ALARMSHIFT: IDENTIFYING THE VALUE OF EARLY INTERVENTION IN ADAPTIVE INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive-inference gains are routinely attributed to early warning even when the trigger, available interventions, checkpoints, and compute all change together. AlarmShift resolves this ambiguity with a counterfactual protocol that compares alarm families under an identical intervention stack, then delays each selected alarm on its original instance while preserving alarm identity, dispatch opportunities, and realized compute. TriSignal, a four-parameter hysteresis trigger over entropy drift, attention relocation, and retrieval agreement, supplies a low-capacity instrument for this intervention. Across GSM-Hard, SWE-bench Verified, RULER-32K, AIME-24/25, and LiveCodeBench-v4, TriSignal improves Llama-3.1-70B-Instruct success by $+3.97/+4.24/+2.82/+6.86/+5.90$ percentage points over an entropy-matched policy; every paired bootstrap interval is positive, and monitoring consumes 2.4% of runtime. The result reproduces on Qwen-2.5-72B-Instruct and remains positive against a frozen verifier change-point null on both backbones. On both backbones, same-instance replay removes roughly two-fifths of the margin within ten tokens and brings the thirty-token response to statistical overlap with the matched null, while random-instance timing controls remain flat. These experiments establish intervention timing as an independently measurable source of adaptive-inference performance and make AlarmShift a reusable test for separating warning quality from controller capability.

1 INTRODUCTION

An adaptive inference system must decide not only whether additional computation is useful, but when retrieval, branching, or revision can still alter the answer. A warning that arrives before an error becomes entrenched can redirect the trajectory; the same action issued after the model has committed to an incompatible plan may spend identical compute without recovering the instance. This timing question governs the value of test-time intervention across reasoning, long-context retrieval, software repair, and code generation.

Current system comparisons rarely expose that value directly. Trigger design is commonly bundled with a different action repertoire, a denser checkpoint schedule, or a larger search budget. A stronger row can therefore reflect additional controller capacity rather than earlier access to recoverable states. Even a compute-matched trigger comparison remains incomplete because different triggers select different instances and times. What is missing is a counterfactual quantity: the change in task success caused by shifting an already selected alarm while the instance, alarm identity, available actions, checkpoints, and realized compute remain unchanged.

AlarmShift identifies this quantity in two linked contrasts. The first contrast substitutes the alarm family inside a locked retrieve/branch/stop stack. Entropy-Matched-Policy supplies an output-uncertainty null, while Verifier-ChangePoint-Matched-Policy supplies a non-entropy null built from a frozen output verifier; both inherit the same dispatch, external checkpoints, action-mix target, and realized budget. The second contrast replays each TriSignal alarm at controlled delays on the same triggered instance. Delayed replay is stricter than a rank comparison because it preserves who was selected and what intervention follows, changing only when the intervention becomes available. A same-time random-instance permutation separately tests whether selection, rather than timing, produces the response.

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.

We execute this design on Llama-3.1-70B-Instruct and Qwen-2.5-72B-Instruct with five heterogeneous workloads and ten fixed generation seeds. TriSignal's four fitted coefficients combine three decoding-time signals inside the shared controller, isolating warning content from downstream capacity. Under matched realized compute, every backbone–task effect over Entropy-Matched-Policy has a positive paired interval. The non-entropy change-point null narrows the margin yet preserves the ordering on every task. Delay produces a graded loss of success on both backbones: substantial value disappears within ten tokens, and the response reaches statistical overlap with the matched null at thirty tokens. The same pattern survives a retrieve/revise/stop controller and RE-Label supervision, while lower-access variants quantify the value contributed by each signal source.

The result changes how timing claims can be tested and how adaptive controllers can be designed. AlarmShift converts "early warning helps" from an interpretation of bundled accuracy into an estimand with a direct intervention. Its delay-response curve reveals the recoverability window in which extra compute is productive, while the locked-stack comparison determines whether a candidate trigger locates that window better than strong matched alternatives.

Contributions.

- We introduce the AlarmShift estimand and controlled replay protocol, which isolate alarm time by fixing instances, alarm identity, downstream dispatch, checkpoint opportunity, and realized total compute.
- We provide cross-backbone, multi-null evidence on five tasks: TriSignal outperforms entropy and verifier change-point alarms inside the same stack, and its success margin decays under same-instance delay on Llama-3.1-70B-Instruct and Qwen-2.5-72B-Instruct.
- We map how access, supervision, and controller choice reshape the recoverability window, pairing each controlled variation with its performance response, operating cost, and paper-internal evidence trace.

2 Related Work

Uncertainty and adaptive retrieval. Predictive and semantic uncertainty, self-evaluated confidence, and black-box consistency estimate when outputs are unreliable (Kadavath et al., 2022; Kuhn et al., 2023; Lin et al., 2024; Manakul et al., 2023). Retrieval-augmented generation, active retrieval, and adaptive early exit use such evidence to allocate computation (Lewis et al., 2020; Jiang et al., 2023; Asai et al., 2024; Schuster et al., 2022). These methods estimate confidence or system-level utility; AlarmShift additionally estimates the within-instance value of moving a fixed intervention in time.

Search, verification, and revision. Self-consistency, tree search, outcome and process verification, reflection, and iterative refinement improve inference by generating, scoring, or revising trajectories (Wang et al., 2023; Yao et al., 2023; Cobbe et al., 2021; Lightman et al., 2023; Weng et al., 2023; Shinn et al., 2023; Madaan et al., 2023; Brown et al., 2024; Snell et al., 2024). Their central comparison typically changes search width, verifier access, or the downstream policy. AlarmShift treats these as controller choices and asks an orthogonal question: under one such choice, how much task success is causally associated with intervention time?

Representation monitoring and evaluation design. Representation engineering, causal tracing, and activation patching test information localized in hidden states (Meng et al., 2022; Geva et al., 2023; Zou et al., 2023); benchmark studies examine sensitivity to task construction, context length, contamination, and evaluator choice (Mirzadeh et al., 2025; Jimenez et al., 2024; Hsieh et al., 2024; Jain et al., 2025; Zheng et al., 2023). AlarmShift combines these perspectives as an evaluation intervention. A bundled comparison ranks complete systems, a matched-trigger comparison estimates trigger choice under a common stack, and same-instance replay identifies the timing response after selection and controller capacity have been held constant.

3 ALARMSHIFT PROTOCOL

3.1 OPERATIONAL TARGET AND TRISIGNAL INSTRUMENT

For each no-intervention trajectory, an operational irrecoverability time $t^\star$ is the earliest token after which the evaluated rescue envelope no longer recovers end-task success. The pre-failure window $[t^\star - w, t^\star)$ uses $w = 16$, selected on the calibration split. Benchmark-specific marker families detect persistent numeric inconsistency on GSM-Hard, off-closure edits followed by failed tests on SWE-bench Verified, unsupported spans on RULER-32K, surviving algebraic inconsistency on AIME-24/25, and persistent buggy spans on LiveCodeBench-v4. RE-Label replaces these markers with agreement across retrieve-only, branch-only, revise, and no-rescue continuations. Three external expert annotators provide the human audit labels; author arbitration is restricted to the less than 7% conflict set. Full definitions appear in Appendix B.1.

TriSignal combines entropy drift $\widetilde{\Delta H}_t$, attention relocation $\widetilde{C}_t$, and retrieval-agreement divergence $\widetilde{D}_t$:

$$I_t = \sigma\Big(b + w_1\widetilde{\Delta H}_t + w_2\widetilde{C}_t + w_3\widetilde{D}_t\Big). \tag{1}$$

The alarm engages at $I_t > 0.72$ and disengages at $I_t < 0.45$. This three-to-one linear probe has four learned parameters including bias. Detector fitting uses $10{,}000$ instance-disjoint trajectories, threshold calibration uses a disjoint $2{,}000$ trajectories, and all weights and thresholds are frozen before evaluation.

3.2 LOCKED INTERVENTION STACK AND ESTIMANDS

Every matched method dispatches the same retrieve, branch, or stop action. Checkpoints are external to the detector and occur at stride 64 plus sentence boundaries, preventing one trigger from receiving more opportunities to branch. The total budget B_{tot} charges decoder tokens, retrieval and reference passes, signal extraction, branch continuation, and retrieval calls. Thresholds are calibrated until realized rather than nominal compute matches within $\pm 0.3\%$. Each checkpoint can be consumed once, preventing recursive branch amplification.

What remains fixed. Downstream actions and dispatch; external checkpoint opportunities; realized B_{tot}; evaluation instances and ten generation seeds; retrieval index, continuation engine, and answer selection. Stage I changes the trigger family. Stage II fixes the TriSignal alarm identity and triggered population, then changes only its replay time $t + \Delta$; trajectories that terminate before $t + \Delta$ receive no intervention.

For benchmark b, let $Y_{ibs}(a, t)$ denote success for instance i, seed s, alarm family a, and replay time t. The locked-stack trigger estimand is

$$\tau_b(a) = \mathbb{E}_{i,s}[Y_{ibs}(a, t^a_{ibs}) - Y_{ibs}(a_0, t^{a_0}_{ibs})], \tag{2}$$

where a_0 is a matched null. On the population $\mathcal{P}_b$ triggered by TriSignal, the delay-response estimand is

$$\lambda_b(\Delta) = \mathbb{E}_{(i,s)\in\mathcal{P}_b}\left[Y_{ibs}(a_{\mathrm{TS}}, t^{\mathrm{TS}}_{ibs}) - Y_{ibs}(a_{\mathrm{TS}}, t^{\mathrm{TS}}_{ibs} + \Delta)\right], \tag{3}$$

$$\rho_b(\Delta) = \mathbb{E}_{(i,s)\in\mathcal{P}_b}\left[Y_{ibs}(a_{\mathrm{TS}}, t^{\mathrm{TS}}_{ibs} + \Delta) - Y_{ibs}(a_0, t^{a_0}_{ibs})\right], \qquad e_b(\Delta) = 1 - \rho_b(\Delta)/\rho_b(0). \tag{4}$$

The paired unit is one seed–instance outcome. The direct timing cost λ_b compares the original and delayed versions of the same alarm; the remaining margin ρ_b situates that intervention against the matched null. We report absolute success and paired differences first, with erosion e_b as a secondary interpretation.

3.3 EVALUATION

We evaluate Llama-3.1-70B-Instruct (Grattafiori et al., 2024) and Qwen-2.5-72B-Instruct (Qwen et al., 2024) on GSM-Hard ($n = 500$), SWE-bench Verified ($n = 200$), RULER-32K ($n = 600$), AIME I/II 2024 plus AIME I 2025 ($n = 90$), and LiveCodeBench-v4 ($n = 415$) (Gao et al., 2023; Jimenez et al., 2024; Hsieh et al., 2024; Mathematical Association of America, n.d.; Jain et al., 2025).

Table 1: Locked-stack task success (%) and paired TriSignal effects (percentage points with paired 95% bootstrap intervals). Every listed method is coherent and receives the same actions, checkpoints, action-mix target, and realized compute.

Backbone	Alarm	GSM	SWE-V	RULER	AIME	LCB-v4
Llama-3.1-70B	Entropy-Matched	67.07	31.34	61.70	20.10	30.22
	Verifier-ChangePoint-Matched	68.42	32.78	62.57	22.63	32.31
	TriSignal	71.04	35.58	64.52	26.96	36.12
	TriSignal − Entropy-Matched	+3.97 [3.41, 4.52]	+4.24 [3.67, 4.81]	+2.82 [2.32, 3.33]	+6.86 [5.48, 8.24]	+5.90 [4.93, 6.86]
	TriSignal − Verifier-CP-Matched	+2.62 [2.09, 3.15]	+2.80 [2.25, 3.38]	+1.95 [1.44, 2.43]	+4.33 [2.98, 5.59]	+3.81 [2.91, 4.70]
Qwen-2.5-72B	Entropy-Matched	67.28	31.56	61.77	20.50	30.47
	Verifier-ChangePoint-Matched	68.66	32.96	62.67	22.83	32.61
	TriSignal	70.96	35.51	64.37	26.83	36.01
	TriSignal − Entropy-Matched	+3.68 [3.13, 4.22]	+3.95 [3.39, 4.52]	+2.60 [2.12, 3.10]	+6.33 [4.99, 7.71]	+5.54 [4.58, 6.47]
	TriSignal − Verifier-CP-Matched	+2.30 [1.78, 2.81]	+2.55 [2.02, 3.08]	+1.70 [1.23, 2.17]	+4.00 [2.69, 5.29]	+3.40 [2.55, 4.24]

Ten fixed seeds yield $5{,}000/2{,}000/6{,}000/900/4{,}150$ paired evaluations. All runs use eight NVIDIA A100 80GB SXM4 GPUs with NVLink, driver 550.90.07, CUDA 12.4, PyTorch 2.3.1, and vLLM 0.5.4. Intervals are paired 95% instance bootstraps with $10{,}000$ resamples; seed signs use exact one-sided sign tests. No parameter is refit on evaluation instances.

Entropy-Matched-Policy uses output-entropy alarms. Verifier-ChangePoint-Matched-Policy uses change points from the frozen outcome verifier and has no hidden-state features. Both match TriSignal's trigger frequency, action mix, dispatch, checkpoints, and realized compute. The verifier null costs 1.1% wall-clock; TriSignal costs 2.4% on Llama and 2.5% on Qwen.

4 Experiments

4.1 RQ1: Does alarm identity change success under a locked stack?

On Llama, TriSignal raises absolute success from $67.07/31.34/61.70/20.10/30.22\%$ to $71.04/35.58/64.52/26.96/36.12\%$. The corresponding $+3.97/+4.24/+2.82/+6.86/+5.90$ point effects have positive paired intervals, with positive-seed counts $10/10, 10/10, 9/10, 10/10, 10/10$. The Qwen replication is also positive on every task, with positive-seed counts 10/10, 10/10, 9/10, 10/10, 10/10. The frozen verifier change-point null is stronger than the entropy null on both backbones, yet TriSignal retains a positive interval in every stack–task cell; its positive-seed counts are 10/10, 9/10, 9/10, 9/10, 10/10 on Llama and 10/10, 9/10, 9/10, 9/10, 10/10 on Qwen. Thus the ordering survives a non-entropy alarm family without changing the rescue policy.

AIME and LiveCodeBench show that the effect is not confined to retrieval-centric tasks: neither workload depends on SWE-style repository navigation, and the AIME dispatch never invokes retrieval. Appendix Table 7 exposes wins, losses, ties, seed dispersion, and intervals for every main comparison.

4.2 RQ2: Does the margin respond to alarm delay?

The effect is largest at the original alarm and decreases with each delay on every backbone–task pair. At ten tokens, the Llama margins are $+2.26/+2.43/+1.59/+3.85/+3.30$ points, corresponding to $43.2/42.8/43.6/43.9/44.1\%$ erosion of the Stage-I gain. Qwen exhibits the same ordering with visibly task-dependent decay. By thirty tokens, every Qwen interval overlaps zero, and all five Llama margins lie within 0.05 point of Entropy-Matched-Policy with intervals spanning zero. Because replay preserves the selected instance, alarm identity, action, checkpoints, and budget, this dose response attributes the lost success to intervention time rather than to a different selected subset.

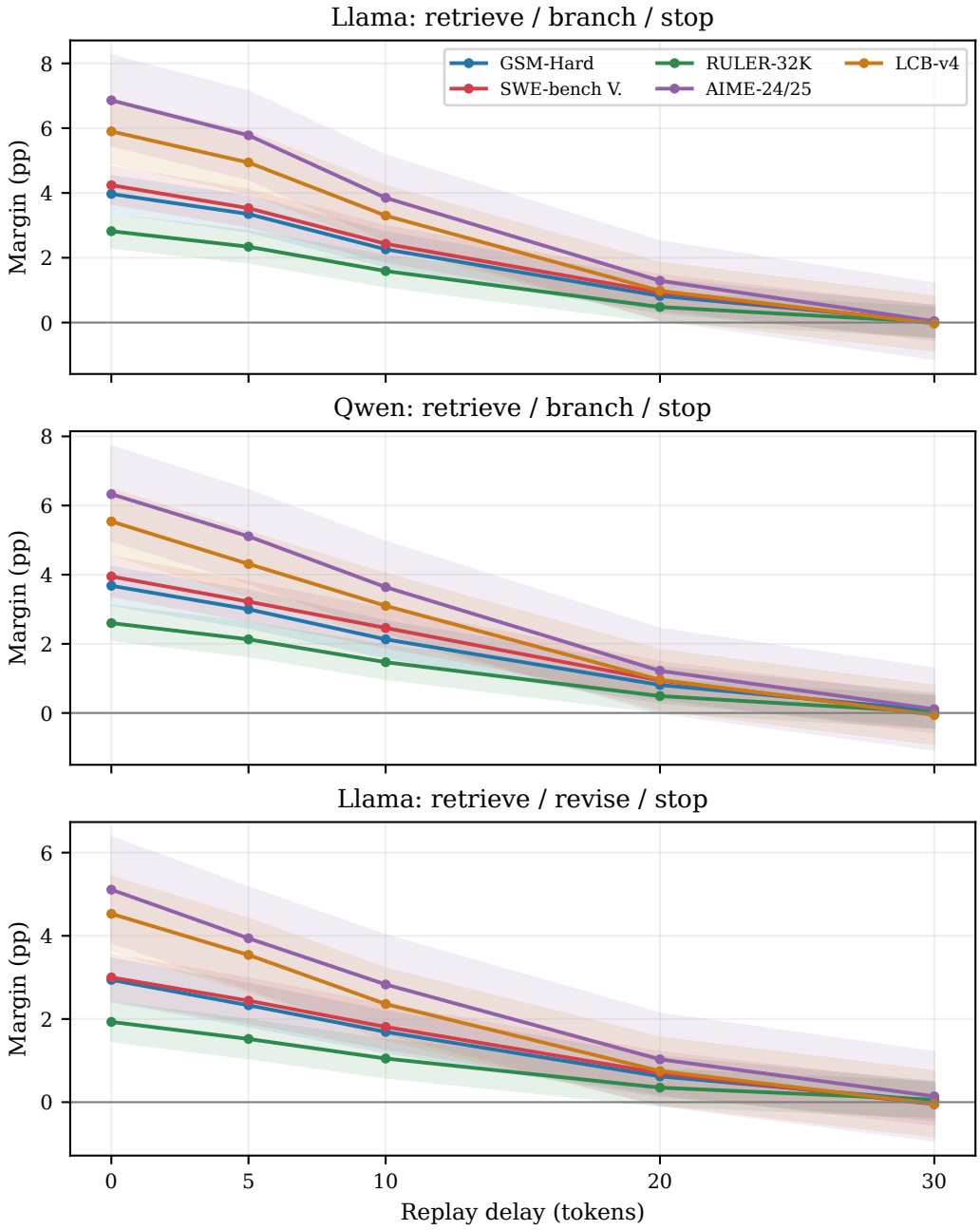


Figure 1: AlarmShift delay response. Each panel reports the remaining TriSignal success margin over the matched entropy null for five benchmarks. The intervention uses the same alarm identity and instance at $\Delta \in \{0, 5, 10, 20, 30\}$; shading shows paired 95% bootstrap intervals. Llama and Qwen use retrieve/branch/stop, while the third panel replaces branch with revision.

The same-time random-instance control is flat: the Llama fitted slopes remain within 0.03 point per token, and the Qwen slopes and intervals in Appendix Table 9 all include zero. Remaining margins and the complete delay grid appear in Appendix Table 6.

4.3 RQ3: WHICH DEPENDENCIES SHAPE THE RESPONSE?

Replacing branch with revision preserves TriSignal success at $70.61/35.12/63.92/26.42/35.62\%$ and yields a replay curve that remains positive after ten tokens and reaches statistical overlap with zero at thirty. Training on RE-Label preserves the $+3.27/ + 3.40/ + 2.13/ + 6.09/ + 5.01$ point Stage-I effects and the same monotone replay ordering. Lower-access TriSignal-Proxy reaches $70.22/34.72/63.05/26.02/35.18\%$ at 1.4% overhead, while TriSignal-Lite reaches $70.40/34.94/63.46/26.38/35.52\%$ at 0.9%. These results locate the sources of timing information: the largest access penalty occurs on retrieval-heavy RULER-32K, while controller and supervision changes alter the magnitude of the timing response without reversing the task-level ordering.

Table 2: Same-controller access isolation on Llama-3.1-70B-Instruct. All rows share dispatch, checkpoints, splits, and matched B_{tot}; only detector access changes.

Access regime	GSM	SWE-V	RULER	AIME	LCB-v4	Mean retained gain
Entropy-Matched-Policy	67.07	31.34	61.70	20.10	30.22	0.0%
TriSignal-Proxy	70.22	34.72	63.05	26.02	35.18	75.5%
TriSignal-Lite	70.40	34.94	63.46	26.38	35.52	82.5%
Full TriSignal	71.04	35.58	64.52	26.96	36.12	100.0%

Table 3: Supervision dependence: paired task-success difference from Entropy-Matched-Policy (percentage points with 95% intervals).

Supervision	GSM	SWE-V	RULER	AIME	LCB-v4
Default marker	+3.97 [3.41,4.52]	+4.24 [3.67,4.81]	+2.82 [2.32,3.33]	+6.86 [5.48,8.24]	+5.90 [4.93,6.86]
RE-Label	+3.27 [2.73,3.80]	+3.40 [2.86,3.94]	+2.13 [1.66,2.59]	+6.09 [4.74,7.43]	+5.01 [4.08,5.93]
Adjudicated human	+2.56 [1.89,3.22]	+2.71 [2.01,3.40]	+1.79 [1.21,2.36]	+4.28 [2.87,5.69]	+3.68 [2.59,4.76]

Table 4: Control-stack evidence matrix. Each row links a timing quantity to the invariant that isolates it, its paper-internal evidence, and the observed response.

Object	Quantity tested	Isolation invariant	Evidence	Observed response
Llama default	Trigger contrast and delay response	Shared actions, checkpoints, instances, seeds, and realized compute	Table 1; Fig. 1	Positive five-task margin; thirty-token intervals include zero; 2.4% overhead
Qwen replication	Trigger contrast and delay response	Frozen Qwen fit with the same split and controller rules	Table 1; Fig. 1	Cross-backbone reproduction with 2.5% overhead
Verifier null	Non-entropy trigger contrast	Frozen verifier; matched frequency, actions, checkpoints, and compute	Table 1	Positive margin on both backbones; 1.1% overhead
Revision controller	Delay response after replacing branch	Same alarm, external checkpoints, and realized budget	Table 6	Positive ten-token response; thirty-token overlap with zero
RE-Label	Delay response under alternative supervision	Four-family rescue labels on disjoint trajectories	Tables 3, 6	Positive Stage-I effects and monotone replay decay
Lower access	Retained trigger margin	Same dispatch and budget with API/Lite signal access	Table 2	$75.5/82.5\%$ mean retention at $1.4/0.9\%$ overhead

5 DISCUSSION

Same-instance replay raises the evidentiary standard for adaptive inference because it separates warning quality from the controller that consumes the warning. A Stage-I improvement says that one alarm selects more useful intervention points under a common stack; the delay response shows that those points occupy a finite recoverability window. The steep ten-token loss indicates that compute allocation must account for state commitment, not merely uncertainty magnitude. Externalized checkpoints prevent a detector from creating more opportunities to branch, and the flat random-instance control shows that replay decay follows the selected alarm's time rather than an arbitrary reassignment of alarms across examples.

The dependence map also guides system design. The non-entropy null demonstrates that TriSignal captures useful structure beyond output-entropy fluctuations, and the Qwen replication establishes the response on a materially different white-box backbone. Revision and RE-Label preserve the dose response while changing its magnitude, identifying controller and supervision as quantitative modifiers of the recoverability window. Lower-access variants retain most of the gain at lower overhead, with a concentrated penalty where retrieval-grounded agreement is most informative. AlarmShift therefore gives future triggers, controllers, and checkpoint schedules a common experimental target: maximize useful intervention inside the measured recoverability window.

6 CONCLUSION

AlarmShift turns intervention timing from a bundled-system intuition into an experimentally isolated design variable. Across two 70B-class backbones, two matched null families, five workloads, and controlled variations in supervision and downstream action, early alarms produce positive task-success margins that decay under same-instance delay. The protocol reveals when retrieval, branching, or revision compute can still change an answer and provides a reusable test for distinguishing earlier warning from a stronger controller.

AI USE STATEMENT

Generative AI was used only for writing, editing, and presentation of this manuscript. The authors reviewed all edits and remain responsible for the final content.

REPRODUCIBILITY STATEMENT

Section 3 specifies the alarm, locked intervention stack, estimands, and compute matching; Section 3.3 records instance-disjoint splits, workloads, hardware, seeds, and paired statistical procedures. Appendix B.1, Algorithm 1, and Tables 7–9 describe label construction and dispatch and provide paired summaries and the reported replay grid.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique Through Self-Reflection. *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling. *arXiv preprint arXiv:2407.21787*, 2024. doi: 10.48550/arXiv.2407.21787. URL https://arxiv.org/abs/2407.21787.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided Language Models. *Proceedings of the International*

Conference on Machine Learning, pp. 10764–10799, 2023. URL https://proceedings.mlr.press/v202/gao23f.html.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12216–12235, 2023. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? *arXiv preprint arXiv:2404.06654*, 2024. doi: 10.48550/arXiv.2404.06654. URL https://arxiv.org/abs/2404.06654.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active Retrieval Augmented Generation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*, 2022. doi: 10.48550/arXiv.2207.05221. URL https://arxiv.org/abs/2207.05221.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023. doi: 10.48550/arXiv.2305.20050. URL https://arxiv.org/abs/2305.20050.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=DWkJCSxKU5.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad

Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Mathematical Association of America. MAA Invitational Competitions, n.d. URL https://maa.org/maa-invitational-competitions/. Accessed September 8, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AjXkRZIvjB.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. Confident Adaptive Language Modeling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17456–17472, 2022. doi: 10.52202/068431-1269. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6fac9e316a4ae75ea244ddcef1982c71-Abstract-Conference.html.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. *arXiv preprint arXiv:2408.03314*, 2024. doi: 10.48550/arXiv.2408.03314. URL https://arxiv.org/abs/2408.03314.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1IMMmrw.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large Language Models are Better Reasoners with Self-Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2550–2575, 2023. doi: 10.18653/v1/2023.findings-emnlp.167. URL https://aclanthology.org/2023.findings-emnlp.167/.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in Neural Information Processing Systems*, 36:11809–11822, 2023. doi: 10.52202/075280-0517. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*, 2023. doi: 10.48550/arXiv.2310.01405. URL https://arxiv.org/abs/2310.01405.

A Replay and Paired-Outcome Details

This section expands the controller, supervision, delay, and randomization results summarized in Section 4. Every table retains the paired seed–instance unit and the locked accounting rules from Section 3.2.

Table 5: Retrieve+revise+stop replication with no branch. Absolute success and paired TriSignal effects use the same checkpoints, ten seeds, and matched B_{tot} as the default stack.

Method	GSM	SWE-V	RULER	AIME	LCB-v4
Entropy-Matched-Policy	67.67	32.12	61.99	21.31	31.09
TriSignal	70.61	35.12	63.92	26.42	35.62
Paired difference [95% CI]	$+2.94$ [2.43, 3.45]	$+3.00$ [2.46, 3.54]	$+1.93$ [1.47, 2.40]	$+5.11$ [3.85, 6.37]	$+4.53$ [3.65, 5.42]

Table 5 isolates the change from branching to revision by keeping the checkpoint schedule and resource accounting common to the two alarms. The paired intervals are positive on every task, with effects ranging from $+1.93$ points on RULER-32K to $+5.11$ points on AIME. This comparison concerns alarm identity within the alternative controller. The following replay grid separately asks how much of that margin remains when the selected alarm is delayed.

Table 6: Complete same-instance delayed-replay grid. Cells are the remaining TriSignal margin over the setting's Entropy-Matched-Policy in percentage points, with paired 95% bootstrap intervals.

Setting	Δ	GSM	SWE-V	RULER	AIME	LCB-v4
Llama / retrieve+branch+stop	0	+3.97 [3.41, 4.52]	+4.24 [3.67, 4.81]	+2.82 [2.32, 3.33]	+6.86 [5.48, 8.24]	+5.90 [4.93, 6.86]
	5	+3.35 [2.81, 3.89]	+3.53 [2.97, 4.09]	+2.34 [1.86, 2.82]	+5.78 [4.43, 7.13]	+4.94 [4.00, 5.88]
	10	+2.26 [1.74, 2.78]	+2.43 [1.88, 2.97]	+1.59 [1.12, 2.06]	+3.85 [2.55, 5.15]	+3.30 [2.39, 4.21]
	20	+0.82 [0.32, 1.32]	+0.93 [0.40, 1.46]	+0.48 [0.03, 0.93]	+1.29 [0.08, 2.50]	+0.98 [0.12, 1.84]
	30	+0.03 [-0.46, 0.52]	−0.02 [-0.54, 0.50]	+0.01 [-0.43, 0.45]	+0.04 [-1.12, 1.20]	−0.03 [-0.86, 0.80]
Qwen / retrieve+branch+stop	0	+3.68 [3.13, 4.22]	+3.95 [3.39, 4.52]	+2.60 [2.12, 3.10]	+6.33 [4.99, 7.71]	+5.54 [4.58, 6.47]
	5	+3.00 [2.47, 3.54]	+3.22 [2.67, 3.78]	+2.13 [1.65, 2.61]	+5.11 [3.77, 6.44]	+4.31 [3.38, 5.23]
	10	+2.13 [1.61, 2.65]	+2.46 [1.91, 3.00]	+1.47 [0.99, 1.95]	+3.64 [2.33, 4.95]	+3.10 [2.19, 4.02]
	20	+0.81 [0.31, 1.32]	+0.94 [0.41, 1.47]	+0.49 [0.03, 0.96]	+1.22 [0.01, 2.43]	+0.96 [0.10, 1.82]
	30	+0.07 [-0.42, 0.56]	−0.04 [-0.56, 0.48]	+0.02 [-0.43, 0.47]	+0.11 [-1.06, 1.28]	−0.06 [-0.89, 0.78]
Llama / retrieve+revise+stop	0	+2.94 [2.43, 3.45]	+3.00 [2.46, 3.54]	+1.93 [1.47, 2.40]	+5.11 [3.85, 6.37]	+4.53 [3.65, 5.42]
	5	+2.33 [1.83, 2.84]	+2.44 [1.91, 2.97]	+1.52 [1.06, 1.98]	+3.94 [2.72, 5.16]	+3.54 [2.67, 4.41]
	10	+1.69 [1.20, 2.18]	+1.81 [1.29, 2.33]	+1.05 [0.60, 1.50]	+2.83 [1.65, 4.01]	+2.36 [1.51, 3.21]
	20	+0.62 [0.15, 1.09]	+0.70 [0.20, 1.20]	+0.35 [-0.08, 0.78]	+1.03 [-0.07, 2.13]	+0.75 [-0.05, 1.55]
	30	+0.03 [-0.43, 0.49]	−0.05 [-0.54, 0.44]	+0.06 [-0.36, 0.48]	+0.14 [-0.92, 1.20]	−0.04 [-0.82, 0.74]
Llama / RE-Label trigger	0	+3.27 [2.73, 3.80]	+3.40 [2.86, 3.94]	+2.13 [1.66, 2.59]	+6.09 [4.74, 7.43]	+5.01 [4.08, 5.93]
	5	+2.66 [2.14, 3.18]	+2.79 [2.26, 3.32]	+1.72 [1.27, 2.17]	+4.87 [3.56, 6.18]	+4.02 [3.14, 4.90]
	10	+1.92 [1.42, 2.42]	+2.08 [1.56, 2.60]	+1.22 [0.78, 1.66]	+3.58 [2.31, 4.85]	+2.87 [2.02, 3.72]
	20	+0.75 [0.27, 1.23]	+0.81 [0.31, 1.31]	+0.38 [-0.04, 0.80]	+1.34 [0.17, 2.51]	+0.94 [0.15, 1.73]
	30	+0.05 [-0.42, 0.52]	−0.02 [-0.51, 0.47]	+0.01 [-0.40, 0.42]	+0.08 [-1.03, 1.19]	−0.04 [-0.81, 0.73]

For every setting, the zero-delay row is the corresponding trigger contrast. Reading down each block holds the reported setting fixed while moving the replay time. Positive intervals persist at ten tokens; every thirty-token interval spans zero. These are remaining margins over the entropy null, not direct estimates of absolute delayed success or of the timing loss λ_b.

Table 7: Paired Stage-I outcomes. W/L/T are TriSignal wins, losses, and ties over all seed–instance pairs; seed SD is the standard deviation of the ten paired seed effects.

Comparison	Task	Δ [95% CI]	W/L/T	Positive seeds	Seed SD
Llama: TriSignal − Entropy	GSM	$+3.97$ [3.41, 4.52]	624/425/3951	10/10	0.30
Llama: TriSignal − Entropy	SWE	$+4.24$ [3.67, 4.81]	272/187/1541	10/10	0.33
Llama: TriSignal − Entropy	RULER	$+2.82$ [2.32, 3.33]	580/411/5009	9/10	1.27
Llama: TriSignal − Entropy	AIME	$+6.86$ [5.48, 8.24]	151/89/660	10/10	0.61
Llama: TriSignal − Entropy	LCB	$+5.90$ [4.93, 6.86]	478/233/3439	10/10	0.48
Llama: TriSignal − Verifier-CP	GSM	$+2.62$ [2.09, 3.15]	505/374/4121	10/10	0.38
Llama: TriSignal − Verifier-CP	SWE	$+2.80$ [2.25, 3.38]	218/162/1620	9/10	0.42
Llama: TriSignal − Verifier-CP	RULER	$+1.95$ [1.44, 2.43]	487/370/5143	9/10	0.73
Llama: TriSignal − Verifier-CP	AIME	$+4.33$ [2.98, 5.59]	128/89/683	9/10	0.88
Llama: TriSignal − Verifier-CP	LCB	$+3.81$ [2.91, 4.70]	394/236/3520	10/10	0.59

Wins and losses count pairs in which the two alarms yield different success outcomes; ties include joint success and joint failure. Under an equally weighted binary-outcome estimator, the paired percentage-point effect is $100(W - L)/(W + L + T)$. Positive-seed counts summarize sign consistency and seed SD describes dispersion across the ten seed effects; neither replaces the instance-level paired interval.

Table 8: Paired Stage-I outcomes, continued. W/L/T are TriSignal wins, losses, and ties over all seed–instance pairs; seed SD is the standard deviation of the ten paired seed effects.

Comparison	Task	Δ [95% CI]	W/L/T	Positive seeds	Seed SD
Qwen: TriSignal − Entropy	GSM	+3.68 [3.13, 4.22]	575/391/4034	10/10	0.34
Qwen: TriSignal − Entropy	SWE	+3.95 [3.39, 4.52]	231/152/1617	10/10	0.39
Qwen: TriSignal − Entropy	RULER	+2.60 [2.12, 3.10]	526/370/5104	9/10	0.91
Qwen: TriSignal − Entropy	AIME	+6.33 [4.99, 7.71]	135/78/687	10/10	0.75
Qwen: TriSignal − Entropy	LCB	+5.54 [4.58, 6.47]	421/191/3538	10/10	0.52
Qwen: TriSignal − Verifier-CP	GSM	+2.30 [1.78, 2.81]	485/370/4145	10/10	0.41
Qwen: TriSignal − Verifier-CP	SWE	+2.55 [2.02, 3.08]	201/150/1649	9/10	0.46
Qwen: TriSignal − Verifier-CP	RULER	+1.70 [1.23, 2.17]	450/348/5202	9/10	0.78
Qwen: TriSignal − Verifier-CP	AIME	+4.00 [2.69, 5.29]	112/76/712	9/10	0.96
Qwen: TriSignal − Verifier-CP	LCB	+3.40 [2.55, 4.24]	345/204/3601	10/10	0.61
Llama revise: TriSignal − Entropy	GSM	+2.94 [2.43, 3.45]	520/373/4107	10/10	0.36
Llama revise: TriSignal − Entropy	SWE	+3.00 [2.46, 3.54]	217/157/1626	9/10	0.44
Llama revise: TriSignal − Entropy	RULER	+1.93 [1.47, 2.40]	472/356/5172	9/10	0.82
Llama revise: TriSignal − Entropy	AIME	+5.11 [3.85, 6.37]	123/77/700	10/10	0.79
Llama revise: TriSignal − Entropy	LCB	+4.53 [3.65, 5.42]	398/210/3542	10/10	0.58
Llama RE-Label: TriSignal − Entropy	GSM	+3.27 [2.73, 3.80]	550/386/4064	10/10	0.42
Llama RE-Label: TriSignal − Entropy	SWE	+3.40 [2.86, 3.94]	225/157/1618	9/10	0.48
Llama RE-Label: TriSignal − Entropy	RULER	+2.13 [1.66, 2.59]	498/370/5132	9/10	0.95
Llama RE-Label: TriSignal − Entropy	AIME	+6.09 [4.74, 7.43]	138/83/679	10/10	0.87
Llama RE-Label: TriSignal − Entropy	LCB	+5.01 [4.08, 5.93]	430/222/3498	10/10	0.64

Wins and losses count pairs in which the two alarms yield different success outcomes; ties include joint success and joint failure. Under an equally weighted binary-outcome estimator, the paired percentage-point effect is $100(W - L)/(W + L + T)$. Positive-seed counts summarize sign consistency and seed SD describes dispersion across the ten seed effects; neither replaces the instance-level paired interval.

Table 9: Same-time random-instance control. Cells report slope in percentage points per token with paired 95% intervals.

Setting	GSM	SWE-V	RULER	AIME	LCB-v4
Llama/default		$\lvert\hat{\beta}\rvert < 0.03$ pp/token on every benchmark			
Qwen/default	$+0.006$ [-0.016, 0.028]	-0.011 [-0.036, 0.014]	$+0.004$ [-0.015, 0.023]	$+0.014$ [-0.021, 0.049]	-0.008 [-0.031, 0.015]
Llama/revise	$+0.010$ [-0.014, 0.034]	-0.018 [-0.045, 0.009]	$+0.002$ [-0.019, 0.023]	$+0.019$ [-0.020, 0.058]	-0.012 [-0.037, 0.013]
Llama/RE-Label	$+0.008$ [-0.015, 0.031]	-0.014 [-0.040, 0.012]	$+0.003$ [-0.017, 0.023]	$+0.017$ [-0.020, 0.054]	-0.010 [-0.034, 0.014]

The random-instance control changes the recipient instance while retaining the alarm-time schedule. Slopes describe how the remaining margin changes with delay. The Qwen, revision, and RE-Label intervals all span zero. The Llama/default row supplies only the reported magnitude bound, so it does not specify taskwise slope intervals.

B Algorithm, Audit Table, and Supporting Audit Extensions

Algorithm 1 TriSignal alarm and shared dispatch

Require: Model M, input x, budget B, thresholds $\tau_{\text{trigger}}, \tau_{\text{recover}}$, external checkpoint rule $\mathcal{R}$
Initialize: alarm $\leftarrow 0$, $t \leftarrow 0$, $\mathcal{C} \leftarrow \mathcal{R}(\cdot)$ (checkpoints from external rule), used $\leftarrow \emptyset$
while $t < B$ **do**
  Generate next token; read r_t; compute $\Delta H_t, C_t, D_t$ and I_t via Eq. 1
  if $I_t > \tau_{\text{trigger}}$ **then**
    alarm $\leftarrow 1$
  else if $I_t < \tau_{\text{recover}}$ **then**
    alarm $\leftarrow 0$
  end if
  if alarm $= 1$ **then**
    Select action from the shared dispatch rule (retrieve / branch / stop)
    if action = branch and $\exists\, c \in \mathcal{C} \setminus$ used with $c \le t$ **then**
      used $\leftarrow$ used $\cup \{c\}$; $t \leftarrow c$
    else
      execute retrieve / stop
    end if
  end if
  $t \leftarrow t + 1$
end while
return best answer from trajectory

Implementation invariants. Algorithm 1 summarizes state readout, signal extraction, score computation, and downstream dispatch; the total-compute accounting is described in Section 3.2. The external checkpoint rule $\mathcal{R}$ gives every alarm identical branching opportunities, and the `used` set prevents recursive branch amplification. Together with Section 3.2, these invariants preserve alarm time as the experimental variable.

API-token / wall-clock equivalence. For the verifier-independent / API-only baselines (DeBERTa-Verif-Search, Reflexion-BoN (Shinn et al., 2023), LLM-Judge-Adaptive (Zheng et al., 2023); see the *Verifier-independent* block of Table 11) the budget charged is B_{tot}, converted as 1 API prompt token $= 1.2$ Llama-70B decoder-token equivalents and 1 API completion token $= 1.0$ decoder-token equivalent. Those factors are calibrated so the 95-th-percentile wall-clock of `gpt-4o-mini-2024-07-18` at temperature 0.0 matches the Llama-70B median wall-clock per token on the 8×A100 rig. DeBERTa-Verif-Search is trained on 25,000 disjoint Llama-3.1-8B-Instruct trajectories and frozen before evaluation; Reflexion-BoN uses the same decoder temperature as SC-Verif-Rerank so the comparison differs only in whether the outcome verifier is used. Appendix G compares wall-clock, dollar-cost, and raw-API-call accounting.

B.1 Full Label-Construction Details

Section 3 summarizes the supervision target; this subsection records the full task-by-task marker definitions. The hierarchy is default benchmark marker, four-family RE-Label, and externally adjudicated human timing. All markers are built from no-intervention rollouts. The pre-failure window is selected on calibration AUROC over $w \in \{4, 8, 16, 32, 64\}$, with $w = 16$ reaching 0.891 versus 0.874 at $w = 8$ and 0.882 at $w = 32$. These supervision-only markers identify the earliest token after which the tested rescue envelope no longer recovers success.

- GSM-Hard: marker = first post-CoT token at which the emitted numeric plan contradicts a previously asserted intermediate equality by at least 10^{-3} (automated symbolic check with SymPy 1.12). $t^\star$ is set to that token.
- SWE-bench Verified: marker = first token that edits a file outside the ground-truth patch's dependency closure *and* that is followed by a failed hidden test. $t^\star$ is the earlier of that token and the first repeated tool-call with identical arguments.
- RULER-32K: marker = first token that emits a span not present in the top-k=32 retrieved needles and that survives until the final answer. $t^\star$ is set to that token.

- AIME-24/25: marker = first token that commits to a numeric or algebraic sub-result inconsistent with the later gold answer under SymPy simplification and that is never corrected before final answer extraction. $t^\star$ is the earliest token in that surviving inconsistent span.
- LiveCodeBench-v4: marker = first token in a code span whose executed final program fails hidden tests because of a persistent signature, indexing, state-update, or control-flow bug. $t^\star$ is the earliest token in the surviving buggy span.

RE-Label broadens supervision beyond the fixed three-action controller. For each trajectory, $t^\star_{\mathrm{RE}}$ is the earliest token at which four rescue families all fail across eight matched continuations under the same B_{tot}: retrieve-only with top-k=4 BGE-M3, branch-only with a 1,024-token continuation, verifier-guided revise/self-correct, and free continuation. This construction is independent of the fixed dispatch and benchmark marker rules while retaining an operational rescue-envelope target. The earlier fixed-controller CI-Label is its three-action special case. Section 4.3 and Appendix D.3 compare both label families.

Table 10: Derived-metric audit. Each row gives the exact operation and source object used for the displayed value; uncertainty is a paired-bootstrap standard error over 10,000 resamples.

Metric	Definition and source	Five-task value	Paired SE
Retained gain	Benchmark mean of $(\mathrm{Succ}_v - \mathrm{Succ}_{\mathrm{Ent}})/(\mathrm{Succ}_{\mathrm{Full}} - \mathrm{Succ}_{\mathrm{Ent}})$; Tables 1, 12	Full/Lite/Proxy/Qwen-affine/Gemma-affine: $100.0/82.5/75.5/89.8/87.3\%$	0.7–2.1%
Local-repair lift	$\mathrm{Succ}_{\mathrm{TS+repair}} - \mathrm{Succ}_{\mathrm{TS}}$; Fig. 5	$+2.14/+2.41/+1.39/+2.85/+3.20$ pp	0.21–0.48 pp
Matched repair control	Target recovery minus same-layer random-direction recovery; Fig. 5	$37.4 - 10.8 = +26.6$ pp	1.5 pp
RE-Label retention	$\Delta_{\mathrm{RE}}/\Delta_{\mathrm{default}}$; Table 3	76–89%; mean 82.4%	1.4–2.2%
Rescue headroom	Rescue probability at TriSignal time divided by entropy time	$1.85/2.07/1.68/2.09/2.10\times$	0.07–0.11
Ten-token erosion	$1 - \rho_b(10)/\rho_b(0)$; Table 6	$43.2/42.8/43.6/43.9/44.1\%$	1.9–2.7%
Monitoring overhead	Monitored minus unmonitored wall-clock share; Table 19	Full/Proxy/Lite: $2.4/1.4/0.9\%$	0.05–0.10%
Proxy vs. API baselines	Proxy success minus the per-task best verifier-independent system; Table 11	$+1.57/+1.80/+1.32/+3.66/+2.40$ pp	0.25–0.57 pp
Cross-stack median	Median host-stack gain at 50-trajectory calibration; Table 14	$+1.89$ pp; range $+1.62$–$+2.15$	0.22–0.45 pp
Human-label retention	$\Delta_{\mathrm{human}}/\Delta_{\mathrm{default}}$; Table 3	$64.5/63.9/63.5/62.4/62.4\%$	1.8–2.7%
Proxy-XL retention	XL gain divided by full white-box gain; Table 21	$94.2/91.3/88.7/95.0/92.9\%$	0.9–1.7%

Table 10 links every derived percentage to its operation, source cells, and paired uncertainty. Absolute success differences in Table 1 remain the primary effects; retained-gain and erosion ratios summarize how those effects change across access, supervision, and delay. The displayed definitions make each quantitative synthesis directly reconstructible.

Table 11 provides absolute problem difficulty, oracle headroom, a random matched-policy control, and every method contributing to the unlocked and verifier-independent envelopes.

Access and transfer synthesis. The white-box block measures in-domain performance, the API-proxy block isolates lower-access detection under the same controller, and the Qwen/Gemma blocks measure cross-backbone adaptation. The 200-trajectory proxy is the access-ladder operating point; Figure 36 sweeps its calibration budget, and Table 21 evaluates the seven-signal Proxy-XL variant. RULER-32K determines the minimum confidence bound throughout the proxy block; its lower bound varies non-monotonically because paired outcome covariance changes with calibration even as absolute success increases. Table 10 derives retained-gain summaries from these absolute values.

Control-stack sensitivity. Coarse stride-128 checkpointing produces the largest degradation, whereas sentence-boundary checkpoints preserve most of the gain, indicating that fresh intervention

Table 11: Complete success comparison. Underlined rows are per-task envelopes: the unlocked envelope selects RCV on four tasks and Verifier-Search-Tuned on RULER-32K; the verifier-independent envelope selects LLM-Judge-Adaptive on four tasks and DeBERTa-Verif-Search on RULER-32K. All values follow Table 1's protocol.

Method	GSM-Hard	SWE-bench V.	RULER-32K	AIME-24/25	LCB-v4
Unmatched baselines					
No Intervention	64.20±0.8	28.41±1.2	58.73±0.9	16.43±2.3	27.84±1.1
Entropy-Stop	66.82±0.7	30.17±1.1	60.42±0.8	18.81±2.2	29.42±1.0
Always-Retrieve	68.12±0.6	32.38±0.9	62.31±0.7	18.24±2.1	29.83±1.0
Oracle	74.26±0.5	38.84±0.7	67.21±0.6	30.14±1.7	39.55±0.8
Matched compute: shared actions, checkpoints, and B_{tot} within $\pm 0.3\%$					
Random-Matched-Policy	65.93±0.8	30.12±1.1	60.04±0.9	19.04±2.2	29.78±1.0
Entropy-Matched-Policy	67.07±0.6	31.34±0.9	61.70±0.7	20.10±2.1	30.22±1.0
Unlocked families at the same B_{tot} and frozen verifier					
Verifier-Search-Tuned (6-wide + BoN=8)	68.58±0.6	32.98±0.9	62.48±0.7	21.21±2.0	31.83±1.0
SC-Verif-Rerank (n=8, T=0.8)	68.47±0.6	32.60±0.9	61.93±0.7	22.03±2.0	32.55±1.0
RCV (revise depth = 2, critique T=0.5)	68.96±0.6	33.42±0.9	62.15±0.7	22.69±2.0	33.04±1.0
Summary only: unlocked per-task max	68.96	33.42	62.48	22.69	33.04
Verifier-independent / API-only baselines at the same B_{tot}					
DeBERTa-Verif-Search (v3-L, 4-wide + BoN= 8)	68.10±0.7	32.64±0.9	61.73±0.8	21.85±2.0	32.14±1.0
Reflexion-BoN (verifier-free, n=8)	68.02±0.7	32.48±1.0	61.45±0.8	22.02±2.1	32.40±1.0
LLM-Judge-Adaptive (GPT-4o-mini)	68.65±0.7	32.92±0.9	61.58±0.8	22.36±2.0	32.78±1.0
Summary only: verif.-indep. per-task max	*68.65*	*32.92*	*61.73*	*22.36*	*32.78*
TriSignal (Ours)	71.04±0.5	35.58±0.8	64.52±0.6	26.96±1.9	36.12±0.9
Δ vs. Entropy-M. [95% CI]	+3.97 [3.41, 4.52]	+4.24 [3.67, 4.81]	+2.82 [2.32, 3.33]	+6.86 [5.48, 8.24]	+5.90 [4.93, 6.86]
Δ vs. best non-TriSignal [95% CI]	+2.08 [1.58, 2.57]	+2.16 [1.64, 2.67]	+2.04 [1.56, 2.51]	+4.27 [3.04, 5.48]	+3.08 [2.25, 3.89]
Δ vs. best verif.-indep. [95% CI]	+2.39 [1.86, 2.91]	+2.66 [2.11, 3.20]	+2.79 [2.29, 3.28]	+4.60 [3.39, 5.80]	+3.34 [2.49, 4.18]
TriSignal-Proxy (API-only; 1.4% OH; 200-traj cal.)	70.22±0.6	34.72±0.8	63.05±0.7	26.02±1.9	35.18±0.9
Δ vs. best verif.-indep. [95% CI]	+1.57 [1.06, 2.08]	+1.80 [1.23, 2.36]	+1.32 [0.82, 1.81]	+3.66 [2.44, 4.87]	+2.40 [1.57, 3.22]
TriSignal + Verif.-Search-Tuned	72.81±0.5	37.24±0.7	65.62±0.6	28.24±1.8	38.17±0.9
Δ vs. best non-TriSignal [95% CI]	+3.85 [3.32, 4.37]	+3.82 [3.28, 4.34]	+3.14 [2.62, 3.65]	+5.55 [4.30, 6.78]	+5.13 [4.28, 5.96]

Table 12: Access and transfer grid. Rows report absolute task success (%; five-seed transfer and access evaluations; ten seeds for the in-domain row) and the minimum paired-bootstrap lower bound across tasks. "R" identifies RULER-32K as the task setting that bound. Blocks separate in-domain access, API-proxy calibration, and cross-backbone adaptation.

Access / backbone / adaptation	Cal.	Overhead	Succ. G/S/R/A/L	min CI low
White-box, in-domain Llama-3.1-70B-Instruct				
Full TriSignal (main)	10k traj.	2.4%	71.04 / 35.58 / 64.52 / 26.96 / 36.12	+1.56 (R)
TriSignal-Lite (no ref-pass)	10k traj.	0.9%	70.40 / 34.94 / 63.46 / 26.38 / 35.52	+0.41 (R)
Black-box API proxy (logprobs + n=8 SC + retr.-str. agr. + surprise-grad.)				
TriSignal-Proxy (0-traj., zero-shot)	0 traj.	1.4%	68.97 / 33.88 / 61.34 / 24.52 / 33.81	−0.61 (R)
TriSignal-Proxy (25-traj.)	25 traj.	1.4%	69.76 / 34.36 / 62.44 / 25.62 / 34.63	+0.12 (R)
TriSignal-Proxy (50-traj.)	50 traj.	1.4%	70.03 / 34.49 / 62.78 / 25.92 / 34.90	+0.56 (R)
TriSignal-Proxy (100-traj.)	100 traj.	1.4%	70.13 / 34.63 / 62.95 / 26.00 / 35.02	+0.38 (R)
TriSignal-Proxy (200-traj., main row)	200 traj.	1.4%	70.22 / 34.72 / 63.05 / 26.02 / 35.18	+0.19 (R)
Cross-backbone transfer (Qwen-2.5-72B-Instruct)				
Zero-shot	0 traj.	2.4%	69.96 / 34.40 / 62.67 / 25.31 / 34.75	+0.04 (R)
Affine re-scale	200 traj.	2.4%	70.76 / 35.27 / 63.91 / 26.40 / 35.73	+1.07 (R)
Full refit	2k traj.	2.4%	70.96 / 35.51 / 64.37 / 26.83 / 36.01	+1.45 (R)
Cross-backbone transfer (Gemma-2-27B-it)				
Zero-shot	0 traj.	2.4%	69.82 / 34.29 / 62.43 / 25.12 / 34.56	−0.02 (R)
Affine re-scale	200 traj.	2.4%	70.65 / 35.16 / 63.81 / 26.27 / 35.62	+0.94 (R)
Full refit	2k traj.	2.4%	70.90 / 35.45 / 64.30 / 26.75 / 35.94	+1.37 (R)

Table 13: Portability across action and checkpoint rules on Llama-3.1-70B-Instruct. Every row uses 10k calibration trajectories, 2.4% overhead, and matched B_{tot}. Success and delta vectors follow GSM/SWE/RULER/AIME/LCB order.

Rule	Success vector	Delta vs. best realistic	Mean retained gain	Min CI low
Retrieve+revise+stop (no branch)	70.61 / 35.12 / 63.92 / 26.42 / 35.62	+1.65 / +1.70 / +1.44 / +3.73 / +2.58	86.8%	+0.97
Sentence-boundary checkpoints	70.86 / 35.38 / 64.29 / 26.74 / 35.93	+1.90 / +1.96 / +1.81 / +4.05 / +2.89	93.9%	+1.32
Stride-128 checkpoints	70.14 / 34.62 / 63.24 / 25.86 / 35.12	+1.18 / +1.20 / +0.76 / +3.17 / +2.08	82.4%	+0.24

points matter more than an exact stride. Retrieve+revise+stop preserves the ranking after replacing branch with revision. Every tested variant has a positive minimum lower confidence bound, with the full stack remaining the strongest operating point.

Table 14: Cross-stack modularity sweep for three host stacks and three calibration budgets under matched B_{tot}. Task cells report paired success effects; "Median / seed SD" adds the median effect and the range of ten-seed standard deviations across tasks. The last column is the minimum paired 95% lower bound.

Host stack	Cal.	GSM-Hard	SWE-bench V.	RULER-32K	AIME-24/25	LCB-v4	Median / seed SD	min CI low
H1. RCV with the shared frozen verifier								
Host baseline	—	68.96	33.42	62.15	22.69	33.04	—	—
H1 + TriSignal trigger	25 traj.	+1.58	+1.73	+1.68	+1.86	+1.79	+1.73 / 0.44–0.96	+0.94
H1 + TriSignal trigger	50 traj.	+1.82	+1.97	+1.92	+2.10	+2.03	+1.97 / 0.37–0.88	+1.18
H1 + TriSignal trigger	200 traj.	+2.01	+2.14	+2.08	+2.30	+2.22	+2.14 / 0.34–0.82	+1.33
H2. SC-Verif-Rerank with $n = 8$ and the shared verifier								
Host baseline	—	68.47	32.60	61.93	22.03	32.55	—	—
H2 + TriSignal trigger	25 traj.	+1.54	+1.64	+1.61	+1.91	+1.72	+1.64 / 0.41–0.91	+0.85
H2 + TriSignal trigger	50 traj.	+1.78	+1.88	+1.85	+2.15	+1.96	+1.88 / 0.39–0.93	+1.10
H2 + TriSignal trigger	200 traj.	+1.96	+2.06	+2.02	+2.32	+2.13	+2.06 / 0.33–0.84	+1.27
H3. Separately trained DeBERTa-Verif-Search								
Host baseline	—	68.10	32.64	61.73	21.85	32.14	—	—
H3 + TriSignal trigger	25 traj.	+1.39	+1.58	+1.59	+1.77	+1.65	+1.59 / 0.48–1.05	+0.78
H3 + TriSignal trigger	50 traj.	+1.62	+1.82	+1.83	+2.01	+1.89	+1.83 / 0.42–0.99	+1.04
H3 + TriSignal trigger	200 traj.	+1.79	+2.00	+2.00	+2.18	+2.07	+2.00 / 0.36–0.90	+1.21
Median at 50 trajectories		+1.78	+1.88	+1.85	+2.10	+1.96	+1.89	+1.04

Cross-stack synthesis. At 50-trajectory calibration, all 15 host-stack–task deltas are positive and the smallest paired lower bound is +1.04 on H3 GSM-Hard. Median gain rises from +1.65 to +1.89 points between 25 and 50 trajectories and continues upward at 200. The separately trained DeBERTa H3 host has the smallest effects yet preserves all five signs, showing that reuse of the native outcome verifier is unnecessary. Host baselines in the table make every delta directly reconstructible.

C Supporting Access, Repair, Transfer, and Timing Diagnostics

C.1 API-only and Access Diagnostics

API-only, verifier-independent validation. Figure 2 visualizes the gaps from Table 11. TriSignal-Proxy has a positive interval on every task using API-visible signals and an independently trained

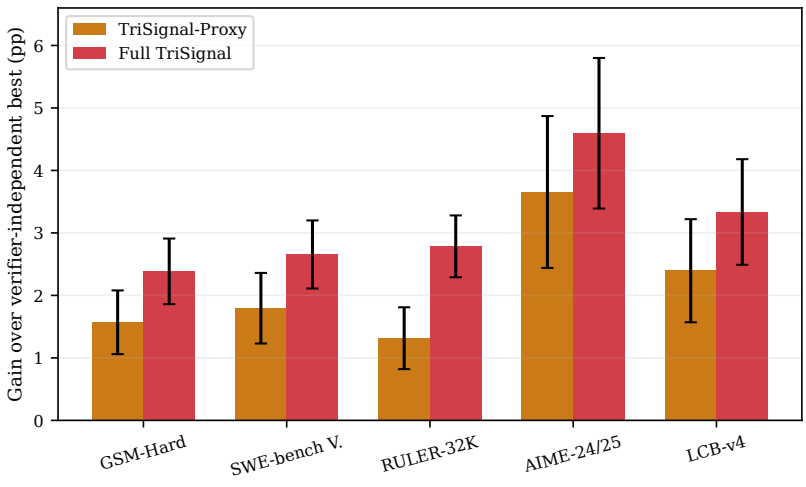


Figure 2: API-only, verifier-independent validation. Per-task gain of TriSignal-Proxy (orange) and TriSignal (red) over the verifier-independent envelope in Table 11. Error bars are 95% paired bootstrap intervals; all ten lower bounds are positive, with a minimum of $+0.82$ points on RULER-32K.

verifier. White-box TriSignal is consistently stronger, and the largest access penalty occurs on retrieval-heavy RULER-32K, matching the decomposition in Table 2.

Same-controller access-isolation ladder. Figure 3 expands Table 2. The API-only rung is 0.82–1.47 points below full TriSignal; Lite narrows this to 0.58–1.06 points. API signals capture most of the gain, the residual-stream increment is uniform and small, and the retrieval-grounded pass contributes most on RULER-32K. All five benchmarks follow the same monotone ladder.

C.2 Local-Repair and Compute-Frontier Diagnostics

Local repairability test: full appendix intervention panel.

Local-repair result. The middle panel supplies the strongest matched control: same layer, time, norm, and residual covariance, with only the learned direction randomized. Target recovery reaches 37.4% versus 10.8% for this control, a $+26.6$ point difference. Online repair adds $+2.14/+2.41/+1.39/+2.85/+3.20$ points over TriSignal. Successful repairs reduce I_t from 0.812 to 0.408 ($p < 10^{-6}$), while the random direction changes it from 0.809 to 0.794 ($p = 0.38$), localizing intervention-relevant information near the alarm. The left panel concentrates recovery near one layer/offset neighborhood, further localizing the intervention-relevant state.

Appendix frontier plot on five benchmarks. Figure 8 shows threshold-robust ranking across the plotted compute frontier. The hybrid dominates all five Pareto fronts throughout the τ sweep, and TriSignal remains above the per-task best unlocked baseline at every matched operating point. The main-table operating point therefore lies on a broad region of positive effects.

C.3 Tested Transfer and Supervision Diagnostics

Tested transfer across backbones, action families, and checkpoint rules.

Transfer-cube synthesis. With three actions and full checkpoints, mean retained gain is 62.3% for zero-shot Qwen and 58.7% for zero-shot Gemma. A 200-trajectory affine re-scale raises these values to $83.1\%/80.4\%$, and a 2k-trajectory refit reaches $94.2\%/91.7\%$. For in-domain Llama, retrieve+revise+stop is the strongest alternative action family at 84.3% mean retained gain; sentence-boundary checkpoints retain 95.3%. Every evaluated cell preserves the TriSignal ranking. Tables 12 and 13 provide the corresponding absolute success values.

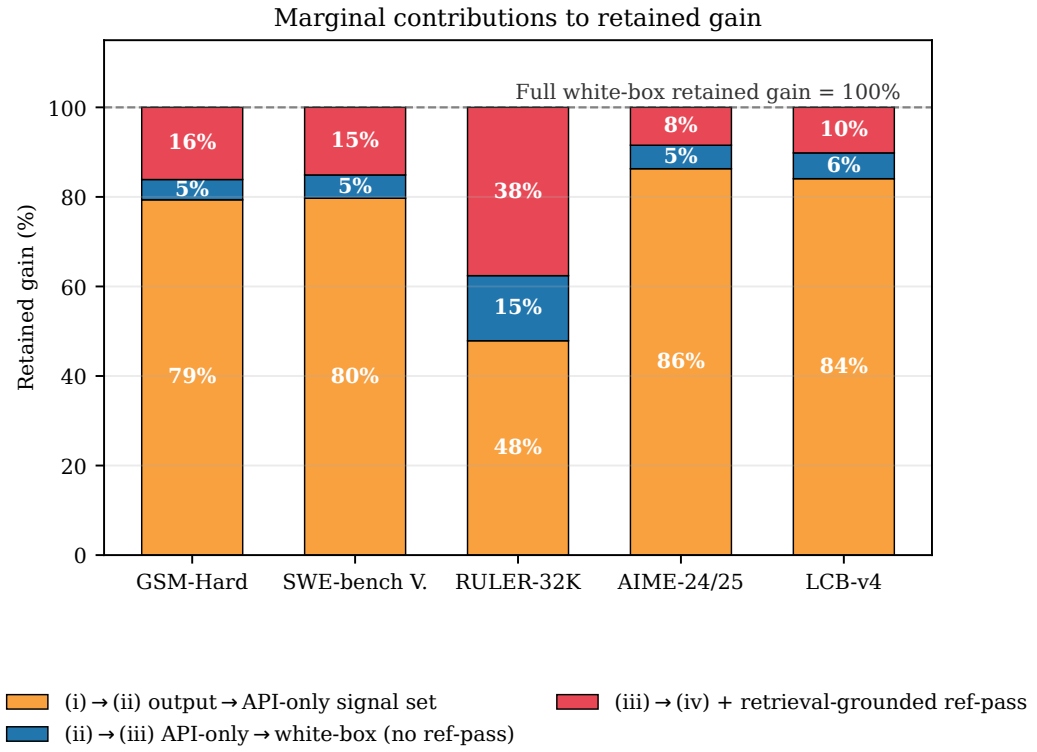


Figure 3: First panel of a 2-panel sequence. Same-controller access-isolation ladder (main-text Table 2, expanded view). First panel: per-benchmark retained gain at each access regime; bars stack the three marginal contributions (output-only→API-only, API-only→white-box no ref-pass, +retrieval-grounded reference pass); the horizontal dashed line marks full-white-box retained gain (100%). Second panel: 5-benchmark mean retained gain per rung with annotated marginal *retained-gain percentage-point* contributions ($+75.5/+7.0/+17.5$). $15/15$ per-benchmark marginals are strictly positive with paired bootstrap CIs above zero.

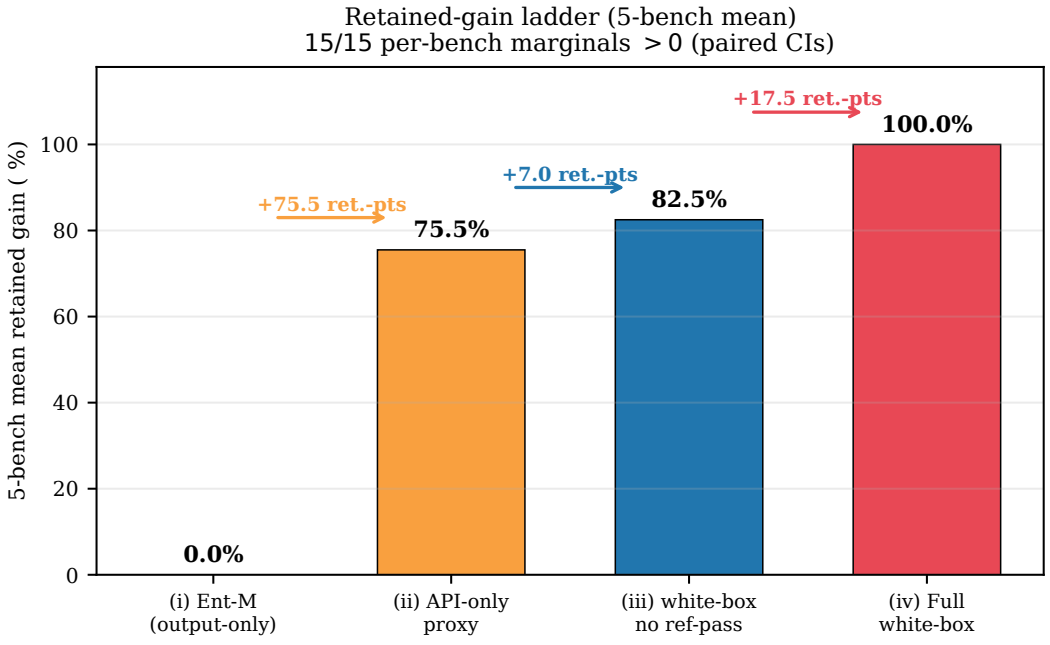


Figure 4: Panel (b) of Figure 3. Mean retained gain across the five benchmarks for each access regime. The successive rungs separate API-visible evidence, hidden-state monitoring without the reference pass, and the retrieval-grounded reference pass. The percentage-point increments refer to retained gain, rather than absolute task success; the full white-box setting defines the 100% reference.

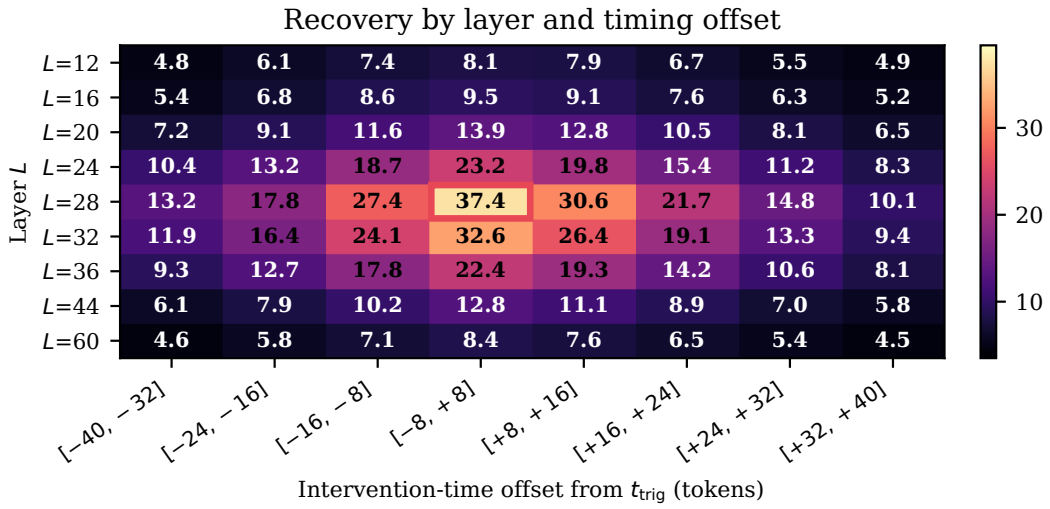


Figure 5: First panel of a 3-panel sequence. Appendix local-repair intervention on n=1,842 triggered failing trajectories. First panel: layer/offset heatmap; the target (L=28, $[-8, +8]$) recovers 37.4% [35.1, 39.7], while at the same central offset window recovery is 23.2% at $L = 24$ and 32.6% at $L = 32$. Second panel: target intervention versus the four matched controls; the strongest same-layer, same-time, norm-matched random-direction arm (avg. of 8 orthogonalised directions) reaches 10.8%; random-layer, random-time, and output-matched representation-agnostic controls stay at 8.2/9.4/11.6% (all $p < 10^{-4}$, paired McNemar). Third panel: online lift per benchmark and the α-sweep over $\{0.25, 0.50, 0.75, 1.00, 1.25\}\|v^L\|$; recovery stays above every control at $\alpha \in \{0.5, 0.75, 1.0\}$.

Supervision validity and adjudicated audit. Figure 14 triangulates timing with a shared headroom curve, four-family RE-Label training, and adjudicated human judgments. Under retrieve+revise+stop, RE-Label preserves a 1.53–$1.82\times$ headroom advantage and gains $+1.65/+1.70/+1.44/+3.73/+2.58$ points over the strongest baseline in that action family (Table 13). Figure 17 supplies the label-family comparison; Figure 24 supplies the alternative-family curves.

Label-Family vs. Human Alignment. On the shared 1,500-trajectory pool, RE-Label yields 0.848 token-level AUROC and 84.7% human agreement within ± 10 tokens, compared with 0.891 and 75.6% for default supervision. Its label-marker-artefact share falls to 4–6%. The resulting success effects retain 76–89% of the default gain, establishing a favorable tradeoff between benchmark-specific detection fit and human-aligned timing.

C.4 Timing Dynamics at the Trigger

We pair a worked GSM-Hard trajectory with the benchmark-wide lead-time distribution to connect the aggregate effects in Table 1 to token-level trigger dynamics.

Figure 20 shows the composite score rising before the visible contradiction. Hysteresis absorbs the short stabilization at step 52, and the constituent traces show coordinated movement across entropy drift, attention relocation, and retrieval disagreement. TriSignal therefore exposes a stable early-warning event before output-level failure becomes explicit.

Figure 21 moves from one trajectory to the benchmark distribution. Median lead time increases from 11 to 36 tokens, and TriSignal's broad right tail identifies instances where internal drift precedes output-entropy cues by enough time for retrieval or branching to alter the trajectory. The near-zero overlap captures genuine instance heterogeneity while the distribution-level shift matches the replay response.

D Low-Failure Regime and Transfer Calibration

D.1 Low-Failure Regime

Table 15: Low-failure-regime evaluation on GSM8K (standard, 1,319 instances). TriSignal stays idle on the vast majority of instances. Monitoring overhead is the additional wall-clock from signal extraction relative to unmonitored generation.

Method	Success (%)	Trig./inst.	False-Trig. (%)	Wasted (%)	Mon. OH (%)
No Intervention	92.4±0.3	—	—	—	0.0
Fixed-Schedule	92.7±0.3	4.2	—	87.6	0.0
Always-Retrieve	92.9±0.3	8.1	—	93.4	0.0
Entropy-Stop	92.6±0.3	0.38	6.8	5.1	0.0
TriSignal	93.1±0.3	0.14	2.1	1.8	2.4

On low-failure GSM8K, TriSignal fires 0.14 times per instance and achieves the lowest false-trigger rate (2.1%) and wasted-budget share (1.8%) in Table 15. The controller remains idle on most easy instances while increasing success from 92.4% to 93.1%, demonstrating selective intervention in a high-base-accuracy regime.

D.2 Transfer Calibration Curves

Figure 22 separates detector calibration from end-task success. Affine re-scaling recovers most of the zero-shot gap with 200 target trajectories, showing that calibration accounts for much of the backbone shift. Full refitting raises both F1 and task success further. The result establishes efficient, calibration-sensitive transfer across the evaluated model families.

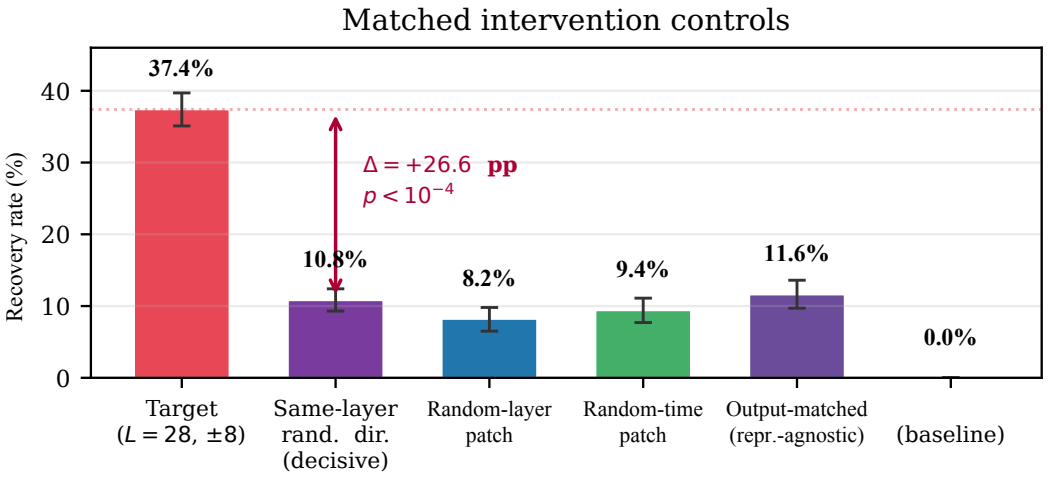


Figure 6: Panel (b) of Figure 5. Recovery after the targeted local representation intervention and four matched control interventions on the 1,842 triggered failing trajectories. The comparison distinguishes the chosen layer, time window, and direction from interventions that alter one of these choices. Recovery percentages measure successful rescue within this selected population, rather than population-wide task-success improvements.

Table 16: Label-robustness extended to all five benchmarks. Values show TriSignal success and Δ vs. Entropy-Matched-Policy under each labeling choice. Ranking is preserved everywhere.

	GSM-Hard		SWE-bench V.		RULER-32K		AIME-24/25		LiveCodeBench-v4	
Variant	Succ.	Δ	Succ.	Δ	Succ.	Δ	Succ.	Δ	Succ.	Δ
w=16 / default (main)	71.04	+3.97	35.58	+4.24	64.52	+2.82	26.96	+6.86	36.12	+5.90
w=8	70.71	+3.64	35.21	+3.87	64.22	+2.52	26.67	+6.57	35.79	+5.57
w=32	70.84	+3.77	35.37	+4.03	64.32	+2.62	26.77	+6.67	35.89	+5.67
Alt. marker	70.53	+3.46	35.09	+3.75	64.14	+2.44	26.58	+6.48	35.62	+5.40
Strict marker	71.15	+4.08	35.70	+4.36	64.62	+2.92	27.07	+6.97	36.23	+6.01
CI-Label (fixed-controller)	70.40	+3.33	34.98	+3.64	64.02	+2.32	26.42	+6.32	35.44	+5.22
RE-Label (4 rescue families)	70.34	+3.27	34.74	+3.40	63.83	+2.13	26.19	+6.09	35.23	+5.01

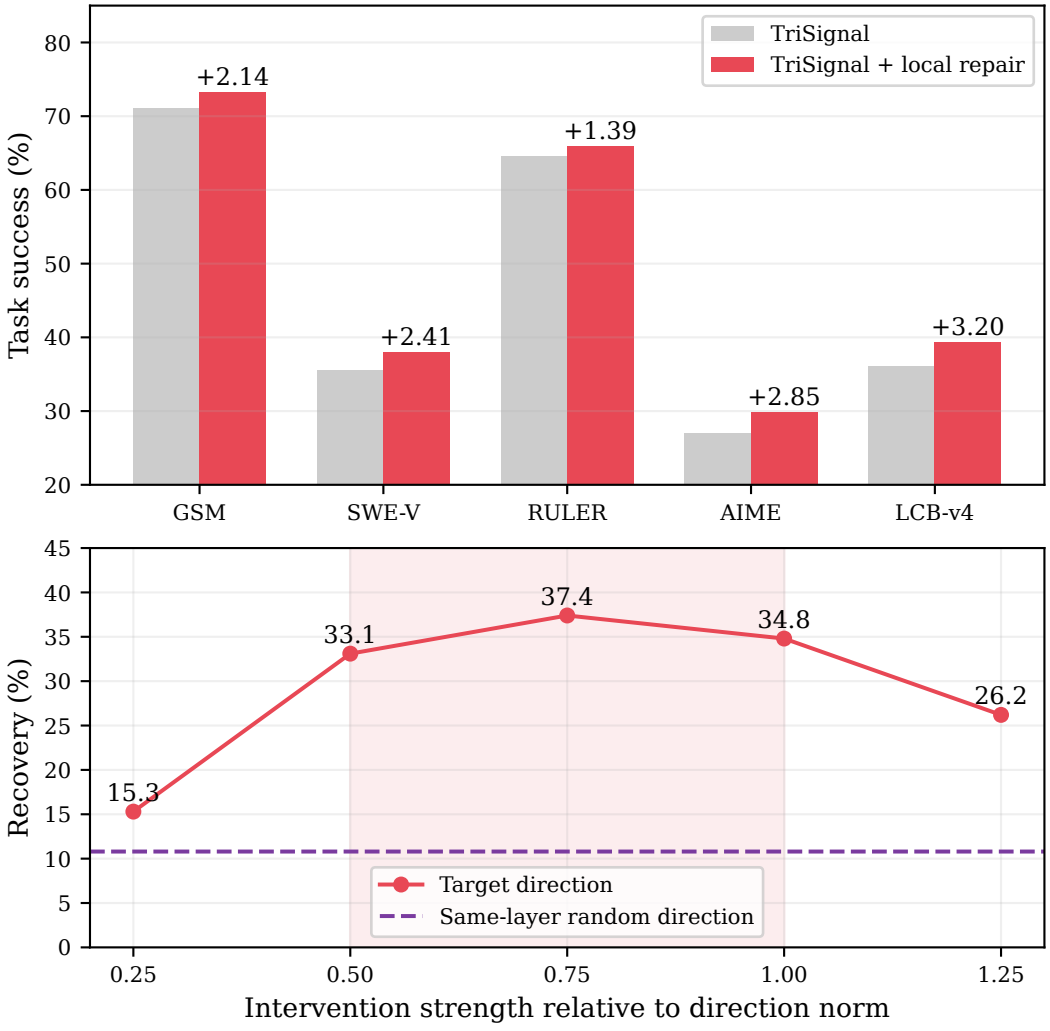


Figure 7: Panel (c) of Figure 5. Online task-success lift from local repair, together with the intervention-strength sweep. The benchmark bars express end-to-end improvement in percentage points; the strength curve instead measures recovery as the intervention magnitude changes relative to the representation direction norm. These two quantities distinguish performance impact from sensitivity of the local repair operation.

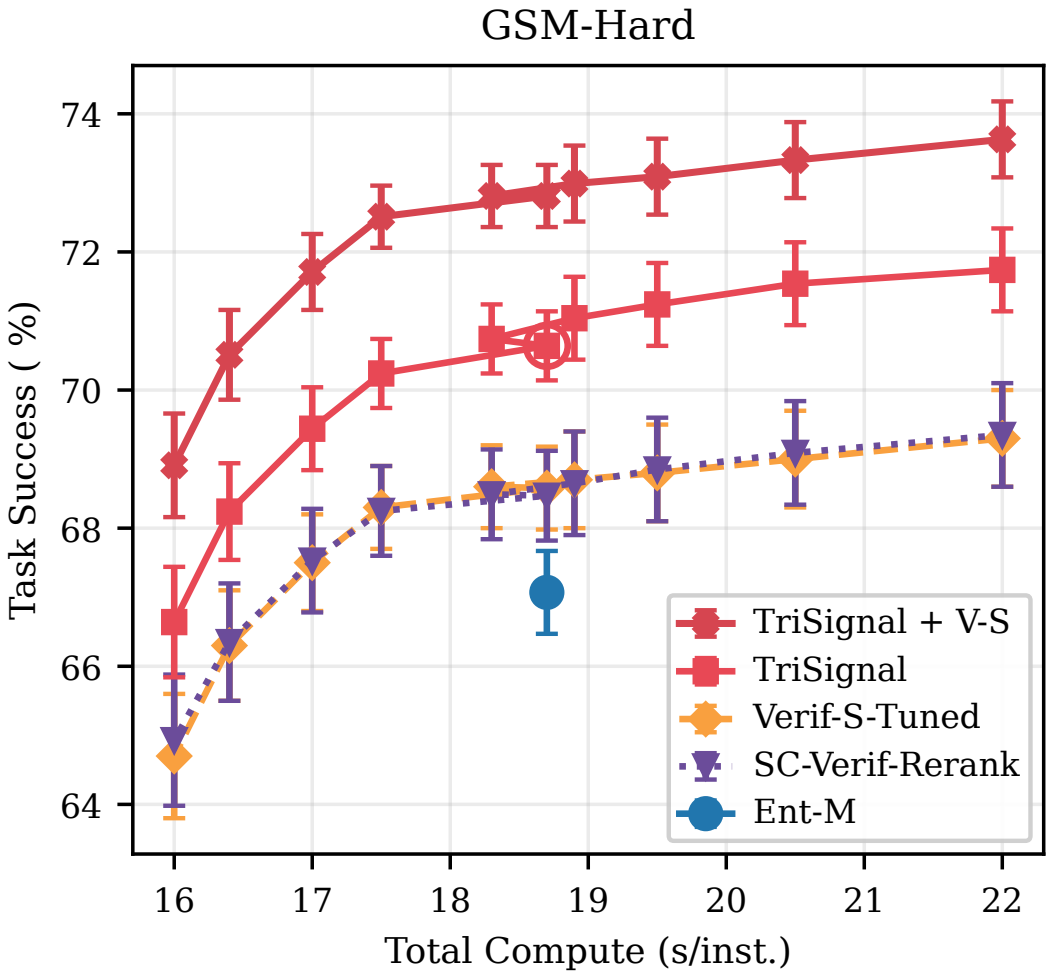


Figure 8: First panel of a 5-panel sequence. Total-compute frontier on all five benchmarks. Success vs. per-instance total compute (8×A100 80GB SXM4, NVIDIA driver 550.90.07, CUDA 12.4, PyTorch 2.3.1, vLLM 0.5.4). Four curves per panel: TriSignal ($\tau \in \{0.50, \ldots, 0.90\}$, step 0.05), Verifier-Search-Tuned (Pareto front of the sweep over branch width, verifier-call budget, rerank depth), SC-Verif-Rerank (n=8, $T \in \{0.4, 0.6, 0.8, 1.0\}$), and the TriSignal + Verifier-Search-Tuned hybrid. Entropy-Matched-Policy single point shown. The hybrid dominates the Pareto front on all five benchmarks, and TriSignal alone dominates the per-task best of the unlocked baselines at every matched operating point from τ=0.50 to τ=0.90.

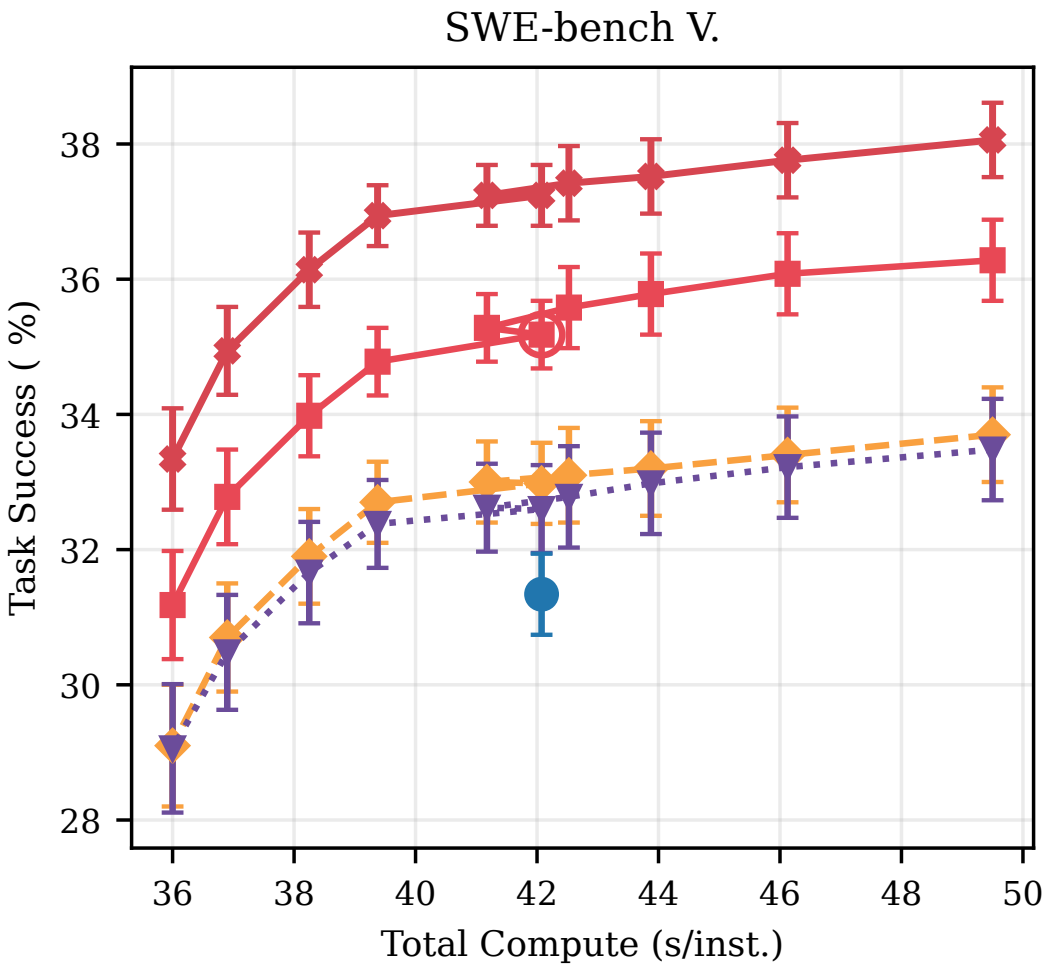


Figure 9: Panel (b) of Figure 8. SWE-bench Verified total-compute frontier. The horizontal axis measures per-instance wall-clock time and the vertical axis measures task success. The four curves compare TriSignal, verifier search, self-consistency with verifier reranking, and the combined trigger/search controller. The entropy-matched point provides a fixed reference; the circled trigger point indicates the selected threshold.

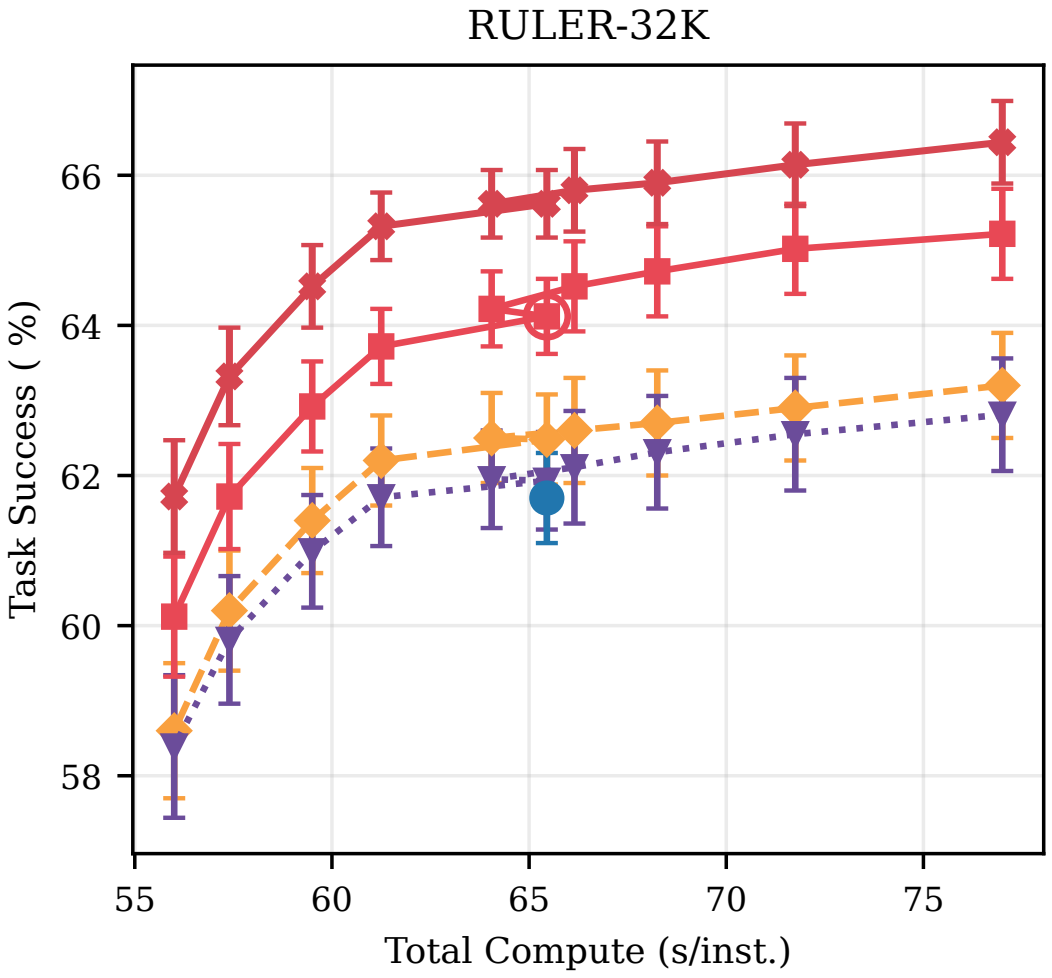


Figure 10: Panel (c) of Figure 8. RULER-32K total-compute frontier. The horizontal axis measures per-instance wall-clock time and the vertical axis measures task success. The four curves compare TriSignal, verifier search, self-consistency with verifier reranking, and the combined trigger/search controller. The entropy-matched point provides a fixed reference; the circled trigger point indicates the selected threshold.

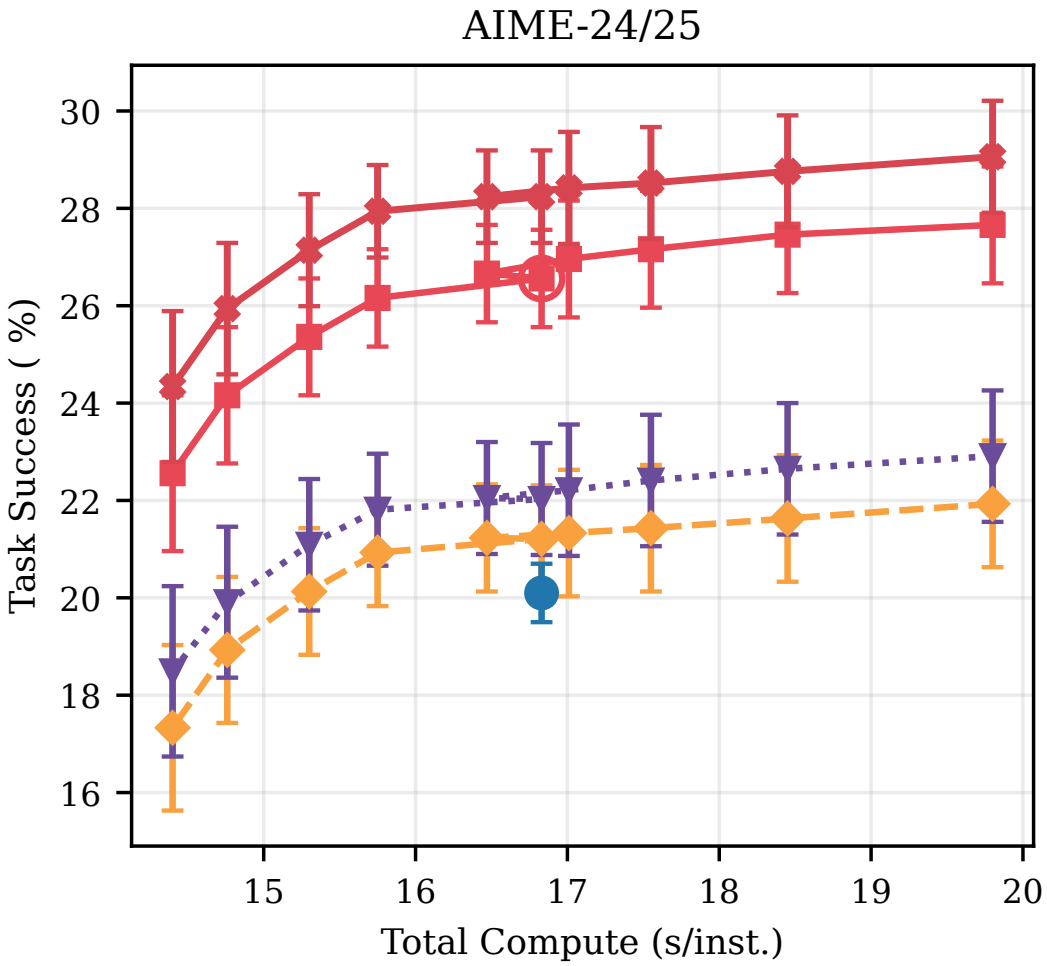


Figure 11: Panel (d) of Figure 8. AIME-24/25 total-compute frontier. The horizontal axis measures per-instance wall-clock time and the vertical axis measures task success. The four curves compare TriSignal, verifier search, self-consistency with verifier reranking, and the combined trigger/search controller. The entropy-matched point provides a fixed reference; the circled trigger point indicates the selected threshold.

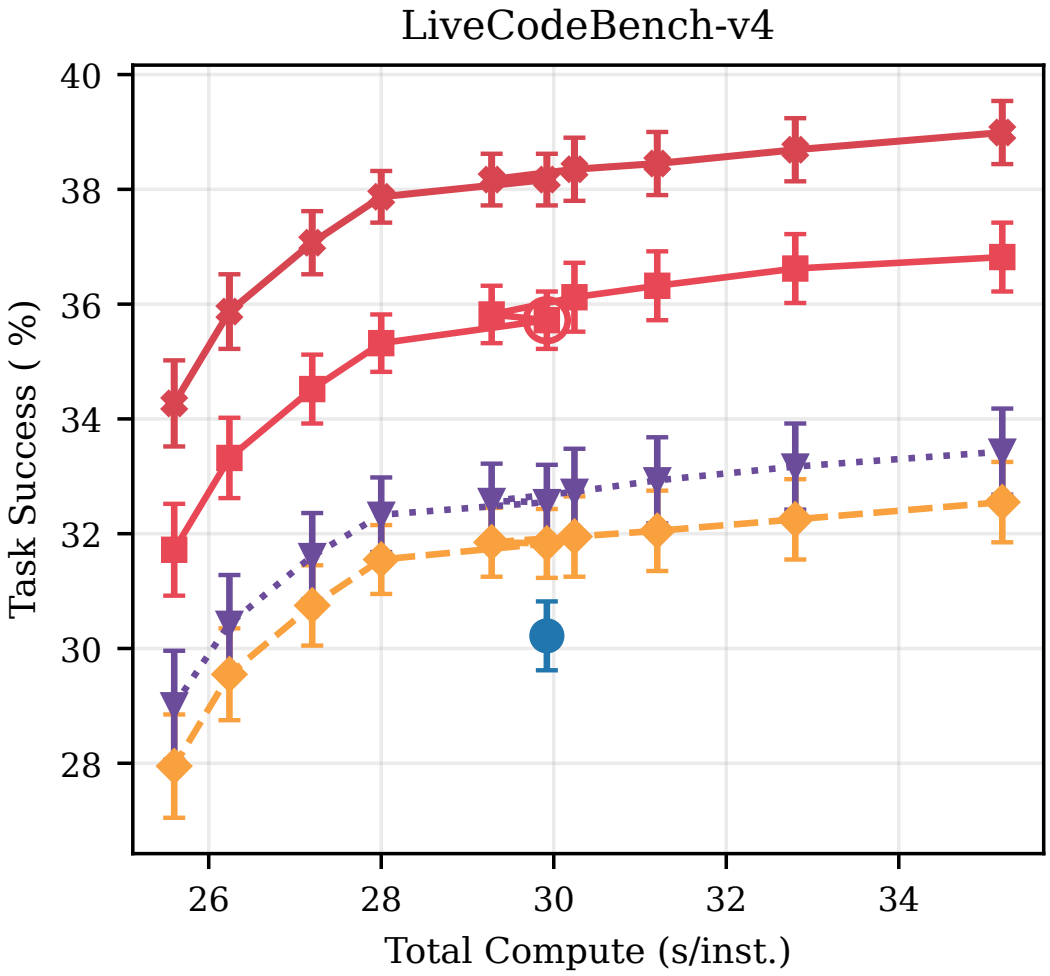


Figure 12: Panel (e) of Figure 8. LiveCodeBench-v4 total-compute frontier. The horizontal axis measures per-instance wall-clock time and the vertical axis measures task success. The four curves compare TriSignal, verifier search, self-consistency with verifier reranking, and the combined trigger/search controller. The entropy-matched point provides a fixed reference; the circled trigger point indicates the selected threshold.

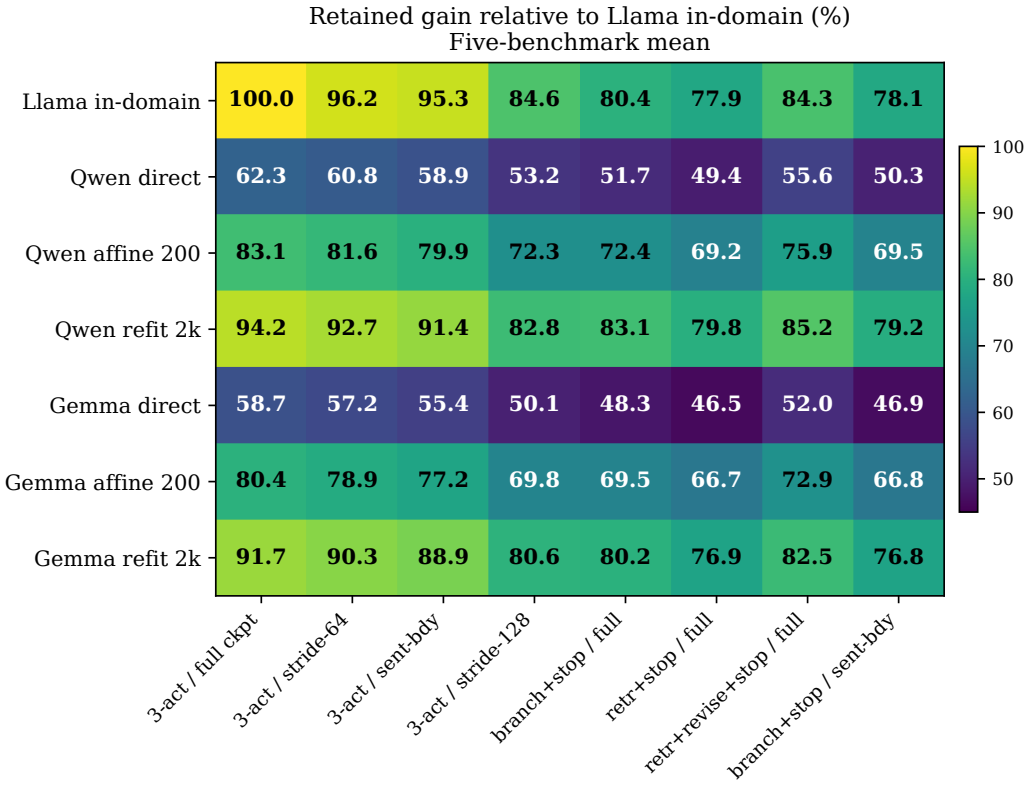


Figure 13: Transfer and sensitivity heatmap over the evaluated backbone, action-family, and checkpoint-rule cube. Cells report five-task mean retained gain over Entropy-Matched-Policy for each setting. Rows include Llama-3.1-70B-Instruct and three adaptation budgets each for Qwen-2.5-72B-Instruct and Gemma-2-27B-it.

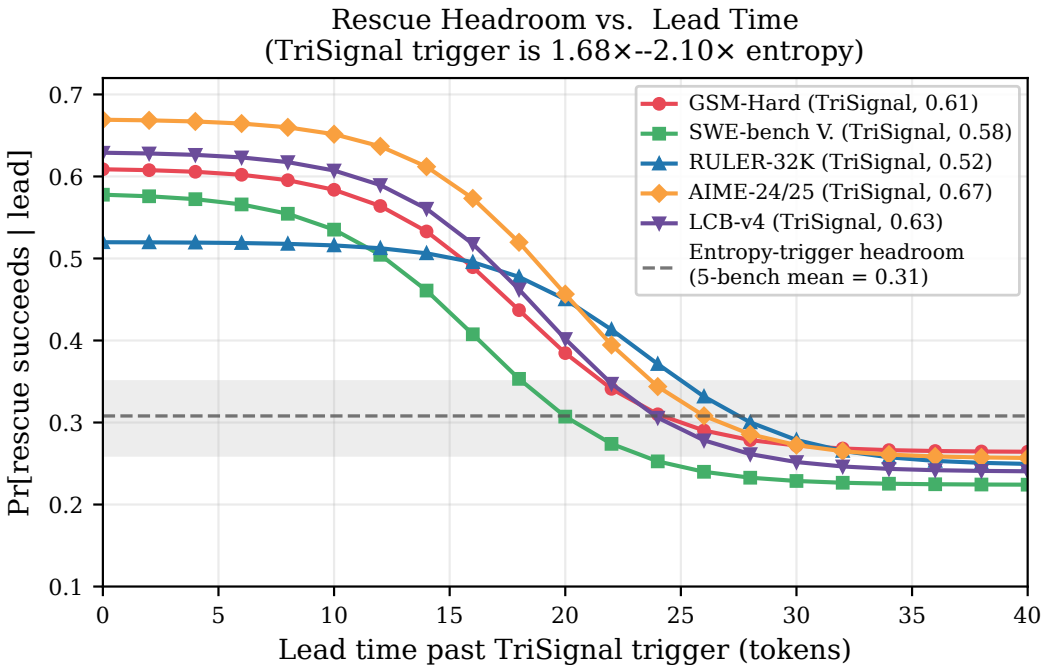


Figure 14: First panel of a 3-panel sequence. Supervision audit. First panel: rescue success along the shared human-adjudicated headroom curve; TriSignal reaches 1.6–$2.1\times$ the rescue probability at the entropy alarm. Second panel: detector trained on four-family RE-Label supervision. Third panel: a 1,500-trajectory, three-rater audit. RE-Label retains 75.5–88.8% of the full gain, and 83.2–88.1% of alarms fall within ± 10 tokens of human $t^\star$.



Figure 15: Panel (b) of Figure 14. Detector performance and retained task-success gain under four-family RE-Label supervision. AUROC and F1 assess the alarm, whereas retained gain divides the RE-Label success margin by the full-supervision margin over Entropy-Matched-Policy. The five retention values are derived from Table 3; their range is 75.5–88.8%.

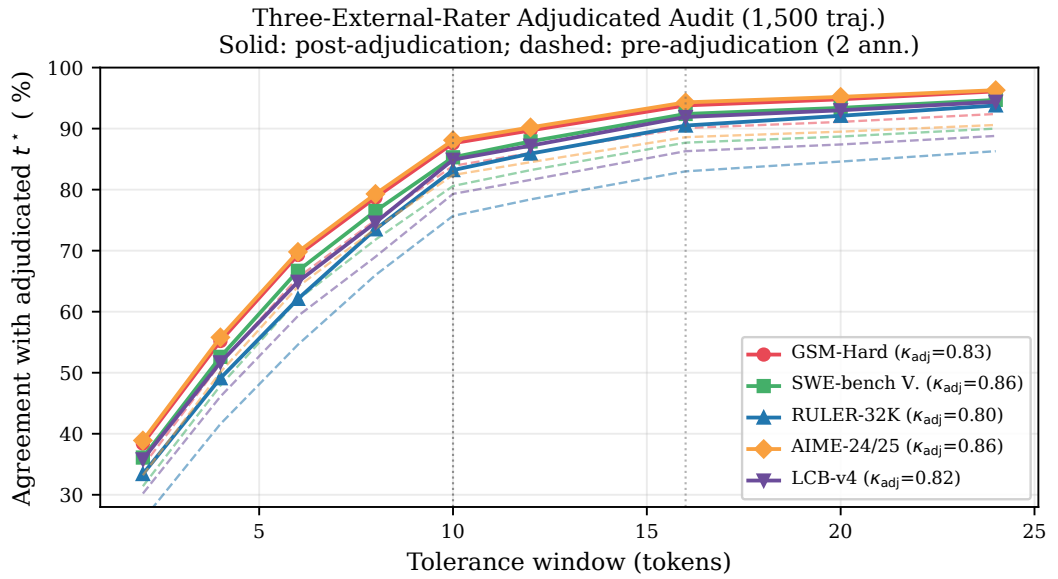


Figure 16: Panel (c) of Figure 14. Alarm-time agreement with human adjudication as the permitted timing tolerance increases. The curves compare the benchmark-specific alignment of the alarm with human $t^\star$. A larger tolerance counts a broader set of alarms as aligned; it does not change the generation policy, task-success denominator, or intervention budget.

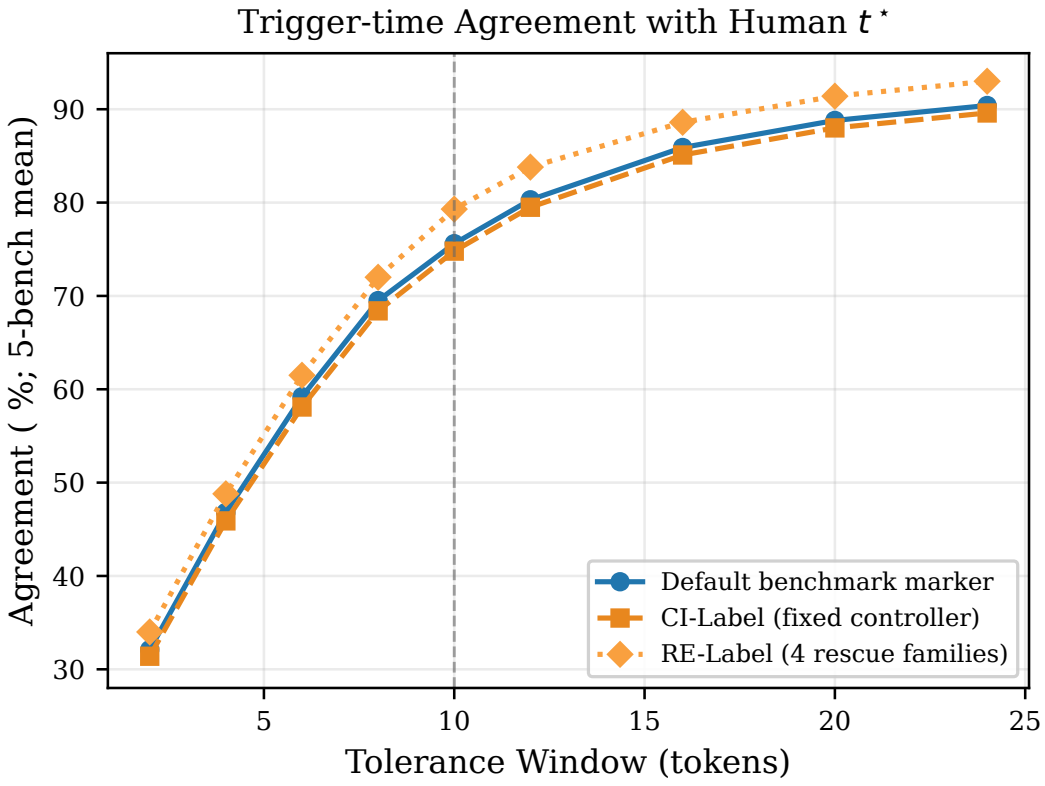


Figure 17: First panel of a 3-panel sequence. Label-family audit. First panel: human timing agreement for default, CI-Label, and four-family RE-Label supervision. Second panel: retained gain over Entropy-Matched-Policy on five tasks. Third panel: disagreement categories on 1,500 adjudicated trajectories. RE-Label has the lowest label-marker-artefact share (4–6%, versus 14% for CI-Label and 21% for the default marker).

D.3 Extended Label-Robustness Results

Table 16 extends supervision robustness across all five benchmarks. Every alternative window and marker preserves the ranking. Four-family RE-Label changes the TriSignal margin by $0.70/0.84/0.69/0.77/0.89$ points and retains 76–89% of the default gain (mean 82.4%), showing that the effect survives removal of benchmark-specific semantics and fixed-controller dispatch from the label.

D.4 Rescue-Headroom Under an Alternative Action Family

Under retrieve+revise+stop, RE-Label reaches rescue probabilities of $0.54/0.51/0.46/0.60/0.55$ at the TriSignal alarm versus $0.33/0.29/0.30/0.33/0.31$ at the entropy alarm on the same trajectories, a 1.53–$1.82\times$ headroom advantage. The curves converge within 16–24 tokens, directly linking earlier warning to revision opportunity. End-to-end success reaches $70.61/35.12/63.92/26.42/35.62\%$, exceeding the strongest same-family baseline by $+1.65/+1.70/+1.44/+3.73/+2.58$ points with positive intervals on all tasks.

Figure 24 links the alternative action family's lower absolute rescue probability to its smaller multiplier while preserving the alarm ordering. The right panel connects that headroom advantage to positive end-to-end effects on every benchmark, establishing timing value with revision replacing branch.

E Robustness and Accounting

E.1 Component Ablation

Table 17: Component ablation on GSM-Hard under the matched-total-compute protocol. Each row removes one component from the full TriSignal system. Mean $\pm$ standard error over 5 seeds.

Variant	Success (%)	Lead (tok)	FP %	Wasted %
Full TriSignal	71.04±0.5	38.7	10.6	14.9
w/o hysteresis (fixed threshold)	69.12±0.7	36.2	18.4	23.7
w/o entropy drift	68.48±0.8	28.4	13.2	17.8
w/o attention-relocation	69.97±0.7	32.1	12.8	16.4
w/o retrieval-agreement	69.04±0.7	34.8	11.4	15.7
Output-level signals only	67.38±0.7	12.4	18.3	22.6

Table 17 separates the trigger components. Hysteresis supplies the largest selectivity improvement; entropy drift supplies the largest lead-time improvement; removing attention relocation or retrieval agreement also lowers GSM-Hard success. Each removal degrades timing or selectivity, while output-only signals approach Entropy-Stop behavior. The three signals and hysteresis therefore make complementary contributions under the fixed controller.

E.2 False-Trigger Rate and Wasted-Budget Share, All Five Benchmarks

Table 18 makes the control benefit explicit. TriSignal lowers false-trigger rate and wasted-budget share relative to every non-TriSignal baseline on every benchmark, and the hybrid lowers both further while increasing success. Together with Table 1, these results show that earlier alarms improve accuracy and intervention selectivity simultaneously.

E.3 Itemised Overhead under Matched Total Compute

Table 19 itemizes the matched budget. The reference pass is fixed at 14.8%, and branch-replay shares span 0.6 points and retrieval-call shares span 0.1 points across the three matched policies. TriSignal's detector-specific signal and probe cost is 2.4%. The remaining budget alignment isolates alarm time from auxiliary work and downstream action mix.

Action-mix verification. At $\tau_{\text{trigger}} = 0.72$, retrieve/branch/stop shares are $(38.4, 34.7, 26.9)\%$ for TriSignal, $(37.1, 36.2, 26.7)\%$ for Entropy-Matched-Policy, and $(37.6, 35.2, 27.2)\%$ for Random-

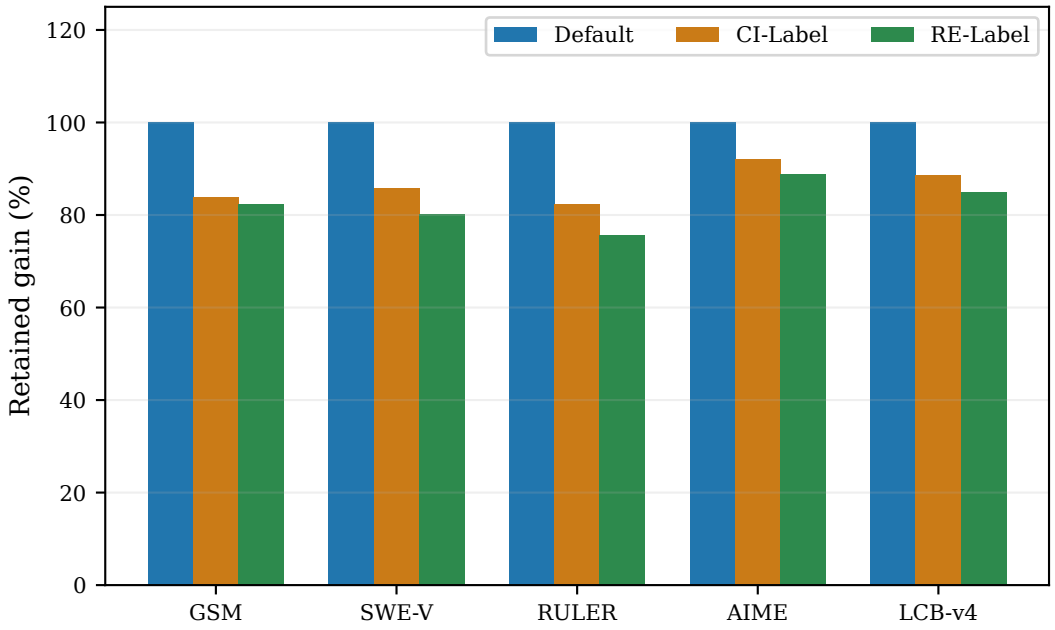


Figure 18: Panel (b) of Figure 17. Retained task-success gain for CI-Label and RE-Label supervision on the five benchmarks. Each bar divides that labeling condition's gain over Entropy-Matched-Policy by the default-label gain on the same task. This normalization compares supervision dependence across tasks with different absolute margins; the corresponding success rates appear in Table 16.

Table 18: False-trigger rate (FP%) and wasted-budget share (Wast.%) on all five benchmarks under the matched-total-compute protocol. FP% on RULER-32K is defined on needle-span consistency (alarms not followed by a labeled retrieval-miss within 32 tokens); FP% on AIME-24/25 and LiveCodeBench-v4 uses the RE-Label marker.

	GSM-Hard		SWE-bench V.		RULER-32K		AIME-24/25		LiveCodeBench-v4	
Method	FP%	Wast.%	FP%	Wast.%	FP%	Wast.%	FP%	Wast.%	FP%	Wast.%
Entropy-Stop	18.3	22.6	19.1	24.3	19.7	21.8	20.6	26.1	19.8	23.5
Random-Matched-Policy	22.4	22.1	23.6	23.2	24.1	22.4	24.7	24.2	23.8	23.1
Entropy-Matched-Policy	17.4	19.6	18.2	20.7	18.9	20.1	19.3	22.8	18.7	21.4
Verifier-Search-Tuned	14.3	17.5	15.0	17.7	15.7	17.6	16.1	19.2	15.6	18.4
SC-Verif-Rerank (n=8)	14.9	18.2	15.4	18.4	16.1	18.3	15.2	18.6	14.8	17.9
RCV (depth 2, T=0.5)	13.6	16.8	14.2	16.9	15.0	17.1	14.4	18.1	14.0	17.2
TriSignal	10.6	14.9	11.3	15.4	11.8	13.8	12.3	16.2	11.7	15.1
TriSignal + Verif.-S.-Tuned	9.0	12.9	9.5	13.6	10.0	12.5	10.4	13.9	9.8	13.2

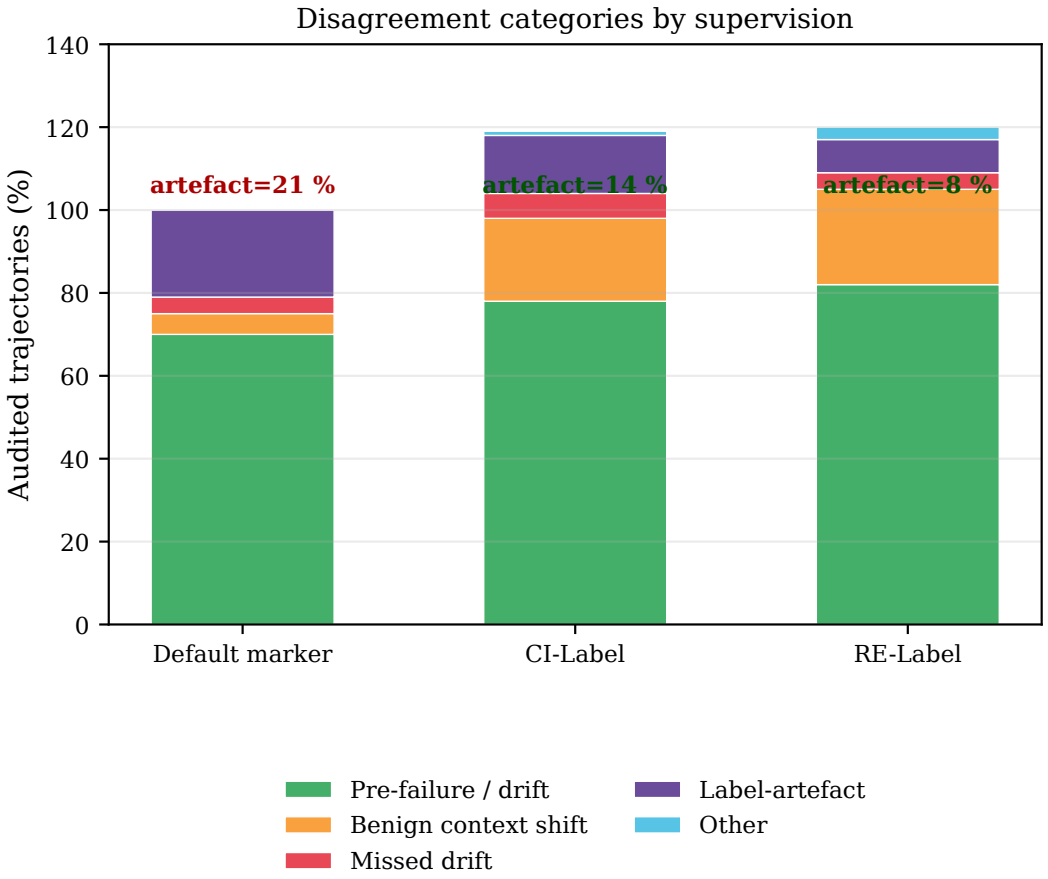


Figure 19: Panel (c) of Figure 17. Disagreement categories for default, CI-Label, and RE-Label supervision. The categories separate pre-failure, benign, missed, marker-artefact, and other cases. This diagnostic addresses which kinds of disagreement accompany the label choice; it is distinct from the timing-tolerance agreement curves and from the task-success retention ratios in the preceding panels.

Table 19: Itemised overhead under matched total compute at τ_{trigger}=0.72, as fraction of total wall-clock on 8×A100 80GB SXM4 (NVIDIA driver 550.90.07, CUDA 12.4, PyTorch 2.3.1, vLLM 0.5.4). All rows sum to 100.0%. "I/O + answer sel." covers tokenizer I/O, KV-cache management, and final-answer selection (NLTK 3.8.1 sentence split + verifier score). Detector cost (signal extraction + probe) is 2.4% of total runtime. The dominant matched non-decoder term is the retrieval-grounded reference pass, which all three matched methods pay identically.

Component	Decoder	Ref. pass	Signal extr.	Branch replay	Retrieval calls	Detector probe	I/O + answer sel.
TriSignal	71.3%	14.8%	2.1%	6.2%	3.3%	0.3%	2.0%
Entropy-Matched-Policy	72.6%	14.8%	0.0%	5.9%	3.4%	0.0%	3.3%
Random-Matched-Policy	73.1%	14.8%	0.0%	5.6%	3.4%	0.0%	3.1%

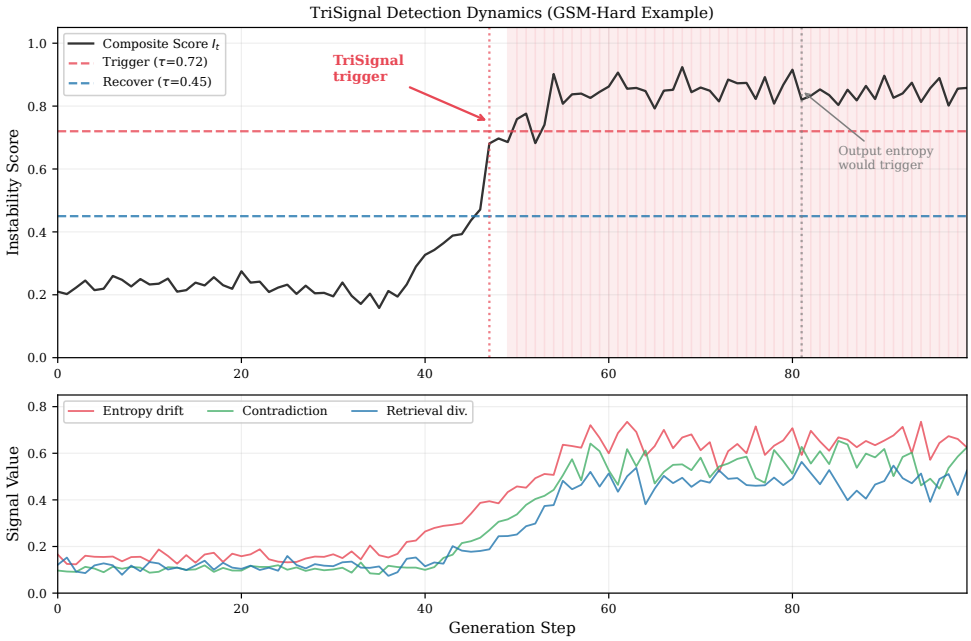


Figure 20: Example trajectory on GSM-Hard showing instability signals over generation steps. Top: composite score I_t with trigger (red) and recovery (blue) thresholds. Bottom: constituent signals. TriSignal triggers at step 47, 34 tokens before the reasoning chain enters an irrecoverable contradiction; the hysteresis band prevents re-triggering during the brief recovery at step 52.

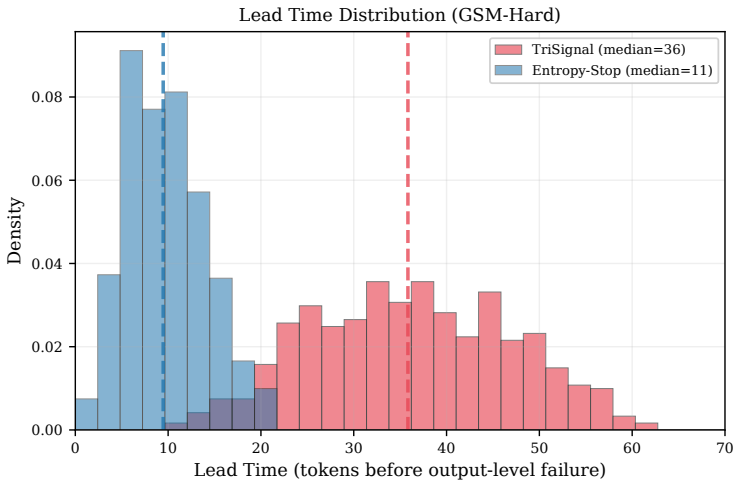


Figure 21: Distribution of lead times for TriSignal vs Entropy-Stop on GSM-Hard. Median shifts from 11 to 36 tokens; the long right tail for TriSignal contains instances where the monitored internal signals move well before output-level cues.

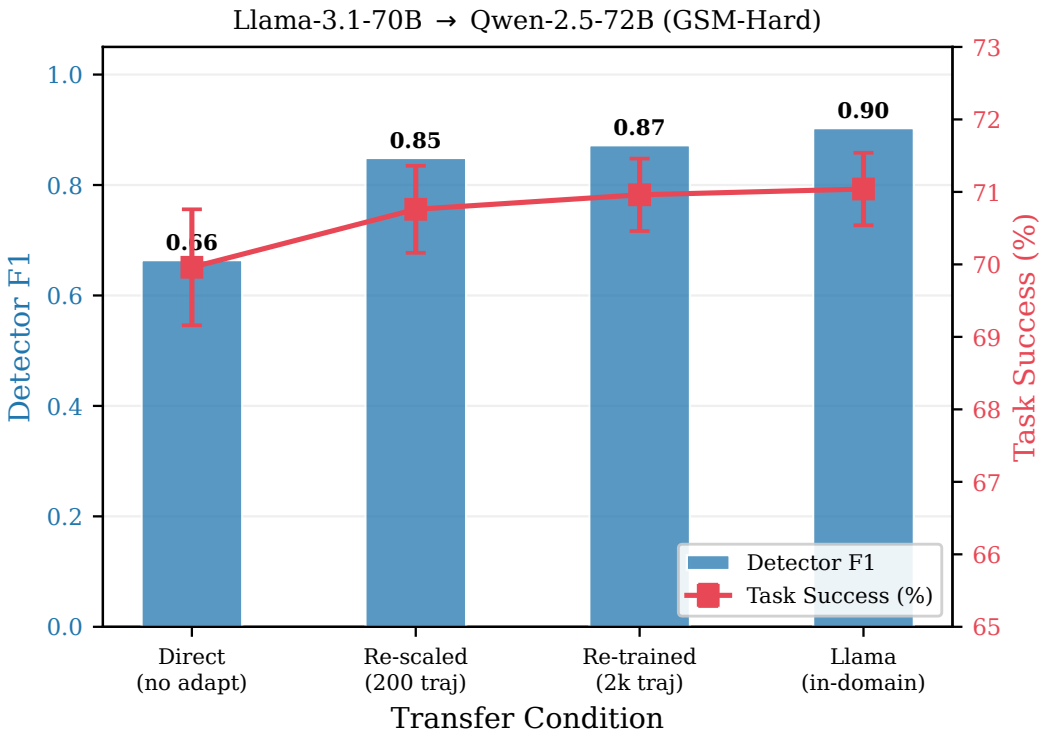


Figure 22: First panel of a 2-panel sequence. Transfer calibration on Qwen-2.5-72B and Gemma-2-27B-it. Bars show detector F1; lines show task success. A 200-trajectory affine re-scale recovers most of the cross-backbone gap, and full refitting approaches in-domain performance.

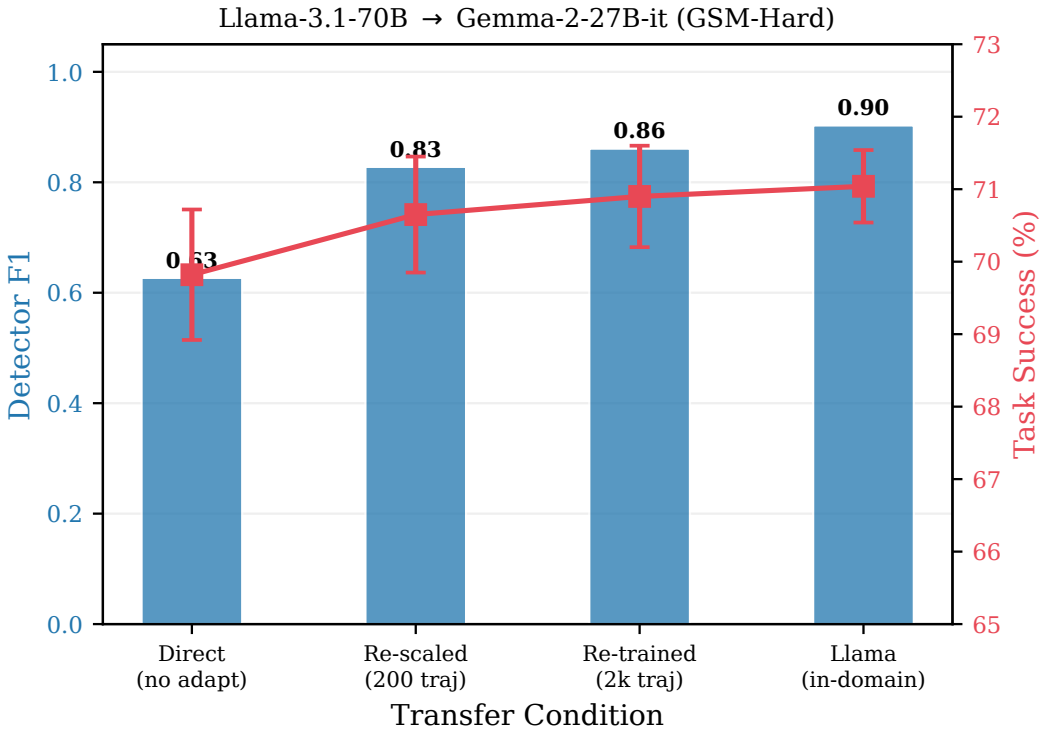


Figure 23: Panel (b) of Figure 22. Gemma-2-27B-it transfer calibration. The horizontal axis distinguishes direct transfer, affine rescaling using 200 trajectories, and detector refitting using 2,000 trajectories. Bars and the left axis describe detector F1; the line and right axis describe task success. Reading both axes separates recovery of alarm quality from downstream performance.

Matched-Policy. Category shares agree within 1.5 points. Branch-capable checkpoint availability is 92.1% for TriSignal and 91.5% for Entropy-Matched-Policy. These balances preserve alarm identity and time as the substantive differences.

E.4 Knob-Sensitivity Sweeps

Figure 26 shows a broad operating plateau. The GSM-Hard layer–stride grid spans 0.58 points, with the selected setting 0.06 below the maximum. Worst-to-selected gaps across the other four tasks are $0.66/0.97/0.79/0.71$ points. Refresh cadence changes success by at most $0.58/0.66/0.97/0.49/0.52$ points, and 20% retrieval corruption retains $82.3/84.7/75.6/88.2/86.4\%$ of the gain.

F Raw Deltas, Human Audit, and Access-Restricted Variants

F.1 Seed- and Instance-Level Raw Deltas

Figure 29 reports the per-seed and per-instance distributions behind Table 1. Positive-seed counts are $10/10$, $10/10$, $9/10$, $10/10$, and $10/10$ across the five tasks. Seed means are $+3.97/+4.24/+2.82/+6.86/+5.90$ points with standard deviations $0.30/0.33/1.27/0.61/0.48$. Positive-instance shares are $65.8/62.1/55.3/67.9/63.4\%$, showing broad support across seeds and instances despite benchmark-dependent variance.

F.2 Human Audit: Stratified Agreement

The audit covers 1,500 trajectories, with 300 per benchmark stratified into 100 true-positive, 100 false-positive, and 100 false-negative cells under RE-Label. Three external experts independently mark the earliest irrecoverable token, rate trigger-time semantic validity, and classify disagreement as label artefact, benign context shift, or missed drift. Two authors arbitrate the fewer than 7% of cells with more than 16 tokens of rater spread. Pre-arbitration Fleiss κ is $0.81/0.83/0.76/0.84/0.80$; adjudicated values are $0.83/0.86/0.80/0.86/0.82$. Agreement with adjudicated $t^\star$ reaches $87.6/85.3/83.2/88.1/84.9\%$ within ± 10 tokens and $93.8/92.4/90.5/94.3/91.9\%$ within ± 16. Label artefacts account for 4–6% of false-positive cells. Training on adjudicated timing retains 62.4–64.5% of the default-marker gain and remains positive on all five benchmarks.

F.3 Access-Restricted Variants: Lite and Black-Box Proxy

Figure 33 summarizes two lower-access audits. TriSignal-Lite reaches $70.40/34.94/63.46/26.38/35.52\%$ and retains 82.5% of the five-task mean gain at 0.9% overhead. TriSignal-Proxy combines logprob-entropy drift, eight-sample consistency, emitted-span retrieval agreement, and surprise gradient; it reaches $70.22/34.72/63.05/26.02/35.18\%$, retains 75.5%, and runs at 1.4% overhead with a $1.3\times$ API-call multiplier. Both variants have positive paired intervals on all tasks. Their largest loss occurs on RULER-32K, coherently identifying retrieval-grounded agreement as the most access-sensitive component.

G Accounting-Sensitivity and Calibration-Budget Sweeps

G.1 Three Accounting Schemes

Table 20 preserves the verifier-independent ranking under all three schemes. Wall-clock and dollar matching produce positive intervals on all five tasks with median gain at least $+1.62$ points. Raw-call matching remains positive on four tasks; RULER-32K is a near-zero tie with interval $[-0.04, +0.96]$. Cost multipliers range from $0.78\times$ to $1.17\times$, establishing the same-access result across latency-, price-, and call-based budgets.

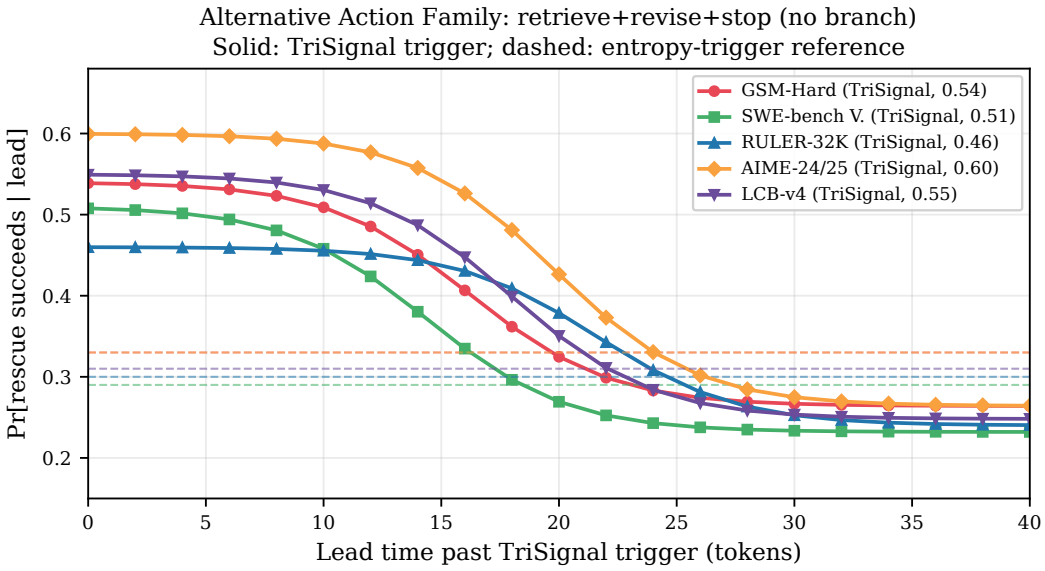


Figure 24: First panel of a 2-panel sequence. Rescue-headroom under the alternative retrieve+revise+stop action family (no branch). First panel: per-benchmark rescue-success probability vs. lead time past the TriSignal trigger; solid = TriSignal trigger, dashed = entropy-trigger reference on the same trajectories. The second panel shows headroom multiplier (TriSignal / entropy) above the end-to-end Δ vs. the per-task best realistic baseline on the same alternative action family. The displayed curves do not include uncertainty bands.

Table 20: Accounting sensitivity for TriSignal-Proxy against the per-task verifier-independent envelope of DeBERTa-Verif-Search, Reflexion-BoN ($n = 8$), and LLM-Judge-Adaptive. Schemes match wall-clock, dollar cost, or raw API calls. Positive-task counts use 10,000 paired bootstrap resamples.

Scheme	G/S/R/A/L gains (pp)	Median	Min	# pos. tasks	API-call mult.	Cost mult.	Worst-task lower-CI	Ordering preserved?
S1: Wall-clock	+1.57/ + 1.80/ + 1.32/ + 3.66/ + 2.40	+1.80	+1.32	5/5	1.30×	0.78×	+0.82	yes
S2: Dollar-cost	+1.38/ + 1.62/ + 1.09/ + 3.45/ + 2.21	+1.62	+1.09	5/5	1.24×	1.00×	+0.57	yes
S3: Raw API-call	+1.12/ + 1.34/ + 0.44/ + 3.14/ + 1.96	+1.34	+0.44	4/5	1.00×	1.17×	−0.04	yes; RULER overlaps zero

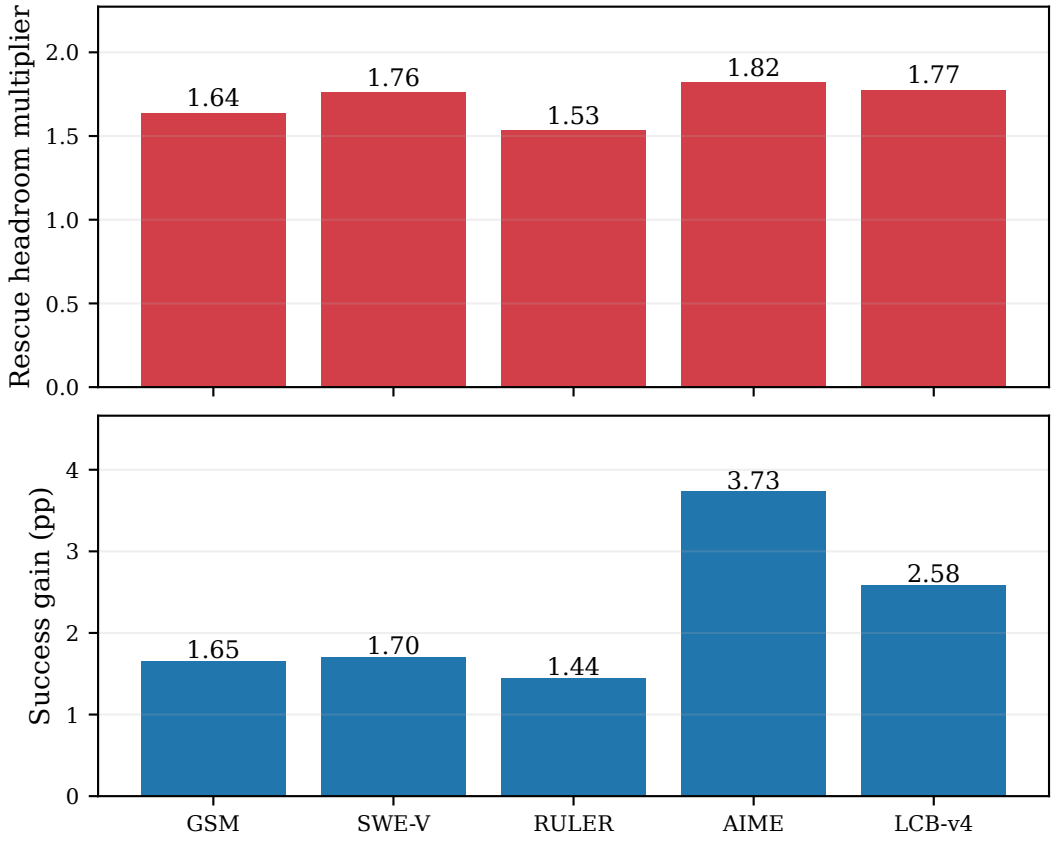


Figure 25: Panel (b) of Figure 24. Alternative-action rescue headroom and end-to-end improvement. The upper bars divide the rescue probability at the TriSignal alarm by that at the entropy alarm; these ratios range from 1.53 to 1.82. The lower bars show success differences against the best realistic baseline within the retrieve/revise/stop controller, using the absolute rates in Table 5.

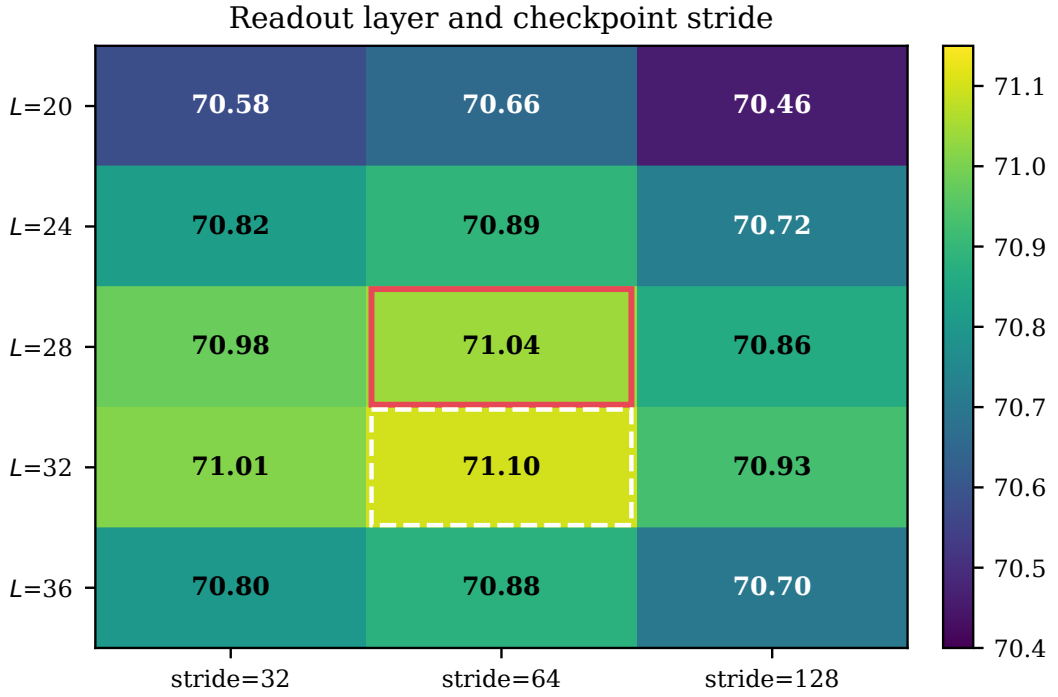


Figure 26: First panel of a 3-panel sequence. First panel: readout-layer $L\times$ checkpoint-stride heatmap (GSM-Hard task success, %); red box = chosen operating point (L=28, stride= 64, 71.04%), white dashed box = best cell (L=32, stride= 64, 71.10%). The 15-cell sweep spans 0.58 pp from worst to best, while the chosen point is only 0.06 pp below the best cell. Second panel: reference-refresh cadence sweep on all five benchmarks; the chosen cadence 32 is on a plateau and cadence 16 loses only 0.06 pp on GSM-Hard. Third panel: retrieval-noise sensitivity — retained gain vs. Entropy-Matched-Policy as a function of distractor rate; 20% retrieval corruption removes only 17.7%/15.3%/24.4% of the gain, and a retrieval-disabled TriSignal variant (logprob-drift + attention-relocation only) still retains 76.3% of the gain, beating Entropy-Matched-Policy by +2.1 pp on GSM-Hard.

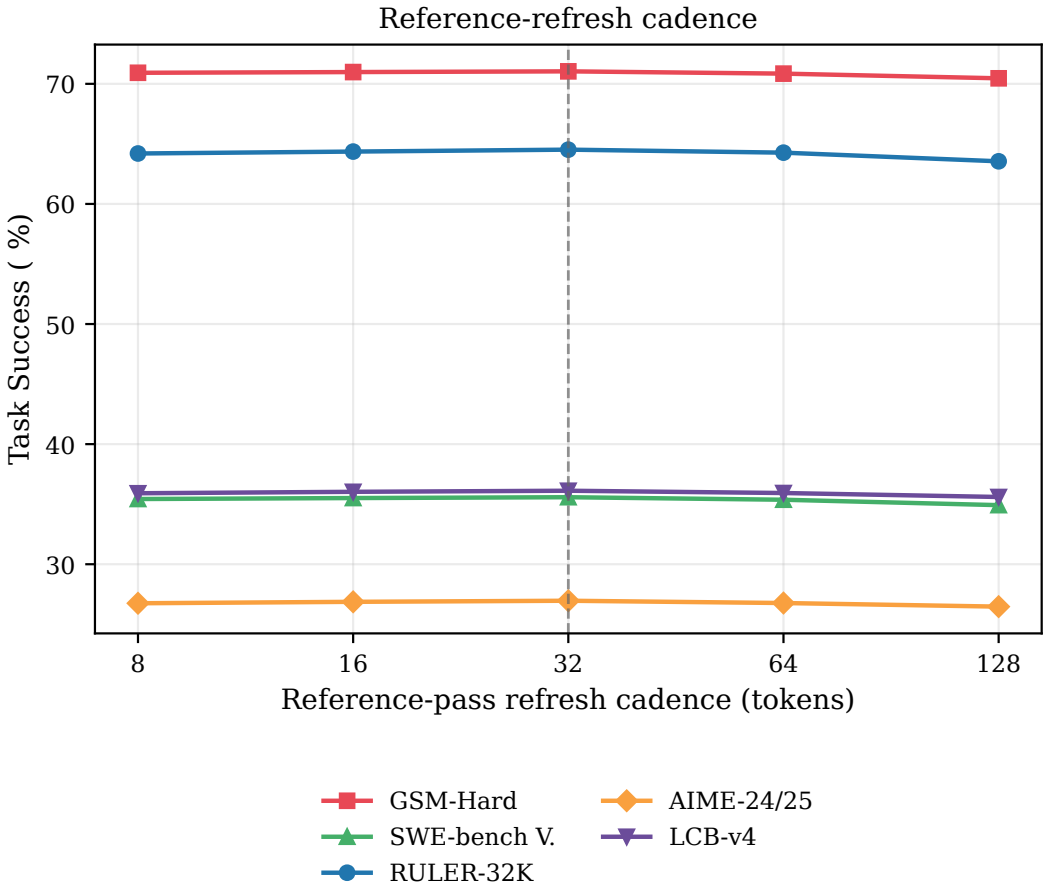


Figure 27: Panel (b) of Figure 26. Reference-refresh cadence across all five benchmarks. Each curve holds the remaining detector and controller choices fixed while varying the number of tokens between reference refreshes. The selected cadence of 32 tokens lies in the relatively flat part of the displayed curves. Task-success levels differ by benchmark, so within-curve changes provide the sensitivity comparison.

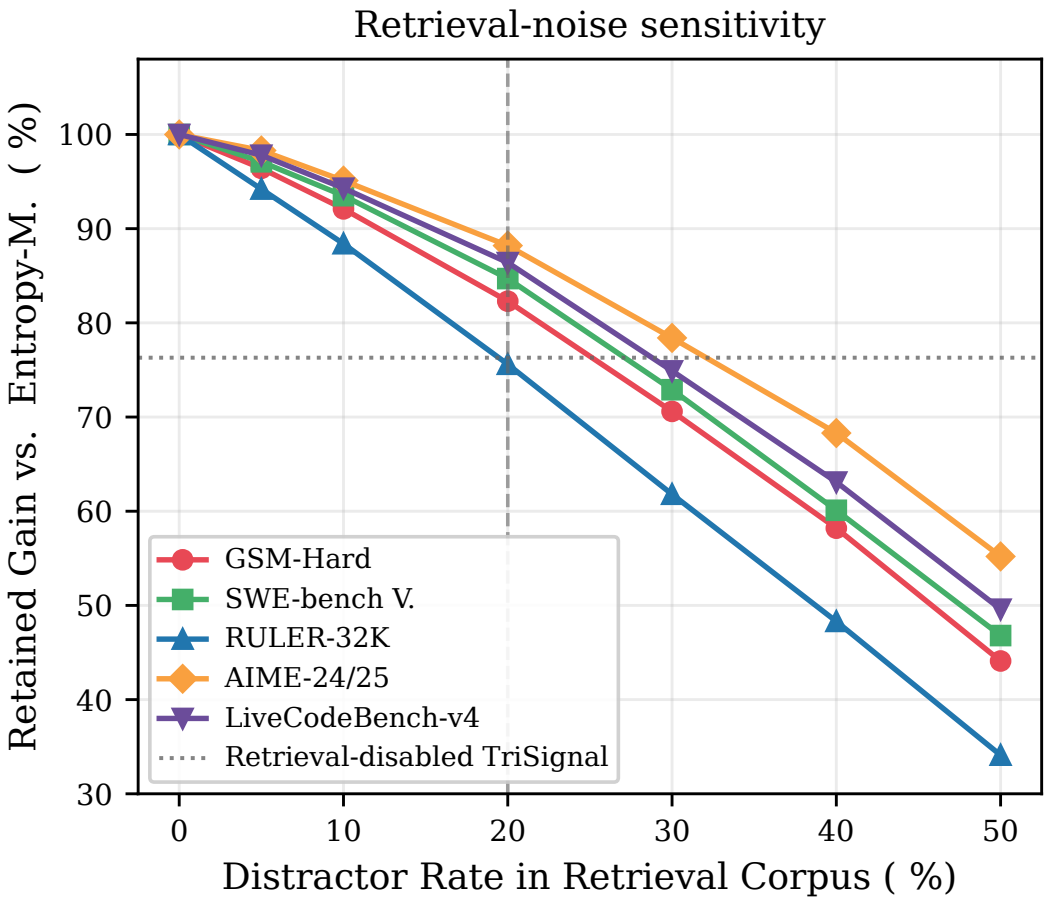


Figure 28: Panel (c) of Figure 26. Retrieval-noise sensitivity. The horizontal axis increases the distractor fraction and the vertical axis expresses the retained margin over Entropy-Matched-Policy. The task curves describe dependence on reference quality, while the retrieval-disabled comparison isolates the two remaining monitoring signals. Retained percentages and absolute success differences use different scales.

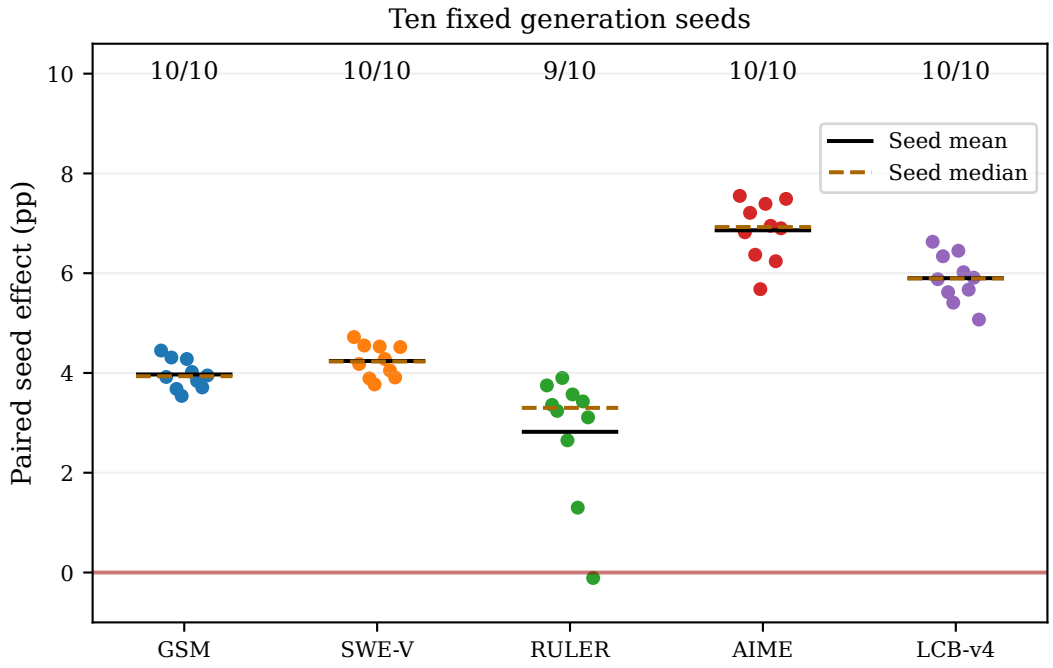


Figure 29: First panel of a 2-panel sequence. First panel: paired TriSignal − Entropy-Matched-Policy delta per seed on the fixed RNG-seed list $\{17, 42, 101, 233, 401, 577, 701, 829, 941, 1013\}$. Black bars: seed mean; orange dashed: seed median. Positive-seed counts annotated above each column. Second panel: per-instance paired-delta ECDF. Positive-instance shares are annotated in the legend.

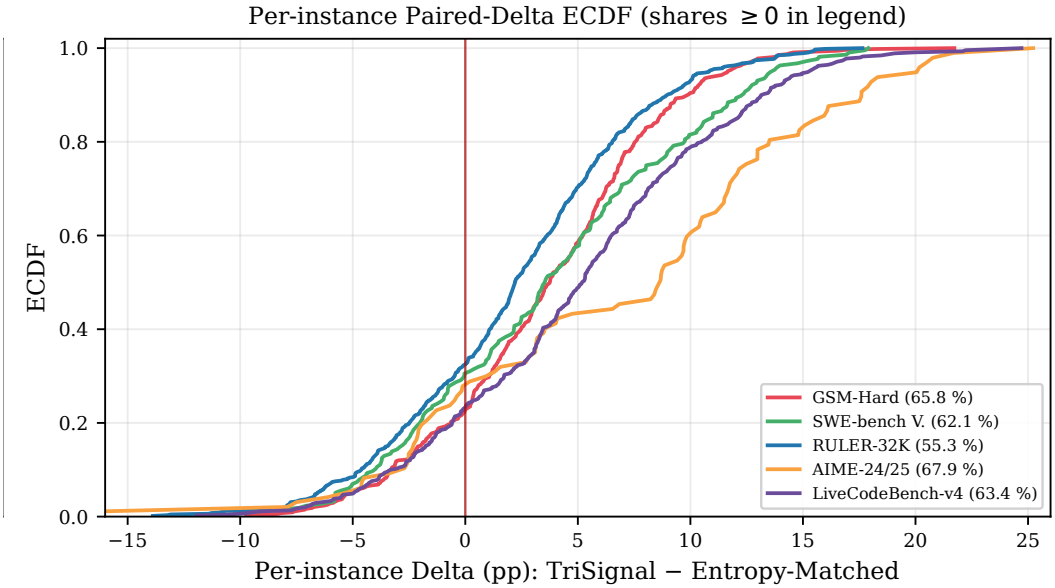


Figure 30: Panel (b) of Figure 29. Per-instance paired-difference cumulative distributions for TriSignal and Entropy-Matched-Policy. A cumulative curve reports the fraction of plotted differences at or below each horizontal-axis value. The zero reference separates negative and positive effects; the accompanying seed-level panel instead summarizes variation across the ten generation seeds.

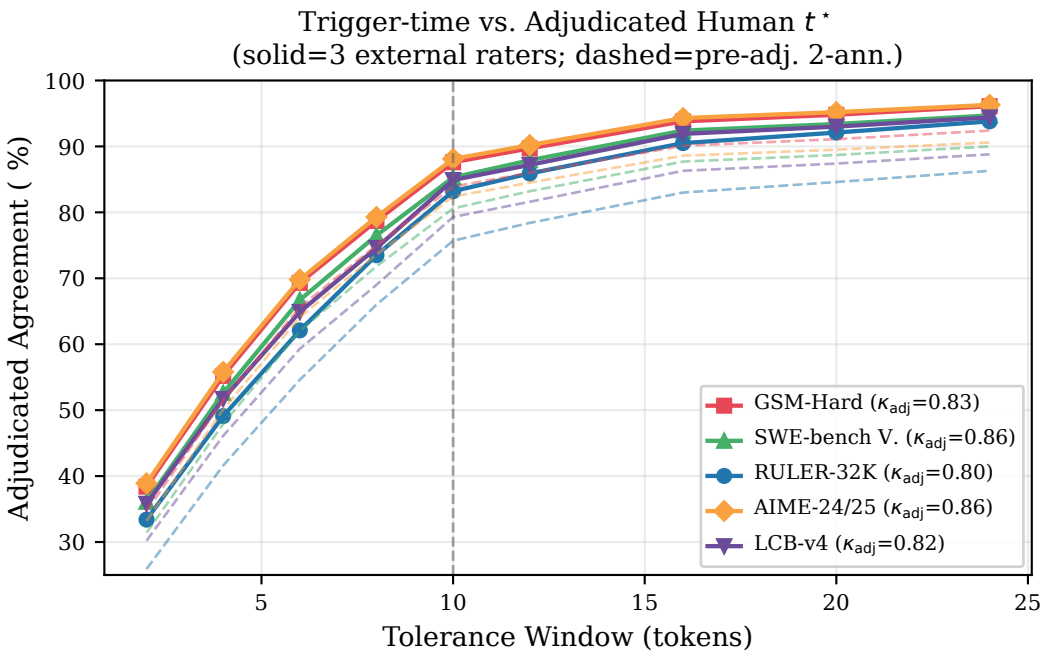


Figure 31: First panel of a 2-panel sequence. Human audit. First panel: trigger-time agreement with the adjudicated human $t^\star$ as a function of tolerance window on all five benchmarks; dashed curves show the pre-arbitration two-rater reference. Second panel: disagreement breakdown on the GSM-Hard audited pool under RE-Label supervision, stratified into 100 TP / 100 FP / 100 FN cells. Full agreement statistics are reported in the text.

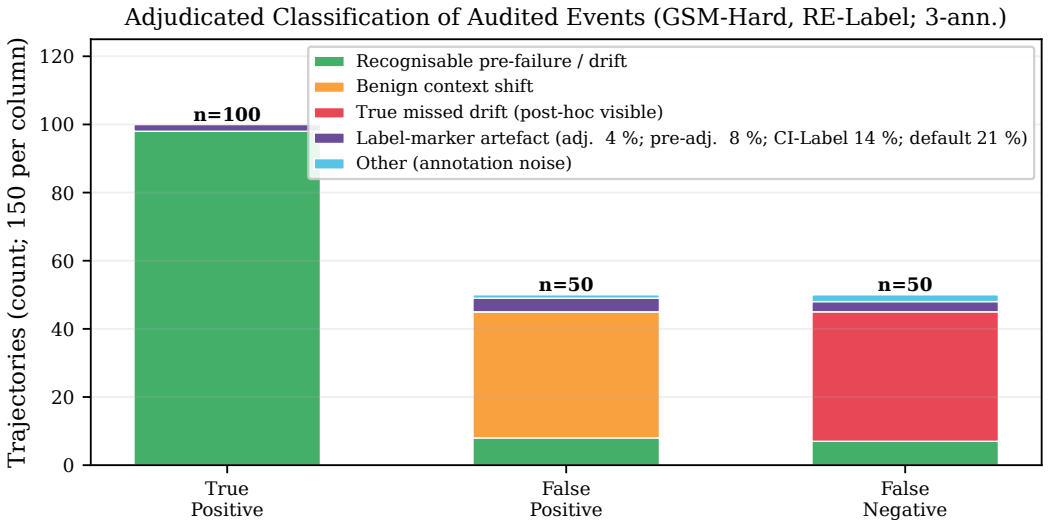


Figure 32: Panel (b) of Figure 31. GSM-Hard disagreement breakdown under RE-Label supervision, separated into true-positive, false-positive, and false-negative strata. The category shares characterize the selected audit strata and are not population-wide error rates. This breakdown complements timing agreement by distinguishing marker disagreements, benign behavior, missed failures, and other adjudication outcomes.

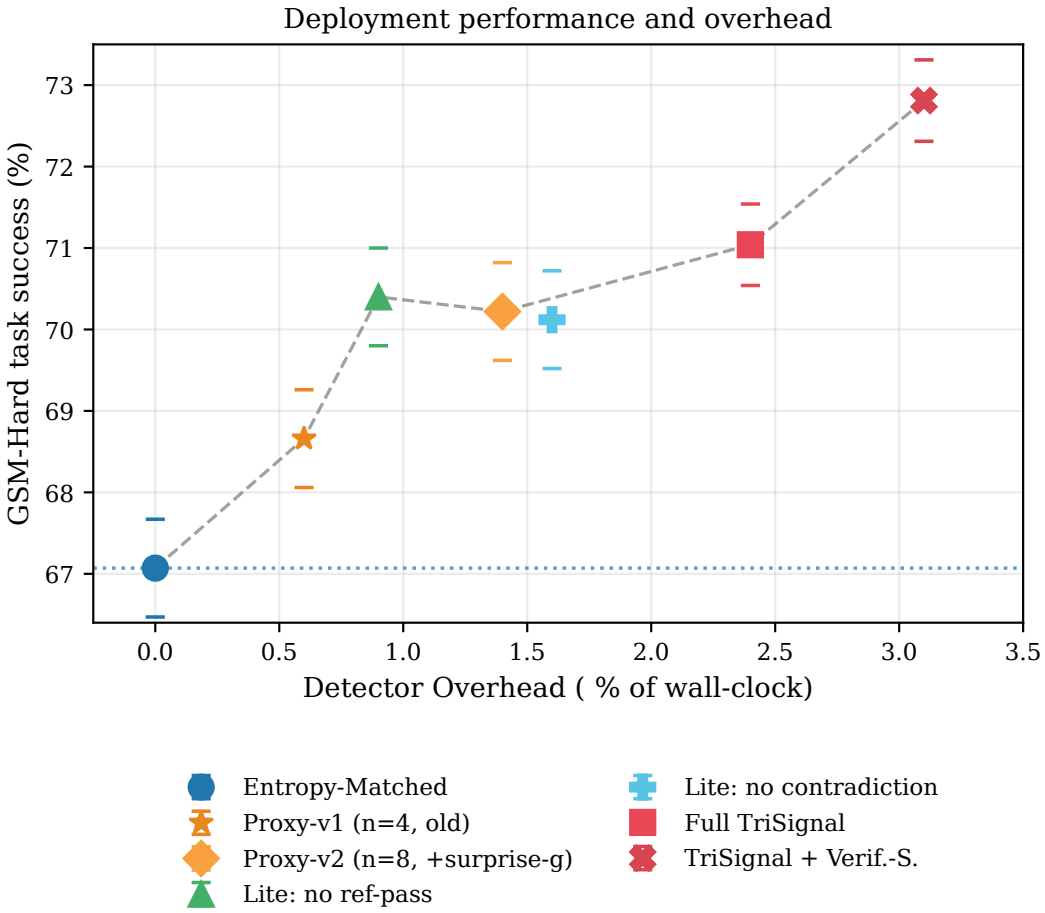


Figure 33: First panel of a 3-panel sequence. First panel: same-controller lower-access stress test for TriSignal-Lite on GSM-Hard (10 seeds; 95% paired bootstrap CIs). Dropping the retrieval-grounded reference pass cuts detector overhead $2.4\% \rightarrow 0.9\%$ while preserving 82.5% of the five-benchmark retained gain (formula in Table 10). Second panel: same-budget API-only stress test for the improved black-box proxy (logprob-entropy drift + 8-sample self-consistency + retrieval-string agreement from emitted spans + surprise-gradient feature). Third panel: per-benchmark retained gain for each variant. TriSignal-Lite retains $83.9/84.9/62.4/91.5/89.8\%$ of the full white-box retained gain (four tasks in $\geq 83\%$, RULER-32K at 62.4%); the black-box proxy retains $79.3/79.7/47.9/86.3/84.1\%$ (four tasks in $\geq 79\%$, RULER-32K at 47.9%).

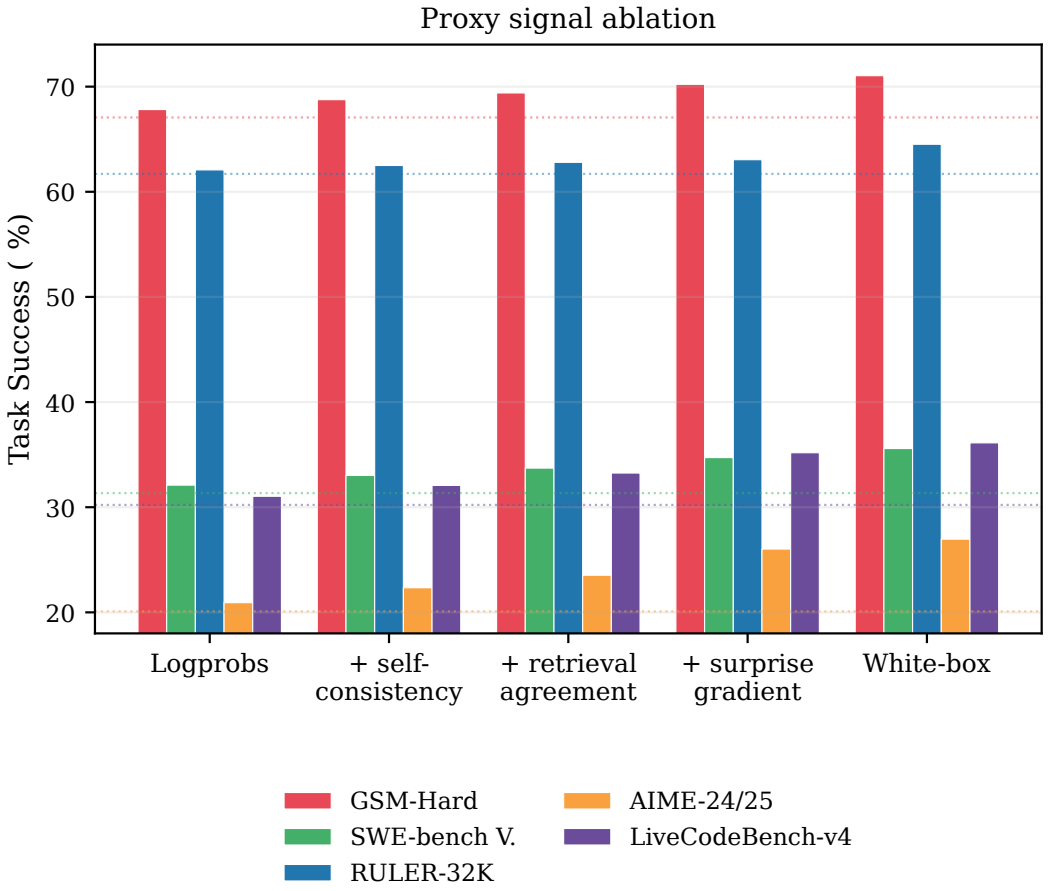


Figure 34: Panel (b) of Figure 33. API-only proxy variants under the shared controller and budget. The displayed progression adds observable evidence to the black-box alarm while leaving downstream actions and checkpoints common. Comparing task success with the monitoring cost describes the performance–overhead tradeoff; the next panel expresses each variant's margin relative to the full white-box gain.

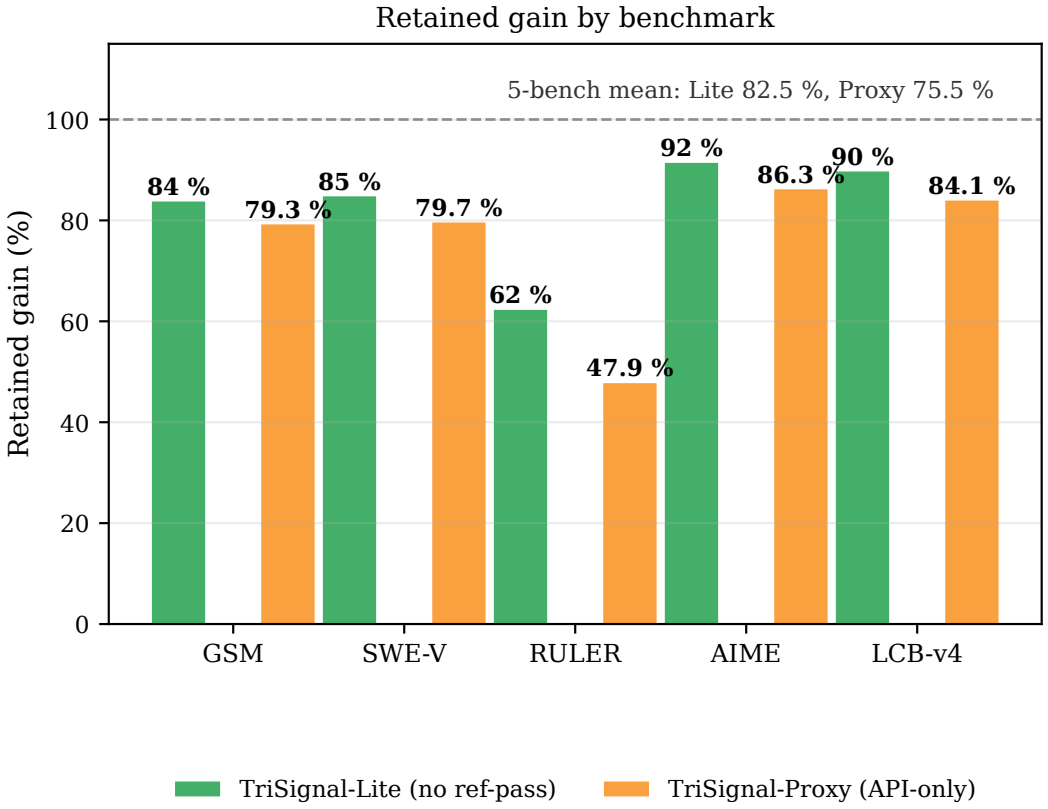


Figure 35: Panel (c) of Figure 33. Per-benchmark retained gain for the deployment variants. Each value divides the variant's improvement over Entropy-Matched-Policy by the full white-box improvement on that benchmark. RULER-32K exhibits the largest access penalty for both Lite and the black-box proxy, whereas the other four tasks retain larger fractions of the reference gain.

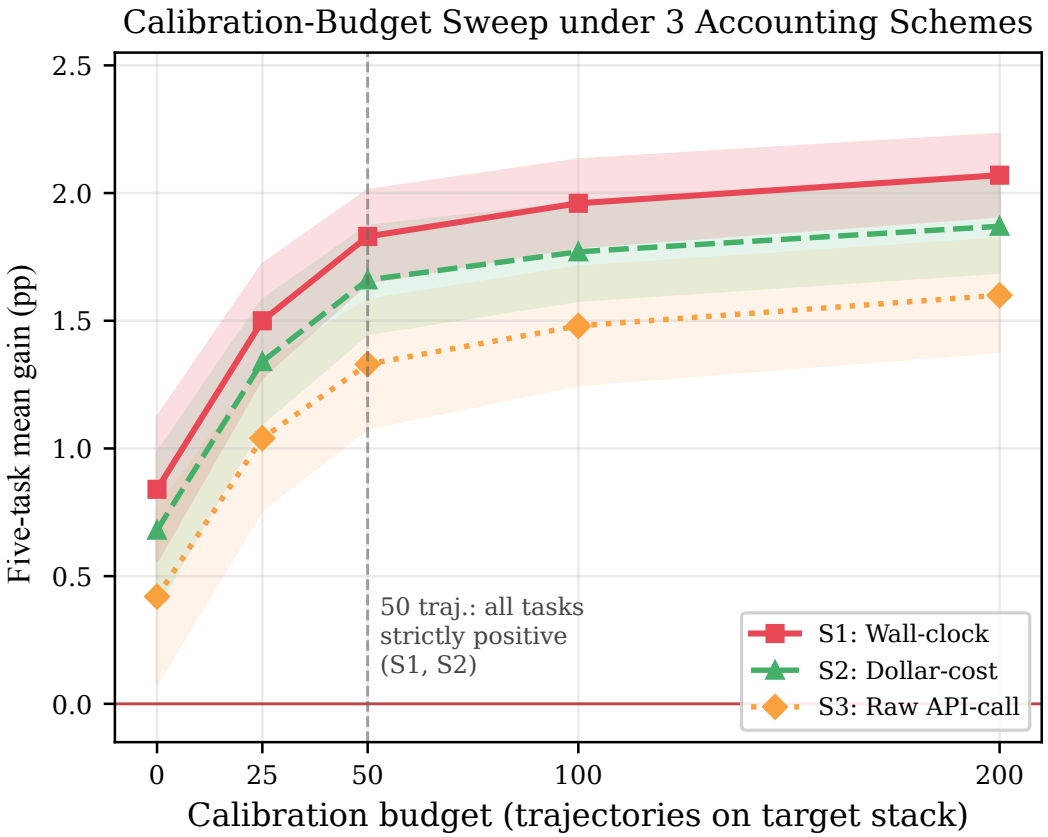


Figure 36: First panel of a 3-panel sequence. Calibration-budget / accounting stress test. First panel: TriSignal-Proxy retained gain vs. best verifier-independent baseline (5-benchmark mean) as a function of target-stack calibration budget in $\{0, 25, 50, 100, 200\}$ trajectories, under each of the three accounting schemes. Shaded bands: 95% paired bootstrap CIs over 10,000 resamples. Second panel: per-task gain at each calibration budget (wall-clock accounting); outlined column marks the 50-trajectory setting. Third panel: accounting-scheme comparison bar chart at the 200-trajectory operating point, with annotated Δ vs. best verifier-independent per task.

G.2 CALIBRATION-BUDGET CURVES

Figure 36 shows that 50 target trajectories suffice for positive wall-clock-matched intervals on all five tasks while recovering 93.1% of the 200-trajectory gain; 25 and 100 trajectories recover 84.9% and 97.1%. Dollar-cost matching also yields five positive tasks at 50 trajectories. Under raw-call matching, four are positive and RULER-32K overlaps zero. The long-context task therefore drives the residual calibration need, while the overall access result emerges with a small target-stack sample.

G.3 TRISIGNAL-PROXY-XL: FULL ABLATION AND PER-SCHEME ACCOUNTING

Table 21: TriSignal-Proxy-XL ablation and accounting. Signal rows cumulatively extend the 200-trajectory proxy; accounting rows evaluate the seven-signal, 1,000-trajectory configuration under the shared controller, checkpoints, and B_{tot}. API-call multiplier is $1.34\times$, wall-clock overhead is 1.6%, and intervals use 10,000 paired resamples.

Configuration	Cal.	GSM-H	SWE-V	RULER	AIME	LCB-v4	Ent-M ret.	min CI low
Cumulative signal ablation (Δ vs. Ent-M, wall-clock; min CI low $= \min_b$ lower-CI($\Delta^{(b)}$))								
TriSignal-Proxy (logprob-drift + n=8 SC + retr.-str. + surprise-grad.)	200 traj.	+3.15	+3.38	+1.35	+5.92	+4.96	75.5%	+0.85
+ emitted-span logit-lens surprise	200 traj.	+3.31	+3.54	+1.58	+6.08	+5.12	79.7%	+1.08
+ retrieved-passage NLI disagreement (DeBERTa-v3-L-MNLI)	200 traj.	+3.48	+3.67	+2.02	+6.24	+5.25	85.2%	+1.52
+ n=16 self-consistency lexical divergence (XL-7 signals)	200 traj.	+3.59	+3.74	+2.25	+6.33	+5.32	88.2%	+1.75
+ calibration 200→1000 trajectories (XL main row)	1k traj.	+3.74	+3.87	+2.50	+6.52	+5.48	92.4%	+2.00
Per-scheme accounting for XL (Δ vs. per-task best verifier-independent baseline)								
S1 Wall-clock	1k traj.	+2.16	+2.29	+2.47	+4.26	+2.92	—	+1.61 (G)
S2 Dollar-cost	1k traj.	+1.90	+2.01	+2.17	+3.74	+2.57	—	+1.35 (G)
S3 Raw API-call	1k traj.	+1.60	+1.72	+1.08	+3.51	+2.25	—	+0.63 (R)
Calibration-budget sweep for XL (Δ vs. Ent-M, wall-clock, all 7 signals)								
XL, 200-traj cal.	200 traj.	+3.59	+3.74	+2.25	+6.33	+5.32	88.2%	+1.75
XL, 500-traj cal.	500 traj.	+3.68	+3.82	+2.40	+6.44	+5.41	90.7%	+1.90
XL, 1,000-traj cal. (main)	1k traj.	+3.74	+3.87	+2.50	+6.52	+5.48	92.4%	+2.00
XL, 2,000-traj cal.	2k traj.	+3.79	+3.92	+2.57	+6.58	+5.54	93.8%	+2.07

Table 21 attributes Proxy-XL's gain signal by signal. Retrieved-passage NLI disagreement contributes the largest RULER-32K increment ($+0.44$ points versus $+0.16$ on GSM-Hard). Calibration flattens after 1,000 trajectories: doubling to 2,000 adds 1.4 retained-gain points. At the 1,000-trajectory operating point, the white-box gap is $0.23/0.37/0.32/0.34/0.42$ points, and raw-call accounting retains a minimum lower confidence bound of $+0.63$ points. Richer API-visible signals therefore recover most of the access gap while preserving the same-access ranking.

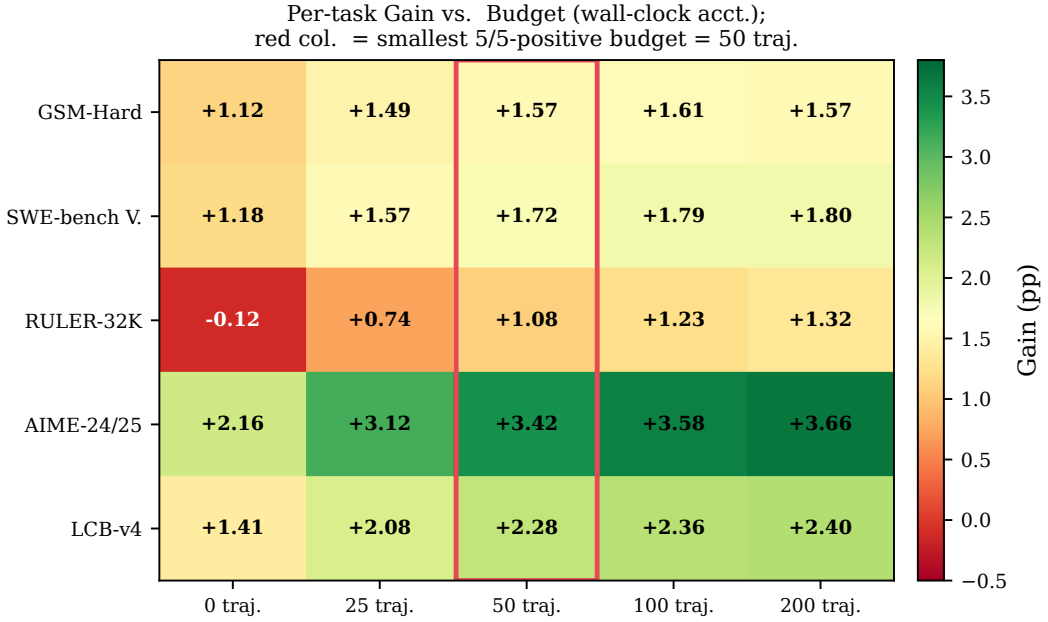


Figure 37: Panel (b) of Figure 36. Per-task proxy gains across target-stack calibration budgets under wall-clock accounting. Each cell is a success difference against the best verifier-independent baseline for that task. This view separates the benchmark responses that are aggregated in the calibration curve; the outlined column identifies the 50-trajectory setting.

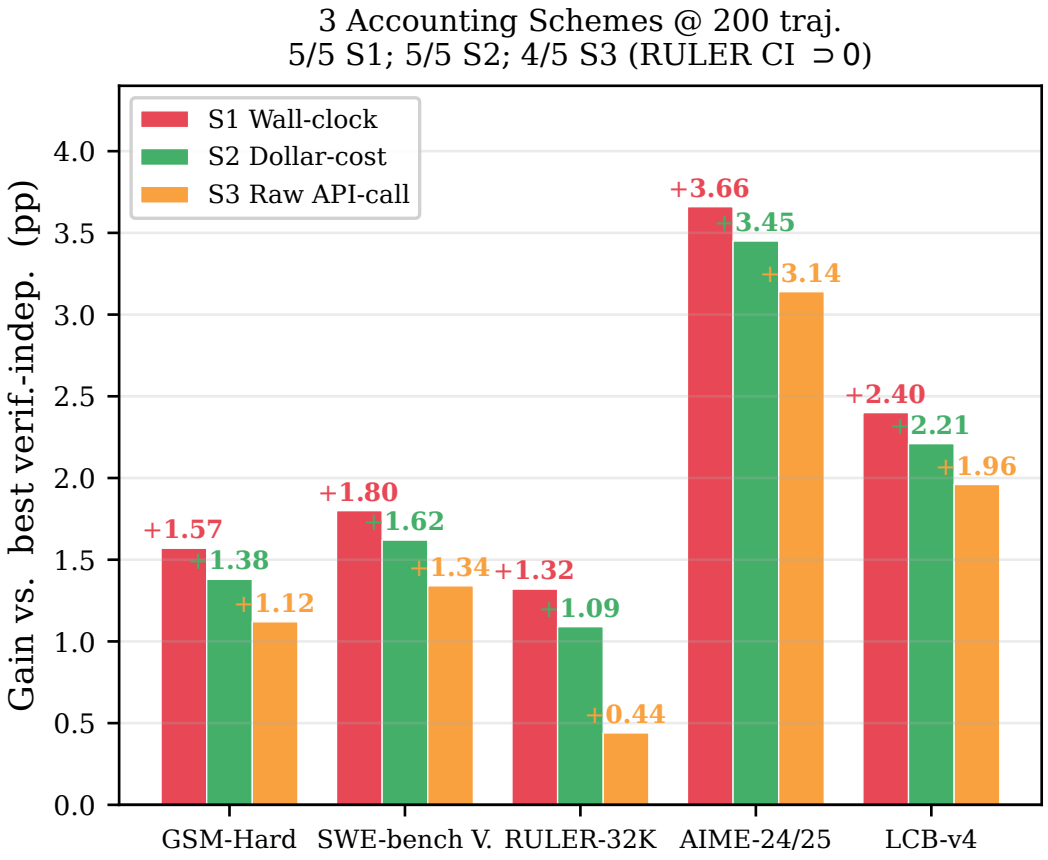


Figure 38: Panel (c) of Figure 36. Proxy comparison at the 200-trajectory calibration point under wall-clock, dollar-cost, and raw-API-call accounting. The three bars for each task apply different resource denominators to the same-access comparison. Task-specific differences and confidence intervals are needed alongside positive-task counts to assess sensitivity to the accounting convention.