive collision witness retained by each, and class and collision counts are monotone under support inclusion.

D Random Support Baseline

Table 6 reports the full random same-budget curve on the frozen class. For $k \leq 4$ all $\binom{36}{k}$ subsets are enumerated exactly (36, 630, 7,140, and 58,905); for $k \geq 5$ probabilities are 50,000-draw uniform Monte Carlo estimates over unordered subsets (seed 20260815) with binomial standard errors. Because frozen policy responses are deterministic, the only randomness is the subset draw; Monte Carlo error approximates the random baseline and is not model uncertainty.

Table 6: Random same-budget baseline, $k = 1$–12. "Subsets" is $\binom{36}{k}$ where enumerated exactly; standard errors use four decimal places. "Mean" and quartiles summarize residual cross-estimand collisions. The certified minimum budget is $k = 2$, where 26 of 630 subsets identify (4.13%).

k	Subsets	Mode	$P(\text{id})$	Mean	q25	Med.	q75
1	36	exact	0.0000	3.72	3	4	5
2	630	exact	0.0413	2.43	1	2	3
3	7,140	exact	0.1755	1.64	1	1	3
4	58,905	exact	0.3364	1.14	0	1	2
5	—	MC	0.4845 ± 0.0022	0.80	0	1	1
6	—	MC	0.6025 ± 0.0022	0.57	0	0	1
7	—	MC	0.6949 ± 0.0021	0.41	0	0	1
8	—	MC	0.7651 ± 0.0019	0.30	0	0	0
9	—	MC	0.8239 ± 0.0017	0.21	0	0	0
10	—	MC	0.8648 ± 0.0015	0.16	0	0	0
11	—	MC	0.8985 ± 0.0014	0.11	0	0	0
12	—	MC	0.9241 ± 0.0012	0.08	0	0	0

E Policy-Class Scaling Details

Family construction. The nested family $\mathcal{H}_7 \subset \mathcal{H}_{10} \subset \mathcal{H}_{20} \subset \mathcal{H}_{30} \subset \mathcal{H}_{50}$ extends the frozen class with deterministic policies from a pre-declared failure-mask grammar: single-axis masks over world roles, mappings, and readouts, plus second-order world×mapping, world×readout, and mapping×readout interactions. Candidates are admitted by a fixed rule order and deduplicated by full 36-cell panel signature, so no model responses inform membership. Each layer declares exactly one τ_{full}=1 policy; every member has a unique panel signature and every cross-estimand pair

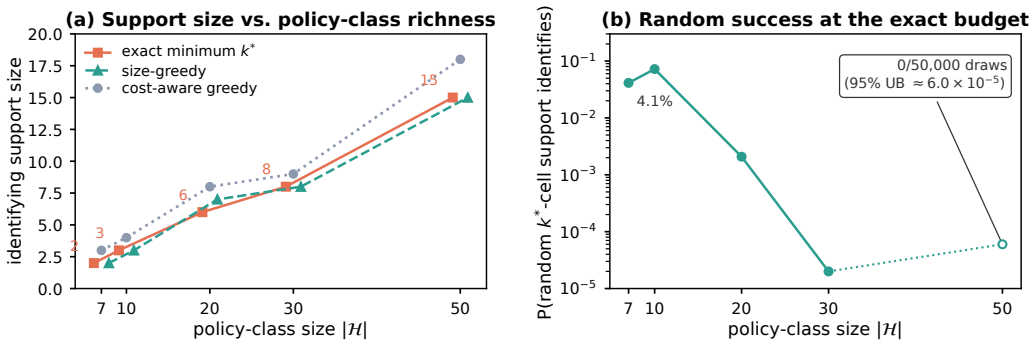


Figure 7: Policy-class scaling. (a) Identifying support size by exact synthesis, size-greedy, and cost-aware greedy on five nested frozen policy classes. (b) Probability that a uniformly random support of the exact minimum budget identifies (log scale); $\mathcal{H}_{50}$ has zero successes in 50,000 draws (open marker: rule-of-three 95% upper bound). The class family is frozen before any model data.

a nonempty distinguishing set, so identification is feasible by construction and the stress test measures *how much support it costs*.

Solvers. Exact minimum sizes are certified by branch-and-bound hitting-set search with an admissible max-cover lower bound; complete enumeration of all minima is attempted only when $\binom{36}{k} \leq 2 \times 10^5$, which holds at $\mathcal{H}_7$ (26 minima) and $\mathcal{H}_{10}$ (516). The size-greedy selector adds the cell covering the most unresolved pairs (coverage-then-index tie-breaking); the cost-aware variant maximizes newly covered pairs per unit cost under the stated cost model, in which a constrained-generation measurement costs 1 and a candidate-scoring measurement costs 3. Random success at the exact budget uses exact enumeration where feasible and 50,000 uniform unordered draws otherwise (seed 20260815).

Tie-break sensitivity. On the seven-policy class, lexicographic tie-breaking yields a size-3 greedy support at cost 7, while the coverage-then-index rule attains the certified optimum (size 2, cost 4); the main text reports the lexicographic variant and Table 7 the coverage-then-index variant. Greedy outcomes on this instance are therefore tie-break sensitive, which is one reason certified optima rather than heuristic sizes serve as the reference.

Table 7: Policy-class scaling, full detail. "Greedy" and "CAG" (cost-aware greedy) give support size (cost); "Cost" is the exact minimum cost at size k^*. "Minima" is the complete count where enumeration is feasible. Random $P(\mathrm{id})$ at the exact budget: exact enumeration ("ex.") or 50,000-draw Monte Carlo ("MC"); the $\mathcal{H}_{50}$ bound is the rule-of-three 95% upper bound.

Class	k^*	Minima	Greedy	CAG	Cost	Random $P(\mathrm{id})$
$\mathcal{H}_7$	2	26	2 (4)	3 (5)	4	4.127% (ex.)
$\mathcal{H}_{10}$	3	516	3 (5)	4 (6)	5	7.227% (ex.)
$\mathcal{H}_{20}$	6	—	7 (9)	8 (10)	8	0.208% (MC)
$\mathcal{H}_{30}$	8	—	8 (14)	9 (13)	12	0.002% (MC)
$\mathcal{H}_{50}$	15	—	15 (27)	18 (30)	27	$<6\times10^{-5}$

F THE 26 MINIMUM SUPPORTS AND SUPPORT-LOCAL SENSITIVITY

Every two-cell minimum on the frozen class pairs one target-world constrained-generation cell with one sham-world candidate-scoring cell (exact cost 4 under the stated cost model); no cell is mandatory and 25 of the 36 cells appear in no minimum. Because the class admits 26 distinct minima, the two-cell conclusions do not depend on the choice of one of them. Structurally, every minimum identifies by construction. Empirically, we recompute two support-local diagnostics on the frozen 24-cluster response panels for each of the 26 minima and both models: the rate at which both cells yield valid responses, and the rate at which both are oracle-correct (Table 8). The diagnostics vary

within a narrow band for one model and remain uniformly low for the other, so the cross-model contrast under a two-cell support is stable across the entire set of minima. The audit thereby separates structural identification from empirical realizability: a support can identify the estimand over the declared class while frozen model responses realize the required observations at very different rates.

Table 8: Support-local diagnostics across all 26 minimum supports (24 clusters per support–model pair; all panels complete).

Model	Diagnostic	Min	Median	Mean	Max
Qwen	both-valid	0.708	0.792	0.811	1.000
Qwen	both-oracle-correct	0.000	0.083	0.083	0.250
Llama	both-valid	0.000	0.000	0.058	0.250
Llama	both-oracle-correct	0.000	0.000	0.000	0.000

G Model-Derived Enrichment Bridge: Freeze, Support, and Witness

The bridge derives behavior classes from frozen predictions of the matched-control rerun; it makes no model calls. The measurement family is the matched 54-cell tensor (three worlds $\times$ six mappings $\times$ three readouts). A unit is one model–cluster pair; its signature encodes, per ordered cell, the response status together with oracle-relative correctness (null when invalid). The estimand is the four-tuple of world-level correct counts (each out of 18) plus the invalid count (out of 54), a finite-valued function of the signature. The derivation split contains 96 units from the balanced source; the holdout contains 48 units from the second controlled source.

Freeze protocol. The seed class takes one modal full-signature per model (two policies). The derived class takes all unique full signatures observed in the derivation split (86 policies, 62 distinct estimands). Both stages, their exact supports, and the analysis seed are fixed before the holdout is read. Table 9 summarizes the stages; the exact solver is the same bitmask minimum hitting-set core used elsewhere in the paper.

Table 9: Bridge stages. "Pairs" counts cross-estimand policy pairs in the frozen class; the exact minimum identifying support grows from one cell to 15 as the class is enriched from seed to derived.

Stage	Policies	Estimands	Pairs	Min. support
Seed (modal)	2	2	1	1 cell
Derived (frozen)	86	62	3,626	15 cells

Selected-support audits. Auditing each stage's support over units rather than policies: the one-cell seed support leaves 2,228 collisions among 4,516 cross-estimand derivation-unit pairs and 551 among 1,125 holdout pairs; the 15-cell derived support separates the derivation split completely (0 of 4,516) and leaves exactly one collision among 1,125 holdout cross-estimand pairs. The 15-cell certificate consists of direct-semantic-generation cells at mappings 0 and 3–5 interleaved with candidate-scoring cells across base, sham, and target worlds.

Same-budget baselines: full detail. Table 10 reports the complete comparison behind the main-text summary (Table 2). All methods use exactly 15 of the 54 cells. Coverage-greedy adds the cell resolving the most unresolved pairs per step (lexicographic tie-break); under uniform costs the cost-aware variant coincides with it on this class. The entropy baseline ranks cells by outcome entropy over the 86 derivation signatures and takes the top 15; it is collision-unaware by construction. Random draws are uniform 15-of-54 subsets (50,000 draws, seed 20260817). For the random family we report the full distribution rather than a point: on derivation unresolved pairs the median is 26 (IQR 19–35, 95% interval 9–65, range 1–238, zero-success fraction $0/50{,}000$); on held-out ambiguous units the median is 22 (IQR 17–27, 95% interval 8–37, range 1–47); on holdout within-source collision pairs the median is 8 (IQR 5–12) and versus the frozen derivation class 24 (IQR 16–36).

The exact support lies at the 0.0 percentile of the random derivation distribution and the 0.022nd percentile on held-out ambiguous units. Worst ambiguity sets reach 2 policies for exact and greedy and 7 for entropy on the derivation class (2, 2, and 4 on the holdout; random median 3). Evaluation metrics follow the audit semantics: a holdout unit is *ambiguous* when it collides either with another holdout unit or with a frozen policy of different estimand under the projected support; the *worst ambiguity set* is the largest undecidable equivalence class; *holdout separation* is the fraction of holdout units whose estimand is uniquely determined under the full-panel re-audit.

Table 10: Full same-budget baseline detail on the model-derived class (derivation: 86 policies, 3,626 cross-estimand pairs; holdout: 48 units, 1,125 cross-estimand pairs). "Within"/"vs. frozen" are held-out collision pairs within the holdout source and against the frozen derivation class; random rows summarize 50,000 draws (seed 20260817).

Method	Deriv. unres.	Within pairs	vs. frozen pairs	Ambig. units	Holdout separated
Exact (frozen)	0	1	0	2	95.8%
Coverage-greedy	1	1	1	3	93.8%
Entropy top-15	72	18	75	31	35.4%
Random median	26	8	24	22	–
Random IQR	19–35	5–12	16–36	17–27	–
Random best	1	0	0	1	–

Exploratory post-holdout augmentation. After inspecting the witness, a single additional measurement at mapping 2 removes the residual collision; six one-cell options exist, spanning base/sham/target crossed with candidate scoring and constrained generation, and one minimum is (base, mapping 2, candidate scoring). Because this choice was made after the holdout was opened, it is reported as exploratory redesign and is not part of the frozen 15-cell certificate.

Scope. The bridge demonstrates class enrichment and transport auditing on real prediction tensors within the solver-generated testbed. It excludes claims of natural-benchmark transport, external natural validation, and membership of observed models in $\mathcal{H}_7$.

H Adversarial Enrichment Stress Test

Deriving the audit class from observed signatures cannot anticipate alternatives that were never observed. This appendix closes that gap constructively: we generate deterministic mixture policies that no source unit realizes, add them to the frozen derived class, and re-run the audit and synthesis pipeline end to end. Zero model calls are required, because a mixture policy is a declared alternative within the same 54-cell measurement family.

Mixture construction. Each mixture combines two observed signatures: world-contingent mixtures let one world (base, sham, or target) follow parent A while the other two follow parent B, and mapping-contingent mixtures let mappings 0–2 follow one parent and 3–5 the other (both directions). Over all 3,655 observed-signature pairs this yields 18,275 candidate mixtures, of which 7,372 are not already realized by an observed unit; the remainder coincide with observed signatures, indicating substantial overlap between the mixture family and the observed class. Random draws (seed 20260805) add k mixtures per stage; Table 11 reports the enrichment schedule and audit results at each stage.

Findings. Three results matter for the framework's semantics. First, identifiability is preserved at every stage: the estimand is a deterministic function of the full signature and the 54-cell family separates every pair, so no empty distinguishing set ever arises; adversarial enrichment raises the cost of identification rather than destroying it. Second, the frozen 15-cell certificate is not closed under enrichment: three collisions appear already at $k{=}25$ and 6,664 at $k{=}5{,}000$. This is the class-relative semantics operating as designed—a certificate certifies exactly the declared class, and an enriched scope demands a re-audit. Third, the repair cost grows mildly: the exact minimum rises from 15 to 19 cells across a doubling of the class, and the greedy bound reaches 29 cells at 5,086 policies and $1.3{\times}10^7$ cross-estimand pairs in under 21 seconds, matching the exact optimum where both are

Table 11: Adversarial enrichment of the frozen derived class. "Frozen collisions" counts cross-estimand pairs colliding under the frozen 15-cell certificate; k^* is the exact minimum identifying support of the enriched class (exact synthesis capped at 186 policies by runtime); "Greedy" is the coverage-greedy upper bound.

Added	Class	Cross-τ pairs	Frozen collisions	k^* (exact)	Greedy
0	86	3,626	0	15	–
25	111	6,073	3	15	18
50	136	9,136	9	16	17
100	186	17,148	17	19	19
200	286	40,626	28	–	21
400	486	117,507	84	–	23
800	886	390,998	232	–	25
5,000	5,086	12,897,632	6,664	–	29

available (19 versus 19 at k=100; 18 versus 15 and 17 versus 16 at smaller stages). Exact synthesis completes in 28–129 seconds on classes of 111–186 policies, which bounds its certification role to moderate scopes. The audit–enrich–re-audit cycle therefore scales through its greedy path.

I Matched-Control Integrity, Readout Breakdown, and Human-Audit Status

The confirmatory rerun executes 7,776 measurements (5,184 balanced; 2,592 second controlled source) with seed 20260805 and fixed model revisions; every bundled artifact passes integrity verification, all combination grids are complete, no rows are reused, and machine audits of formal-oracle labels and surface matching pass. Response validity is 1.0 for all three readouts on both sources. Table 12 reports the balanced-source readout breakdown behind the pooled descriptive rates (accuracy 0.553, selective 0.226) that underlie the accuracy–selective gap diagnostic retained in Table 1; uncertainty for the gap uses 2,000-replicate cluster bootstrap (unit: cluster; seed 20260805).

Table 12: Balanced-source readout breakdown (48 clusters, pooled over models and mappings). The strict selective contract is hardest under opaque readouts and least hard when the model can state the semantic label directly; the accuracy–selective gap holds in every readout.

Readout	Accuracy	Selective	Validity
Candidate scoring	0.459	0.111	1.0
Constrained opaque generation	0.459	0.109	1.0
Direct semantic generation	0.740	0.458	1.0
Pooled (descriptive)	0.553	0.226	1.0

Item-level human semantic audit status: the second controlled source is complete (48 items; two independently randomized blinded reviews, raw agreement 46/48 $= 0.958$, Cohen's $\kappa = 0.936$; two disagreements and three post-unblinding shared-label items adjudicated; final formal-oracle match 48/48). The matched balanced source is complete (96 items; two independently randomized blinded reviews, raw agreement 93/96 $= 0.969$, Cohen's $\kappa = 0.952$; three disagreements and five post-unblinding shared-label items adjudicated; final formal-oracle match 96/96). All eight adjudicated items attribute FALSE where the formal semantics gives UNKNOWN: reviewers occasionally read one affirmed property as denying another, an exclusivity inference the formal semantics does not license.