EVERY FAILURE IS A TIMEOUT: FAILURE-MODE HOMOGENEITY MAKES ROBOT POLICY EVALUATORS UNVERIFIABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic robot policy evaluators are commonly validated by their agreement with ground-truth success labels. We show that such agreement can provide no evidence of genuine evaluation capability when the evaluation protocol itself reveals the outcome. Under stop-on-success evaluation, policies without a termination channel fail exclusively by reaching the time horizon, allowing episode duration alone to distinguish success from failure. We establish this failure mode through a controlled intervention on the termination channel and construct a frame-blind witness that exactly reproduces a vision–language judge's evaluation metrics. We further formalize an evidence budget based on success–failure pairs misordered by duration and derive a randomization floor below which evaluator certification is impossible. Across 47 screened conditions, only two satisfy this floor, and neither remains admissible after suppressing the termination channel. To address this identifiability failure, we investigate fixed-horizon logging, which eliminates duration-based outcome leakage without increasing the number of rollouts. This protocol also distinguishes achievement from terminal success, revealing discrepancies concealed by stop-on-success evaluation. However, removing duration leakage does not guarantee evaluator verifiability: the first success time retains predictive information, while some terminal success labels depend on historical states inaccessible to frame-based evaluators. Together, our findings establish that reliable evaluator validation requires both sufficient evidence beyond protocol-induced shortcuts and alignment between ground-truth labels and observable information.

1 INTRODUCTION

Automatic evaluators are becoming integral to robot policy evaluation. Vision–language judges and learned success classifiers increasingly replace manual inspection of rollouts, providing success estimates that inform checkpoint selection, benchmark comparisons, and subsequent policy development. Their reliability is typically established through agreement with ground-truth labels on held-out rollouts, measured by accuracy, AUROC, or calibration. Yet high agreement does not necessarily establish that an evaluator can recognize task success from the observations it receives. If the evaluation protocol itself reveals the outcome through an incidental feature, the same agreement may be achieved without interpreting robot behavior. Beyond evaluator accuracy, this raises the question of whether an evaluation set contains sufficient evidence to distinguish genuine task recognition from protocol-induced shortcuts.

We identify a structural source of this problem in the widely used **stop-on-success** evaluation protocol. Under this protocol, a rollout terminates when the task's success predicate fires, when the time horizon is reached, or when the policy explicitly requests termination. For policies without a termination channel, failures can end only at the horizon, whereas successful episodes typically terminate earlier. When no success reaches the horizon, episode duration perfectly separates the two outcomes. Rather than a shortcut that an evaluator must learn, this is a property of the evaluation data imposed by the interaction between the policy's action space and the rollout harness. Indeed, the absence of a termination channel makes this failure mode predictable from the policy and harness specifications before any rollout is scored.

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.

A concrete example illustrates the resulting validation failure. On the Octo-Small / `close_drawer` condition, all 23 failed episodes terminate at the horizon of 114 steps, whereas all 17 successes terminate earlier. Episode duration alone therefore achieves perfect outcome discrimination. A vision–language judge obtains an AUROC of 0.893, yet a comparator that never observes a frame can reproduce its reported evaluation metrics exactly. We construct this **frame-blind witness** by permuting the judge's own scores within duration-defined outcome classes. Because duration already determines the outcome, the permutation preserves the class-conditional score distributions and consequently leaves AUROC, the complete ROC curve, calibration, Brier score, and expected calibration error unchanged. Across three conditions, the construction reproduces all 138 ROC operating points while substantially changing the assignment of scores to individual episodes (Figure 1). The witness is conditional on the judge's observed score multiset and is not a deployable alternative evaluator. Rather, it demonstrates that conventional agreement statistics alone cannot distinguish the judge's predictions from scores assigned without visual evidence.

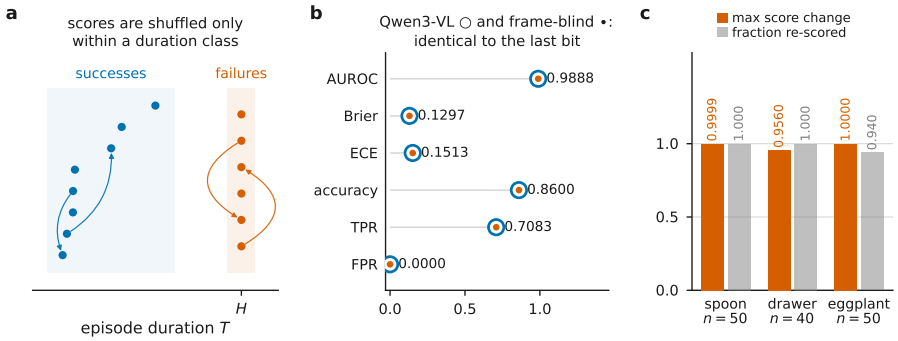

Figure 1: **Identical evaluation reports without visual evidence.** **(a)** When episode duration perfectly separates success from failure, a judge's scores can be permuted within duration-defined outcome classes while preserving the class-conditional score distributions. **(b)** The original judge and the frame-blind witness obtain identical conventional evaluation statistics. **(c)** Despite this agreement, the permutation substantially changes the scores assigned to individual episodes, demonstrating that agreement statistics alone cannot establish frame-derived evaluation capability.

We formalize this limitation through an **evidence budget**, defined as the success-failure pairs that episode duration orders incorrectly and on which an evaluator could therefore demonstrate an advantage over the clock. When duration perfectly separates outcomes, this budget is zero, regardless of the number of additional rollouts collected. To quantify the finite-sample requirement for evaluator validation, we derive a randomization floor using pairs that share neither an episode nor an initial state. Under the associated randomization framework, no valid test can reject at level δ with fewer than $\lceil \log_2(1/\delta) \rceil$ matched units. At $\delta = 0.1$, at least four such units are required. This floor is a necessary condition for certification under the specified test, not a guarantee of evaluator competence, and follows from the randomization structure rather than a new general statistical result.

The prevalence of this limitation is evident across our evaluation census. In 27 SimplerEnv conditions, all 646 failures from two policies without termination channels reach the horizon, whereas RT-1-X terminates 179 of its 296 failures early. Among 47 screened conditions, 14 lack cross-label pairs, 27 have zero evidence budget, and four fall below the randomization floor, leaving only two admissible conditions. Neither remains admissible after we suppress RT-1-X's termination component on the same seeds. All 34 resulting failures reach the horizon, and duration becomes a perfect outcome classifier, establishing the termination channel's causal role in evaluator identifiability.

Since this failure arises from the evaluation protocol, we investigate **fixed-horizon logging** rather than attempting to correct the collected data. Running all episodes to the same horizon eliminates duration-based outcome leakage across nine conditions and 1,800 episodes while preserving the first success time and terminal state. It also reveals 93 episodes in which achievement and terminal success disagree, a distinction concealed by stop-on-success evaluation.

Fixed-horizon logging nevertheless does not resolve all identifiability failures. A clock reconstructed from the first success time outperforms the judges in 14 of 18 arms, while 28 of 38 terminal successes in an audited condition depend on a non-decaying history counter inaccessible to frame-based evaluators. These findings show that removing protocol-induced shortcuts restores opportunities for

validation but cannot guarantee it when ground-truth labels depend on information unavailable to the evaluator.

We make three contributions. First, we formulate evaluator validation as an identifiability problem and introduce a conditional frame-blind witness, an evidence budget, and a randomization floor that characterize when conventional evaluation statistics can support claims of evaluator skill beyond duration. Second, we identify failure-mode homogeneity as a structural consequence of the interaction between policy action spaces and stop-on-success evaluation. We establish its prevalence through a multi-policy census and its mechanism through a controlled termination-channel intervention, providing a diagnostic that can be applied before evaluator scoring. Third, we investigate fixed-horizon logging as a collection-level remedy and characterize its limitations. By separating achievement from terminal success and identifying residual dependence on predicate history, we show that reliable evaluator validation requires both sufficient evidence beyond protocol-induced shortcuts and ground-truth labels aligned with the evaluator's observable information.

2 Related work

Automatic evaluators and their validation. Automatic judges are increasingly validated through agreement with ground-truth labels (Zheng et al., 2023), and vision–language models have been adopted as success detectors for robot rollouts (Du et al., 2023). Statistical approaches further provide distribution-free guarantees on prediction and disagreement risks (Bates et al., 2021; Angelopoulos et al., 2024; Jung et al., 2025). These methods establish reliability with respect to their specified targets but do not necessarily demonstrate that an evaluator uses task-relevant visual evidence. Our work examines this distinction by asking whether the evaluation set can distinguish genuine success recognition from predictions supported by information leaked through the evaluation protocol.

Benchmark shortcuts and evaluation protocols. Models can exploit unintended features of benchmarks (Geirhos et al., 2020), and prior work has investigated how much of a video benchmark can be solved without temporal evidence (Buch et al., 2022). We extend this line of inquiry from benchmark performance to evaluator validation by constructing a frame-blind witness that exactly reproduces conventional evaluation statistics. Existing mitigations include matched subsets and length-controlled evaluation (Dubois et al., 2024). However, in robot rollouts, stop-on-success evaluation determines both episode duration and the terminal observation, so matching existing episodes by duration cannot recover the observations lost through early termination. Previous studies have also demonstrated how time limits and environment conventions influence agent learning and evaluation (Pardo et al., 2018; Machado et al., 2018). We instead investigate how the interaction between policy action spaces and termination rules determines whether the resulting data can support claims about evaluator capability.

Statistical identifiability and robot evaluation. Our evidence budget is closely related to the overlap requirement in causal inference, where violations of positivity can make a target unidentifiable regardless of the estimator (Rosenbaum & Rubin, 1983; Petersen et al., 2012; D'Amour et al., 2021). The analogy concerns the absence of comparable observations rather than a direct application of causal identification, since episode duration is generated by the rollout outcome and termination protocol rather than being a pre-treatment covariate. We characterize this finite-sample limitation through clock-discordant success–failure pairs and derive a certification floor under matched-unit randomization, alongside a more conservative criterion based on e-values (Vovk & Wang, 2021; Ramdas et al., 2023). Existing work on empirical robot policy evaluation emphasizes reliable performance comparisons and sufficient rollout budgets (Kress-Gazit et al., 2024), while simulation benchmarks provide reproducible environments for such comparisons (Li et al., 2024; Liu et al., 2023). Our analysis addresses a complementary question concerning whether the collected rollouts contain the evidence required to validate an automatic evaluator in the first place.

3 Evaluator validation as an identifiability problem

This section asks what an evaluation set can say about an evaluator. It defines the diagnostics, states the evidence floors and the corollary, and records what fixed-horizon logging changes.

3.1 SETUP AND DIAGNOSTICS

An *evaluation condition* is one policy on one task, with n rollouts that each carry an outcome $Y_i \in \{0, 1\}$ and a duration T_i. Under stop-on-success, the standard harness rule (Li et al., 2024; Liu et al., 2023), a rollout ends when the success predicate first fires, at the horizon H, or at a termination signal of the policy. Under fixed-horizon logging, every rollout runs to H and is labeled by the predicate at H, with the same seeds and rollout count, and its first firing step t_i^* is kept. A *termination channel* is an action dimension through which a policy can end its own episode.

The *early-failure fraction* φ is the share of failures that end before H, and the *clock score* is the AUROC of the duration-only rule. The *evidence budget* b counts the success–failure pairs with $T_{\text{success}} \geq T_{\text{failure}}$, the only pairs on which an evaluator can beat the clock. Ties count as discordant. This decides b when a success lasts exactly H, and under a fixed horizon every cross-label pair ties. The *matched-unit count* m^* is the largest number of such pairs sharing no episode, and m_{c}^* also forbids a shared initial state. A condition is *admissible* at level δ when m_{c}^* reaches the floor of Proposition 3, a property of the evaluation set. An evaluator is *certified* on an admissible set when a valid level-δ test rejects the null that it is no better than the clock. Admissibility is necessary for certification but not sufficient. Skill beyond duration is *identifiable* when the set separates the evaluator from the clock.

3.2 NON-IDENTIFIABILITY AND EVIDENCE FLOORS

Proposition 1 (Conditional non-identifiability)**.** *Fix a batch in which no success lasts exactly H. Condition on the labels Y, the durations T and the scorer's* own *observed score multiset within each outcome class. Suppose that the duration supports of the two outcome classes are disjoint, so that $Y_i = g(T_i)$ for some g. Let $\tilde{s}$ assign to episode i the score $s_{\sigma(i)}$, where σ is drawn uniformly from the permutations that map each level set of g onto itself. Then, conditionally on those multisets, $\tilde{s}$ is a function of (T, U) alone for exogenous randomness U, and the joint empirical distributions of $(\tilde{s}, Y)$ and (s, Y) are identical. Every functional that depends on the scores only through the class-conditional multisets therefore takes the same value for both. This includes accuracy, AUROC, the ROC curve, calibration bins, the Brier score, expected calibration error, and any test built on them.*

The permutation changes which episode receives which score, not the scores that each class holds. Built from the scorer's own output, the comparator is a witness, not a deployable rival evaluator, and statistics based on episode identity or on the score–duration association tell the two apart. The negative claim is that reported statistics are functions of duration on such batches (Appendix A).

Proposition 2 (Evidence-budget floor)**.** *Consider a test of the null that an evaluator is no better than the clock, based on the signs of the disagreements over the b clock-discordant pairs. An evaluator that wins all b pairs attains the largest possible statistic, at exact one-sided p-value 2^{-b}. Certification at level δ therefore requires*

$$b \;\geq\; b_{\min}(\delta) \;=\; \lceil \log_2(1/\delta) \rceil \,. \tag{1}$$

Proposition 3 (Randomization floor)**.** *Condition on the observed score multiset, the labels and the durations. Let M be a set of clock-discordant pairs in which no two pairs share a* unit*, and write $m = |M|$. Consider the sharp null under which exchanging which member of a pair carries the success label leaves the joint distribution of scores and labels invariant. Under this null, the sign vector on M is uniform on $\{-1, +1\}^m$, a randomization group with 2^m elements. Any level-δ test measurable with respect to that group has a rejection region of at most $\lfloor \delta\, 2^m \rfloor$ points. A non-empty rejection region therefore requires*

$$m \;\geq\; \lceil \log_2(1/\delta) \rceil \,. \tag{2}$$

When $m = 0$ the group is trivial, so no such test attains power exceeding its own level, whatever the evaluator and whatever n.

At $\delta = 0.1$ the floor is four. Neither the evaluator nor n enters Equation 1, so the bound belongs to the evaluation set. It is not an artifact of the sign test (McNemar, 1947), and the e-value criterion (Vovk & Wang, 2021) raises it to seven. The unit is the episode for m^* and the initial state for m_{c}^*, and for $m = 0$ no assumption is needed. Both floors are necessary, not sufficient, and follow from the randomization structure, not from a new general result.

3.3 THE ACTION-SPACE COROLLARY

Corollary 1 (Action-space corollary)**.** *Consider a harness whose only terminations are the success predicate, the horizon H, and any termination signal the policy emits. If the policy's action space contains no termination channel, then $\varphi = 0$, and every failed episode ends at H. The failure duration support is then the single point H. Unless some success also lasts exactly H, the supports are disjoint and Proposition 1 applies. The evidence budget is then $b = 0$, and by Proposition 2 no evaluator can be certified at any $\delta < 1$, at any sample size.*

The check needs only the model card and the harness rule, so it is decidable before collection. It works in one direction, since a reserved channel guarantees nothing until φ is measured. If a platform also aborts on collision or operator intervention, φ is a lower bound. The screen then stays conservative for impossibility but becomes *anti*-conservative for admissibility.

3.4 FIXED-HORIZON LOGGING AND ITS RESIDUAL LIMITS

Fixed-horizon logging removes the harness rule that makes a missing channel fatal.

Remark 1. Under fixed-horizon logging $T \equiv H$, so the realized clock is 0.500 and every cross-label pair ties. The witness fails, because its permutation then destroys the class-conditional multisets.

Remark 2. The minority outcome class, an episode count, replaces the clock as the evidence.

Remark 3. The retained t^* gives a *derived clock*, the duration that the episodes would have shown under stop-on-success. The fixed-horizon pool itself does not show it.

Remark 4. An episode *achieves* the task if the predicate ever fires, and it is a *terminal success* if the predicate holds at H. A *reversal* achieves without terminal success. The labels coincide under stop-on-success, and we take no position on which to report. *If the predicate's truth at H depends on non-decaying episode history, no frame-reading evaluator can recover the terminal label.*

Sections 4.2 to 4.4 test Proposition 1 and Corollary 1, fixed-horizon logging, and Remarks 3 and 4.

4 EXPERIMENTS

4.1 EXPERIMENTAL DESIGN

Policies. Whether a policy can stop itself is readable from its model card. RT-1-X tokenizes eight action dimensions, one of them termination (Brohan et al., 2022; O'Neill et al., 2023), and Octo predicts seven continuous dimensions with a diffusion head (Octo Model Team et al., 2024). Decoder family and quantization do not decide this. OpenVLA (Kim et al., 2024) inherits RT-1's discretization but drops the eighth dimension, SpatialVLA (Qu et al., 2025) discretizes into adaptive spatial grids, and both land on seven dimensions with no channel, like the environment (Li et al., 2024) and the LIBERO policies OpenVLA-OFT and $\pi_{0.5}$.

Collections. Table A1 registers all pools. S1 holds 27 SimplerEnv conditions of Octo-Base, Octo-Small and RT-1-X under stop-on-success, and S2 and S3 hold ten LIBERO cells each for OpenVLA-OFT and $\pi_{0.5}$. *Our nine conditions* are nine S1 cells fixed before scoring, which W1 re-collects under fixed-horizon logging at 200 episodes each. T1 is the masked run, P1, P2 and F1 serve the correction study, C1 adds OpenVLA and SpatialVLA, and M1, D1 and D2 serve appendix analyses.

Evaluators and provenance. An *evaluator* is the system being validated, a *judge* is a vision–language evaluator, and a *scorer* is any score-producing function, including the witness. Qwen2.5-VL-7B is open-weight and runs locally. Qwen3-VL-Plus (2025-12-19) and GPT-5.6-sol are vendor-served, and GPT-4o (`gpt-4o-2024-11-20`) runs through a third-party gateway. No pool is scored by all four. We name the scorer behind every certification claim, and this provenance fixes the scope of each claim in advance. GPT-4o is descriptive only, and GPT-5.6-sol is gate-blocked on the committed-verdict batch D2. Results on D1 and D2 are scoped to pools we can describe but not name, and the scorer protocol for W1, C1 and D2 is unestablished (Appendix C).

Statistics. A certification ladder sweeps a calibration setting over 26 rungs at $\delta = 0.1$ on 50 re-calibration splits, and $N1$ is the fraction of non-degenerate splits that certify. It tests a calibration-majority null at margin $\gamma_0 = 0$ with Holm control at $\alpha = 0.1$ (Holm, 1979). A matched-unit recom-

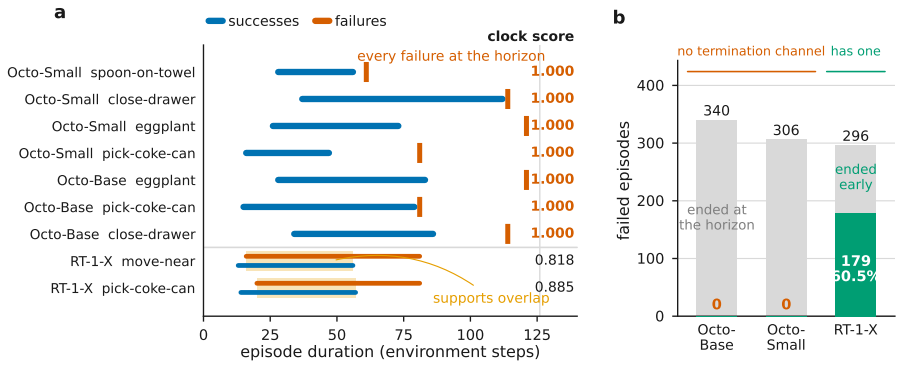

Figure 2: **Every failure is a timeout for policies without a termination channel. (a)** Under stop-on-success, the seven Octo conditions among our nine have every failure at the horizon and a clock score of 1.000, while the RT-1-X supports overlap on 21 and 27 episodes (shaded). **(b)** Across 27 SimplerEnv conditions, no Octo failure ends early, but RT-1-X ends 179 of 296 early.

putation uses an exact sign test on initial-state-disjoint units (Appendix C.4). Deduplication changes no admissibility verdict (Appendix B), and a learned success head appears only in Appendix N.

4.2 Stop-on-success makes evaluators unverifiable

The frame-blind witness. A scorer that never sees a frame reproduces every reported statistic of a judge (Figure 1). The judge is Qwen3-VL-Plus, labeled Qwen3-VL in Figure 1b, on three Octo-Small conditions of S1 with a clock score of 1.000. The witness matches its AUROC of 0.9888, 0.8926 and 0.9951 bit for bit (Table A2). The ROC tables with 138 operating points, the calibration tables, the confusion matrices and the sorted score multisets are identical. At the episode level the two scorers agree on nothing else, with maximum score differences of 0.9999, 0.9560 and 1.0000 (Figure 1c). Conditioning on duration separates them, since 6,152 of the 19,370 ladder statistics differ (Appendix C.3). The ladder certifies neither, so it is a diagnosis and not yet a remedy.

Every failure is a timeout. Policies without a termination channel fail only at the horizon (Figure 2). Across the 27 conditions of S1, none of the 646 Octo failures ends early (340 Octo-Base, 306 Octo-Small). RT-1-X ends 179 of its 296 failures early. The census thus splits on the binary property that Corollary 1 names, and its side condition holds, as no success ends at the failure horizon. In the seven Octo conditions of our nine, the supports share no episode and the clock score is exactly 1.000. Since $\varphi = 0$ is forced, resampling adds only timeouts, and the test has power exactly zero at every sample size.

Only two of the 47 screened cells are admissible (Table A3). Fourteen lack one outcome class. Of the other 33, 27 supply zero evidence because duration separates them perfectly, 26 of them from policies without a channel. Four of the remaining six fall below the floor once pairs share neither an episode nor an initial state. The two that clear it are RT-1-X cells at $m_c^* = 5$ against a floor of four. Under the e-value criterion on matched units none of the 47 is admissible. Under the b criterion both RT-1-X cells clear the threshold, and we report both verdicts wherever admissibility is claimed. On LIBERO, an OpenVLA-OFT checkpoint degraded in four perturbation variants has a near-balanced pooled success rate of 0.55375, yet none of its 357 failures ends early. The policy $\pi_{0.5}$ leaves three minority-outcome episodes against a floor of four. All 20 LIBERO cells supply zero matched units. This is a fact about the policies run on that benchmark rather than about the benchmark, and it holds by deduction for any policy without a channel (Appendix D). A validation result on such data cannot be interpreted, whatever agreement figure accompanies it.

The termination channel is the cause. Suppressing one decoded dimension removes this paper's entire admissible set (Figure 3). The masked run T1 repeats the two admissible RT-1-X cells and forces only the decoded `terminate_episode` component to zero. Early failure disappears in all 34 failures. The clock score rises from 0.8177 and 0.8846 to exactly 1.000000, and b and m^* fall to zero. The policy is not damaged, since success rises from 43 to 46 of 80 and all three label changes run from failure to success. Nor does the mask act through a diverged trajectory, because every label-changing pair has a bit-identical action prefix (Appendix E). It removes the ability to stop, not

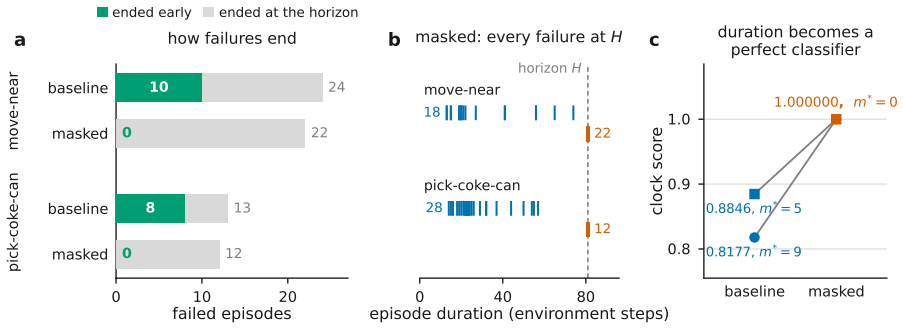

Figure 3: **One decoded dimension decides whether the evaluation set can validate anything.** RT-1-X runs on its two admissible cells with the same checkpoint, 80 seeds and simulator, and only `terminate_episode` is forced to zero. **(a)** The policy's own early terminations, 18 of 37 failures, disappear. **(b)** All 34 remaining failures sit at the horizon. **(c)** The clock score becomes exactly 1.000000, and b (72 and 41) and m^* (9 and 5) fall to zero.

Table 1: **Fixed-horizon logging restores evidence on all nine conditions at unchanged rollout count.** The left block is stop-on-success (S1). The right block is fixed-horizon logging (W1), whose realized clock is 0.500, so the informative clock is the *derived* one. *Min.* is the minority class, *rev.* and *frac.* give reversals, and *ach.* and *term.* are the achievement and terminal rates. Admissibility is a screening verdict, not a certification. †This count is one below the gate of 30 (Appendix G).

		stop-on-success ($n = 40$–50)				fixed horizon ($n = 200$)					
Policy	Task	φ	clock	m_c^*	adm.	derived	min.	rev.	frac.	ach.	term.
Octo-Base	close-drawer	0.00	1.000	0	no	0.9894	81	4	0.020	0.425	0.405
Octo-Base	pick-coke-can	0.00	1.000	0	no	0.9758	29†	13	0.065	0.210	0.145
Octo-Base	eggplant	0.00	1.000	0	no	0.9961	90	2	0.010	0.460	0.450
Octo-Small	close-drawer	0.00	1.000	0	no	0.9951	71	1	0.005	0.360	0.355
Octo-Small	pick-coke-can	0.00	1.000	0	no	0.9379	38	27	**0.135**	0.325	0.190
Octo-Small	eggplant	0.00	1.000	0	no	0.9824	93	3	0.015	0.550	0.535
Octo-Small	spoon-on-towel	0.00	1.000	0	no	0.9431	100	15	0.075	0.575	0.500
RT-1-X	move-near	0.42	0.818	5	yes	0.9438	79	11	0.055	0.450	0.395
RT-1-X	pick-coke-can	0.62	0.885	5	yes	0.8609	68	17	0.085	0.745	0.660
Total, 1,800 episodes								93	0.052	0.456	0.404

skill. The result concerns two 40-episode screening cells and the channel, not a benchmark. As a decoding-time surrogate, it shows that removing a channel removes the evidence, not that adding one would restore it.

4.3 Fixed-horizon logging

Post hoc correction does not recover the estimate. Duration matching recovers only about a tenth of the protocol gap (Appendix F). On RT-1-X / pick-coke-can under both protocols (P1 and P2, $n = 200$ each), Qwen3-VL-Plus scores 0.91384 as reported and 0.90521 after length matching (Dubois et al., 2024). The fixed-horizon reference is 0.82977 and the early-pool clock score 0.78788, so matching closes 0.00863 of the 0.08407 gap. On move-near it removes 3.6%, and with two conditions we give no interval on either fraction. Stopping at success does not merely shorten an episode. It also changes the judge's last frames. Changing only the observed frames of fixed episodes (F1) moves AUROC by $+0.1049$ $[0.0527, 0.1624]$ on move-near and $+0.0187$ $[-0.0274, 0.0615]$ on pick-coke-can (95% intervals), establishing this second channel on one condition of two. Switching labels moves AUROC by only 0.011 and 0.018, so the success definition is not the cause. A clock correction addresses one channel of two.

Fixed-horizon logging restores evidence. Fixed-horizon logging supplies evidence on all nine conditions at unchanged rollout count (Table 1). Duration is constant across the 1,800 episodes of W1, so the realized clock is exactly 0.500, and each condition supplies 29 to 100 minority-outcome episodes against a floor of four. The change is not that a zero budget became certifications. Stop-on-success left no pair for assessing skill on seven of the nine. The null is unchanged, but the clock

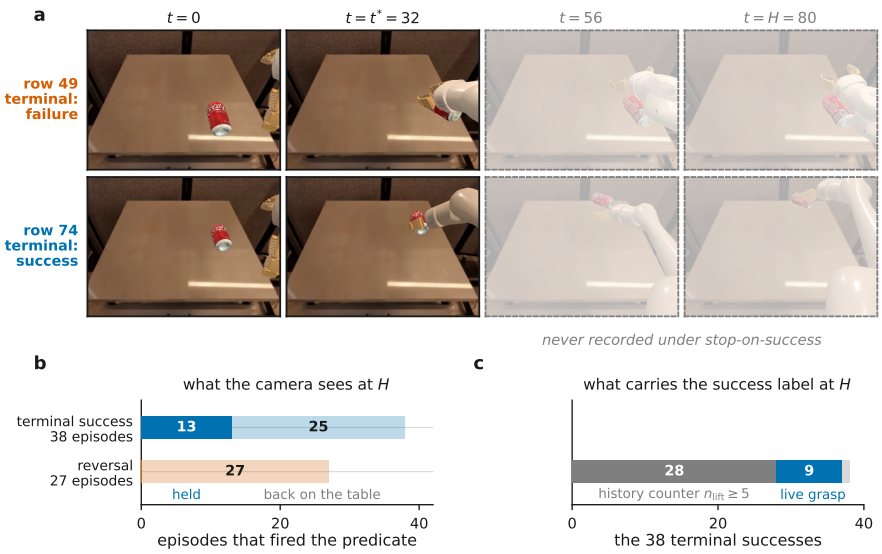

Figure 4: **The terminal label is not in the terminal frame.** Octo-Small runs on pick-coke-can with $H = 80$. **(a)** Rows 49 and 74 both fire at $t^* = 32$ and end with the can on the table, yet only row 74 is a terminal success. **(b)** Of the 38 terminal successes, 25 end with the can back on the table, as do all 27 reversals. **(c)** The history counter $n_{\text{lift}} \geq 5$ carries the label for 28 of the 38 (Appendix H).

now ties on every pair. At the episode level, Qwen2.5-VL-7B and Qwen3-VL-Plus certify 12 of 18 evaluator–condition arms against a calibration-majority null under Holm control at $\alpha = 0.1$. The episode is the wrong unit (Section 4.4), and we read the 26-setting ladder as descriptive. GPT-4o, excluded by the provenance rule, passes four further arms. Ten of the twelve arms fall on five Octo conditions that had a clock of 1.000 and zero pair supply, so conditions that could answer nothing now supply most of the evidence. On the certification matrix, margin over the clock rather than raw AUROC separates the scorers, and only Qwen3-VL-Plus certifies, on 3 of 3 conditions (Appendix K).

Achievement versus terminal success. One collection recovers both labels, which differ on 93 of the 1,800 episodes, with 820 fired and 727 terminal successes (Table 1). Reversals reach 13.5% in one condition, and achievement labels inflate the reported rate by up to 1.711×. On pick-coke-can the gap between the two Octo policies is 11.5 points under achievement labels and 4.5 under terminal labels, with rankings preserved on all three task families (Appendix H). Reporting one number is thus a label choice that the protocol may not expose.

4.4 What fixed-horizon logging cannot fix

A fixed horizon supplies a budget, but it does not guarantee that any evaluator wins. Octo-Base / pick-coke-can supplies 29 minority-outcome episodes (Table 1), yet neither judge certifies on any of its 44 non-degenerate splits (Appendix I). Four results run against our own case.

The first success time is still a clock. A clock rebuilt from the first success time still beats the judges in 14 of 18 arms. The derived clock (Remark 3) scores 0.861 to 0.996 (Table 1). It uses the predicate's own trace, so beating it was never the bar, and its margins bound what a judge adds to that trace. Part of the gap is an observability bound (Figure 4), and a judge may also consume duration directly, which we do not model. On two policy families we had not studied, all nine arms of C1 fall below the derived clock, by -0.0178 to -0.4647, although the conditions were selected before scoring on the two criteria that most favor a judge (Appendix J). This result is descriptive, and C1 does not test these families for $\varphi = 0$.

The terminal label is not in the terminal frame. Predicate memory rather than scene content carries most terminal labels in the audited condition (Figure 4). In Octo-Small / pick-coke-can, the highest-reversal condition, 65 episodes fired the predicate. Only 13 of the 38 terminal successes still hold the can at H, and the other 25 put it back on the table, where all 27 reversals also end. The non-

decaying history counter $n_{\text{lift}} \geq 5$ carries the label for 28 of the 38, a live grasp for 9 and both for 1. Holding the can is what the camera sees, and a live grasp is a predicate term, so the counts need not match. This second identifiability failure is independent of duration. Fixed-horizon logging makes the terminal label observable to the logger but only partly to a frame-reading evaluator (Remark 4). Predicates should therefore be state-free at H, or a benchmark should publish their state dependence beside its success rate.

Clustering and the conservative criterion. Clustering by initial state removes most of the positive result. The $1{,}800$ episodes come from 287 initial states, with a design effect of 6.636. Collapsing to one episode per initial state lowers $N1$ in all twelve arms on the six conditions that repeat states, at worst from 0.980 to 0.041, and the count with $N1 > 0.5$ falls from 12 to 10 of 18 (Appendix I). The ladder's Holm count has not been recomputed at that level, so no clustered ladder count should be read from this paper. The initial-state-disjoint recomputation certifies 16 of 18 arms, but with another unit and test it does not revise the 12, and we report the smaller number (Appendix C.4). Under the e-value criterion on matched units no cell is admissible. On D1 and D2, which we can describe but not name, a transferred duration rule $T \leq 66$ errs below two of three scorers on unseen content and above the strongest (Appendix L). Fixed-horizon logging is necessary, not sufficient.

5 Recommendations

A policy with no termination signal on a stop-on-success harness yields $\varphi = 0$, a clock score of 1.000 and no admissible pair, so Corollary 1 settles before collection that no evaluator run on it can be validated. After collection, report φ, the clock score beside the evaluator AUROC, the matched-unit count and the admissibility verdict. Below $\lceil \log_2(1/\delta) \rceil$ matched units, withhold agreement rather than publish it with an interval, since an interval on an unidentifiable quantity is a different kind of error, not a weaker claim. Where safe, log to the horizon with t^* and the terminal state, deduplicate content, and report both labels with the reversal fraction. Make predicates state-free at the horizon or publish their state dependence (Appendix M).

6 Limitations and conclusion

Limitations. We do not measure the overhead of running successes past t^*, and on hardware continued actuation may disturb the achieved state, so deployment needs qualification beyond simulation. The screen is an admissibility test, not a magnitude estimator. It assumes a per-step predicate, simulator-derived labels and tasks whose objective is not duration, and it does not directly apply otherwise. Three corrections remain unrun, namely the ladder's Holm count at the initial-state level, the run-to-run variation of $N1$ and a split-sample protocol for γ^*. The provenance of D1 and D2 and the scorer protocol for W1, C1 and D2 remain open (Appendices B, C).

Conclusion. A policy that cannot end its own episode can fail only by running out of time, so its evaluation set contains no pair on which an evaluator could outrank the clock. Masking one decoded dimension on matched seeds makes the admissible set inadmissible, which shows that the channel causes this condition. On the model card, no reserved channel forces this failure mode, whereas a reserved channel guarantees nothing until φ is measured. One line of code can remove the condition, but not the second failure, in which the predicate carries state that the camera cannot observe.

AI USE STATEMENT

All scientific work in this paper was done by the authors. Generative AI tools were used only for limited supporting tasks in the three areas described below. The authors reviewed every AI-assisted output and take full responsibility for the content of this paper.

Writing. The authors wrote the manuscript, and AI tools were not used to draft it. AI tools were used only to polish the authors' text. This covered language editing, rewording for clarity, moving existing text within the outline that the authors specified, and checks that terminology, numbers and cross-references are consistent. They did not add results or claims.

Retrieval and discovery. AI tools helped search for related work and retrieve the bibliographic records of cited papers. The authors chose every reference and wrote the discussion of related work themselves. Every citation was checked against its source.

Research ideation and execution. AI tools were not used for research ideation. The research question, the theoretical results, the experimental design, the policy rollouts, the evaluator scoring runs and the analyses are the authors' own work. In the experiments, AI tools were used only to help test the authors' code under different random seeds. They did no core experimental work, and every number in this paper comes from the authors' experimental records.

The vision–language models evaluated in this paper (Section 4.1) are objects of study, and this statement does not concern them.

REPRODUCIBILITY STATEMENT

The central checks need no model. The admissibility screen uses only an outcome label and an episode duration per rollout, and Equations 1 and 2 are closed-form, so it runs without a GPU. Appendix A proves Propositions 1 to 3 and Corollary 1, and Sections 3.1 and **??** state the collection change in full. On acceptance we will release the per-episode `(label, duration, initial-state hash)` tables for all 47 screened cells, the screen, the m^* and m_c^* matchers, the harness-side surrogate run of Appendix E, and the scripts that produce every table and figure here.

ETHICS STATEMENT

This work evaluates automatic evaluators on simulated robot rollouts. It involves no human subjects and no personally identifying data. One scorer was served through a third-party API gateway rather than its official endpoint. This is disclosed in Section 4.1, and no claim in the paper rests on that scorer alone.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, et al. RT-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2917–2927, 2022.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi. Vision-language models as success detectors. In *Conference on Lifelong Learning Agents (CoLLAs)*, volume 232, pp. 120–136, 2023.

Yann Dubois, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple debiasing of automatic evaluators. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.04475.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.18370.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, 2024. arXiv:2406.09246.

Hadas Kress-Gazit, Kunimatsu Hashimoto, Naveen Kuppuswamy, Paarth Shah, Phoebe Horgan, Gordon Richardson, Siyuan Feng, and Benjamin Burchfiel. Robot learning as an empirical science: Best practices for policy evaluation. *arXiv preprint arXiv:2409.09491*, 2024.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In *Conference on Robot Learning (CoRL)*, 2024. arXiv:2405.05941.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 44776–44791, 2023.

Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems (RSS)*, 2024. arXiv:2405.12213.

Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open X-embodiment: Robotic learning datasets and RT-X models. *arXiv preprint arXiv:2310.08864*, 2023.

Fabio Pardo, Arash Tavakoli, Vitaly Levdik, and Petar Kormushev. Time limits in reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 4045–4054, 2018. arXiv:1712.00378.

Maya L. Petersen, Kristin E. Porter, Susan Gruber, Yue Wang, and Mark J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Jiayuan Gu, Zhigang Wang, Yan Ding, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring spatial representations for visual-language-action models. In *Robotics: Science and Systems (RSS)*, 2025. arXiv:2501.15830.

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

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Bernhard Rüger. Das maximale Signifikanzniveau des Tests: "Lehne h_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen". *Metrika*, 25(1):171–178, 1978. doi: 10.1007/BF02204362.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric 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 (NeurIPS), Datasets and Benchmarks Track*, 2023.

A Proofs and remarks for Section 3

This appendix proves the results of Section 3 and expands the remarks that the main text states in one sentence each.

Proof of Proposition 1. The supports are disjoint, so $Y_i = g(T_i)$ for every episode. The permutation σ maps each level set of g onto itself, so $Y_{\sigma(i)} = Y_i$ for every i. The pairs $(\tilde{s}_i, Y_i) = (s_{\sigma(i)}, Y_{\sigma(i)})$ are therefore a rearrangement of the pairs (s_j, Y_j), and the two joint empirical distributions coincide. Because duration already determines the outcome, the permutation preserves the class-conditional score distributions. It therefore leaves AUROC, the complete ROC curve, calibration, the Brier score and expected calibration error unchanged, together with every other functional of the class-conditional multisets. Given the class-conditional score multisets, $\tilde{s}$ places the multiset of each class on the episodes of that class in a uniformly random order. This holds for any original assignment of scores to episodes. It therefore depends on the data only through the level sets of g, which T determines, and through σ, whose randomness U is exogenous. Hence $\tilde{s}$ is a function of (T, U) alone.

The witness under a fixed horizon. The construction cannot survive a fixed horizon. With duration constant, the level sets of g collapse to a single class. The permutation then becomes one of the whole sample, and it destroys the class-conditional multisets rather than preserving them.

Proof of Proposition 2. Under the null, the signs of the disagreements over the b clock-discordant pairs define an exact one-sided test. An evaluator that wins all b pairs attains the largest possible statistic, at exact one-sided p-value 2^{-b}. No evaluator can reach a smaller p-value on this evidence. Certification at level δ therefore needs $2^{-b} \leq \delta$, and because b is an integer this is Equation 1. At $\delta = 0.1$ the floor is $b_{\min} = 4$. Four discordant pairs give $p = 0.0625$ and pass, and three give $p = 0.125$ and cannot, however good the evaluator is.

The floor under other criteria. The floor is not an artifact of the paired sign test (McNemar, 1947). Under the uniform-mixture e-value (Vovk & Wang, 2021), winning all b pairs yields $e = 2^b/(b+1)$. At $\delta = 0.1$ the requirement $e \geq 1/\delta$ then needs $b \geq 7$. The b pairs are not independent, so treating b wins as b independent trials is anti-conservative. This matters for accepting an evaluator and not for impossibility. A bound under an over-optimistic test is still necessary, and a condition that fails it fails under every valid test. Proposition 3 removes the asymmetry instead of relying on it.

Proof of Proposition 3. Under the sharp null, exchanging the success label within any pair of M leaves the joint distribution of scores and labels invariant. The pairs of M share no unit, so the exchanges on different pairs can be made independently. Conditionally on the observed score multiset, the labels and the durations, the sign vector on M is therefore uniform on $\{-1, +1\}^m$. Each of the 2^m elements of this randomization group has probability 2^{-m} under the null. A level-δ

test that is measurable with respect to the group can therefore reject on at most $\lfloor \delta\, 2^m \rfloor$ elements. A non-empty rejection region needs $\lfloor \delta\, 2^m \rfloor \geq 1$, which means $2^m \geq 1/\delta$, and for integer m this is Equation 2. When $m = 0$ the group has one element, and a level-δ test with $\delta < 1$ cannot reject on it with probability above its level. This ranges over every test valid under the randomization, not over one test family.

The choice of unit. The randomization does not remove all dependence, and the assumption lies in the choice of unit. The episode unit gives m^* and assumes that episodes are exchangeable under the null. Two episodes from one initial state are not, and Section 4.4 shows that such episodes are common. We therefore also take the unit to be an initial state, which gives m_c^*. Neither choice is assumption-free. When $m = 0$, however, no assumption is needed, because a one-element group admits no test.

Proof of Corollary 1. The harness ends an episode only at a predicate firing, at the horizon, or at a termination signal of the policy, and a failed episode has no predicate firing before its end. Without a termination channel every failed episode therefore ends at H, so $\varphi = 0$ and the failure duration support is the single point H. If no success lasts exactly H, every success ends strictly before every failure, so no success–failure pair has $T_{\text{success}} \geq T_{\text{failure}}$ and $b = 0$. The supports are then disjoint, and Proposition 1 applies. With $b = 0$ the smallest attainable p-value in Proposition 2 is $2^0 = 1$, which exceeds every $\delta < 1$. No evaluator can therefore be certified at any $\delta < 1$, at any sample size.

B Experimental details and pool registry

This appendix supports the collections of Section 4.1. Every result in the paper is computed on one of the collections in Table A1. The one exception is the window ablation of Appendix N, which has no entry here. Where a record does not identify a pool's policy or task, we say so rather than infer it. In the table, a dash means that the field does not apply to that pool, while `[not recorded]` means that the record does not contain it. The two were previously conflated and are now distinguished.

Table A1: **Every result in the paper is computed on one of these pools.** A dash means that the field does not apply to the pool, and `[not recorded]` means that the record does not contain it. The value n is per cell for S1, M1, C1 and T1 and a pool total elsewhere, and T1 masks the termination component. The content-unique column is a pool total except for M1 and C1, where it is listed per cell. Within W1 the duplication is concentrated on one condition, RT-1-X / move-near, which holds 80 distinct contents in 200 episodes. Q2.5 is Qwen2.5-VL-7B, Q3 is Qwen3-VL-Plus, G5.6 is GPT-5.6-sol and G4o is GPT-4o.

ID	Policy	Task scope	Protocol	n	dedup	Scorers	Used in
S1	Octo-B, Octo-S, RT-1-X	27 SimplerEnv	stop-on-success	40–50	1,076 / 1,200	Q3 (3 cells)	§4.2, App. D
S2	OpenVLA-OFT, 4 variants	10 LIBERO	stop-on-success	800	800	—	§4.2, App. D
S3	$\pi_{0.5}$	10 LIBERO	stop-on-success	200 / 204	204 / 204	—	§4.2, App. D
W1	Octo-B, Octo-S, RT-1-X	9 conditions	fixed-horizon	1,800	1,679 / 1,800	Q2.5/Q3/G4o	§4.3, §4.4, App. I
T1	RT-1-X	move-near, pick-coke	stop-on-success	40	[not recorded]	—	§4.2, App. E
P1	RT-1-X	pick-coke-can	stop-on-success	200	200	Q3	§4.3
P2	RT-1-X	pick-coke-can	fixed-horizon	200	200	Q3	§4.3
F1	RT-1-X	pick-coke, move-near	frame counterfactual	400	—	Q3	§4.3, App. F
M1	RT-1-X	3 Google-Robot tasks	stop-on-success	200	177/181/200	Q3/G5.6/G4o	§4.3, App. K
D1	[not recorded]	4 cells	[not recorded]	800	521 contents	Q3/G5.6/G4o	App. K, App. L
D2	[not recorded]	[not recorded]	[not recorded]	200	184	Q3/G5.6/G4o	§4.4, App. L
C1	OpenVLA, SpatialVLA	3 conditions	fixed-horizon	200	94/73/80	Q2.5/Q3/G4o	§4.4, App. J

Identification of D1, D2 and M1. Two pools are incompletely identified, and we mark them rather than guess. For D1 the record gives four cells of 200 episodes, with per-cell unique counts of 177, 181, 75 and 200. Those cell-level sets overlap. The 633 cell-unique episodes collapse to 521 distinct contents pool-wide, so 112 memberships repeat a content held by another cell. Three of the per-cell counts coincide with the three conditions of Appendix K. This is suggestive but is not an identification. The fourth cell holds 75 distinct contents in 200 episodes, and its count matches none of those conditions. It carries no certificate anywhere in this paper, and it is excluded on the same content criterion that we apply to every other pool. For D2 the record gives the deduplicated size and the four error rates, and no cell name. M1, in contrast, is identified. Its three conditions

are RT-1-X on Google-Robot tasks, and its 200-episode condition is the early-terminated pool P1 of Section 4.3. This is why the two share a clock score of 0.788. Every claim that rests on D1 or D2 is scoped to pools whose provenance we can describe but not name. Recovering those two is the most valuable item on our outstanding list, because together they carry the committed-verdict result of Appendix L.

Deduplication. Deterministic policies in finite-initial-state environments can produce byte-identical rollouts, so the registry records the content-unique count of every pool. Per-episode statistics on the stop-on-success pools scored for Appendices K and L are computed on the content-unique set. The fixed-horizon pool W1 is analyzed as collected. What content duplication touches is decidable rather than a matter of degree. Byte-identical episodes share an initial state, and m_{c}^* admits no two pairs that share one. A maximum matching therefore survives deduplication by substituting the retained duplicate. Hence m_{c}^*, and with it every admissibility verdict in Table A3, is invariant. We checked this on all nine RT-1-X cells, where deduplication moves m^* on three and m_{c}^* on none. The episode counts and the episode-disjoint count m^* are reported as collected. The census that carries Corollary 1 is unaffected for a second reason. Duplication in S1 is confined to the RT-1-X cells, and the 340 of 340 and 306 of 306 come from Octo cells whose contents are fully unique. The condition audited in Section 4.4 is content-unique.

C Evaluators and the certification ladder

This appendix supports the provenance rule of Section 4.1, the witness of Section 4.2, and the certification counts of Sections 4.3 and 4.4.

C.1 Scorers and score keys

Qwen2.5-VL-7B scores a fixed frame index and writes `p_hat_fixed_frame`. The served scorers score by log-probability, and more than one of them writes `p_hat_logprob`. The key therefore does not identify the model, and the pool-by-pool assignment of Table A1 does. Table A1 records the assignment pool by pool, and this section records each scorer's configuration and the caveats on its input. We name the scorer that carries every certification claim rather than a position in a panel. GPT-4o is descriptive throughout, in both directions. It has no episode-level ladder on the fixed-horizon pool W1. Qwen2.5-VL-7B was run at a deterministic rescaling to a 50,176-pixel cap, so its values are scoped to that implementation, and no invariance to native-resolution input is claimed.

For the served scorers, the input handling is recorded wherever the pool was scored under one of our frozen scoring work orders. Such a pool uses four frames at indices $[0, T-3, T-2, T-1]$ of the native 512×640 episode record, a single prompt frozen by hash, and greedy decoding. Its $\hat{p}$ is the softmax over the first-token logits of `Yes` and `No`. That covers S1 and D1. For the fixed-horizon pool W1, the coverage pool C1 and the committed-verdict batch D2, we could not establish from the records whether the same protocol was used. We list those three as unestablished and outstanding rather than assume it. GPT-4o is served through a third-party gateway, because the official endpoint returned a connection error. The fixed-horizon and coverage pools are scored by Qwen2.5-VL-7B, Qwen3-VL-Plus and GPT-4o. The early-terminated certification matrix of Appendix K and the committed-verdict batch of Appendix L are scored by Qwen3-VL-Plus, GPT-5.6-sol and GPT-4o.

Two further caveats apply. Qwen2.5-VL-7B's frame index is clamped on short episodes, so its input may encode duration. Repeat-pass stability is recorded for GPT-4o only. There a second pass through an independent gateway agrees on 97% of a 100-episode sample, at Spearman $\rho = 0.967$ and mean absolute score difference 0.046.

C.2 The certification ladder

A *rung* is one calibration setting λ. The ladder sweeps 26 rungs at $\delta = 0.1$, with $\lambda \in \{0.00, 0.02, \ldots, 0.50\}$, on each of 50 recalibration splits that draw $k_{\mathrm{split}} = 10$ calibration episodes apiece. At each rung the evaluator and the comparator are compared on cross-label pairs. The wins w and the losses ℓ are counted, and the exact one-sided p-value of Proposition 2 is taken. The sweep on a split runs from $\lambda = 0.50$ downward and stops at the first rung that fails, so a split certifies exactly when its first rung passes. A split is *degenerate* when its ten calibration episodes share one outcome,

so that the score cannot be recalibrated on it. The degenerate rate reaches 0.14 at the episode level and 0.18 on the collapsed ladder. The certification rate $N1$ is the fraction of non-degenerate splits that certify. It is therefore a ratio whose denominator varies by arm, and Table A6 gives it as k of K for every arm of both ladders. Per-rung counts are taken over the b clock-discordant pairs, while the floors of Section 3.2 govern admissibility. The two protocols test the same null but differ in what the comparator can use. Under stop-on-success the comparator preserves duration and realizes the clock, whereas under a fixed horizon the clock ties on every pair.

The p-value that an arm contributes to the Holm count of Section 4.3 is the median of its first-rung p-values over non-degenerate splits. When their number is even, it is the upper of the two central values. Merging the 18 arm-level values by Rüger's order-statistic rule (Rüger, 1978) uses $k_{\mathrm{arm}} = 9$. This is an order-statistic index and not a sample size like k_{split}, which is why we subscript the two. Ties are handled by the policy `all_cross_label_pairs_and_calibration_majority_null`. Where the comparator ties on a pair, the pair is retained and the null becomes the calibration majority. This makes the ladder well defined under fixed-horizon logging, where the clock ties on every pair, and it is why the comparator differs between protocols. A tie on the scorer's side is handled separately and identically in both ladders. An episode whose recalibrated score is exactly 0.5 is not committed and does not enter the counts. This is what *committed* means wherever we use it (Remark 2).

Two parameters, not one. The calibration setting swept over the rungs is λ. The margin that an evaluator must clear is γ, and the null is that it wins each discordant comparison with probability at most $\frac{1}{2} + \gamma$. The two parameters are distinct, and the paper keeps them so. The sweep starts at $\lambda = 0.50$ and stops at its first failure, so every certification is decided at $\lambda = 0.50$. The recorded λ^* is therefore 0.50 wherever it is defined, and no calibration setting is tuned. Every ladder count in the paper is at $\gamma_0 = 0$, where the null is chance, with one exception. Table A8 adds a column at $\gamma^* = 0.05$, a point of the grid $\{0, 0.05, 0.10, 0.15, 0.20\}$ fixed in advance. The choice of γ^* from that grid was frozen after rollouts pooled from two of the table's three conditions, scored by Qwen3-VL-Plus, had been examined. On those two Qwen3-VL-Plus rows the column is therefore a selected quantity, and the nominal exactness of the per-rung p-value does not carry over to it. On every other row, γ^* was fixed before the scores existed. We rest no conclusion on the column. Making it inferential would require a split-sample protocol, which we did not run, and the column is not used to order the scorers.

C.3 The witness under the ladder

The three witness conditions are spoon-on-towel, close-drawer and eggplant, with 50, 40 and 50 episodes. Each has a clock score of 1.000, so the judge's best AUROC already sits below what duration alone achieves. The complete ROC tables hold $50 + 40 + 48 = 138$ operating points. The ten-bin calibration tables and the confusion matrices at threshold 0.5 are identical. On spoon-on-towel, both scorers report Brier 0.129679, expected calibration error 0.151303, accuracy 0.86, true-positive rate 0.7083 and false-positive rate 0.0 (Figure 1b). Every episode is re-scored on two of the three conditions, and no condition draws the identity permutation.

Section 4.2 compares the judge and the witness field by field across the 26 rungs on its three conditions. Each rung has five fields, namely the accepted pairs, the wins w, the losses ℓ, the exact p and the verdict, on each of 149 non-degenerate recalibration splits. This gives $19{,}370$ recorded statistics in all, of which $13{,}218$ are equal and $6{,}152$ differ. Table A2 gives the full-precision AUROC values that Section 4.2 rounds.

Table A2: **The witness matches the judge to the last bit on three conditions.** The conditions are Octo-Small cells of S1, scored by Qwen3-VL-Plus. The witness permutes the judge's own scores within duration-defined classes, and both scorers attain the AUROC shown with identical ROC tables. The last column is the maximum per-episode score difference between the two scorers.

Condition	n	AUROC, judge and witness	ROC points	Max. score difference
spoon-on-towel	50	0.988782051282	50	0.9999
close-drawer	40	0.892583120205	40	0.9560
eggplant	50	0.995073891626	48	1.0000

C.4 Computing matched units and the matched-unit recomputation

Computing m^*. The discordance relation is a threshold on duration. The largest set of clock-discordant pairs that share no episode is therefore obtained by sorting both classes and matching each failure, in increasing duration, to the shortest unused success that is at least as long. We verified this against exact augmenting-path matching on 400 random instances, with no disagreement.

Acceptance-side recomputation. Recomputing on an initial-state-disjoint matching certifies 16 of 18 arms under Holm. It is not a filtered version of the 12 of Section 4.3, and we do not present it as one. The recomputation differs from the ladder in the unit and in the test. Within each matched success–failure unit, it asks whether the evaluator's raw score ranks the success above the failure, using an exact sign test against chance with ties dropped. The ladder instead asks whether recalibrated verdicts beat the calibration-majority comparator. No calibration setting enters the recomputation, so its 16 does not revise the ladder's 12. What it establishes is that 16 of 18 arms retain enough disjoint units to reach a rejectable exact p. The two arms that do not are both on RT-1-X / move-near, which supplies six units against 12 to 68 elsewhere.

C.5 Null replay

An end-to-end null replay calibrates a third procedure. It fixes the setting in advance at $\lambda_0 = 0.50$ and $\gamma_0 = 0$, merges an arm's 50 split-level p-values by Rüger's rule, and applies Holm across the 18 arms. The scores are replaced by uniform draws and by across-label permutations, at $1{,}000$ replicates each. The procedure then rejects 0 of $1{,}000$ family-wise at $\alpha = 0.1$ and 0 of $1{,}000$ on every arm. This is consistent with correct calibration and with heavy conservatism alike, and this replay cannot tell which. It calibrates neither the twelve nor the sixteen, which come from different procedures. Suppose instead that an arm could reject at any of the 26 settings, which that procedure does not allow. The per-arm rate would then reach at most 0.010 under uniform draws and 0.007 under permutation, against the nominal 0.1. Some arm would reject in 25 and 26 of $1{,}000$ replicates respectively.

D The census and all 47 cells

This appendix supports the census and the 47-cell screen of Section 4.2. The three diagnostics φ, the clock score and b follow from the outcome and duration columns alone. Under stop-on-success, successes last until they succeed, whereas failures run to H unless the policy stops them. In survival terms, the horizon right-censors the time to success, and the censoring indicator is the outcome label itself. This is why two columns of the log determine each other. The precondition of Proposition 1 holds in seven of our nine conditions. A screen that consumes only labels and durations therefore returns 0.0 there by deduction, not by prediction. The matched-unit count leaves these seven conditions unchanged, because $\varphi = 0$ forces $m_c^* = 0$ just as it forces $b = 0$. The side condition of Corollary 1 is checked rather than assumed. Across all 27 conditions, zero successful episodes end at the failure horizon.

All 47 cells. Over all 47 cells, the arithmetic divides into three layers that fail for different reasons, and only one layer is the mechanism of this paper. Table A3 lists every screened cell with both matched-unit counts. The ten OpenVLA-OFT rows each pool four perturbation variants of one checkpoint at 20 episodes per variant, which is not the one-policy-one-task condition of Section 3.1. Of the 47 screened *cells*, 37 are single-policy conditions and 10 pool four variants, and we use "cell" wherever the 47 denominator appears. The ten $\pi_{0.5}$ rows are the 204-episode audit pool. It is four episodes larger than the frozen 200-episode formal pool behind Section 4.2 and the second objection below, because one cell was collected at 24. The three failures, and hence the minority count, are identical in both. Forty-one cells have $m_c^* = 0$, and they divide into two kinds that should not be pooled. Twenty-seven contain both classes and separate perfectly on duration, which is the mechanism of this paper. Twenty-six of them involve policies that reserve no termination channel. The last one involves RT-1-X, which reserves a channel but still separates the outcomes perfectly on that task. Fourteen contain no episode of one class at all. Seven of them are $\pi_{0.5}$ rows with no failure, and three Octo-Base, two Octo-Small and two RT-1-X rows have no success. In these cells no cross-label pair exists, for reasons that have nothing to do with termination channels. No cell

here has $m^* > 0$ collapsing to $m_c^* = 0$. Two cells clear the floor. Both are RT-1-X cells at $m_c^* = 5$, against floors of four at $\delta = 0.1$ and five at $\delta = 0.05$. The admissibility call depends on the unit for one of the 47 cells. RT-1-X / carrot-on-plate passes at $m^* = 4$ and fails at $m_c^* = 2$, and we report the stricter verdict. Under the e-value floor of seven applied to the matched-unit count, none of the 47 cells clears. Applied to b, both RT-1-X cells clear it. All 20 LIBERO cells have $m_c^* = 0$, so no evaluator can be certified on any condition of the second simulator. This is a fact about the policies run on that benchmark rather than about the benchmark.

Table A3: **Only two of the 47 screened cells are admissible.** The count m^* counts clock-discordant pairs that share no episode, and m_c^* also forbids a shared initial state. Admissibility is $m_c^* \geq 4$ at $\delta = 0.1$. GR marks the Google-Robot tasks and bridge the Bridge tasks of SimplerEnv, n_+ and n_- count successes and failures, and *states* counts distinct initial states.

Policy / task	n	n_+	n_-	states	m^*	m_c^*	adm.
SimplerEnv							
Octo-B / GR / close-drawer	40	12	28	20	0	0	no
Octo-B / GR / move-near	40	1	39	10	0	0	no
Octo-B / GR / open-drawer	40	0	40	9	0	0	no
Octo-B / GR / pick-coke-can	40	13	27	40	0	0	no
Octo-B / GR / place-apple-in-closed-top-drawer	40	0	40	40	0	0	no
Octo-B / bridge / carrot-on-plate	50	7	43	21	0	0	no
Octo-B / bridge / eggplant	50	18	32	23	0	0	no
Octo-B / bridge / spoon-on-towel	50	9	41	21	0	0	no
Octo-B / bridge / stack-cube	50	0	50	20	0	0	no
Octo-S / GR / close-drawer	40	17	23	20	0	0	no
Octo-S / GR / move-near	40	4	36	11	0	0	no
Octo-S / GR / open-drawer	40	0	40	8	0	0	no
Octo-S / GR / pick-coke-can	40	14	26	40	0	0	no
Octo-S / GR / place-apple-in-closed-top-drawer	40	0	40	40	0	0	no
Octo-S / bridge / carrot-on-plate	50	5	45	21	0	0	no
Octo-S / bridge / eggplant	50	29	21	19	0	0	no
Octo-S / bridge / spoon-on-towel	50	24	26	23	0	0	no
Octo-S / bridge / stack-cube	50	1	49	22	0	0	no
RT-1-X / GR / close-drawer	40	36	4	18	1	1	no
RT-1-X / GR / move-near	40	16	24	10	9	5	yes
RT-1-X / GR / open-drawer	40	9	31	9	0	0	no
RT-1-X / GR / pick-coke-can	40	27	13	40	5	5	yes
RT-1-X / GR / place-apple-in-closed-top-drawer	40	9	31	40	3	3	no
RT-1-X / bridge / carrot-on-plate	50	4	46	22	4	2	no
RT-1-X / bridge / eggplant	50	0	50	22	0	0	no
RT-1-X / bridge / spoon-on-towel	50	3	47	20	3	1	no
RT-1-X / bridge / stack-cube	50	0	50	21	0	0	no
LIBERO							
OpenVLA-OFT / libero-10 / task00	80	49	31	20	0	0	no
OpenVLA-OFT / libero-goal / task00	80	50	30	20	0	0	no
OpenVLA-OFT / libero-goal / task01	80	50	30	20	0	0	no
OpenVLA-OFT / libero-object / task00	80	37	43	20	0	0	no
OpenVLA-OFT / libero-object / task01	80	30	50	20	0	0	no
OpenVLA-OFT / libero-object / task02	80	57	23	20	0	0	no
OpenVLA-OFT / libero-spatial / task00	80	35	45	20	0	0	no
OpenVLA-OFT / libero-spatial / task01	80	41	39	20	0	0	no
OpenVLA-OFT / libero-spatial / task02	80	47	33	20	0	0	no
OpenVLA-OFT / libero-spatial / task03	80	47	33	20	0	0	no
$\pi_{0.5}$ / libero-10 / task00	20	20	0	20	0	0	no
$\pi_{0.5}$ / libero-goal / task00	20	19	1	20	0	0	no
$\pi_{0.5}$ / libero-goal / task01	20	20	0	20	0	0	no
$\pi_{0.5}$ / libero-object / task00	20	20	0	20	0	0	no
$\pi_{0.5}$ / libero-object / task01	20	20	0	20	0	0	no
$\pi_{0.5}$ / libero-object / task02	20	20	0	20	0	0	no
$\pi_{0.5}$ / libero-spatial / task00	24	24	0	20	0	0	no
$\pi_{0.5}$ / libero-spatial / task01	20	20	0	20	0	0	no
$\pi_{0.5}$ / libero-spatial / task02	20	19	1	20	0	0	no
$\pi_{0.5}$ / libero-spatial / task03	20	19	1	20	0	0	no

The screen, condition by condition. Table A4 holds two columns that are *not* the same estimand, and they should not be read as a prediction against its own outcome. *Predicted* is an extrapolation of the committed-pair supply to $n = 200$, made from outcome labels and durations alone. *Realized* is the median of the per-rung supply across the certification ladder, not a count of pairs in one pool. On RT-1-X / pick-coke-can that median is 49.5, with interquartile range 41.5 to 56.0. On move-near it is 36.0, with interquartile range 31.0 to 40.0. Because it is a median, a realized supply can be a half-integer. Printed side by side, the two columns do disagree, and we print the size of the

disagreement rather than only the caveat. The screen was built to close a single condition. On pick-coke-can, a supply extrapolation of 29.0 met a realized median of 49.5, which has the same order of magnitude and lies far above the floor of four. The point estimates miss on both non-trivial rows, and in opposite directions. The prediction 29.0 with $[20.0, 39.0]$ met 49.5, and 47.0 with $[39.0, 56.0]$ met 36.0, each outside its own interval. The screen is assessed only on the binary admissibility call, which is correct nine times out of nine. The move-near row is a screening result, and no certificate in this paper rests on it.

The seven zero rows are not estimates. Where the policy reserves no termination channel, Corollary 1 forces $\varphi = 0$. The bootstrap interval is then degenerate because the quantity is determined, not because the sample is large.

Table A4: **The screen makes the correct admissibility call on all nine conditions.** The table compares the predicted committed-pair supply at $n = 200$ with the realized median over ladder rungs, beside the matched-unit counts m^* and m^*_c. The two supply columns are different estimands and are read as described above. The call does not change at $\delta = 0.05$.

			Committed-pair supply, $n = 200$				Admissible?	
Policy	Task	Clock	Predicted [5–95%]	Realized	m^*	m^*_c	δ=0.1	δ=0.05
Octo-Small	spoon-on-towel	1.000	0.0 [0.0, 0.0]	0.0	0	0	no	no
Octo-Small	close-drawer	1.000	0.0 [0.0, 0.0]	0.0	0	0	no	no
Octo-Small	eggplant	1.000	0.0 [0.0, 0.0]	0.0	0	0	no	no
Octo-Small	pick-coke-can	1.000	0.0 [0.0, 0.0]	0.0	0	0	no	no
Octo-Base	eggplant	1.000	0.0 [0.0, 0.0]	0.0	0	0	no	no
Octo-Base	pick-coke-can	1.000	0.0 [0.0, 0.0]	0.0	0	0	no	no
Octo-Base	close-drawer	1.000	0.0 [0.0, 0.0]	0.0	0	0	no	no
RT-1-X	move-near	0.818	47.0 [39.0, 56.0]	36.0	9	5	yes	yes
RT-1-X	pick-coke-can	0.885	29.0 [20.0, 39.0]	49.5	5	5	yes	yes

The harness assumption. The harness assumption is verified for SimplerEnv by construction and for LIBERO from the source. In the implementation we ran, the task environment overwrites the parent done with the success predicate before returning. A collision or joint-limit termination therefore cannot present as a success or end an episode on its own, and the horizon is enforced by the collection loop rather than by the environment. The observation agrees. None of the 360 LIBERO failures ended before its horizon, with 357 from the OpenVLA-OFT perturbation variants and three from $\pi_{0.5}$. The reading is scoped to that dependency version, since the collector equates done with success rather than calling the predicate itself. The premise of Corollary 1 concerns the harness. A platform that aborts on collision or operator intervention admits a second route to early failure. There φ is a lower bound, and the two verdicts diverge. The screen stays conservative for impossibility and becomes *anti*-conservative for the admissibility verdict that a practitioner would act on.

The two standard objections. The first objection is that the failure mode is an artifact of weak policies. A single OpenVLA-OFT checkpoint run under four action-perturbation variants spans a wide competence range. The four variants are the clean policy, two noise levels and a 25% action shuffle. The clean variant succeeds at 0.925 on its own 200-episode collection and at 0.915 within the 800-episode perturbation pool, where the four variants were merged. Pooled over ten task conditions and 800 episodes, the success rate is 0.55375, with per-condition base rates from 0.375 to 0.7125. This is the near-balanced regime that an evaluation set should most favor. Not one of the 357 failures ended before its horizon. In ten of ten conditions, the longest success is strictly shorter than the horizon at which every failure sits, so the duration-only rule ranks every pair correctly. Degrading one checkpoint four ways leaves the failure mode untouched, because the mode is a property of the action space and not of competence. The second objection is that the base rate is too extreme. The policy $\pi_{0.5}$ answers it from the opposite side. It succeeds on 197 of 200 episodes and so leaves three minority-outcome episodes against a floor of four. That pool is descriptive only and supports no claim about the skill of $\pi_{0.5}$. Balance does not help when every failure is a timeout, and near-perfection fails for want of failures.

E THE TERMINATION-MASK INTERVENTION

This appendix supports the causal claim of Section 4.2.

Design. The RT-1-X checkpoint, simulator, environment seeds, policy state and physical action dimensions are unchanged from the formal baseline run. At every step, the decoded `terminate_episode` component alone is forced to zero before the harness reads the termination signal, and the raw token is retained in the trace. Each cell has forty paired episodes, 80 in all. Every seed of the baseline run is matched in the masked run, and the match is verified programmatically. The checkpoint is `rt_1_x_tf_trained_for_002272480_step`. Its `saved_model.pb` hash and the baseline manifest hash are recorded in the release bundle.

What the mask changed. In the masked run, every one of the 34 remaining failures sits exactly at the horizon, while successes keep their spread.

What the mask suppressed. The mask suppressed 213 raw termination firings across 19 of the 80 episodes. These are 85 firings in 11 move-near episodes and 128 firings in 8 pick-coke-can episodes. Of the 80 pairs, 19 changed duration, for a total of $+747$ environment frames, and 3 changed label, all from failure to success. No label changed from success to failure.

The three label changes. Two label changes are move-near episodes 1 and 81, with seeds 716000001 and 716000081. Their durations go from 24 to 65 and from 25 to 74, and their first raw terminations occur at steps 23 and 24. The third is pick-coke-can episode 25, with seed 716000025, whose duration goes from 27 to 50 with a first firing at step 26. In each case the baseline episode ended at the policy's own termination signal, and the masked episode continued to a success.

Action-prefix audit. We compare the physical action vectors through the first raw termination prediction, or through the full baseline sequence where none was predicted. Prefixes are exact to 10^{-7} on 37 of 40 move-near pairs and 35 of 40 pick-coke-can pairs, which is 72 of 80 overall. All three label-changing pairs are exact, at maximum absolute difference 0.0. The outcome changes are therefore attributable to the suppressed token and not to a diverged trajectory. The remaining eight pairs are not bit-exact, so duration changes outside the label-changing set should not all be read as pure token effects.

A surrogate in the other direction. We cannot give Octo a termination channel, so we approximate one on the harness side and report the result as an illustration. Each failed episode is truncated at the first point where the full simulator state has been static for five steps. Across the eighteen stop-on-success Octo cells of the census, this lifts φ from 0 to 0.085 over 646 failures. The initial-state-disjoint matched-unit count rises from 0 to between 1 and 5 on five of the eighteen, and two of those reach the floor of four. We do not pool this count across these heterogeneous cells. Three limits come with the run, and none is incidental. The stopper reads simulator state that a policy would not have, so it is neither an object-relative termination nor a policy one could deploy. The episodes are truncated post hoc rather than re-collected, so this is not a second collection. The stopper is also a heuristic rather than a competent stop, so the result shows that the budget moves with φ, not what a trained termination head would do.

F CORRECTION AGAINST COLLECTION

This appendix supports the correction analysis of Section 4.3. The natural response to the confound is to keep the data and repair the estimate, and length-matched evaluation is the first repair a reader would try. In Figure A1, panel (a) is the estimate ladder on RT-1-X / pick-coke-can, one of the two conditions collected under both protocols. Panel (b) plots one point per episode for the 400 episodes of the frame counterfactual. The episodes are held fixed and only the frames that the judge observes change, so the shift is attributable to the terminal observation rather than to the clock. The early-terminated pick-coke-can pool of panel (a) is a different collection from the $n = 40$ screening pool of Table 1. This is why its clock score is 0.78788, against 0.885 in that table.

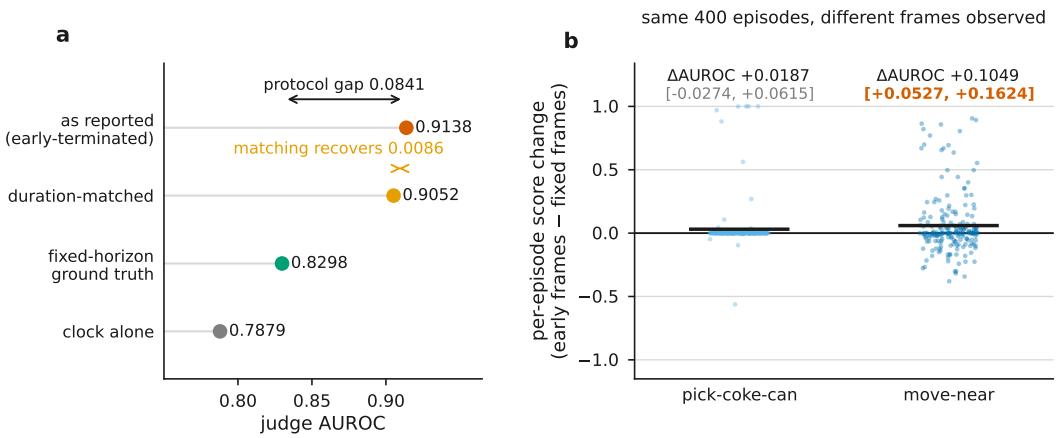

Figure A1: **Duration matching does not recover the fixed-horizon estimate.** **(a)** RT-1-X on pick-coke-can was collected under both protocols, with content-unique episodes at $n = 200$ each. Matching the early-terminated pool on duration moves the estimate by 0.00863, about a tenth of the 0.08407 gap to the fixed-horizon reference of 0.82977, which the panel labels ground truth and prints to four digits. **(b)** The 400 fixed-horizon episodes of RT-1-X on pick-coke-can and move-near are each scored twice with only the observed frames changed. Each point is one episode's score change, and the bar is the mean. Judge AUROC moves by $+0.0187$ on pick-coke-can, with an interval containing zero, and by $+0.1049$ on move-near, where the interval excludes zero. Early termination perturbs the terminal observation as well as the clock, so a correction on the clock addresses one channel of two.

Three further details qualify the pick-coke-can contrast. The fixed-horizon reference is taken with the duration column held constant rather than with duration removed. The two collections share no initial state, so they are two independent draws under the two rules rather than an intervention on one. The frame counterfactual holds the episodes themselves fixed, so it does not rest on that two-draw contrast. Two conditions are not a replication study. The two paired conditions agree only on the sign and the order of magnitude of what matching leaves behind. A control rules out the alternative explanation that the pools differ in their success definition, since switching between achievement and terminal labels changes judge AUROC by 0.011 and 0.018, an order of magnitude below the protocol gap.

All five digits are shown because the differences do not reproduce from four.

The second paired condition is RT-1-X / move-near. There the judge scores 0.91935 as reported, 0.90952 matched and 0.64897 as the fixed-horizon reference. Matching therefore removes 0.00982 of a 0.27038 gap, which is 3.6 percent, with the differences taken before rounding. The two move-near collections share twelve initial states, of which one yields a byte-identical action prefix. The early-terminated move-near pool holds 75 distinct contents in 200 episodes. Like its fixed-horizon counterpart in W1, it is analyzed as collected.

G FIXED-HORIZON DETAILS

This appendix supports the evidence counts of Section 4.3 and gives the detail behind Table 1. Fixed-horizon logging follows from Corollary 1, whose premise concerns the harness rather than the policy. We cannot give Octo a termination channel, but we can remove the rule that makes its absence fatal. Under stop-on-success, the table reports four numbers for each condition. They are φ, the clock score, the matched-unit count m_c^* and the admissibility verdict. Once the horizon is fixed, the minority count and the reversal fraction replace the clock. *Minority* is the smaller outcome class, which is the evidence that a fixed horizon supplies in place of the clock. That is an episode count, not on the same scale as the committed-pair supply, and the two should not be differenced. The two nonzero values of φ come from the baseline run of Figure 3. The seven zeros are forced by Corollary 1, not estimated. The admissibility column is a screening verdict on the evidence a condition supplies. It is not a statement that any scorer was certified on that condition.

In Figure A2, panel (a) is the clock score for the same nine conditions under the two termination rules, and panel (b) is the minority-outcome evidence available afterwards. The panels count different things, and their axes are not comparable. Octo-Base / pick-coke-can supplies 29 minority-outcome episodes and so misses by one the collection-side gate of 30 that we used when admitting cells. We retain it and mark it in Table 1. Retaining it enlarges the Holm family of Section 4.3 from sixteen arms to eighteen and can only make that correction stricter. Neither of its two arms is among the twelve that reject. Duration is constant across W1 with zero alignment failures. On the Octo conditions, stop-on-success had removed the ability to show that an evaluator used the signal, not the signal itself. A remedy that makes the benchmark unusable does not solve the problem, and fixed-horizon logging keeps the benchmark usable.

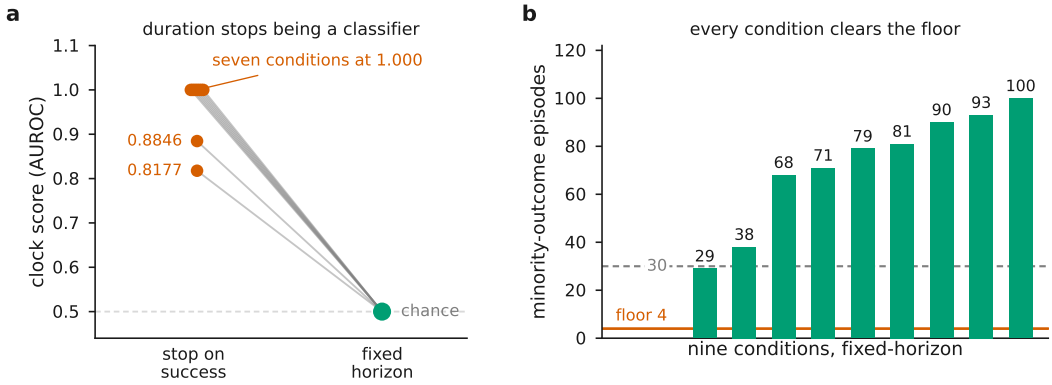

Figure A2: **Collection repairs the evidence at unchanged rollout count. (a)** The panel shows the clock score of the same nine policy–task conditions under the two termination rules. Seven leave a perfect 1.000, two leave 0.8177 and 0.8846, and all nine arrive at chance. **(b)** After fixed-horizon logging, every condition supplies minority-outcome evidence, from 29 to 100 episodes, against the floor of four. The dashed line at 30 is the collection-side gate we used when admitting cells. One condition misses it by one episode, and it is retained and marked in Table 1. Panel (b) counts episodes and is not on the same scale as the committed-pair supply of Table A4.

H ACHIEVEMENT AGAINST TERMINAL SUCCESS

This appendix supports the label comparison of Section 4.3 and the audit of Section 4.4.

Row selection for Figure 4. The frames are as recorded during the fixed-horizon collection of Octo-Small on pick-coke-can, with $H = 80$. An open-loop control reproduces all 81 frames of each episode bitwise. The two rows were selected arithmetically before any frame was viewed. Row 49 is the reversal at the median post-firing gap, and row 74 is the terminal success matched on t^*. The two rows therefore differ in outcome and not in when the predicate fired. Under stop-on-success, neither row would have been recorded past t^*.

Both labels for all nine conditions. Table A5 gives both labels for all nine fixed-horizon conditions. *Fired* counts episodes in which the predicate fired at any point, which is recoverable from the retained first-firing index. *Terminal* counts those still satisfying it at the horizon. Under stop-on-success only the first is observable, and the two coincide by construction. Which label a benchmark should report is a question of task semantics, on which we take no position. Our claim concerns which protocol can distinguish the two labels.

I CERTIFICATION UNDER CLUSTER-LEVEL SAMPLING

This appendix supports the clustering result of Section 4.4.

The nine fixed-horizon conditions hold 287 distinct initial states across 1,800 episodes. Initial states are shared across conditions of the same task and counted once pool-wide. This is why three conditions of 200 distinct states each are compatible with 287 in total. The per-condition counts are

Table A5: **The reported success rate measures achievement, not success.** The table covers all nine fixed-horizon conditions, with $n = 200$ each. *Fired* counts episodes in which the success predicate fired at any point, and *terminal* counts those still satisfying it at the horizon. Reversal reaches 13.5% of episodes on one condition and inflates the reported rate by up to 1.711× relative, 1.128× overall. Policy rankings are preserved under both labeling rules on all three task families. On pick-coke-can, the measured gap between the two Octo policies is 11.5 points under achievement labels and 4.5 points under terminal labels, a factor of 2.6.

Policy	Task	Fired	Terminal	Rev.	Rev. frac.	Achiev. rate	Term. rate
Octo-Base	close-drawer	85	81	4	0.020	0.425	0.405
Octo-Base	pick-coke-can	42	29	13	0.065	0.210	0.145
Octo-Base	eggplant	92	90	2	0.010	0.460	0.450
Octo-Small	close-drawer	72	71	1	0.005	0.360	0.355
Octo-Small	pick-coke-can	65	38	27	0.135	0.325	0.190
Octo-Small	eggplant	110	107	3	0.015	0.550	0.535
Octo-Small	spoon-on-towel	115	100	15	0.075	0.575	0.500
RT-1-X	move-near	90	79	11	0.055	0.450	0.395
RT-1-X	pick-coke-can	149	132	17	0.085	0.745	0.660
Total, 1,800 episodes		820	727	93	0.052	0.456	0.404

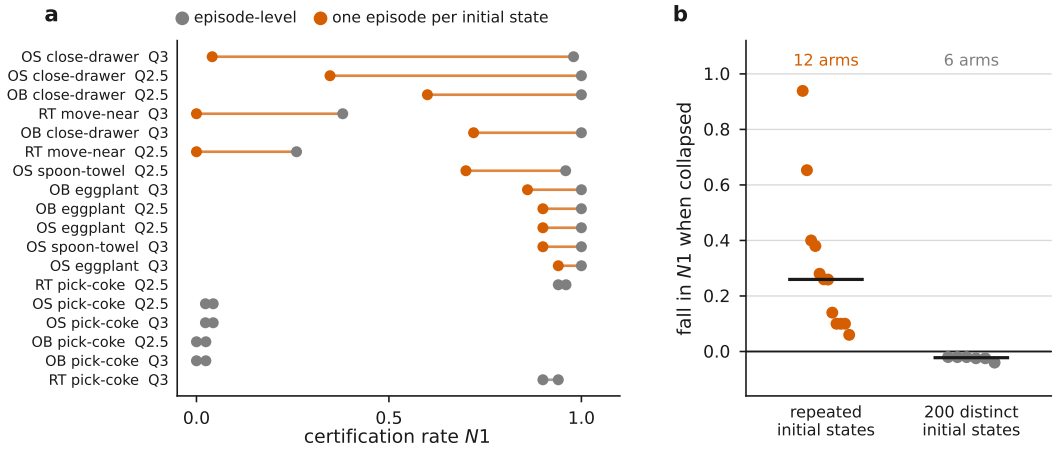

Figure A3: **The positive result is smaller under cluster-level sampling. (a)** The panel shows the certification rate $N1$ for the 18 arms with an episode-level ladder, at the episode level (gray) and after collapsing to one episode per initial state (vermilion where the condition has repeated initial states). **(b)** The panel shows the fall in $N1$ on collapsing, split by whether the condition repeats initial states. Gray is the episode level and vermilion the collapsed one, and conditions with repeated initial states are marked. Labels abbreviate Octo-Small (OS), Octo-Base (OB), RT-1-X (RT), Qwen2.5-VL-7B (Q2.5) and Qwen3-VL-Plus (Q3).

200, 200, 200, 27, 27, 24, 24, 24 and 12. On the three conditions with 200 distinct states apiece, collapsing is a no-op. On the other six, within-condition cluster sizes reach 27 and the design effect reaches 6.636, with a worst effective sample size of 30.1 of 200. Collapsing and re-running the ladder lowers $N1$ in all twelve arms on those six conditions. It falls most sharply from 1.000 to 0.347 and from 0.980 to 0.041, and it falls to 0.000 on both arms of RT-1-X / move-near. Taking $N1 > 0.5$ as the decision rule, 12 of 18 arms certify at the episode level and 10 of 18 after collapsing. Cluster bootstrap intervals over 4,000 resamples accompany every arm in the source artifact. Table A6 gives every $N1$ in the figure as k of K.

The exposure is specific to the ladder's unit. The matched-unit recomputation of Appendix C.4 is initial-state-disjoint by construction and certifies 16 of 18 arms. Fourteen of them fall on the seven conditions whose stop-on-success clock is 1.000 and whose pair supply is zero. The three counts come from three procedures and are not three estimates of one quantity. The two failures of the recomputation are the two arms of RT-1-X / move-near. That condition draws its 200 episodes from twelve initial states and 80 distinct contents, and it supplies six matched units against twelve to

sixty-eight elsewhere. It is the one condition whose two arms fail under both recomputations, and the other eight are not in that position.

Table A6: **Collapsing to one episode per initial state lowers the counts behind Figure A3.** For each of the 18 arms, $N1 = k/K$, where K is the number of non-degenerate recalibration splits out of 50 and k the number that certify. K is shared by the two scorers of a condition. Q2.5 is Qwen2.5-VL-7B and Q3 is Qwen3-VL-Plus.

		Episode level			One episode per initial state		
Policy	Task	K	k, Q2.5	k, Q3	K	k, Q2.5	k, Q3
Octo-Small	spoon-on-towel	49	47	49	50	35	45
Octo-Small	close-drawer	49	49	48	49	17	2
Octo-Small	eggplant	50	50	50	50	45	47
Octo-Small	pick-coke-can	43	1	1	46	2	2
Octo-Base	eggplant	50	50	50	50	45	43
Octo-Base	pick-coke-can	44	0	0	41	1	1
Octo-Base	close-drawer	50	50	50	50	30	36
RT-1-X	move-near	50	13	19	50	0	0
RT-1-X	pick-coke-can	50	47	45	50	48	47

One reading caution. On the three conditions that already hold 200 distinct states, collapsing is a no-op on the pool, yet the two ladders differ by up to 0.04. They come from separate artifacts, whose calibration splits are drawn under two different recorded seed schemes. One scheme is keyed to the cell index and the other to its sorted position. Even their non-degenerate-split denominators therefore differ on two of those three conditions (Table A6). The gap is a method difference and not unseeded randomness, because each ladder is reproducible from its own recorded seed. A run-to-run standard deviation of $N1$ over S seeds remains outstanding, since each ladder has a single recorded run. Differences of this size between the two columns therefore carry no meaning, including in Table A8 and Figure A3.

J TWO POLICY FAMILIES UNDER THE RECOMMENDED PROTOCOL

This appendix supports the coverage result of Section 4.4, which is wholly negative. All nine arms fall below the derived clock, by -0.0178 at the closest and -0.4647 at the widest, on conditions selected before any scoring on the two criteria that most favor a judge.

Three conditions were collected fresh for OpenVLA and SpatialVLA under fixed-horizon logging, at $n = 200$ each and score-blind, on seeds drawn independently of every other pool in this paper. Under the byte-level criterion that we apply to the SimplerEnv pools, the per-cell content-unique counts are 94, 73 and 80 of 200. The dedup field recorded at collection time hashed the episode directory tree rather than the episode contents. It therefore certified uniqueness that it had not tested, and we correct it here. These conditions are themselves fixed-horizon SimplerEnv collections, and they duplicate content for the same reason our own pools do. The AUROC values below are computed on the 200 rows as collected, and the bootstrap intervals resample those rows. On a content-unique sample the intervals would be wider.

The conditions were chosen before any scoring, on the two criteria that most favor a judge. The first is a balanced terminal success rate in $[0.25, 0.75]$ at selection time. The second is a derived clock score strictly below 1.000, so that discordant pairs exist at all. A negative result obtained under selection that favors the alternative is harder to explain away than one obtained under neutral selection, which is why the selection rule is stated here rather than left implicit. The band was applied to the selection-time estimate, where the SpatialVLA condition sat at 0.6333. On the re-collected $n = 200$ it realized 0.755, just outside the band, which we report rather than re-select.

The comparator is the *derived* clock. It is the counterfactual duration that these same episodes would have handed a judge under stop-on-success, reconstructed from the retained predicate trace. On the episodes as actually collected, the duration is constant and the realized clock score is 0.500. The margins below are therefore not a comparison against anything observable in the fixed-horizon pool.

Three scoping conditions bound what Table A7 can support. First, the results are descriptive, since no family-wise correction is applied and no certification is claimed on these cells. Second, the Qwen2.5-VL-7B values are scoped to its 50,176-pixel implementation (Appendix C.1). Third, three conditions across nine arms support no population-level claim in either direction. What they show is that on two further policy families no judge beats a clock derived from the predicate trace. Whether those families exhibit $\varphi = 0$ is *not* tested here. C1 is fixed-horizon only, so no failure in it can end early. The selection rule also required a derived clock strictly below 1.000, which is the regime in which duration does *not* separate perfectly.

Table A7: **The same result holds on two policy families we had not studied.** The three conditions were collected fresh under fixed-horizon logging at $n = 200$ each. No judge exceeds the derived clock in any of the nine arms. That clock is not observable in this pool, where the realized clock is 0.500. Brackets are 95% bootstrap intervals over $5{,}000$ resamples. The selection rule and the scoping conditions are stated above.

Scorer	AUROC [95% CI]	Margin vs. derived clock
OpenVLA / close-drawer (terminal success rate 0.580, derived clock 0.988814)		
Qwen2.5-VL-7B (open-weight)	0.908405 [0.854797, 0.954947]	-0.080408
Qwen3-VL-Plus (vendor API)	0.925954 [0.888590, 0.956566]	-0.062859
GPT-4o (gateway)	0.524066 [0.445040, 0.605296]	-0.464748
OpenVLA / move-near (terminal success rate 0.575, derived clock 0.869156)		
Qwen2.5-VL-7B (open-weight)	0.681586 [0.600957, 0.756378]	-0.187570
Qwen3-VL-Plus (vendor API)	0.448338 [0.367252, 0.528846]	-0.420818
GPT-4o (gateway)	0.550997 [0.471058, 0.629923]	-0.318159
SpatialVLA / move-near (terminal success rate 0.755, derived clock 0.811056)		
Qwen2.5-VL-7B (open-weight)	0.442289 [0.351464, 0.530885]	-0.368766
Qwen3-VL-Plus (vendor API)	0.793215 [0.723040, 0.855028]	-0.017840
GPT-4o (gateway)	0.688066 [0.601072, 0.772044]	-0.122990

K THE CERTIFICATION MATRIX

This appendix supports the statement of Section 4.3 that margin over the clock, and not raw AUROC, separates the scorers on the certification matrix. Qwen3-VL-Plus has the highest raw AUROC on all three conditions, and it clears its condition's clock by $+0.077$, $+0.058$ and $+0.126$. GPT-5.6-sol and GPT-4o sit below it on two of three conditions each, with margins from -0.113 to $+0.026$ that are negative in four of their six arms. Under length control, the ordering reverses on the 177-episode condition. There Qwen3-VL-Plus is the lowest of the three (0.803 against 0.807 and 0.842), and all three sit below that condition's clock of 0.864. This panel is not the one that scores the nine fixed-horizon conditions. It shares Qwen3-VL-Plus and GPT-4o with them but not Qwen2.5-VL-7B, and it has GPT-5.6-sol in its place (Table A1). Nothing here therefore transfers across pools by position in a panel.

Table A8 covers three early-terminated, content-deduplicated conditions, each scored by all three scorers. *Raw AUROC* is the scorer against the outcome label. *LC AUROC* is the same quantity after length control, computed on the duration-matched subset of Section 4.3, so that the two columns differ only in whether duration has been held fixed. $N1$ is the fraction of non-degenerate recalibration splits that certify against the clock, at margin $\gamma_0 = 0$ and at $\gamma^* = 0.05$. On the two Qwen3-VL-Plus rows of the first two conditions, the γ^* column is a selected quantity (Appendix C.2), and no conclusion rests on that column anywhere. As k of K splits in the row order of the table, the γ_0 column reads 39, 1 and 7 of 50, then 32, 10 and 4 of 50, and then 40 of 50, 22 of 49 and 20 of 49. The γ^* column reads 37, 1 and 5 of 50, then 26, 6 and 1 of 50, and then 36 of 50, 16 of 49 and 14 of 49. GPT-5.6-sol and GPT-4o draw their splits under a different recorded seed from Qwen3-VL-Plus, and on the 200-episode condition one of their 50 splits is degenerate. With the counts printed, the rows can be read as k of K. We still give the comparison in the main text only as 3 of 3 against 0 of 3 within a judge, and never as a ranking of $N1$ values across rows. Each certificate is relative to its own condition's clock, and the scorers' splits are not the same draws. These three conditions form pool M1, a different collection from our nine conditions, and their clock scores are 0.864, 0.876 and 0.788. The third has $n = 200$ and a clock of 0.788 because it is pool P1 of Section 4.3 (Appendix B).

Table A8: **Certification, not agreement, is what separates scorers.** The three early-terminated, content-deduplicated conditions are RT-1-X on Google-Robot tasks (pool M1, Appendix B), with three scorers each. $N1$ is the certification rate against the clock at margin $\gamma_0 = 0$ and at $\gamma^* = 0.05$, and the γ^* column is carried for completeness only. Qwen3-VL-Plus certifies on 3 of 3 conditions, GPT-5.6-sol on 0 of 3, and GPT-4o, excluded from our claims by the provenance rule, on 0 of 3. Their AUROC columns overlap. On the 200-episode condition GPT-4o's raw value of 0.800 sits just above the clock's 0.788, and under length control Qwen3-VL-Plus is the lowest of the three on the 177-episode condition. The text gives every $N1$ as k of K, and the column is not a cross-row ranking.

n	Scorer	$N1$ at γ_0	$N1$ at γ^*	Raw AUROC	LC AUROC	Clock
177	Qwen3-VL-Plus (vendor API)	**0.780**	0.740	0.941	0.803	0.864
177	GPT-5.6-sol (vendor API)	0.020	0.020	0.751	0.807	0.864
177	GPT-4o (gateway)	0.140	0.100	0.798	0.842	0.864
181	Qwen3-VL-Plus (vendor API)	**0.640**	0.520	0.934	0.865	0.876
181	GPT-5.6-sol (vendor API)	0.200	0.120	0.791	0.857	0.876
181	GPT-4o (gateway)	0.080	0.020	0.809	0.833	0.876
200	Qwen3-VL-Plus (vendor API)	**0.800**	0.720	0.914	0.905	0.788
200	GPT-5.6-sol (vendor API)	0.449	0.327	0.814	0.801	0.788
200	GPT-4o (gateway)	0.408	0.286	0.800	0.809	0.788

Three disclosures belong with this table. First, it applies no multiplicity correction. If the arms were independent, the chance that at least one of a null arms certifies at $\delta = 0.1$ would be $1 - 0.9^a$, which is 0.41 at five arms and 0.61 at the nine shown. The matrix is therefore a description and not a certified ranking. Second, the 177- and 181-episode conditions are not independent collections. The second was collected as a replication of the first and shares 27 contents with it. Third, on the committed-verdict batch of Appendix L the GPT-5.6-sol column failed a pre-registered parse-rate gate, at 0.025 against a 0.02 ceiling. The error rate we report for it there is a counterfactual computed under a relaxation that we did not adopt.

One subset of that batch runs against the reading of this table. Consider the 34 episodes whose durations are atypical for their label. On them the clock is wrong, so by our own account they are the only episodes on which the question is decided. There the uncertified GPT-5.6-sol errs at 0.147 against 0.294 for the certified Qwen3-VL-Plus, discordant six to one. On the batch as a whole the two stand at parity, 0.2389 against 0.2391, with the GPT-5.6-sol column gate-blocked and descriptive. On pool D1, which holds 800 episodes with 521 distinct contents, the three scorers err on the content-unique set at 0.19 (Qwen3-VL-Plus), 0.2668 (GPT-5.6-sol) and 0.2937 (GPT-4o). The paired advantage of Qwen3-VL-Plus there is 0.1036 over GPT-4o ($p = 3.84 \times 10^{-8}$) and 0.0768 over GPT-5.6-sol ($p = 4.65 \times 10^{-5}$). On all 800 episodes the latter is 0.08875, with interval $[0.0575, 0.11875]$ and $p = 1.27 \times 10^{-8}$. Three annotations bind these figures. First, the pool is not blind. Its per-cell counts match the three conditions of Table A8, so we treat its conditions and episodes as having taken part in the certification of Qwen3-VL-Plus, and the advantage inherits that selection. Second, the three score columns were collected in different time windows. Third, the parity on the fresh batch is the counterpoint to the second difference. We report all of this for two reasons. Certification is a statement about what an evaluation set can license, not a guarantee about downstream error. Our own framework also says that the length-atypical subset is where that distinction is tested.

L THE COMMITTED-VERDICT CONSEQUENCE

This appendix supports the scoped committed-verdict statement of Section 4.4. An evaluator's output is ultimately used as a decision, since a checkpoint is kept or a policy is reported as better. The relevant error is therefore a thresholded verdict rather than a ranking. We measure committed-verdict error on a fresh batch of 184 deduplicated episodes, with 104 successes and 80 failures, on which the majority-class predictor errs at 0.4348. This batch and the pool on which the rule is fit are the two collections whose policy, task and protocol the registry marks as not recorded. The result is therefore scoped to pools we can describe but not name, and we keep it out of the abstract and the contributions until that is repaired (Appendix B). Figure A4 gives the batch in full.

Qwen3-VL-Plus errs at 0.2391 on all 184 episodes. GPT-5.6-sol errs at 0.2389 on the 180 it scored, but its column failed the pre-registered parse-rate gate, so that value is descriptive and not a result. Counting its four unscored episodes as all right or all wrong would place it in $[0.2337, 0.2554]$. GPT-4o errs at 0.3665 on the 161 episodes it scored in the first pass. The other 23 were transmission failures. They were recovered by a transport-only retry, under a rule frozen in advance, that re-sent the identical requests. After the retry GPT-4o errs at 0.3804 on all 184, which is the value that the four tiers below use. All three scorers beat the trivial baseline. Since the denominators differ, the paired comparison on a common subset is the appropriate comparison.

A one-parameter rule on episode duration, in which $T \leq 66$ predicts success, was fit once on a separate 800-episode pool and transferred without adaptation. It errs at 0.1848 on all 184 episodes, with AUROC 0.8694, against an in-sample best achievable error of 0.1576. The batch is held out by collection but not by content. Of its 184 episodes, 52 are byte-identical to episodes in the fit pool and 132 are not. A single threshold fit on 800 episodes is not in a position to memorize 52 of them. The overlap still belongs in the text rather than behind a word like separate, because it carries the result. The rule errs at 0.0962 on the 52 recurrent-content episodes and at 0.2197 on the 132 unseen ones. On those 132, Qwen3-VL-Plus errs at 0.1742 and GPT-4o at 0.3636, and GPT-5.6-sol errs at 0.2403 on the 129 of them it returned a score for. Both decompositions close exactly on the batch. The rule makes $29 + 5 = 34$ errors, and Qwen3-VL-Plus makes $23 + 21 = 44$, out of 184. On the common subsets the rule errs at 0.1889 on the 180 and 0.1925 on the 161. The comparison that we call appropriate has therefore been run. On each scorer's own episodes, the transferred rule errs less than that scorer does. After stratifying by content, however, the rule no longer errs less than every scorer. We therefore draw the weaker conclusion that Section 4.4 states.

A relative comparison among the judges is also available, and it does not hold between the two scorer columns that we ran ourselves. On the 161 episodes that both scored, Qwen3-VL-Plus improves on GPT-4o by 0.1429, with interval $[0.068, 0.217]$. On the 180 episodes that both scored, Qwen3-VL-Plus improves on GPT-5.6-sol by 0.0000, with interval $[-0.078, 0.083]$. This second comparison exists only by admitting the gate-blocked GPT-5.6-sol column, a relaxation that we did not adopt. The whole apparent benefit of choosing a certified scorer is therefore Qwen3-VL-Plus against the gateway-served GPT-4o, which the provenance rule of Section 4.1 keeps out of our claims. The comparison that carries weight is the one against the duration rule, which neither self-run scorer beats on the batch as a whole. On the 132 unseen episodes the ordering is not uniform. Qwen3-VL-Plus, at 0.1742, beats the rule's 0.2197, while GPT-5.6-sol at 0.2403 and GPT-4o at 0.3636 do not.

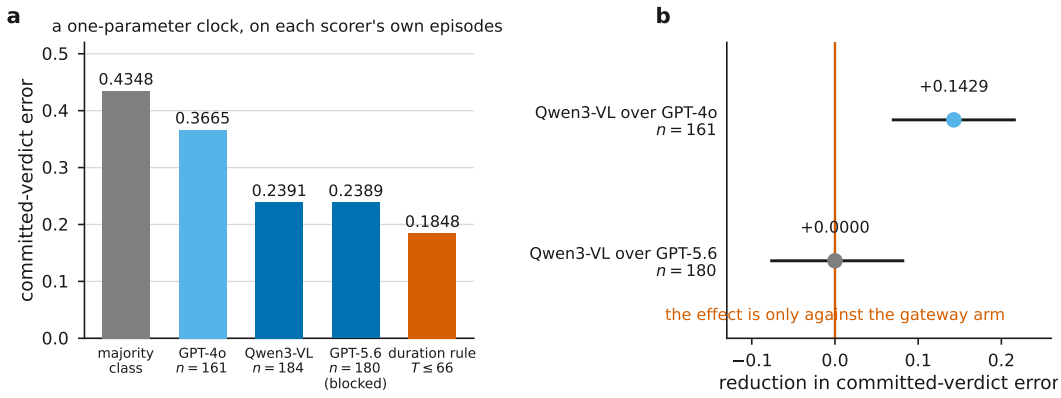

Figure A4: **The committed-verdict consequence is scoped to pools we can describe but not name.** **(a)** On a 184-episode batch, a transferred one-parameter duration rule errs less than each scorer does on the episodes that scorer scored. The three scorer denominators are 184, 180 and 161, so the bars are not a common-subset comparison, and the GPT-5.6-sol bar is its gate-blocked column. **(b)** The panel shows two paired contrasts on the common subsets, Qwen3-VL-Plus over GPT-4o ($n = 161$) and over GPT-5.6-sol ($n = 180$). The GPT-4o contrast is shown for completeness, and the provenance rule of Section 4.1 keeps it out of our claims. The GPT-5.6-sol contrast, the one we ran ourselves end to end, is 0.0000 and exists only by admitting that scorer's gate-blocked column, a relaxation we did not adopt. The panels label Qwen3-VL-Plus as Qwen3-VL and GPT-5.6-sol as GPT-5.6.

The archive also records a four-tier decomposition, by content novelty, of the difference between Qwen3-VL-Plus and GPT-4o. The four tiers are to be read together. The difference is $\Delta = +0.1500$ on all 200 episodes ($p = 1.19 \times 10^{-5}$). It is $\Delta = +0.1413$ on the 184 deduplicated episodes, with 95% interval $[0.0707, 0.2120]$ and $p = 7.82 \times 10^{-5}$. This is the citable primary result, although the resampling unit, number of draws and seed behind that interval were not recorded with it. It is $\Delta = +0.1894$ on the 132 episodes whose content had not been seen before ($p = 3.51 \times 10^{-5}$). Finally, it is $\Delta = +0.0192$ on the 52 recurrent-content episodes, where the two error rates are 0.4038 and 0.4231 and $p = 0.5$. The concentration on unseen content is real and interpretable. Every tier, however, is a comparison of Qwen3-VL-Plus against GPT-4o. The whole decomposition therefore inherits the provenance of the gateway-served scorer and supports no claim in this paper. It is reported here for completeness only. The transferred duration rule errs at 0.1848, below all four committed-verdict error rates of Figure A4(a). These are the majority-class predictor and each of the three scorers on the episodes it scored.

M THE FULL REPORTING CONTRACT

This appendix supports the recommendations of Section 5. Table A9 is a reporting contract, not a claim about which quantity a benchmark ought to target. Every row names a field that the harness already computes or could compute without new rollouts. The change is in what gets written down rather than in what gets run. The clock score is on the same scale as the evaluator AUROC and should be printed beside it.

Table A9: **The reporting contract needs no new rollouts.** Every row names a field that the harness already computes or could compute without new rollouts, so the change is in what gets written down. The cost is two columns per episode, an outcome label and a duration.

Desideratum	Current practice	Recommendation
Termination	Ends at the first predicate firing	Run to the horizon where safe; keep t^* and the terminal state
Failure mode	Not reported	Check the model card before collection; report φ, since zero means no evaluator can be validated here
Evaluator warrant	AUROC against ground truth	The same statistic beside the clock score, and the margin over it
Sample size	The episode count n	The matched-unit count; withhold agreement below $\lceil \log_2(1/\delta) \rceil$
Repeated episodes	Distinct seeds assumed distinct	Content-level dedup at collection; report effective n
Success	The predicate fired at some point	Both labels with the reversal fraction; a predicate that is state-free at H, or its published state dependence

Two rows deserve a word. The *success* row asks for both labels and the reversal fraction, because a protocol that records only one makes the choice unrecoverable. It takes no position on which label is the right target, which is a question about the task and not about logging. The *sample size* row asks for the matched-unit evidence count rather than the episode count. The episode count does not bound what a certification test can conclude, and the matched count does.

The cost is two columns per episode, an outcome label and a duration. It needs no model, no GPU and no change to the number of rollouts collected. The one thing the table asks for that is not free is running every episode to the horizon. Its simulator-step overhead falls on the successful episodes that now run past t^*, and not on the failures, which already ran to H wherever $\varphi = 0$. We have not measured this overhead, and its safety implications on hardware are discussed in Section 6.

N THE WINDOW ABLATION

This appendix records the window ablation mentioned in Section 4.1, which has no entry in the pool registry. It uses three LIBERO evaluation folds, two window variants (terminal and prefix), three seeds (42, 43 and 44), $K_{\mathrm{abl}} = 64$ and 3,000 steps, for 18 rows in total. The setting K_{abl} belongs to this ablation only and is unrelated to the split count K of Appendix C. Each fold holds out one policy's 200 episodes and trains on the other two. A learned success head that reads only the prefix

window reaches a fold-mean AUROC of 0.630 to 0.763. This is below the clock score of 0.898 to 0.943 on the same folds. The same head on the terminal window reaches 0.899 to 0.992.