ASSUMPTION-STRATIFIED AUDITING OF DETERMINISTIC EVALUATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

When audit queries are scarce, every choice of where to look commits to a failure model. We study deterministic evaluators on a finite counterfactual frontier. Assumption-stratified auditing begins with broad coverage (Q), then uses executable faulty evaluators (H) to focus probability on probes that expose modeled bugs. Its core, STABLE-H, asks whether that focus survives the removal of one modeled fault story: delete each distinct H hit pattern, recompute the best Q/H audit, and seek one randomized policy that remains near-optimal across all leave-one worlds. The required slack $\epsilon^{\star}_{H}$ is a design-time diagnosis: zero means those worlds share an exact optimum, while positive slack marks an allocation conflict that must be reconciled. In a 13,500-task controlled study, all 27 trajectories move in the predicted direction. On the predefined 19-task operational cohort, the common solve strengthens deletion stability and worst-case H assurance on 19/19 tasks. When H itself is scarce, complementary 50% H matches typical random 75% H on full-library certificate quality. An exact shared-hit-event decomposition certifies nonzero-stability instances with up to 1.77×10^{8} implicit plans. Stronger semantic specialization remains optional and helps only when its geometry transfers.

1 INTRODUCTION

Suppose a contract says that an answer may contain at most five sentences. The observed answer has four. Two implementations return PASS: the correct checker and a broken checker that always accepts. A six-sentence answer separates them because the contract requires FAIL.

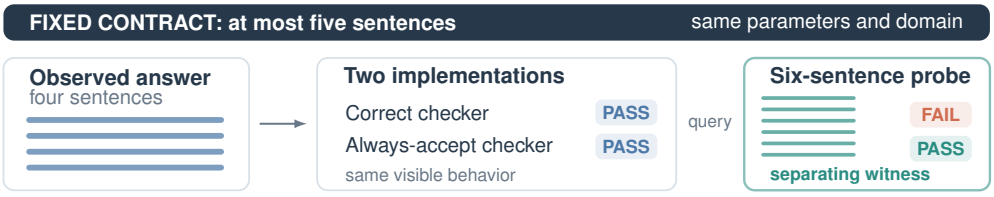

Figure 1: **A correct observation can hide a functional fault.** The implementations agree on the observed answer but disagree on a six-sentence counterfactual.

Now give the auditor only two additional queries. Candidate probes include answers with 2, 5, 6, or 8 sentences, malformed punctuation, and reordered structure. Testing one location means not testing another.

With no failure model, spending both queries near the five-sentence boundary is dangerous. Q therefore keeps every probe alive. Executable off-by-one, parser, and ordering bugs then provide a reason to concentrate: H moves probability toward the probes that expose those modeled faults. The resulting audit depends on the stories in H. STABLE-H makes that dependence visible by removing each distinct hit pattern in turn and recomputing the best audit. If the leave-one audits disagree, it constructs one randomized policy that reconciles them. S can specialize further when independently defined semantic mechanisms agree, but it is an optional stronger assumption.

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.

Existing testing and design tools provide coverage objectives, executable mutations, and robust optimization primitives. Here the central object is the audit policy induced by those tools: how stronger information reallocates a scarce query budget, whether removing one modeled story changes the optimal allocation, and what broad protection remains. These are design-time properties of the audit, distinct from its later detection rate on held-out faults.

STABLE-H is the methodological center. Its slack $\epsilon_H^\star$ diagnoses whether the leave-one worlds share an exact optimum or require reconciliation. Constructing this common randomized certificate appears combinatorial because its variables are complete audit plans. The same probe- and H-hit events recur across all leave-one worlds, which permits exact certification without enumerating the plan universe. When the H library is itself limited, complementary hit patterns also use that information budget more efficiently.

We study the deterministic, stateless base case, where a fault is exposed by hitting a fixed witness set on a fixed frontier. Stochastic evaluators require repeated evidence; stateful evaluators replace fixed plans with adaptive policies. Our three contributions are:

- **Assumption-aware allocation.** A Q/H/S ladder starts from broad coverage and makes each stronger concentration an explicit commitment with a recorded fallback.
- **Self-diagnosing stability.** STABLE-H deletes each distinct H pattern, tests whether the induced audits agree, and introduces positive common slack only when their requirements conflict.
- **Practical certification.** Complementary hit geometry makes a limited H library more certificate-efficient, while shared hit events make the full STABLE-H certificate exactly computable over large implicit plan spaces.

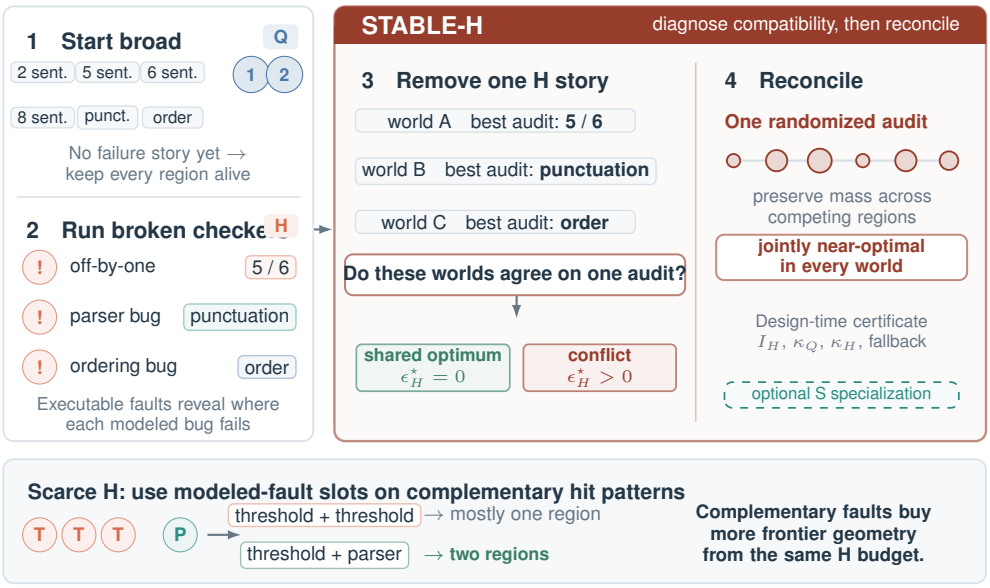

Figure 2: **Start broad, diagnose the modeled stories, then stabilize only when they disagree.** Executable faults reveal where particular bugs fail. STABLE-H asks whether the audits induced by removing those stories are compatible and, when necessary, reconciles them in one randomized policy. The inset shows why complementary patterns are useful when H is scarce.

2 Auditing as Hitting Unseen Witnesses

Fix a finite frontier Q of valid counterfactual probes and an independently justified verdict $A(q)$ for each probe. A deterministic fault g has witness set

$$D_g = \{q \in Q : g(q) \neq A(q)\}. \tag{1}$$

D_g is simply the set of probes that expose g. For the always-accept checker, a six-sentence answer belongs to D_g; a two-sentence answer does not.

An audit plan P contains exactly B probes. A policy λ randomizes over the feasible plans Π and detects a witness set D with probability

$$\phi_\lambda(D) = \Pr_{P\sim\lambda}[P \cap D \neq \varnothing]. \tag{2}$$

This is the chance that a randomized B-query audit touches at least one revealing probe. With $Q = \{2, 5, 6, 8, \text{punctuation}, \text{order}\}$ and $B = 2$, always querying 5 and 6 detects a threshold fault but gives a punctuation-only fault zero chance. Randomization can keep broad protection while still favoring a structured region.

For a witness family $\mathcal{D}$, define $\kappa_{\mathcal{D}}(\lambda) = \min_{D\in\mathcal{D}} \phi_\lambda(D)$. This protects the hardest-to-hit set in the family. Q uses singleton probes; H uses rows produced by executable faults; S uses regions inferred from separately specified semantic mechanisms. None is guaranteed to contain the unknown fault. Each is an assumption about where probability should move.

3 From Broad Coverage to Self-Diagnosing Stability

3.1 Start broad: Q

If the checker's failure mode is unknown, punctuation, ordering, short answers, and the 5/6 boundary should all retain some chance of being tested. Q therefore maximizes the worst probe-inclusion probability before trusting a narrower story.

Let $N = |Q|$. If every B-subset is feasible and an unknown fault has at least w witnesses, uniform plan sampling has exact optimal assurance

$$\kappa^\star_w = \begin{cases} 1 - \binom{N-w}{B}/\binom{N}{B}, & w \le N - B, \\ 1, & w > N - B. \end{cases} \tag{3}$$

For singleton witnesses, every probe is included with probability B/N; in the six-probe example, broad assurance is $2/6$. Restricted fixtures or ordering constraints simply replace the complete universe with the actual finite plan set (Appendix A). Once a structured policy drives any probe below the attainable Q guarantee, it has spent broad fallback.

3.2 Use executable evidence, then diagnose it: H and STABLE-H

Now deliberately implement broken checkers. An off-by-one checker may fail at five or six sentences, a parser bug on malformed punctuation, and an ordering bug on reordered structure. Running each checker over the frontier produces an H hit row: the probes that expose that modeled fault. H maximizes assurance against the hardest-to-hit row, concentrating the audit where the executable faults provide evidence.

STABLE-H asks how much that allocation depends on any one modeled story. We first merge faults with identical hit rows. Then we remove each distinct row in turn and recompute the best Q/H audit. After deleting one checker, the best audit may still query 5 and 6 sentences. After deleting another, it may move a query to punctuation; a third deletion may favor order. These are counterfactual leave-one worlds induced entirely by design-time information. The question is whether one randomized audit can remain close to the optimum in every world.

Q and H have different best attainable assurances, so raw losses are not comparable. Normalized regret measures the fraction of each layer's own guarantee that a policy gives up. For an active layer θ with positive optimum $\kappa^\star_\theta$, define

$$\rho_\theta(\lambda) = 1 - \frac{\kappa_\theta(\lambda)}{\kappa^\star_\theta}, \qquad \rho_\Theta(\lambda) = \max_{\theta\in\Theta} \rho_\theta(\lambda). \tag{4}$$

Regret zero means no loss relative to that layer's optimum; larger regret means protection was spent elsewhere. In the checker example, this puts lost broad probe coverage and lost modeled-fault coverage on a common scale.

Let $\rho^\star_{-u}$ be the best Q/H regret after deleting distinct H row u. STABLE-H chooses the policy requiring the smallest common relaxation:

$$\epsilon^\star_H = \min_{\lambda \in \Delta(\Pi)} \max_u \big[\rho_{-u}(\lambda) - \rho^\star_{-u}\big]. \tag{5}$$

The inner maximum is the policy's deletion-instability score $I_H(\lambda)$; lower is a stronger common leave-one certificate, and STABLE-H minimizes it. The optimum $\epsilon^\star_H$ is therefore a design-time conflict diagnosis. No usable H capacity leaves the audit at Q. If $\epsilon^\star_H = 0$, the leave-one worlds share an exact optimum. If $\epsilon^\star_H > 0$, no policy satisfies all of their optimal requirements exactly, and the value is the minimum slack needed to reconcile them. It is not a predictor of held-out fault detection.

Proposition 1 (Common leave-one certificate and Q fallback)**.** *Let $\mathcal{A}_H$ be the distinct H hit rows over a finite feasible plan universe. The STABLE-H set*

$$F^{\mathrm{stab}}_H = \big\{\lambda : \rho_{-u}(\lambda) \le \rho^\star_{-u} + \epsilon^\star_H \text{ for every } u \in \mathcal{A}_H\big\} \tag{6}$$

is a nonempty compact convex polytope. Every $\lambda \in F^{\mathrm{stab}}_H$ satisfies

$$\kappa_Q(\lambda) \ge \kappa^\star_Q \left[1 - \epsilon^\star_H - \min_{u \in \mathcal{A}_H} \rho^\star_{-u}\right]_+ . \tag{7}$$

Duplicate raw faults with the same hit row do not change this certificate.

Thus the common certificate records both proximity to every leave-one optimum and a broad Q fallback. The fallback applies to nonempty witness sets on the declared frontier; it is not a held-out generalization guarantee. The proof is in Appendix A.7.

A simpler construction is to solve every leave-one world independently and average the resulting policies; we call it AVG-REDUCED. STABLE-H asks the stronger joint question: can one policy be simultaneously near-optimal for every world? An average of separate optima need not have that common guarantee.

Leave-one is a first-order test: correlated misspecification requires deleting several related patterns.

The H library can itself be scarce. Suppose four broken checkers are available but only two can be retained. Three threshold variants light up the same 5/6 probes, while another exposes punctuation. Keeping two threshold variants spends two modeling slots on nearly one region; keeping one threshold and one parser fault covers two. *Two similar modeled bugs can buy one piece of geometry; two complementary bugs can buy two.*

All leave-one problems constrain the same probe- and H-hit events. The exact solver in Section 4.4 shares generated plans across them, separates only violated original constraints, and reuses one weighted-coverage pricing oracle instead of enumerating Π.

3.3 Optional stronger specialization and explicit fallback

S asks whether independently labeled training faults expose the same semantic region. In the sentence checker, several faults labeled "sentence counting" may all fail at 5 and 6 sentences even if their remaining witnesses differ. If $R_i \subseteq Q$ is the exposure signature of training instance i assigned to semantic mechanism m, its strict core is

$$C_m = \bigcap_{i \in \mathcal{I}^{\mathrm{train}}_m} R_i. \tag{8}$$

For exposure sets $\{5, 6, 8\}$, $\{5, 6\}$, and $\{5, 6, \text{punctuation}\}$, the core is $\{5, 6\}$. S may concentrate there, but an interpretable mechanism label need not transfer to an unseen witness set. We therefore place S after H stability as an optional stronger specialization, not as the paper's primary guarantee.

Suppose the auditor is willing to lose at most a declared amount of weaker protection in exchange for stronger specialization. δ is that risk budget, not a value selected from hidden detection. If Q is in the weaker baseline, the selected policy satisfies

$$\kappa_Q(\lambda_\delta) \ge \kappa^\star_Q(1 - \rho^\star_0 - \delta). \tag{9}$$

The auditor may set $\delta = 0$, predeclare an acceptable loss, or publish the whole frontier. If the stronger layer cannot improve its own certified assurance within that budget, the weaker policy is retained. Over a finite plan universe the certified stronger-assumption frontier is nondecreasing, concave, continuous, and piecewise linear; real downstream exchange rates are task-specific (Appendix A.4 and Appendix D.4).

3.4 CERTIFY A RANDOMIZED AUDIT

The output is a distribution, not one privileged test set. Before held-out faults are revealed, the certificate records the frontier and plan rules, deduplicated H rows, any S cores, $(\kappa_Q, \kappa_H, \epsilon_H^\star, \kappa_S, \delta)$, active constraints, policy support, and a public seed. These are design-time guarantees under the declared assumptions.

Held-out detection is computed only after evaluation faults reveal their witness sets. In the worked `detectable_format:title` example, minimum and mean detection are approximately .204 and .435 per draw, while one public-seed plan detects one of three faults. The probabilities describe the randomized policy; the sampled plan is one realization. Appendix B.2 gives the complete record.

4 WHAT THE ASSUMPTION LADDER BUYS

We evaluate two different objects: design-time certificate quantities that the audit explicitly optimizes, and held-out detection observed only after evaluation faults are revealed. The former include κ_Q, κ_H, $\epsilon_H^\star$, I_H, and certified fallback; the latter measures downstream usefulness but is not the STABLE-H objective. The evidence follows four questions, with protocols and complete tables in Appendix C.

4.1 DOES STABLE-H REACT CORRECTLY WHEN A MODELED STORY BECOMES UNRELIABLE?

The controlled experiment contains 13,500 synthetic tasks. Its hidden-fault generator does not inspect the audit policies and varies how strongly one singled-out H pattern represents hidden faults. When that pattern is less representative, reducing reliance on it should help; as it becomes more representative, the advantage should shrink. All 27/27 endpoint trajectories move in the predicted direction (Figure 3a). This is mechanism evidence that STABLE-H responds as intended when one modeled story becomes less trustworthy.

4.2 WHEN IS STABLE-H NEEDED, AND WHAT DOES THE COMMON SOLVE BUY?

Among the 30 development tasks, 11 have no usable joint H capacity under the full-library definition and therefore remain at Q. The remaining 19 form the predefined operational H cohort. All 19 have $\epsilon_H^\star > 0$: whenever H supports the joint objective in this cohort, the leave-one worlds require genuine reconciliation (Figure 3b).

Executable faults are informative on these tasks. Moving from Q to the ordinary Q/H audit raises worst-case modeled-fault assurance by approximately .298, with all 19/19 tasks moving in the same direction. Moving from that ordinary Q/H policy to STABLE-H reduces leave-one exposure by about .047.

A simpler alternative, AVG-REDUCED, solves every leave-one world independently and averages the resulting policies. It nearly matches STABLE-H in held-out detection, but it does not optimize one policy jointly across all worlds. Relative to AVG-REDUCED, STABLE-H lowers deletion instability I_H by mean .0556 on 19/19 tasks and raises worst-case H assurance by mean .0365 on 19/19 (Figure 3c). This consistent strengthening of the common leave-one certificate is the central empirical result for STABLE-H.

Held-out detection is evaluation-only. STABLE-H minus the ordinary Q/H policy is $+.0053$ on this cohort; STABLE-H minus AVG-REDUCED is $+.0018$ with interval $[-.0040, +.0088]$. Thus the separate averaging construction nearly matches downstream detection, while the common solve supplies the stronger certificate it explicitly targets. Additional Q-matched and diversity controls are in Appendix C.5.

4.3 WHAT IF MODELED-FAULT INFORMATION IS ALSO SCARCE?

Half a complementary H library roughly matches a typical three-quarter random library on the certificate axis. At 50% H, complementary selection has certificate gap .1400 and full-library H coverage .5256; typical random 75% H gives .1452 and .4936 (Figure 3d). The selection uses only

design-time hit signatures, so this is an information-efficiency result rather than hidden-fault tuning. Same-budget comparisons and other H definitions are in Appendix D.2.

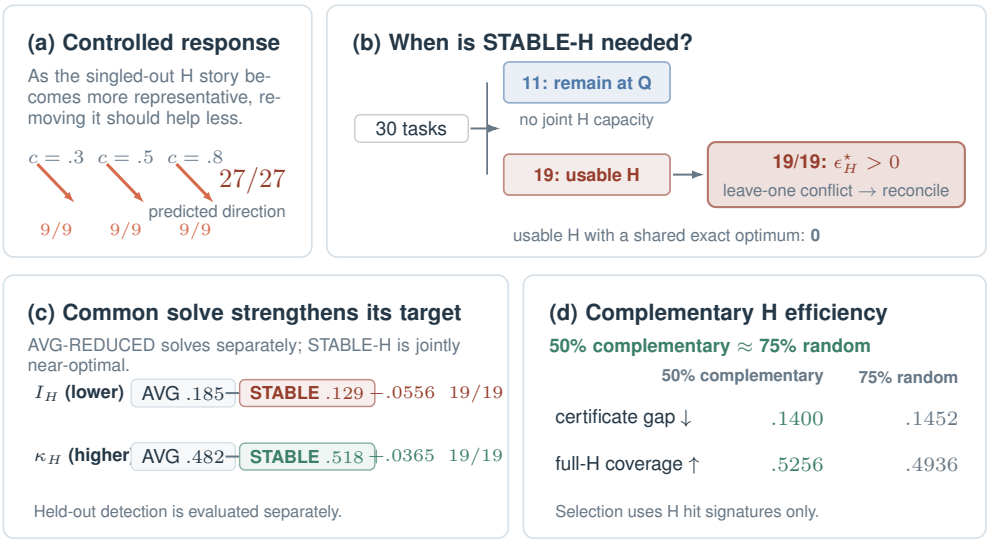

Figure 3: **Evidence follows the method's decision path.** Controlled misspecification moves STABLE-H in the predicted direction; the full-library diagnosis sends 11 tasks to Q and identifies leave-one conflict on the remaining 19; the common solve strengthens both certificate measures on 19/19 tasks; and complementary H recovers similar certificate quality with less modeled-fault information.

4.4 Can the full certificate be computed at scale?

The largest closed instance contains 177,232,627 implicit plans and receives an exact full-STABLE-H certificate. This is possible because every leave-one world uses the same probe- and H-hit events. The decomposition shares generated plans across the reduced problems and the common master, adds only violated original constraints, and uses exact weighted maximum coverage to find new plans. The optimization machinery is classical; the specialization exploits the shared events of this certificate.

Proposition 2 (Exact termination certificate)**.** *Assume the feasible plan universe is finite, the reduced quantities are solved exactly, every original deletion constraint is separated exactly, and plan pricing is exact. If the decomposition terminates with no violated deletion row and no improving plan, its restricted-master solution is optimal for the original finite STABLE-H problem.*

The proof is in Appendix C.11.

The decomposition matches dense full-STABLE-H on all 19 operational tasks and on 24/24 lifted comparisons where dense evaluation is feasible. It closes five positive-slack cells from 296,010 to 177,232,627 implicit plans. The largest has $|Q| = 44$, $B = 8$, 32 distinct H rows, and $\epsilon_H^\star = .028852$, requiring about 22.6 s and 6 MB peak memory. Other attempted positive-slack cells remain unclosed; the complete 17-cell accounting and the proof of Proposition 2 appear in Appendix C.11.

Stronger assumptions remain conditional. S completes the ladder but does not dominate it. A narrow seven-task cohort has a $+.0293$ semantic increment, whereas the predefined expanded 26-task effect is $-.0041$ with interval $[-.0247, +.0161]$ and changes sign across evaluator/archetype families. An interpretable mechanism is therefore not automatically transferable geometry. Further semantic and transfer results are in Appendix C.8 and Appendix D.

5 Related Work

Active and discriminating design ask how to allocate observations to separate candidates (Atkinson & Fedorov, 1975; Naghshvar & Javidi, 2013; Nitinawarat et al., 2013). Mutation and metamorphic testing provide executable fault structure under limited oracles (Jia & Harman, 2011; Chen et al., 2018; Barr et al., 2015), while behavioral testing and LLM-judge work study coverage and evaluator reliability (Ribeiro et al., 2020; Zheng et al., 2023; Raina et al., 2024; Tan et al., 2025). Robust design, minimax regret, distributionally robust optimization, and group DRO protect objectives under model or group uncertainty (Dette et al., 2013; Rahimian & Mehrotra, 2019; Sagawa et al., 2020).

These traditions primarily ask which tests, observations, or modeled faults are informative, or how to protect a fixed objective. STABLE-H instead studies whether the optimal audit policies induced by counterfactual removal of modeled fault assumptions are mutually compatible, and constructs one randomized policy when they are not. Maximin optimization and generic decomposition remain classical; the contribution is the leave-one audit contract, its common certificate, and the shared-event specialization that computes it exactly.

6 Limitations

Guarantees condition on a fixed frontier, oracle, and plan set; faults without frontier witnesses are undetectable. Q and the oracle, H libraries, and S ontologies remain human-authored, and authoring time was not measured. Leave-one is first-order robustness: $\epsilon_H^\star$ diagnoses conflict among declared H rows, not future detection. No ex-ante rule identifies a transferable S mechanism.

The real cohorts are small, and secondary H-definition analyses reuse development data. Large-scale exact certification covers the tested full-STABLE-H instances, not the S/bounded-loss composition, and some positive-slack cells remain unclosed. Stochastic or stateful evaluators require repeated evidence or richer adaptive policy classes.

7 Conclusion

A scarce audit budget forces a choice about where to look. Assumption- stratified auditing starts broad, then treats executable and semantic information as explicit allocation commitments. STABLE-H asks the central design-time question: after removing one modeled fault story, do the induced audits still agree? If not, it constructs one randomized policy that is jointly near-optimal across the leave-one worlds. On the operational tasks, that common solve consistently strengthens the certificate it targets; held-out detection remains a separate evaluation quantity. Complementary hit geometry uses a limited H budget more efficiently, and shared hit events make the certificate exactly computable over very large implicit plan spaces.

REFERENCES

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

Earl T. Barr, Mark Harman, Phil McMinn, Muzammil Shahbaz, and Shin Yoo. The oracle problem in software testing: A survey. *IEEE Transactions on Software Engineering*, 41(5):507–525, 2015. doi: 10.1109/TSE.2014.2372785.

Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, T. H. Tse, and Zhi Quan Zhou. Metamorphic testing: A review of challenges and opportunities. *ACM Computing Surveys*, 51(1):4:1–4:27, 2018. doi: 10.1145/3143561.

Holger Dette, Viatcheslav B. Melas, and Petr Shpilev. Robust T-optimal discriminating designs. *The Annals of Statistics*, 41(4):1693–1715, 2013. doi: 10.1214/13-AOS1117.

Daniel H. Gottlieb. A certain class of incidence matrices. *Proceedings of the American Mathematical Society*, 17(6):1233–1237, 1966. doi: 10.1090/S0002-9939-1966-0205871-6.

Yue Jia and Mark Harman. An analysis and survey of the development of mutation testing. *IEEE Transactions on Software Engineering*, 37(5):649–678, 2011. doi: 10.1109/TSE.2010.62.

Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013. doi: 10.1214/13-AOS1144.

Sirin Nitinawarat, George K. Atia, and Venugopal V. Veeravalli. Controlled sensing for multihypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013. doi: 10.1109/TAC.2013.2261188.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Vyas Raina, Adian Liusie, and Mark Gales. Is LLM-as-a-judge robust? investigating universal adversarial attacks on zero-shot LLM assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7499–7517. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.427.

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

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

REPRODUCIBILITY STATEMENT

The appendix gives the proofs, finite procedures, protocol, uncertainty, failed-cell, construction-burden, and scope details. Manuscript-facing summaries were reconciled against the available frozen or versioned records. The claim audit distinguishes locally reverified results from outcomes supplied only through locked revision records, and no missing plan-level certificate fields were reconstructed.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, experimental engineering, code and infrastructure, and reproducibility and audit workflows. The authors defined the research questions, protocol freezes, evaluation boundaries, and claims; reviewed the resulting artifacts and analyses; and take responsibility for the final content.

A WITNESS-FAMILY PROOFS AND FINITE PROCEDURES

This appendix gives the finite proofs behind Section 3. The subset-inclusion rank, finite parametric-LP structure, matching facts, and Hitting Set reduction are classical. Their role is to calibrate audit-design commitments exactly, not to claim standalone mathematical innovations.

For every active protection family θ with positive standalone capacity, we use

$$\rho_\theta(\lambda) = 1 - \frac{\kappa_\theta(\lambda)}{\kappa^\star_\theta}, \qquad \rho_\Theta(\lambda) = \max_{\theta\in\Theta} \rho_\theta(\lambda), \qquad \rho^\star_\Theta = \min_\lambda \rho_\Theta(\lambda).$$

Zero-capacity families are reported as frontier insufficiency rather than assigned a normalized regret with a zero denominator. Every reduced STABLE-H problem recomputes its active-family capacities before applying this definition.

Theorem 3 (Complete fixed-budget audits)**.** *Let $1 \le B \le N$ and $1 \le w \le N$. In the complete plan universe, when every allowed witness set has at least w probes,*

$$\kappa^\star_w = \begin{cases} 1 - \dfrac{\binom{N-w}{B}}{\binom{N}{B}}, & w \le N - B, \\ 1, & w > N - B. \end{cases} \tag{10}$$

For $w \le N - B$,

$$\operatorname{affdim} \Lambda^\star_w = \binom{N}{B} - \min\left\{ \binom{N}{w}, \binom{N}{B} \right\}, \tag{11}$$

while for $w > N - B$ every policy is optimal and $\operatorname{affdim} \Lambda^\star_w = \binom{N}{B} - 1$.

A.1 PROOF OF THE COMPLETE-UNIVERSE THEOREM

Average miss identity and value. Index the rows of M_w by w-subsets D of Q and its columns by B-subsets S, with $(M_w)_{D,S} = 1$ exactly when $S \cap D = \emptyset$. For a policy λ, $M_w\lambda$ is its vector of miss probabilities. Every fixed B-plan misses exactly $\binom{N-B}{w}$ of the $\binom{N}{w}$ possible w-witness sets. Consequently,

$$\frac{1}{\binom{N}{w}} \mathbf{1}^\top M_w \lambda = \frac{\binom{N-B}{w}}{\binom{N}{w}} = \frac{\binom{N-w}{B}}{\binom{N}{B}} = c_w.$$

Thus some w-set has detection probability at most $1 - c_w$. Uniform sampling over $\binom{Q}{B}$ has miss probability c_w for every w-set and attains that bound. If $w > N - B$, no B-plan can avoid a w-set, so every policy detects every admissible set with probability one.

Equality and the affine dimension. Suppose $w \le N - B$. A policy is optimal exactly when every w-set has detection probability at least $1 - c_w$. The average identity then forces every miss probability to equal c_w. Hence

$$\Lambda^\star_w = \{\lambda \ge 0 : M_w\lambda = c_w \mathbf{1}\}. \tag{A.1}$$

The displayed equalities already imply normalization. Indeed, summing them gives

$$\binom{N-B}{w} \mathbf{1}^\top \lambda = \binom{N}{w} c_w = \binom{N-B}{w},$$

and $\binom{N-B}{w} > 0$ in this regime.

Complementing each B-set column identifies M_w with the classical inclusion matrix whose rows are w-sets and whose columns are $(N-B)$-sets. Its real rank is

$$\operatorname{rank}(M_w) = \min\left\{ \binom{N}{w}, \binom{N}{N-B} \right\} = \min\left\{ \binom{N}{w}, \binom{N}{B} \right\} \qquad (w \le N - B),$$

by the classical inclusion-matrix rank theorem (Gottlieb, 1966). The uniform policy is a strictly positive solution of Equation (A.1). It is therefore in the relative interior of the nonnegative solution polytope, so nonnegativity removes no affine directions. Rank–nullity proves Equation (11). In the saturated regime, $M_w = 0$ and the full simplex is optimal, with dimension $\binom{N}{B} - 1$.

Edge-case audit. The argument includes $B = 1$: for $w = 1$ it gives the unique uniform one-probe policy, while $w = N$ is saturated and all distributions over probes are optimal. For $w = 1$ generally, equal miss probabilities are precisely equal probe-inclusion probabilities B/N; at $B = 2$ this recovers the Q-optimal marginal statement. For $w = B \le N - B$, the rank is $\binom{N}{B}$ and uniform sampling is unique. For $w = N - B$, the same full-column rank makes the uniform policy unique. For every $w > N - B$, including the case $B = N$, every feasible plan hits every admissible witness set; the stated simplex dimension also gives zero when there is only one plan.

Mechanical checks. We constructed the miss matrices exactly and checked their ranks over $\mathbb{R}$ for every $1 \le B, w \le N \le 9$ in the nonsaturated regime. Every rank was $\min\{\binom{N}{w}, \binom{N}{B}\}$, and every column had the common positive row sum $\binom{N-B}{w}$ required for normalization redundancy. These finite checks include $B > N/2$; the arbitrary-N statement follows from the cited classical rank theorem, not from enumeration.

A.2 The fixed-fault two-probe illustration

For the complete unordered distinct-pair universe, write $x_{ij} = \Pr_\lambda(\{i, j\} \text{ is queried})$ and $d_i = \sum_{j \neq i} x_{ij}$. The $B = 2, w = 1$ specialization of Theorem 3 says that Q-optimality is equivalent to $d_i = 2/N$ for every i.

For a fixed $D \subseteq Q$, let $E(D)$ be the edges internal to D and let $\delta(D)$ be its cut. Then

$$\sum_{i \in D} d_i = 2x(E(D)) + x(\delta(D)), \qquad \phi_x(D) = x(E(D)) + x(\delta(D)).$$

On the Q-optimal face, if $w = |D|$,

$$\phi_x(D) = \frac{w}{N} + \frac{1}{2} x(\delta(D)). \tag{A.2}$$

Consequently,

$$\frac{w}{N} \le \phi_x(D) \le \min\left\{1, \frac{2w}{N}\right\}. \tag{12}$$

Nonnegativity gives the lower endpoint in Equation (12). For $2 \le w \le N - 2$, it is attained by assigning uniform weight $2/[N(w-1)]$ to edges inside D and uniform weight $2/[N(N-w-1)]$ to edges inside D^c.

The cut mass is at most the total degree on either side: $x(\delta(D)) \le \min\{2w/N, 2(N-w)/N\}$. If $w \le N/2$, assigning each crossing edge weight $2/[N(N-w)]$ and spreading the remaining degree uniformly inside D^c attains this bound. If $w \ge N/2$, the symmetric construction assigns each crossing edge weight $2/(Nw)$ and spreads the remaining degree inside D. Substitution into Equation (A.2) gives the upper endpoint. For $w = 1$, detection is $2/N$; for $w = N-1, N$, every pair intersects D. These constructions are separately chosen for a fixed D and do not assert simultaneous endpoints for all D.

Pair requirements and matching. For a possible pair witness $e = \{i, j\}$ on the Q-optimal face,

$$\phi_x(e) = d_i + d_j - x_{ij} = \frac{4}{N} - x_{ij}.$$

If E_M is the collection of mechanism-supported pair requirements, then $\kappa_{S,2}(x) = 4/N - \max_{e \in E_M} x_e$. Achieving $4/N$ is equivalent to setting $x_e = 0$ on E_M. Under $y_{ij} = (N/2)x_{ij}$, the remaining constraints are $\sum_{j \neq i} y_{ij} = 1$ on the complement graph. This is the standard fractional-perfect-matching feasibility condition. It concerns only the clean Q-only problem; H can shrink the feasible face.

A.3 Restricted plan universes

For a restricted plan universe, let M_w record whether a plan misses each w-witness set and let $c_w = \binom{N-w}{B}/\binom{N}{B}$. The complete-design bound is attainable exactly when the following system is feasible:

$$M_w \lambda = c_w \mathbf{1}, \qquad \lambda \ge 0, \qquad \mathbf{1}^\top \lambda = 1. \tag{13}$$

Proposition 4 (Restricted-plan attainability)**.** *Let every $S \in \Pi \subseteq \binom{Q}{B}$ have size B. The complete-universe value in Equation* (10) *is an upper bound on $\kappa^\star_{\mathcal{F}_w}$ over Π. It is attained if and only if there is a policy satisfying Equation* (13)*.*

Proof. A uniformly chosen w-set is hit by every fixed size-B plan with probability $1 - c_w$. Averaging first over the plan sampled by any policy gives an average detection probability of $1 - c_w$, hence a minimum no larger than that value. If $M_w\lambda = c_w\mathbf{1}$, every w-set is detected with probability $1 - c_w$, proving sufficiency. Conversely, if the upper bound is attained, every w-set must have detection at least its average $1 - c_w$. Equality of the average forces equality for every row, proving necessity. The saturated case has $c_w = 0$ and is included.

If the test is infeasible, solve the finite minimax LP directly. To report an optimal-set dimension, first find its maximal support $J = \{S : \max_{\lambda \in \Lambda^\star} \lambda_S > 0\}$ by one LP per plan. Compute the affine hull of the resulting optimal polytope on J (for example, from its equality closure or an affine basis of vertices), then report its dimension. The rank shortcut applies only when the equal-miss system is feasible with positive mass on every feasible plan.

A.4 Bounded-loss stronger assumptions

Theorem 5 (Bounded-loss stronger-assumption assurance)**.** *For a finite plan universe, finite baseline witness families, and finite optional family $\mathcal{G}$, the function $V_\mathcal{G}$ is nondecreasing, concave, continuous, and piecewise linear on $\delta \geq 0$.*

Corollary 6 (Joint ASRA on the bounded-loss frontier)**.** *For baseline Θ_0 and optional family $\mathcal{G}$,*

$$\rho^\star_{\Theta_0 \cup \{\mathcal{G}\}} = \min_{\delta \geq 0} \max\left\{\rho^\star_0 + \delta,\, 1 - V_\mathcal{G}(\delta)\right\}. \tag{14}$$

If Q belongs to the baseline collection, every policy feasible at loss budget δ satisfies

$$\kappa_Q(\lambda_\delta) \geq \kappa^\star_Q(1 - \rho^\star_0 - \delta). \tag{15}$$

Finite LP form. For each active baseline family $\theta \in \Theta_0$ and witness set $D \in \theta$, the constraint $\rho_0(\lambda) \leq \rho^\star_0 + \delta$ is equivalent to

$$\phi_\lambda(D) \geq \kappa^\star_\theta(1 - \rho^\star_0 - \delta).$$

Consequently $V_\mathcal{G}(\delta)$ is the optimum of the finite LP

$$\begin{aligned} \max_{\lambda, z} \quad & z \\ \text{s.t.} \quad & z \leq \phi_\lambda(D)/\kappa^\star_\mathcal{G} && (D \in \mathcal{G}), \\ & \phi_\lambda(D) \geq \kappa^\star_\theta(1 - \rho^\star_0 - \delta) && (\theta \in \Theta_0,\ D \in \theta), \\ & \lambda \geq 0, \qquad \mathbf{1}^\top\lambda = 1. \end{aligned} \tag{A.3}$$

Proof of Theorem 5. Monotonicity follows because increasing δ only relaxes the baseline constraints. Let λ_i be feasible at δ_i and set $\lambda_t = t\lambda_1 + (1-t)\lambda_2$, $\delta_t = t\delta_1 + (1-t)\delta_2$. Every baseline constraint in Equation (A.3) is linear, so λ_t is feasible at δ_t. Since $\phi_\lambda(D)$ is linear,

$$\kappa_\mathcal{G}(\lambda_t) = \min_{D \in \mathcal{G}} \phi_{\lambda_t}(D) \geq t\kappa_\mathcal{G}(\lambda_1) + (1-t)\kappa_\mathcal{G}(\lambda_2).$$

Taking optimal policies proves concavity. Equation (A.3) is a finite parametric LP with an affine right-hand side. Its value is piecewise linear and continuous on the feasible half-line (and constant once the baseline constraint is redundant).

Proof of Corollary 6. For any policy λ, set $\delta = \rho_0(\lambda) - \rho^\star_0 \geq 0$. Then $\kappa_\mathcal{G}(\lambda)/\kappa^\star_\mathcal{G} \leq V_\mathcal{G}(\delta)$, so its joint regret is at least the right-hand expression in Equation (14). Conversely, a policy attaining $V_\mathcal{G}(\delta)$ has baseline regret at most $\rho^\star_0 + \delta$ and stronger-family regret $1 - V_\mathcal{G}(\delta)$. Minimizing over δ proves equality.

Fallback proof. If Q is in Θ_0, its individual normalized regret is at most $\rho^\star_0 + \delta$. Rearranging gives Equation (15). For nonempty D_g, hitting any fixed $q \in D_g$ detects g, so $\phi_{\lambda_\delta}(D_g) \geq \min_{q \in Q} \Pr(q \in P) = \kappa_Q(\lambda_\delta)$.

A.5 SINGLE-MECHANISM SPECIALIZATION

Theorem 7 (Single-mechanism specialization under a Q floor)**.** *Let $M \subset Q$ have size K, let the complete fixed-B plan universe satisfy a Q marginal floor t, and assume $1 \le w \le K$. Put*

$$f_{K,w}(j) = 1 - \frac{\binom{K-w}{j}}{\binom{K}{j}}, \qquad c(t) = \min\{\min(B, K),\ B - t(N-K)\}.$$

If $\bar{f}_{K,w}$ is the chord interpolation of $f_{K,w}$ and the Q floor is feasible, the exact optimal mechanism assurance is

$$\Gamma_{K,w}(t) = \bar{f}_{K,w}(c(t)).$$

When $c(t)$ is nonintegral, the optimum randomizes between its adjacent integer values.

A.6 EXACT ASSURANCE REGION FOR DISJOINT MECHANISM POOLS

Theorem 8 (Disjoint block assurance region)**.** *Partition the frozen frontier as $Q = M_0 \dot\cup M_1 \dot\cup \cdots \dot\cup M_r$, with $K_m = |M_m|$. Every plan selects exactly B probes and satisfies block quotas $L_m \le |P \cap M_m| \le U_m$ and broad floors t_m. For mechanism $m \ge 1$, suppose at least w_m witnesses lie in M_m, and set*

$$f_m(j) = 1 - \frac{\binom{K_m - w_m}{j}}{\binom{K_m}{j}},$$

with $\bar{f}_m$ its chord interpolation. For a target assurance γ_m, let

$$r_m(\gamma_m) = \inf\{x : \bar{f}_m(x) \ge \gamma_m\}, \qquad a_m(\gamma_m) = \max\{L_m, K_m t_m, r_m(\gamma_m)\},$$

and let $a_0 = \max\{L_0, K_0 t_0\}$. The vector $(\gamma_1, \ldots, \gamma_r)$ is attainable if and only if

$$a_m(\gamma_m) \le U_m \quad \textit{for all } m, \qquad \sum_{m=0}^{r} a_m \le B \le \sum_{m=0}^{r} U_m.$$

Proof. Average any policy over permutations within each disjoint block. This preserves quotas, broad floors, and worst-case block assurance, and reduces the policy to the integer occupancy vector $J_m = |P \cap M_m|$. Conditional on $J_m = j$, sample the j probes uniformly inside M_m. A fixed w_m-witness subset is then hit with probability $f_m(j)$, and every probe in M_m has marginal $E[J_m]/K_m$.

For necessity, put $x_m = E[J_m]$. Exact budget and quotas give $\sum_m x_m = B$ and $L_m \le x_m \le U_m$; the broad floor gives $x_m \ge K_m t_m$. Since $\bar{f}_m$ is concave and agrees with f_m at integer arguments, Jensen's inequality gives $\gamma_m \le E[f_m(J_m)] \le \bar{f}_m(x_m)$, hence $x_m \ge r_m(\gamma_m)$. Therefore $x_m \ge a_m(\gamma_m)$ and the displayed conditions follow.

For sufficiency, the displayed conditions permit real numbers $x_m \in [a_m, U_m]$ with $\sum_m x_m = B$: begin at the lower bounds and distribute the remaining mass without exceeding the upper bounds. Consider

$$\mathcal{O}(x) = \{y : \lfloor x_m \rfloor \le y_m \le \lceil x_m \rceil,\ \sum_m y_m = B\}.$$

This polytope contains x and has only integral extreme points. Indeed, a vertex cannot have two coordinates strictly between their integer bounds, because one can move a small amount from one to the other; and it cannot have exactly one fractional coordinate because its coordinate sum is the integer B. Thus x is a convex combination of feasible integer occupancy vectors j, each satisfying $j_m \in \{\lfloor x_m \rfloor, \lceil x_m \rceil\}$, the quotas, and $\sum_m j_m = B$. Randomize according to this convex combination and, conditional on j, sample uniformly within each block. Then every block-m probe has marginal $x_m/K_m \ge t_m$, while $E[f_m(J_m)] = \bar{f}_m(x_m) \ge \gamma_m$. This realizes all targets simultaneously.

Overlapping-pool complexity. Define UNIT-OVERLAP-AUDIT: given a probe set Q, arbitrary pools $R_1, \ldots, R_s \subseteq Q$, and an integer $B \le |Q|$, decide whether an exactly-B plan intersects every pool. A proposed plan is checked in polynomial time, so the problem is in NP. For NP-hardness, take a HITTING SET instance with universe U, subsets $C_1, \ldots, C_s$, and budget $B \le |U|$; set $Q = U$

and $R_i = C_i$. A hitting set of size at most B can be padded to an exactly-B plan and intersects every R_i, while every feasible plan is itself a hitting set of size B. The construction is identity-sized and polynomial. A randomized policy has unit assurance only if every plan in its support intersects every pool, so randomization does not change feasibility. Hence unit-assurance feasibility for arbitrarily overlapping pools is NP-complete. Such overlap is natural: one probe can test both bullet-count and sentence-order requirements, or both CSV quoting and field-count requirements. This is the tractability boundary.

Discrete concavity. Write $q_j = \binom{K-w}{j}/\binom{K}{j}$ with the zero convention. Then $f_{K,w}(j) = 1 - q_j$ and, where $q_j > 0$,

$$f_{K,w}(j+1) - f_{K,w}(j) = q_j \frac{w}{K-j}.$$

The ratio of consecutive nonzero increments is

$$\frac{f_{K,w}(j+2) - f_{K,w}(j+1)}{f_{K,w}(j+1) - f_{K,w}(j)} = \frac{K-w-j}{K-j-1} \leq 1.$$

After q_j reaches zero, the increments remain zero. Thus $f_{K,w}$ is nondecreasing and discretely concave.

Proof of Theorem 7. Permuting probes within M and within $Q \setminus M$ preserves the Q floor and the mechanism witness family. Averaging a policy over these permutations cannot lower its worst mechanism assurance, so it suffices to consider a symmetric policy. Conditional on $J = j$, every j-subset of M is uniform, and the mechanism assurance is $f_{K,w}(j)$.

The outside probes have total marginal mass at least $t(N-K)$, hence $E[J] \leq B - t(N-K)$. Feasible plans also have

$$\max\{0, B-(N-K)\} \leq J \leq \min\{B, K\}.$$

Thus $E[J] \leq c(t)$. Discrete Jensen for the chord interpolation gives

$$E[f_{K,w}(J)] \leq \bar{f}_{K,w}(E[J]) \leq \bar{f}_{K,w}(c(t)).$$

For achievability, mix $\lfloor c(t) \rfloor$ and $\lceil c(t) \rceil$ to obtain mean $c(t)$, choose uniformly within M and within its complement conditional on J, and query exactly B probes. The inside marginal is $c(t)/K \geq t$ and the outside marginal is $(B - c(t))/(N-K) \geq t$; the endpoint cases follow by the same inequalities. This policy meets the Q floor and achieves the displayed bound.

Finite check. The completed frozen analysis mechanically compared the formula with the finite optimization on 30,095 parameter settings with $N \leq 14$ and found no discrepancy. This check supports the implementation; the preceding argument is the proof.

A.7 STABLE-H FINITE PROCEDURE AND Q FALLBACK

Proof of Proposition 1. The simplex over a finite plan universe is compact, and every $\rho_{-u}(\lambda)$ is the maximum of finitely many affine functions. The minimum defining $\epsilon_H^\star$ is therefore attained. The set in Equation (6) is a nonempty intersection of the simplex with finitely many closed halfspaces, hence a compact convex polytope.

For each stable policy, Q is one term of every reduced H/Q objective. Thus $\rho_Q(\lambda) \leq \rho_{-u}^\star + \epsilon_H^\star$ for every u. Choosing the smallest reduced optimum and rearranging gives Equation (7). Every nonempty witness set contains a probe whose inclusion detects it, so its detection probability is at least the same Q fallback. Identical hit rows induce identical constraints; deleting duplicate raw fault identifiers therefore changes neither the optimization nor the stable set. All statements hold for any finite feasible plan universe and fixed budget.

A.8 SEMANTIC-ORACLE UNCERTAINTY

The guarantees above treat the semantic verdict A as fixed. A limited post-design noise calculation helps mark, but does not remove, this boundary. Let A^* be the true binary verdict, let $D_g^* = \{q :$

$g(q) \neq A^*(q)\}$, and suppose a policy λ is fixed before corruption. Hypothetically, let each queried verdict then be independently flipped relative to A^* with probability η. Conditional on a selected plan hitting $r \geq 1$ true witnesses, at least one of those witnesses remains correctly labeled with probability $1 - \eta^r \geq 1 - \eta$. Averaging over plans therefore gives

$$\Pr(\text{detect a correctly labeled true witness}) \geq (1 - \eta)\, \phi_\lambda(D_g^*).$$

Here η is only a hypothetical perturbation parameter; it is not estimated empirically in this work.

This calculation does not imply that the full design procedure degrades only by a factor $1-\eta$ when a perturbed oracle is used during design. Such a perturbation can change D_g, modeled detection rows, retained H structure, $\epsilon_H^\star$, the pools $Q_\tau = \{q : A(q) \neq \tau(A)(q)\}$, S support, and the optimized policy itself. Without additional margin or stability assumptions, robustness to design-time semantic-oracle misspecification is outside the present guarantees.

A.9 Preregistered local mechanisms and frozen-pool evaluation

For each frozen task, a preregistered library supplies deterministic one-feature perturbations τ. The contract feature and semantic mechanism label are chosen without inspecting Q_τ, hidden faults, or downstream detection. After semantic and audit-row deduplication, each retained nondegenerate pool is $Q_\tau = \{q \in Q : A(q) \neq \tau(A)(q)\}$. The primary S condition is $m_g = \max_\tau |D_g \cap Q_\tau| \geq 2$.

The evaluation uses 30 frozen development tasks and 164 hidden faults with nonempty frontier signatures. A joint frontier permutation applied to every pool supplies the geometry-matched control. Same-family-held-out analyses remove all pools from the hidden fault's known H mutation/operator family. Frozen H-operator attribution is available for 126 faults; same-family removal is undefined for the 38 UNKNOWN-family faults, which are excluded from the formal interpretation without imputation. The available frozen records do not support a more specific provenance claim for those labels. Terminal detection is computed exactly as $p_\lambda(g) = \Pr_{S\sim\lambda}[S \cap D_g \neq \varnothing]$. The cohort supports mechanism diagnostics and internal falsification. Independent confirmation remains necessary to establish generalization beyond the frozen development benchmark.

B Notation and system labels

Table 1: **Core notation and empirical system labels.**

Symbol or label	Meaning
Q, P, B	Candidate frontier probes, one feasible audit plan, and its fixed query budget.
$D_g(Q)$, $\phi_\lambda(D)$	Witness set for fault g and probability that policy λ detects at least one member of D.
OLD	Original H/Q-guided development policy.
STABLE-H	Policy constrained by one-pattern-removal H stability.
SR-STABLE-V1	Frozen procedure combining H stability and S specialization in the development and QuixBugs comparisons.
NO-S	Corresponding H-stable baseline without an S objective in the independent-S benchmark.
CONCRETE-S	S objective using individual S-train exposure signatures.
MECHANISM-CORE-S	S objective using intersections within named semantic mechanisms.
ORACLE-S	Non-deployable upper bound using held-out S information.

B.1 Construction burden and exported certificate

Table 2: **Inputs required by each assumption layer.** Authoring time was not measured.

Layer	Source	Human versus automatic	Recorded artifact
Q	contract-derived frontier	mixed: domain design plus programmatic enumeration	probes and feasible plans
A	semantic verdict	task-specific human or external oracle, then executable	one verdict per probe
H	executable mutation library	human-authored operators; automatic execution, signatures, and deduplication	raw faults → distinct rows
S	mechanism specification	human-authored ontology; automatic splitting, signatures, and intersections	mechanisms, S-train signatures, and cores

S mechanism specification is a human input in the current implementation. The strict core is activated only when a nondegenerate intersection remains; otherwise the exported policy is the H-stable baseline.

B.2 Certificate record and archive boundary

The V2 evidence memo records three worked certificates: an ordinary R2 task, an actionable-H task with an empty-witness modeled row, and an R2 task with a valid S core. Each underlying certificate carries the frontier, plan universe, distinct H rows and H mutation/operator families, S mechanisms, $\kappa_Q, \kappa_H, \kappa_S, \epsilon_H^\star, \delta$, active constraints, the complete policy support, a public seed, and one realized plan.

The fields in the preceding paragraph are design-time certificate quantities: they are computable before held-out faults are revealed and state guarantees under the declared Q/H/S assumptions. The attached manuscript-facing record also supplies a held-out evaluation excerpt for `detectable_format:title`: after the evaluation faults are revealed, the policy's minimum detection probability across those witness sets is approximately .204, its mean is approximately .435, and the public-seed sampled plan detected one of three faults. These are evaluation diagnostics, not ex-ante guarantees for arbitrary future faults. The probabilities attach to the policy; the realized plan is one Bernoulli outcome, not an interval. For r independent draws, a fault with one-draw detection probability p is detected at least once with probability $1 - (1 - p)^r$.

The detailed certificate JSON and case-study file named by the V2 memo are not present among the attached manuscript files in this workspace. Consequently, frontier size, budget, support probabilities, the realized probe identities, and other task-specific fields for this example are not reconstructed here. The predecessor row-41 record remains separately available: $|Q| = 19$, $B = 2$, 171 feasible pairs, six preregistered candidate faults, two retained H rows, $\kappa_Q^\star = 2/19$, $\kappa_H^\star = 1$, and historical deployed capacities $\kappa_Q = .054422$, $\kappa_H = .517007$. It is not relabeled as the current V2 certificate.

C Empirical Protocols and Complete Supplied Summaries

C.1 Development-cohort provenance

Three H-related definitions appear in the empirical record. R1, the registered H-active rule, holds on 29/30 development tasks. R2, the operational full-H-capacity gate $k_H > 10^{-9}$, selects 19/30 tasks. Under that original full-capacity definition, 11 tasks have no usable joint H capacity, none has usable H with $\epsilon_H^\star = 0$, and all 19 R2 tasks have $\epsilon_H^\star > 0$. The historical 27/30 positive-slack count is the later actionable-H diagnostic described below, not the original full-capacity domain.

Exactly ten tasks satisfy R1 but fail R2. Every one has at least one retained modeled fault with a nonempty frozen-frontier witness set and at least one retained modeled fault with an empty witness set. The latter is a zero row in the complete H detection matrix, forcing $k_H = 0$. If empty-witness modeled faults are removed, the nonzero-row capacity is $k_{H,\mathrm{nz}} = 1.0$ on all ten. Thus R1 asks whether any usable H signal exists, whereas R2 requires the entire retained H collection to have nonzero joint capacity. No downstream outcome was used to choose R2, but the two domains are not equivalent.

Table 3: **Development sensitivity at $B = 2$.** The R1 SR contrasts are non-estimable under the frozen procedure.

Rule	Tasks	STABLE-H−OLD	SR-STABLE-V1−STABLE-H	SR-STABLE-V1−OLD
R1: registered H-active	29	$+.0046$ $[-.0066, +.0133]$; $19/1/9$; $p \simeq .44$	non-estimable	non-estimable
R2: full-H-capacity	19	$+.0053$ $[+.0002, +.0108]$; $12/0/7$; $p = .070$	$+.0105$ $[-.0009, +.0218]$; $13/0/6$; $p = .097$	$+.0158$ $[+.0024, +.0295]$; $12/0/7$; $p \simeq .041$
R3: actionable $\epsilon^\star_H > 0$	27	exploratory	exploratory	exploratory

On R1, the STABLE-H–OLD median is approximately $+.0040$ and the minimum leave-one-task-out mean is approximately $+.0028$. The positive average is attenuated and substantially more uncertain than on R2. The comparison is also heterogeneous: OLD combines H and Q on the 19 R2 tasks but collapses to Q on the ten R1-only tasks because $k_H = 0$. The frozen SR procedure was instantiated only on the nonzero-H-capacity domain and therefore supplies no policy for those ten tasks. Its R1 contrasts are non-estimable under the frozen procedure; this is a domain limitation, not a measured null effect. The development $B = 4$ aggregate remains incomplete and is not reconstructed.

C.2 Actionable-H prospective method extension

The actionable-H extension changes the operational objective by removing only H rows with no witness on the frozen frontier. This produces a defined policy for all 29 R1 tasks. The experiment is classified as `PROSPECTIVE_METHOD_EXTENSION_ON_EXISTING_DEVELOPMENT_DATA`: the method definition is prospective, but the evaluation domain is the existing development set rather than an independent confirmation cohort.

Table 4: **Actionable-H effects on the 29-task R1 domain at $B = 2$.** These contrasts use the extended method definition.

Contrast	Mean difference	Primary interpretation
SR-STABLE-V1−OLD	$+.0113$	SR direction remains positive
SR-STABLE-V1−STABLE-H	$+.0108$	positive added S-stage direction
STABLE-H−OLD	$+.0005$	STABLE-H benefit attenuates

The extension shows that a policy can be defined over the full R1 domain under a principled witnessed-row interpretation. It does not retroactively define the frozen SR policy on the ten R1-only tasks, and its $p = .044$ calculation is not treated as new confirmatory significance.

C.3 Design-time conflict diagnosis and certificate increments

The sign of $\epsilon^\star_H$ separates three operational cases before any held-out fault is revealed. A task with no positive joint H capacity remains at Q. If H is usable and $\epsilon^\star_H = 0$, all leave-one reduced problems share an exact optimum, so no positive common-relaxation slack is required. If $\epsilon^\star_H > 0$, their optimal requirements conflict and STABLE-H solves a genuine reconciliation problem.

Table 5: **Design-time STABLE-H diagnosis at $B = 2$.** The actionable-H row is a sensitivity on the existing development tasks.

H definition	Q only	Usable H, $\epsilon^\star_H = 0$	Usable H, $\epsilon^\star_H > 0$
Original full capacity	11	0	19
Actionable-H	1	2	27

For the two actionable-H zero-slack cases, the observed difference between STABLE-H and the unstabilized Q/H policy in κ_Q is at most approximately 1.2×10^{-8}. Across 319 R2 partial-H cells

with $\epsilon_H^\star = 0$, deletion exposure is approximately 4.3×10^{-9} and the corresponding κ_Q change is numerically zero. These checks support the no-positive-slack interpretation; they do not imply that a particular tie-broken STABLE-H policy must always equal OLD.

On the frozen 19-task R2 cohort, moving from Q to the ordinary Q/H policy raises worst-case modeled-fault assurance by approximately .298, with all 19/19 tasks moving in the same direction. Moving from the ordinary Q/H policy to STABLE-H reduces leave-one exposure by approximately .047. The comparison with AVG-REDUCED in Appendix C.5 isolates the additional value of the final common solve.

C.4 Budget-matched development baselines

All six strategies use the same frontier, semantic oracle, hidden faults, and $B = 2$ budget on the 19 R2 tasks.

Table 6: **Task-balanced hidden-fault detection for simple and structured development strategies.**

Method	Information	Mean	Approx. 95% interval
Uniform feasible-plan sampling	none	.5416	[.4813, .6027]
Q-maximin	Q	.5416	[.4811, .6029]
Greedy H coverage	H	.4818	[.3698, .5894]
OLD	H,Q	.5350	[.4702, .6007]
STABLE-H	H,Q	.5404	[.4762, .6063]
SR-STABLE-V1	H,Q,S	.5508	[.4853, .6203]

Against Uniform, the task-balanced SR-STABLE-V1 difference is $+.0093$ $(10/0/9)$. Against Q-maximin it is also $+.0093$ $(10/0/9)$, and against STABLE-H it is $+.0105$ $(13/0/6)$. Relative to Uniform, Greedy-H is approximately $-.0598$ $(7/0/12)$, OLD is approximately $-.0065$, and STABLE-H is approximately $-.0012$. These comparisons rule out a trivial explanation by naive H concentration, but do not establish blanket superiority over all baselines.

C.5 Simple controls for the STABLE-H construction

The Q-matched mixture asks whether STABLE-H helps only because it retreats toward broad Q coverage. For each task,

$$\lambda_{\text{mix}} = (1 - \alpha)\lambda_{\text{OLD}} + \alpha\lambda_Q, \tag{16}$$

where α is the smallest design-time value for which the mixture attains at least STABLE-H's Q assurance. No hidden-fault outcome enters this choice, and the Q requirement is verified on all 19 tasks. Its mean held-out detection at $B = 2$ is .5333.

AVG-REDUCED averages the outcome-blind optimum policies from the distinct leave-one-row reduced problems. DIVERSITY is a parameter-free canonical-row H-diversity heuristic. The reported I_H is a deletion-instability score, so smaller values indicate a stronger stability certificate.

Table 7: **Simple controls on the frozen 19-task cohort at $B = 2$.** Paired effects are STABLE-H minus the named control. Dashes mark assurance values not supplied in the baseline report.

Control	Paired detection	95% interval	κ_H	I_H
Q-matched OLD/Q mix	$+.0071$	$[+.0016, +.0133]$	—	—
AVG-REDUCED	$+.0018$	$[-.0040, +.0088]$	.482	.185
DIVERSITY	$-.0237$	$[-.0426, -.0047]$	.353	.328
STABLE-H reference	—	—	.518	.129

The Q-matched comparison shows that STABLE-H is doing more than adding broad coverage back to OLD. AVG-REDUCED nearly matches its average held-out detection, suggesting that much of the empirical effect comes from deletion itself; STABLE-H nevertheless supplies the stronger common minimax certificate. Define $\Delta I_t = I_H(\lambda_{\text{AVG,t}}) - I_H(\lambda_{\text{STABLE,t}})$ and $\Delta\kappa_{H,t} =$

$\kappa_H(\lambda_{\mathrm{STABLE,t}}) - \kappa_H(\lambda_{\mathrm{AVG,t}})$. Both are positive on 19/19 tasks. ΔI_t has mean .05557, median .04902, and leave-one-task-out mean range $[.0514, .0570]$; $\Delta\kappa_{H,t}$ has mean .03650, median .03386, and leave-one-task-out range $[.0313, .0385]$. This shows that the common certificate advantage is broad on the frozen cohort, but it is descriptive rather than independent generalization evidence because STABLE-H optimizes deletion stability. DIVERSITY achieves higher average held-out detection while accepting substantially lower worst-case modeled-fault assurance and worse deletion stability. These controls compare different allocation contracts rather than identify a single best detector.

C.6 CONTROLLED STABLE-H VALIDATION

The preregistered controlled experiment contains 13,500 synthetic tasks. The hidden generator does not read Q, OLD, or STABLE policies. It varies whether a singled-out modeled H pattern represents the hidden faults. STABLE-H tends to help more when the pattern is unrepresentative; its advantage shrinks as the pattern becomes more representative, and all 27 registered endpoint trajectories decrease from $\pi = 0$ to $\pi = 1$. This validates the intended misspecification response, not the realism or production prevalence of the synthetic fault generator.

C.7 DEVELOPMENT SEMANTIC-LOCALIZATION DIAGNOSTICS

For each development task, a preregistered library supplies deterministic one-feature perturbations τ. After semantic and audit-row deduplication, each retained pool is $Q_\tau = \{q \in Q : A(q) \neq \tau(A)(q)\}$. The pool construction, local-support criterion, joint frontier-permutation control, and pooled comparison precede the later equal-task reaggregation.

The evaluation contains 30 development tasks and 164 hidden faults with nonempty frontier signatures. Registered local support occurs for 116/164 (70.7%) fault rows, compared with 57.4% under the geometry-matched control; the corresponding pooled downstream shifts are $+.0533$ and $+.0004$. Task-balanced deletion and bootstrap summaries are post-hoc descriptive sensitivities, not independent confirmation.

Same-family holdout removes all pools belonging to the hidden fault's known H mutation/operator family. Frozen H-operator attribution exists for 126 faults. It is undefined for the 38 UNKNOWN-family rows, so those rows are excluded from the formal same-family interpretation without imputation. This diagnostic tests direct family reuse inside the development benchmark; it is not the independent H/S construction used below.

C.8 ORIGINAL NARROW INDEPENDENT-S BENCHMARK

The independent-S cohort contains seven evaluator tasks. H is derived from generic executable or surface perturbations, including comparison, threshold, and weight changes. S is independently authored from semantic-mechanism information and does not reuse H labels or mutation families. S-train instances construct the Concrete-S signatures and Mechanism-Core-S intersections; held-out S_{test} faults are used only for evaluation. The systems share H geometry, stability constraints, solver, and evaluation protocol.

Table 8: **Independent-S systems.** Semantic names replace opaque implementation codes.

System	S information	Role
NO-S	none	Matched H-stable baseline
CONCRETE-S	individual S-train signatures	Deployable concrete-signature diagnostic
MECHANISM-CORE-S	intersections within named mechanisms	Primary deployable S system
ORACLE-S	held-out S information	Non-deployable assurance upper bound

Table 9: **Held-out independent-S contrasts.** The primary contrast is MECHANISM-CORE-S–NO-S.

Contrast	Mean	95% interval	Sign-flip p	Holm p	W/T/L
MECHANISM-CORE-S–NO-S	$+.0293$	$[+.020, +.038]$	.008	.016	$7/0/0$
MECHANISM-CORE-S–CONCRETE-S	$+.007$	—	.25	—	$2/5/0$
ORACLE-S–MECHANISM-CORE-S	$+.075$	—	.008	—	—

The primary comparison supports additional held-out detection from an independently constructed mechanism-level S objective on this cohort. The Concrete-S contrast is directionally positive but inconclusive. A preceding concrete-signature experiment produced an approximately $+.015$ improvement with $p = .11$; it is a supporting diagnostic rather than the headline result. The Oracle-S gap shows partial transfer and unresolved headroom, not a deployable alternative.

Table 10: **Independent-S 2×2 attribution diagnostic.** Entries are mean held-out detection.

H condition	S off	S on
H unstable	OLD $= .6831$	OLD+S $= .7331$
H stable	STABLE-H $= .6948$	SR-STABLE $= .7241$

The S marginal is $+.0500$ when H is unstable and $+.0293$ when H is stable. The H-stability marginal is $+.0117$ with S off and $-.0090$ with S on, for an interaction of approximately $-.0206$. This is an attribution diagnostic: S is positive in both H conditions, whereas the H-stability marginal is smaller and context-dependent on this mechanism benchmark. It neither makes H stability unnecessary nor overturns the controlled STABLE-H validation.

The cohort is small, the main comparison has seven task-level units, and some judge specifications are near-duplicates. Effective task diversity is therefore lower than the raw number of generated cases suggests. Mechanism cores often contain only two probes. These limitations make the experiment a targeted identification test rather than a large-scale population benchmark.

C.9 PROSPECTIVE EXPANDED INDEPENDENT-S BENCHMARK

The expanded benchmark contains 26 tasks across seven evaluator families: clause-sum, threshold-gate, order-monotonic, set-membership, multi-condition, aggregation-score, and boundary-interval. The primary comparison remains Mechanism-Core-S minus the matched No-S policy. Its mean is $-.0041$, with an approximate 95% interval $[-.0247, +.0161]$, $p \approx .70$, and task directions $12/1/13$.

Table 11: **Selected family means from the expanded independent-S benchmark.** The supplied summary gives exact displayed values for these five families; the main aggregate uses all seven.

Evaluator family	Mechanism-Core-S − No-S
`clause_sum`	$\approx +.027$
`order_monotonic`	$\approx +.022$
`boundary_interval`	$\approx -.072$
`aggregation_score`	$\approx -.034$
`multi_condition`	$\approx -.035$

The expanded result changes the population-level interpretation of the narrow study. Independent semantic information can improve allocation in some families, but the same specialization can be neutral or harmful when the mechanism core is misaligned with held-out witnesses. The old seven-task effect remains a valid narrow-cohort result; it is not used as the general S claim.

C.10 EXTERNAL QUIXBUGS COHORT

Table 12: **QuixBugs external procedure-level comparison.**

Contrast	Mean	95% interval	Positive/tied/negative	Exact p
SR-STABLE-V1−STABLE-H	$+.0208$	$[+.0092, +.0333]$	$8/5/0$	.0078

The 13-task median difference is $+.0081$, and the minimum leave-one-task-out mean remains positive. QuixBugs is disjoint from the development task family and therefore supplies external generalization of the frozen procedure. It is not an independent S validation: its S pools reuse the generic H mutation families NUM$\pm$1, COMP2, ARITH, and BOOL.

C.11 Certified scaling for core H capacity and full STABLE-H

Core H capacity. The simpler core maximin uses a restricted master over generated plans and an exact weighted-coverage pricing problem. Pricing is exhaustive for $B = 2$ and a maximum-coverage MILP for larger budgets. On exact-small instances it recovers k_H to machine precision and $k_Q = B/|Q|$; all 18 tested cells are certified. The largest reaches $|Q| = 100, B = 10$ without materializing $\binom{100}{10} \approx 1.73 \times 10^{13}$ plans. This result establishes the pricing primitive but is distinct from full deletion stability.

Exact decomposition for STABLE-H. Every leave-one scenario changes its required capacity and regret level, but all scenarios use the same probe- and H-hit events. The solver therefore shares one generated plan pool across the reduced problems and the final common-stability master.

For deletion scenario u, let i index the surviving H rows. Its reduced H capacity is obtained from the exact dual cutting-plane problem

$$h_u = \min_{y,\mu} \mu, \qquad y \geq 0, \quad \sum_i y_i = 1, \tag{17}$$

subject, for every feasible plan $P \in \Pi$, to

$$\mu \geq \sum_i y_i \, \mathbf{1}[P \cap H_i \neq \varnothing]. \tag{18}$$

Separating a violated plan in Equation (18) is exactly weighted maximum coverage. The same shared columns compute the reduced Q/H regret optimum $r_u = \rho^\star_{-u}$ for every scenario. The target remains $\epsilon^\star_H = \min_\lambda \max_u [\rho_{-u}(\lambda) - r_u]$.

The final restricted master minimizes the common excess ϵ. For each scenario u and surviving H row H_j, it retains the original constraint

$$\phi_\lambda(H_j) + h_u \epsilon \geq h_u(1 - r_u). \tag{19}$$

All original H and Q deletion constraints are checked and violated rows are added. After aggregating master dual weights on the shared H- and Q-hit events, plan pricing again reduces to exact weighted maximum coverage. Certification terminates only when no original deletion constraint is violated and exact pricing proves that no improving plan exists. This is an exact decomposition of the original STABLE-H objective, not a relaxation or a new audit rule.

Proof of Proposition 2. No violated original row makes the restricted solution primal feasible for the full problem, with zero mass on ungenerated plans. Exact pricing finding no improving plan makes the restricted-master dual feasible for every omitted plan column and hence for the full dual. The restricted primal and dual have the same objective value; finite-LP strong duality therefore certifies that this value is the optimum of the original STABLE-H problem.

Dense equivalence. On the frozen 19-task R2 cohort at $B = 2$, the decomposition matches dense STABLE-H on 19/19 tasks: the maximum absolute differences are 2.25×10^{-15} for $\epsilon^\star_H$, 1.11×10^{-16} for h_u, and 3.55×10^{-15} for r_u. Generated policies are nonnegative, sum to one to machine precision, and satisfy every original deletion constraint. No frozen detection result or maximum-entropy policy is changed by this computational validation.

A deterministic dense search selected and froze the first five small conflicting H motifs in a fixed order, each with $0 < h_u < 1$ in every deletion regime and $\epsilon^\star_H \approx .095$–$.120$, and lifted them to larger frontiers. Where dense computation remained feasible, decomposition and dense agreed on 24/24 comparisons with zero disagreements.

Table 13: **Exactly certified nontrivial full-STABLE-H cells.** Every row has $0 < h_u < 1$, $\epsilon^\star_H > 0$, complete original-constraint verification, and a closed exact-pricing gap.

$\|Q\|$	B	Distinct H	Implicit plans	$\epsilon_H^\star$	Runtime	Peak memory
27	6	24	296,010	.039431	11.2 s	2.0 MB
32	6	24	906,192	.027815	8.7 s	2.0 MB
35	6	24	1,623,160	.038882	10.0 s	2.5 MB
38	6	30	2,760,681	.031479	23.0 s	4.2 MB
44	8	32	177,232,627	.028852	22.6 s	6.0 MB

The large-lift study attempted 17 cells in total. Five positive- $\epsilon_H^\star$ cells closed exact certificates and appear above. Six had $\epsilon_H^\star = 0$ and therefore did not test conflicting reduced optima. The remaining six had positive $\epsilon_H^\star$ but did not close the exact certificate; their retained gaps were approximately 1.2–1.7. Tolerances were not loosened to turn those failures into successes.

For the same full-STABLE-H computation, dense enumeration solves 142,506 plans in about 391 s with 1.8 GB peak memory. A 2,118,760-plan dense case does not complete after 780 s and reaches approximately 9.3 GB. Unclosed cells are reported as failures rather than approximate certificates. Thus the result demonstrates substantial exact capability, not convergence on every large instance. It does not certify the S or bounded-loss composition at these scales.

D V2 Sensitivity, Engineering, and Construction Records

D.1 Budget dependence

The budget analysis is a method extension on existing data, not a collection of independent cohorts. Every budget uses the same tasks and hidden faults. The frozen $B = 2$ values remain the manuscript-facing values for the original experiments. The V2 values below are used only for the budget-sensitivity analysis.

Table 14: **Paired mean held-out detection differences across budgets.** Cohort A is the frozen 19-task R2 development cohort; cohort C is the 26-task expanded Independent-S benchmark.

Contrast	$B = 1$	2	3	4	5	8	10
A: STABLE−OLD	.0013	.0053	.0027	.0026	.0015	.0003	.0001
A: semantic−STABLE	.0025	.0105	.0221	.0244	.0205	.0050	.0022
A: semantic−OLD	.0038	.0158	.0248	.0271	.0220	.0053	.0023
B: STABLE−OLD	.0070	.0005	.0020	.0025	.0014	.0003	.0001
B: semantic−OLD	.0087	.0113	.0240	.0257	.0212	.0071	.0019
C: Core-S−No-S	.0085	−.0033	−.0022	−.0051	−.0109	—	—

On cohort A, the Q-floor binding fraction rises from approximately .11 to .37 over the primary budget range. The structured policies remain concentrated, but the available audit can cover more of the frontier and their incremental detection advantage approaches zero at the largest tested budgets. On cohort C, the semantic effect is non-positive on average for $B \geq 2$ under the V2 tie-break. The original frozen confirmatory value at $B = 2$ remains $-.0041$, not $-.0033$.

D.2 H-information budget and geometry-aware selection

This frozen-data sensitivity uses the 19-task R2 cohort at $B = 2$ and the existing 29-task actionable-H extension. It retains 25, 50, 75, or 100% of the pre-dedup raw H library. The random baseline averages 20 fixed, outcome-independent sublibraries per task and fraction. Evaluation is against the complete H library: we report its worst-row H coverage normalized by the full-library k_H, and excess over the full-library STABLE-H certificate. Standalone partial-library k_H is not an information measure because a smaller library can remove competing rows and raise its own maximin value.

ROW-DIVERSE uses only design-time hit signatures. For raw H rows $R_i, R_j \subseteq Q$, let

$$d_J(R_i, R_j) = 1 - \frac{|R_i \cap R_j|}{|R_i \cup R_j|}.$$

The deterministic farthest-first rule starts with the row having greatest average Jaccard distance from the rest, then repeatedly adds the candidate maximizing its minimum distance to the retained set;

ties follow the frozen row order. No hidden outcome enters selection. Intuitively, duplicate or near-duplicate faults receive little extra budget.

On R2, random 25/50/75/100% libraries recover normalized full-library worst-row coverage $.292/.342/.494/.661$, while their excess certificate gap falls from .310 to .277, .145, and .000. The table compares ROW-DIVERSE with the random mean at the same raw H budget; dashes mark quantities not supplied.

Table 15: **Certificate efficiency of complementary H selection.** Lower gap and higher full-library worst-row coverage are better. The final column reports paired p for gap/coverage.

Domain	H	Gap random	Gap diverse	Coverage random	Coverage diverse	$p_{\text{gap}}/p_{\text{cov}}$
R2	25%	.3102	.2608	.2920	—	.065/—
R2	50%	.2767	.1400	.3424	.5256	.0002/.0002
R2	75%	.1452	.0414	.4936	.6101	.0006/.002
Actionable-H	25%	.2770	.2405	—	—	—/—
Actionable-H	50%	.2398	.1180	.3881	.5364	$< 10^{-4}/ < 10^{-4}$
Actionable-H	75%	.1163	.0358	.5323	—	.0001/—

On R2, ROW-DIVERSE's certificate gap is better than roughly 92, 88, and 80% of the 20 random draws at 25, 50, and 75% H. Most memorably, 50% diverse H roughly matches random 75% H: gap .1400 versus .1452 and coverage .5256 versus .4936. The actionable-H sensitivity has the same ordering: gap .1180 versus .1163 and coverage .5364 versus .5323. This is an operational-domain sensitivity on existing data, not an independent cohort.

A design-time-only robustness check removes every task–budget cell where the ROW-DIVERSE subset appears among the random draws. At 50% H, the remaining R2 cells improve gap by $-.2224$ and coverage by $+.2979$ (both $p = .016$); actionable-H improves them by $-.2236$ and $+.3003$ (both $p = .008$). Within-task operator labels are singleton, so the frozen labels do not support a nontrivial balancing comparison. These results concern certificate recovery, not held-out detection or human authoring time.

D.3 Stronger H deletion sets

The primary STABLE-H method removes one distinct modeled row at a time. Leave-k and H mutation/operator-family removal are sensitivity extensions. Groups are defined by pre-existing H operator provenance, never by hidden outcomes.

Table 16: **Synthetic hidden detection under stronger H misspecification.**

Regime	OLD	STABLE-1	STABLE-2	STABLE-3	group removal
clean	.3720	.3777	.3832	.3964	.3811
one bad pattern	.3198	.3345	.3445	.3541	.3332
two correlated bad patterns	.3180	.3219	.3350	.3350	.3336
bad operator family	.3204	.3243	.3368	.3418	.3356
partial overlap	.3383	.3444	.3544	.3604	.3510

For the two-correlated-pattern regime, leave-one recovers about 23% of the gain achieved by leave-two/three. In the bad-family regime it recovers about 22% of the leave-three gain. On the 19-task real cohort, $\epsilon_H^{(1)}, \epsilon_H^{(2)}, \epsilon_H^{(3)}$ are approximately $.118, .164, .194$. Stronger protection has a nominal cost: leave-three reduces κ_H from about .503 to .408, and supplied real-task detection can fall below OLD. Moreover, every distinct real-task H row is its own H mutation/operator family, so group removal is identical to leave-one there.

D.4 Bounded-loss frontier endpoints

The V2 study evaluates the declared grid $\delta \in \{0, .01, .025, .05, .10, .20\}$ and reports the full frontier; no operating point is selected from hidden detection. All five studied tasks are still improving in structured assurance at $\delta = .20$, so saturation is not claimed. The attached manuscript-facing summary supplies the following endpoint examples.

Table 17: **Contrasting real-task endpoint changes from $\delta = 0$ to .20.** Dashes denote values not stated in the attached summary, not zero change.

Task	κ_Q	κ_H	hidden detection
`words:prime_lengths`	$.111 \to .078$	$.667 \to .467$	$.417 \to .422$
`bi_int5`	—	—	$.694 \to .953$

These tasks illustrate why δ is a risk declaration rather than a benchmark-tuned hyperparameter. A comparable willingness to spend fallback can buy almost nothing or a large downstream gain, depending on alignment.

D.5 Exploratory S-core representations

The expanded benchmark contains 26 tasks from seven evaluator/archetype families and 109 task–semantic-mechanism instances. Ninety-two strict cores are retained; 17 mechanism instances yield no retained core. Among the 92, 17 have size two, 75 have size greater than two, and none is a singleton. Every task retains at least one core.

Table 18: **Exploratory S-core sensitivity on the expanded benchmark.** All representations are constructed from S-train only.

Representation	Mean Core-S$-$No-S	Median	Negative tasks
strict intersection	$-.0033$	$+.0015$	$12/26$
75% frequency core	$-.0031$	$+.0015$	$12/26$
50% frequency core	$+.0094$	$+.0104$	$10/26$
weighted core	$-.0003$	$+.0032$	$11/26$

Softening the representation can improve the average, but it does not remove negative transfer. The 50%-frequency rule is not selected as a preferred method. The result points to alignment between semantic grouping and hidden witness geometry, not merely empty or singleton intersections.

D.6 Pipeline runtime and dense-plan boundary

On the existing real tasks at $B = 2$, the complete pipeline has median runtime about .18 seconds per task and maximum about 3.4 seconds in the supplied profile. Selected moderate synthetic cells are shown below.

Table 19: **Selected complete-pipeline profiles.**

Cell	Runtime	Peak memory
$\lvert Q\rvert = 20, B = 5$	63.6 s	266 MB
$\lvert Q\rvert = 50, B = 3$	71.7 s	537 MB
$\lvert Q\rvert = 30, B = 5$	dense plan-universe cap exceeded	
$\lvert Q\rvert = 50, B = 5$	dense plan-universe cap exceeded	

The dense cap is reported rather than bypassed. These complete-pipeline profiles complement the exact full-STABLE-H decomposition above; they do not certify the S or bounded-loss stages at the same scale.

D.7 Minimal stochastic proof of concept

The exploratory independent-repeat model reduces to the deterministic witness-set problem when per-probe error probabilities are zero or one. When the call budget is smaller than $|Q|$, broad coverage dominates repetition in the supplied toy calculation. Repetition becomes useful after coverage saturates: at $C = 80 > |Q| = 40$, the selected allocation has two repeats over 40 frontier locations. Aggregation is itself part of the model. This is an extension sketch, not validation on stochastic or LLM-based judges.

D.8 NUMERICAL TIE-BREAK PROVENANCE

The prespecified mathematical tie-break is maximum entropy over the selected policy face. The planned SLSQP implementation under-converged and became row-order sensitive at larger plan universes (entropy 3.193970791 versus 5.039650399 for the exact optimum at 462 plans). V2 therefore uses an exact dual active-set implementation of the same objective. At $B = 2$, development contrasts move by less than 10^{-4}; the expanded Independent-S contrast moves by about $+.0008$, from the frozen $-.00412$ to $-.0033$.

Two aborted pre-amendment runs read some V2 hidden outcomes in memory before writing no results. The filesystem therefore does not establish that the implementation amendment preceded the first hidden-outcome read. No such frozen-implementation claim is made; tie-break-dependent V2 analyses are treated as engineering implementation sensitivities. Frozen V1 results are unchanged and reproduce under their frozen path.

D.9 CONSTRUCTION BURDEN

The audit automates execution, signature computation, deduplication, train/test splitting, and optimization. Its modeling content remains human-authored: task-specific Q/oracle definitions, H mutation/operator libraries, and S mechanism ontologies or admissibility rules. The development record contains 336 mutation-library entries, 153 modeled classes after no-op removal, and 90 distinct H detection rows after deduplication. Artifact read/rebuild overhead is not authoring effort. No authoring-time record exists: `AUTHORING_TIME_NOT_MEASURED`.

Traceability boundary. The manuscript reports supplied frozen or versioned summaries. The V2 maximum-entropy implementation change and its ordering limitation are recorded above rather than described as prospectively frozen. Formal guarantees remain conditional on the fixed oracle, frontier, modeled class, and feasible-plan universe; empirical interpretations are limited to the stated cohorts and comparisons.