N×N E-VALUATION: HYPOTHESIS CERTIFICATION WITHOUT PER-HYPOTHESIS CONDITIONAL RESAMPLING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

High-throughput discovery pipelines produce hypotheses faster than a conditional null can be designed for each, making certification a bottleneck. Held-out validation shows whether a hypothesis predicts but not how: the same score can come from population prevalence, a pattern shared across the scope, a unit-matched restriction, or an association carried by a minority. Conditional randomization tests must specify, for each hypothesis, its explanatory variables, conditioning set and resampling law. We introduce N×N e-valuation, in which the units governed by one hypothesis serve as references for one another. It tests whether held-out success is carried by the hypothesis's scope, by its restriction, or by which unit receives which restriction, with no resampling model written per hypothesis, and reports which of these carries the prediction. E-values give finite-sample, per-hypothesis control of false certification under stated exchangeability nulls, and an asymptotically valid equivalence test certifies sharedness only when restriction differences above a declared noise ceiling are rejected. On planted and randomized-intervention data, N×N certifies shared associations that a CRT and an exact permutation test reject; on these and on real data with expert-audited labels, it rejects minority-carried associations they certify: their nulls ask whether any unit-level dependence exists, not whether an association holds across its scope. Removing any of its statistics merges a verdict that certifies with one that rejects. The certificates concern predictive association, not causality.

1 INTRODUCTION

Large language models now propose hypotheses faster than anyone can check them, and many are wrong (Alkan et al., 2025). An LLM-based system needs a certification step that asks of the proposer only what it does reliably—stating a hypothesis—and returns feedback, since LLMs rarely correct their own reasoning unaided (Huang et al., 2024). Hence we propose three requirements for an LLM-compatible certification algorithm—machine-automatable, attributable and diagnostic—and N×N e-valuation, an algorithm that meets them.

A certification algorithm is *machine-automatable* if a machine can carry out every per-hypothesis step with the guarantee intact. A conditional randomization test (CRT) is not, in general: it needs a resampling model written for each hypothesis (Candès et al., 2018), which an LLM can write but cannot ensure is right, and a wrong model silently voids the guarantee. An algorithm is *attributable* if what it certifies is a stable, scope-level hypothesis association: across the scope, the hypothesis, not prevalence, carries the prediction—an association, not causality. Held-out validation is not: a held-out gain can be a restated base rate (Ferrari Dacrema et al., 2019) or popularity bias (Abdollahpouri et al., 2019). An algorithm is *diagnostic* if its output helps the proposer revise a hypothesis or keep the part that works, for instance by saying which part of it—the scope, the restriction or their pairing to units—carries the association. A CRT and a permutation test are not: both return only certify or reject. None of these meets all three.

N×N e-valuation is proposed to meet all three. It is machine-automatable because the units under a hypothesis serve as nulls for one another, replacing the one step of a CRT that is not: the construction of a null for each hypothesis. Following the CRT's idea, it is attributable: altering a component of

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.

the hypothesis—its scope, or which unit receives which restriction—changes the prediction, and its formulation (Table 1) makes the association scope-level. It is diagnostic because its verdict is richer than certify or reject: its verdicts distinguish the failures most common in the field—prevalence, a scope-level effect read as individual (Robinson, 1950), a latent common cause (Greenland et al., 1999) and a lift carried by a subgroup (Simpson, 1951; Jadidinejad et al., 2021)—and each suggests how to revise the hypothesis (Table 12).

Our contributions, for datasets whose units are drawn independently, are two: the requirements and N×N e-valuation, to our knowledge the first certification algorithm meeting all three. Because the requirements are ours, no existing algorithm targets them; the best we found, a model-X CRT and an exact permutation test, return only certify or reject, so only certify decisions are compared. On planted, randomized-intervention and expert-labeled real data N×N issues no false certificate; it certifies planted shared effects that both baselines—each valid for its own null—reject, and rejects subgroup-carried ones they certify. Its verdicts, which the baselines do not produce, are scored against the audited types for N×N alone, and Table 12 shows how each suggests a next step.

2 Related work

Conformal e-values. The e-value is the safe-testing construction (Grünwald et al., 2024; Ramdas et al., 2023), and scoring an observation against an exchangeable pool is its conformal instance (Vovk & Wang, 2021; Vovk, 2015; Bates et al., 2023). Cross-conformal prediction and the jackknife+ (Vovk, 2015; Barber et al., 2021) reuse a pool to calibrate a predictor, not to test a hypothesis; conformal test martingales (Vovk, 2021) test whether a *stream* stays exchangeable, not a fixed pairing on a set in hand. E^{sig} is that standard construction, and E^{uniq} is the group-invariance e-value for a permutation null (Koning, 2023; Pérez-Ortiz et al., 2022)—the softmax likelihood ratio when the group is the symmetric group.

Machine proposers and their verifiers. Discovery benchmarks supply labelled instances to *measure* verification against, not a way to perform it on a new proposal (Majumder et al., 2025). Self-verification is cheapest and inherits the proposer's failure mode: models do not reliably correct their own reasoning without an external signal (Huang et al., 2024; Kamoi et al., 2024). Trained verifiers (Lightman et al., 2024) supply one, but for *reasoning steps* rather than a claim about a population. Both remain unbounded model judgements. We therefore compare against tests that take the same inputs we do.

Randomisation tests. The permutation test (Fisher, 1935; Pitman, 1937) draws its reference distribution from exchangeability rather than from a model. The model-X CRT (Candès et al., 2018) resamples the covariate from $P(X_j \mid X_{-j})$; the conditional permutation test (Berrett et al., 2020) relaxes exactness of that law, and sequential model-X (Shaer et al., 2023) recasts the family as betting. Some restriction on the covariate law is unavoidable—without one, no test of conditional independence has nontrivial power against all alternatives (Shah & Peters, 2020)—and model-X supplies it by assumption. All three need that conditional structure per hypothesis; the pairing null needs none, and no such test can express a unit-invariant hypothesis at any budget (Section B.2).

Matrix nulls. Mantel (1967) permutes the pairing between two distance matrices; Xu et al. (2026) build a $K \times K$ matrix from a learned critic and tests its diagonal mean against column permutations. Both share the *mutual-null* construction, samples serving as nulls for each other. The same idea was reinvented by us without a hypothesis constraint, and E^{uniq} is their diagonal score up to a normaliser. They use the off-diagonal only for their own purposes; we read it with three statistics: E^{uniq} against the in-scope row mean, the invariance call on the variance structure, cover on the per-unit split. Table 10 derives which verdicts merge without each; a permutation test on the same matrix, lacking all three, rejects most *Shared* samples. To our knowledge that combination is unoccupied.

Other work. Significant pattern mining asks the same question with multiplicity control: Tarone's testability bound (Tarone, 1990) excludes patterns that cannot reach significance, LAMP (Terada et al., 2013) makes that a corrected threshold, and Westfall–Young light (Llinares-López et al., 2015) reaches FWER control by permutation at scale. All of them are coupled to the enumeration.

Table 1: Objects in the certification problem, with the running recommendation example. Both apply and the restriction it returns may be non-deterministic and noisy.

Object	Notation	Meaning	Example
Unit	i	a member of the population	a user
Test set, scope	T; $U = \{i : \text{in_scope}(x_i)=1\} \subseteq T$	T is the evaluation population	all users; those matching it
Input	x_i	what is known about a unit	interaction history
Outcome	y_i	the held-out target: a *set* of $\lvert y\rvert$ elements	the next $\lvert y\rvert$ interactions
Hypothesis	$H = (\text{in_scope}, \text{apply})$	a scope test plus apply	"a phone-buyer buys a case"
Apply	$r = \text{apply}(x)$	normally the *expensive* step; returns a restriction	compile the pattern to a filter
Restriction	$r, r(y) \geq 0$	what apply returns; its *cheap* score rates y	how well y fits it
Collection	$A_B = \{a_i\}_{i \in B}$	$Y_U = \{y_i\}_{i \in U}$; similarly R_U	outcomes over U

Algorithm 1 $N \times N$ E-valuation: certify hypothesis H at level α

Require: $H = (\text{in_scope}, \text{apply})$, data $\{(x_i, y_i)\}_{i \in T}$, level α, thresholds $\tau_{\text{cov}}, \gamma$, noise ceiling k

$U \leftarrow \{\, i \in T : \text{in_scope}(x_i) = 1 \,\}$ ▷ in-scope subset $U \subseteq T$

if $|U| > \gamma|T|$ **return** *base rate* ▷ a scope over γ of T is the population

$V \leftarrow |V|$ units drawn uniformly from $T \setminus U$ ▷ out-of-scope reference, $|V| = |U|$ by default

for $i \in U$ **do**

  $r_i \leftarrow \text{apply}(x_i)$ ▷ restriction; normally the expensive call, once per unit ($O(|U|)$)

end for

for $i \in U,\ j \in U$ **do**

  $M[i][j] \leftarrow r_i(y_j)$ ▷ cheap score; diagonal $M[i][i] = r_i(y_i)$ ($O(|U|^2)$)

end for

$E^{\text{sig}} \leftarrow \frac{1}{|U|} \sum_{i \in U} r_i(y_i) / g_0^{(r_i)}$ ▷ significance: diagonal vs. out-of-scope rate

$E^{\text{uniq}} \leftarrow \frac{1}{|U|} \sum_{i \in U} r_i(y_i) \big/ \big(\frac{1}{|U|} \sum_{j \in U} r_i(y_j)\big)$ ▷ uniqueness: diagonal vs. in-scope row mean

$\widehat{D},\ \widehat{\text{se}} \leftarrow \widehat{S}_{\text{r}} - (k\bar{M})^2$ off the diagonal, and its jackknife error ▷ Section 4; $O(|U|^2)$

$\text{shared} \leftarrow [\, \widehat{D} + z_{1-\alpha}\widehat{\text{se}} < 0 \,]$ ▷ the invariance call (Equation (8))

$\text{cover} \leftarrow \frac{1}{|U|}\big|\{j \in U : r_j(y_j) > r_j(y_{v_j})\}\big|$, $p_{\text{cov}} \leftarrow \Pr\big(\text{Bin}(|U|, \tau_{\text{cov}}) \geq |U|\,\text{cover}\big)$ ▷ Equation (4)

return $\text{verdict}(E^{\text{sig}}, E^{\text{uniq}}, \text{shared}, p_{\text{cov}})$ per Section 3.2 ▷ certify iff *Matched* or *Shared*

3 Method: The $N \times N$ E-Value Matrix Certifier

A hypothesis is a pair $H = (\text{in_scope}, \text{apply})$ of functions of the input: $\text{in_scope}(x)$ decides membership, apply returns a *restriction* r scoring an outcome as $r(y) \geq 0$. Table 1 fixes the notation and Algorithm 1 the procedure. Every quantity here can be statistical—restrictions may be sampled, scores estimated, outcomes noisy—so the thresholds are tolerance dials the user sets (Section D).

3.1 Concepts

Two exchangeability statements replace the per-hypothesis resampling model: no test is constructed per hypothesis, but something is assumed of the data. Whether either holds is domain-specific—temporal drift, for instance, breaks H_0^{sig}—so the method removes the burden of *constructing* a null, not the burden of *checking* that one applies.

Significance null. Let the *reference* V be units drawn uniformly from outside the scope, $T \setminus U$, without reading an outcome—$|U|$ of them by default. H_0^{sig}: *for every* $j \in U$, given r_j and $Y_{V \cup \{j\}}$, the outcome y_j is *exchangeable* with the members of $Y_{V \cup \{j\}}$—in scope, the restriction fits no better than outside it.

Uniqueness null. H_0^{uniq}: given R_U and Y_U, the pairing of restrictions to outcomes is exchangeable.

Significance e-value. For unit j, with $r_j = \mathrm{apply}(x_j)$:

$$E_j^{\mathrm{sig}} \;=\; \frac{r_j(y_j)}{g_0^{(r_j)}}, \qquad g_0^{(r_j)} \;=\; \frac{1}{|V|+1} \sum_{k \in V \cup \{j\}} r_j(y_k), \tag{1}$$

with one reference shared by every j. A restriction scoring 0 on every outcome in its pool, its own included, makes the ratio $0/0$; we set both e-values to the null value 1 there: with $g_0^{(r_j)} = 0$ the numerator is zero too, since j is in the pool, so the convention is forced rather than chosen.

Uniqueness — the $N \times N$ matrix. Over the claimed population (size $|U|$), compute each restriction *once*, then form

$$M[i][j] \;=\; r_i(y_j). \tag{2}$$

The per-unit *conformal uniqueness e-value*—the group-invariance e-value for a permutation null (Koning, 2023; Pérez-Ortiz et al., 2022)—is

$$E_j^{\mathrm{uniq}} \;=\; \frac{M[j][j]}{\frac{1}{|U|}\sum_{k \in U} M[j][k]}. \tag{3}$$

The invariance call is the third test, an equivalence test at asymptotic level α; the coverage test is the fourth, exact at level α (Section D.2). Table 2 states what each certificate requires. The α bound on the lift comes from the e-values alone: cover and the call only shrink the sets they admit, and neither can promote a hypothesis the e-values did not clear. What the call adds to a *Shared* certificate is an attribution—one restriction for the whole scope—which it bounds separately (Section 4).

Invariance. A hypothesis is *invariant* over its scope when its units share one restriction; every column of M then applies that restriction to one outcome and is constant up to the restriction's own variation across units. The user declares a ceiling on that variation, k times the mean off-diagonal score, which fixes what counts as one restriction: restrictions whose scores on a common outcome spread by less than the ceiling are one restriction with noise. The statistic is that spread less the squared ceiling, $\widehat{S}_{\mathrm{r}} - (k\bar{M})^2$, with $\widehat{S}_{\mathrm{r}}$ the mean variance of M's columns and $\bar{M}$ the mean cell, both with the diagonal left out (Section 4); the call is *shared* when its upper confidence bound falls below zero, *variant* when its lower bound exceeds zero, and *undecided* otherwise.

Coverage. Each unit j is paired with a comparator v_j, a distinct member of V drawn at random:

$$\mathrm{cover} \;=\; \frac{1}{|U|}\,\big|\{\, j \in U : r_j(y_j) > r_j(y_{v_j}) \,\}\big|, \qquad p_{\mathrm{cov}} \;=\; \Pr\big(\mathrm{Bin}(|U|, \tau_{\mathrm{cov}}) \geq |U|\,\mathrm{cover}\big). \tag{4}$$

cover can also be used to improve hypothesis proposing. For a hypothesis rejected for low coverage but high significance, the proposer may try to restrict the scope to make the hypothesis valid. To avoid information leak, the coverage information and significance details in the feedback should be re-computed on the development set the proposer always sees. The raw information from the certification set should be held unseen by proposers, like a validation or test set.

Reading the four. Figure 1 shows M under the four regimes. Both e-values ask whether a restriction scores its own unit's outcome above a random one's, and differ only in the pool that random unit is drawn from. E^{sig} draws from outside the scope, making it a randomization test against the base rate among units the hypothesis does not claim: a hypothesis with no lift gets $E^{\mathrm{sig}} \approx 1$, and only signal beyond that base rate accumulates evidence. The scope is no part of its own baseline, so its width does not cap E^{sig} (Section D.3). E^{uniq} draws from the scope U, randomizing the *pairing* rather than the covariate, so unlike the model-X CRT (Candès et al., 2018) it needs no model of $P(X_j \mid X_{-j})$ (Section 5 compares the two); a hypothesis that fails it is not *Matched*. The invariance call asks whether the restrictions in scope U score a common outcome within the declared ceiling of one another, so that one restriction explains every unit. And cover reads the diagonal per unit rather than in aggregate: unless $p_{\mathrm{cov}} \leq \alpha$ shows that more than a fraction τ_{cov} of units outscore their comparators, it withholds the certificate and names the case *Sparse*, giving the proposer a hypothesis to *re-scope* and re-propose.

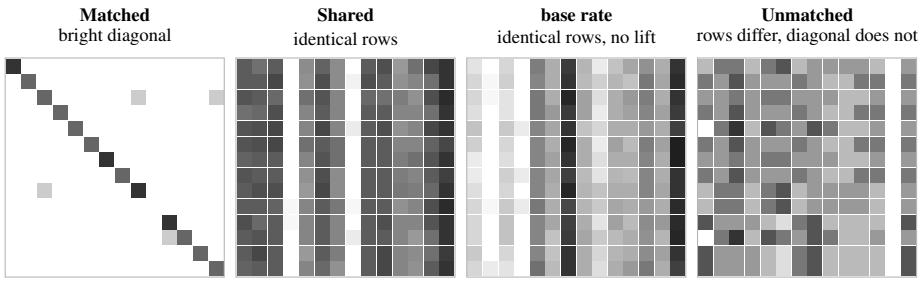

Figure 1: The cross-prediction matrix $M[i][j] = r_i(y_j)$ under the four regimes, on 14 in-scope units; darkness is the score, normalised per panel, and all cells may carry noise. The regimes are defined in Table 2. *Matched*: each unit's own restriction fits its own outcome, so only the diagonal is dark. *Shared*: one restriction over the whole scope—identical rows, and in scope real lift. *Base rate*: identical rows too, but satisfied as often outside the scope, so no lift. *Unmatched*: restrictions differ, none matched to its own outcome, so no diagonal stands out.

3.2 CERTIFICATION

Proposition 1 (Significance validity)**.** *E_j^{sig} is an exact e-value for the exchangeability null H_0^{sig}: under a* random *(no-lift) hypothesis $\mathbb{E}[E_j^{\mathrm{sig}}] = 1$, so only genuine lift over the base rate drives it above* 1.

Proposition 2 (Uniqueness validity, conformal)**.** *E_j^{uniq} is an exact e-value for the exchangeability null H_0^{uniq}: under a* random *(no-mechanism) hypothesis $\mathbb{E}[E_j^{\mathrm{uniq}}] = 1$, so only genuine per-unit structure drives it above* 1.

Both nulls are statements about the scope, not about one unit: H_0^{sig} quantifies over $j \in U$ and H_0^{uniq} is a single claim about the pairing. That is the right shape, because the verdict is a claim about the scope—and a hypothesis true of some units and false of others is not what *certify* asserts, but what *Sparse* exists to name. Aggregate the per-unit e-values by their mean rather than their product, which would require an independence they do not have:

$$E^{\mathrm{sig}} = \frac{1}{|U|}\sum_{j\in U} E_j^{\mathrm{sig}}, \qquad E^{\mathrm{uniq}} = \frac{1}{|U|}\sum_{j\in U} E_j^{\mathrm{uniq}}. \tag{5}$$

Proposition 3 (Dependence-robust aggregation)**.** *The mean of e-values is an e-value under arbitrary dependence; hence E^{sig} and E^{uniq} are valid despite inter-unit dependence from shared outcomes.*

Table 2 gives the decision rule in full.

Theorem 1 (Certification error)**.** *Every certificate requires $E^{\mathrm{sig}} \geq 1/\alpha$, so under H_0^{sig} the per-hypothesis probability of certifying is $\leq \alpha$. A* Matched *certificate also requires $E^{\mathrm{uniq}} \geq 1/\alpha$, and its probability under $H_0^{\mathrm{sig}} \cup H_0^{\mathrm{uniq}}$ is $\leq \alpha$. A* Shared *certificate also asserts that one restriction serves the scope, an attribution the invariance call bounds separately, at level α as $|U|$ grows (Section 4). Every certificate also asserts coverage, and under H_0^{cov}—at most a fraction τ_{cov} of in-scope units expected to outscore a random out-of-scope unit—its probability is $\leq \alpha$ given independent units (Section D.2); a minority-carried effect is withheld this way as* Sparse*.*

A hypothesis with $|U| < 1/\alpha$ is *undecided*: $E^{\mathrm{uniq}} \leq |U|$ puts *Matched* out of arithmetic reach and, at the default $|V| = |U|$, $E_j^{\mathrm{sig}} \leq |V| + 1$ leaves *Shared* at best on the bar, so we decline the space at that size. $|U| \geq 1/\alpha$ is necessary and not sufficient—the mean in Equation (5) reaches $1/\alpha$ only if nearly every unit sits near that cap. At $\alpha = 0.05$ we suggest $|U| = 400$, with 100 the minimum; a larger scope is subsampled uniformly, its reference still drawn from outside the whole scope. E^{sig} is asked of every certificate and carries its α bound; E^{uniq} decides which certificate a hypothesis can earn; cover gates both, the invariance call also gates *Shared*, and γ can only reject. Standing conditions and proofs are in Section C; Table 2 states what each reads under every class, and Section 4's variance model says what the invariance call reads under each.

Table 2: Every value combination of the four statistics with its verdict, for a scope covering at most a fraction γ of T; a wider one is *base rate* before any statistic is read. The invariance call is consulted only where $E^{\mathrm{uniq}} < 1/\alpha$, and separates *Shared* from *Unmatched*, and *base rate* from *Unsupported*: under the declared noise ceiling k it is *shared* when the restriction spread is shown to lie below the squared ceiling, and *not shared* otherwise, including when the data cannot place it. "—" marks a statistic the verdict does not consult, the columns to its left having settled it.

E^{sig}	E^{uniq}	call	p_{cov}	Type / verdict	In plain terms
$< 1/\alpha$	$< 1/\alpha$	shared	—	Base rate *reject*	Same claim for everyone, and it fits in scope no better than outside. Uninformative rather than false.
$< 1/\alpha$	$< 1/\alpha$	not shared	—	Unsupported *reject*	Not shown to be one claim, and it predicts nothing.
$< 1/\alpha$	$\geq 1/\alpha$	—	—	Subthreshold-Matched *reject*	Fits its own unit better than others, yet in scope no better than outside.
$\geq 1/\alpha$	$< 1/\alpha$	not shared	—	Unmatched *reject*	Not shown to be one claim; it does predict, but shuffling who gets which changes nothing.
$\geq 1/\alpha$	$< 1/\alpha$	shared	$\leq \alpha$	Shared *certify*	One claim covering the scope, and it holds there. Real, but about the scope rather than the unit.
$\geq 1/\alpha$	$< 1/\alpha$	shared	$> \alpha$	Sparse	One claim, but its lift is not shown to reach enough of the units.
$\geq 1/\alpha$	$\geq 1/\alpha$	—	$> \alpha$	*reject*	Per-unit prediction, not shown beyond a minority—the aggregate may say the rule holds where most units say it does not.
$\geq 1/\alpha$	$\geq 1/\alpha$	—	$\leq \alpha$	Matched *certify*	The restriction carries information the scope alone does not: shuffle the assignment and prediction breaks.

Shared and base rate. A *unit-invariant* hypothesis is one whose $\mathrm{apply}(x)$ is constant over its scope, where both apply and the restriction it generates may be non-deterministic. If its lift clears $1/\alpha$ it is *Shared*, or *Sparse* if $p_{\mathrm{cov}} > \alpha$; if not, or if its scope covers more than a fraction γ of T, it is *base rate*. γ is a user input, 0.9 by default, marking where a claim about a scope becomes one about the population (Section D.3). Such a hypothesis is not mechanism-free: its mechanism is scope membership itself. For a phone-buyer scope, an apply returning "buys a case" hands every in-scope unit the *same* restriction, so $E^{\mathrm{uniq}} = 1$ exactly and every column of M is constant: the spread vanishes, and the call is *shared* whatever the ceiling, provided $\bar{M}$ is clear of zero (Remark 1); returning "buys a case for this *specific* phone" makes it *Matched*.

3.3 Cost

$\mathrm{apply}(x)$—normally the expensive step—runs once per in-scope unit ($O(|U|)$ calls), not $O(|U|^2)$. The rest is cheap scoring: the matrix, which carries both the uniqueness e-value and the invariance call, costs $O(|U|^2)$, and the significance baseline, which scores each restriction on the $|V|$ reference outcomes, $O(|U||V|)$. The only pass over all of T is the scope test. The second cost is human, and it is the one that does not scale. A per-hypothesis resampling model must be *written*: someone decides, per hypothesis, which covariates define scope membership and must be held fixed and which carry the mechanism and must be redrawn—get that backwards and the test is invalid. Ours needs no such model, at any number of hypotheses; what a batch of them costs in error control is Section F.

4 Distinguishing unit-invariant hypotheses

Two pairs of verdicts turn on one question: over its scope, is the hypothesis **invariant**—all units under one restriction, within a noise tolerance—or **variant**? *Shared* against *Unmatched* asks it where the lift clears $1/\alpha$; *base rate* against *Unsupported*, where it does not. Under invariance every column of M is one restriction scored on one outcome, so every column is constant up to noise; a variant hypothesis's restrictions differ, sometimes only slightly, since restrictions can differ yet share one hidden factor (Figure 1).

In the datasets we target, units are **i.i.d.** (Section C)—the common case in ads and recommendation, where a live system serves independent requests—so for distinct units $i \neq j$ the restriction r_i and the outcome y_j are independent: the off-diagonal cells of M are draws from the product of the restriction and outcome laws, and the diagonal, where a unit meets its own outcome, carries the matching signal E^{uniq} reads. Under that product law each score splits as

$$M[i][j] = \mu + a_i + b_j + c_{ij}, \tag{6}$$

with μ the mean score, a_i how highly restriction i scores overall, b_j how easily outcome j is fitted, and c_{ij} which restriction fits which outcome beyond both; c averages to zero over either argument, so the three terms are uncorrelated, with variances σ_{r}^2, σ_{y}^2 and σ_{ry}^2. Any randomness in how apply produces a restriction belongs to that restriction, so no separate cell noise enters. What separates one shared restriction from several is how differently the restrictions score a common outcome,

$$S_{\mathrm{r}} = \tfrac{1}{2}\,\mathbb{E}\big[(M[i][j] - M[k][j])^2\big] = \sigma_{\mathrm{r}}^2 + \sigma_{\mathrm{ry}}^2, \qquad i, k, j \text{ distinct}, \tag{7}$$

which is 0 for one restriction and grows with every way restrictions differ, in level or in pattern (Section C.1).

The call. The declared ceiling puts sharedness at $S_{\mathrm{r}} < (k\mu)^2$. $\widehat{S}_{\mathrm{r}}$, the mean over columns of the variance of each column's off-diagonal cells, averages $\frac{1}{2}(M[i][j] - M[k][j])^2$ over every distinct triple and so estimates S_{r} without bias, and $\bar{M}$, the mean off-diagonal cell, estimates μ. The call tests $\widehat{D} = \widehat{S}_{\mathrm{r}} - (k\bar{M})^2$, with standard error $\widehat{\mathrm{se}}$ from the delete-one-unit jackknife of $\widehat{D}$ itself—a unit removing its row and its column from both terms, so the ceiling's estimate carries its own error. With $z = \Phi^{-1}(1-\alpha)$,

$$\text{call} = \begin{cases} \text{shared}, & \widehat{D} + z\,\widehat{\mathrm{se}} < 0, \\ \text{variant}, & \widehat{D} - z\,\widehat{\mathrm{se}} > 0, \\ \text{undecided}, & \text{otherwise}, \end{cases} \tag{8}$$

and only *shared* admits *Shared* or *base rate*. It is an equivalence test: sharedness is claimed only when the data place the spread below the ceiling, and a scope they cannot place is declined. Every quantity is a row or column sum, so the call costs no more than forming M.

What it bounds. If the restrictions truly spread by the ceiling or more, the call is *shared* with probability at most α; if they spread by less, it is *variant* with probability at most α (Section C.1). Both bounds are asymptotic in $|U|$, through the normal approximation to $\widehat{D}$'s error, and need units drawn independently and the regularity conditions of Section C.1; nothing else is assumed of the scores' law or of independence between cells. Only the first feeds a certificate. What k decides is where one restriction ends: set too high, restrictions that genuinely differ fall under it and an *Unmatched* scope certifies as *Shared*; set too low, a shared restriction whose per-unit variation exceeds it is declined.

5 Experiments

5.1 Setup

A verdict can only be scored against a known class, and no real corpus carries one. We therefore plant the classes, each defined by how its data are generated, not by our statistics (Table 8), and since planting only targets a class, experts audit every planted sample against that definition; the audited class is the label. We evaluate at three tiers. First, a generator we write: both the world and the mechanism are ours, and every verdict is scored. Second, an *intervention* on Amazon Video-Games—scope membership drawn at random and a mechanism planted, so the world is real and only the mechanism is ours—again scored. Third, 101 hypotheses written over that same corpus so neither the world nor the mechanism is ours; each carries an human-audited label, and every verdict is scored. The corpus also supplies what needs no label, null calibration at deployment scale (Section E).

Four of the six types plant the documented confusions of Section 1: *base rate* the restated base rate and popularity bias, *Shared* the scope-level effect, *Unmatched* the shared latent cause and *Sparse* the subgroup-carried lift. Exhausting the space is available to no dataset, so the reachable standard is coverage of the common cases. All six appear in every tier; on the generator, three also carry instances from generators written from scratch, sharing no popularity model with the main one: a latent cause (*Matched*), constant effects (*Shared*) and continuous windows (*Unmatched*). Within each tier, every method is scored on the same samples. All three tiers were built before the invariance call was derived, to reflect real situations rather than to satisfy any condition it needs.

The populations. Throughout, $\alpha = 0.05$, so an e-value must reach $1/\alpha = 20$ to certify, and the planted tiers read $|U| = 400$ units of each scope, drawn uniformly. The generator ($|T| = 16{,}000$, $|y| = 6$) gives 300 samples, 50 per type over seventeen rules (Table 3); the intervention plants the

six types into a scope of 610 Amazon users ($|y| = 3$), 18 samples per type, 108 in all (Table 4); Table 5 pools the two, 408. The real corpus adds 101 hypotheses (Table 6). Every count is over samples. Generator, certifier and hypotheses are in Section B, Table 8 and Table 9, and the real corpus's hypotheses in Section E.

5.2 RESULTS

The model-X CRT and the exact permutation test are the best existing procedures we found for certifying a hypothesis at the scope level, but both are designed for other nulls and return only certify or reject, so the certify decision, which our guarantees also cover, is what we compare. The comparison covers all three tiers, whose labels expert audited: per rule Table 3, per type Table 11, the intervention tier Table 4 and the real corpus Table 6. Each method's decision is compared with the human label implies—certify *Matched* or *Shared*, reject the rest—and the verdict names, which only ours returns, are reported as diagnostic categories (Section B.4). Only the model-X CRT needs a resampling model written by hand, one per rule family. Both remain valid for their intended nulls, which do not answer whether an association holds across its declared scope (Section B.3).

Table 3: Per-rule certification over the 300 generator samples. Every method certifies or rejects; ours also names a verdict (Table 2), and the last column counts the samples whose verdict matches the planted type, as a reference rather than a comparison. *call* is the invariance call at $k = 0.3$: *shared*, *variant* or *undecided*.

Rule	E^{sig}	E^{uniq}	call	cover	our verdict	matches target				verdict matches
						ours	CRT	perm	base	ours
brand	59.7	47.7	variant	0.91	*Matched*	10/10	10/10	10/10	10/10	10/10
category	41.2	32.8	variant	0.91	*Matched*	10/10	10/10	10/10	10/10	10/10
complement	189.5	142.5	variant	0.90	*Matched*	10/10	10/10	10/10	10/10	10/10
complement_cat	189.2	125.0	variant	0.91	*Matched*	10/10	10/10	10/10	10/10	10/10
latent cause	95.5	28.4	variant	0.91	*Matched*	10/10	10/10	10/10	10/10	10/10
constant, 3/50/80%	38.5	1.0	shared	0.91	*Shared*	37/37	0/37	1/37	37/37	37/37
seasonal	173.1	1.0	shared	0.90	*Shared*	13/13	0/13	0/13	13/13	13/13
heavy_pop	1.0	1.0	shared	0.41	*base rate*	25/25	25/25	25/25	0/25	25/25
popularity	1.0	1.0	shared	0.00	*base rate*	25/25	25/25	24/25	0/25	25/25
simpson_cat	57.2	42.0	variant	0.27	*Sparse*	50/50	0/50	0/50	0/50	50/50
random restriction	1.1	1.0	variant	0.32	*Unsupported*	50/50	50/50	48/50	0/50	50/50
confound_seg	95.7	1.0	variant	0.90	*Unmatched*	13/13	12/13	13/13	0/13	13/13
continuous, 3/3.5/4%	34.2	1.0	variant	0.81	*Unmatched*	37/37	34/37	33/37	0/37	37/37
					over all 300 *samples*	**300**	196	194	100	300
					resampling models written by hand	**0**	1/family	0	0	

Table 4: Certify decisions on the intervention tier, 108 samples, as target matches / of; the last column counts the samples whose verdict matches the planted type, ours only, as a reference rather than a comparison. Ours names seventeen of the base rates *Unsupported*, where the call is undecided.

type	matches target / of				verdict matches
	ours	CRT	perm	held-out	ours
Matched	18/18	18/18	18/18	18/18	18/18
Shared	18/18	2/18	0/18	18/18	18/18
base rate	18/18	18/18	17/18	0/18	1/18
Sparse	18/18	0/18	0/18	0/18	18/18
Unsupported	18/18	18/18	18/18	14/18	18/18
Unmatched	18/18	18/18	18/18	0/18	18/18
all	**108/108**	74/108	71/108	50/108	91/108
types matched	6/6	4/6	3/6	2/6	

Table 5: Certify decisions over the 408 generator and intervention samples, compared between our algorithm, our metrics with and without our scope on perm, and the baselines.

	match	f. decline	f. cert.
$N \times N$ stats, $N \times N$ M	**408**	0	**0**
$N \times N$ stats, perm M	404	0	4
perm with $N \times N$ stats over T	340	0	68
perm baseline	265	67	76
CRT baseline	270	66	72
held-out baseline	150	0	258

Needing no per-hypothesis resampling model comes from the mutual-null construction, in which the units under one hypothesis are used as nulls for each other. That construction is also what makes E^{uniq} and the invariance call expressible, and together with E^{sig} and cover these four statistics separate cases the CRT and the permutation test are not designed to separate. An unmodified CRT

detects conditional dependence and a pairing permutation test detects unit-specific matching, not scope coverage or a shared restriction: both certify every *Sparse* hypothesis and reject nearly every *Shared* one (Table 4). Held-out validation, which asks only whether a hypothesis predicts, certifies every one of the harder cases. Beyond the decisions, the seven verdicts (Table 2) are diagnostic categories intended to support hypothesis refinement. The permutation test shares the mutual-null construction, so it covers almost all cases when patched with our three statistics, but it still differs from $N \times N$ (Table 5): it certifies four confounds, because shuffling the pairing across the matrix destroys the within-group structure E^{uniq} measures. Removing the hypothesis scope degrades the patched test substantially as well (Table 5), which supports the hypothesis-scope contribution.

On the real corpus, Table 6 scores 101 hypotheses over the Amazon Video-Games data, but LLM proposed less *Shared* ones. The baselines' minority failure recurs: both certify nearly every *Sparse* hypothesis, all of which ours rejects. The hypothesis was built by an LLM with no context of our algorithm, but label by human, and Table 6 scores the audited type. Our 4 misses are all declines: three with on-boarder coverage, and one *Shared* hypotheses, whose restrictions have genuine on-boarder differences from user to user. It should be noted, though, that human decisions may still miss the real underlying mechanism so that our 4 misses were actually the right choice.

Table 6: Hypotheses on Amazon Video-Games ($|U| \leq 800$, $|V| = 2000$), scored as in Table 3; the three statistics are medians over each type.

					decision matches target			verdict matches
type	E^{sig}	E^{uniq}	cover	n	ours	CRT	perm	ours
Matched	143.75	25.82	0.576	18	15	**18**	**18**	15
Shared	339.48	1.00	0.550	3	2	0	**3**	2
base rate	1.08	1.00	0.247	20	**20**	**20**	16	20
Sparse	83.30	31.76	0.247	22	**22**	2	2	21
Unsupported	0.94	0.89	0.011	20	**20**	19	19	20
Unmatched	80.46	1.11	0.317	18	**18**	15	10	18
total				101	**97**	74	68	96
false certificates					**0**	24	33	

The formal guarantee is a separate matter, and it holds where it is claimed: Theorem 1 is checked directly on permutation nulls of the real corpus, which make each null true by construction: over 8,325 draws neither e-value reaches $1/\alpha$ (Section E). The invariance call's level is checked by simulation at the ceiling (Section C.1), and on the planted cases it calls shared 119 of the 136 scored samples whose units share a restriction, the other seventeen undecided, and none of the 272 whose units do not. That bound is per-hypothesis, and cover only narrows it (Section 3.1). What a *batch* of hypotheses admits is a separate question, and Section F answers it: at the same level e-BH keeps the lift of all 17 certificates.

6 Conclusion

$N \times N$ e-valuation extends a core idea it shares with the permutation test (Fisher, 1935; Pitman, 1937) and Xu et al. (2026)—using the units under one hypothesis as nulls for each other—with four statistics: E^{sig}, E^{uniq}, the invariance call and cover. Each is read by a test that certifies, and together they route every hypothesis to a diagnostic verdict. Its decisions match the scope-certification target on all six planted types, including a minority-carried and a scope-level effect. The CRT's answers on those two are insufficient for scope-level certification: its null asks whether unit-level dependence exists, so it rejects the scope-level effect and certifies the minority-carried one. The exact permutation test shares the same cross-prediction matrix, so the four statistics are our only addition to it; patched with them it still issues four false certificates (Table 5). No resampling model is written per hypothesis at any point, by construction: Algorithm 1 takes the hypothesis, the data and four constants—α, two thresholds and a noise ceiling—and has nowhere to put one. One thing bounds what that buys: restrictions that nearly coincide read as shared, because on the observed outcomes they nearly are. The demonstration is a single corpus, so what the restriction spread reads for a real-valued restriction, and whether the split survives outside set-valued outcomes, we have not shown.

REPRODUCIBILITY STATEMENT

Every number here is produced by a script over a synthetic generator or over the public Amazon Reviews 2023 corpus (Hou et al., 2024), so reproduction needs the generator, the corpus split, the certifier and the baselines. The world model and all of its parameters—catalogue size, population, categories, brands, popularity exponent, outcome length, follow rate and scope weights—are stated in Section B, the construction sequence in the order the generator executes it is Section B.1, and each planted rule is given in Tables 8 and 9; the corpus split and the intervention are in Section B.1, and the real-corpus hypotheses in Section E. The certifier is Algorithm 1 together with the four statistics of Section 3.1, and the tolerances are in Section D. Every table states its own denominator. The model-X CRT is handed an oracle sampling law; Section B.2 describes what it can and cannot express.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods (the certifier, the synthetic generator, and the model-X CRT baseline of Section 5); to generate the synthetic data sets those scripts produce; to help *refine* the conceptual framework—surfacing the Simpson's-paradox failure mode that E^{uniq} alone cannot reject, and helping refine the coverage signal (Equation (4)) that resolves it, which the authors designed—and to refine hypotheses; to supply ingredients for the proofs of our mathematical claims and assist in writing those proofs (Section C and the appendix bounds); to give feedback on the experiments; and to interpret results. We have not used generative AI tools to design or propose the algorithm, to formulate the mathematical claims, to design the experiments, to assist with translation, to clean or reformat a dataset, or to support qualitative or thematic data analysis. Additionally, we used generative AI tools to write and edit software code, to draft parts of this paper, to edit it for readability and suggest its structure, to identify and summarise related literature, and to suggest experimental parameters.

We have reviewed all AI-assisted work. Every theorem statement and proof was re-derived and checked by the authors. Every number reported in this paper is produced by a script in our experiment artifact rather than by a model, and each was checked against that script's raw output; arithmetic stated in prose—ratios, break-even points—was recomputed from those outputs rather than carried over. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

AUTHOR CONTRIBUTIONS

Except all the authors in the author list, ********** also contributed to this work as the project intern. She implemented the initial version of the system and should be considered as an author too. Unfortunately, we couldn't find her open review profile thus wasn't able to add it to the author list before deadline. Thus acknoledge her contribution here.

REFERENCES

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The unfairness of popularity bias in recommendation. In *RecSys Workshop on Recommendation in Multistakeholder Environments*, 2019.

Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, et al. A survey on hypothesis generation for scientific discovery in the era of large language models. *arXiv preprint arXiv:2504.05496*, 2025.

James N. Arvesen. Jackknifing U-statistics. *The Annals of Mathematical Statistics*, 40(6):2076–2100, 1969.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24 (4):295–300, 1982.

Thomas B. Berrett, Yi Wang, Rina Foygel Barber, and Richard J. Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B*, 82(1):175–197, 2020.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577, 2018.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *ACM Conference on Recommender Systems (RecSys)*, 2019.

Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

Sander Greenland, James M. Robins, and Judea Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46, 1999.

Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(5):1091–1128, 2024.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024.

Amir H. Jadidinejad, Craig Macdonald, and Iadh Ounis. The simpson's paradox in the offline evaluation of recommendation systems. In *ACM Transactions on Information Systems*, 2021. arXiv:2104.08912.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics (TACL)*, 12:1417–1440, 2024.

Nick W. Koning. Measuring evidence against exchangeability and group invariance with e-values. *arXiv preprint arXiv:2310.01153*, 2023.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

Felipe Llinares-López, Mahito Sugiyama, Laetitia Papaxanthos, and Karsten Borgwardt. Fast and memory-efficient significant pattern mining via permutation testing. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 725–734, 2015. doi: 10.1145/2783258.2783363.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220, 1967.

Muriel Felipe Pérez-Ortiz, Tyron Lardy, Rianne de Heide, and Peter Grünwald. E-statistics, group invariance and anytime valid testing. *arXiv preprint arXiv:2208.07610*, 2022.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357, 1950.

Shalev Shaer, Gal Maman, and Yaniv Romano. Model-X sequential testing for conditional independence via testing by betting. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2054–2086, 2023.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B*, 13(2):238–241, 1951.

Robert E. Tarone. A modified bonferroni method for discrete data. *Biometrics*, 46(2):515–522, 1990. doi: 10.2307/2531456.

Aika Terada, Mariko Okada-Hatakeyama, Koji Tsuda, and Jun Sese. Statistical significance of combinatorial regulations. *Proceedings of the National Academy of Sciences*, 110(32):12996–13001, 2013. doi: 10.1073/pnas.1302233110.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74 (1–2):9–28, 2015.

Vladimir Vovk. Testing randomness online. *Statistical Science*, 36(4):595–611, 2021.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 2022.

Nathaniel Xu, Feng Liu, and Danica J. Sutherland. Learning representations for independence testing. *Transactions on Machine Learning Research*, 2026. arXiv:2409.06890.

A NOTATION

Symbols appear in the order the body introduces them, each listed once at the definition that governs its later uses; Table 1 holds the objects they are derived from.

Table 7: Every derived symbol, with where it is defined.

Symbol	Defined	Meaning
Sets and sizes		
A_B	Table 1	collection convention: $A_B = \{a_i\}_{i \in B}$ for any symbol a and index set B; hence Y_U, Y_V (outcomes) and R_U (restrictions)
U, T	Table 1	in-scope set and test set, $U \subseteq T$
V	Section 3.1	reference: units drawn uniformly from $T \setminus U$, $\lvert U \rvert$ of them by default, all of them if fewer
$M[i][j]$	Equation (2)	cross-prediction matrix, $r_i(y_j)$; the diagonal is each unit's own restriction
The variance model		
σ^2_{r}	Equation (6)	variance of the restrictions' overall levels; zero when restrictions are shared
σ^2_{y}	Equation (6)	variance of the outcomes' overall ease; the signal a shared restriction produces
σ^2_{ry}	Equation (6)	restriction × outcome interaction: which restriction fits which outcome; zero when restrictions are shared
S_{r}	Equation (7)	restriction spread on a common outcome, $\sigma^2_{\mathrm{r}} + \sigma^2_{\mathrm{ry}}$; 0 for one restriction
μ	Equation (6)	mean score off the diagonal; the noise ceiling is $k\mu$
$\bar{M}$	Section 3.1	the mean off-diagonal cell of M, which estimates μ
Baselines the e-values divide by		
$g_0^{(r_j)}$	Equation (1)	restriction r_j's mean score over V and unit j—the significance null
The four statistics		
E^{sig}	Equation (1)	diagonal against the base rate outside the scope; an exact e-value
E^{uniq}	Equation (3)	diagonal against the in-scope row mean; an exact e-value, no covariate model
call	Equation (8)	the invariance call: *shared* when $\widehat{D} + z\,\widehat{\mathrm{se}} < 0$, with $\widehat{D} = \widehat{S}_{\mathrm{r}} - (k\bar{M})^2$, read at α
cover	Equation (4)	fraction of units whose outcome outscores their comparator's, v_j being a distinct random member of V
p_{cov}	Equation (4)	the coverage test: exact binomial p-value that more than a fraction τ_{cov} of units outscore their comparators; $p_{\mathrm{cov}} \leq \alpha$ is required to certify, never sufficient
User inputs: the level, two thresholds and a noise ceiling		
α, $1/\alpha$	Section 3	level, and the threshold every e-value must clear to certify
k	Section 3.1	noise ceiling: restrictions whose scores on a common outcome spread by less than $k\mu$ count as one. Declared; it fixes what counts as one restriction, so report it with any *Shared* certificate
γ	Table 2	share of T above which a scope is the population and reads as *base rate*; 0.9 by default
τ_{cov}	Table 2	share of units the coverage test must show outscoring their comparators, set by the user; if it is not shown the certificate is withheld (*Sparse*)

B EXPERIMENT DETAILS

Generator. A catalog of 1500 items, each with a category (of 40), a brand (of 60), and a Zipf popularity (s=1.05). Each of $|T|$=16,000 users draws a scope from {complement, brand, category, holiday, exposed, base} with weights $(0.12, 0.03, 0.03, 0.03, 0.03, 0.76)$, so every rule scope holds at least 3% of T and the certifier draws $|U| = 400$ of its units; a history x_i unioning the user's recent item, favourite-brand and recent-category picks with popularity draws, so $|x_i|$ varies (Section B.1); and a held-out outcome set y_i (6 items) whose mechanism is the scope's own with probability p_{follow}=0.9 and popularity noise otherwise; Section B.1 gives the per-item draw within a mechanism. The structural confound is planted separately: each "segment" covers a different ∼60% subset of a *low-popularity* promoted set P—a structural confound must carry a scope-level lift, which requires its target set to be rare outside the scope—exposed users' $y_i \subset P$, and the segment→outcome score factors as (segment overlap with P)×(item appeal) with no pairing term, so no unit's restriction is matched to its own outcome, as the definition requires. Each shared re-

striction is returned with per-unit noise, every item replaced with a probability drawn per unit from $[0, 0.02]$, and a further rule hands a 3.9% slice of base users each its own popularity-weighted random restriction (*Unsupported*).

Certifier. $r_i(y)$ = recall of restriction i on outcome set y ($|\text{retained}_i \cap y|/|y|$); E^{sig}, E^{uniq}, the invariance call and cover exactly as in Section 3, with $g_0^{(r_i)}$ taken over $V \cup \{i\}$ and the matrix over all of U, so both self-inclusions hold, as Propositions 1 and 2 need. The generator predates the certifier and was written to mimic a real corpus and nothing else: none of Section C.1's conditions was considered in building it—with $|y| = 6$ each cell takes one of seven values, and restrictions of different breadth carry different cell variance.

Tolerances. α=0.05, $\tau_{\text{cov}} = 0.5$, $\gamma = 0.9$ and $k = 0.3$ are held fixed across every world and every rule. τ_{cov} is a declaration rather than a measurement: it is the share of in-scope units a user requires to outscore a random out-of-scope unit, and a certificate needs the coverage test to show more than that share. At 0.5 we ask for a majority, which is stricter than the usual reading of *minority-carried*; raising it rejects real rules and never miscertifies (Table 13).

Baselines. Three comparators, each implemented from its source paper. *Naive held-out validation* is a one-sided z-test that the rule's mean held-out recall exceeds that of a random same-size set; it certifies anything predictive, all four fakes included. The *model-X CRT* (Candès et al., 2018) tests the diagonal mean, given the generator's sampling law as an oracle. The *exact within-scope permutation test* is the sharpest of the three: it reads the same M we do, under the permutation of the pairing that Mantel (1967) tests two matrices with, so it differs from our certifier in the statistics taken off the matrix and not in the null, and isolates what those statistics contribute.

Scoring. On the two planted tiers the label is the planted type as audited by experts, and every method's certify decision is compared with the target: it matches when it certifies a *Matched* or *Shared* sample or rejects any other. The decisions are the CRT's mechanism p-value at α, with per-family resamplers on the generator and a corpus-wide redraw on the intervention tier; the permutation test's $p \leq \alpha$; held-out validation's one-sided z; and whether our verdict certifies. Our verdict names are counted separately, as a reference (Section B.4).

Seeds. The planting routine draws while iterating a set of string identifiers, so the world it builds depends on the interpreter's string-hash seed; we pin that seed for every run.

B.1 World construction

With 1500 items, $16{,}000$ users, 40 categories, 60 brands, in this order:

1. Category and brand for each item, uniform and independent.
2. Popularity: draw a uniform permutation of the items and give the item at rank ρ weight $\rho^{-1.05}$, normalised. **Rank is assigned to a shuffled order**, so popularity is independent of item index; an implementation that ranks by index correlates popularity with every other index-derived quantity.
3. *gift_items*: 6 items sampled from category 0.
4. *promo*: 40 items sampled from the *lower-popularity half* of the non-gift items ordered ascending by popularity. A structural confound must carry a strong scope-level lift, which requires its target set to be rare outside the scope.
5. Per-item complements: 4 items sampled uniformly from the other 1499, per item.
6. Segments: for each of 10 segments, a different 24-item subset of *promo* (60%). All segments predict *promo* about equally well but disagree on which of its items: restrictions that differ across units with none matched to its own outcome, as the definition of a structural confound requires.
7. *top_pop*: the 40 highest-popularity items.
8. Per user, in order: scope label from {complement, brand, category, holiday, exposed, base} with weights $(0.12, 0.03, 0.03, 0.03, 0.03, 0.76)$; a recent item; a favourite brand; a recent category; a segment; a heavy-user flag at rate 0.25. Then the mechanism label: the scope's own rule with probability 0.9 if the scope is not *base*, otherwise *popularity*.

9. Input x_i: the recent item, 3 items of the favourite brand, 3 of the recent category, and 15 popularity-weighted draws, unioned as a *set*—so $|x_i|$ varies (19–22 at seed 0) rather than being fixed.

Outcomes are drawn last, after every user record exists, because the mechanism pools are read from the completed world. For a unit with a mechanism, draw 6 items, each from the mechanism's pool with probability 0.95 and popularity-weighted otherwise; for a *popularity* unit, draw 6 popularity-weighted. Accumulate into a *set* and top up with popularity draws until it holds exactly 6, so $|y_i| = 6$ for every unit.

Weighted sampling without replacement is by rejection—draw one at a time, keep if new, give up after $60k$ attempts for a target of k. With a Zipf tail this matters: sampling by an exact without-replacement algorithm gives a measurably different popularity mix.

The Amazon corpus. The intervention tier and Section E run on the same corpus: the Video-Games category of Amazon Reviews 2023 (Hou et al., 2024), taken from the user-review and item-metadata dumps at parent-ASIN granularity. We keep users with at least eight interactions and at least eight distinct parent-ASINs after de-duplication, retaining one occurrence per ASIN at its latest timestamp, and split them into a proposal half of $20{,}462$ users and a disjoint certification half of $|T| = 20{,}323$. Outcomes are the last three purchases by timestamp, so $|y| = 3$; the de-duplication is what makes that exactly three, since a user may touch one ASIN repeatedly. A unit's history and its outcome are disjoint in both item and time—the outcome is the final three distinct ASINs, the history everything strictly before them—so a restriction compiled from the history is never scored on an item it was compiled from. Popularity, per-period popularity and co-occurrence are computed on the proposal half alone, and every claim is relative to that subpopulation rather than to Amazon as a whole. The intervention on it plants the six types of Table 8.

Table 8: The planted hypotheses.

Type	Rule	Recommendation rule (and, for fakes, why it is fake)
Matched *real*	complement	"a user who just bought X buys X's complement next": `apply` returns the per-item complement of the user's *own* recent purchase, so r_i fits unit i's outcome and few others'
Matched *real*	brand	"a user keeps buying from their favourite brand": `apply` returns that user's favourite-brand items
Matched *real*	category	"a user re-engages a recently browsed category": `apply` returns that user's recent-category items
Shared *real*	seasonal	"holiday shoppers buy the gift-wrap line": `apply` returns *one* gift set, near-identical across in-scope users—real but constant
base rate *fake*	popularity	"everyone buys the best-sellers": `apply` returns the globally most-popular items (same for all). *Why fake:* it merely restates the base rate—in-scope users buy them no more than the population does
base rate *fake*	heavy_pop	"the best-sellers, claimed of a sub-population": the same globally popular items, claimed only of a random quarter of the users. *Why fake:* popular items carry no lift over the base rate even there—a base rate dressed up as a targeted rule
Unmatched *fake*	confound_segment	"each user buys items matching their market segment": `apply` returns a per-user segment set and *looks* personalized. *Why fake:* purchases are driven by a shared, exogenous promoted campaign, not the segment; every segment overlaps it, so r_i predicts others as well as its own
Sparse *fake*	simpson_cat	"users buy the complement of their recent item," but the mechanism holds only when the recent item's category lies in a fixed 30% of categories. *Why fake:* as stated it claims the whole scope, and it is predictive in aggregate only because that subgroup carries it. Because the subgroup is a *covariate* the hypothesis is *refinable*
Matched *real*	complement_cat	*simpson_cat* after adding the condition the planted subgroup supplies directly. A genuine per-unit rule on the narrowed scope
Unsupported *fake*	random restriction	each user in a 3.9% slice of the base population is handed its own popularity-weighted random restriction. *Why fake:* the restrictions are unrelated to the outcomes, so nothing in them predicts
Planted by intervention on the Amazon corpus		
Matched	co-purchase kNN	each unit's restriction is drawn from its own item-KNN neighbourhood, so every restriction is matched to the unit holding it
Matched	latent set	a latent 10-item set per unit, of which the restriction exposes 4 while the outcome is drawn from all 10—the diagonal is diluted before p_{follow} reaches it
Shared	one shared set	one restriction, near-identical across the whole scope
base rate	popularity head	the popularity head, with nothing planted, so in- and out-of-scope users hit it at the same rate
Unmatched	group subset	units split into eight latent groups; `apply` returns the unit's group subset of a promoted set P while outcomes are drawn from P whole, so every group predicts about equally and none is matched to its own unit
Sparse	30% subgroup	the mechanism reaches only 30% of the scope; the rest keep their real outcomes
Unsupported	random per unit	outcomes left real; each in-scope unit is handed its own popularity-weighted random restriction

Table 9: The added-threat instances Table 3 scores, in the register of Table 8. Three generators written from scratch, sharing no popularity model with the main one: out-of-scope outcomes are uniform, with no categories, brands, histories or mechanism pool. Each scope holds at least 3% of T and the certifier reads 400 of its units; together they give 84 of the 300 generator samples.

Type	Rule	Recommendation rule (and, for fakes, why it is fake)
Matched *real*	latent readout	"a user buys within the interest group they belong to": each unit joins one of 40 latent groups, each owning a 10-item set from which that unit's outcome is drawn. apply returns the 10-item set of the group a *noisy readout* names, correct with probability 0.9. The group drives both the observed covariate and the outcome, and nothing else does. Not fake: the association is real and matched to its unit, and the readout noise is what keeps it from collapsing into *complement*. Scope 3%
Unmatched *fake*	continuous window	"each user buys the items in their own slice of a quality ranking": one exogenous score ranks the catalogue, in-scope outcomes are drawn from its top 40, and apply returns a per-unit 10-item window of that ranking at a uniformly drawn offset. *Why fake:* the shared ranking drives the purchases, not the unit's window; every window overlaps the same head, so r_i predicts other units as well as its own. The shared factor is continuous where *confound_segment*'s is a discrete segment. Scopes 3%, 3.5% and 4%
Shared *real*	constant effect	"a promotion lands on an audience": one 30-item set, near-identical for every in-scope unit and followed with probability 0.9. Real and constant: its lift is measured against units outside the scope, who do not follow it. Scopes 3%, 50% and 80% of T, all under the $\gamma = 0.9$ line where a scope becomes the population

B.2 What each signal buys

Table 10: Verdicts of Table 2 that become indistinguishable when one statistic is removed. All four are load-bearing: † marks a merge that joins a *certify* verdict with a *reject* one, and every removal makes one.

removed	verdicts it merges	lost
E^{uniq}	Base rate = Unsupported = Subthreshold-Matched = Shared = Unmatched = Matched = Sparse†	**5**
E^{sig}	Base rate = Shared = Sparse† = Subthreshold-Matched = Matched = Sparse† = Unsupported = Unmatched	**5**
call	Base rate = Unsupported = Shared = Unmatched = Sparse†	3
cover	Matched = Sparse† = Shared = Sparse†	2

E^{uniq} separates a per-unit mechanism from a unit-invariant one but is blind inside the latter: a shared restriction makes every row of M equal, forcing $E^{\mathrm{uniq}} = 1$ by Proposition 2. The invariance call reads that degeneracy as the signature of sharing rather than as a failure, which is what admits *Shared* and *base rate* at all—and why dropping it merges *Shared* into *Unmatched*. cover is what separates *Sparse*: *simpson_cat*'s carrying minority pushes E^{uniq} past $1/\alpha$, so the case never reaches the branch the call gates.

On the planted tiers the call says shared for every *Shared* sample and every generator *base rate*, but for only one of the eighteen intervention base rates, whose restrictions spread by 0.24–$0.31\bar{M}$ against the ceiling of $0.3\bar{M}$; the other seventeen are undecided and named *Unsupported*, a right decision under another name. On the real corpus the call is shared for 2 of the 3 *Shared* hypotheses.

This is what the CRT misses. Such a test carries information only through apply's dependence on the covariates, so on a unit-invariant hypothesis—every *Shared* and every *base rate* rule—resampling changes nothing and the test is not weak but absent. Its resampling model also has to be written by hand for each hypothesis: our own CRT baseline needed one for a single rule in nine, and a stream of machine-proposed hypotheses would demand the same judgement for each.

Table 11: The certify decision by planted type—the per-type split of Table 3's certify column. The permutation test given our statistics certifies four of the fifty *Unmatched* samples.

planted type	ours	CRT	perm	perm + cover + call	held-out
Matched	50/50	50/50	50/50	50/50	50/50
Shared	50/50	**0/50**	**1/50**	50/50	50/50
base rate	50/50	50/50	49/50	50/50	**0/50**
Sparse	50/50	**0/50**	**0/50**	50/50	**0/50**
Unsupported	50/50	50/50	48/50	50/50	**0/50**
Unmatched	50/50	46/50	46/50	46/50	**0/50**
types matched	**6/6**	3/6	1/6	5/6	2/6

B.3 Limitations of the CRT and the exact permutation test

Both baselines ask whether unit-level dependence exists: the model-X CRT whether the outcome depends on the covariate that carries the mechanism, given the scope's covariates, and the exact permutation test whether each unit's own restriction scores its outcome better than a randomly paired one. Both answer that question validly. Neither answer is a scope-level certificate, and read as one it misleads in two opposite ways.

A minority-carried association. It has unit-level dependence, so both tests reject their nulls: each certifies all fifty *Sparse* samples of Table 11. The rejection is right for their nulls and silent on whether the association holds across the claimed scope. $N \times N$ withholds the certificate through the coverage test, exact at level α given independent units, and it withheld every planted *Sparse* sample here (Section D.2).

A restriction shared by every unit. It gives both tests nothing to move: the CRT's resampling cannot change a restriction that is the same for everyone (Section B.2), and permuting the pairing changes little when every unit holds a score-equivalent restriction. The CRT certifies none of the fifty *Shared* samples and the permutation test one. $N \times N$ tests the scope-level lift against out-of-scope references (E^{sig}) and the sharedness with the invariance call.

The burden. Specialised tests could recover both distinctions, a subgroup test for the minority and a scope-level comparison for the shared restriction, but each needs a conditional or subgroup structure specified per hypothesis, as the CRT's resampling model already is (Section B.2), and that does not scale to machine-proposed hypotheses. $N \times N$ draws both distinctions from one matrix, with no per-hypothesis design.

B.4 What each verdict tells the user

A certificate says that a hypothesis's lift is real; the verdict adds where that lift appears to come from, which is what a proposer needs to decide what to try next. Table 12 gives one reading of each verdict and a step it suggests. The readings are examples rather than the only way to interpret a verdict, and none is covered by Theorem 1: the guarantee concerns certification, and the name is feedback.

C Guarantees: standing conditions and proofs

Propositions 1 to 3 and Theorem 1 are stated in Section 3.2; this section gives the conditions they are proved under and the proofs. The three propositions and the e-value bounds of Theorem 1 are exact, finite-sample, and valid under arbitrary dependence between units. The coverage claim of Theorem 1 is exact and finite-sample but needs the i.i.d. assumption below, and so does the invariance call, whose bounds are asymptotic and which also needs a declared noise ceiling; it is derived in Section C.1.

The i.i.d. assumption. Conditional on the proposal data, the hypothesis and its scope, the in-scope pairs $Z_i = (r_i, y_i)$ are independent and identically distributed, and independent of the reference units V, which are themselves i.i.d. from the out-of-scope population. Only the invariance call

Table 12: One way to read each verdict, and a next step it suggests. Examples, not definitions.

Verdict	One reading	A step it suggests
Matched	The predictive factor comes from each unit: the hypothesis extracts a property of the unit that predicts that unit's own outcome.	Keep it; the per-unit restriction does the work.
Shared	The predictive factor lies in the hypothesis itself, or in a property the units in its scope have in common, rather than in anything specific to one unit.	Keep it as a claim about the scope, not about each unit.
base rate	The prediction comes from the population the data are drawn from, not from anything the hypothesis says.	Drop it; it restates what holds everywhere.
Sparse	The hypothesis may hold for only part of its scope, or the scope is too small to show that most units follow it.	Narrow the scope to where it holds, or re-test on more units.
Unmatched	The scope carries a real gain, but the factor behind it is not yet revealed: the restrictions differ and none is matched to its own unit.	Look for the common factor behind the gain.
Unsupported	Neither one shared claim nor any gain: nothing the hypothesis names predicts.	Drop it or rethink the restriction.
Subthreshold-Matched	Each unit's restriction fits its own outcome best, yet in scope no better than outside: a per-unit signal the scope does not sharpen, or one too weak to show at this size.	Widen or drop the scope, or test on more units.

and the coverage test use it: the call's two U-statistics need independent units (Section C.1), and the coverage test needs each unit's pairing with its comparator to be independent of the others' (Section D.2). The e-values rest on their exchangeability nulls instead, and Proposition 3 allows any dependence among the per-unit e-values. Clustered or time-dependent units violate the assumption; datasets where a live system serves independent requests satisfy it. Throughout, each restriction $r(\cdot)$ is a nonnegative score returned by apply from the input and, when apply is stochastic, from proposal randomness independent of the outcomes. Two distinct ingredients are used in those proofs. *Measurability*: the proofs condition on r_j itself, which *fixes* it and makes $g_0^{(r_j)}$ and the row mean constants. They also condition on the outcomes as a *multiset*: duplicates are kept, so a pool of n units has n members even where two units share an outcome, and ordering is dropped, so conditioning does not reveal which member is y_j. The first keeps those averages well defined; the second is what leaves y_j random. *Exchangeability*: what drives each e-value to 1 is the null, not the form of r. The null concerns the *pairing*. When the hypothesis is real r_j and y_j are dependent—both are downstream of x_j—and what E^{uniq} tests is whether that dependence is matched to the unit.

Exchangeability is also what makes Proposition 1 *exact*. A weaker null still gives $\mathbb{E}[E_j^{\mathrm{sig}}] \le 1$, a valid but conservative test, if it is stated in distribution rather than in mean: an in-scope score stochastically no larger than an out-of-scope one, units independent. E_j^{sig} increases in its unit's own score, so replacing that score by one drawn like the references, which is no smaller, can only raise the mean, and that mean is exactly 1. Equal means alone do not suffice, because the denominator is random and includes the unit's own score. Exchangeability, or that dominance, is therefore the condition to check in a new domain.

Remark 1 (Identities under a unit-invariant hypothesis)**.** For a unit-invariant hypothesis, $r_i \equiv r$, so $M[i][j] = r(y_j)$ does not depend on i: every row of M equals $r(y_\cdot)$, every row-mean equals the same $\overline{r} = \frac{1}{|U|}\sum_{k\in U} r(y_k)$, and every column is constant. Hence

$$E_j^{\mathrm{uniq}} = \frac{r(y_j)}{\overline{r}}, \qquad E^{\mathrm{uniq}} = \tfrac{1}{|U|}\sum_{j\in U} E_j^{\mathrm{uniq}} = 1, \\ \widehat{S}_{\mathrm{r}} = 0, \qquad \widehat{D} = -(k\bar{M})^2. \tag{9}$$

The last two follow with no approximation: every column is constant, so each column's variance vanishes, in $\widehat{S}_{\mathrm{r}}$ and in every jackknife replicate, leaving $\widehat{D}$ and its replicates at $-k^2$ times $\bar{M}^2$ and its replicates. The call is then *shared* exactly when $\bar{M}^2$ exceeds z times its own jackknife error; k cancels, so on a unit-invariant hypothesis the call never turns on its value. The premise $r_a = r_b$ holds exactly for a deterministic apply and only up to sampling agreement otherwise.

Proof of Proposition 3. E^{sig} and E^{uniq} are *means* of the per-unit e-values. By linearity of expectation,

$$\mathbb{E}\left[\frac{1}{|U|}\sum_{j\in U} E_j\right] = \frac{1}{|U|}\sum_{j\in U}\mathbb{E}[E_j] \le \frac{1}{|U|}\sum_{j\in U} 1 = 1,$$

using $\mathbb{E}[E_j] \leq 1$, which Propositions 1 and 2 give with equality. The per-unit e-values are dependent—E^{uniq}'s share the multiset Y_U, E^{sig}'s the reference Y_V—but linearity never invokes independence, so the mean is a valid e-value regardless. □

Proof of Theorem 1. Both certificates require $E^{\text{sig}} \geq 1/\alpha$. Under H_0^{sig}, E^{sig} is a valid e-value (Propositions 1 and 3), so Markov's inequality gives $\mathbb{P}(E^{\text{sig}} \geq 1/\alpha) \leq \alpha$, and since every certificate lies in that event, $\mathbb{P}(\text{certify}) \leq \alpha$. A *Matched* certificate also requires $E^{\text{uniq}} \geq 1/\alpha$, and under H_0^{uniq} the same argument via E^{uniq} bounds it by α, so it has probability $\leq \alpha$ under either component of $H_0^{\text{sig}} \cup H_0^{\text{uniq}}$: the *intersection–union* principle (Berger, 1982). The remaining conditions—$p_{\text{cov}} \leq \alpha$, a *shared* call and a scope within $\gamma|T|$—only shrink the sets these bounds are taken over: a one-sided gate that can only remove inherits the guarantee whatever value it is set to, and needs no calibration of its own (Ramdas et al., 2023, Fact 1.3(vi)). The coverage claim is bounded by its own test: under H_0^{cov} every certificate lies in $\{p_{\text{cov}} \leq \alpha\}$, of probability at most α (Section D.2). The attribution a *Shared* certificate adds is bounded by the call, asymptotically (Section C.1). □

Proof of Proposition 1. Fix a unit $j \in U$, write $V_j = V \cup \{j\}$ for its pool, and condition on V, on Y_{V_j} and on r_j. V is drawn without reading an outcome, so conditioning on it leaves H_0^{sig} intact. The denominator $g_0^{(r_j)} = \frac{1}{|V_j|}\sum_{k\in V_j} r_j(y_k)$ is a symmetric function of Y_{V_j}, hence a constant given the conditioning. *Exchangeability* of y_j with Y_{V_j} under H_0^{sig} is stated conditionally on r_j, and that is how it is used: conditionally on (V, Y_{V_j}, r_j), y_j is equally likely to be any of the $|V_j|$ members of Y_{V_j}—a uniform draw. Conditioning on r_j is what makes $g_0^{(r_j)}$ constant, and conditioning on the outcomes as an unordered multiset is what leaves y_j random given it. So the conditional mean equals the average over Y_{V_j}:

$$\mathbb{E}\big[r_j(y_j) \mid V, Y_{V_j}, r_j\big] = \frac{1}{|V_j|}\sum_{k\in V_j} r_j(y_k) = g_0^{(r_j)}.$$

Dividing by the constant $g_0^{(r_j)}$ gives $\mathbb{E}[E_j^{\text{sig}} \mid V, Y_{V_j}, r_j] = 1$, and by the tower property $\mathbb{E}[E_j^{\text{sig}}] = 1$. It is *exact* rather than ≤ 1 because $j \in V_j$: the pool already contains y_j, the conformal self-inclusion that removes the off-by-one bias. Where $g_0^{(r_j)} = 0$ the ratio is $0/0$: r_j scores zero on every member of Y_{V_j}, and $j \in V_j$, so the numerator vanishes too. Setting $E_j^{\text{sig}} = 1$ there is the only convention that keeps the proposition *exact*—0 is equally consistent but makes $\mathbb{E}[E_j^{\text{sig}}] < 1$. □

Proof of Proposition 2. Identical recipe with $Y_U = \{y_k\}_{k\in U}$ in place of Y_{V_j}; that substitution is the only difference between the two propositions. Fix $j \in U$ and condition on Y_U and on r_j. The denominator, the row mean $\frac{1}{|U|}\sum_{k\in U} r_j(y_k)$, is a symmetric function of Y_U, hence constant given the conditioning—in particular invariant to the within-scope pairing that H_0^{uniq} randomizes. Under H_0^{uniq} that pairing is exchangeable, so y_j is conditionally a uniform draw from Y_U, and

$$\mathbb{E}[M[j][j] \mid Y_U] = \mathbb{E}[r_j(y_j) \mid Y_U] = \frac{1}{|U|}\sum_{k\in U} r_j(y_k),$$

which is exactly the denominator. Hence $\mathbb{E}[E_j^{\text{uniq}} \mid Y_U] = 1$, and by the tower property $\mathbb{E}[E_j^{\text{uniq}}] = 1$. □

C.1 Deriving the invariance call

The product law. Conditional on the proposal data and the scope, the unit pairs (r_i, y_i) are drawn independently from one law, so for $i \neq j$ the pair (r_i, y_j) follows the product of its marginals, and $h(r, y) = r(y)$ decomposes under it as in Equation (6), with $a(r) = \mathbb{E}_y[h] - \mu$, $b(y) = \mathbb{E}_r[h] - \mu$ and $c = h - \mu - a - b$. Since c has mean zero over r for every y and over y for every r, the cross terms in $\frac{1}{2}\mathbb{E}\big[(a(r_1) - a(r_2) + c(r_1, y) - c(r_2, y))^2\big]$ vanish, leaving $\sigma_{\text{r}}^2 + \sigma_{\text{ry}}^2$, which is Equation (7). The diagonal is left out because (r_i, y_i) follows the joint law, not the product. For a unit-invariant hypothesis a and c vanish and $S_{\text{r}} = 0$ (Remark 1).

The estimate and its error. For column j let v_j be the unbiased variance of $\{M[i][j] : i \neq j\}$; then $\widehat{S}_{\mathrm{r}} = \frac{1}{m}\sum_j v_j$, with $m = |U|$, is the U-statistic of $\frac{1}{2}(M[i][j] - M[k][j])^2$ over distinct triples, unbiased for S_{r}, and $\bar{M}$, the mean of the $m(m-1)$ off-diagonal cells, is the U-statistic of $\frac{1}{2}(M[i][j] + M[j][i])$ over distinct pairs, unbiased for μ. Deleting unit u removes row u and column u; each v_j updates in $O(1)$ from its column's sum and sum of squares, and $\bar{M}$ from the row and column sums, so the m replicates $\widehat{D}^{(-u)} = \widehat{S}_{\mathrm{r}}^{(-u)} - (k\bar{M}^{(-u)})^2$ cost $O(m^2)$ together, and $\widehat{\mathrm{se}}^2 = \frac{m-1}{m}\sum_u \left(\widehat{D}^{(-u)} - \overline{D}\right)^2$ with $\overline{D}$ their mean. $\widehat{D}$ is a smooth function of two U-statistics of independent units. When the scores have finite fourth moments, so both kernels have finite second moments, and $\widehat{D}$'s first-order projection has positive variance, it is asymptotically normal (Hoeffding, 1948) and this jackknife estimates its variance consistently (Arvesen, 1969); when $S_{\mathrm{r}} \geq (k\mu)^2$ the probability that $\widehat{D} + z\,\widehat{\mathrm{se}} < 0$ therefore tends to at most α, and likewise for the variant call.

A finite-sample alternative. For scores in $[0, 1]$, as recall is, the kernel lies in $[0, \frac{1}{2}]$, so Hoeffding's inequality for an order-3 U-statistic (Hoeffding, 1963) gives $\widehat{S}_{\mathrm{r}} \pm \frac{1}{2}\sqrt{\ln(1/\alpha)/(2\lfloor m/3 \rfloor)}$ with no approximation at all. At our ceiling that half-width alone is a median 172 times $(k\bar{M})^2$ on the real corpus, so no real *Shared* hypothesis could certify; the call uses the jackknife.

How close the asymptotics are. We simulated the call at its boundary, $S_{\mathrm{r}} = (k\mu)^2$, with independent units and set-valued restrictions and outcomes—$|U|$ from 50 to 800, $|y| \in \{3, 6\}$, hit rates 0.1 and 0.5, 60,000 draws. It called *shared* 1.2–4.7% of the time per setting against $\alpha = 5\%$, and at most 0.9% at 1.1 times the ceiling; it called *variant* 2.9–8.5% of the time, 5.5% pooled. The two sides differ because the jackknife error moves with $\widehat{D}$: a high $\bar{M}$ raises the ceiling and the error together, so at these sizes the call leans away from *shared* and toward *variant*.

Conditions, and what is not proved. The bounds need units drawn independently given the scope and the proposal data, restrictions that depend only on their unit's input and on proposal randomness independent of the outcomes, and a hypothesis and scope chosen without reading a certification outcome; datasets whose units cluster or drift are outside our target. The ceiling is k times the mean off-diagonal score μ; $\bar{M}$ estimates it, and that error enters the jackknife with the spread's. The one approximation is the normal one behind the jackknife bound, roughest at small $|U|$ and small α. None of this assumes a law for the scores or independence between cells. A per-hypothesis resampling model is one modelling assumption per hypothesis; this is one declared ceiling, once.

D SETTING THE TOLERANCES

The certifier takes four numbers from its user, and they are not alike: one is an error budget, two define decision boundaries, and one is the noise ceiling the invariance call is tested under. Table 13 summarises them; the subsections say what each decides, which way its errors run, and how to pick it.

D.1 THE LEVEL α

α is the per-comparison error budget of Theorem 1 and enters as the single threshold $1/\alpha$ against which both e-values are read; it is per-comparison and not a batch quantity. α is also the level of the invariance call's equivalence test and of the coverage test, so one number sets every bar. One consequence is a *precondition* rather than a tolerance: $E^{\mathrm{uniq}} \leq |U|$ always, because the diagonal is one of the summands of its own row mean, and $E_j^{\mathrm{sig}} \leq |V| + 1$ for the same reason, its own outcome being one of the $|V| + 1$ in its pool. So a *Matched* certificate at level α is unreachable unless $|U| \geq 1/\alpha$—20 units at $\alpha = 0.05$, 100 at 0.01, 1000 at 0.001—and a *Shared* one unless $|V| \geq 1/\alpha - 1$. Neither floor involves the scope's width (Section D.3). At small scopes the invariance call loses power: its jackknife error widens as $|U|$ shrinks, so more scopes are undecided and declined rather than certified *Shared*, and the normal approximation behind its bound is roughest there.

Table 13: The four user inputs. None carries a scale. α, τ_{cov} and γ gate certification and err toward rejecting, so **none of the three has a failure direction that produces a false certificate**. The ceiling is the exception: it fixes what counts as one restriction, so too high a ceiling lets restrictions that genuinely differ certify as one *Shared* restriction, and it must be reported alongside any *Shared* certificate.

	decides	err in this direction	how to set it	transfers
α (as $1/\alpha$)	certify or not, on both e-values	too small: nothing certifies; needs $\lvert U\rvert \geq 1/\alpha$	your error budget for a single comparison	yes
τ_{cov}	whether enough units show the lift individually to certify	too high: real rules rejected—never miscertified	the share of units you require to outscore a random out-of-scope unit	yes
γ	whether a scope is the population: wider than $\gamma\lvert T\rvert$ reads as *base rate*	too low: a real effect on a wide scope rejected—never miscertified	the share of T beyond which a claim is about everyone: 0.9 by default	yes
k	*invariant* vs *variant*: what counts as one restriction (Section 4)	too high: an *Unmatched* scope certified *Shared*; too low: a shared restriction whose per-unit variation exceeds it is declined	the spread, as a fraction of the mean score, below which two restrictions count as one: 0.3 by default	declared

D.2 τ_{cov}: THE COVERAGE TEST

A certificate claims the lift for most of the scope, and the e-values cannot carry that claim: E^{sig} rejects a null under which *no* unit has lift, so a small subgroup can reject it alone, and E^{uniq} concerns the pairing, not how widely it holds. The coverage test carries it. The comparators v_j of Equation (4) are assigned uniformly at random, one member of V to a unit and without reading an outcome, and $B_j = \mathbf{1}\{r_j(y_j) > r_j(y_{v_j})\}$, ties counting as losses; cover is the mean of the B_j. With units i.i.d., in scope and in the reference, the pairs (Z_j, y_{v_j}) are i.i.d. across j, so the B_j are independent draws with a common rate π: the probability that a unit's own outcome outscores a random out-of-scope outcome under the unit's restriction. Then $\lvert U\rvert\,\text{cover} \sim \text{Bin}(\lvert U\rvert, \pi)$, and under $H_0^{\text{cov}} : \pi \leq \tau_{\text{cov}}$ the binomial tail p_{cov} of Equation (4) satisfies $\mathbb{P}(p_{\text{cov}} \leq \alpha) \leq \alpha$ exactly, since that tail increases with π. No asymptotics and no model of the scores enter, and π is a population rate that does not depend on which reference was drawn.

A comparator per unit is what makes the count binomial. Comparing every unit with the mean of one shared reference would make the B_j independent only given V, at a rate that moves with V: averaged over references the count is a mixture of binomials, and a reference that happens to be unrepresentative can certify coverage the population lacks—one shared restriction whose in-scope scores sit just below rare large reference scores certifies with probability $1/2$ when the population rate is exactly $1/2$. On our tiers the two constructions differ in one real certificate, and the random assignment moves no decision across five independent draws.

What it certifies is a rate of predictive success, not the share of units with a genuine effect: a unit can outscore its comparator by chance, and with one outcome per unit no test separates a persistent unit-level effect from a lucky hit without further assumptions. The rate is what *Sparse* turns on, since a lift carried by a minority keeps π near that minority's share. A failed test is not proof of a minority, though—a scope too small to show $\pi > \tau_{\text{cov}}$ fails it too—so *Sparse* reads as coverage not shown. τ_{cov} stays the user's declaration of how many units must show the lift, and being a fraction it transfers across scales. Like every gate it can overrule an overwhelming mechanism statistic: *simpson_cat* is rejected at $E^{\text{uniq}} = 28.4$ (Section B.2).

A stronger target, that at least $\lceil \tau_{\text{cov}} \lvert U\rvert \rceil$ units violate their own no-lift null, also has a valid e-value: the mean of the $\lvert U\rvert - \lceil \tau_{\text{cov}} \lvert U\rvert \rceil + 1$ smallest E_j^{sig}, a partial-conjunction bound that holds under any dependence. But a unit whose outcome misses scores $E_j^{\text{sig}} = 0$, so that mean reaches $1/\alpha$ only when most units hit strongly: taken with $\min(E_j^{\text{sig}}, E_j^{\text{uniq}})$ for *Matched*, it certifies 43 of the 68

planted *Matched* samples and none of the 18 real ones, where the coverage test certifies all 68 and 16.

On the planted tiers coverage is shown for all 136 samples of the two certifying types and for none of the 68 *Sparse* ones. Of the 49 real hypotheses whose other tests pass, 22 have cover below one half and ten more between 0.51 and 0.58, eight of them with 50 to 105 units, where a majority cannot be shown at that size; coverage is shown for the other 17.

D.3 γ: WHERE THE POPULATION BEGINS

Taking the baseline over all of T would put the scope inside its own baseline: a unit-invariant hypothesis supplies a fraction $|U|/|T|$ of it, so $E^{\text{sig}} \leq |T|/|U|$, and a real shared effect on any scope wider than α of T could never certify. Measured that way, a real scope-level effect planted at matched scope sizes held $E^{\text{sig}} = 3.17$ at $|U|/|T| = 0.30$, against a pure base rate's 1.00, and 2.41 at 0.40—real lift, far below $1/\alpha$. Drawing the reference from outside the scope removes that cap: $E^{\text{sig}}_j \leq |V| + 1$ at any width, the size floor E^{uniq} already carries (Section D.1).

What the reference leaves open is where a scope stops being one. A claim about nearly all of T is a claim about the population, which is what *base rate* means, and what remains outside it is too little of T to stand for that population. γ marks the point: a scope covering more than $\gamma|T|$ is *base rate* whatever its statistics say. It is a definition rather than a measurement, set by that question alone and 0.9 by default. Like cover it can only reject, so it has no failure direction that produces a false certificate, and being a fraction of T it transfers across corpora unchanged.

D.4 SENSITIVITY OF THE REPORTED RESULTS

The level. α is the one constant Theorem 1 depends on, so we sweep it. The full rule is barely sensitive to it:

		decisions matching the target, planted (408) / real (101)		
α	$1/\alpha$	full rule	E^{sig} and E^{uniq} only	baseline CRT
0.20	5	**408 / 99**	272 / 76	251 / 70
0.10	10	**408 / 97**	272 / 76	264 / 72
0.05	20	**408 / 97**	272 / 79	270 / 74

Every gate reads α: the invariance call through its confidence bounds and the coverage test through its level. Over this range the level changes at most nine of our 408 planted verdicts and none of their decisions, and six of the 101 real verdicts, four of them decisions. Agreement with the target is not the quantity Theorem 1 bounds, so it cannot show whether the bound binds; the permutation nulls of Section E check that directly.

The noise ceiling. k is declared, so what matters is what a different declaration would change. Recomputing the call from the stored matrices over $k \in [0.2, 0.5]$: every *Shared* plant on both planted tiers is called shared throughout, and what k moves is the confounds. None is called shared at $k \leq 0.3$; from 0.35 the thirteen generator segment confounds are, their restrictions spreading by 0.32–0.34$\bar{M}$, and all eighteen intervention confounds at 0.5. On the real corpus 2 of the 3 *Shared* hypotheses certify at every point of the grid, while four hypotheses of the audited *Unmatched* type are called shared from 0.45 and five at 0.5.

τ_{cov} binds on the real corpus, where 32 of the 101 have cover inside $[0.5, 0.7]$. It moves both certificates: from 49 at the loosest settings, the first is lost at 0.15, 17 remain at 0.5, none is *Shared* from 0.55, and none is *Matched* from 0.7. Decisions matching the target peak at 106 at 0.4, but with two false certificates; there are none from 0.45, and 97 match at the stated 0.5.

Tighter levels are not an operating point at these scope sizes: $E^{\text{uniq}} \leq |U|$, and $\alpha = 0.01$ would need $E^{\text{uniq}} \geq 100$ against a real-corpus maximum of 52.8 (Section D.1).

Coverage. Sweeping τ_{cov} with the others at their stated values, the 408 planted certify decisions all match the target across $[0.3, 0.6]$, so the reported result does not depend on where in that range the gate sits. That is not evidence that the value *transfers*: the flat region is the gap between the planted classes.

E THE REAL-CORPUS TIER

Table 6 scores the certifier on 101 hypotheses over the Amazon Video-Games corpus. This appendix says how they were built and labelled, and checks what needs no label.

Setup. Corpus, split and outcome definition as in Section B.1, with $\alpha = 0.05$, $\tau_{\text{cov}} = 0.5$ and the noise ceiling at $k = 0.3$, as in Section 5; the certifier reads at most 800 units of each scope and draws a reference of 2000 units from outside it, all of them where fewer remain. All are declared rather than fitted.

Both exchangeability e-values have mean 1 on a null true by construction. From the stored score matrices—each restriction scored on its scope's outcomes and on the reference's—we permute the outcomes over scope and reference together, which makes H_0^{sig} true, and separately the pairing within the scope, which makes H_0^{uniq} true: 75 draws of each for every one of the 101 hypotheses. E^{sig} pools to 0.991 and E^{uniq} to 1.001, standard errors 0.007 and 0.003, and no draw reaches $1/\alpha$: the largest E^{sig} is 15.2 and the largest E^{uniq} 4.8. What a permutation cannot decide is whether the real outcomes were exchangeable to begin with; this checks the estimator, not the assumption.

E.1 HOW THE 101 HYPOTHESES WERE BUILT

The hypothesis were written by a LLM with no context of our algorithm, and labeled by human. The 101 are 18 *Matched*, 3 *Shared*, 20 *base rate*, 22 *Sparse*, 20 *Unsupported* and 18 *Unmatched* hypotheses.

The statistics separate rather than degenerate. Over the 101, E^{sig} spans 0.43 to 1101, E^{uniq} 0.43 to 52.8, the restriction spread $\sqrt{\widehat{S}_{\text{r}}}/\bar{M}$ 0 to 55.5 against a ceiling of 0.3, and cover 0 to 0.74, across scopes of 45 to 19,467 users.

F BATCH CONTROL OVER A STREAM OF HYPOTHESES

Theorem 1 bounds the error of *one* certification, and a proposer emitting many accumulates them in proportion.

E^{sig} is an e-value, which is what makes the batch question cheap to answer: e-BH (Wang & Ramdas, 2022) takes e-values directly and controls the false discovery rate under *arbitrary* dependence—the regime here, since all 101 hypotheses of Table 6 are read off one shared outcome multiset. It selects the top k^*, $k^* = \max\{k : e_{(k)} \geq m/(k\alpha)\}$. That bar is far above $1/\alpha$ except in the limit: at $m = 101$ and $\alpha = 0.05$ it uses all 101 e-values, 30.15 of the sixty-seventh, and reaches 20 only at $k = 101$.

	clears $1/\alpha$	survives e-BH	largest e-value
E^{sig}	71	**67**	1101.0
by verdict: *Matched* / *Shared* / *Unmatched* / *Sparse*	15/2/22/32	**15**/**2**/20/30	

The lift of every certificate survives the batch bar. All 15 *Matched* and both *Shared* certificates clear the e-BH threshold on E^{sig}, as do 20 of the 22 *Unmatched* and 30 of the 32 *Sparse* rejects; the 20 *base rate* readings do not, as a base rate should not.