COMPOSITIONAL COORDINATION CONFIDENCE: ANYTIME-VALID MONITORING FOR MULTI-AGENT LLM TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent LLM systems can carry an erroneous intermediate step many turns before a final output is inspected. We study online monitoring under an operational requirement: bound, uniformly over time, the probability of halting a trajectory before any coordination fault occurs. Compositional Coordination Confidence (CCC) calibrates per-turn evidence rather than a decision threshold: an LLM judge's rubric score becomes a role-conditional factor with null mean at most one, so products and prespecified mixtures of restarted or tempered products are test supermartingales, and one factor calibration, fit on fault-free or pre-fault turns, yields anytime-valid false-alarm control for each prespecified member of the family without refitting. Calibrating at the factor makes the guarantee's premise, a history-conditional moment restriction, a per-turn quantity that can be tested directly. On AFTraj (math, coding, agentic) and on Who&When, the primary restarted mixture has observed false-alarm rates at or below 3.5% on all five evaluation subsets at nominal 5%, with restricted post-onset detection of 25.9% on Who&When HC and 6.9% and 2.9% on AFTraj math and agentic. The premise test does not reject on coding or agentic; on math it finds a conditional factor of 1.35 under one judge and 1.24 under a second, with cluster-bootstrap intervals excluding 1 in both cases, an excess the forward product exposes on 15.5% of outcome-safe-labelled traces while the primary statistic stays below nominal. An exploratory state-conditioned recalibration reduces that excess while held-out math detection falls from $4/58$ to $0/58$. CCC therefore provides a reusable anytime-valid construction together with an audit that shows, per deployment, when the theorem's certificate does not apply.

1 INTRODUCTION

A multi-agent LLM pipeline can run for many turns before anyone sees its output; when one agent returns a confident wrong result mid-run every downstream step is built on a false premise. Cemri et al. (2025) document this failure mode across over 1,600 multi-agent traces, and Zhang et al. (2025b) report that post-hoc attribution reaches only 53.5% on the responsible agent and 14.2% on the responsible step. Neither figure stops the run.

The operational question is when to halt, and the decision is asymmetric: an operator needs a bound, fixed in advance, on the probability of ever stopping a trajectory in which no coordination fault has occurred. That event — whether any past turn introduced an uncorrected coordination fault — is what we ultimately protect (informally, the "semantic null"); it cannot be observed at runtime, so we test against a moment restriction on the per-turn judge score that we assume it implies (informally, the "statistical null"). Formal definitions are in §3.2 and the bridge assumption in §4.1; §6 reports the approximations that enter how the guarantee is measured on our corpora.

CCC calibrates per-turn factors, exposing a shared conditional-moment premise that supports reusable monitoring and direct checks for violations (§2). At each turn the judge's rubric score becomes a nonnegative role-conditional factor whose expectation under the null is at most one. Multiplying these factors along the trajectory gives a nonnegative process whose probability of ever crossing $1/\alpha$ is bounded by α (Ville's inequality). Every statistic assembled from those factors

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.

under a fixed construction — running product, mixtures of restarted products, tempered versions — inherits the same guarantee without refitting. Section 3 gives the construction and Section 4 its proof.

Three properties of real traces shape the construction. Agent roles have systematically different null distributions, so the null is conditioned on the acting role. We evaluate two core statistics, together with their fixed mixture: a forward product W_{fwd} that accumulates evidence from the first turn and is most sensitive to errors that appear early, and an anchored mixture W_{loc} that restarts the product at fixed positions and remains sensitive to errors that appear later (§3.4). Rubric-based LLM judges emit only the rubric's anchors — clustering documented as a loss of agreement with human raters (Wang et al., 2026b) — so we fix a monotone weighting of those anchors in advance, bounded away from ruin by Proposition 2.

Standard supermartingale closure under mixture, restart, and convex combination (Ramdas et al., 2023; Vovk & Wang, 2021) does the heavy lifting in the construction.

On math, role-conditional calibration does not deliver the history-conditional property: the mean per-turn factor after a turn below the top rubric levels is 1.35 $[1.17, 1.51]$ under one judge and 1.24 $[1.07, 1.42]$ under a second, and W_{fwd} alarms on $9/58 = 15.5\%$ of held-out outcome-safe traces, surviving multiplicity control. A state-conditioned refit, exploratory because its state map was informed by the same diagnostic, brings that mean to 0.99 $[0.87, 1.11]$, while held-out W_{loc} detection falls from $4/58$ to $0/58$. On math development data, out of fold, the refit's Type-I is 2.5–$3.8\times$ lower than that of a reference inflated by the same average amount, so its effect there is state-dependent rather than uniform conservatism; on held-out math a false-alarm-matched inflation reproduces the detection loss under the earlier anchor budget, and the agentic replication is weaker (Section 5, Appendix R).

Contributions. (i) An empirical diagnosis of when role-based calibration fails: a direct conditional-factor test, replicated under two judges, with matched-null and matched-inflation controls that separate state-dependent miscalibration from uniform conservatism; on math the test flags a premise failure that the primary statistic's nominal-looking false-alarm rate ($2/58$) does not reveal. (ii) A reusable monitoring construction: per-turn factors whose single calibration covers every pre-specified statistic in the family without refitting, via standard supermartingale closure, estimated with a betting confidence sequence under a stated constant-conditional-mean assumption rather than an iid bound. (iii) An evaluation of detection costs across thresholds and attempted corrections: under its own calibration assumptions a PAC-fitted threshold raises coding detection from 0% to 27.0% and agentic from 2.9% to 16.2% but leaves math at 6.9%; the math refit reduces the observed conditional-factor excess while W_{loc} detection falls from $4/58$ to $0/58$.

2 Related work

2.1 Diagnosing failure in multi-agent systems

Our corpora — the MAST failure taxonomy (Cemri et al., 2025; 2026) and the Who&When attribution benchmark (Zhang et al., 2025b) — document coordination failures in multi-agent traces. Post-hoc attribution and pre-deployment stress tests (Wang et al., 2026a; Zhang et al., 2025a; Zhu et al., 2026; Liu et al., 2026a; Qiao et al., 2026; Bonagiri et al., 2026; Banerjee et al., 2025; Mazumder et al., 2026) act outside runtime and cannot halt a working run. Online detectors, prefix monitors, trajectory-anomaly models, and runtime-defence and graph-anomaly frameworks (Zhang et al., 2026; Huang et al., 2026; Advani, 2026; Liu et al., 2026b; Mao et al., 2025; Wang et al., 2025; He et al., 2025; Zhou et al., 2025; Miao et al., 2025) act during a run without a formal bound on halting a working one.

2.2 Statistical wrappers on verifier scores

Closest to our setting: Sadhuka et al. (2025) frame black-box verifier monitoring as sequential hypothesis testing on an estimated density ratio, calibrating a PAC threshold because a plain Ville threshold on an estimated ratio need not control size; related sequential monitors on single-model generation (Wasserman et al., 2020; Chen & Wang, 2025; Ota et al., 2026; Dhillon et al., 2026) route through universal inference (which requires an alternative-side model we do not need). Li

(2026) run parallel anytime-valid e-processes for LLM evaluation-pipeline drift — one on live, one on a frozen re-scored anchor set — with a guard-window verdict; the setting differs in what is anchored (across items vs. within a single trajectory) and in the null. General constructions of e-processes from sequential e-values are given by Vovk & Wang (2024), whose martingale merging supplies the closure our factor calibration builds on; what differs here is the object we calibrate — a role-conditional categorical factor on an LLM judge's finite support, rather than a merge of pre-supplied e-values — and its coupling to a plug-in reference under an upper confidence bound. Gauthier et al. (2026) use betting e-processes to detect strategic deviations from equilibrium in repeated and stochastic games; our null is instead a per-turn role-conditional moment on LLM-judge evidence. The calibration-good-event device that yields our $\alpha + \delta$ statement is also used by Dubey & Huo (2026); what differs here is where calibration attaches — per-turn factor rather than rejection threshold (Remark 1). Schirmer et al. (2026) show a single calibrated threshold on the raw verifier score is competitive on single-model generation; we implement their monitors on our scores via CRC (Angelopoulos et al., 2024; Hultberg et al., 2026; Khosravi & Huo, 2026), noting that Ville-style first-crossing degenerates on a five-level rubric while count variants, though themselves calibratable and false-alarm-controlled, would need recalibration whenever the rule changes. Missed-detection control (theirs) remains open here.

2.3 Scoring with LLM judges and sequential testing

Our per-turn signal is a rubric-prompted LLM judge (Zheng et al., 2023); graded-rubric judges return the anchors motivating the finite-support categorical construction (Wang et al., 2026b; Kwok et al., 2026; Shi et al., 2025), whose closure we calibrate at the factor level from fault-free calibration data (§4.4), inherited by every pre-specified statistic under $\mathrm{N2}$. We draw on the sequential-testing tradition (Wald, 1947; Ville, 1939; Howard et al., 2021; Vovk & Wang, 2021; Ramdas et al., 2023; Grünwald, 2024; Koning & van Meer, 2025; Waudby-Smith & Ramdas, 2023; Larsson et al., 2025; Grünwald et al., 2024; Hore et al., 2026; Vovk et al., 2003) and on mixing over unknown change locations (Shin et al., 2023; Page, 1954; Lorden, 1971; Farran, 2026).

3 Methods

This section defines the objects the monitor sees — a trace and its per-turn judge scores — and the two nulls the theory attaches to. The construction is small: at each turn the LLM judge emits a rubric score; a role-conditional map turns that score into a scalar factor whose expected value under the null is at most one; any fixed-weight mixture of running products of those factors is a nonnegative supermartingale, so by Ville's inequality its crossing probability at threshold $1/\alpha$ is at most α under the null. Each of the three statistics we evaluate (W_{fwd}, W_{loc}, M_{dual}) is one such fixed mixture and is monitored independently at that threshold; we do not compose them into a joint OR-alarm, which would not inherit a single α for free. §3.1 sets up the filtration; §3.2 gives the two nulls; §3.3 defines the factor map; §3.4 defines the monitor. Section 4 proves that every construction in §3 controls the false-alarm rate.

3.1 Traces, roles, scores, and the filtration

A *trace* is a sequence of turns $\xi = (\xi_1, \ldots, \xi_T)$; turn t is produced by an agent in *role* $R_t \in \mathcal{R}$ (finite). A *judge* assigns turn t a score S_t in a finite ordered set $\mathcal{L} = \{\ell_1 < \cdots < \ell_L\}$ (higher = more conformant); S_t is $\sigma(\xi_1, \ldots, \xi_t, R_1, \ldots, R_t)$-measurable (no lookahead).

The filtration includes the realised roles and scores,

$$\mathcal{G}_t \;=\; \sigma(\xi_1, \ldots, \xi_t,\; R_1, \ldots, R_t,\; S_1, \ldots, S_t), \qquad \mathcal{G}_t^{+} \;=\; \sigma(\mathcal{G}_t,\; R_{t+1}).$$

3.2 Two nulls

Definition 1 (Semantic null, N1)**.** $\mathrm{N1}_t$ *is the event that none of* $\xi_1, \ldots, \xi_t$ *was incorrect at the point of production and passed downstream rather than corrected within the same turn.*

Informally, N1 is the operational event we protect: no uncorrected coordination fault has occurred so far.

Definition 2 (Statistical null, N2)**.** N2 *is the moment constraint that for every t,*

$$\mathbb{E}\big[\, e_{R_t}(S_t) \,\big|\, \mathcal{G}_{t-1}^{+} \,\big] \;\le\; 1 \qquad \textit{almost surely,}$$

where the factor maps $e_r : \mathcal{L} \to [0, \infty)$ are the role-conditional factors defined in Equation 3 (§3.3).

Informally, $e_r(\ell)$ is the per-turn evidence factor attached to a turn with role r and rubric score ℓ; larger factors mean stronger evidence against N1 at that turn. N2 is the moment restriction the monitor actually tests against; the bridge assumption in §4.1 states when a no-fault trajectory (N1) satisfies it.

Let π_r be a probability vector on $\mathcal{L}$ for each role r, fixed before deployment, and let

$$\mathcal{P} \;=\; \big\{\, P :\; P\big(S_t = \ell_i \,\big|\, \mathcal{G}_{t-1}^{+}\big) = \pi_{r,i} \text{ on } \{R_t = r\},\ \text{for all } t, r, i \,\big\} \tag{1}$$

be the class of laws under which the conditional score distribution depends on the past only through the acting role. Informally, $\mathcal{P}$ says the score at any turn depends on the past only through the acting role; this is the class the per-role calibration reference is fit against.

3.3 Per-turn factors on a finite support

A *factor map* for role r assigns a nonnegative number to each level, $e_r(\ell_i) = \varepsilon_{r,i}$, and is *calibrated for* π_r when

$$\sum_{i=1}^{L} \pi_{r,i}\, \varepsilon_{r,i} \;=\; 1. \tag{2}$$

Informally, calibration says the null-average factor per turn equals 1 — a null turn neither builds nor erodes evidence in expectation. We take a weight vector $w = (w_1, \dots, w_L)$ with $w_1 \ge \cdots \ge w_L > 0$, fixed in advance and shared across roles, and set

$$\varepsilon_{r,i} \;=\; \frac{w_i}{S_r}, \qquad S_r \;=\; \sum_{j=1}^{L} \pi_{r,j}\, w_j. \tag{3}$$

3.4 The monitor

Let $K_{\max}$ be a turn budget and m an anchor spacing, both declared before the run, with anchor positions $\mathcal{A} = \{1, 1+m, \dots\} \cap [1, K_{\max}]$ and nonnegative weights $(c_j)_{j \in \mathcal{A}}$ satisfying $\sum_j c_j \le 1$. Then

$$E_j(t) \;=\; \prod_{s=j}^{t} e_{R_s}(S_s) \;\; (t \ge j), \qquad E_j(t) \;=\; 1 \;\; (t < j), \qquad M_t \;=\; \sum_{j \in \mathcal{A}} c_j\, E_j(t), \tag{4}$$

and the monitor halts at $\sigma = \inf\{t : M_t \ge 1/\alpha\}$.

We evaluate two anchored-mixture statistics from the same family $M_t = \sum_{j \in \mathcal{A}} c_j E_j(t)$ under different fixed c-priors. The forward product W_{fwd} concentrates all mass at $j = 1$ ($c_1 = 1$, $c_j = 0$ otherwise), so it multiplies every factor from the first turn. The anchored restart mixture W_{loc} uses uniform weights $c_j = 1/K'$ across the K' anchor positions fixed in advance. We also report $M_{\mathrm{dual}} = \frac{1}{2}(W_{\mathrm{fwd}} + W_{\mathrm{loc}})$, which since $1 \in \mathcal{A}$ is again in the same family under a top-heavy prior: $c_1 = (K' + 1)/(2K')$ and $c_j = 1/(2K')$ for $j \ne 1$. Not a distinct construction — a second c-prior on the same anchored mixture, kept for continuity with our pre-registration (Section 6).

4 Theory

All proofs are in Appendix A, whose opening paragraph gives a short roadmap.

4.1 Assumptions

Assumption 1 (Predictable anchors and weights)**.** *$\mathcal{A}$ and (c_j) are $\mathcal{G}_0$-measurable: fixed before the run and independent of the realised trace. For the anchored restart mixture W_{loc} this includes the*

anchor count K'; setting K' from the realised trace length would violate this assumption. Our census rule (§5.1) fixes K' per stratum from the 95th-percentile chain length on the dev-safe pool.

Assumption 2 (N1 implies N2)**.** *The sequence $(\mathrm{N1}_t)_{t\geq 0}$ is $(\mathcal{G}_t)$-adapted (with $\mathrm{N1}_0 = \Omega$), and for every $t \geq 1$, on the event $\mathrm{N1}_{t-1}$,*

$$\mathbb{E}\big[\, e_{R_t}(S_t) \;\big|\; \mathcal{G}^+_{t-1} \,\big] \;\leq\; 1 \qquad \textit{a.s.}$$

This is a modelling premise, not a theorem (Remark 7 argues some link of this form is unavoidable).

4.2 Validity

Lemma 1 (Anchored processes)**.** *Under* N2 *and Assumption 1, each $(E_j(t))_t$ is a nonnegative $(\mathcal{G}_t)$-supermartingale with $E_j(0) = 1$.*

Lemma 2 (Fixed-weight mixtures)**.** *For (c_j) $\mathcal{G}_0$-measurable with $c_j \geq 0$ and $\sum_j c_j \leq 1$, $M_t = \sum_j c_j E_j(t)$ is a nonnegative supermartingale with $\mathbb{E}[M_0] \leq 1$.*

Ville's inequality then gives:

Theorem 1 (Time-uniform false-alarm control)**.** *Under* N2 *and Assumption 1, for every $\alpha \in (0, 1)$,*

$$\Pr\big[\, \exists t \geq 1 : M_t \geq 1/\alpha \,\big] \;\leq\; \alpha.$$

In plain terms: if the per-turn moment restriction holds throughout the trajectory, the monitor's chance of ever crossing the threshold $1/\alpha$ is at most α, uniformly over the trajectory's length.

Reading the bound. A single calibrated process supports testing at any prospectively chosen level α, through the threshold $1/\alpha$, without refitting the factor map (Grünwald, 2024). A level chosen as a function of the observed process is not covered by this guarantee.

Corollary 1 (Alarm before the first fault)**.** *Let $\nu = \inf\{t \geq 1 : \text{a coordination fault occurs at turn } t\}$, with $\nu = \infty$ on trajectories where no fault ever occurs; and let $\tau = \inf\{t \geq 1 : M_t \geq 1/\alpha\}$. Under Assumptions 1 and 2, for every $\alpha \in (0, 1)$,*

$$\Pr[\tau < \nu] \;\leq\; \alpha.$$

That is: the alarm probability is at most α before an eventual fault (turn ν), and at most α ever on a fault-free trajectory ($\nu = \infty$). On finite-length trajectories we measure $\{\tau \leq \min(H, \nu) - 1\}$. The bound is unconditional; empirical rates on selected subpopulations inherit it only under the same assumptions. Post-fault alarms ($\nu \leq \tau \leq H - 1$) do not count here; they count as detection under §5.4's restricted convention, following the finite-window change-detection formulation of Tartakovsky et al. (2014).

4.3 Inherited validity

Proposition 1 (Closure)**.** *Call a process* admissible *if it is a nonnegative supermartingale with initial value at most one, built from factor maps satisfying* N2*. Admissible processes are closed under each of the following, with no further calibration:*

- *(i) restart, $E_j(t) = \prod_{s=j}^{t} e_{R_s}(S_s)$ for $\mathcal{G}_0$-measurable j;*
- *(ii) fixed-weight mixture, $\sum_j c_j E_j$ with $\mathcal{G}_0$-measurable $c_j \geq 0$, $\sum_j c_j \leq 1$, and more generally $\int E_\theta \, d\pi(\theta)$ for a prior fixed in advance;*
- *(iii) tempering, $\prod_s e_{R_s}(S_s)^\lambda$ for $\lambda \in (0, 1]$;*
- *(iv) betting, $\prod_s \big[(1 - \lambda_s) + \lambda_s \, e_{R_s}(S_s)\big]$ for predictable $\lambda_s \in [0, 1]$.*

Remark 1 (The contrast). Closure concerns validity, not power. The contrast is with wrappers that calibrate a rejection threshold, which certifies one statistic and must be refitted for another (post-hoc selection is not covered; a fixed mixture over candidates is the admissible alternative).

A companion proposition (Appendix K, Proposition 3) bounds the false-alarm probability under short adversarial bursts of length $|B|$ by $\alpha\kappa^{|B|}$ with $\kappa = w_{\max}/w_{\min}$; at $\kappa = 5$ the bound is vacuous from $|B| = 2$ and binds only on single-turn bursts.

4.4 ESTIMATED NULLS

S_r is estimated on a fault-free calibration sample $\mathcal{D}$ as $\widehat{S}_r = \min(w_{\max}, \mathrm{UCB}_r)$, with UCB_r the upper endpoint of the betting confidence set of Waudby-Smith & Ramdas (2023) at coverage $1 - \delta_r$ (construction in App. B.11; the fixed-n iid empirical-Bernstein bound of Maurer & Pontil (2009) is reported alongside as comparator). Coverage requires the calibration law Assumption 3 (constant role-r conditional mean on the calibration subsequence, stated in App. B.11): a modelling premise on the calibration draw, stronger than the trajectory-level Assumption 2 (a one-sided inequality) and stated, not derived. With δ_r split evenly across roles and $A = \{\widehat{S}_r \geq S_r \ \forall r\}$, $\Pr[A^c] \leq \delta := \sum_r \delta_r$. We fix $\delta = 0.005$ (one-time estimation budget), giving unconditional level $\alpha(1-\delta) + \delta = 0.05475$ at $\alpha = 0.05$.

Theorem 2 (Control under an estimated null)**.** *Suppose the monitored trajectory is independent of $\mathcal{D}$, its law lies in $\mathcal{P}$, Assumption 1 holds, and each $\widehat{S}_r$ is constructed as a one-sided $(1-\delta_r)$-CS upper bound for S_r under Assumption 3 (App. B.11) with $\sum_r \delta_r \leq \delta$. Then*

(a) with probability at least $1-\delta$ over the draw of $\mathcal{D}$, $\Pr[\exists t : M_t \geq 1/\alpha \mid \mathcal{D}] \leq \alpha$;

(b) unconditionally, $\Pr[\exists t : M_t \geq 1/\alpha] \leq \alpha(1-\delta) + \delta$.

Proof in Appendix A.6.

4.5 ESTIMATING THE REFERENCE: SHIPPING REGIME

All three AFTraj domains ship the bounded reference (WSR UCB at $\delta = 0.005$); Who&When ships a shrunk plug-in on pre-mistake windows (its corpus contains no fault-free trajectories). Per-role losses, the coding illustration, the pooling caveat, and the Who&When shrunk-plugin construction are in Appendix L.

5 EXPERIMENTS

We ask three questions in order. Do observed false-alarm rates stay at nominal across statistics we never recalibrated for? What does the monitor detect? And what does it cost, in false alarms spent and in compute?

5.1 SETUP

Corpora. Two benchmarks with different failure semantics. AFTraj contains 2,276 trajectories across three task domains — math (793), coding (608), agentic (875) — with both outcome-safe and failing traces (AFTraj labels are outcome-based; see limitations). Failing trajectories arise two ways: 728 by controlled injection of a decisive error into a verified outcome-safe trace, and 386 diagnosed from naturally failed runs by a propose-verify pipeline. Mix by domain: coding entirely injected, math 84% injected, agentic 69% naturally failed. Who&When contains 184 real annotated failures across a hand-crafted subset (HC, 58 traces, drawn from Magentic-One (Fourney et al., 2024)) and an algorithmically generated one (AG, 126 traces, drawn from AutoGen (Wu et al., 2023)).

Who&When has no fault-free trajectories, so we measure false alarms on pre-mistake windows of eligible annotated failures. The resulting rate is conditional on that selection: Corollary 1 applies to it only if its assumptions hold under the selected population's law, and the annotation is a proxy for the true onset. If the annotation is correct, a prefix preceding the fault satisfies N1 under the process-level null of Section 3 (it would not under an outcome-level null).

The reference is estimated from the same windows, leave-one-out across trajectories, so no trajectory contributes to the distribution against which it is scored. Admitting this pool as a null reference requires pre-mistake prefixes to satisfy N2, i.e. Assumption 2 for this population; neither the annotation nor leave-one-out estimation establishes it, and the pool consists of prefixes of trajectories that later failed rather than of trajectories that did not. The per-role empirical is shrunk toward the subset-pooled distribution as $(n_r\, \pi_r^{\mathrm{emp}} + \kappa\, \pi_{\mathrm{pool}})/(n_r + \kappa)$ with $\kappa = 5$. An unseen role is assigned a point mass on the top-weight rubric cell, which bounds its per-turn factor by 1 and prevents an unrecognised role from alarming on its own (Remark 4).

Median chain lengths differ substantially: 12, 11 and 7 turns across the AFTraj domains against 31 and 10 on HC and AG.

Splits. Three disjoint pools. The per-role reference is estimated on development safe-labeled traces only, and those traces are never scored. False-alarm rates are measured on held-out safe-labeled traces (58, 54, 57 by domain); detection on held-out failing traces (58, 37, 68). Denominators are never mixed.

Parent–child audit. 725/728 injected AFTraj failing trajectories match a safe-labeled parent; zero land with parent in dev and child in held-out (the split is pair-aware). $106/163$ share a pre-mistake prefix with a held-out safe-labeled trajectory under a different `conv_id`; restricting the safe-labeled pool to non-parents leaves Type-I inside every full-set Wilson 95% interval (Appendix P).

Scoring. One LLM judge scores every turn against a five-level rubric with no access to ground truth (full prompt in Appendix S; verified by `sha256` against the shipped p-hat caches). Both benchmarks use the same judge and rubric.

Configuration. The per-turn factor is the rubric-level categorical construction of Section 3 with a fixed monotone weighting $w = (5, 4, 3, 2, 1)$ chosen a priori. The threshold is $1/\alpha = 20$ at $\alpha = 0.05$, supplied by Theorem 1 rather than fitted. Anchor spacing is $m = 3$; the anchor budget $K' = \lceil K_{\max}/m \rceil$ is fixed per stratum from a corpus census before any trace is scored, with $K_{\max}$ set at the ceiling of the 95th percentile of realized chain length on the census pool (dev-safe on AFTraj; the same pool used to fit the reference). This is a function of trajectory lengths only, not of the judge scores, so $(\mathcal{A}, c_j)$ is $\mathcal{G}_0$-measurable as Assumption 1 requires. Values: $K_{\max} = 128$ ($K' = 43$) for HC and 10 ($K' = 4$) for AG on Who&When; $K_{\max} = 36$ ($K' = 12$) on AFTraj math, 11 ($K' = 4$) on coding, and 20 ($K' = 7$) on agentic. No hyperparameter is selected against held-out outcomes.

Reference regime. All three AFTraj domains ship the *bounded reference* (one-sided WSR UCB at $\delta = 0.005$); Who&When ships a shrunk plug-in on pre-mistake windows (its corpus carries no fault-free trajectories). Theorem 2 applies under its conditional-law (N2) and calibration assumptions: on the good event the conditional false-alarm level is α; averaging over calibration draws gives $\alpha(1 - \delta) + \delta$ (0.05475 at $\alpha = 0.05$). Whether those assumptions hold at a deployed cell is a separate empirical question — small estimated loss is a marginal-mean statement, not a conditional-law statement (Appendix Q). A complete list of departures from pre-registration appears in the reproducibility statement.

Estimand. AFTraj reports alarm rates on held-out outcome-safe-labelled trajectories. Reading them as process-null error rates bounded by Corollary 1 requires both the semantic premise that an outcome-safe trace contains no coordination fault (N1) and the conditional-moment premise (N2); §5.2 and Appendix Q find evidence against the latter on math. Who&When rates are conditional on the 54 HC and 72 AG eligible failures (App. N). Reported rates inherit Theorem 2's bound only where its assumptions hold. Post-fault alarms count as detection under the restricted convention (§5.4).

5.2 Empirical false-alarm behaviour across pre-specified processes

The claim of Section 3 is compositional: calibrate the per-turn factor once, and every pre-specified statistic assembled from those factors inherits time-uniform false-alarm control under N2. Appendix C gives the full grid — one reference estimation per domain, three statistics built from it, no per-statistic calibration of any kind.

Under the two calibration regimes of §5.1, the nine shipped cells fall at or below the nominal level with one formal exception and one point-estimate observation: the forward product on math is elevated at $9/58 = 15.5\%$ (one-sided binomial $p = 0.0022$ uncorrected, BH $q = 0.020$ across the nine cells — survives multiplicity control), and the dual mixture on math sits at $5/58 = 8.6\%$ at point estimate but is not formally elevated ($p = 0.164$; Clopper–Pearson 95% CI $[0.029, 0.190]$ covers 5%). The restart mixture on math and every coding and agentic cell hold at or below nominal. Appendix Q diagnoses the W_{fwd} elevation. Section 5.4 reports detection under the restart mixture, chosen as primary a priori for the burst-structure reason of Section 3.4.

Direct N2 diagnostic on the math elevation. On held-out math safe ($n = 196$), $\mathbb{E}[\varepsilon_{R_t}(S_t) \mid \text{prev turn not L4-L5}] = 1.35$ (cluster BCa 95% CI $[1.17, 1.51]$; math is the only domain whose CI excludes 1). A tuned matched-null shows the mechanism can generate 15.5% but does not explain the observed value quantitatively. Full audit in Appendix Q.

5.3 BASELINE COMPARISON, AND THE ASSUMPTION-FAILURE DIAGNOSIS ACROSS JUDGES

CCC-PAC matches a labelled per-role likelihood-ratio baseline (SLR) on coding detection (27.0% each) at strictly lower Type-I (0.0% vs 5.6%); on agentic the two gaps have the same sign; on math SLR is a trajectory ahead on each axis at $n = 58$, with fully overlapping Clopper–Pearson intervals. Failure labels do not deliver a decisive detection edge at any of the three achieved rates. Full head-to-head numbers (Table 4) and the threshold-portability illustration are in Appendix H.

The role-only calibration exhibits a diagnosed assumption failure on math held-out safe (App Q); a state-conditioned refit — exploratory on this data because its state map was informed by that same diagnostic — reduces the observed excess and yields an interval containing 1 (App R). A second judge (Anthropic Sonnet 5) rescored the same trajectories, and a matched-inflation control on the state-conditioned refit separates state structure from raw conservatism. Table 1 consolidates the four substantive findings; details in Apps Q, R, and T.

Table 1: Consolidated findings. Rows (3a) and (3b) are two matched-inflation controls: (3a) matches arm 2b (dev-FA) to arm 3 held-out; (3b) matches arm 2a (average log-inflation) to arm 3 dev out-of-fold. Intervals: cluster BCa 95% on rows (1) and (2); row (3b) uses percentile bootstrap-with-refit. Procedural detail in cited appendices.

finding	numeric evidence	source
(1) Conditional-factor excess under both judges	$\mathbb{E}[\varepsilon \mid \text{prev not top}]$ on math held-out safe: GPT-4o-mini 1.35 $[1.17, 1.51]$; Sonnet 5 1.24 $[1.07, 1.42]$. Both exclude 1.	App Q; App T
(2) State-conditioned refit	GPT arm 3 0.99 $[0.87, 1.11]$; Sonnet arm 3 0.92 $[0.81, 1.04]$. Held-out det. W_{loc} $4/58 \to 0/58$; M_{dual} $10/58 \to 0/58$.	App R; App T
(3a) Math held-out: FA-matched inflation reproduces the collapse	Arm 2b vs arm 3 on math held-out ($n = 58$) under per-trace K': McNemar $b = c = 0$ on both W_{loc} and M_{dual}.	App R (Stage 4 v2)
(3b) Math dev out-of-fold	Math dev OOF Type-I ratio (arm 2a vs arm 3) on 338 dev trajectories: $2.5\times$ / $3.75\times$ / $3.0\times$ on $W_{\text{fwd}}/W_{\text{loc}}/M_{\text{dual}}$; bootstrap-with-refit ($B = 1000$) 95% CI on paired Type-I difference excludes zero on every statistic. Agentic frozen-map replication is reported separately in App R.	App R (Stage 6 OOF)
(4) Judge swap moves the Who&When HC operating point	W_{loc} restricted detection $15/58 \to 21/58$ $(+6)$; pre-mistake Type-I $1/54 \to 4/54$ $(+3)$.	App T

Row 3b's discordant pairs are all in arm 3's favour: arm 3 never adds an alarm arm 2a keeps.

5.4 DETECTION

Table 2 reports detection for W_{loc}, the anchored restart mixture, which we take as the primary statistic on the a-priori grounds of Section 3.4: coordination failures arrive as short bursts at unknown onset, and a single forward product from the first turn is poorly matched to that structure. The level sweep of Appendix D is consistent with that choice but did not motivate it.

Safe-only points sit at or below $\alpha = 5\%$ but on 54–72 denominators the CP upper bounds reach 6.6–11.9% (*consistent with nominal* rather than validated). Pooled rates deflate because pre-mistake prefixes are 2–3$\times$ shorter than safe trajectories; the exposure-weighted rates in the caption are near-invariant under pooling. The price of fixing the threshold at the formal Ville level $h = 1/\alpha = 20$

Table 2: W_{loc} across the five reported subsets, Clopper–Pearson 95% intervals. **References:** bounded WSR on math, coding, and agentic ($\delta = 0.005$), shrunk plug-in on Who&When. Restricted convention: alarms before annotated onset u are Type-I on the pre-onset prefix, not detections; detection counts alarms at-or-after u. **FA denominators:** AFTraj rows on complete safe trajectories; Who&When rows (†) on pre-onset windows of failing trajectories, N2-safe only under Assumption 2. Exposure-weighted primary/pooled: math 0.20%/0.23%, coding 0.00%/0.00%, agentic 0.19%/0.15%. Who&When intervals approximate (overlapping LOO). Cells are empirical operating characteristics, not per-cell attachments of Theorem 2 (App. Q).

	FA primary ↓ (95% CI)	FA pooled ↓ (95% CI)	detection ↑ (95% CI)
Who&When HC†	1/54 = 1.9% [0.0, 9.9]	—	15/58 = 25.9% [15.3, 39.0]
Who&When AG†	1/72 = 1.4% [0.0, 7.5]	—	0/126 = 0.0% [0.0, 2.9]
AFTraj math	2/58 = 3.4% [0.4, 11.9]	3/116 = 2.6% [0.5, 7.4]	4/58 = 6.9% [1.9, 16.7]
AFTraj coding	0/54 = 0.0% [0.0, 6.6]	0/91 = 0.0% [0.0, 4.0]	0/37 = 0.0% [0.0, 9.5]
AFTraj agentic	1/57 = 1.8% [0.0, 9.4]	1/125 = 0.8% [0.0, 4.4]	2/68 = 2.9% [0.4, 10.2]

varies by subset. HC W_{loc} reaches $15/58 = 25.9\%$ under both Ville and PAC ($h \approx 14.4$), at Type-I $1/54 = 1.9\%$ and $2/54 = 3.7\%$; math gives the same detection under both ($h_{\text{PAC}} \approx 19$, detection $4/58 = 6.9\%$ at Type-I $2/58 = 3.4\%$). Agentic ($h_{\text{PAC}} \approx 2.4$) moves from $2/68 = 2.9\%$ to $11/68 = 16.2\%$ at Type-I $1/57 = 1.8\%$ and $2/57 = 3.5\%$; coding ($h_{\text{PAC}} \approx 2.9$) moves from $0/37 = 0.0\%$ to $10/37 = 27.0\%$ at unchanged Type-I $0/54 = 0.0\%$. Coding trajectories therefore carry enough evidence to alarm at a PAC-fitted cutoff and nowhere near enough to reach $1/\alpha$; the earlier reading that coding *detects nothing* was a statement about the Ville threshold that omitted its own scope. AG sits at the floor either way ($K - u \leq 7$). We do not pool across benchmarks.

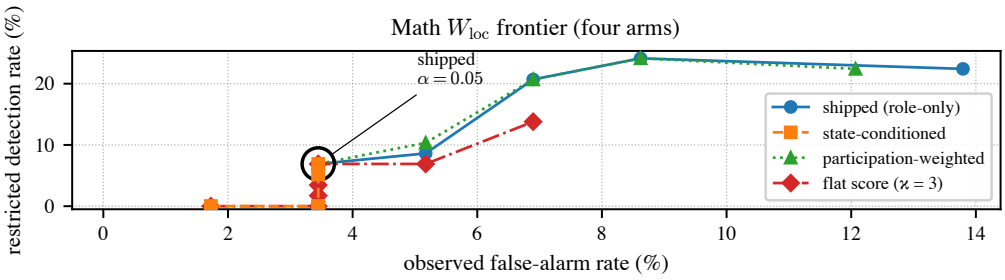


Figure 1: At $\alpha = 0.05$ no alternative exceeds shipped detection at comparable Type-I; not uniform across the grid — participation reaches $6/58$ vs shipped $5/58$ at equal $3/58$ Type-I when $\alpha = 0.10$. Math W_{loc}, seven α; four arms: shipped, state-conditioned (App. R), participation-weighted (App. M), flat-score ($\kappa = 3$; App. B.3). HC: Fig. 2.

Delay, opportunity, oracle, cost, and level sweep in Apps N and I.

6 Discussion

Low detection has different sources in these results. On coding and agentic, the same scores and statistic carry detection information that the Ville threshold does not use at the reported operating point: a PAC-fitted threshold, under its own calibration assumptions, reaches $10/37$ at Type-I $0/54$ and $11/68$ at $2/57$ (Section 5.4). On math the threshold change gives no benefit ($4/58$ under both); the math refit reduces the observed conditional-factor excess while W_{loc} detection falls from $4/58$ to $0/58$ (Appendix R). The second judge moves HC detection ($15/58 \to 21/58$) and pre-mistake Type-I ($1/54 \to 4/54$) together (Appendix T). These are properties of the tested thresholds, interventions, and judge-score representation, not a statement that useful evidence and false alarms are inseparable in general. On math the history-conditional property does not hold at the operating point, so Theorem 2 does not certify that deployment; math rates are empirical operating characteristics.

Limitations: sub-seventy held-out failing trajectories per AFTraj cell; Who&When rates and reference both from pre-mistake windows (§5.1); AFTraj safe labels are outcome-based rather than direct annotations of the process-level null, and a single-annotator blind audit found possible recoverable process faults in some safe-labeled trajectories (App. Q); fleet false-alarm bounded only by $N_{\text{deploy}}\alpha$; the one-sided bound is trivial for a never-firing monitor; numbers are judge- and length-specific.

REPRODUCIBILITY STATEMENT

This paper reports experiments as run under the pre-registration `manuscript/CCC_experiments_v3.tex` (24 August 2026), with deviations itemized in Appendix V. Two post-pre-registration corrections were applied: the AFTraj anchor budget K' moved from per-trace to a per-domain census rule to restore $\mathcal{G}_0$-measurability of the mixture weights, and agentic's reference moved from the raw plug-in to the bounded WSR reference (deviations 21 and 1). These corrections regenerated all primary within-domain results. Secondary analyses that retain an earlier configuration, whether an earlier reference regime or the earlier per-trace anchor budget, are identified at the point of use in Appendices J (off-diagonal OOD cells), R (arm-2b matched-inflation and the out-of-fold Type-I comparison on agentic), E (the fixed-vs-randomized Ville comparator), M (the E2 evaluation), and L (the math pooling caveat), and are not part of the shipped operating-point evaluation. The shipped judge is `gpt-4o-mini` with a second judge run at `claude-sonnet-5` (App. T); both are provider aliases rather than dated snapshots, so exact rescoring from our prompts is not guaranteed reproducible across future dates but recomputation from the released rubric values is (prompt `sha256` and truncation caps in Appendix V). The release archive contains per-numeric-claim tables and a manifest naming each table with the sections it supports and the script that produces it.

AI USE STATEMENT

In this work, we used generative AI tools (LLM chat models and coding assistants) to search for and summarize related work, draft and revise exposition, edit LaTeX, implement and run analysis code under the authors' direction, check and revise proofs, audit numerical consistency, and provide review feedback on the manuscript and its experiments. The authors designed the methods and experiments, verified every theorem statement and proof, and re-derived every reported number from committed artifacts and source data rather than accepting it as generated. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Laksh Advani. Trajectory guard: A lightweight, sequence-aware model for real-time anomaly detection in agentic AI, 2026. URL https://arxiv.org/abs/2601.00516.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2208.02814.

Adi Banerjee, Anirudh Nair, and Tarik Borogovac. Where did it all go wrong? a hierarchical look into multi-agent error attribution, 2025. URL https://arxiv.org/abs/2510.04886.

Akash Bonagiri, Devang Borkar, Gerard Janno Anderias, Setareh Rafatirad, and Houman Homayoun. CausalFlow: Causal attribution and counterfactual repair for LLM agent failures, 2026. URL https://arxiv.org/abs/2605.25338.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei A. Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2503.13657. Datasets and Benchmarks Track.

Mert Cemri, Andrei Cojocaru, Melissa Pan, Shu Liu, Shubham Agarwal, Alexander Krentsel, Jay Tang, Kannan Ramchandran, Joseph E. Gonzalez, Matei Zaharia, Alex Dimakis, and Ion Stoica. Fantastic adaptive taxonomies and how to use them, 2026. URL https://arxiv.org/abs/2607.16387. Best Paper Award, ICML 2026 FAGEN Workshop.

Can Chen and Jun-Kun Wang. Online detection of LLM-generated texts via sequential hypothesis testing by betting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9231–9276, 2025. URL https://arxiv.org/abs/2410.22318.

Guneet S. Dhillon, Javier Gonzalez, Teodora Pandeva, and Alicia Curth. E-scores for (in)correctness assessment of generative model outputs. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://arxiv.org/abs/2510.25770. arXiv:2510.25770, version 2, 1 Apr 2026.

Prasanjit Dubey and Xiaoming Huo. Anytime-valid federated conformal RAG for LLM swarms, 2026. URL https://arxiv.org/abs/2605.29139.

Tristan Farran. When your model stops working: Anytime-valid calibration monitoring, 2026. URL https://arxiv.org/abs/2603.13156.

Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Erkang Zhu, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, Peter Chang, Ricky Loynd, Robert West, Victor Dibia, Ahmed Awadallah, Ece Kamar, Rafah Hosn, and Saleema Amershi. Magentic-one: A generalist multi-agent system for solving complex tasks, 2024. URL https://arxiv.org/abs/2411.04468.

Etienne Gauthier, Francis Bach, and Michael I. Jordan. Anytime detection of strategic deviations in multi-agent systems. *arXiv preprint arXiv:2601.05427*, 2026. URL https://arxiv.org/abs/2601.05427.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(5):1091–1128, 2024. doi: 10.1093/jrsssb/qkae011. URL https://arxiv.org/abs/1906.07801.

Peter D. Grünwald. Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, 121(39):e2302098121, 2024. doi: 10.1073/pnas.2302098121. URL https://arxiv.org/abs/2205.00901.

Xu He, Di Wu, Yan Zhai, and Kun Sun. SentinelAgent: Graph-based anomaly detection in multi-agent systems, 2025. URL https://arxiv.org/abs/2505.24201.

Rohan Hore, Ruodu Wang, and Aaditya Ramdas. Online monotone density estimation and log-optimal calibration, 2026. URL https://arxiv.org/abs/2602.08927. Version 3, 21 May 2026.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, 49(2):1055–1080, 2021. URL https://arxiv.org/abs/1810.08240.

Xinmiao Huang, Jinwei Hu, Rajarshi Roy, Changshun Wu, Yi Dong, and Xiaowei Huang. PrefixGuard: From LLM-agent traces to online failure-warning monitors, 2026. URL https://arxiv.org/abs/2605.06455.

Bror Hultberg, Dave Zachariah, and Antônio H. Ribeiro. Anytime-valid conformal risk control. *arXiv preprint arXiv:2602.04364*, 2026. URL https://arxiv.org/abs/2602.04364.

John L. Kelly, Jr. A new interpretation of information rate. *Bell System Technical Journal*, 35(4): 917–926, 1956. doi: 10.1002/j.1538-7305.1956.tb03809.x.

Hamed Khosravi and Xiaoming Huo. Conformal selective acting: Anytime-valid risk control for RLVR-trained LLMs, 2026. URL https://arxiv.org/abs/2605.20270.

Nick W. Koning and Sam van Meer. Anytime validity is free: Inducing sequential tests, 2025. URL https://arxiv.org/abs/2501.03982.

Jacky Kwok, Shulu Li, Pranav Atreya, Yuejiang Liu, Yixing Jiang, Chelsea Finn, Marco Pavone, Ion Stoica, and Azalia Mirhoseini. LLM-as-a-Verifier: A general-purpose verification framework, 2026. URL https://arxiv.org/abs/2607.05391.

Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *Annals of Statistics*, 53(3):1015–1043, 2025. URL https://arxiv.org/abs/2402.18810.

Yitao Li. Who drifted: the system or the judge? Anytime-valid attribution in LLM evaluation pipelines, 2026. URL https://arxiv.org/abs/2606.15474.

Yang Liu, Hongjiang Feng, Junsong Pu, and Zhuangbin Chen. MASPrism: Lightweight failure attribution for multi-agent systems using prefill-stage signals. In *Proceedings of the 41st IEEE/ACM International Conference on Automated Software Engineering*, 2026a. URL https://arxiv.org/abs/2605.07509. arXiv:2605.07509.

Yibing Liu, Chong Zhang, Zhongyi Han, Hansong Liu, Yong Wang, Yang Yu, Xiaoyan Wang, and Yilong Yin. TrajAD: Trajectory anomaly detection for trustworthy LLM agents, 2026b. URL https://arxiv.org/abs/2602.06443.

Gary Lorden. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

Junyuan Mao, Fanci Meng, Yifan Duan, Miao Yu, Xiaojun Jia, Junfeng Fang, Yuxuan Liang, Kun Wang, and Qingsong Wen. AgentSafe: Safeguarding large language model-based multi-agent systems via hierarchical data management, 2025. URL https://arxiv.org/abs/2503.04392.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009. URL https://arxiv.org/abs/0907.3740.

Aritra Mazumder, Shubhashis Roy Dipta, Nusrat Jahan Lia, Tanzila Khan, Kainat Raisa Hossain, Nehaa Shri, Shubhrangshu Debsarkar, Humayra Tasnim, Gour Gupal Talukder Shawon, Debjoty Mitra, Sumaiya Ahmed Rani, Al Jami Islam Anik, and Al Nafeu Khan. AgentCollabBench: Diagnosing when good agents make bad collaborators, 2026. URL https://arxiv.org/abs/2605.08647.

Rui Miao, Yixin Liu, Yili Wang, Xu Shen, Yue Tan, Yiwei Dai, Shirui Pan, and Xin Wang. BlindGuard: Safeguarding LLM-based multi-agent systems under unknown attacks, 2025. URL https://arxiv.org/abs/2508.08127.

Hirofumi Ota, Naoto Iwase, Yuki Ichihara, Junpei Komiyama, and Masaaki Imaizumi. CITE: Anytime-valid statistical inference in LLM self-consistency, 2026. URL https://arxiv.org/abs/2605.05873.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954.

Hezhe Qiao, Hanghang Tong, Ee-Peng Lim, Bing Liu, and Guansong Pang. VerifyMAS: Hypothesis verification for failure attribution in LLM multi-agent systems, 2026. URL https://arxiv.org/abs/2605.17467.

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with E-values. *arXiv preprint arXiv:2410.23614*, 2025. URL https://arxiv.org/abs/2410.23614. Also as monograph, Now Publishers, 2025.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Shuvom Sadhuka, Drew Prinster, Clara Fannjiang, Gabriele Scalia, Bonnie Berger, Aviv Regev, and Hanchen Wang. E-valuator: Reliable agent verifiers with sequential hypothesis testing, 2025. URL https://arxiv.org/abs/2512.03109. Version 2, 28 May 2026.

Mona Schirmer, Metod Jazbec, Alexander Timans, Christian Naesseth, Maja Waldron, and Eric Nalisnick. Online safety monitoring for LLMs, 2026. URL https://arxiv.org/abs/2607.02510.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 14th International Joint Conference on Natural Language Processing*, 2025. URL https://arxiv.org/abs/2406.07791.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. E-detectors: A nonparametric framework for sequential change detection. *The New England Journal of Statistics in Data Science*, 2023. URL https://arxiv.org/abs/2203.03532.

Alexander Tartakovsky, Igor Nikiforov, and Michèle Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Chapman and Hall/CRC, 2014.

Jean Ville. *Étude Critique de la Notion de Collectif*. Gauthier-Villars, Paris, 1939.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020.

Vladimir Vovk and Ruodu Wang. Merging sequential e-values via martingales. *Electronic Journal of Statistics*, 18(1):1770–1791, 2024. doi: 10.1214/24-EJS2244.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. Technical report, Royal Holloway, University of London, 2003.

Abraham Wald. *Sequential Analysis*. Wiley, 1947.

Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. G-Safeguard: A topology-guided security lens and treatment on LLM-based multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://arxiv.org/abs/2502.11127.

Yawen Wang, Wenjie Wu, Junjie Wang, and Qing Wang. From flat logs to causal graphs: Hierarchical failure attribution for LLM-based multi-agent systems, 2026a. URL https://arxiv.org/abs/2602.23701.

Yidong Wang, Yunze Song, Tingyuan Zhu, Xuanwang Zhang, Zhuohao Yu, Hao Chen, Chiyu Song, Qiufeng Wang, Cunxiang Wang, Zhen Wu, Xinyu Dai, Yue Zhang, Wei Ye, and Shikun Zhang. TrustJudge: Inconsistencies of LLM-as-a-judge and how to alleviate them. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2509.21117. arXiv:2509.21117.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020. doi: 10.1073/pnas.1922664117. URL https://arxiv.org/abs/1912.11436.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B*, 85(1):1–27, 2023. URL https://arxiv.org/abs/2010.09686.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation, 2023. URL https://arxiv.org/abs/2308.08155.

Boxuan Zhang, Jianing Zhu, Zeru Shi, Dongfang Liu, and Ruixiang Tang. AgentForesight: Online auditing for early failure prediction in multi-agent systems, 2026. URL https://arxiv.org/abs/2605.08715.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. AgenTracer: Who is inducing failure in the LLM agentic systems?, 2025a. URL https://arxiv.org/abs/2509.03312.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. Which agent causes task failures and when? On automated failure attribution of LLM multi-agent systems. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2025b. URL https://arxiv.org/abs/2505.00212. Spotlight.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05685.

Jialong Zhou, Lichao Wang, and Xiao Yang. GUARDIAN: Safeguarding LLM multi-agent collaborations with temporal graph modeling. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.19234. arXiv:2505.19234.

Taiyu Zhu, Yifan Wu, Weilin Jin, Ying Li, and Gang Huang. StepFinder: A temporal semantic framework for failure attribution in multi-agent systems. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. URL https://arxiv.org/abs/2606.03467. arXiv:2606.03467.

A Deferred proofs

Roadmap. §A.1 shows that each restarted product $(E_j(t))_t$ is a nonnegative supermartingale under N2 (Lemma 1). §A.2 shows that fixed-weight mixtures of supermartingales inherit the same property with initial value at most one (Lemma 2). §A.3 applies Ville's inequality to give the time-uniform false-alarm bound (Theorem 1) and, via a killed process $Z_t = M_t \cdot \mathbf{1}[\mathrm{N1}_t]$, the pre-fault variant (Corollary 1). §A.4 proves the closure properties of Proposition 1. §A.5 proves Proposition 3; §A.6 handles the estimated-null theorem (Theorem 2) via the calibration good event $A = \{\widehat{S}_r \geq S_r \ \forall r\}$.

A.1 Proof of Lemma 1

Lemma (Lemma 1, restated: Anchored processes)**.** *Under* N2 *and Assumption 1, each* $(E_j(t))_t$ *is a nonnegative supermartingale with respect to* $(\mathcal{G}_t)$*, with* $E_j(0) = 1$*.*

Proof. Nonnegativity and adaptedness are immediate. Fix $t \geq 1$. If $t < j$ then $E_j(t) = E_j(t-1) = 1$. If $t \geq j$ then $E_j(t) = E_j(t-1)\, e_{R_t}(S_t)$, and $E_j(t-1)$ is $\mathcal{G}_{t-1}$-measurable, hence $\mathcal{G}^+_{t-1}$-measurable, and nonnegative. Therefore

$$\mathbb{E}\big[E_j(t) \,\big|\, \mathcal{G}^+_{t-1}\big] \;=\; E_j(t-1)\,\mathbb{E}\big[e_{R_t}(S_t) \,\big|\, \mathcal{G}^+_{t-1}\big] \;\leq\; E_j(t-1)$$

by N2, and towering down to $\mathcal{G}_{t-1} \subseteq \mathcal{G}^+_{t-1}$,

$$\mathbb{E}\big[E_j(t) \,\big|\, \mathcal{G}_{t-1}\big] \;=\; \mathbb{E}\Big[\,\mathbb{E}\big[E_j(t) \,\big|\, \mathcal{G}^+_{t-1}\big] \,\Big|\, \mathcal{G}_{t-1}\Big] \;\leq\; E_j(t-1). \qquad \square$$

A.2 Proof of Lemma 2

Lemma (Lemma 2, restated: Fixed-weight mixtures)**.** *Let* (c_j) *satisfy Assumption 1 with* $c_j \geq 0$ *and* $\sum_j c_j \leq 1$*. Then* $M_t = \sum_j c_j E_j(t)$ *is a nonnegative supermartingale with* $\mathbb{E}[M_0] \leq 1$*; under the paper's anchor indexing* $\mathcal{A} \subseteq [1, K_{\max}]$ *the pointwise bound* $M_0 \leq 1$ *in fact holds a.s.*

Proof. $M_0 = \sum_j c_j\, E_j(0) = \sum_j c_j \leq 1$ a.s., using $E_j(0) = 1$ for every $j \geq 1$ (all anchors are strictly positive-indexed under Section 3.4's convention $\mathcal{A} \subseteq [1, K_{\max}]$); the expectation bound follows. The weights are $\mathcal{G}_{t-1}$-measurable, so by linearity and Lemma 1, $\mathbb{E}[M_t \mid \mathcal{G}_{t-1}] \leq \sum_j c_j E_j(t-1) = M_{t-1}$. For an infinite index set apply conditional monotone convergence. $\square$

A.3 Proof of Theorem 1

Proof. Lemma 2 gives M a nonnegative supermartingale with $\mathbb{E}[M_0] \leq 1$. Ville's inequality bounds $\Pr[\sup_t M_t \geq c] \leq \mathbb{E}[M_0]/c$; take $c = 1/\alpha$. $\square$

A.3.1 Proof of Corollary 1

Proof. Work component by component. For each anchor $j \in \mathcal{A}$ define the killed component

$$Z_j(t) \;:=\; c_j\, E_j(t)\, \mathbf{1}[\mathrm{N1}_t],$$

so that $\sum_j Z_j(t) = M_t\,\mathbf{1}[\mathrm{N1}_t] =: Z(t)$. Each Z_j is $(\mathcal{G}_t)$-adapted ($c_j, j \in \mathcal{G}_0$ by Assumption 1; $\mathrm{N1}_t \in \mathcal{G}_t$ by Assumption 2; $E_j(t) \in \mathcal{G}_t$ by construction) and nonnegative.

Inactive anchor ($t+1 < j$): $E_j(t+1) = E_j(t) = 1$ by the convention $E_j(s) = 1$ for $s < j$, so $Z_j(t+1) = c_j\,\mathbf{1}[\mathrm{N1}_{t+1}]$. Since a fault cannot un-occur, $\mathrm{N1}_{t+1} \subseteq \mathrm{N1}_t$, so $\mathbf{1}[\mathrm{N1}_{t+1}] \le \mathbf{1}[\mathrm{N1}_t]$ pointwise, and $c_j\,\mathbf{1}[\mathrm{N1}_t] \in \mathcal{G}_t \subseteq \mathcal{G}_t^+$, giving $\mathbb{E}[Z_j(t+1) \mid \mathcal{G}_t^+] \le c_j\,\mathbf{1}[\mathrm{N1}_t] = Z_j(t)$. The current-turn factor plays no role in this case: an inactive anchor moves only through its indicator.

Active anchor ($t+1 \ge j$): the product update gives $E_j(t+1) = E_j(t)\,e_{R_{t+1}}(S_{t+1})$ uniformly on this branch (at the activation step $t+1 = j$: $E_j(t) = E_j(j-1) = 1$ by the inactive convention, so $E_j(t+1) = e_{R_j}(S_j)$ matches $\prod_{s=j}^{j} e_{R_s}(S_s)$). On $\{\mathrm{N1}_t^{\mathrm{c}}\}$, $\mathbf{1}[\mathrm{N1}_{t+1}] = 0$ by the nesting above, so $Z_j(t+1) = 0 = Z_j(t)$. On $\{\mathrm{N1}_t\}$, $c_j\,E_j(t) \in \mathcal{G}_t \subseteq \mathcal{G}_t^+$ factors out and $\mathbf{1}[\mathrm{N1}_{t+1}] \le 1$ pointwise, so

$$\mathbb{E}[Z_j(t+1) \mid \mathcal{G}_t^+] \;\le\; c_j\,E_j(t)\,\mathbb{E}[e_{R_{t+1}}(S_{t+1}) \mid \mathcal{G}_t^+] \;\le\; c_j\,E_j(t) \;=\; c_j\,E_j(t)\,\mathbf{1}[\mathrm{N1}_t] \;=\; Z_j(t),$$

where the second inequality is Assumption 2 on $\{\mathrm{N1}_t\}$.

Both cases give $\mathbb{E}[Z_j(t+1) \mid \mathcal{G}_t^+] \le Z_j(t)$, so each Z_j is a nonnegative $(\mathcal{G}_t^+)$-supermartingale. By linearity, using that the c_j are $\mathcal{G}_0$-measurable (hence $\mathcal{G}_t^+$-measurable for every t), the mixture $Z(t) = \sum_j Z_j(t)$ is a nonnegative $(\mathcal{G}_t^+)$-supermartingale. Ville's inequality requires $\mathbb{E}[Z(0)] \le 1$; this holds because $Z(0) = M_0\,\mathbf{1}[\mathrm{N1}_0] = M_0$ (using $\mathrm{N1}_0 = \Omega$ by convention: no fault has occurred before turn 1), and $M_0 = \sum_j c_j\,E_j(0) = \sum_j c_j \le 1$ almost surely under the paper's anchor indexing $\mathcal{A} \subseteq [1, K_{\max}]$ (every $j \ge 1$ has $E_j(0) = 1$). Ville then gives $\Pr(\sup_{t\ge 0} Z(t) \ge 1/\alpha) \le \alpha$.

Since $\{t < \nu\} = \{\mathrm{N1}_t\}$, $\{\sup_t Z(t) \ge 1/\alpha\} = \{\exists\, t : M_t \ge 1/\alpha \text{ and } t < \nu\} = \{\tau < \nu\}$, giving $\Pr[\tau < \nu] \le \alpha$. When $\nu = \infty$ almost surely, $\mathrm{N1}_t = \Omega$ for every t, $Z(t) = M_t$, and the bound reduces to Theorem 1. □

A.4 Proof of Proposition 1

Proof. (i) and (ii) are Lemmas 1 and 2, the integral form by Fubini. For (iii), conditional Jensen on the concave $x \mapsto x^\lambda$ gives $\mathbb{E}[e^\lambda \mid \mathcal{G}_{t-1}^+] \le (\mathbb{E}[e \mid \mathcal{G}_{t-1}^+])^\lambda \le 1$, then tower as in Lemma 1. For (iv), λ_s leaves the conditional expectation by predictability, giving $(1-\lambda_s) + \lambda_s\,\mathbb{E}[e \mid \mathcal{G}_{s-1}^+] \le 1$. Initial values are unchanged and each operation returns a process of the same type. □

A.5 Proof of Proposition 3

The proof uses the Bounded Multiplier auxiliary (Proposition 2 below), whose proof we record first.

Proposition 2 (Bounded multiplier)**.** *Under the construction of Section 3.3, for every role r and level i, $w_{\min}/w_{\max} \le \varepsilon_{r,i} \le w_{\max}/w_{\min} =: \kappa$.*

Proof. S_r is a convex combination of the w_j, so $w_{\min} \le S_r \le w_{\max}$, whence w_i/S_r lies in the stated interval. □

Probabilistic alarm bound.

Proof. Define the modified factors $\tilde{\varepsilon}_t := \varepsilon_t \mathbb{1}[t \notin B] + \mathbb{1}[t \in B]$. By (H1), $\mathbb{1}[t \notin B]$ and $\mathbb{1}[t \in B]$ are $\mathcal{G}_0$-measurable and therefore $\mathcal{G}_{t-1}^+$-measurable, so by (H2) $\mathbb{E}[\tilde{\varepsilon}_t \mid \mathcal{G}_{t-1}^+] \le 1$ for every t. For each anchor j, set $P_j(t) = \prod_{s=j}^{t} \tilde{\varepsilon}_s$ for $t \ge j$ and $P_j(t) = 1$ for $t < j$; by Proposition 1(i), P_j is a nonnegative supermartingale from 1.

For $t \ge j$, split $[j,t]$ over B and B^c:

$$E_j(t) \;=\; \prod_{s=j}^{t} \varepsilon_s \;=\; \Big(\prod_{s\in[j,t]\cap B^c} \varepsilon_s\Big)\Big(\prod_{s\in[j,t]\cap B} \varepsilon_s\Big) \;=\; P_j(t)\,Q_j(t),$$

where $Q_j(t) = \prod_{s\in[j,t]\cap B} \varepsilon_s$. By (H3), $Q_j(t) \le \kappa^{|B\cap[j,t]|} \le \kappa^{|B|}$, so $E_j(t) \le \kappa^{|B|}\, P_j(t)$ pointwise. Setting $\bar{P}(t) := \sum_j c_j\, P_j(t)$,

$$M_t \;=\; \sum_j c_j\, E_j(t) \;\le\; \kappa^{|B|}\, \bar{P}(t).$$

By Proposition 1(ii), $\bar{P}$ is a nonnegative supermartingale with initial value at most $\sum_j c_j \le 1$.

Assume $\alpha\kappa^{|B|} \le 1$, so the level $1/(\alpha\kappa^{|B|}) \ge 1$. Ville's inequality on $\bar{P}$ gives

$$\Pr\Big(\sup_{t\le T} M_t \ge 1/\alpha\Big) \;\le\; \Pr\Big(\sup_{t\le T} \bar{P}(t) \ge 1/(\alpha\kappa^{|B|})\Big) \;\le\; \alpha\,\kappa^{|B|}.$$

The single-anchor case is $c_{j_0} = 1$ and $c_j = 0$ otherwise, and reduces to the same argument applied to P_{j_0}. □

A.6 Proof of Theorem 2

Proof. Condition on $\mathcal{D}$. By independence, $\widehat{S}_r$ is $\mathcal{G}_0$-measurable for the monitored trajectory. Under a law in $\mathcal{P}$, $\mathbb{E}[\widehat{e}_{R_t}(S_t) \mid \mathcal{G}^+_{t-1}] = S_{R_t}/\widehat{S}_{R_t}$; on $A = \{\widehat{S}_r \ge S_r\ \forall r\}$ this is at most 1, giving N2, so Theorem 1 applies conditionally, which is (a). For (b), write $P(\mathcal{D})$ for that conditional probability; $\mathbb{E}[P] = \mathbb{E}[P\,\mathbf{1}_A] + \mathbb{E}[P\,\mathbf{1}_{A^c}] \le \alpha\Pr[A] + \Pr[A^c] \le \alpha(1-\Pr[A^c]) + \Pr[A^c] \le \alpha(1-\delta)+\delta$, maximising the affine bound in $\Pr[A^c] \in [0,\delta]$. □

A.7 Proof of Lemma 3

Proof. Since $\widehat{S}_r$ is $\mathcal{G}_0$-measurable and $\mathcal{G}_0 \subset \mathcal{G}^+_{t-1}$, both $\log\widehat{S}_r$ and $\log S_r$ are $\mathcal{G}^+_{t-1}$-measurable and come out of the conditional expectation as constants. Writing $q^{(t)}_{r,i} := Q(S_t = i \mid \mathcal{G}^+_{t-1}, R_t = r)$ for the history-conditional Q-law of the score,

$$\begin{aligned}
\mathbb{E}_Q\big[\log e_r(S_t) \;\big|\; \mathcal{G}^+_{t-1}, R_t = r\big] &= \sum_i q^{(t)}_{r,i}\big(\log w_i - \log\widehat{S}_r\big) = \sum_i q^{(t)}_{r,i}\log w_i - \log\widehat{S}_r \\
&= \Big(\sum_i q^{(t)}_{r,i}\log w_i - \log S_r\Big) - \big(\log\widehat{S}_r - \log S_r\big) \\
&= \mathbb{E}_Q\big[\log e^*_r(S_t) \;\big|\; \mathcal{G}^+_{t-1}, R_t = r\big] - \log(\widehat{S}_r/S_r).
\end{aligned}$$

On A, $\widehat{S}_r \ge S_r$ so the loss term is ≥ 0. When the clip $\widehat{S}_r = w_{\max}$ binds, the identity still holds with loss $\log(w_{\max}/S_r)$. When the unobservable S_r is proxied by the calibration mean $\bar{x}_r$ for reporting purposes, the reported loss is $\log(\widehat{S}_r/\bar{x}_r)$. □

B Extended remarks

The following remarks elaborate on the main-body statements. Numbering follows their appearance order in Sections 3 and 4.

B.1 Why a weight vector rather than a score transform

Remark 2 (Why a weight vector rather than a score transform). The natural construction on a $[0,1]$-valued score is $e = (1-S_t)/(1-p_r)$ with p_r the null mean score. It is calibrated, and is the special case of equation 3 with $w_i = 1-\ell_i$. Its difficulty is numerical rather than logical. When the judge places most of its null mass on the top level, the map assigns that level a weight of $1-\ell_L$, which for a rubric whose top anchor is 1 is zero exactly: a single top-scored turn would multiply the process by zero and end it permanently. Implementations avoid this only by clamping the top anchor away from 1, after which the weight is a small positive number and the factor on a typical null turn is a small quantity divided by a small quantity: the median factor collapses toward zero even as the mean stays at one by construction.

This is the case $w_L \to 0$, in which the bounded-multiplier guarantee of Proposition 2 degenerates. Choosing w freely rather than inheriting it from the score scale removes both the need for a clamp and the degeneracy it papers over.

B.2 Concentration on the top level is harmless here

Remark 3 (Concentration on the top level is harmless here). When the null concentrates on the top level, $S_r \to w_L = w_{\min}$, so the factor applied to a typical null turn is $\varepsilon_{r,L} = w_L/S_r \to 1$. The process is near-lossless on ordinary null turns precisely in the regime that destroys the linear map. The clustering that costs the linear construction everything costs this construction almost nothing, which is the mechanism behind the empirical repair rather than merely its description.

B.3 Choosing w, and unseen roles

Remark 4 (Choosing w, and unseen roles). Any decreasing positive w yields a calibrated map, so w is a design parameter. The principled choice maximises expected log growth under the alternative subject to equation 2, which on a finite support with one linear constraint is a small convex program in L variables. We fix w in advance, before any faulted trajectory is inspected, and treat it as an approximation to that solution rather than claiming optimality.

A role unseen during calibration has no estimated normaliser. Since S_r is a convex combination of the w_j, the choice $S_r = w_{\max}$ is valid with certainty and contributes nothing to the error budget of Theorem 2; it yields $\varepsilon_{r,i} = w_i/w_{\max} \leq 1$, so an unrecognised role can only decrease the process and never triggers an alarm on its own. Unknown roles are therefore handled conservatively at no cost to the guarantee.

An empirical note on the log-growth objective. We prespecified a six-candidate family — linear (shipped), geometric, steep, flat, low-penalty, high-reward — and ran a search under the Kelly-log-optimal criterion named above: mean post-onset log ε on dev failing trajectories (prespec at `results/score_evidence_search_prespec_20260917.txt`; driver at `code/score_evidence_search.py`). On both AFTraj math and Who&When HC the winner was the flattest candidate, $w = (3, 2.5, 2, 1.5, 1)$ with $\kappa = w_{\max}/w_{\min} = 3$; it beat the shipped linear weight on the mean without concentrating evidence in a few turns (post-onset log-ε standard deviation 0.39 vs linear's 0.59 on math; 0.38 vs 0.57 on HC). Held-out detection did not follow: on math restricted detection fell (baseline linear $4/58$ under Path-A vs the flat-score winner's $2/58$ under per-trace K'; we do not re-run the flat-score search under fixed K'); on Who&When HC $15/58 \to 13/58$; pre-mistake Type-I was unchanged on both. The mechanism is a criterion mismatch. Detection at a fixed threshold $h = 1/\alpha$ is a peak-crossing event and depends on the tail of the per-turn factor distribution; the log-growth criterion is a statement about that distribution's mean under the alternative. Flattening the mapping raises the mean by shrinking the losses on the top-scored turns that dominate on our judge's rubric (about 57–65% at the top level on Who&When HC), and it shrinks the peaks that push wealth over h in the process. The two objectives coincide asymptotically as $\alpha \to 0$, but at $\alpha = 0.05$ with chains of 30–50 turns we are far from that regime. This is a finding about the criterion, not a defence of our weight vector: we kept linear because the search did not beat it, and the reason it did not is informative.

B.4 Wrapping an external detector

Remark 5 (Wrapping an external detector). Nothing in equation 3 refers to a judge. The construction applies to any per-turn signal taking finitely many values, or to any bounded signal discretised to finitely many bins, provided a reference law π_r for that signal is available from fault-free trajectories. An existing online auditor that emits a per-prefix score can therefore be wrapped without modification: its scores replace S_t, and the guarantee attaches to the wrapper rather than to the auditor.

B.5 N2 is genuinely an assumption

Remark 6 (N2 is genuinely an assumption). Calibration in the sense of equation 2 is a statement about the *role-marginal* null law of S_t. N2 is the stronger statement that the constraint holds con-

ditionally on the entire history. The two coincide exactly on the class $\mathcal{P}$ of equation 1, and diverge otherwise: if score behaviour depends on task difficulty, on position in the chain, or on whether earlier turns were themselves degraded, a factor calibrated to the role marginal need not satisfy N2 pointwise. A calibration sample identifies the marginal; it cannot certify the conditional. We therefore state N2 explicitly rather than deriving it, and Section 5 reports a stress test on the subset where our diagnostics find residual heterogeneity beyond role.

A design consequence is worth naming. A judge that scores turn t in light of the entire prefix produces scores whose conditional distribution depends on more than the acting role, which is precisely the property that Equation 1 — not N2 itself — excludes. N2 is a bound on the conditional expectation of a factor and permits serial dependence in the underlying scores; it is Equation 1, the role-conditional stability assumption, that rules out history-dependent score laws. A more turn-local judge makes Equation 1 more plausible at some cost in context. The scoring interface is therefore a lever on the assumption, not only a source of noise.

Remark 7 (Why some bridge from N1 to a score-distribution condition is unavoidable). If the semantic null placed no restriction on the score distribution, any valid factor would satisfy $e \leq 1$ at every level: choosing a null law concentrated on level ℓ forces π_r to be a point mass and $\mathbb{E}[e] = e(\ell) \leq 1$ for validity, so no factor could ever exceed one at ℓ. Repeating this across levels forces $e \leq 1$ everywhere. Assumption 2 is therefore not a matter of convenience: some link from fault-free semantics to a per-role score distribution is required for any e-value construction on this signal to have power at all.

B.6 Provenance

Remark 8 (Provenance). Convex combination and countable stitching of nonnegative supermartingales, and the betting form, are established results in the e-value literature. We record them here only to make the scope of the calibration explicit.

B.7 Why mix rather than maximise

Remark 9 (Why mix rather than maximise). The natural alternative to mixing over restarts is to take their running maximum, by analogy with CUSUM. A running maximum of nonnegative supermartingales is not in general a supermartingale, and Ville does not apply to it. Mixing is the operation that preserves the guarantee; maximising is not.

B.8 Spacing sweeps do not compare grids on equal footing

Remark 10 (Spacing sweeps do not compare grids on equal footing). Coarsening the anchor grid scales the dilution constant K' applied to every anchored process, so a raw comparison of two spacings mixes an arithmetic component with any true empirical effect on either error rate. We therefore report spacing comparisons only at matched Type-I. Empirical anchor-spacing behaviour on Who&When and on a synthetic $K \times M$ grid is available at `code/anchor_spacing_curve.py` and `code/anchor_kxm_sweep.py`; we do not use those runs as a quantitative illustration here.

B.9 What the grid is for, and what it is not

Remark 11 (What the grid is for, and what it is not). Spacing being neutral on average does not make the mixture pointless. A single forward process started at turn one has decayed by roughly ν^{u-1} by the time a burst at onset u arrives, so its detection is strongly dependent on where the burst falls: early onsets are favoured and late ones heavily penalised. The mixture equalises across onset position at no average cost. This predicts that restarts earn their keep on long chains, where late onsets are common and forward decay is severe, and matter little on short chains where ν^{u-1} remains near one throughout. Whether that holds on our corpora is an empirical question we report in Section 5, not one settled here.

B.10 The bad event is shared, not resampled

Remark 12 (The bad event is shared, not resampled). Conclusion (b) can be misread as granting each trajectory an independent budget of $\alpha + \delta$. It does not. The calibration sample is drawn once

and reused for every trajectory the monitor sees, so the event A^c is common to all of them. Across N monitored trajectories, the expected number of false alarms is at most $N\alpha$ on A, and may be as large as N on A^c. This is why we treat (a) as the operationally meaningful statement: it says that with high probability the deployment as a whole is calibrated, rather than that each run is calibrated on average over deployments that never happened.

B.11 Constructing $\widehat{S}_r$, and the role of δ

Assumption 3 (Constant conditional mean on calibration)**.** *The calibration sample $\mathcal{D}$ is drawn once, prior to deployment. Along the role-r subsequence with filtration $(\mathcal{F}_s)$, $\mathbb{E}[w(S_s) \mid \mathcal{F}_{s-1},\ R_s = r] = S_r$.*

Assumption 3 is a modelling premise on the calibration draw, stronger than Assumption 2: the bridge is a one-sided inequality on the monitored trajectory, while Assumption 3 pins the conditional mean on the calibration subsequence and permits dependence, not arbitrary history-dependent conditional means. Both are stated, neither is derived; the theorem of Waudby-Smith & Ramdas (2023) requires the conditional-mean equality on the tested subsequence.

Remark 13 (Constructing $\widehat{S}_r$, and the role of δ). $S_r = \mathbb{E}[w(S_t) \mid R_t = r]$ is a bounded mean under the process-level null. The Maurer–Pontil empirical Bernstein bound (Maurer & Pontil, 2009) is fixed-n and assumes iid observations. Dev-pool turns within a trajectory are not independent (Verifier lag-1 rubric autocorrelation is $\rho_1 = 0.44$, computed on the rubric-level index — an affine function of w, so the same correlation as on the weights $w(S_t)$ — on dev math, pooling pairs (prev turn, current Verifier turn) with prev turn of any role; Appendix Q). We therefore use the betting-based one-sided confidence sequence of Waudby-Smith & Ramdas (2023), which is valid under the calibration-law Assumption 3 above (constant role-r conditional mean on the calibration subsequence) and does not require independence. It does require a constant conditional mean, which makes $w(S_s)$ uncorrelated with its past; the nonzero lag-1 autocorrelation above is evidence against that premise on this role, so on math the WSR bound is a better-motivated reference rather than a certified one. It is also empirically tighter than Maurer–Pontil at every role in our tables (widths shrink by 25–61% across math and coding). Bounding S_r directly is preferable to bounding each $\pi_{r,i}$ and combining, since S_r is linear in π_r and a combination of per-cell worst cases is not attainable. The width scales as $n_r^{-1/2}$, so δ buys validity and calibration sample size buys it cheaply. Roles observed too sparsely for a useful bound may either be pooled by a rule fixed in advance or assigned the certain bound $S_r = w_{\max}$ of Remark 4.

δ is a reported quantity, not a hidden constant. Larger δ tightens $\widehat{S}_r$ and raises power at a weaker guarantee; smaller δ deflates every factor uniformly. We fix $\delta = 0.005$ throughout; a δ sweep is on the deferred list in the reproducibility statement.

Betting confidence-sequence construction, made explicit. For each role r with $n_r \geq 2$ calibration observations $w_1, \ldots, w_{n_r}$ (per-turn weights $w_s = w(S_s)$ on turns whose role equals r; the calibration pool is the development safe-labeled split), rescale to $[0,1]$ via $X'_s = (w_s - w_{\min})/R$ with $R = w_{\max} - w_{\min}$ (here $w_{\max} = 5$, $w_{\min} = 1$, so $R = 4$). For a candidate lower bound $m' \in (\overline{X'}, 1)$ and a nonneg predictable schedule (λ_s), form the reverse-side wealth

$$K^-_{n_r}(m') \;=\; \prod_{s=1}^{n_r} \big(1 + \lambda_s\,(m' - X'_s)\big). \tag{5}$$

Under $H_0(m')$: $\mathbb{E}[X'_s \mid \mathcal{F}_{s-1}] \geq m'$, K^- is a nonneg supermartingale (conditional multiplier $1 + \lambda_s(m' - \mathbb{E}[X'_s \mid \mathcal{F}_{s-1}]) \leq 1$ whenever $\lambda_s \geq 0$); Ville's inequality gives $\mathbb{P}(K^-_{n_r}(m') \geq 1/\delta_r) \leq \delta_r$. The one-sided upper CB at level $1 - \delta_r$ is the upper endpoint of the non-rejected set, unrescaled to the original weight scale:

$$m^\star_r \;=\; \sup\big\{m' \in (\overline{X'_r},\, 1) :\; K^-_{n_r}(m') < 1/\delta_r\big\}, \qquad \widehat{S}_r \;=\; \min\big\{w_{\max},\; w_{\min} + R \cdot m^\star_r\big\}, \tag{6}$$

with the conventions $m^\star_r = \overline{X'_r}$ if every m' in the interval is rejected and $m^\star_r = 1$ if none is. For a candidate-independent nonneg predictable schedule the rejection set is monotone in m' by construction: $\partial \log K^-(m')/\partial m' = \sum_s \lambda_s/(1 + \lambda_s(m' - X'_s)) \geq 0$ whenever the nonnegativity constraint $1 + \lambda_s(m' - X'_s) > 0$ holds, so $K^-(\cdot)$ is non-decreasing in m' and the smallest-rejected form $\inf\{m' : K^-_{n_r}(m') \geq 1/\delta_r\}$ agrees with equation 6 identically. No grid check is required (Waudby-

Smith & Ramdas 2023, App. B.1); the candidate-dependent aGRAPA schedule of Waudby-Smith & Ramdas (2023, App. B.3), whose sublevel sets need not be intervals (Waudby-Smith & Ramdas, 2023, App. E.4), is a different construction and does not apply here.

Testing $m = S_r$ on the original scale corresponds to testing $m' = \mu'_r := (S_r - w_{\min})/R$ on the rescaled scale, and requires $\mathbb{E}[X'_s \mid \mathcal{F}_{s-1}] \geq \mu'_r$ along the role-r calibration subsequence, equivalently $\mathbb{E}[w(S_s) \mid \mathcal{F}_{s-1}, R_s = r] \geq S_r$, so that $K^-_{n_r}(\mu'_r)$ is a supermartingale. Assumption 3 supplies exactly this: with $\mathbb{E}[w(S_s) \mid \mathcal{F}_{s-1}, R_s = r] = S_r$ we get $\mathbb{E}[\mu'_r - X'_s \mid \mathcal{F}_{s-1}] = 0$, so $K^-_{n_r}(\mu'_r)$ is a martingale. Assumption 2 on its own supplies only the one-sided $\mathbb{E}[w(S_s) \mid \mathcal{G}^+_{s-1}, R_s = r] \leq S_r$ on N1, which is the opposite direction of what K^- needs; it does not by itself justify coverage at $m = S_r$. Assumption 3 is therefore introduced as a stronger, explicitly stated premise on the calibration draw, not derived from the trajectory-level bridge.

Predictable schedule and admissible range. We use the candidate-independent predictable plug-in empirical-Bernstein schedule of Waudby-Smith & Ramdas (2023, §3.2 and App. B.1), $\lambda_s = \min\big(\sqrt{2\log(1/\delta_r)/(n_r \widehat{V}_{s-1})},\ 0.90\big)$, with running mean and variance $\widehat{\mu}_s, \widehat{V}_s$ built from $X'_1, \ldots, X'_{s-1}$ only (initialised at $\widehat{\mu}_0 = \frac{1}{2}$, $\widehat{V}_0 = \frac{1}{4}$). The schedule does not depend on the candidate m'; the candidate-dependent aGRAPA schedule of Waudby-Smith & Ramdas (2023, App. B.3) is a different construction. The nonnegativity restriction $1 + \lambda_s(m' - X'_s) \geq 0$ for every $m', X'_s \in [0,1]$ is necessary and sufficiently ensured by $\lambda_s \in [0, 1)$ (worst case $m' = 0,\ X'_s = 1$ gives $1 - \lambda_s$; the opposite corner is automatically nonneg for $\lambda_s \geq 0$). The cap $\lambda_s \leq 0.90$ is a sufficient choice with margin, keeping $\log(1 + \lambda_s(m' - X'_s))$ bounded away from $-\infty$ for numerical stability. The scalar n_r used to tune the schedule is the size of the role-r subsequence in the fixed dev-fold (a partition of trajectories decided before scoring), so it is not a stopping time on turn-level scores and no routing loop feeds back into the calibration counts. We take $\delta_r = \delta/K$ with K the number of distinct roles per domain, giving overall calibration failure probability $\leq \delta$ on the good event of Theorem 2. Roles with $n_r < 2$ or observed range $= 0$ receive the certain bound $\widehat{S}_r = w_{\max}$ of Remark 4. The observations w_s are per-turn within a trajectory, not per-trajectory averages: the calibration target is the turn-level normaliser $S_r = \mathbb{E}[w(S_t) \mid R_t = r]$ under Assumption 3.

Remark 14 (Sampling design underlying predictability). The calibration pool $\mathcal{D}$ is scored exhaustively in an order fixed before scoring begins (trajectories in ID order, turns within a trajectory in time order). n_r denotes the total count of role-r turns in $\mathcal{D}$ and is $\mathcal{F}_0$-measurable by this design. Under the calibration filtration $(\mathcal{F}_s)$ induced by the enumeration, λ_s is $\mathcal{F}_{s-1}$-measurable (prefix-statistic + constant horizon), so Assumption 1 in the sense of Waudby-Smith & Ramdas (2023) is satisfied. A sequential-only alternative that avoids referencing n_r (e.g. $\lambda_s = \min\big(\sqrt{2\log(s+1)/((s+1)\widehat{V}_{s-1})},\ 0.9\big)$) would widen $\widehat{S}_r$ on smaller roles by discarding the known horizon; the assumption is retained because it matches the sampling design as executed.

B.12 Relation to prior work

Remark 15 (Relation to prior work). The calibration-good-event device applied in Theorem 2 is standard: Section 2 names its appearances in PAC-thresholded sequential monitoring (Sadhuka et al., 2025) and anytime-valid conformal retrieval (Dubey & Huo, 2026), and Remark 1 states the contrast in application.

B.13 Constructions considered and set aside

Proposition 1 makes several alternatives admissible, and we evaluated them. The negative results are informative about where the difficulty lies.

Betting. Proposition 1(iv) is the standard form in the sequential testing literature and protects against ruin. By Proposition 2 that protection is already present, and empirically every betting rate below one reduced power without improving false-alarm control. We use $\lambda \equiv 1$.

Time-averaged statistics. A running average $t^{-1}\sum_{s \leq t} e_s$ is not a supermartingale: its conditional increment is bounded by $\frac{t-1}{t} A_{t-1} + \frac{1}{t}$, which exceeds A_{t-1} whenever $A_{t-1} < 1$, the regime

the process occupies under the null. A horizon-normalised sum $U_t = K^{-1}\sum_{s\le t} g_s$ with each g_s an e-value is nondecreasing with $\mathbb{E}[U_K] \le 1$, so $\max_{t\le K} U_t = U_K$ and Markov's inequality gives $\Pr[\max_{t\le K} U_t \ge 1/\alpha] \le \alpha$ directly. That bound is specific to the declared horizon and is not anytime-valid beyond it, which is the trade such constructions make. Neither improved on the product in our experiments.

Role-selective weighting. Since roles differ in how informative their scores are, one may concentrate the mixture on high-information roles; any such weighting fixed in advance is valid by Proposition 1(ii). Empirically this produced no improvement surviving correction for multiple comparisons, and we report it as evidence that role structure is informative rather than as a component of the method.

B.14 Loss-limiting property of the restart mixture

Remark 16 (W_{loc} dilutes per-turn e-power loss, empirical). For anchors $\mathcal{A}$ of spacing m over a trajectory of length T, each sub-process $E_j(T) = \prod_{s=j}^{T} e_{R_s}(S_s)$ inherits loss only over turns $t \ge j$; its expected log-loss is $\sum_{t\ge j} \log(\widehat{S}_{R_t}/S_{R_t})$, so late anchors carry less loss than early ones. The mixture wealth $W_{\mathrm{loc}}(T) = K'^{-1}\sum_j E_j(T)$ inherits a loss bounded by some average of the E_j losses, but the natural bound is not clean: the identity

$$\log W^*_{\mathrm{loc}}(T) - \log W_{\mathrm{loc}}(T) \;=\; \log\Big(\sum_j \pi_j(T)\,\exp L_j(T)\Big) \quad \text{with } \pi_j(T) = \frac{E_j(T)}{\sum_k E_k(T)},$$

and $L_j(T) = \log E^*_j(T) - \log E_j(T)$, gives only the loose bound $\max_j L_j = L_1$ (the earliest-anchor loss) directly, and a Jensen bound on the same expression signs the wrong way. We therefore report the divisor ratio empirically rather than as a proven bound. Verification on the three agentic held-out trajectories where the bounded reference kills detection (the third, `hotpotqa_0160_unsafe_diagnosed`, becomes a lost trajectory under the tighter WSR reference; the two already lost under Maurer–Pontil are still lost, with reduced margins):

trajectory	K	role mix	W_{fwd}-max divisor	W_{loc} divisor
`wrong_count_propagation`	21	13/7/1 (task/env/user)	15.60	9.79
`stale_db_data_propagation`	25	16/8/1 (task/env/user)	29.13	11.59
`hotpotqa_0160_unsafe_diagnosed`	7	4/1/1/1 (Manager/user/search/env)	1.93	1.69

On all three trajectories the mixture divisor is 1.1–2.5× smaller than the forward product's max (`results/mixture_loss_limiting_v4_wsr_20260910.csv`); the smallest factor (1.14× on `hotpotqa_0160`) reflects the shorter $K = 7$ trajectory, where late-anchor loss dilution has less room to run. We report this as an empirical finding on the trajectories in the table and do not extrapolate; it is a third empirical reason to prefer W_{loc} under an estimated reference, independent of burst structure (Section 3.4) and of held-out performance.

C Full nine-cell AFTraj grid

Plain Ville threshold $1/\alpha = 20$, shipped configuration (categorical 1b; math, coding, and agentic all under the WSR bounded reference at fixed $\delta = 0.005$, fixed per-domain anchor budget $K' = 12, 4, 7$ from the census rule of §5.1). Detection under the restricted convention (alarm at $t \ge$ onset counts as detection), matching Table 2. The pooled-unsafe column carries a rate on the full failing pool as a check on whether the held-out failing subset is representative; it is not a reported quantity in Section 5. Numbers from the AFTraj shipped-configuration table.

domain	stat	held-out safe (Type-I) k/n	rate	held-out unsafe k/n	rate	pooled unsafe k/n	rate
math	W_{fwd}	9/58	15.5%	10/58	17.2%	34/397	8.6%
math	W_{loc}	2/58	3.4%	4/58	6.9%	17/397	4.3%
math	M_{dual}	5/58	8.6%	10/58	17.2%	29/397	7.3%
coding	W_{fwd}	0/54	0.0%	0/37	0.0%	15/247	6.1%
coding	W_{loc}	0/54	0.0%	0/37	0.0%	7/247	2.8%
coding	M_{dual}	0/54	0.0%	0/37	0.0%	11/247	4.5%
agentic	W_{fwd}	0/57	0.0%	4/68	5.9%	20/470	4.3%
agentic	W_{loc}	1/57	1.8%	2/68	2.9%	14/470	3.0%
agentic	M_{dual}	0/57	0.0%	2/68	2.9%	18/470	3.8%

D ALPHA SWEEP

Rejection rate as α varies, $1/\alpha$, shipped configuration. AFTraj rows from the level-sweep table (math, coding, and agentic all under the WSR bounded reference at fixed $\delta = 0.005$, fixed per-domain K'; one factor process serves every α). Who&When rows from the Who&When level-sweep table (shrunk plug-in). *Sweep tables use the unrestricted (any-alarm) count for continuity with the shipped artifact; primary detection numbers in Table 2 use the restricted convention of §5.4.* All rates are k/n percentages of trajectories on which the running-max reaches $1/\alpha$.

Two caveats on reading the sweep. Below $\alpha \approx 0.02$ it is uninformative at these sample sizes: a single alarm in 54 trajectories already exceeds the level, so an empirical zero at $\alpha = 0.01$ is a statement about sample size rather than about the method. Multiplicity correction is applied within each level rather than across the grid, since tests at different levels on the same cell are the same alarm indicators read at different thresholds and no decision changes under correction. The forward product W_{fwd} drifts above nominal on math most strongly at $\alpha = 0.05$ (point estimate $9/58 = 15.5\%$, one-sided binomial $p = 0.0022$ uncorrected; the elevation survives Benjamini–Hochberg correction across the nine AFTraj cells (BH $q = 0.020$; reproducibility statement item 4)) and remains marginally elevated at $\alpha = 0.10$ ($10/58 = 17.2\%$, $p = 0.060$); at $\alpha \geq 0.15$ the drift is within Clopper–Pearson coverage. The dual mixture M_{dual} sits at $5/58 = 8.6\%$ on math at $\alpha = 0.05$ (point estimate above nominal, $p = 0.164$; Clopper–Pearson 95% CI $[0.029, 0.190]$ covers 5%, so not formally elevated). The restart mixture W_{loc} math sits at or below nominal at every $\alpha \geq 0.05$ in the grid; at the tight levels $\alpha \in \{0.01, 0.02\}$ its cell sits at 3.4%, above nominal. We report the mixture as primary for the a-priori burst-structure reason of Section 3.4, and this sweep is consistent with that choice.

D.1 TYPE-I ON HELD-OUT SAFE (AFTRAJ) AND PRE-MISTAKE (WHO&WHEN)

		α						
cell	n	.01	.02	.05	.10	.15	.20	.25
math W_{fwd}	58	3.4%	6.9%	15.5%	17.2%	19.0%	20.7%	24.1%
math W_{loc}	58	3.4%	3.4%	3.4%	5.2%	6.9%	8.6%	13.8%
math M_{dual}	58	3.4%	5.2%	8.6%	15.5%	17.2%	17.2%	19.0%
coding W_{fwd}	54	0.0%	0.0%	0.0%	0.0%	0.0%	1.9%	1.9%
coding W_{loc}	54	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
coding M_{dual}	54	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
agentic W_{fwd}	57	0.0%	0.0%	0.0%	3.5%	3.5%	3.5%	7.0%
agentic W_{loc}	57	0.0%	0.0%	1.8%	1.8%	1.8%	1.8%	1.8%
agentic M_{dual}	57	0.0%	0.0%	0.0%	1.8%	3.5%	3.5%	3.5%
hc W_{fwd}	54	1.9%	1.9%	3.7%	5.6%	7.4%	7.4%	7.4%
hc W_{loc}	54	1.9%	1.9%	1.9%	5.6%	7.4%	7.4%	7.4%
ag W_{fwd}	72	1.4%	2.8%	2.8%	4.2%	4.2%	5.6%	5.6%
ag W_{loc}	72	0.0%	1.4%	1.4%	2.8%	2.8%	2.8%	2.8%

		α						
cell	n	.01	.02	.05	.10	.15	.20	.25
math W_{fwd}	58	3.4%	13.8%	22.4%	25.9%	31.0%	46.6%	48.3%
math W_{loc}	58	1.7%	1.7%	8.6%	10.3%	22.4%	27.6%	27.6%
math M_{dual}	58	3.4%	5.2%	20.7%	24.1%	27.6%	29.3%	39.7%
coding W_{fwd}	37	0.0%	0.0%	0.0%	8.1%	10.8%	18.9%	24.3%
coding W_{loc}	37	0.0%	0.0%	0.0%	2.7%	5.4%	18.9%	21.6%
coding M_{dual}	37	0.0%	0.0%	0.0%	2.7%	10.8%	21.6%	24.3%
agentic W_{fwd}	68	0.0%	1.5%	5.9%	10.3%	14.7%	16.2%	19.1%
agentic W_{loc}	68	0.0%	0.0%	2.9%	5.9%	8.8%	8.8%	8.8%
agentic M_{dual}	68	0.0%	1.5%	2.9%	8.8%	11.8%	14.7%	16.2%
hc W_{fwd}	58	13.8%	17.2%	22.4%	24.1%	27.6%	29.3%	31.0%
hc W_{loc}	58	19.0%	22.4%	27.6%	32.8%	34.5%	34.5%	34.5%
ag W_{fwd}	126	0.8%	1.6%	1.6%	2.4%	4.0%	4.0%	4.8%
ag W_{loc}	126	0.8%	0.8%	0.8%	1.6%	1.6%	1.6%	2.4%

D.2 Power on held-out unsafe (AFTraj) and full-trace (Who&When)

E Randomized Ville: a negative result

The Ramdas–Manole randomized-Ville inequality (Ramdas & Wang, 2025) strengthens the deterministic bound: for a nonnegative supermartingale M with $\mathbb{E}[M_0] \le 1$ and any stopping time τ,

$$\Pr(\exists t < \tau : M_t \ge 1/\alpha \text{ or } M_\tau \ge Z/\alpha) \;\le\; \alpha, \qquad Z \sim \mathrm{Unif}(0,1) \text{ indep. of the filtration.}$$

Applied at end-of-trace stopping time, the sequential rule fires as before, and additionally an alarm is declared at trace end if $W_T \ge Z/\alpha$. Since $Z < 1$ that threshold is below 20; traces that accumulate evidence without reaching 20 can still reject. Precondition status: W_{fwd} is a nonnegative supermartingale under N2 (Proposition 1). The mixture statistics W_{loc} and M_{dual} satisfy the closure result when the anchor grid is $\mathcal{G}$-predictable; the comparator retained here uses a per-trace anchor budget whose normaliser depends on the trace length K, so the mixture weights are not $\mathcal{G}$-predictable and Proposition 1 does not apply to those rows.

Stopping-time condition. The inequality requires τ to be a stopping time of the monitor's filtration. On AFTraj complete safe trajectories the trace end T is a stopping time, so the W_{fwd} rows below are licensed by that condition. The AFTraj W_{loc} and M_{dual} rows retain the per-trace anchor budget whose normaliser depends on K (see the Preconditions clause above), so their mixture weights are not $\mathcal{G}$-predictable and Proposition 1 does not apply: we report those entries as empirical comparisons rather than randomised-Ville-controlled rates. On Who&When pre-mistake windows the window end is the annotated onset $u-1$; annotation is not $\mathcal{G}_{u-1}$-measurable (the oracle looks at the trajectory globally, not at the monitor's filtration), so those rows evaluate M_{u-1} against Z/α at what is not a stopping time and the randomised bound is not licensed there. We mark those rows ‡ in Table 3; their rates lie outside the randomised bound's scope and are reported as empirical rates only.

Power gains, and the false-alarm cost. Analytical expected rejection under a fresh Z per trace is $(1/n)\sum_i\{\mathbf{1}[\max_i \ge h]+\mathbf{1}[\max_i < h]\cdot\min(1, \alpha\cdot\mathrm{final}_i)\}$. Numbers below are that expectation; the Monte Carlo mean over 1000 draws of n uniform Zs (one per trace, per draw) matches to sampling precision (fixed-vs-randomized Ville sweep table in the release archive).

Empirical rates, without a causal account. W_{loc} controls under the fixed rule at every reported α on non-math AFTraj cells and on both Who&When subsets (also at $\alpha = 0.25$; the two lowest levels sit at the sample-size limit); on math it sits marginally above nominal at $\alpha = 0.10$ (10.3%) under the retained per-trace anchor budget of this table; the same cell under the shipped Path-A fixed $K' = 12$ is 5.2% (App D), and inside nominal at every $\alpha \ge 0.05$ in the grid other than 0.10, under either budget. Under randomization the empirical Type-I on math (AFTraj complete safe trajectories) sits above nominal at every $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$ under the shipped bounded reference (rand rates 6.8%, 12.8%, 15.6%, 22.2% against nominal 5%, 10%, 15%, 20%). Under the per-trace anchor budget of this table the mixture weights are not $\mathcal{G}$-predictable (Preconditions clause above), so the randomised bound is not licensed on the W_{loc} or M_{dual} rows here; we report the excess as

Table 3: Fixed and randomized Ville, $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$. This table retains the earlier per-trace anchor budget for the fixed-vs-randomized comparator; the shipped Path-A fixed-per-domain K' gives different absolute cell values (see App D) but the fix-vs-rnd comparison is not sensitive to that choice. Levels outside this range are omitted for readability: the two lowest ($\alpha = 0.01, 0.02$) sit at the sample-size limit flagged in the Appendix D preamble; $\alpha = 0.25$ is dropped to keep the row within `textwidth`. Column pairs abbreviate `fix` (fixed Ville) and `rnd` (randomized Ville). Rows marked (any-rej.) are unsafe any-rejection rates over all observed turns including the terminal turn, not restricted-detection estimates as in Table 2. Underlying counts, all seven levels, in the fixed-vs-randomized Ville sweep table.

			$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.15$		$\alpha = 0.2$	
subset (regime)	stat	n	fix	rnd	fix	rnd	fix	rnd	fix	rnd
math (Type-I)	W_{fwd}	58	15.5%	16.0%	17.2%	18.1%	19.0%	20.1%	20.7%	21.7%
	W_{loc}	58	3.4%	6.8%	10.3%	12.8%	12.1%	15.6%	19.0%	22.2%
HC pre (Type-I)[‡]	W_{fwd}	54	3.7%	6.1%	5.6%	10.2%	7.4%	11.7%	7.4%	13.2%
	W_{loc}	54	1.9%	8.4%	5.6%	16.4%	7.4%	21.7%	7.4%	26.5%
math (any-rej.)	W_{fwd}	58	22.4%	26.8%	25.9%	32.9%	31.0%	40.0%	46.6%	52.8%
	W_{loc}	58	13.8%	23.1%	22.4%	34.2%	34.5%	45.7%	39.7%	51.3%
HC full (any-rej.)	W_{fwd}	58	22.4%	24.5%	24.1%	26.7%	27.6%	28.9%	29.3%	31.0%
	W_{loc}	58	27.6%	32.6%	32.8%	38.6%	34.5%	43.0%	34.5%	45.8%
AG full (any-rej.)	W_{fwd}	126	1.6%	3.5%	2.4%	5.3%	4.0%	7.3%	4.0%	8.5%
	W_{loc}	126	0.8%	3.7%	1.6%	6.9%	1.6%	9.6%	1.6%	12.2%

[‡] Pre-mistake window end is the annotated onset, which is not a stopping time of the monitor's filtration; the randomised inequality is not licensed on these rows.

an empirical observation whose cause is unresolved in this configuration and do not attribute it to N2. The Who&When HC and AG pre-mistake rows in the release archive fail the stopping-time condition above, so their rates lie outside the guarantee's scope; they are not evidence for or against the rule.

We therefore report randomized Ville here for completeness rather than as a recommended configuration. End-of-trace rejection adds detections but no early-stopping compute savings: the halting-cost figures in Section 5 are tied to the sequential-crossing rule and are unaffected.

F SPACING-NEUTRALITY SIMULATION

Simulation of the construction with the level distribution measured on our own calibration sample: plant a burst of length b at a uniformly random onset with turns drawn from a shifted distribution, sweep m, and match the threshold separately for each m so that empirical Type-I equals $\alpha = 0.05$. Reproduced from `manuscript/CCC_methods_sec24_rev4.pdf` p. 2 (dated 2026-08-23, the archived methods note this table was carried across from). Unlike the other simulations in the paper, the driver was not committed to the release archive; the PDF is the recoverable source, and this appendix carries a weaker provenance than the driver-backed appendices around it (`ARTIFACT_INDEX.md` notes the caveat). The simulation is the empirical companion to Proposition 3: it shows that at matched Type-I, spacing is neutral and burst length dominates detection, so the anchor grid choice is justified empirically rather than by a burst-length threshold.

	m=1	m=2	m=3	m=5	m=8	m=13	m=21
K=17, b=3	0.114	0.115	0.110	0.113	0.106	0.105	0.100
K=17, b=5	0.196	0.191	0.183	0.188	0.181	0.166	0.168
K=50, b=5	0.113	0.121	0.119	0.115	0.115	0.111	0.106
K=50, b=8	0.208	0.206	0.208	0.200	0.198	0.198	0.178

No interior optimum appears at any chain length or burst length; burst length dominates. At $K = 50$, moving $b = 5$ to $b = 8$ takes power from 0.113 to 0.208 at $m = 1$ (1.84×); by contrast, the twentyfold spacing change from $m = 1$ to $m = 21$ moves power by roughly one to three points on each row (K=17 b=3: $0.114 \to 0.100$; K=17 b=5: $0.196 \to 0.168$; K=50 b=5: $0.113 \to 0.106$;

K=50 b=8: $0.208 \to 0.178$). Burst length dominates; at matched Type-I, spacing is approximately neutral over the range swept.

G AN ADAPTIVE REFERENCE, AND WHY WE DO NOT USE IT

The raw pooled plug-in of Section 4.4 (without the bounded reference's δ-UCB) gives math held-out Type-I 17.2% against nominal 5% (unemitted diagnostic run of `code/aftraj_canonical_v4_wsr_regen.py` with the WSR inflation disabled); the cause is departure of the conditional score law from the role marginal, not marginal-estimation error. A different remedy from the one deployed in Section 4.5 is the shrinkage-toward-pooled estimator

$$\widehat{\pi}_{r,t} \;=\; \lambda_{r,t}\,\widehat{\pi}^{\mathrm{prefix}}_{r,t} + (1-\lambda_{r,t})\,\widehat{\pi}^{\mathrm{pooled}}_{r}, \qquad \lambda_{r,t} \;=\; \frac{n^{\mathrm{prefix}}_{r,t}}{n^{\mathrm{prefix}}_{r,t} + c},$$

built from each trajectory's strict prefix. Selecting $c = 1$ on math and $c \to \infty$ on coding and agentic (collapsing to the pooled plug-in outside math) gives math held-out Type-I 3.4% at 5.2% power (W_{fwd}).

The estimator is predictable, which allows it to factor out of the conditional expectation, but the null additionally requires the estimate to sit above the true normaliser and we do not establish that for a shrinkage blend. Enforcing a one-sided bound on the prefix component removes all of math's detection under both a time-uniform Hoeffding and a predictable-plug-in empirical-Bernstein construction, because role prefixes on math run two to five turns with variance near 1.5. The bounded pooled reference of Section 4.5 carries an explicit calibration guarantee under Assumption 3 (App. B.11); that guarantee is what Theorem 2 attaches to the deployment, and N2 must additionally hold on the monitored trajectory for the bound to bind (App. Q reports the observed failure on math). We report the pooled reference instead.

H BASELINE THRESHOLD-TRANSFER ILLUSTRATION

Head-to-head comparison of CCC against a per-role likelihood-ratio (SLR) baseline that uses failure labels, under CCC's Ville threshold and a PAC-fitted threshold per domain.

Table 4: Held-out Type-I (safe-only, lower is better ↓) and restricted detection (higher is better ↑; §5.4). *SLR* = per-role likelihood ratio, a common failure-label baseline. *PAC* = a fixed threshold chosen on a labelled calibration split with a distribution-free PAC coverage bound rather than the Ville $1/\alpha$ threshold. CCC (Ville) uses $1/\alpha = 20$; CCC (PAC) and SLR use PAC-fitted thresholds per domain. Reference regime: bounded (WSR UCB at $\delta = 0.005$) on all three AFTraj domains. Empirical operating characteristics, not attached to Theorem 2 (Appendix Q). Shared calibration does not imply matched achieved Type-I. Safe/fail n: math $58/58$, coding $54/37$, agentic $57/68$.

	math		coding		agentic	
	Type-I ↓	detect ↑	Type-I ↓	detect ↑	Type-I ↓	detect ↑
	bounded ref.		*bounded ref.*		*bounded ref.*	
CCC W_{loc} (Ville, $1/\alpha$)	3.4%	6.9%	0.0%	0.0%	1.8%	2.9%
CCC W_{loc} (PAC threshold)	3.4%	6.9%	0.0%	27.0%	3.5%	16.2%
SLR, per-role likelihood ratio (PAC)	1.7%	8.6%	5.6%	27.0%	7.0%	17.6%

On coding CCC-PAC matches SLR's 27.0% detection at strictly lower Type-I; on agentic the two gaps have the same sign; on math SLR is a trajectory ahead on each axis at $n = 58$, with fully overlapping Clopper–Pearson intervals. Failure labels do not deliver a decisive detection edge at any of the three achieved rates.

Threshold-portability illustration: moving a threshold between two different baselines on the same held-out sample. Numbers from the threshold-transfer table, family `baseline-PAC-swap`, off-diagonal (source $\neq$ target).

Carried one way the threshold saturates and alarms on everything; carried the other it goes inert and alarms on nothing. Neither is a failure an operator can detect from the alarm stream, which is

	math	coding	agentic
baseline threshold → another baseline (max off-diag)	100%	100%	100%
baseline threshold, other direction (min off-diag)	0%	0%	0%

what makes scale-dependence a safety property rather than an inconvenience. We report it as an illustration of scale-dependence rather than as a finding.

I FURTHER EXPERIMENTAL RESULTS

This appendix contains material moved from Section 5 verbatim: the window-aware oracle reference on Who&When HC, the halting-cost analysis against post-hoc auditors, the level-sweep robustness check, and the Who&When HC companion to Figure 1. Their headline numbers are summarised in the pointer paragraph at the end of Section 5.4.

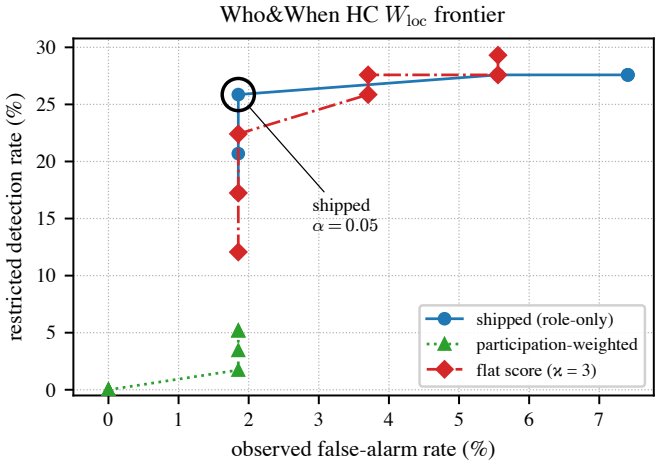


Figure 2: Who&When HC companion to Figure 1. Same axes, same α grid, three arms (state-conditioned is math-only — Stage 3 does not define an HC counterpart). Same qualitative finding as math within the same scalar-score representation: at the prespecified nominal 5% operating point no alternative exceeds shipped detection at a comparable observed false-alarm rate. The improvement is not consistent across the evaluated α-grid: at $\alpha = 0.25$ the flat-score arm ($\kappa = 3$) attains $17/58$ detection against shipped's $16/58$ at the same $3/54$ Type-I. Participation-weighted zeros the two Orchestrator roles that carry HC's signal (Appendix M's frozen E_2 selection), which is why its detection stays near zero across the alpha range. Data: `results/frontier_sweep_20260918.csv`.

I.1 HOW MUCH DETECTION IS AVAILABLE

Modest detection invites the question of whether the statistic is the constraint. As an unmatched reference, an oracle that knows the fault window and scores the same per-turn factors detects 46.6% on HC — against our 25.9% — while raising false alarms at 14.8%, three times our operating level. The oracle is not an admissible monitor under the guarantee at any $\alpha < 0.148$; its rate at unmatched false alarms is an informative reference on the same-factor pipeline, not an information-theoretic ceiling and not evidence that no alternative statistic could do better at 5% false alarms. We do not report a ratio of 25.9% to 46.6% because the ranges are not comparable.

The pattern — oracle detection at 46.6% where reweighting the rubric levels does not recover, and the domains with the shortest chains detecting least — is consistent with a hypothesis that judge-side separation between fault-free and failing turns bounds detection more than the choice of statistic does. The unmatched oracle and single-judge experiments support that hypothesis but do not establish it as a conclusion. A second judge on the same corpora (Appendix T) moves the W_{loc} HC operating point on both axes — restricted detection $15/58 \to 21/58$, pre-mistake Type-I

$1/54 \to 4/54$ — first-cut evidence that the judge matters; two operating points cannot say whether that is a change in separation or a change in threshold.

I.2 THE GUARANTEE IS NET COST-SAVING ON HC UNDER OUR COST MODEL

Cost model. Per turn t with content c_t, the trajectory input at turn t is $\mathrm{sys}_{\mathrm{ag}} + n_{\mathrm{tok}}(q) + \sum_{s<t} n_{\mathrm{tok}}(c_s)$ (agent system stub $\mathrm{sys}_{\mathrm{ag}} = 200$ tokens; q = user question), and the agent-output count is $n_{\mathrm{tok}}(c_t)$. The judge input at turn t is $\mathrm{sys}_{\mathrm{jd}} + n_{\mathrm{tok}}(q) + \sum_{s \in P_t} n_{\mathrm{tok}}(\widetilde{c}_s) + n_{\mathrm{tok}}(\widetilde{c}_t)$ with $P_t = \{\max(0, t-\texttt{MAX_PRIOR_TURNS}), \ldots, t-1\}$ with $\texttt{MAX_PRIOR_TURNS} = 6$ (path_b_judge.py), $\widetilde{c}$ truncates content to $\texttt{MAX_CONTENT_CHARS} = 1500$ characters, $\mathrm{sys}_{\mathrm{jd}} = 175$ tokens, and judge output is fixed at 100 tokens. Prices (USD per 10^6 tokens, GPT-4o family): agent 2.50 in / 10.00 out, judge 0.15 in / 0.60 out. Per-turn agent cost A_t and judge cost J_t are the sums of these token counts times the corresponding prices. For a trajectory of length K with alarm at τ (or $\tau = \infty$ if no alarm), let $A_{a,b} = \sum_{t=a}^{b-1} A_t$ and analogously $J_{a,b}$, and let C be the CHIEF analyzer cost per case (held constant across arms). The dollar cost per trajectory is:

$$\text{baseline} = A_{0,K} + C$$
$$\text{embedded (halt at } \tau\text{, no correction)} = A_{0,\tau+1} + J_{0,\tau+1} + C$$

The premature-halt-corrected embedded cost on a failing trajectory with onset u is

$$\text{embedded}^{\dagger} = \begin{cases} A_{0,K} + J_{0,K} + C & \text{if no alarm } (\tau = \infty) \\ A_{0,\tau+1} + J_{0,\tau+1} + C & \text{if } \tau \geq u \text{ (detection)} \\ A_{0,K} + J_{0,\tau+1} + C & \text{if } \tau < u \text{ (pre-onset halt: no skip credit)} \end{cases}$$

On safe trajectories a false halt at τ_{s} is treated the same way as a pre-onset halt: the operator continues past the false halt (variant A, body headline). Variant B (halt-and-rerun) adds an $A_{0,\tau_{\mathrm{s}}+1}$ waste term; both variants are reported below. CHIEF analyzer tokens per case are pulled from Wang et al. (2026a) Table 2 (§6.2): 55,085 HC, 19,504 AG, at $90/10$ input/output split at the agent-tier price. Held constant on both baseline and embedded sides so it cancels in the delta and only the agent-side savings drive the comparison.

HC headline. Placed ahead of a post-hoc auditor, the monitor halts the trajectory when it alarms and the auditor examines what exists. Under the premature-halt correction (alarms strictly before annotated onset earn no skip credit; $1/16$ HC alarms is pre-onset), we report both aggregation targets, which answer different questions:

- **Mean-of-per-trace-ratio (MOR) saving on failing HC: $\mathbf{11.2}\%$** (uncorrected 12.0%; bootstrap 95% CI $[4.8, 17.8]$, $B = 1000$ percentile, seed 20260826, from `results/compute_cost_wnw_20260909.csv`). *Answers:* on a typical failing HC trajectory, what fraction of its cost does the monitor save?
- **Ratio-of-means (ROM) saving on the failing HC corpus: $\mathbf{24.1}\%$** corrected (24.7% uncorrected). *Answers:* of the total agent-cost dollars spent on failing HC trajectories, what fraction does the monitor save in aggregate?

The ROM reading is higher than the MOR because alarming trajectories are longer than the corpus average, so their savings weight more heavily under a turn-weighted aggregation. The MOR saving against E-valuator is 13.0% under the same accounting. All stopping times come from the shipped scoring pipeline (`results/onset_delay_20260903.csv`), the same source used by Table 7 and the AFTraj summary that follows it.

Prevalence sensitivity. The corpus is failures-only, so the deployment-relevant number is net saving as a function of failure prevalence p across a mixed traffic stream. Safe trajectories add monitoring cost $J_{0,K}$ (mean \$0.0139) with no offset when the monitor stays quiet; false alarms occur on safe traces at rate $q = 1/54 \approx 1.9\%$ (HC CCC W_{loc} Ville safe-only, Table 2). Since Who&When contains no safe trajectories with recorded stopping times, we proxy $\mathbb{E}[\tau_{\mathrm{s}} \mid \text{false alarm}]$ by the mean at/after-onset τ on failing HC traces (same corpus and length distribution), giving 51.9 turns and judge cost \$0.0143.

Table 5: Net saving (%) vs failure prevalence p, under the two safe-side variants. **A (body headline).** Operator continues past a false halt on a safe trace; safe-embedded cost $= A_{0,K} + J_{0,\tau_s+1} + C$ with $\tau_s = \infty$ on no-alarm safe traces. **B (harsher, reported alongside).** Operator discards work and reruns unmonitored after a false halt; extra A_{0,τ_s+1} charged on false-alarm safe traces. Both apply the premature-halt correction on failing traces uniformly. Source: `results/cost_prevalence_sensitivity_20260908.csv`.

prevalence p	0.00	0.02	0.05	0.10	0.20	0.30	0.50	1.00
A (continues)	-0.67	-0.18	$+0.57$	$+1.81$	$+4.28$	$+6.76$	$+11.72$	$+24.11$
B (halt-and-rerun)	-2.55	-2.02	-1.22	$+0.12$	$+2.78$	$+5.45$	$+10.78$	$+24.11$

Break-even: $p^*_A = 2.7\%$, $p^*_B = 9.6\%$ (*modelled*: the safe-run cost per trajectory uses the HC pre-onset-window false-alarm rate $1/54 \approx 1.9\%$ as a proxy for the alarm rate on complete safe runs, since Who&When contains no fault-free traces; sensitivity to that proxy and to safe-run length is reported below). A and B are different operator behaviours, not a modelling choice the data settles; the body headline uses A and states the operator-continues assumption explicitly.

What drives the number. Monitoring costs about $\$0.014$ per trajectory against a mean net saving of roughly $\$0.50$ per failing HC trajectory (averaged across all 58 failing traces, including the 42 no-alarms that pay pure judge overhead and the 1 pre-onset halt that earns no skip credit; per-detection the saving is about $\$1.93$ divided across the 15 at/after-onset detections but that is not the quantity break-even uses). Break-even p^* is built on the per-failing-trace expected saving because that is what enters the mixed-traffic net-saving equation $p \cdot \mathbb{E}[\text{saving} \mid \text{failing}] - (1-p) \cdot \mathbb{E}[\text{cost} \mid \text{safe}]$. The judge-to-agent ratio is about 1:36; break-even is low because the judge (GPT-4o-mini) is cheap relative to the agent (GPT-4o). A more expensive judge or a longer judge context (larger `MAX_PRIOR_TURNS` or `MAX_CONTENT_CHARS`) would raise p^* in both variants. That is a property of the price ratio, not of the method.

Sensitivity to the safe-run proxy and safe-run length. Under variant-A accounting, the safe-side per-trajectory cost is $C_{\text{safe},A}(q) = (1-q)\,\mathbb{E}[J_{0,K} \mid \text{no alarm}] + q\,\mathbb{E}[J_{0,\tau_s+1} \mid \text{alarm}]$: the $(1-q)$ term is judge cost on quiet safe traffic — the monitor scores every turn to detect faults — and dominates because $\mathbb{E}[J_{0,K}] \approx \mathbb{E}[J_{0,\tau_s+1}] \approx \$0.014$ per trace on the shipped proxy. So $C_{\text{safe},A}$ is essentially flat in q: p^*_A swings from 2.72% at $q = 0$ to 2.79% at $q = 1$, and halving or doubling q around $1/54$ moves it by under 0.01 percentage points (`results/cost_sensitivity_recompute_20260916.txt`). The sensitivity that matters is safe-run length: $\mathbb{E}[J_{0,K}]$ scales linearly with $\bar{K}_{\text{safe}}$, so halving safe-run length moves p^*_A to 1.4% and doubling it moves it to 5.2%. The 2.7% headline is a modelled point estimate under the pre-onset-window proxy and depends principally on $\bar{K}_{\text{safe}}$, not q.

Sensitivity to price ratio and other subsets. Rescaling agent prices to hit target agent-to-judge cost ratios, the corrected failing-only mean-of-per-trace saving is 0.07%, 11.04%, 12.02% and 12.75% at ratios 1.16, 8, 17 and 100 under the uncorrected accounting of `results/compute_cost_sensitivity_20260907.csv` (the premature-halt correction applied elsewhere to the primary cost figures would lower these); the curve is nearly flat once the agent costs at least $8\times$ the judge, and stays positive down to close-to-parity price ratios. AG at shipped prices loses 1.7% under the same corrected accounting for the scope reason its detection is bounded (0 at-or-after-onset alarms in 126 trajectories). We do not report a self-monitoring-at-parity number here because it depends on the specific judge/agent pair, and the AG figure already illustrates the loss regime.

I.3 Robustness to the level

We fix $\delta = 0.005$ once and build one factor process at that δ; only the threshold $1/\alpha$ then varies across the grid. Ville's inequality therefore holds simultaneously at every α for the α-component of the guarantee, and a reader may apply whatever tolerance their deployment calls for; a PAC-fitted-threshold baseline cannot produce this curve because its threshold would require refitting at every level. The total unconditional level is $\alpha(1-\delta) + \delta$, a one-time δ paid at factor construction

rather than per-α. The cost is additive rather than proportional, so it bites hardest at tight levels: at $\alpha = 0.01$ the total is 0.015, a 50% relative loosening, against 0.055 (10%) at $\alpha = 0.05$. The restart mixture W_{loc} math sits at or below nominal α at every $\alpha \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ under the shipped Path-A fixed K'; the other AFTraj W_{loc} cells and both Who&When W_{loc} cells also stay at or below nominal from $\alpha = 0.05$ upward. At the tight-level tail $\alpha \in \{0.01, 0.02\}$ math W_{loc} sits at 3.4%, above nominal there. Appendix D reports the per-α numbers and the caveats on reading them.

J CROSS-DOMAIN OOD TRANSFER

We fit each method on a *source* AFTraj domain's development split and apply it, unchanged, to a *target* domain's held-out split. Restricted detection convention (§5.4); Ville $1/\alpha = 20$ for CCC; PAC-fitted source-domain threshold for SLR. Roles are matched across domains under the functional map of the six-bucket functional map (`solver`, `verifier`, `orchestrator`, `tool`, `environment`, `user`), so `MathSolver/CodeWriter/TaskSolver` all count as `solver` on the target and are scored under the source's `solver` reference. Numbers from the OOD transfer table. Diagonal cells (source = target) match the Path-A shipping configuration; off-diagonals were computed against the pre-Path-A source references (agentic under plug-in) and are not regenerated under bounded agentic here. The direction claim below (Type-I degrades off-diagonal, with worst cell above nominal for both methods) is a statement about mismatch magnitude that is not sensitive to that choice.

Table 6: Cross-domain OOD transfer. Off-diagonal cells apply a source-domain reference to a target-domain split. Restricted detection.

		SLR		CCC W_{loc}	
source	target	Type-I	detect	Type-I	detect
math	math	1.7%	8.6%	3.4%	6.9%
math	coding	0.0%	0.0%	0.0%	0.0%
math	agentic	0.0%	0.0%	0.0%	0.0%
coding	math	**50.0%**	51.7%	**27.6%**	43.1%
coding	coding	5.6%	27.0%	0.0%	0.0%
coding	agentic	**17.5%**	27.9%	1.8%	2.9%
agentic	math	**51.7%**	44.8%	**19.0%**	29.3%
agentic	coding	**14.8%**	13.5%	0.0%	0.0%
agentic	agentic	5.3%	17.6%	1.8%	2.9%

Both methods degrade. SLR's Type-I reaches 51.7% in the worst cell (agentic→math), exceeding its own detection rate (44.8%) there; four of six off-diagonals exceed 14%. CCC's Type-I reaches 27.6% (coding→math), also above nominal; one other off-diagonal sits above 2% (19.0%, agentic→math), one at 1.8% (coding→agentic), and three at 0.0%. Off-diagonal averages are 22.3% (SLR) versus 8.1% (CCC), and off-diagonal worst-case is 51.7% versus 27.6%. *Our guarantee does not transfer.* A reference fitted on the source domain does not dominate the target-domain per-role normaliser, so the good event of Theorem 2 is not on the target's law; the operator cannot use $1/\alpha$ as a proven bound in the transferred regime. What survives is that the loss term of Lemma 3, $\log(\widehat{S}_r/S_r)$, is still computable against the target's calibration data, so the degradation is diagnosable pre-deployment.

Name-keyed contrast, reported for honesty. Under raw role names — `MathSolver`, `CodeWriter`, `TaskSolver` etc. treated as distinct — CCC's off-diagonal Type-I is $0/n$ everywhere, because unrecognised roles fall to the $S = w_{\max}$ fallback of Remark 4 and the factor is capped at 1: the monitor cannot fire on target-only roles. That result is real but is safety-by-inertness rather than safety-with-power. The functional-role transfer above removes the fallback and reveals the actual degradation; we report it in preference to the name-keyed version.

Reconciliation with Table 4. The diagonal cells of the functional-role table above and Table 4 agree on CCC agentic Type-I at $1/57$ under Path-A. SLR agentic Type-I differs by one alarm across the two tables ($3/57$ here vs $4/57$ in Table 4), and SLR agentic detection agrees ($12/68$ on both). The Type-I difference is a consequence of the functional-role map collapsing agentic's twelve raw

roles into six buckets, changing the estimated per-role likelihood ratios; Table 4 uses raw role names throughout.

K SHORT BURSTS ARE HARD TO DETECT

Proposition 3 (Short bursts are hard to detect)**.** *Let $\kappa = w_{\max}/w_{\min}$ and fix a subset $B \subseteq \{1, \dots, T\}$ of turn indices satisfying*

(H1) *B is $\mathcal{G}_0$-measurable;*

(H2) *$\mathbb{E}[\varepsilon_t \mid \mathcal{G}_{t-1}^+] \le 1$ for every $t \notin B$;*

(H3) *$\varepsilon_t \le \kappa$ almost surely for every $t \in B$ (automatic by Proposition 2).*

Then for any $\alpha \in (0,1)$ with $\alpha\kappa^{|B|} \le 1$,

$$\Pr\left(\sup_{t \le T} M_t \ge 1/\alpha\right) \;\le\; \alpha\,\kappa^{|B|},$$

for the single anchored process $M_t = E_j(t)$ (j $\mathcal{G}_0$-measurable) and for any fixed-weight mixture $M_t = \sum_j c_j E_j(t)$ with $\mathcal{G}_0$-measurable $c_j \ge 0$, $\sum_j c_j \le 1$.

Hypothesis (H2) is Assumption 2 restricted to non-burst turns. Hypothesis (H1) fixes the burst set a priori; the intended reading is adversarial, so the bound holds for the worst $\mathcal{G}_0$-measurable choice of B. The same proof covers a $\mathcal{G}$-predictable B, while a fully data-dependent B would require a union bound over subsets that we do not attempt.

At $\alpha = 0.05$ and $\kappa = 5$ the bound reads 0.05 at $|B| = 0$, 0.25 at $|B| = 1$, and is vacuous from $|B| = 2$: the proposition binds on very short bursts and says nothing about longer ones. The bound does not depend on the anchor grid or the mixing weights. Proof in Appendix A.5.

L PER-ROLE E-POWER LOSSES AND CALIBRATION REGIME DETAILS

This appendix collects the material moved out of Section 4.5 for length: the per-turn e-power bookkeeping lemma and its expected-growth headroom diagnostic, followed by the per-role losses and shipped-regime constructions those objects motivate.

Per-turn e-power decomposition and expected-growth headroom.

Lemma 3 (Per-turn e-power decomposition under an estimated reference)**.** *Let $e_r(k) = w_k/\widehat{S}_r$ be the categorical factor of Section 3.3 with $\widehat{S}_r$ estimated on independent calibration data $\mathcal{D}$, and $e_r^*(k) = w_k/S_r$ the same factor at the true normaliser. Under Assumption 1 ($\widehat{S}_r$ is $\mathcal{G}_0$-measurable), positivity of $\widehat{S}_r, S_r$, and on the good event $A = \{\widehat{S}_r \ge S_r \ \forall r\}$, for every alternative law Q,*

$$\mathbb{E}_Q\big[\log e_r(S_t) \,\big|\, \mathcal{G}_{t-1}^+, R_t = r\big] \;=\; \mathbb{E}_Q\big[\log e_r^*(S_t) \,\big|\, \mathcal{G}_{t-1}^+, R_t = r\big] \;-\; \log(\widehat{S}_r/S_r). \tag{7}$$

Bookkeeping identity, proof in Appendix A.7: a scalar-normaliser specialisation of the log-wealth framework of Kelly (1956) and the log-optimality of the numeraire e-variable (Larsson et al., 2025), stated pointwise by neither. The per-turn loss $\log(\widehat{S}_r/S_r) \ge 0$ is calibration-computable pre-deployment; Corollary 2's headroom additionally needs the operator's per-turn e-power target.

Corollary 2 (Expected-growth diagnostic)**.** *Under the conditions of Lemma 3, define the* per-turn expected-growth headroom *of role r against an alternative Q at trajectory length K and level α as*

$$h_r(Q, K, \alpha) \;:=\; \mathbb{E}_Q\big[\log e_r^*(S_t) \,\big|\, \mathcal{G}_{t-1}^+, R_t = r\big] \;-\; \log(\widehat{S}_r/S_r) \;-\; \frac{\log(1/\alpha)}{K}. \tag{8}$$

Headroom is neither necessary nor sufficient *for detection — variance carries lucky paths across $1/\alpha$ even when $h_r < 0$, and trace-level wealth is a role-weighted sum; we report loss and hurdle as a pre-deployment diagnostic against operator-supplied Q, not as a condition that licenses detection.*

Per-role losses vs alarm hurdle. At $K = 20$, $\alpha = 0.05$ the alarm hurdle is $\log(20)/20 = 0.150$ nats/turn. Per-role losses $\log(\widehat{S}_r/\overline{w}_r)$ under the betting confidence sequence fall below that hurdle on every math role (0.053–0.073) and every coding role (0.032–0.080); on agentic under WSR widths, loss exceeds hurdle on eight of twelve roles. `Manager` (0.087), `environment` (0.115), `user` (0.021), and `Computer_terminal` (0.142) sit below on agentic; the other eight roles (`Verification_Expert` 0.164, `search_agent` 0.174, `TaskSolver` 0.170, `Statistics_Expert` 0.313, `Movie_Expert` 0.592, `Geography_Expert` 0.605, `Research_Expert` 1.199, `Sports_Expert` 1.535) sit above. The TaskSolver-heavy lost trajectories of Remark 16's table accumulate their loss on the failing roles.

Coding illustration and role-subset ablation. On coding all four roles sit below hurdle. Switching to the bounded reference costs no detection on the primary statistic W_{loc} ($0/37$ either way); two of its three secondary cells lose one alarm each (W_{fwd} and M_{dual} each drop from $1/37$ to $0/37$; full grid in Appendix C). Loss below hurdle does not imply detection — the alternative signal on this corpus is too small to accumulate above $1/\alpha$ regardless of reference. A role-subset ablation confirms the mechanism on agentic: pooling agentic's eight low-n roles into one bucket removes the tail-role clip artifact but recovers zero traces, since widths on the top-4 roles are near-invariant under the pooling.

Pooling caveat on math. Pooling is not free even where it helps: math with two roles + pooled tail lifts W_{loc} detection $13.8\% \to 17.2\%$ at Type-I moving $3.4\% \to 5.2\%$ (earlier per-trace K' configuration; shipped W_{loc} detection is $4/58$), and raises W_{fwd} Type-I from 15.5% to 17.2%, because the flatter reference costs the wide-variance `MathSolver` factor its discriminating power.

W_{loc} mixture loss reduction (empirical). The restart mixture W_{loc} dilutes the per-turn loss by 1.1–$2.5\times$ against W_{fwd} on the three tabulated lost trajectories (Remark 16), an empirical finding we do not extrapolate.

Who&When shrunk plug-in. On Who&When the corpus contains no fault-free trajectories, so the per-role reference is estimated from pre-mistake windows of failing trajectories, leave-one-out across traces, and shrunk toward the subset-pooled distribution with weight $\kappa = 5$ chosen a priori to match the backoff threshold used elsewhere in our code. Theorem 2 does not cover it, for two reasons: the estimate is a point plug-in rather than a one-sided bound, and the pool consists of prefixes assumed to satisfy N2 under the selected population's law rather than trajectories known to be fault-free. Control there is measured, as it is on agentic.

M Predictable participation weights (attempted performance intervention)

We tried an intervention on the per-turn factor that stops uninformative roles from spending wealth without adding signal, following the performance direction sketched by Remark 4. Both variants preserve false-alarm validity by construction; the empirical results were non-positive and the reason is instructive.

Construction and validity. Let $\mathcal{G}^+_{t-1}$ denote the filtration through turn $t-1$ plus turn t's role R_t. For any $\gamma_t \in [0,1]$ that is $\mathcal{G}^+_{t-1}$-measurable (a function of R_t, $\widehat{S}$, and $S_1, \ldots, S_{t-1}$ only, never of S_t), define

$$e^{\gamma}_r(t) \;:=\; \gamma_t\, e_r(t) + (1-\gamma_t). \tag{9}$$

Under any null with $\mathbb{E}[e_r(t) \mid \mathcal{G}^+_{t-1}, R_t = r] \le 1$,

$$\mathbb{E}[e^{\gamma}_r(t) \mid \mathcal{G}^+_{t-1}, R_t = r] \;=\; \gamma_t \cdot \mathbb{E}[e_r(t) \mid \cdot] + (1-\gamma_t) \;\le\; 1,$$

so every statistic in the closure of Section 3 inherits validity. The turn clock and anchor grid are unchanged: $\gamma_t = 0$ inserts a factor of 1 into the product at position t, not a hole in the sequence.

Experiment 1: neutral factor on shipped fallback roles. The shipped scoring path assigns $S_r = w_{\max}$ (AFTraj) or a point mass on the top-weight rubric cell (Who&When via the `pi_hat_shrunk`

routine in `code/who_and_when_1b.py`) when the role's dev-pool count falls to the fallback threshold, producing a per-turn factor that is ≤ 1 and can only shrink the running product. Setting $\gamma_r = 0$ on such roles removes that loss without compromising validity. On our shipped configuration every AFTraj role clears the dev threshold on math, coding, and agentic (per-role counts in `code/neutral_fallback_diagnostic.py`), so the fallback fires on zero held-out turns and the intervention is a no-op there. On Who&When the fallback fires only on `Orchestrator (termination condition)` at HC (28 post-mistake turns); those turns do not sit at alarm-crossing positions and neither Type-I nor detection changes on HC or AG.

Experiment 2: dev-selected participation policy. We enumerate a candidate role list per corpus from role function: `user` on math and coding; `user` and `Manager` on agentic; `Orchestrator (thought)` and `Orchestrator (termination condition)` on Who&When HC; no functional candidates on Who&When AG. For each candidate we compute two per-role dev statistics, $L_r^{\text{safe}} = \mathbb{E}_{\text{dev safe}}[\log e_r]$ and $L_r^{\text{fail}} = \mathbb{E}_{\text{dev fail}}[\log e_r]$. Selection rule: $\gamma_r = 0$ iff r is a candidate and $L_r^{\text{fail}} \leq 0.02$. The rule is a function of role name and frozen dev summaries only and is set before any held-out trace is scored. It selects `user` on math and coding, `user` and `Manager` on agentic, and both HC orchestrator candidates.

Evaluation. Held-out W_{loc} at $h = 1/\alpha = 20$, anchor spacing $m = 3$, Clopper–Pearson 95% intervals; the "shipped" column reports the pre-Path-A configuration (plug-in agentic reference, per-trace K'), and both columns of the table use that baseline for consistency:

corpus	Type-I (shipped $\to$ E2)	detection (shipped $\to$ E2)	Δ
AFTraj math	$2/58 \to 3/58$	$7/58 \to 8/58$	$+1$ TI, $+1$ det
AFTraj coding	$0/54 \to 0/54$	$0/37 \to 0/37$	0
AFTraj agentic	$1/57 \to 1/57$	$6/68 \to 2/68$	0 TI, -4 det
Who&When HC	$1/54 \to 0/54$	$15/58 \to 0/58$	-1 TI, -15 det
Who&When AG	$1/72 \to 1/72$	$0/126 \to 0/126$	0

The conditional-factor diagnostic moves under E2 by direct substitution: at every turn whose role is in the selection set the effective factor becomes 1; state assignment (based on the previous turn's rubric level, unaffected by γ) is unchanged, so the diagnostic-conditional subsample is fixed. The observed drop on agentic ($1.13 \to 1.05$) and Who&When HC ($1.05 \to 1.02$) reflects that on the specific "prev not top" turns the selected roles carried per-turn factor above 1 — $\mathbb{E}[e \mid \text{prev not top}]$ is 1.10 over the 78 `Manager` turns on agentic and 1.06 over the 107 `Orchestrator (thought)` turns on HC, out of 96 and 210 diagnostic-conditional turns respectively — so zeroing those roles removes their contribution to the observed excess rather than removing separately-tallied wealth-losing turns. Math and coding show no change: `user` occupies too few "prev not top" turns on those corpora to move the mean.

Reading. The empirical finding is about the selection statistic, not the two roles that drove the losses on Who&When HC and agentic. L_r^{fail} averaged to -0.005 over the 1,151 post-mistake `Orchestrator (thought)` turns on HC and to $+0.005$ over the 225 post-mistake `Manager` turns on agentic — essentially zero either way — yet individual turns within those roles carried enough factor mass to push trajectories over $h = 20$ at their peaks, while averaging across the role summed to nothing. Zeroing the role removed those peaks; the per-turn mean statistic that selected the role never saw them. On math the intervention moves both Type-I and detection by one alarm ($2/58 \to 3/58$ and $7/58 \to 8/58$), matching the §6 reading that tested corrections tend to move detection and false alarms together; `user` still occupies too few prev-not-top turns to shift the diagnostic itself. The detection loss lands on Who&When HC and agentic, where E2 had per-turn evidence to remove, and the direction reproduces the trade Appendix R's state-conditioned refit makes on math independently.

Why we did not pursue a variance-based selection rule. The E2 selection above uses only frozen development summaries ($L_r^{\text{safe}}, L_r^{\text{fail}}$) and is set before any held-out trace is scored; that is a legitimate use of the dev/held-out split, and the negative result above is a genuine test of the prespecified L_r^{fail} rule. What would be improper is observing E2 fail on held-out, proposing a variance-based selection rule as a response, and evaluating that second rule on the same held-out

signal — a decision informed by held-out outcomes without fresh evaluation. A variance-based rule is a natural reaction to the peak-driven failure this table shows, but running it here would collapse dev and held-out into a single selection stage. We record the E2 negative result and leave a variance-based selection rule to independent evaluation on a future corpus.

N DELAY, OPPORTUNITY, AND WHERE THE MISSED DETECTIONS SIT

Table 7 classifies every held-out failing Who&When trajectory as pre-onset, at-onset, post-onset, or no-alarm; the AFTraj rollup follows in the sentence after it. All numbers use the shipped W_{loc} + Ville $1/\alpha = 20$. Onset uses the annotated `mistake_step`; AFTraj records it 1-indexed but semantics differ by source — injected onsets mark where an incorrect step was inserted, diagnosed onsets mark where an oracle diagnosed failure. 4 HC (3 out-of-bounds annotations, 1 human-role turn) and 20 AG (plan-level with no specific turn) trajectories have missing or ambiguous onset and cannot be classified pre/at/post; we fold them into the "no alarm" bucket of Table 7 so its denominator matches Table 2's 58 and 126. Pre-onset Type-I denominators (54 HC, 72 AG) further require a nonempty pre-mistake window (mistake index ≥ 1 in the scored sequence): the same 4 HC traces excluded above and the same 20 AG plan-level traces, plus 34 additional AG traces whose annotated mistake falls at the first scored turn.

Table 7: Alarm-vs-onset classification, Who&When only. Delay and σ/K are medians (IQR) over at-or-after-onset alarms only. AFTraj row-level detail dropped for space; per-domain summary in the sentence below.

subset	n	pre	at	post	no alarm	delay (med, IQR)	σ/K med
Who&When HC	58	1	1	14	42	31.0 [15.5, 43.0]	0.70
Who&When AG	126	1	0	0	125	—	—

On AFTraj under Path-A, restricted detection on the primary statistic is sparse enough that per-bin opportunity slicing is uninformative: math 1 pre / 0 at / 4 post / 53 no-alarm (median post-onset delay 7.5 turns); agentic 0 pre / 0 at / 2 post / 66 no-alarm (median delay 8.5); coding 0/37 across every opportunity bin including $[7, 15)$, so coding's zero detection is a per-turn evidence problem rather than an opportunity one. Full per-trajectory classification in `results/kprime_regen_all_20260918.csv`.

O THRESHOLD-SWAP ILLUSTRATION

Full presentation of what Section 5.3 summarises in one sentence. First, fitted thresholds do not agree with one another across nine cells (span 2.44–52.89) despite $1/\alpha = 20$ being a single number for all nine; a threshold is a functional of the whole score path and is therefore scale-dependent by construction. Second, they do not agree with the threshold the theorem supplies, and not in a consistent direction: on coding the fitted values fall well below $1/\alpha$ so the rule alarms more readily, on math they land at or above so the PAC-fitted rule is more conservative. A practitioner fitting a threshold has no way to know in advance which case they are in — it is a property of the calibration sample rather than of the method. Holding the statistic fixed at W_{loc} under the restricted convention, the fitted threshold raises coding $0.0 \rightarrow 27.0\%$ and agentic $2.9 \rightarrow 16.2\%$; on math it changes nothing (both thresholds give 6.9%).

Table 8: Held-out false alarms as a threshold moves. Row 1: across-statistic range for the CCC construction at fixed $1/\alpha = 20$ (math, coding, and agentic all under the bounded WSR reference). Row 2: worst-case off-diagonal in the CCC-PAC threshold swap. Baseline threshold-transfer illustrations in Appendix H.

	math	coding	agentic
CCC, $1/\alpha = 20$ fixed, no refit	3.4–15.5%	0.0%	0.0–1.8%
CCC threshold swapped between statistics	15.5%	5.6%	10.5%

The first row is the licensed use. The second is deliberate misuse of our own construction — fit a threshold on one statistic, apply it to another — which degrades but stays bounded at 15.5% in the

worst case. Moving a threshold between two different baselines is more dramatic and less informative: it saturates at 100% or goes inert at 0% depending on direction, because the two statistics live on scales an order of magnitude apart. Appendix H reports it as an illustration of scale-dependence rather than as a finding.

P AFTRAJ PARENT–CHILD LEAKAGE AUDIT

AFTraj's 728 injected failing trajectories are each derived from a verified outcome-safe parent, so the split matters: a parent in dev with its child in held-out would let the child's pre-mistake turns shape the reference it is scored against. We match every injected child to its parent by $(\text{task}, \text{gold answer})$ with a turn-prefix tiebreak, recovering 725 pairs (333/333 math, 244/247 coding, 148/148 agentic), 703 with the pre-mistake prefix byte-identical to the parent. Zero pairs land with parent in dev and child in held-out; the released split is pair-aware.

A second, weaker fact: 106 of the 163 held-out failing trajectories share a byte-identical pre-mistake prefix with a held-out safe-labeled trajectory under a different `conv_id` (100% coding, 84% math, 29% agentic), so the Type-I and detection denominators are not independent trajectory populations on their pre-mistake portions. Restricting the held-out safe-labeled pool to non-parent trajectories (9, 17, 37 by domain) gives Type-I of $0/9$, $0/17$, and $1/37$ (Ville) or $2/37$ (PAC); every restricted rate lies inside the full-set Wilson 95% interval (`results/aftraj_leakage_audit_summary_20260908.csv`).

Q CONDITIONAL VALIDITY: MECHANISM AND RESIDUAL

Section 6 states the guarantee's premises: both levels of Theorem 2 presuppose N2 on the deployment population, and averaging over calibration draws pays δ for a bad calibration draw of $\widehat{S}_r$, not for a violated conditional-moment premise. The upper confidence bound does not itself establish N2, and Section 5.2's math W_{fwd} elevation ($9/58 = 15.5\%$ at nominal 5%; $p = 0.0022$ uncorrected, BH $q = 0.020$ across nine cells) invites the direct check below. Appendix R then tests a state-conditioned refit on math.

Mechanism. On the held-out math safe pool (998 turns across 58 trajectories), $\mathbb{E}[\varepsilon_{R_t}(S_t) \mid \text{prev turn not L4-L5}] = \mathbf{1.346}$ (cluster bootstrap, $B = 10{,}000$ replicates resampling trajectories with replacement, BCa 95% CI $[\mathbf{1.17},\ \mathbf{1.51}]$; $n = 196$ turns from 55 of 58 trajectories; zero empty replicates). The CI excludes 1, so the conditional-mean violation survives the within-trajectory dependence the point estimate ignores. Early positions give $\mathbb{E}[\varepsilon] = 1.078$ and long trajectories ($K \geq 24$) give 1.118 (point estimates; the per-position and per-K subsets are smaller, so their CIs would be correspondingly wider). *Held-out role marginals* are 0.86–1.01, with the upper bound coming from math's `user` role ($n_{\text{ho}} = 58$, mean 3.17 against $\widehat{S} = 3.13 \Rightarrow 1.014$); these are a diagnostic on observed data, not Null A's sampling distribution (see *Simulation* below).

Simulation. *Null A* draws rubric levels iid$\mid R_t$ from the *dev* role marginals — the pool the reference was fit on — so that on this sampling distribution the per-role expected factor $\overline{w}_r^{\text{dev}}/\widehat{S}_r$ is bounded by 0.98 (max: math's `user`, 0.9793; other math roles ≤ 0.95; every coding and agentic role ≤ 0.98). N2 holds by construction, and Null A alarms at W_{fwd} Type-I 0.80% (Wilson $[0.70, 0.91]$; 500 replicates $\times$ 58 structures), which is a Ville-conservative statement under the assumption. Drawing the null from the held-out role marginals instead would put math's `user` at expected factor $1.01 > 1$ and Null A would no longer be a null; we keep the dev-drawn null because that is the one that satisfies N2. A matched-marginal Null B with trajectory-level log-linear tilt $z_j \sim \mathcal{N}(0, \sigma)$ reaches 13.2% at $\sigma = 0.40$ and 18.0% at $\sigma = 0.50$; the tuning that produces the observed 15.5% is $\sigma \approx 0.45$, which shows the mechanism *can* generate that level of inflation, not that it explains it quantitatively.

Residual. The empirical trajectory tilt on dev math (variance ratio 1.88 vs iid $\mid R_t$) gives $\sigma \approx 0.22$ under a tilt-model estimator; empirical lag-1 rubric autocorrelation is $\rho_1 = 0.25$ overall (0.44 on Verifier). Combining both at empirical parameters yields 3.83% W_{fwd} Type-I — $\sim$3.8 of the 15.5 observed points. The remaining $\sim$11.7 points are not explained by this analysis. We report this as evidence of conditional misspecification with substantial residual inflation unexplained, not as a quantitative account. Candidates include tilt-model misspecification (heavier-tailed heterogene-

ity than log-linear), longer-range dependence than lag-1, and heavy-tailed factor values that our marginal parametrisation cannot reproduce. We do not pin down which.

Symmetric annotation of 60 safe traces (2026-08-25). A single-annotator blind read of 60 held-out safe trajectories labelled each trace for *process_fault_present* (a semantic fault visible in the turns even though the outcome is correct) and *score_pattern* (`clean`, `localized_dip`, `mixed`, `sustained_low`). The sample is stratified. The math stratum (30 traces) comprises 26 traces drawn uniformly at random from the held-out safe pool with four forced traces excluded (`random_math`), plus two prespecified forced substrata of 2 traces each — one the residual math trajectories that alarm under W_{fwd} at adaptive $c = 1$, the other a matched non-alarming control set. Per the labels file's `case_description_disclaimer_20260825` the forced substrata are pre-registered as case descriptions only; with $n = 2$ per side they cannot support a rate estimate and are excluded from the rates below. Coding and agentic (15 each) are uniform-random with no forcing. Corpus-wide: $5/60$ traces carry a visible process fault (8.3%; Wilson 95% $[3.6\%, 18.1\%]$); score patterns are 52 clean, 3 localized dip, 4 mixed, 1 sustained-low. By estimable substratum: `random_math` is $1/26 = 3.85\%$ ($[0.7\%, 18.9\%]$); coding is $0/15$ ($[0\%, 20.4\%]$); agentic is $4/15 = 26.7\%$ ($[10.9\%, 52.0\%]$). All five fault carriers are annotated as recovering before the final answer (four `benchmark_plausibly_missed`, one `benchmark_should_have_caught`), consistent with the outcome-safe benchmark labels. Cohen's κ was not computed; a second annotator was not run. Two observations. (i) The math point estimate (3.85%) is far below the 15.5% W_{fwd} math elevation, but on $n = 26$ with a single event the interval $[0.7\%, 18.9\%]$ is wide enough to be consistent with nearly any population rate in that range, so it does not bear on the question in either direction; the annotation is underpowered to distinguish an outcome-versus-process explanation from its absence, and 26 traces with one event cannot separate the two hypotheses. (ii) The agentic point estimate is the highest of the three and its interval sits above the math one, yet the paper treats agentic as a domain where W_{fwd} safe Type-I is at or below nominal on outcome-labelled data — if the population process-fault rate is anywhere near the agentic point estimate, the outcome-safe versus process-safe divergence would be larger on agentic than on math and would not be visible in the outcome-labelled detection numbers reported elsewhere in the paper. Sealed labels at `code/aftraj_sym_annotation_labels_20260825.json`; blinding key at `code/aftraj_sym_worksheet_key_20260825.json`.

Cross-domain. The prev-turn conditional mean is $\mathbf{1.346}$ on math (cluster BCa 95% CI $[1.17, 1.51]$; $n = 196$ turns from 55 of 58 trajectories), $\mathbf{1.133}$ on agentic (BCa CI $[0.9989, 1.32]$; $n = 96$ from 57), and $\mathbf{1.028}$ on coding (BCa CI $[0.92, 1.19]$; $n = 62$ from 46 of 54). Only math's CI excludes 1; agentic's and coding's CIs both cover 1 (agentic's lower bound is 0.9989). The diagnostic is robustly detected only on math. Agentic and coding point estimates sit above 1 but the intervals do not exclude it, so the data does not distinguish mild conditional misspecification from the assumption holding on those domains at these sample sizes. Math is where the mechanism would compound if present, for reasons we can name: it has both the longest and most variable chains and the strongest within-trajectory role autocorrelation (Verifier $\rho_1 = 0.44$); coding's trajectories are of fixed length ($K = 11$) and its rubric distribution is concentrated at the top level, features that would either satisfy the assumption or bound how much any violation could accumulate. We do not claim more than that structural difference from these data.

On W_{loc}. The mixture reports Type-I $2/58 = 3.4\%$ on math. The same per-turn factor sequence drives both statistics; the mixture's lower rate reflects damping of persistent streaks by the anchored restart, not absence of the conditional violation. W_{loc}'s observed rate stays below nominal under this dependence; we do not treat that as validation of Equation 1 for the whole family.

Full audit per trace, factor-mean diagnostic across all conditioning cells, empirical lag-1 estimates by role, and the null-A / null-B simulation output at 500 replicates each: reproduced by `code/cond_validity_wsr.py`. (An earlier CSV `cond_validity_audit_math_wfwd_20260909.csv` carries the pre-WSR MP figures ($8/58$ alarms rather than $9/58$); the WSR figures in this appendix are computed by the driver from the shipped W_{fwd} alarms directly and do not read from that CSV.)

R STATE-CONDITIONED REFERENCE: EXPLORATORY MATH REFIT WITH INDEPENDENT CHECKS

The diagnostic in Appendix Q identifies a named conditioning cell — turns following a low-scored turn — where the role-only reference undershoots. The direction of that undershoot suggests a state-conditioned reference $\widehat{S}(R_t, C_t)$ using a predictable state C_t computed from the prefix through turn $t-1$. We report the fit and its held-out evaluation on math (Stages 2 and 3), a matched-inflation conservatism control on the same math held-out (Stage 4 v2), a frozen-map replication on agentic held-out (Stage 5), and a dev out-of-fold three-arm comparison (Stage 6). Stages 5 and 6 are the independent checks; Stages 2–4 share the math held-out that motivated the intervention and are exploratory on that data (see chronology below).

Prespecification. State $C_t \in \{\text{top}, \text{not_top}, \text{first}\}$ from turn $t-1$'s rubric level: top if $S_{t-1} \in \{0.8, 0.999\}$, not_top if $S_{t-1} \in \{0.001, 0.2, 0.5\}$, first if $t = 1$. Budget $\delta_{\text{cell}} = 0.005/(R \cdot |\mathcal{C}|) = 0.005/12$. Thin-cell fallback $n_{\min} = 30$: if a cell has fewer than 30 dev observations, the cell inherits the shipped role-only $\widehat{S}_r$. State map, δ split, and fallback were fixed before Stage 2 measured the fit and before Stage 3 evaluated on held-out.

Chronology and scope of "prespecified". The top-versus-not_top split above was chosen after Appendix Q's conditional-factor diagnostic on math held-out reported $\mathbb{E}[\varepsilon \mid \text{prev not top}] = 1.35$ under the role-only reference: that diagnostic named the cell the split then isolates. The prespecification paragraph describes what was fixed before Stages 2–3 ran, not before the data that motivated the split was inspected. Stage 3's held-out evaluation on math is exploratory in that sense — the split boundary is informed by the very diagnostic Stage 3 corrects, and Stage 4 v2's conservatism control (below) also runs on the same held-out trajectories. The independent evidence in this appendix is Stage 5's replication on agentic held-out with the math state map *frozen* out of sample and Stage 6's dev out-of-fold three-arm comparison; neither depends on the held-out math trajectories the map was designed around.

Arm naming. Throughout this appendix we compare two matched uniform-inflation controls to the state-conditioned refit (arm 3). *Arm 1* is the shipped role-only reference. *Arm 2a* is arm 1 scaled by a constant chosen to match state conditioning's average change in log per-turn factors on the dev pool. *Arm 2b* is arm 1 scaled by a per-statistic constant chosen to match state conditioning's dev false-alarm operating point on that statistic. Arms 2a and 2b answer different questions and appear in the Stage 4 v2, Stage 5, and Stage 6 paragraphs below under these fixed labels.

Conditional-factor diagnostic: interval now covers 1. Under the state-conditioned reference on math held-out safe, $\mathbb{E}[\varepsilon \mid \text{prev turn not top}] = 0.99$, cluster BCa 95% CI $[0.87, 1.11]$ on 55 contributing trajectories over 196 qualifying turns — versus the shipped role-only 1.35 with BCa CI $[1.17, 1.51]$ that motivated the intervention. The new interval covers 1 where the old one excluded it. Coverage of 1 does not by itself establish that the population conditional mean is at most 1, and — as the conservatism control below reports — does not by itself distinguish state structure from raw conservatism: arm 2b matched on either mixture statistic (uniform inflation $c = 1.20$ or $c = 1.19$) also yields point estimate $\approx$1.12 with BCa CI covering 1. Arm 3's 0.99 against arm 2b's 1.12 is a difference in degree, not in kind on this held-out data.

False-alarm inflation reduced (held-out safe, $n = 58$, Clopper–Pearson 95% intervals):

statistic	shipped role-only	state-conditioned
W_{fwd}	$9/58 = 15.5\%$ $[7.3, 27.4]$	$\mathbf{2/58 = 3.4\%}$ $[0.4, 11.9]$
W_{loc}	$2/58 = 3.4\%$ $[0.4, 11.9]$	$2/58 = 3.4\%$ $[0.4, 11.9]$
M_{dual}	$5/58 = 8.6\%$ $[2.9, 19.0]$	$\mathbf{2/58 = 3.4\%}$ $[0.4, 11.9]$

W_{fwd} point estimate drops from above nominal (BH-surviving under the shipped reference) to 3.4%, though the Clopper–Pearson upper bound extends to 11.9%; M_{dual} point estimate drops from above nominal to at nominal. These are reduced observed inflations, not established validity: on $n = 58$ the intervals cannot separate an at-nominal from a moderately elevated rate.

Detection on math falls sharply (restricted convention, held-out unsafe, $n = 58$):

statistic	shipped role-only	state-conditioned
W_{fwd}	$10/58 = 17.2\%$ $[8.6, 29.4]$	$\mathbf{2/58 = 3.4\%}$ $[0.4, 11.9]$
W_{loc}	$4/58 = 6.9\%$ $[1.9, 16.7]$	$\mathbf{0/58 = 0.0\%}$ $[0.0, 6.2]$
M_{dual}	$10/58 = 17.2\%$ $[8.6, 29.4]$	$\mathbf{0/58 = 0.0\%}$ $[0.0, 6.2]$

W_{loc} falls from 6.9% to 0.0% ($4/58 \to 0/58$) and M_{dual} falls from 17.2% to 0.0% ($10/58 \to 0/58$). W_{fwd} retains 3.4 points.

Conservatism control: state versus a matched inflation. Appendix Q's refit changes conditional-mean estimates, cell sizes, and the per-cell δ-budget together, so the detection collapse in the previous table could reflect either state-specific tightening or plain raising of the reference. We separate them with a three-arm control on the same held-out trajectories, using arm 2b as defined above (matching constants computed under 5-fold trajectory-level cross-fitting to remove in-sample bias: $c_b^{\text{CF}} = 1.14, 1.20, 1.17$ for $W_{\text{fwd}}, W_{\text{loc}}, M_{\text{dual}}$ respectively; matching constants, design check, and full four-arm output at `results/stage4_conservatism_control_v2_math_heldout_20260917.txt`, computed under per-trace K'). Arm 2b and arm 3 had zero discordant held-out trajectories on both mixture statistics under per-trace K': they removed exactly the same 7 trajectories on W_{loc} and the same 10 on M_{dual} (McNemar $b = c = 0$). Under Path-A fixed $K' = 12$ the arm 3 removal counts drop to 4 (W_{loc}) and 10 (M_{dual}); we do not re-run the arm 2b match under Path-A so the $b = c = 0$ headline is stated as a per-trace-K' result rather than a fixed-K' one. The observed mixture-statistic detection collapse is reproducible on this data by a suitably-scaled role-only reference; the identical alarm sets establish sample-level agreement but do not identify the mechanism behind the collapse.

An initial version of this control used a single W_{fwd}-matched constant applied to all three statistics (`stage4_conservatism_control_math_heldout_20260915.txt`, marked superseded in `ARTIFACT_INDEX.md`); per-statistic matching under cross-fitting changed the mixture-statistic conclusion from a small residual state effect to zero discordant pairs.

Stratified check: change concentrates on the high-not_top stratum. Pooled held-out trajectories binned by fraction of not_top turns into tertiles (cutpoints 0.150 and 0.333); columns are per-trace K'. The false-alarm reduction and the detection change both concentrate on the high-not_top stratum; the low-not_top stratum shows no change either way. This localises the effect to the cells the state map isolates. M_{dual} is absent from the table: the shipped stratified artifact (`stage3_state_cond_eval_math_heldout_20260915.csv`) predates the M_{dual} correction in App V item 6 (its M_{dual} column was computed as $\frac{1}{2}(W_{\text{fwd}}+W_{\text{rev}})$), and the 2026-09-17 recompute carries aggregates only.

stratum	n_{safe}	W_{fwd} Type-I (role→state)	n_{unsafe}	W_{fwd} det.	W_{loc} det.
low (< 0.15)	29	$0 \to 0$	9	$0 \to 0$	$0 \to 0$
mid (0.15–0.33)	21	$4 \to 0$	17	$0 \to 0$	$0 \to 0$
high (> 0.33)	8	$5 \to 2$	32	$10 \to 2$	$7 \to 0$

Residual violation. In-sample check on held-out safe: per (role, not_top) cell, $\mathbb{E}[\varepsilon \mid R{=}r, C{=}\text{not_top}]$ under the state-conditioned reference.

cell	shipped role-only	state-conditioned
MathSolver × not_top	1.23	$\mathbf{0.99}$
Verifier × not_top	1.36	$\mathbf{0.79}$
environment × not_top	1.60 ($[1.23, 1.93]$)	$\mathbf{1.15}$ ($[0.89, 1.39]$, covers 1)

Two of three cells now sit at or below 1 in point estimate on held-out; environment × not_top has point estimate 1.15 (cluster BCa 95% CI $[0.89, 1.39]$; $B = 10{,}000$ replicates, seed matched to the app-P bootstrap; 24 contributing trajectories over 48 qualifying turns;

`results/stage3_env_nottop_bca_math_heldout_20260915.txt`). The environment interval covers 1, so on the two-level state map on this data the diagnostic no longer excludes N2 in any cell; that is a reduction in observed inflation, not evidence that N2 holds. Any tuning of the state map after seeing the environment cell would have been selection on the outcome, so we report the residual as observed.

What the corrected control shows. State conditioning does one thing on this data that a matched uniform inflation cannot reproduce: it delivers a dev Type-I reduction disproportionate to its average magnitude. Under 5-fold trajectory-level cross-fitting, so that every evaluated trajectory is excluded from the reference scoring it (seed matched to the arm 2b target; `results/stage6_oof_comparison_math_20260916.txt`), arm 2a (role-only reference scaled per fold by $\exp(\bar{\Delta}_k)$; across folds mean $\bar{\Delta} = +0.0135$, sd 0.0019, mean $c_a = 1.014$) has out-of-fold Type-I $0.059/0.044/0.053$ on $W_{\text{fwd}}/W_{\text{loc}}/M_{\text{dual}}$; arm 3 has $0.024/0.012/0.018$; ratios $2.5\times$ / $3.8\times$ / $3.0\times$. Cross-fitting excludes each trajectory from its own reference but the fitted references still share training trajectories, so the exact-binomial reading of a McNemar p-value on per-trajectory alarm indicators is not automatically justified. A trajectory-cluster bootstrap-with-refit ($B = 1000$, base seed 20260917; per replicate resamples the 338 dev trajectories with replacement and runs the full $K = 5$ OOF procedure; `code/stage6_bootstrap_refit.py`, results at `stage6_bootstrap_refit_math_20260917.txt`) gives percentile 95% CIs on the paired Type-I difference (arm 2a − arm 3) that exclude zero on every statistic: $[+0.0148, +0.0562]$ on W_{fwd}, $[+0.0148, +0.0473]$ on W_{loc}, $[+0.0178, +0.0533]$ on M_{dual}; $1000/1000$ replicates place arm 2a strictly above arm 3. Discordance counts on the shipped sample are $b = 12, 11, 12$ against $c = 0$; no bootstrap replicate produces $c > 0$ on any statistic. Bootstrap degeneracy is nil on math: no replicate loses a role from a fold's training set, and no (r, c) cell falls back in a resample at fold k where the shipped Stage 6 at fold k did fit. In-sample rates were $0.065/0.047/0.053$ vs $0.033/0.012/0.018$ ($2\times/4\times/3\times$). The out-of-fold W_{fwd} rate for arm 3 sits slightly below its in-sample value rather than above. The mechanism is counterintuitive: a smaller per-fold training pool widens the WSR bound and raises $\widehat{S}(r, c)$, dropping marginal running-product trajectories rather than lifting them; the mixture statistics are dominated by well-above-threshold alarms and are unaffected. That gap survives the control; we refer to this dev-pool gap between arm 2a and arm 3 at matched average log-inflation as Claim A, and test whether it holds on an independent corpus in the frozen-map replication below. What state conditioning does not do on this held-out sample is remove any additional trajectory beyond a matched uniform inflation on the mixture statistics. Whether a state-conditioned reference with a richer state map, or a matched-inflation control under a different judge (Appendix T runs a second judge but not this specific control under it), would separate more clearly from a matched uniform inflation is an open question this experiment cannot settle.

Frozen-map replication. With the state map from the math prespec frozen out of sample and the reference $\widehat{S}(r, c)$ refit on the replication corpus's own dev pool — the split that distinguishes a replication from a further refit — the direction of Claim A replicates on agentic on W_{fwd} and M_{dual}, is untestable on W_{loc}, and the magnitude weakens. The average log-inflation on agentic dev is $\bar{\Delta}_{\text{agentic}} = -0.0033$ across folds under the pre-Path-A plug-in reference (essentially zero), so arm 2a is indistinguishable from arm 1 in that setting. Under the same 5-fold OOF protocol as the math control, arm 3's dev Type-I is $1.71\times$ / $1.00\times$ / $1.50\times$ lower than arm 2a's on $W_{\text{fwd}}/W_{\text{loc}}/M_{\text{dual}}$ over the full 348-trajectory dev pool (in-sample was $2.00\times$ / $1.20\times$ / $1.80\times$; `results/stage6_oof_comparison_agentic_20260916.txt` plus the fixed-K' recompute at `code/kprime_stages_456_math.py`). Shipped-sample discordance on the OOF W_{loc} comparison is $b = c = 0$: arm 2a and arm 3 alarm on exactly the same six dev trajectories, so no paired McNemar test on W_{loc} can move against the null direction — the state effect on this statistic is not distinguishable from zero on this corpus. Under the trajectory-cluster bootstrap-with-refit on agentic ($B = 1000$; `stage6_bootstrap_refit_agentic_20260917.txt`), percentile 95% CIs on the paired Type-I difference are $[-0.0001, +0.0259]$, $[+0.0000, +0.0201]$, and $[+0.0000, +0.0201]$ on W_{fwd}, W_{loc}, and M_{dual} respectively: the lower bound touches zero on every statistic. Direction survives one-sided under bootstrap resampling on W_{fwd} and M_{dual} (90% and 84% of replicates place arm 2a strictly above arm 3, shipped-sample discordance $b = 5, 3$ against $c = 0$), but on W_{loc} the 78% figure sits over a $b = c = 0$ shipped sample, so it reads as bootstrap variation around a null rather than direction survival. Two readings are compatible with a CI that touches zero: either the state-conditioning effect on agentic really is close to zero on this corpus, or the

state-conditioned construction is unstable under trajectory resampling on a corpus where only 11–13 of 36 role×state cells clear $n_{\min} = 30$ on the shipped run and resampling pushes marginal cells into fallback. Our own bootstrap points at the second: $956/1000$ replicates have at least one (r, c) cell that falls back in the resample at fold k where the shipped Stage 6 at fold k did fit, for 4,351 extra-fallback events out of a possible 180,000 (2.4% of cell-fold slots), so bootstrap replicates score arm 3 under a systematically less flexible reference than the shipped run. Role-missing folds are nil. Our bootstrap favours the instability reading over a small-effect reading, but the two are compatible: a 2.4% cell-fold fallback rate is consistent with instability and does not exclude a genuinely small state-conditioning effect on agentic. Separating them would need a corpus where more of agentic's role×state cells clear $n_{\min}$. On the primary statistic W_{loc} the state effect is untestable here. On agentic, arm 3 differs from arm 1 in only three of twelve `not_top` cells ($n_{\text{dev}} = 481, 56, 56$); the remaining nine fall back to the role-only $\widehat{S}_r$ (six at $n \in \{1, 3, 5, 6, 6, 6\}$, three empty; `results/stage5_replication_agentic_design_20260915.txt`), so arm 3 equals arm 1 there by construction and any state-conditioning effect on agentic is measured over three cells.

Frozen-map replication under Path-A bounded agentic. Under Path-A the shipped agentic reference is bounded WSR rather than plug-in, and the dev log-inflation direction flips: $\bar{\Delta}_{\text{agentic}} = +0.0406$ across the same 5 folds with $c_a = 1.041$. The bounded arm 1 has held-out W_{loc} detection $2/68$, arm 3 (state-cond bounded) has $0/68$; arm 3 removes two trajectories that arm 1 alarms on. On W_{fwd} and M_{dual} arm 3 removes further alarms relative to arm 1 ($4/68 \to 1/68$ and $2/68 \to 1/68$ on the same held-out unsafe pool); these are detection numbers on unsafe trajectories and do not speak to Claim A. We do not re-run the OOF Type-I comparison or the bootstrap-with-refit under bounded agentic, so Claim A — a dev-pool false-alarm gap at matched log-inflation — is untested at this Path-A configuration. The frozen-map replication result on the primary statistic under Path-A is therefore *arm 3 removes two agentic detections that arm 1 keeps*, without a paired Type-I comparison to attach it to.

Coding at its shipped $\alpha = 0.05$ produces $0/307$ dev alarms under every arm on every statistic (`results/stage5_replication_coding_design_20260915.txt`); with no signal to compare, Claim A is not testable on coding at this operating point, and we report the null-signal case rather than a refutation. Coding's conditional-factor diagnostic under arm 1 is $1.03\ [0.92, 1.19]$ (covers 1), so the diagnostic identifies no assumption failure to fix; applying the state-conditioned correction there yields arm $3 = 0.86\ [0.76, 0.98]$. A point estimate below 1 is compatible with the null-side bound $\mathbb{E}[\varepsilon] \leq 1$; the correction consumes null-side headroom rather than removing an observed excess.

S JUDGE PROMPTS, VERBATIM

The judge is a single `gpt-4o-mini` call at `temperature=0` with `response_format=json_object`, per turn. The system prompt is fixed across corpora; the user message is templated on {task, prior context, current agent's response}. The corpus gold answer is not part of the prompt: an earlier revision inserted a "Correct answer: …" line into the user message when a gold answer was passed, which we retired on 2026-08-17 and re-scored the Who&When deployment corpus under the corrected prompt. The parameter that enabled the insertion has since been removed from `code/path_b_judge.py`. Truncation caps are `MAX_CONTENT_CHARS=1500` on each turn's content (both the current turn and each prior turn) and `MAX_PRIOR_TURNS=6` on the depth of prior context.

The AFTraj calibration and evaluation phat caches (`code/aftraj_phats_{math,coding,agentic}_{safe,unsafe}_withthought.json`) are the output of the same `code/path_b_judge.py` pipeline after the gold-answer parameter was removed; the prompt shown below is the one every cached AFTraj score was produced under, and no cached AFTraj score originated from a prompt that included a "Correct answer: …" line.

System prompt (shared by all corpora, verbatim).

```
You are an expert evaluator of multi-agent AI pipeline responses.
```

```
You are given a task, prior context, and one agent's response.  Your
job is to score whether this response is a correct step that moves
the pipeline toward the right answer.
Follow this procedure before scoring:
1.  In one short sentence, state what THIS agent's turn is trying to
do (search, plan, dispatch, compute, produce final answer, etc.).
2.  In one short sentence, judge whether it succeeds at that
sub-goal, given prior context.  A search/plan/dispatch turn should
be scored on whether it is a correct step, NOT on whether it
produces a final answer.
3.  Assign a score from the rubric.
Scoring rubric:
- 1.0:  Correct step; factually right and clearly useful for
advancing the task
- 0.8:  Mostly correct step with minor errors or unnecessary content
- 0.5:  Reasonable attempt but makes a factual or reasoning mistake
- 0.2:  Substantially wrong or contradicts prior evidence
- 0.0:  Completely wrong, refuses the task, or derails the pipeline
Use the full range.  Do not cluster scores near 0.8.
Return JSON only:  {"purpose":  "<one sentence>", "assessment":
"<one sentence>", "score":  <float 0.0-1.0>}
```

User message, one real turn per corpus. Each shown as sent, with content truncated to 1500 characters as noted (elided beyond that with "...").

Who&When HC (trace 0, turn 0).

```
Task:  I read a paper about multiwavelength observations of fast
radio bursts back in March 2021 on Arxiv, and it had a fascinating
diagram of an X-ray time profile.  There was a similar burst-1
diagram in another paper from one of the same authors about
fast radio bursts back in July 2020, but I can't recall what the
difference in seconds in the measured time span was.  How many more
seconds did one measure than the other?  Just give the number.
Prior context:
(none)
Current agent (Orchestrator (thought)) response:
Initial plan:
We are working to address the following user request:  I read
a paper about multiwavelength observations of fast radio bursts
...[truncated to 1500 characters]
Score this response.
```

Who&When AG (trace 0, turn 0).

```
Task:  How much did I save by purchasing a season pass instead
of daily tickets for California's Great America in San Jose, if I
planned to visit once a month in June, July, August, and September
during the summer of 2024?
Prior context:
(none)
Current agent (ProblemSolving_Expert) response:
You are given:  (1) a task and advises from your manager with a
specific plan and (2) a general task.  Collect information from the
general task, follow the suggestions from manager to solve the task.
...[truncated to 1500 characters]
Score this response.
```

AFTraj math (trace 0, turn 0).

```
Task:  Convert the point $(0,3)$ in rectangular coordinates to polar
coordinates.  Enter your answer in the form $(r,\theta),$ where $r >
0$ and $0 \le \theta < 2\pi.$
```

```
Prior context:
(none)
Current agent (user) response:
CONTENT: Solve this math problem step by step using the compute
tool.
Problem:  Convert the point $(0,3)$ in rectangular coordinates to
polar coordinates.  Enter your answer in the form $(r,\theta),$
where $r > 0$ and $0 \le \theta < 2\pi.$
Score this response.
```

AFTraj coding (trace 0, turn 0).

```
Task:  from typing import List
def has_close_elements(numbers:  List[float], threshold:  float) ->
bool:  """ Check if in given list of numbers, are any two numbers
closer to each other than given threshold.
Prior context:
(none)
Current agent (user) response:
CONTENT: Implement the following Python function.  Test it using the
run_tests tool.  ...[truncated to 1500 characters]
Score this response.
```

AFTraj agentic (trace 0, turn 0).

```
Task:  Were Scott Derrickson and Ed Wood of the same nationality?
Prior context:
(none)
Current agent (user) response:
CONTENT: Were Scott Derrickson and Ed Wood of the same nationality?
Score this response.
```

The user message template is `"Task: {q}\n\nPrior context:\n{prior}\n\nCurrent agent ({agent}) response:\n{content}\n\nScore this response."` on every corpus, with `{prior}` formatted by `code/path_b_judge.py:_format_prior` (agent name plus truncated content per turn, joined by blank lines).

T JUDGE-SWAP CONCORDANCE: GPT-4O-MINI VS SONNET 5

To quantify how conditional the paper's empirical results are on the shipped LLM judge (`gpt-4o-mini`, App S), we ran the same prompt under Anthropic `claude-sonnet-5` on the same 10,956 turns spanning four splits: Who&When HC (2,935), math dev-safe (6,188), math held-out safe (998), math held-out unsafe (835). Sonnet 5 does not support the shipped judge's assistant-message prefill or the temperature parameter, so the two judges differ on those protocol details in addition to the model. We did not run a controlled comparison isolating the protocol contribution, so those protocol differences are an uncontrolled confound in the substantive differences reported below.

Marginal rubric-level distributions differ sharply. Per-judge marginals on each judge's own non-null turns:

corpus / judge	0.001	0.2	0.5	0.8	0.999
HC — GPT	8.0%	7.5%	14.1%	5.3%	**65.0%**
HC — Sonnet	7.6%	25.4%	39.9%	17.4%	**9.7%**
math held-out safe — GPT	2.4%	6.4%	10.9%	3.6%	**76.7%**
math held-out safe — Sonnet	8.0%	10.0%	8.0%	7.1%	66.8%
math held-out unsafe — GPT	4.7%	12.6%	16.9%	4.8%	**61.1%**
math held-out unsafe — Sonnet	29.1%	25.4%	11.2%	6.6%	27.7%

Sonnet is systematically more conservative than GPT at the top level: on Who&When HC, GPT places 65% of turns at 0.999 against Sonnet's 9.7%; on math failing traces, 61% against 28%.

Confusion on Who&When HC. Restricted to the paired subset (2,890 turns where both judges have a non-null score):

	S=0.001	S=0.2	S=0.5	S=0.8	S=0.999
G=0.001	72	77	58	27	1
G=0.2	40	72	71	26	5
G=0.5	35	125	175	62	11
G=0.8	6	45	74	24	6
G=0.999	**68**	**416**	**775**	**363**	**256**

Of the 1,878 HC turns GPT sends to top, only 256 (14%) stay at top under Sonnet; 1,259 (67%) drop by two or more rubric levels. Sonnet is not disagreeing at the margin — it re-scores GPT's top mass across the full range.

Level-match and Spearman by corpus.

corpus	paired n	level match	Spearman ρ
Who&When HC	2,890	20.7%	+0.29
math dev-safe	6,002	61.5%	+0.44
math held-out safe	971	63.2%	+0.44
math held-out unsafe	798	30.7%	+0.25

Per-role pattern. On math dev-safe, agreement rises with the syntactic determinism of the role's output: *environment* (programmatic tool responses) 75.1%; *Verifier* 69.8%; *MathSolver* (reasoning text) 53.2%; *user* (initial question) 15.1%. The pattern is consistent with judges agreeing when there is little interpretive latitude and disagreeing on reasoning text where the rubric admits many mappings.

Null pattern. Sonnet 5 returned no score on 295 of the 10,956 turns (2.7%): 45 on HC, 213 on math safe (dev + held-out combined), 37 on math unsafe. On Who&When HC the withhold pattern is associated with prior observed scores: 40 of the 42 nulls whose immediate prev is non-null follow a not-top prev turn (95.2%), against a Sonnet-side base rate of 72.6%; trajectory-clustered BCa 95% CI on the null-conditional rate is $[86.7\%, 100.0\%]$ across 27 contributing trajectories ($B = 10{,}000$, seed matched to App P; `results/second_judge_null_pattern_cluster_20260916.txt`). The upper endpoint is at the boundary of the parameter space on 27 clusters — 100% is a saturation of the bootstrap, not a point estimate — but the interval excludes the base rate. This is an observed association between missingness and the previous observed score, not evidence that missingness depends on the unobserved score itself. On math_unsafe the analogous rate is $24/36 = 66.7\%$ against base 64.1%, cluster BCa $[50.0\%, 81.5\%]$ covers the base rate; the elevation is HC-specific. Rule for downstream handling (prespecified before any analysis): drop the null turn from both fit and eval. Operationally, the null turn keeps its original index in the trajectory record; the eval sequence is compressed (indices $t = 0, 1, \ldots$ over non-null turns only), which shifts anchor positions and the effective onset. For onset comparisons we use $\text{scored_pos_at_eval}[\text{first_alarm_eval}] \geq \text{onset_scored}$, mapping the compressed alarm position back to the original scored index (`code/second_judge_hc_restricted_precise.py`), which correctly handles nulls dropped before onset. The turn immediately after a dropped null enters state-conditioned analyses with $C_t = \text{unknown}$; on the shipped data $C_t = \text{unknown}$ falls under the role-only fallback because no state-specific reference is fit for it.

Within-judge coherence checks. Two follow-up analyses on the Sonnet-scored corpora, framed as coherence checks not numerical replications (the marginal differences above rule out the latter). *First*, the App R phenomenon appears under Sonnet: refitting the whole pipeline on Sonnet math (Sonnet-side $\widehat{S}_r$ and $\widehat{S}(r, c)$) gives arm 1 conditional-factor mean 1.24, cluster BCa 95%

CI $[1.07, 1.42]$ (excludes 1), and arm 3 mean 0.92 ($[0.81, 1.04]$, covers 1). A second, differently-scoring judge sees the arm 1 exceedance and the state-conditioned move. *Second*, Sonnet moves the HC W_{loc} operating point on both axes. Restricted detection changes from $15/58 = 25.9\%$ to $21/58 = 36.2\%$ ($[24.0, 49.9]$) — an absolute increase of 6 trajectories, ratio $1.4\times$. Pre-mistake Type-I changes from $1/54 = 1.9\%$ to $4/54 = 7.4\%$ ($[2.1, 17.9]$) — an absolute increase of 3 trajectories, ratio $4\times$. Two operating points cannot establish improved separation or identify why they moved; we report the shift without further interpretation. Full artifacts: `results/second_judge_coherence_20260915.txt`.

Every empirical rate in this paper — Type-I, detection, and the App Q conditional-factor diagnostic — is a property of the paper's shipped judge, and would move under a differently-behaved instrument by at least the marginal shift shown above.

U AGENTFORESIGHT COMPARISON

The closest published prior work in trained online auditors for AFTraj-style multi-agent traces is AgentForesight-7B (Zhang et al., 2026): an RL-fine-tuned online auditor (Qwen2.5-7B-Instruct trained on AFTraj-2K, ≈2.3K trajectories, coarse-to-fine BPPO+GRPO recipe). Its authors report performance under a different metric, a held-out split whose membership against ours cannot be confirmed (see (ii) below), and a different training pipeline than ours, so a direct comparison to Table 2 is not available on either paper's numbers.

What Zhang et al. report. On the AFTraj-2K held-out split (their 15% trajectory-grouped split; membership against our `TEST_SAFE_IDS/TEST_UNSAFE_IDS` cannot be confirmed — see item (ii) below), AgentForesight-7B achieves 66.44 overall Exact-F1 (77.36 math, 78.87 coding, 48.70 agentic; their Table 1) at empirical false-alarm rate $\text{FAR} = 2.4\%$ on their $\mathcal{D}_{\text{safe}}$ and Step-Acc $= 59.5\%$ on their $\mathcal{D}_{\text{unsafe}}$ (their Figure 4). On the external Who&When benchmark they report Step-Acc 57.69 and Agent-Acc 73.08 (their Table 2). The 2.4% FAR is an empirical rate, not a formal bound.

Why the comparison is not apples-to-apples. Three axes differ. (i) *Metric.* Exact-F1 is a harmonic mean of step-level recall and precision on decisive-error step localisation. Our reported quantities are trajectory-level Type-I (on complete safe trajectories for AFTraj; on pre-onset windows for Who&When) with a formal stopping-time bound (Table 2), and restricted trajectory detection under the alarm-at-or-after-onset convention. Ours does not penalise mislocalisation of the alarm past onset within a trajectory; theirs does. (ii) *Safe-pool split.* Their FAR is measured on their AFTraj-2K 15% trajectory-grouped held-out split; ours is a 15% trajectory-grouped held-out split of AFTraj at 332 trajectories (91 coding, 116 math, 125 agentic; 169 safe, 163 unsafe). Their reported counts match ours exactly across every domain and safe/unsafe cell under the same protocol name on the same upstream corpus, which is consistent with the two evaluations drawing on the same partition; AgentForesight-7B's held-out conv-id list is not released with the paper, so membership cannot be confirmed either way from our artifacts. (iii) *Training pipeline.* They RL-fine-tune 7B parameters on ≈2.3K trajectories with a 2×H200 setup; the CCC monitor we ship trains no auditor parameters and fits only a per-role reference on dev-safe.

The one coordinate that lines up: FAR versus W_{loc} Type-I. The closest we can put the two numbers is false-alarm rate on labelled-safe trajectories: AgentForesight-7B reports $\text{FAR} = 2.4\%$ on its $\mathcal{D}_{\text{safe}}$ (their Fig. 4); our shipped W_{loc} reports Type-I $2/58 = 3.4\%$ (math), $0/54 = 0.0\%$ (coding), $1/57 = 1.8\%$ (agentic) on `TEST_SAFE_IDS` (Table 2). We report both and claim neither direction of superiority. For the comparison to bear on relative performance the two rates would need to be measured on the same trajectory pool and with a matched notion of "safe" (their $\mathcal{D}_{\text{safe}}$ carries their filtering pipeline's outcome-plus-integrity-plus-coherence gates; our `TEST_SAFE_IDS` carries AFTraj's outcome-based safe label plus our parent–child leakage audit). Neither of the two is confirmed in matching form here, so the two numbers do not compose into a ratio or a gap that means anything. Even if the pools coincide, the safe-definition mismatch remains, so the conclusion does not rest on pool identity.

Head-to-head and superiority claims. A head-to-head requires running AgentForesight-7B on our `TEST_SAFE_IDS/TEST_UNSAFE_IDS` split (or CCC on their held-out split) and comparing under

a common metric — either their Exact-F1 or our Type-I / restricted-detection pair. That experiment is out of scope for this paper and stands as future work. We make no superiority claim on any axis we have not measured: the two methods target different operational contracts (a formal per-trajectory bound versus an empirical FAR from a trained auditor).

V PRE-REGISTRATION DEVIATIONS, ITEMIZED

Sites at which the shipped paper departs from the pre-registration (`manuscript/CCC_experiments_v3.tex`, 24 August 2026). Referenced from the Reproducibility Statement in the body.

1. All three AFTraj domains are deployed with the δ-level bounded WSR upper confidence bound that Theorem 2 requires (fixed $\delta = 0.005$), including agentic. Agentic was initially shipped under the raw plug-in because per-role loss under WSR widths exceeds the $\alpha = 0.05$ hurdle on eight of twelve roles; we later shipped bounded on agentic anyway because Corollary 2's loss-vs-hurdle diagnostic is not a licence to mix references at deployment time (Appendix Q shows it is neither necessary nor sufficient for conditional-law validity). Every primary AFTraj result in the shipped paper reflects bounded agentic under Path-A. Two secondary analyses retain the earlier plug-in reference and are marked as such where they appear: the off-diagonal cells in Appendix J and the out-of-fold Type-I comparison and bootstrap-with-refit in Appendix R. The arm-2b matched-inflation comparison in Appendix R additionally retains the per-trace anchor budget.
2. No δ sweep.
3. Development/held-out split rather than ten-fold cross-fitting.
4. No Benjamini–Hochberg correction in the original analysis; applied post hoc, no decision flips: math W_{fwd} at $\alpha = 0.05$ (raw $p = 0.0022$, BH $q = 0.020$ across the nine AFTraj cells and $q = 0.029$ across all thirteen) survives multiplicity control. We keep the uncorrected reading primary because a reader checking whether the per-cell guarantee holds wants the per-cell answer; BH is a secondary check on false-discovery across a chosen family. Full BH table in the release archive.
5. The third statistic is M_{dual} (Section 3.4), not the pre-registered "tempered" variant (`manuscript/CCC_experiments_v3.tex` lists "tempered" only as a label without a formal definition; we ship the concrete M_{dual} instead).
6. M_{dual} correction (2026-09-17). Earlier code computed a forward-reverse variant $\frac{1}{2}(W_{\mathrm{fwd}} + W_{\mathrm{rev}})$ instead of the paper's $\frac{1}{2}(W_{\mathrm{fwd}} + W_{\mathrm{loc}})$; the reverse product reads future turns and is not anytime-valid, so the paper's validity claim did not hold for what was computed. The M_{dual} statistic is now $\frac{1}{2}(W_{\mathrm{fwd}} + W_{\mathrm{loc}})$ everywhere (Section 3.4: a top-heavy prior on the same anchored-mixture family). Every M_{dual} number in the paper has been regenerated under the corrected construction — code `aftraj_inheritance_check.py` and `onset_regen_wsr.py`; artifacts `canonical_table_v4_wsr_20260917.csv`, `onset_delay_v4_wsr_20260912.csv` (overwritten), `stage3_state_cond_eval_math_heldout_20260917.csv`, `stage4_conservatism_control_v2_math_heldout_20260917.txt`. Type-I numbers on complete safe trajectories are unchanged; the detection numbers that changed are catalogued in the diff artifact `results/canonical_v4_wsr_diff_20260917.md`.
7. Judge-prompt cost equation correction. An earlier draft of the judge-cost equation in Appendix I.2 included a gold-answer conditioning term. That term was removed from the judge prompt before the shipped p-hat caches were produced (no cached AFTraj score originated from a prompt containing a "Correct answer:" line, per Appendix S); the prompt shown in Appendix S is the one every shipped score was scored under. The removal does not change any reported number.
8. Appendix Q symmetric-annotation paragraph correction (2026-09-17). An earlier version of the paragraph gave the math process-fault rate as $1/30 = 3\%$, folding the four pre-registered forced math traces (two alarm-under-adaptive-$c = 1$, two matched controls) into the denominator. The sampling script

`code/aftraj_sym_annotation_sample.py` draws only `random_math` ($n = 26$) uniformly from the held-out safe pool; the four forced traces are case descriptions per the labels file's `case_description_disclaimer_20260825`, not draws from that pool. The estimable math rate on this sample is `random_math` $1/26 = 3.85\%$ (Wilson 95% $[0.7\%, 18.9\%]$). Appendix Q now reports rates by estimable substratum with Wilson intervals; the fold-in figure $1/30$ no longer appears in the paper.

9. The α grid is extended at both ends: 0.02 rather than 0.025, and adds 0.15 and 0.25.
10. Symmetric annotation of 60 annotated-safe trajectories was run on 2026-08-25 (labels sealed in `code/aftraj_sym_annotation_labels_20260825.json`, key at `code/aftraj_sym_worksheet_key_20260825.json`) as a single-annotator blind read; Cohen's κ deferred (a second annotator was not run in time). Counts are reported in Appendix Q. (The fault-present denominator is reported: Table 4 and the AFTraj summary in App. N score held-out failing trajectories.)
11. The clustering test on alarm times is not run. (Median lag, σ/K, and censoring rates are reported in Table 7 and the AFTraj sentence beneath it.)
12. Per-domain per-role total-variation separation between safe turns and turns at or after onset is not reported.
13. The fraction of unsafe trajectories on which the annotated turn itself receives a factor above one is not reported. (Post-onset horizon $K - u$ is reported in App. N.)
14. Full-context versus output-only judge configurations: only one run.
15. Wrapping an external detector: not demonstrated.
16. Fisher's exact is not run. Paired arm-vs-arm alarm comparisons in App R report McNemar discordance counts (b, c) as descriptives; the paired Type-I difference is intervalled by a trajectory-cluster bootstrap-with-refit ($B = 1000$; `code/stage6_bootstrap_refit.py`) because cross-fitting does not remove the dependence between alarm indicators induced by shared training trajectories, so the exact-binomial reading of McNemar is not automatically justified. Primary detection intervals stay Clopper–Pearson and multiple-comparison correction stays BH.
17. Judge provenance: recorded in the Reproducibility Statement rather than declared unavailable.
18. Mechanism decomposition by failure type is not run.
19. Judge-prompt pilot re-analysis. The earlier prompt-variant pilot (`code/judge_pilot.py`) held p_{null} fixed at the shipped baseline value across variants and so measured $\widehat{p}$-distribution drift, not each variant's own separation between safe and failing turns. Re-analysed with per-variant p_{null} via LOO (`code/judge_pilot_reanalyze.py`) the chain-of-thought variant led on the 20-trace pilot under the pilot's own continuous factor construction. Scaled to all 58 HC traces (`code/judge_cot_hc58.py`) that advantage narrowed to overlapping Clopper–Pearson intervals; when re-scored under the SHIPPED categorical shrunk-plug-in-$1b$ path that Table 2 uses (`code/hc_shipped_path_rescore.py`), the baseline reproduces Table 2's $15/58$ HC detection exactly and CoT changes it by one trajectory: restricted detection $16/58 = 27.59\%$ [16.7, 40.9], pre-mistake Type-I unchanged at $1/54$, conditional-factor diagnostic $\mathbb{E}[\varepsilon \mid \text{prev not top}]$ $1.053 \to 1.026$. We kept the shipped prompt because under the deployed scoring path the intervention has no measurable effect either way.
20. Score-to-evidence mapping search under the Kelly-log-optimal criterion: prespecified six-candidate family run on AFTraj math and Who&When HC; the winner (flattest, $\kappa = 3$) did not beat the shipped linear weight on held-out detection. Full finding reported in the extension to Remark 4 of Appendix B.3; artifacts at `results/score_evidence_search_prespec_20260917.txt` and `results/score_evidence_search_stage1_2_20260917.txt`.
21. AFTraj anchor budget K' was originally implemented per-trace as $\lceil K/3 \rceil$ (with K the realised chain length); the resulting anchor set and $1/K'$ mixture weights are not $\mathcal{G}_0$-measurable and violate Assumption 1. Corrected to a per-domain census rule $K' = \lceil K_{\max}/3 \rceil$ with $K_{\max}$ the 95th-percentile chain length on dev-safe (§5.1): math $K' =$

12, coding 4, agentic 7; Who&When already used this convention (HC 43, AG 4). On Who&When the census pool contains the evaluated traces; a leave-one-out census leaves K' unchanged for every trace (HC 43, AG 4). Every primary result in Table 2 and Section 5 reflects the fixed per-domain K'; the secondary analyses identified in the Reproducibility Statement (Appendices J, R's arm-2b matched-inflation, E, M, and L) retain the earlier per-trace anchor budget at the point of use. Diff artifact `results/kprime_shipped_vs_post_20260918.csv`.