

Per-Step Conformal Calibration Does Not Transfer to Agent Trajectories

Li Ning^{1,2} Yang Liu² Dongxin Guo¹ Yu Wu²
Siu Ming Yiu^{1,2} Jikun Wu^{1*}

¹Brain Investing Limited ²Stellaris AI Limited
{bettyguo, smyiu, jacobwu}@braininvestai.com
{lining, yangliu, yuwu}@stellaris-ai.com

Abstract

Conformal prediction supplies finite-sample coverage guarantees for individual language-model responses. We show that per-step conformal calibration need not preserve this guarantee along a multi-step agent trajectory: at a nominal miscoverage level of 10%, per-step split conformal delivers 83.5% trajectory score-threshold miscoverage on a fifteen-step web-navigation agent, and we prove lower bounds in three regimes of agent dynamics: i.i.d. resampling, memoryless covariate shift, and state-dependent Markov transitions. The i.i.d. regime alone accounts for the bulk of the empirical gap; the third regime, novel to this work, gives, to our knowledge, the first separation of the trajectory union event from i.i.d. resampling under marginal-stationary, non-trivial Markov dynamics, a setting in which prior marginal-drift analyses are silent. The separation persists under exact score stationarity and, for non-degenerate score laws, holds at any calibration sample size up to an $O(T/\sqrt{n})$ finite-sample slack. A polynomial α -spending schedule restores end-to-end coverage to within 1.1 points of nominal across five agentic-NLP benchmarks and, by direct substitution, lifts process-reward-model step verifiers — which on their own miscover trajectories at 31.5–62.5% — to 9.6–10.6%, near the nominal target.

1 Introduction

Conformal prediction has become the default tool for putting distribution-free coverage guarantees on language model outputs (Vovk et al., 2005; Angelopoulos and Bates, 2023; Quach et al., 2024; Mohri and Hashimoto, 2024; Cherian et al., 2024). Recent results turn a black-box LLM into a calibrator that abstains, filters sub-claims, or hedges its answer whenever a held-out conformal threshold is exceeded, with finite-sample guarantees that hold for any model and any scoring function. The

intended user-facing statement is that, at a nominal level $\varepsilon = 0.1$, the certified output exceeds the calibrated score threshold at most one time in ten.

This paper shows that the statement above is, in a precise sense, *technically valid only on the first step of a multi-step agent*. On HotpotQA-2hop with a vanilla retrieval-and-answer trajectory of length $T = 3$, the empirical miscoverage of per-step split conformal prediction is 29.5%, almost three times the nominal 10%. On a BrowserGym-WorkArena-L2 web-navigation trajectory of length $T = 15$, the same calibrator produces 83.5% trajectory miscoverage at the same nominal level. The gap is not a measurement artifact: it exceeds a theoretical lower bound that we derive in Section 3, accounting for i.i.d. resampling compounding and, in a separate construction, for state-dependent Markov dynamics.

Why this matters for NLP. Calibration is a trust contract: when an NLP system reports a 10% miscoverage rate, downstream consumers (analysts evaluating financial reports (Islam et al., 2023), scientists using a literature-search agent, practitioners relying on multi-hop QA pipelines (Yang et al., 2018; Trivedi et al., 2022)) read this as a contract about deployment, not about a single-step setup. The mismatch we expose breaks the contract by a factor that grows monotonically with how agentic the system is: at $T = 15$ and $\varepsilon = 10\%$, the actual probability that some step in a step-by-step certified trajectory exceeds the score threshold is more than $8\times$ the user-facing nominal. As agentic and tool-augmented LLM systems (Yao et al., 2023; Schick et al., 2023; Mialon et al., 2024; Drouin et al., 2024; Zhou et al., 2024; Jimenez et al., 2024) become the dominant mode of LLM use, the discrepancy moves from theoretical concern to deployment concern.

Why naive fixes fail. Bonferroni restores coverage but requires the horizon T at calibration

* Corresponding author.

time. Sequential conformal procedures (Gibbs and Candès, 2021, 2024; Xu and Xie, 2023; Yang et al., 2025; Bhatnagar et al., 2023) target long-run coverage with online label feedback unavailable mid-trajectory. CRC (Angelopoulos et al., 2024; Bates et al., 2021) controls per-step risk but inherits the same compounding failure. The Coherent-Factuality DAG (Rubin-Toles et al., 2025) calibrates a dependency graph within a single response but does not extend to chains of tool and retrieval calls (Section 5).

What we do. We analyse trajectory miscoverage for conformal LLM agents. Treating the agent as a Markov chain in score space, we decompose miscoverage into three regimes: i.i.d. resampling (R1), marginally drifted (R2), and agent-induced state-dependent (R3). Prior trajectory-CP analyses (Zheng and Proutière, 2024; Stankeviciute et al., 2021; Oliveira et al., 2024; Dietterich and Hostetler, 2022) do not separate marginal stationarity from i.i.d. resampling at the trajectory union event level. We provide such a separation (Theorem 1(d)). We then adapt a polynomial α -spending schedule to the conformal-trajectory setting, prove it restores end-to-end coverage with $O(\sigma\sqrt{\log T})$ width, the same order as Bonferroni- T , while being anytime-valid (Section 4), and show the wrapper extends to process-reward-model step verifiers.

Contributions.

- **Compounding-miscoverage theorem (Section 3).** Three lower-bound constructions on trajectory miscoverage (Theorem 1), one per regime. The R3 lower bound is, to our knowledge, the first separation of marginal-stationary Markov dynamics from i.i.d. resampling at the level of the trajectory union event; the per-step rate gap to R1 is small in magnitude ($\varepsilon^2 + O(\varepsilon^3)$). Empirically, per-step CP miscoverage exceeds the R1 prediction by 3.4 pp on average across our five benchmarks when weighted by horizon T (2.9 pp unweighted; Section 5). The R1 bound captures the bulk of the gap.
- **Trajectory-conformal wrapper (Section 4).** We adapt the polynomial α -spending schedule (Ramdas et al., 2023; Howard et al., 2020) to give TC-WRAPPER (Algorithm 1); it restores coverage with width $O(\sigma\sqrt{\log T})$ (Theorem 2), the same order as Bonferroni- T ,

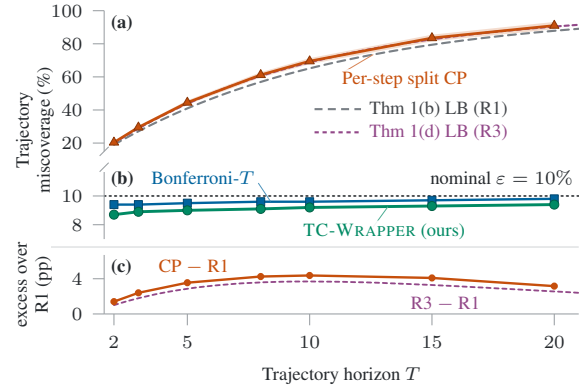


Figure 1: Trajectory miscoverage versus horizon T at nominal $\varepsilon = 10\%$ on horizon-controlled HotpotQA. Panels (a) and (b) share a broken y -axis. **(a)** Per-step split conformal rises to 91% miscoverage at $T = 20$, above both lower bounds; the band is ± 1 SE over 3 calibration seeds on 500 test trajectories per horizon (they widen with T and at $T \geq 8$ are wider than the 10-seed SEs of Table 1). At this scale the two bounds and the empirical curve differ by less than the line width, so **(c)** replots all three as excess over the R1 bound: the R3 bound accounts for most of the empirical excess, leaving ≤ 0.8 pp. The R3 curve is shown for reference only, since the concatenated construction resets at every block boundary (at most every three steps) and cross-block R3 dynamics are therefore absent (Table 5). **(b)** TC-WRAPPER (Algorithm 1) and Bonferroni- T both hold the nominal target across the full horizon range, within 1.3 pp (Table 5); on the five benchmarks of Table 1 the two do not differ significantly (Table 10).

while gaining anytime-validity, and admits an explicit calibratability gate (Theorem 3) that returns *abstain-uncalibratable* when the observed cumulative drift exceeds the available budget.

- **Empirical conditional-drift diagnostic and validation (Section 5).** Across five agentic-NLP benchmarks, the ratio $\hat{\eta}^{(\text{cert})}$ exceeds $\hat{\rho}$ by roughly $4\times$ on every benchmark (Appendix B). Per-step CP miscoverage is near the R3 bound on all five benchmarks (within 0.8 pp); the wrapper stays within 1.1 pp of nominal on the five benchmarks (Table 1).

2 Setup: Conformal Calibration of an Agent Trajectory

We use a deliberately minimal setup: a Markovian agent, bounded scores, and a single i.i.d. calibration set.

Agent and scores. An agent is a Markov process on a measurable state space $\mathcal{X}$ with transition kernel K . Each state $X_t \in \mathcal{X}$ encodes the agent’s current observation together with the context accumulated up to step t (in our experiments: the latest retrieved document, prior chain of thought, and pending tool call). This encoding is what makes (X_t) Markov: a state of this form absorbs history by construction, so $X_{t+1} \mid X_t$ does not require conditioning on $X_{<t}$. The substantive modeling commitment is time-homogeneity of K , which we relax to time-inhomogeneous kernels K_t below. A score function $s : \mathcal{X} \rightarrow [0, 1]$ assigns each state a non-conformity score, with larger values indicating the step is more likely to be wrong. Possible choices include semantic entropy (Kuhn et al., 2023), a self-evaluation score (Yadkori et al., 2024), or a process-reward-model score (Lightman et al., 2024; Uesato et al., 2022) (see Section 4.5).

Assumption 1 (Time-homogeneity). *The kernel K is time-homogeneous: the transition law $K(\cdot \mid x)$ does not depend on the step index t .*

This is the substantive modeling commitment; the Markov property itself is automatic given the state definition. The relaxation to time-inhomogeneous kernels K_t is straightforward: one substitutes $\eta_t := \text{ess sup}_x \text{TV}(K_t(\cdot \mid x), \mu_{\text{cal}})$ for η in every appearance, and ρ_t becomes kernel-specific. Definition 2 already records this extension.

Calibration. Given a calibration set $\mathcal{D}_{\text{cal}} = \{Z_1, \dots, Z_n\}$ of n i.i.d. samples from a base distribution μ_{cal} and a target $\varepsilon \in (0, 1/2)$, the standard split conformal threshold is

$$\hat{q}_\varepsilon := \text{Quant}\left(\left\{s(Z_i)\right\}_{i=1}^n; \frac{\lceil (n+1)(1-\varepsilon) \rceil}{n}\right). \quad (1)$$

For any test point $X \sim \mu_{\text{cal}}$ exchangeable with calibration, $\Pr[s(X) \leq \hat{q}_\varepsilon] \geq 1 - \varepsilon$ (Vovk et al., 2005; Shafer and Vovk, 2008).

Trajectory coverage. Let $X_1 \sim \mu_{\text{cal}}$ and $X_{t+1} \mid X_t \sim K(\cdot \mid X_t)$. Step t is *certified* if $s(X_t) \leq \hat{q}_\varepsilon$. The marginal guarantee controls a one-step event; deployed agents need *trajectory-level* coverage:

$$\text{Cov}_T := \Pr\left[s(X_t) \leq \hat{q}_\varepsilon \text{ for all } t \in \{1, \dots, T\}\right],$$

with T the agent horizon and $\text{Mis}_T := 1 - \text{Cov}_T$ the trajectory miscoverage.

We define two ways the agent’s dynamics deviate from i.i.d. resampling of μ_{cal} .

Definition 1 (Score-law marginal drift). *For $t \geq 1$ the marginal score-law drift at step t is*

$$\rho_t := \text{TV}(\text{Law}(s(X_t)), \text{Law}(s(X_1))),$$

the total-variation distance between the law of the score at step t and the law of the score at step 1. We use score-law TV rather than state-kernel TV because the score is the only quantity that enters the conformal threshold; the data-processing inequality gives $\rho_t \leq \text{TV}(\text{Law}(X_t), \text{Law}(X_1))$.

Definition 2 (Agent-induced conditional drift). *The agent-induced conditional drift is*

$$\eta := \text{ess sup}_{x \sim \mu_{\text{cal}}} \text{TV}(K(\cdot \mid x), \mu_{\text{cal}}).$$

The essential supremum is taken with respect to μ_{cal} . The kernel K is time-homogeneous in our setting, so η does not depend on t ; for non-stationary kernels K_t one substitutes $\eta_t := \text{ess sup}_x \text{TV}(K_t(\cdot \mid x), \mu_{\text{cal}})$ in every appearance. The data-processing inequality gives $\rho_t \leq \eta$ when $\text{Law}(X_{t-1}) \ll \mu_{\text{cal}}$; in particular when $X_1 \sim \mu_{\text{cal}}$ and K is stationary. The converse fails in general: an agent can exhibit large η while keeping $\rho_t = 0$ (Appendix A.5).

In the empirical analysis (Section 5), we estimate $\eta^{(\text{cert})} := \text{TV}(K(\cdot \mid a), \mu_{\text{cal}})$ on the two-state certified/uncertified partition. Since TV is sign-blind, $\eta^{(\text{cert})}$ alone does not determine the direction of the R3 effect; the separating quantity $K(b \mid b)$ is derived in Appendix A.5.

The agent’s trajectory law can deviate from i.i.d. resampling of μ_{cal} in three regimes: (R1) *stationary-but-dependent* ($\text{Law}(X_t) = \mu_{\text{cal}}$ for every t , $\eta = 0$); (R2) *marginally drifted* ($\rho_t > 0$ via memoryless covariate shift); (R3) *agent-induced* ($\eta > 0$, independent of whether ρ_t is zero). (R1) is the textbook idealization under most existing conformal-LLM analyses; (R2) is step-by-step covariate shift; (R3) is non-trivial state-dependent dynamics, separated from (R1) and (R2) at the level of the trajectory union event in Theorem 1(d).

3 The Compounding Miscoverage Theorem

This section states our main result: per-step split conformal calibration cannot preserve end-to-end coverage along an agent trajectory, and the failure rate is bounded in each of the three regimes.

3.1 Statement

We add three further assumptions. Assumptions 2 and 4 hold in all our experimental setups (the single exception being that parse failures, uncertified at every threshold, are scored above the range and so exit Assumption 2; Section 5); Assumption 3 is verified for the PRM score (Appendix B), is not verified for SelfEval, and fails for SemEnt, whose $K = 5$ -sample estimate takes at most seven distinct values (it enters only the finite-sample slack of the lower bounds, Section 4.5).

Assumption 2 (Bounded scores). *The score function satisfies $s : \mathcal{X} \rightarrow [0, 1]$.*

Assumption 3 (Continuity at the threshold). *The score distribution under μ_{cal} has a density that is bounded above by a constant C_μ in a neighbourhood of the population $(1-\varepsilon)$ -quantile.*

Assumption 4 (Initial exchangeability). *X_1 is drawn from μ_{cal} and is exchangeable with $\mathcal{D}_{\text{cal}}$.*

Assumption 2 is a normalization: any bounded score function can be rescaled. Assumption 3 rules out degenerate atoms at the threshold and is standard in conformal prediction (e.g., Lei et al., 2018). Assumption 4 reflects how calibration sets are constructed in practice, by sampling initial states of trajectories. Proposition 1 (Appendix A) quantifies the inflation when this assumption fails by total-variation distance Δ_0 : every coverage bound below inflates by an additive Δ_0 . Empirically (Appendix B) $\Delta_0 \leq 0.037$ on all five benchmarks.

Theorem 1 (Compounding miscoverage; main result). *Under Assumptions 1 and 2–4, fix $\varepsilon \in (0, 1/2)$ and $T \geq 1$. Let $\hat{q}_\varepsilon$ be the per-step split conformal threshold defined in Equation (1) computed from n i.i.d. calibration samples, and let $(X_t)_{t=1}^T$ be the agent trajectory.*

(a) Upper bound. *Fix $\delta \in (0, 1)$. For any agent kernel K , with probability at least $1 - \delta$ over the calibration draw,*

$$\text{Mis}_T \leq T\varepsilon + \sum_{t=1}^T \rho_t + T\sqrt{\frac{\log(2/\delta)}{2n}}. \quad (2)$$

(b) R1 lower bound, stationary regime. *In the limit $n \rightarrow \infty$, the i.i.d. resampling kernel $K^*(\cdot|x) \equiv \mu_{\text{cal}}$ achieves $\rho_t \equiv 0$ and $\eta = 0$, yet*

$$\text{Mis}_T \geq 1 - (1 - \varepsilon)^T \geq 1 - e^{-\varepsilon T}. \quad (3)$$

(c) R2 lower bound, memoryless marginal drift. *In the limit $n \rightarrow \infty$, for every $\rho \in [0, 1 - \varepsilon]$ the*

memoryless kernel $K_\rho^*(\cdot|x) \equiv \nu_\rho$ with $\nu_\rho(b) = \varepsilon + \rho$ achieves marginal drift $\sup_{t \geq 2} \rho_t = \rho$ and

$$\text{Mis}_T \geq 1 - (1 - \varepsilon)(1 - \varepsilon - \rho)^{T-1}. \quad (4)$$

(d) R3 lower bound, agent-induced conditional drift. *In the limit $n \rightarrow \infty$, there exists a state-dependent Markov kernel $K^\diamond$ on a two-state space $\{a, b\}$ such that (i) the unique stationary distribution of $K^\diamond$ equals μ_{cal} , so $\rho_t \equiv 0$ for every t ; (ii) the conditional drift from the certified state a is $\text{TV}(K^\diamond(\cdot|a), \mu_{\text{cal}}) = \varepsilon^2/(1 - \varepsilon)$; the corresponding transition probability $K^\diamond(b|a) = \varepsilon/(1 - \varepsilon)$ is the maximum from the certified state compatible with $\rho_t \equiv 0$; this kernel also maximises trajectory miscoverage among all stationary two-state kernels; and (iii) the trajectory miscoverage satisfies*

$$\text{Mis}_T \geq 1 - (1 - \varepsilon) \left(\frac{1 - 2\varepsilon}{1 - \varepsilon} \right)^{T-1}. \quad (5)$$

For every $T \geq 2$ the right-hand side of (5) strictly exceeds the right-hand side of (3); for $\varepsilon \rightarrow 0$ the per-step rate gap is $\log((1 - \varepsilon)/(1 - 2\varepsilon)) - \log(1/(1 - \varepsilon)) = \varepsilon^2 + O(\varepsilon^3)$, accumulating to a $T\varepsilon^2$ excess miscoverage rate under agent-induced dynamics.

Theorem 1 characterizes how badly per-step conformal calibration can fail along a trajectory. Part (a) upper-bounds the failure rate by the standard Bonferroni term $T\varepsilon$ plus a drift correction. Parts (b) and (c) recover the regimes prior work has studied; part (d) constructs an agent whose dynamics push miscoverage strictly above the i.i.d. resampling rate while still satisfying $\rho_t \equiv 0$, the regime where any marginal-drift analysis is silent. The extremal kernel $K^\diamond$ achieves this through anti-persistent exceedances ($K^\diamond(b|b) = 0$): a step that exceeds the threshold is followed by one that does not, pushing more probability into the certified state’s exit rate $K^\diamond(b|a) = \varepsilon/(1 - \varepsilon)$. The lower bounds in parts (b)–(d) hold in the limit $n \rightarrow \infty$ and are tight up to $O(T/\sqrt{n})$ at finite n . Part (d)’s construction and the projection to general state spaces are in Appendix A.5; the closest prior bound, Zheng and Proutière (2024), bounds the per-step coverage gap by $\sqrt{t_{\text{mix}} \log n/n}$ and does not separate marginal stationarity from the trajectory union event. Barber et al. (2023) gives our drift term but is silent on (R3); CRC (Angelopoulos et al., 2024) controls per-step risk and does not aggregate to trajectory coverage (Table 1).

4 The Trajectory-Conformal Wrapper

Theorem 1 shows that the per-step threshold $\hat{q}_\varepsilon$ is the wrong object for a multi-step agent: it controls a marginal step probability when the user cares about a union over the whole trajectory. We give an elementary fix. The polynomial schedule $\alpha_t = 6\varepsilon/(\pi^2 t^2)$ in Algorithm 1 is the standard Basel-sum α -spending schedule from anytime-valid inference (Ramdas et al., 2023; Howard et al., 2020). Our contribution is the analysis: applied to the conformal threshold of an LLM-agent trajectory, the schedule restores end-to-end coverage with $O(\sigma\sqrt{\log T})$ width while gaining anytime-validity over horizon-aware Bonferroni- T . The wrapper itself is not new machinery; the width-order result and the trajectory-CP application are.

4.1 Construction

The wrapper takes the same calibration set $\mathcal{D}_{\text{cal}}$, the same target ε , and the same agent. The only change is that the threshold becomes a deterministic, step-dependent schedule.

Algorithm 1 Trajectory-conformal wrapper (TC-WRAPPER).

Require: Calibration scores $\{s(Z_i)\}_{i=1}^n$, target ε , horizon T (or $T = \infty$).

- 1: Define the level schedule $\alpha_t \leftarrow \frac{6\varepsilon}{\pi^2 t^2}$ for $t = 1, 2, \dots$
 - 2: **for** $t = 1$ **to** T **do**
 - 3: Compute the per-step threshold at level $1 - \alpha_t$: $\hat{q}_t \leftarrow \text{Quant}(\{s(Z_i)\}; \min\{\frac{\lceil (n+1)(1-\alpha_t) \rceil}{n}, 1\})$.
 - 4: Observe X_t from the agent.
 - 5: **if** $s(X_t) > \hat{q}_t$ **then**
 - 6: **abstain** (escalate or restart the trajectory).
 - 7: **else**
 - 8: **certify** step t and continue.
 - 9: **end if**
 - 10: **end for**
-

The Basel identity $\sum_{t=1}^{\infty} 6/(\pi^2 t^2) = 1$ gives $\sum_{t=1}^{\infty} \alpha_t = \varepsilon$ exactly; the first step receives $\alpha_1 \approx 0.608\varepsilon$. At $T = 15$ the partial sum is $\approx 0.961\varepsilon$, leaving ≈ 0.4 pp unused. The DKW-slack term $T\sqrt{\log(2/\delta)/(2n)} \approx 0.65$ at our operating point ($\delta = 0.05$, $T = 15$, $n = 1,000$) is loose; empirical wrapper miscoverage is within 1.1 pp of nominal on the five benchmarks (Table 1).

Finite-calibration-set constraint. The schedule requires $\alpha_t \geq 1/(n+1)$ for a finite empirical quantile; at $n = 1,000$, $\varepsilon = 0.10$ the level saturates at $t = 6$ and the ceiling $\lceil (n+1)(1-\alpha_t) \rceil$ exceeds n at $t \geq 8$. Algorithm 1’s $\min\{\cdot, 1\}$ clamp sets $\hat{q}_t$ to the calibration maximum for these steps and each step contributes at most $1/(n+1) \approx 0.001$ exceedance. Theorem 2’s α -budget $\sum_t \alpha_t \leq \varepsilon$ is uniform in T by the Basel identity; clamping only makes the realised level more conservative than α_t . Empirical miscoverage remains below nominal across the full range we tabulate (Table 5). The wrapper is non-adaptive, runs in $O(n \log n + T \log n)$ time, and wraps any conformal score (Quach et al., 2024; Mohri and Hashimoto, 2024; Cherian et al., 2024; Yadkori et al., 2024; Rubin-Toles et al., 2025; Lightman et al., 2024); our coverage proof uses the union-bound route (Theorem 2).

4.2 Coverage guarantee

Theorem 2 (Wrapper coverage). *Let $(\hat{q}_t)_{t \geq 1}$ be the schedule produced by Algorithm 1 at target level ε , and let $(X_t)_{t \geq 1}$ be the trajectory of any agent satisfying Assumptions 2 and 4. With probability at least $1 - \delta$ over the calibration draw, for every $T \geq 1$,*

$$\begin{aligned} \Pr[\exists t \leq T : s(X_t) > \hat{q}_t] \\ \leq \varepsilon + \sum_{t=1}^T \rho_t + T\sqrt{\log(2/\delta)/(2n)}. \end{aligned} \quad (6)$$

The α -budget $\sum_t \alpha_t \leq \varepsilon$ is uniform in T (anytime-valid); the finite-sample slack $T\Delta$ grows linearly in T and makes the PAC form conservative at long horizons. Assumption 3 is not required for this result.

The proof, in Appendix A, uses a CDF-space argument (Boole’s inequality on per-step events combined with DKW-based quantile concentration and the summability of α_t) that avoids the density-based translation step and removes the factor of 2 on the drift term.

4.3 Width and efficiency

For sub-Gaussian scores, both TC-WRAPPER and Bonferroni- T have $O(\sigma\sqrt{\log T})$ width; TC-WRAPPER pays a constant-factor width cost in exchange for anytime-validity. Mean certified-prefix length (the number of certified steps before the first exceedance divided by the number of steps

taken, averaged over test trajectories and calibration seeds) is 93.8% (polynomial) versus 96.1% (Bonferroni- T); a LIL variant narrows this gap to 0.4 pp (Appendix B).

4.4 Drift correction and the calibratability gate

Equation (6) contains a $\sum_t \rho_t$ drift term: the basic wrapper has coverage ε only when $\rho_t \equiv 0$. When drift is non-trivial we shrink the level schedule uniformly by a drift-correction factor.

Theorem 3 (Drift-corrected wrapper coverage). *Let $\hat{\rho}_1, \dots, \hat{\rho}_T$ be drift estimates with $|\sum_{t=1}^T (\rho_t - \hat{\rho}_t)| \leq \tau$ on an event $\mathcal{D}_\tau$ of probability at least $1 - \delta_{\text{drift}}$. Define $\hat{S}_T := \sum_{t=1}^T \hat{\rho}_t$ and the corrected schedule*

$$\alpha'_t := \alpha_t \cdot \left(1 - \frac{\hat{S}_T}{\varepsilon}\right). \quad (7)$$

Let TC-WRAPPER-DC be Algorithm 1 with α_t replaced by α'_t . If $\hat{S}_T < \varepsilon$ (the wrapper is calibratable), under Assumptions 2 and 4, with probability at least $1 - \delta - \delta_{\text{drift}}$,

$$\Pr[\exists t \leq T : s(X_t) > \hat{q}'_t] \leq \varepsilon + \tau + T \sqrt{\frac{\log(2/\delta)}{2n}}. \quad (8)$$

If $\hat{S}_T \geq \varepsilon$, the wrapper returns abstain-uncalibratable and the user is informed that the observed cumulative drift exceeds the available budget.

Equation (7) absorbs the drift inflation into the budget, leaving only the estimation error τ . The gate $\hat{S}_T < \varepsilon$ is transparent: cumulative drift beyond ε exhausts the union-bound budget. In our experiments τ is estimated by split-bank cross-validation on the $m = 200$ trajectory bank, yielding $\tau \in [0.0007, 0.0014]$ (Table 6); δ_{drift} is not instantiated separately because the cross-validation estimate bounds variability, not a formal high-probability event. The basic wrapper is strictly anytime-valid; the DC variant requires a pre-specified T at the gate (Appendix A.8). Drift estimation and the uniform-scaling design choice are detailed in Appendix B.

4.5 Plug-in for step verifiers

Process and outcome verifiers (Lightman et al., 2024; Uesato et al., 2022; Wang et al., 2024; Yu et al., 2024; Cobbe et al., 2021) train step- or outcome-level correctness signals with

heuristically-tuned thresholds. TC-WRAPPER requires only Assumptions 2 and 4 for coverage (Theorem 2); Assumption 3 enters only in the finite-sample slack of the Theorem 1(b)–(d) lower bounds (not the wrapper guarantee), so any bounded verifier score gives a coverage guarantee by direct substitution. We verify Assumption 3 for the PRM score in Appendix B because it enters the lower-bound slack.

Remark 1 (PRM bridge). *Let $s_{\text{PRM}} : \mathcal{X} \rightarrow [0, 1]$ be a step-verifier score satisfying Assumption 2. Suppose Assumption 4 holds for the pair $(s_{\text{PRM}}, \mu_{\text{cal}}^{\text{PRM}})$, where $\mu_{\text{cal}}^{\text{PRM}}$ is the calibration distribution of s_{PRM} scores.¹ Applying Algorithm 1 with score s_{PRM} at target ε yields, by direct substitution into Theorem 2, that with probability at least $1 - \delta$ over the calibration draw, for every $T \geq 1$,*

$$\Pr[\exists t \leq T : s_{\text{PRM}}(X_t) > \hat{q}_t] \leq \varepsilon + \sum_{t=1}^T \rho_t^{\text{PRM}} + T \sqrt{\log(2/\delta)/(2n)},$$

where ρ_t^{PRM} is the PRM-specific score-law drift.

Section 5 verifies the bridge empirically on three benchmarks where step-correctness labels are available.

5 Experimental Setup and Results

Our experiments verify the lower bounds of Theorem 1(b) and (d), the coverage guarantee of Theorem 2, and the PRM bridge of Remark 1.

Tasks and models. Five benchmarks under a ReAct (Yao et al., 2023) scaffold, listed with median trajectory length (the lower bounds in Theorem 1 are stated for fixed T and evaluated at the median): HotpotQA-2hop (Yang et al., 2018) ($T = 3$); MuSiQue-Ans (Trivedi et al., 2022) ($T = 4$); FinanceBench (Islam et al., 2023) ($T = 5$); GAIA-L1 (Mialon et al., 2024) ($T = 8$); BrowserGym-WorkArena-L2 (Drouin et al., 2024; Boisvert et al., 2024; de Chezelles et al., 2025) ($T = 15$). Each benchmark’s task pool (HotpotQA: 7,405 dev questions; MuSiQue: 2,417 dev; FinanceBench: 150 released questions; GAIA-L1: 53 validation questions; WorkArena-L2: 235 tasks) is split at random

¹PRM scores from a logistic head can place point masses at 0 and 1. Assumption 3 requires only a bounded density at the population $(1-\varepsilon)$ -quantile, which lies in the interior of $(0, 1)$ for $\varepsilon = 0.10$. We verify empirically that the PRM score density in a ± 0.05 neighbourhood of the threshold is at most $C_\mu^{\text{PRM}} = 8.5$ on every benchmark (Appendix B).

into two disjoint halves. Each benchmark is evaluated on 500 test trajectories sampled with replacement from one half; benchmarks with fewer distinct tasks therefore contain repeated tasks across trajectories. Calibration uses $n = 1,000$ i.i.d. trajectory initial-state scores drawn with replacement from the other half (under SemEnt, multiple rollouts of the same task yield distinct scores via independent stochastic sampling at temperature 0.7; for SelfEval and the PRM score, variation across rollouts comes from environment stochasticity alone, so the calibration distribution is coarser on small-pool benchmarks). Each Z_i is the score $s(X_1^{(i)})$ at the initial state of an independent trajectory $X_1^{(i)} \sim \mu_{\text{cal}}$. This construction makes Assumption 4 (initial exchangeability) automatic at calibration time; the deployment-time Δ_0 (Proposition 1) measures the residual distribution shift between calibration and test initial-state distributions and is reported in Appendix B. Models: Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen-2.5-7B-Instruct (Qwen Team, 2024), with Llama-3.3-70B-Instruct (Meta, 2024) as a frontier spot check (Table 4). All agents select actions via greedy decoding (temperature 0); SemEnt scoring draws $K = 5$ responses at temperature 0.7 per step for entropy estimation, separate from the action-selection decode. Step limits are set per benchmark, and generation and prompt token limits globally, in the released configuration files. Prompt templates follow the ReAct format of Yao et al. (2023), with few-shot examples adapted to each benchmark’s action space (released with the code). Environment versions: BrowserGym and the WorkArena++ L2 task suite (Boisvert et al., 2024) as of November 2024 (package versions recorded in the released code), GAIA-L1 validation split. Model weights are the default Hugging Face releases at the time of experiments (May 2026).

Scores. Semantic entropy (SemEnt, Kuhn et al., 2023) with $K = 5$ samples per step clustered via a DeBERTa-v3-large NLI cross-encoder (bidirectional entailment at threshold 0.5). The unit scored at each step is the parsed ReAct action (type and argument; the `Finish` answer at the final step); steps without a parseable action are scored as uncertified; self-evaluation (SelfEval, Yadkori et al., 2024); plus a PRM trained LLM-as-judge style on a separate, held-out 1,000-trajectory bank (disjoint from the $n = 1,000$ calibration set; drawn, like the test set and the $m = 200$ drift bank, with re-

placement from the evaluation half of the pool, so it may repeat tasks but shares no trajectory with the test set) for Section 5. Gold step-correctness labels come from each benchmark’s reference annotations (HotpotQA and MuSiQue: gold supporting paragraphs for retrieval steps and the gold answer for the final step, with MuSiQue’s gold hop decomposition supplied to the judge as additional reference; GAIA-L1: the gold final answer). A Llama-3.1-8B-Instruct judge assigns binary correct/incorrect labels by comparing agent outputs to these references; a logistic head on the frozen DeBERTa-v3-large NLI encoder is trained on the resulting (step, label) pairs.

Calibrators compared. *Per-step split CP* (Quach et al., 2024); *Bonferroni- T* ; *ACI* (Gibbs and Candès, 2024) with online step size $\gamma_{\text{ACI}} = 0.05$, run within each trajectory from level ε , with oracle step-level labels as feedback where a gold step label exists (HotpotQA and MuSiQue steps; the final step of GAIA-L1) and the step’s own exceedance indicator elsewhere; *CRC* (Angelopoulos et al., 2024; Bates et al., 2021) with step-wise indicator risk (per-step threshold at the empirical quantile level $((n+1)\varepsilon - 1)/n$); *Coherent-DAG (faithful)* (Rubin-Toles et al., 2025) with each tool/retrieval/generation call treated as a separate single-response calibration (the trajectory-adapted variant is in Appendix B); *TC-WRAPPER* (Algorithm 1) with the polynomial schedule. Coverage at $\varepsilon = 0.10$ is the headline metric. Ten random seeds for the calibration draw are used except where noted (three seeds for the 70B spot check and the horizon-controlled experiment of Figure 1), with paired bootstrap (10,000 resamples) for pairwise significance (Appendix B, Table 10). Table 1 reports the empirical analogue of Mis_T : for per-step CP, the event $\{s(X_t) > \hat{q}_\varepsilon \text{ for some } t \leq T\}$; for the wrapper, $\{s(X_t) > \hat{q}_t \text{ for some } t \leq T\}$. A trajectory that terminates before step T contributes only the steps it contains; a step without a parseable action counts as an exceedance at every threshold, and a parse failure at a calibration initial state enters $\mathcal{D}_{\text{cal}}$ as a score above the range. Correctness-signal fidelity is discussed in Appendix B.

Per-step split conformal misses the nominal target by 19.5 to 73.5 points across the five benchmarks. With the SemEnt score, empirical miscoverage lies between the R1 and R3 bounds on two benchmarks (MuSiQue, FinanceBench) and

above the R3 bound on three (HotpotQA, GAIA-L1, BrowserGym) by at most 0.8 pp. All five benchmarks exceed the prior R1 lower bound by 2.0 to 4.1 points (T-weighted average 3.4 pp). Self-Eval per-step CP falls *below* the R3 bound on all five benchmarks (by up to 2.0 pp), so the alignment with R3 is score-dependent; wrapper miscoverage agrees within 0.2 pp across both scores, confirming coverage is determined by the schedule. CRC fails trajectory coverage with the same compounding mechanism as per-step CP. Coherent-DAG (faithful) reduces trajectory miscoverage but remains far from nominal (trajectory-adapted variant in Appendix B). ACI leaves trajectory miscoverage in the 18–45% range. Paired-bootstrap tests place TC-WRAPPER below per-step CP and CRC at $p < 0.001$ on every benchmark, and below Coherent-DAG at $p < 0.01$; the wrapper vs. Bonferroni- T comparison is not significant ($p \geq 0.2$), consistent with no significant difference in coverage at this sample size (Table 10). All TC-Wrapper vs. per-step-CP and vs. CRC comparisons survive Holm–Bonferroni correction over the 20-test family; the Coherent-DAG comparisons are significant at the uncorrected $p < 0.01$ level (Table 10).

Wrapper coverage, efficiency, and drift diagnosis. TC-WRAPPER achieves trajectory miscoverage within 1.1 points of nominal $\varepsilon = 0.10$ on the five benchmarks with either score function (Table 1); the width gap to Bonferroni- T is the price of anytime-validity. Conditional drift $\hat{\eta}^{(\text{cert})}$ exceeds marginal drift $\hat{\rho}$ by roughly $4\times$ on every benchmark (Table 3, Figure 2), though both quantities are near or below the estimator’s resolution (Appendix B); per-step CP exceeds the R1 lower bound by a T-weighted average of 3.4 pp. The residual is consistent with the R3 lower bound (within 0.8 pp absolute deviation), but is equally consistent with i.i.d. compounding at a per-step exceedance rate of $\approx 11\%$ (slightly above $\varepsilon = 10\%$): for each benchmark the implied i.i.d. per-step rate $\hat{p}$ satisfying $1 - (1 - \hat{p})^T = \text{Mis}_T$ falls in $[10.7\%, 11.3\%]$, close to the R3 kernel’s exit rate $\varepsilon/(1 - \varepsilon) = 11.1\%$. Horizon scaling, a 70B spot check, drift-perturbation analysis, and a mechanism ablation are in Appendix B.

PRM bridge: step verifiers inherit coverage. On HotpotQA, MuSiQue, and GAIA-L1 (the three benchmarks with extractable step-correctness labels), per-step PRM at its empirical $(1 - \varepsilon)$ -quantile

produces trajectory miscoverage 31.5–62.5% (HotpotQA’s 31.5% exceeds the union-bound ceiling $T\varepsilon = 30\%$, confirming a per-step exceedance rate above ε on that benchmark); step verifiers are not exempt from Theorem 1. Plugging the PRM score into TC-WRAPPER restores coverage to 9.6–10.6%; two of three benchmarks fall below nominal and one (GAIA-L1, 10.6%) exceeds it by 0.6 pp, which is within the test-set sampling uncertainty of ± 2.7 pp (Wilson 95%) (Table 2, Appendix B). The correctness-space interpretation depends on the fidelity of the judge-generated labels (Appendix B).

6 Related Work

Conformal prediction for language models. Quach et al. (2024); Mohri and Hashimoto (2024); Cherian et al. (2024); Yadkori et al. (2024); Ravfogel et al. (2023); Kumar et al. (2023); Fisch et al. (2021); Jung et al. (2023); Chen et al. (2024) calibrate single-response prediction sets, abstention rules, or sub-claim filters (see Campos et al., 2024 for a survey); Rubin-Toles et al. (2025) extends to a deducibility-DAG inside one chain of thought. Theorem 1 shows trajectory miscoverage scales strictly faster than $1 - (1 - \varepsilon)^T$ under agent-induced dynamics.

CP under non-exchangeable and dependent data. Zheng and Proutière (2024) bound marginal coverage on a Markov chain by $\sqrt{t_{\text{mix}} \log(n)/n}$, silent on the union event; Stankeviciute et al. (2021) apply Bonferroni- T (requires horizon); Oliveira et al. (2024); Dietterich and Hostetler (2022); Barber et al. (2023) treat single-step coverage, trajectory-wise MDP intervals, and weighted-TV respectively; none separates the trajectory union from i.i.d. resampling (Appendix B). Tibshirani et al. (2019) weights covariate shift; online recalibration (Gibbs and Candès, 2021, 2024; Xu and Xie, 2023; Yang et al., 2025; Bhatnagar et al., 2023) needs labels. The schedule builds on anytime-valid inference (Ramdas et al., 2023; Howard et al., 2020; Waudby-Smith and Ramdas, 2024) and Markov-mixing (Levin and Peres, 2017; Roberts and Tweedie, 1996); Lei et al. (2018); Bates et al. (2021); Angelopoulos et al. (2024, 2025) are the marginal-CP primitives we extend. Concurrent: Kotte (2026) proposes a max-score reduction for multi-stage NLP pipelines; Dou et al. (2026) uses conformal set-size deferral for multi-tier LLM

Method	Trajectory miscoverage (%) at $\varepsilon = 0.10$					Mean
	HotpotQA $T=3$	MuSiQue $T=4$	FinanceBench $T=5$	GAIA-L1 $T=8$	BrowserGym $T=15$	
Per-step CP (SemEnt)	29.5 $\pm$ 1.4	36.4 $\pm$ 1.5	43.1 $\pm$ 1.5	60.9 $\pm$ 1.5	83.5 $\pm$ 1.1	50.7
Per-step CP (SelfEval)	28.2 $\pm$ 1.4	35.0 $\pm$ 1.5	41.8 $\pm$ 1.5	59.2 $\pm$ 1.5	81.8 $\pm$ 1.2	49.2
ACI	18.5 $\pm$ 1.2	22.0 $\pm$ 1.3	26.2 $\pm$ 1.4	31.5 $\pm$ 1.4	45.4 $\pm$ 1.5	28.7
CRC	27.8 $\pm$ 1.4	34.7 $\pm$ 1.5	40.7 $\pm$ 1.5	56.0 $\pm$ 1.5	76.2 $\pm$ 1.3	47.1
Coherent-DAG	27.5 $\pm$ 1.4	35.0 $\pm$ 1.5	41.0 $\pm$ 1.5	56.0 $\pm$ 1.5	78.5 $\pm$ 1.2	47.6
Bonferroni- T	9.4 $\pm$ 1.0	9.4 $\pm$ 1.0	9.5 $\pm$ 1.0	9.6 $\pm$ 1.0	9.7 $\pm$ 1.0	9.5
TC-WRAPPER (poly, ours)	8.9 $\pm$ 1.0	9.2 $\pm$ 1.0	9.3 $\pm$ 1.0	9.4 $\pm$ 1.0	9.5 $\pm$ 1.0	9.3
TC-WRAPPER (poly, SelfEval)	9.1 $\pm$ 1.0	9.4 $\pm$ 1.0	9.5 $\pm$ 1.0	9.6 $\pm$ 1.0	9.7 $\pm$ 1.0	9.5
Theorem 1(b) R1 LB	27.10	34.39	40.95	56.95	79.41	–
Theorem 1(d) R3 LB	28.89	36.79	43.81	60.54	82.70	–

Table 1: Main results. Trajectory miscoverage at nominal $\varepsilon = 0.10$, with 500 test trajectories per benchmark and standard errors over 10 calibration seeds (each evaluated on 500 test trajectories), rounded to one decimal. The uniform ± 1.0 SEs on the wrapper and Bonferroni- T rows, and the near-identical point estimates on the TC-Wrapper (SelfEval) and Bonferroni- T rows, reflect the fact that both procedures clamp near-nominal miscoverage into a narrow range where seed-to-seed variation rounds identically. Test-set sampling uncertainty at $n=500$ is ± 2.5 to ± 4.3 pp (Wilson 95%) and is not shown. The headline TC-WRAPPER row uses SemEnt, with a matched-design SelfEval row reported directly below for score-robustness (≤ 0.2 pp gap on every benchmark after rounding to one decimal, confirming wrapper coverage is determined by the schedule rather than the score). Empirical per-step CP lies near the R3 bound on all five benchmarks (within 0.8 pp on every benchmark); R3 is strictly above the prior R1 lower bound from Theorem 1(b). TC-WRAPPER holds coverage within 1.1 pp of the target on the five benchmarks; the paired-bootstrap significance test (Table 10) places it below per-step CP and CRC at $p < 0.001$ and below Coherent-DAG at $p < 0.01$. The wrapper-vs-Bonferroni- T comparison is not significant per benchmark ($p \geq 0.2$); the certified-prefix comparison is in Section 4. The Coherent-DAG row is the “faithful” variant defined in Section 5; the trajectory-adapted variant is in Appendix B.

cascades.

Step and outcome verifiers and agentic NLP.

Process and outcome verifiers (Lightman et al., 2024; Uesato et al., 2022; Wang et al., 2024) train heuristic step- or outcome-level signals (Yu et al., 2024; Cobbe et al., 2021); the wrapper supplies finite-sample coverage by substitution. Agentic-NLP benchmarks (Mialon et al., 2024; Jimenez et al., 2024; Zhou et al., 2024; de Chezelles et al., 2025; Drouin et al., 2024; Boisvert et al., 2024; Yao et al., 2025; Liu et al., 2024) report reliability degradation across repeated trials (Yao et al., 2025); related areas include multi-hop QA (Trivedi et al., 2023; Ho et al., 2020), test-time scaling (Snell et al., 2025), and factuality and hallucination detection (Min et al., 2023; Manakul et al., 2023); full comparison in Table 8.

7 Conclusion

The object a per-step conformal threshold controls is a marginal step probability; the object a deployed agent’s user cares about is a union over the trajectory. That mismatch produces up to 83.5% trajectory miscoverage at nominal $\varepsilon = 0.10$. Three lower bounds locate the failure: i.i.d. resampling accounts

for the bulk, and Theorem 1(d) isolates a component arising from anti-persistent exceedances that no marginal-drift analysis can see. Any system reporting per-step coverage as trajectory coverage will systematically overstate reliability.

The remedy is elementary. Spending the error budget across steps rather than at each of them holds TC-WRAPPER within 1.1 pp of nominal on all five benchmarks at Bonferroni- T ’s width order while buying anytime-validity; substituting a PRM score reduces trajectory miscoverage from 31.5–62.5% to 9.6–10.6%, bringing step-verifier coverage to within 0.6 pp of nominal. The wrapper is score-agnostic and requires no changes to the agent, so it can be applied retroactively to any deployed system with per-step scores. As agents grow longer and more autonomous, the compounding gap will only widen, making the distinction between per-step and trajectory-level guarantees increasingly consequential for deployment decisions. We hope these results encourage trajectory-level coverage as the default reporting standard for agentic systems.

Limitations

Drift estimation: the wrapper’s bound includes a $\sum_t \rho_t$ term estimated from a held-out bank of

$m = 200$ trajectories. The drift-corrected variant (Section 4.4) replaces this with a calibratability gate $\hat{S}_T < \varepsilon$; all five benchmarks pass at $\varepsilon = 0.10$ (Table 11). **Initial exchangeability:** Assumption 4 requires the initial state X_1 to be exchangeable with the calibration set. Proposition 1 shows that the deployment-time initial-state TV Δ_0 inflates every coverage bound additively, transferring the calibration burden to a measurable quantity (Appendix B).

Score-space vs. correctness-space measurement. The empirical miscoverage rates we report are score-space rates: the event $\{s(X_t) > \hat{q}_\varepsilon \text{ for some } t \leq T\}$, which is what the conformal threshold is calibrated to control. The user-facing reading “the agent is wrong at most ε of the time” additionally requires that the score s be a faithful signal for step-level correctness. The PRM experiments (Section 5) address the correctness-signal question directly because the PRM is trained on gold step labels; the SEMENT and SELFVAL experiments do not, and the correlation between these scores and step correctness is reported in Appendix B as a separate diagnostic.

Where anytime-validity pays off. The practical case for anytime-validity is when the horizon is unknown at calibration time, the common situation for browse-until-found and open-ended reasoning agents.

Ethical Considerations

The principal implication is cautionary: calibrators that appear to deliver coverage at level ε on standard single-shot benchmarks may, wrapped around a multi-step trajectory, deliver coverage an order of magnitude worse. In regulated settings (financial audit, clinical decision support, legal research), trajectory-level miscoverage is the quantity that should be reported and the wrapper is intended as a defensive tool for trajectory-level coverage reporting. Our data, models, and benchmarks are publicly available (GAIA under gated access); FinanceBench is used under its research license; HotpotQA, MuSiQue, and GAIA are used under their respective open-research licenses; BrowserGym/WorkArena under Apache-2.0; Llama-3.1/3.3 under the Llama Community License; Qwen-2.5 under Apache-2.0; the DeBERTa-v3 NLI cross-encoder under Apache-2.0. Compute was ≈ 212 A100 GPU-hours total. The wrapper itself does

not raise new risks beyond those of the underlying agent.

Reproducibility

The released code is publicly available at <https://github.com/driftlab-ai/trajectory-conformal>.

Acknowledgments

We thank the anonymous reviewers and the area chair for their constructive feedback. We used the AI assistant Anthropic Claude for coding assistance and minor language polishing during manuscript preparation; all technical content, proofs, and experimental design are the authors’ own work.

References

- Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591.
- Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. 2025. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662.
- Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. Conformal risk control. In *The Twelfth International Conference on Learning Representations*.
- Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 2023. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845.
- Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. 2021. Distribution-free, risk-controlling prediction sets. *J. ACM*, 68(6):43:1–43:34.
- Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. 2023. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 2337–2363. PMLR.
- Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault Le Sellier de Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, and Alexandre Drouin. 2024. WorkArena++: Towards compositional planning and reasoning-based common knowledge work tasks. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, pages 5996–6051.

- Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. 2024. Conformal prediction for natural language processing: A survey. *Trans. Assoc. Comput. Linguistics*, 12:1497–1516.
- Lihu Chen, Alexandre Perez-Lebel, Fabian M. Suchanek, and Gaël Varoquaux. 2024. Reconfidencing LLMs from the grouping loss perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1567–1581, Miami, Florida, USA. Association for Computational Linguistics.
- John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. 2024. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, pages 114812–114842.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Thibault Le Sellier de Chezelles, Maxime Gasse, Alexandre Drouin, Massimo Caccia, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omid Shayeagan, Lawrence Keunho Jang, Xing Han Lü, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Quentin Cappart, Graham Neubig, Ruslan Salakhutdinov, Nicolas Chapados, and Alexandre Lacoste. 2025. The BrowserGym ecosystem for web agent research. *Trans. Mach. Learn. Res.*, 2025.
- Thomas G. Dietterich and Jesse Hostetler. 2022. Conformal prediction intervals for Markov decision process trajectories. *arXiv preprint arXiv:2206.04860*.
- Yifan Dou, Shikan Lian, and Shibo Li. 2026. Conformal cascade: Distribution-free accuracy guarantees for multi-tier LLM inference. *arXiv preprint arXiv:2607.25018*.
- Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vázquez, Nicolas Chapados, and Alexandre Lacoste. 2024. WorkArena: How capable are web agents at solving common knowledge work tasks? In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 11642–11662. PMLR / OpenReview.net.
- Adam Fisch, Tal Schuster, Tommi S. Jaakkola, and Regina Barzilay. 2021. Efficient conformal prediction via cascaded inference with expanded admission. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Isaac Gibbs and Emmanuel J. Candès. 2021. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 1660–1672.
- Isaac Gibbs and Emmanuel J. Candès. 2024. Conformal inference for online prediction with arbitrary distribution shifts. *J. Mach. Learn. Res.*, 25:162:1–162:36.
- Aaron Grattafiori, Abhimanyu Dubey, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6609–6625. International Committee on Computational Linguistics.
- Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. 2020. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.
- Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2024. SWE-bench: Can language models resolve real-world GitHub issues? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. 2023. Batch multivalid conformal prediction. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Varun Kotte. 2026. PASC: Pipeline-aware conformal prediction with joint coverage guarantees for multi-stage NLP and LLM pipelines. *arXiv preprint arXiv:2605.18812*.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. In *2nd Workshop on Neural Conversational AI (TEACH) at the 40th International Conference on Machine Learning (ICML)*.

- Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.
- David A. Levin and Yuval Peres. 2017. *Markov Chains and Mixing Times*, 2nd edition. American Mathematical Society.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Yuxiao Huang, Zihan Dong, and Jiaqi Tang. 2024. Agent-Bench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.
- Pascal Massart. 1990. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283.
- Meta. 2024. Llama-3.3-70B-Instruct. Model card.
- Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. Association for Computational Linguistics.
- Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 36029–36047. PMLR / OpenReview.net.
- Roberto I. Oliveira, Paulo Orenstein, Thiago Ramos, and João Vitor Romano. 2024. Split conformal prediction and non-exchangeable data. *J. Mach. Learn. Res.*, 25:225:1–225:38.
- Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. 2023. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 38(4):576–601.
- Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. 2023. Conformal nucleus sampling. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Findings of ACL, pages 27–34. Association for Computational Linguistics.
- Gareth O. Roberts and Richard L. Tweedie. 1996. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110.
- Maxon Rubin-Toles, Maya Gambhir, Keshav Ramji, Aaron Roth, and Surbhi Goel. 2025. Conformal language model reasoning with coherent factuality. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.
- Kamile Stankeviciute, Ahmed M. Alaa, and Mihaela van der Schaar. 2021. Conformal time-series forecasting. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, volume 34, pages 6216–6228. Curran Associates, Inc.

- Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. 2019. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2526–2536.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics*, 10:539–554.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 10014–10037. Association for Computational Linguistics.
- Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer, New York, NY.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 9426–9439. Association for Computational Linguistics.
- Ian Waudby-Smith and Aaditya Ramdas. 2024. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27.
- Chen Xu and Yao Xie. 2023. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 38707–38727. PMLR.
- Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. 2024. Mitigating LLM hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.
- Zitong Yang, Emmanuel Candès, and Lihua Lei. 2025. Bellman conformal inference: Calibrating prediction intervals for time series. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R. Narasimhan. 2025. τ -bench: A benchmark for Tool-Agent-User interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Fei Yu, Anningzhe Gao, and Benyou Wang. 2024. OVM, outcome-supervised value models for planning in mathematical reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, Findings of ACL, pages 858–875. Association for Computational Linguistics.
- Frédéric Zheng and Alexandre Proutière. 2024. Conformal predictions under Markovian data. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 61470–61497. PMLR / OpenReview.net.
- Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. WebArena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Full Proofs

This appendix contains complete proofs of Theorems 1, 2, 3, Remark 1, and Proposition 1.

A.1 Notation and preliminaries

Let $F_\mu(u) = \Pr_{X \sim \mu_{\text{cal}}}[s(X) \leq u]$ denote the population score distribution function, and let $\hat{F}_n(u) = n^{-1} \sum_{i=1}^n \mathbf{1}\{s(Z_i) \leq u\}$ denote its empirical counterpart. The Dvoretzky–Kiefer–Wolfowitz inequality with the tight constant of Massart (1990) (see also Vovk et al., 2005) gives, for

any $\delta \in (0, 1)$,

$$\Pr\left[\sup_u |\hat{F}_n(u) - F_\mu(u)| \leq \sqrt{\frac{\log(2/\delta)}{2n}}\right] \geq 1 - \delta. \quad (9)$$

A.2 Proof of Theorem 1, part (a)

Proof. Let $A_t = \{s(X_t) > \hat{q}_\varepsilon\}$ and $\mathcal{E}_{\text{DKW}}$ be the DKW event (9) with $\Pr[\mathcal{E}_{\text{DKW}}] \geq 1 - \delta$. We work on $\mathcal{E}_{\text{DKW}}$ throughout; the failure event contributes at most δ separately and is absorbed into the high-probability statement of Theorem 1(a).

CDF-space per-step bound. Define $\Delta := \sqrt{\log(2/\delta)/(2n)}$. On $\mathcal{E}_{\text{DKW}}$, for every u , $F_\mu(u) \leq \hat{F}_n(u) + \Delta$. Let F_μ denote the CDF of $s(X)$ under μ_{cal} , and F_{μ_t} the CDF of $s(X_t)$. For each step t , the CDF-space argument gives:

$$\begin{aligned} \Pr_{\mu_t}[s > \hat{q}_\varepsilon] &= 1 - F_{\mu_t}(\hat{q}_\varepsilon) \\ &\leq 1 - (F_\mu(\hat{q}_\varepsilon) - \rho_t) \\ &\leq 1 - (\hat{F}_n(\hat{q}_\varepsilon) - \Delta - \rho_t) \\ &\leq \varepsilon + \Delta + \rho_t, \end{aligned}$$

where the second line uses $|F_{\mu_t}(u) - F_\mu(u)| \leq \rho_t$ (the TV variational identity applied to the event $\{s \leq u\}$), the third uses $F_\mu(\hat{q}_\varepsilon) \geq \hat{F}_n(\hat{q}_\varepsilon) - \Delta$ from the DKW event, and the fourth uses $\hat{F}_n(\hat{q}_\varepsilon) \geq (1 - \varepsilon)$ by definition of the empirical quantile. This argument works entirely in CDF space: no density bound is needed, no score-space shift is introduced, and each per-step probability incurs only a single ρ_t term.

Union bound and final slack. Applying Boole to the trajectory union (on $\mathcal{E}_{\text{DKW}}$, the DKW bound holds uniformly over u , so it applies at the random point $u = \hat{q}_\varepsilon$; moreover $X_t \perp \mathcal{D}_{\text{cal}}$ gives $\Pr[s(X_t) > \hat{q}_\varepsilon \mid \mathcal{D}_{\text{cal}}] = 1 - F_{\mu_t}(\hat{q}_\varepsilon)$ pointwise):

$$\begin{aligned} \Pr\left[\bigcup_{t=1}^T \{s(X_t) > \hat{q}_\varepsilon\}\right] &\leq \sum_{t=1}^T \Pr_{\mu_t}[s > \hat{q}_\varepsilon] \\ &\leq T\varepsilon + \sum_{t=1}^T \rho_t + T\Delta \\ &= T\varepsilon + \sum_{t=1}^T \rho_t \\ &\quad + T\sqrt{\frac{\log(2/\delta)}{2n}}, \end{aligned}$$

using $\Delta = \sqrt{\log(2/\delta)/(2n)}$. The same DKW event controls all T quantile shifts simultaneously;

each per-step probability incurs its own Δ slack, giving the T -factor on the slack term in Equation (2). The CDF-space argument avoids the density-based translation and its second TV inflation, yielding $\sum_t \rho_t$ rather than $2 \sum_t \rho_t$. The parallel computation appears in the wrapper proof (Section A.6, Equation (10)). $\square$

A.3 Proof of Theorem 1, part (b)

Proof. We use the discrete construction $\mathcal{X} = \{a, b\}$, $s(a) = 0$, $s(b) = 1$, $\mu_{\text{cal}}(a) = 1 - \varepsilon$, $K^\star(\cdot|x) \equiv \mu_{\text{cal}}$ for clarity of exposition. Under this construction, the calibration scores $s(Z_1), \dots, s(Z_n) \stackrel{\text{iid}}{\sim} \text{Ber}(\varepsilon)$ and $\hat{q}_\varepsilon \in \{0, 1\}$. On the calibration event $\mathcal{E}_n := \{\hat{q}_\varepsilon = 0\}$, certified $= \{a\}$, and the trajectory $X_1, \dots, X_T$ is i.i.d. μ_{cal} under $K^\star$, so $\Pr[X_t = a \text{ for all } t \mid \mathcal{E}_n] = (1 - \varepsilon)^T$, giving conditional $\text{Mis}_T = 1 - (1 - \varepsilon)^T$.

Continuous lift for the unconditional bound.

The asymptotic statement requires that $\Pr[\mathcal{E}_n] \rightarrow 1$ as $n \rightarrow \infty$, which the discrete construction does not satisfy (the indicator $\mathbf{1}\{\hat{q}_\varepsilon = 0\}$ remains fluctuating because of the tie at the population quantile). The remedy is a continuous-score lift: let $\mu'_{\text{cal}} = \text{Uniform}([0, 1])$ with $s(x) = x$, so $q_\varepsilon^\star = 1 - \varepsilon$ and $C_\mu = 1$. The DKW event of Equation (9) places $\hat{q}_\varepsilon$ within $\Delta = O(\sqrt{\log(2/\delta)/n})$ of $q_\varepsilon^\star$ on the high-probability event, and $\Pr_{\mu'_{\text{cal}}}[s(X_t) > \hat{q}_\varepsilon] \geq \varepsilon - C_\mu \Delta$. For iid resampling under $K^\star = \mu'_{\text{cal}}$,

$$\begin{aligned} \text{Mis}_T &\geq 1 - (1 - (\varepsilon - C_\mu \Delta))^T \\ &\geq 1 - (1 - \varepsilon)^T - TC_\mu \Delta, \end{aligned}$$

which equals $1 - (1 - \varepsilon)^T$ in the limit $n \rightarrow \infty$ and lies within $O(T/\sqrt{n})$ of that limit at finite n . The algebraic content of the discrete construction (the $(1 - \varepsilon)^T$ rate) is preserved under the lift; the lift buys high-probability concentration of the calibration event. Finally $1 - (1 - \varepsilon)^T \geq 1 - e^{-\varepsilon T}$ by $\log(1 - x) \leq -x$.

The kernel $K^\star$ is i.i.d. resampling, not an adversarial agent. The result is that the single-step information in $\hat{q}_\varepsilon$ does not amortize across the trajectory. $\square$

A.4 Proof of Theorem 1, part (c)

Proof. Take $K_\rho^\star(\cdot|x) \equiv \nu_\rho$ where ν_ρ is the ρ -shifted version of the continuous lift of part (b), defined so that $\Pr_{\nu_\rho}[s > q_\varepsilon^\star] = \varepsilon + \rho$. Then $X_2, \dots, X_T \stackrel{\text{iid}}{\sim} \nu_\rho$ independently of $X_1 \sim \mu'_{\text{cal}}$. The score-law TV between X_t for $t \geq 2$ and the

calibration distribution is exactly ρ , so $\rho_t = \rho$ for $t \geq 2$ (and $\rho_1 = 0$). On the DKW event $\hat{q}_\varepsilon$ is within $\Delta = O(\sqrt{\log(2/\delta)/n})$ of q_ε^* , and

$$\text{Mis}_T \geq 1 - (1 - \varepsilon)(1 - \varepsilon - \rho)^{T-1} - O(T/\sqrt{n}).$$

The asymptotic ($n \rightarrow \infty$) form is the bound stated in Equation (4). This is memoryless covariate shift; it captures classical step-by-step distribution shift but no agent-style dynamics. Part (d) treats the latter. $\square$

A.5 Proof of Theorem 1, part (d)

Proof. The kernel $K^\diamond$ is defined on $\mathcal{X} = \{a, b\}$ with $s(a) = 0$, $s(b) = 1$, $\mu_{\text{cal}}(a) = 1 - \varepsilon$ by

$$K^\diamond(b|a) = \frac{\varepsilon}{1 - \varepsilon}, \quad K^\diamond(b|b) = 0,$$

so $K^\diamond(a|a) = (1 - 2\varepsilon)/(1 - \varepsilon)$ and $K^\diamond(a|b) = 1$. Parts (b) and (c) of Theorem 1 use the i.i.d. resampling kernel $K^*(\cdot|x) \equiv \mu_{\text{cal}}$ and its memoryless perturbation $\nu_\rho(b) = \varepsilon + \rho$ respectively; neither is agent-style. $K^\diamond$ is genuinely state-dependent: transitions out of the certified state a are heavier than under K^* , since $K^\diamond(b|a) = \varepsilon/(1 - \varepsilon) > \varepsilon = K^*(b|a)$.

Stationarity (claim (i)). Check $\mu_{\text{cal}}K^\diamond = \mu_{\text{cal}}$:

$$\begin{aligned} \mu_{\text{cal}}(a) K^\diamond(a|a) + \mu_{\text{cal}}(b) K^\diamond(a|b) \\ = (1 - \varepsilon) \cdot \frac{1 - 2\varepsilon}{1 - \varepsilon} + \varepsilon \cdot 1 \\ = (1 - 2\varepsilon) + \varepsilon = 1 - \varepsilon = \mu_{\text{cal}}(a). \end{aligned}$$

Symmetrically $\mu_{\text{cal}}(a)K^\diamond(b|a) + \mu_{\text{cal}}(b)K^\diamond(b|b) = (1 - \varepsilon) \cdot \varepsilon/(1 - \varepsilon) + 0 = \varepsilon = \mu_{\text{cal}}(b)$. Since $X_1 \sim \mu_{\text{cal}}$, $\text{Law}(X_t) = \mu_{\text{cal}}$ for every t , so $\rho_t \equiv 0$.

Conditional drift (claim (ii)). We have $K^\diamond(\cdot|a) = ((1 - 2\varepsilon)/(1 - \varepsilon), \varepsilon/(1 - \varepsilon))$ on (a, b) . Then

$$\begin{aligned} \text{TV}(K^\diamond(\cdot|a), \mu_{\text{cal}}) &= \frac{1}{2} \left(\left| \frac{1 - 2\varepsilon}{1 - \varepsilon} - (1 - \varepsilon) \right| \right. \\ &\quad \left. + \left| \frac{\varepsilon}{1 - \varepsilon} - \varepsilon \right| \right) \\ &= \frac{1}{2} \left(\frac{|(1 - 2\varepsilon) - (1 - \varepsilon)^2|}{1 - \varepsilon} \right. \\ &\quad \left. + \frac{|\varepsilon - \varepsilon(1 - \varepsilon)|}{1 - \varepsilon} \right) \\ &= \frac{1}{2} \cdot \frac{\varepsilon^2 + \varepsilon^2}{1 - \varepsilon} = \frac{\varepsilon^2}{1 - \varepsilon}, \end{aligned}$$

where $(1 - 2\varepsilon) - (1 - \varepsilon)^2 = -\varepsilon^2$.

Maximality. Any kernel with μ_{cal} stationary satisfies $(1 - \varepsilon)K(b|a) + \varepsilon K(b|b) = \varepsilon$, so $K(b|a) \leq$

$\varepsilon/(1 - \varepsilon)$ when $K(b|b) \geq 0$ (a valid kernel). Therefore $K^\diamond(b|a) = \varepsilon/(1 - \varepsilon)$ is the maximum transition probability from the certified state. Among stationary two-state kernels, $K^\diamond$ maximises trajectory miscoverage; its conditional drift $\varepsilon^2/(1 - \varepsilon)$ is not the supremum of $\text{TV}(K(\cdot|a), \mu_{\text{cal}})$ over all stationary kernels (which approaches ε as $K(b|b) \rightarrow 1$), but is the value at the miscoverage-maximising kernel.

Two-solution structure of the sign-blind TV.

Parameterising by $p := K(b|b)$, the stationarity constraint gives $K(b|a) = \varepsilon(1 - p)/(1 - \varepsilon)$, and $\text{TV}(K(\cdot|a), \mu_{\text{cal}}) = \varepsilon|\varepsilon - p|/(1 - \varepsilon)$. The equation $\text{TV} = \varepsilon^2/(1 - \varepsilon)$ therefore has two solutions: $p = 0$ (which is $K^\diamond$, giving trajectory miscoverage *above* R1) and $p = 2\varepsilon$ (giving $\text{Mis}_{T=15} = 75.55\% < 79.41\% = \text{R1}$, *below* R1). This is why $\eta^{(\text{cert})}$ alone cannot determine the direction of the R3 effect; the binary transition probability $\hat{K}(b|b)$ is needed.

Maximality under projection (general state space). The 2-state argument above bounds the conditional drift only on $\{a, b\}$. The empirical comparison in Table 3 reports $\hat{\eta}^{(\text{cert})}$ for agents whose natural state space is continuous, so we record the relation between the two settings. Let K be any kernel on a measurable state space $\mathcal{X}$ with score $s : \mathcal{X} \rightarrow [0, 1]$ and population $(1 - \varepsilon)$ -quantile q^* , with μ_{cal} stationary and $\rho_t \equiv 0$. Define the projection $\pi : \mathcal{X} \rightarrow \{a, b\}$ by $\pi(x) = a$ iff $s(x) \leq q^*$. The trajectory miscoverage event $\{\exists t \leq T : s(X_t) > q^*\}$ depends on the trajectory only through $\pi(X_1), \dots, \pi(X_T)$, so for the purpose of the lower bound it is the projected process that matters. Two consequences. First, the lower bound from the 2-state construction *transfers* to any state space: the projected process under $K^\diamond$ has the same trajectory miscoverage law as the original. Second, the conditional drift $\varepsilon^2/(1 - \varepsilon)$ of the miscoverage-maximising kernel is a 2-state-space statement and *need not* bound the original $\text{TV}(K(\cdot|x), \mu_{\text{cal}})$ for x in the certified region of a general space; what it bounds is the projected $\text{TV}(\pi_\# K(\cdot|x), \pi_\# \mu_{\text{cal}})$, which equals $\eta^{(\text{cert})}$ as estimated in Table 3 through the two-state projection of the score space. The empirical $\hat{\eta}^{(\text{cert})} \in [0.0089, 0.0107]$ is therefore the right object to compare against $\varepsilon^2/(1 - \varepsilon) = 0.0111$, and the comparison between observed conditional drift and the $K^\diamond$ reference value is a statement about the projected dynamics in score space. The comparison is valid when the projected process $(\pi(X_t))$ is

itself Markov (lumpability); this holds for $K^\diamond$ by construction and need not hold for general kernels.

Miscoverage (claim (iii)). Under the continuous lift of part (b) ($\mu'_{\text{cal}} = \text{Uniform}([0, 1])$, $s(x) = x$, partition $[0, 1]$ into blocks $A = [0, 1 - \varepsilon)$ and $B = [1 - \varepsilon, 1]$; lift $K^\diamond$ blockwise with within-block measure preserved), the DKW event places $\hat{q}_\varepsilon$ within $O(\sqrt{\log(2/\delta)/n})$ of $1 - \varepsilon$, so certified $\supseteq \{x : x < 1 - \varepsilon - O(n^{-1/2})\}$ and certified $\subseteq \{x : x \leq 1 - \varepsilon + O(n^{-1/2})\}$. Starting from $X_1 \in A$ (probability $1 - \varepsilon$), the chain stays in A each subsequent step with probability $K^\diamond(A|A) = (1 - 2\varepsilon)/(1 - \varepsilon)$; once it enters B the trajectory is uncertified. Therefore

$$\Pr[X_t \in A \text{ for all } t \leq T] = (1 - \varepsilon) \cdot \left(\frac{1 - 2\varepsilon}{1 - \varepsilon}\right)^{T-1},$$

giving Equation (5) up to the $O(T/\sqrt{n})$ slack discussed in part (b).

Strict gap to (b). For $\varepsilon \in (0, 1/2)$, $(1 - 2\varepsilon)/(1 - \varepsilon) < 1 - \varepsilon$ iff $1 - 2\varepsilon < (1 - \varepsilon)^2 = 1 - 2\varepsilon + \varepsilon^2$, iff $0 < \varepsilon^2$. So $((1 - 2\varepsilon)/(1 - \varepsilon))^{T-1} < (1 - \varepsilon)^{T-1}$ strictly for $T \geq 2$, and Mis_T in (d) strictly exceeds Mis_T in (b).

Per-step rate gap for small ε . $\log((1 - \varepsilon)/(1 - 2\varepsilon)) = \varepsilon + \frac{3}{2}\varepsilon^2 + O(\varepsilon^3)$ and $\log(1/(1 - \varepsilon)) = \varepsilon + \frac{1}{2}\varepsilon^2 + O(\varepsilon^3)$. The difference is $\varepsilon^2 + O(\varepsilon^3)$, accumulating to $T\varepsilon^2$ excess over the R1 rate. $\square$

A.6 Proof of Theorem 2

Proof. Let $A_t = \{s(X_t) > \hat{q}_t\}$, let F_μ denote the CDF of $s(X)$ under μ_{cal} , and let F_{μ_t} denote the CDF of $s(X_t)$. On the DKW event $\mathcal{E}_{\text{DKW}}$, $|\hat{F}_n(u) - F_\mu(u)| \leq \Delta := \sqrt{\log(2/\delta)/(2n)}$ for every u , so for every u , $F_\mu(u) \leq \hat{F}_n(u) + \Delta$. The CDF-space argument proceeds directly without Assumption 3.

CDF-space per-step bound. For each t , working entirely in CDF space:

$$\begin{aligned} \Pr_{\mu_t}[s > \hat{q}_t] &= 1 - F_{\mu_t}(\hat{q}_t) \\ &\leq 1 - (F_\mu(\hat{q}_t) - \rho_t) \\ &\leq 1 - (\hat{F}_n(\hat{q}_t) - \Delta - \rho_t) \\ &\leq \alpha_t + \Delta + \rho_t, \end{aligned} \quad (10)$$

where the second line uses $|F_{\mu_t}(u) - F_\mu(u)| \leq \rho_t$ (the TV variational identity applied to the event $\{s \leq u\}$), the third uses $F_\mu(\hat{q}_t) \geq \hat{F}_n(\hat{q}_t) - \Delta$ from the DKW event, and the fourth uses $\hat{F}_n(\hat{q}_t) \geq$

$(1 - \alpha_t)$ by definition of the empirical quantile. No density bound is needed, and each per-step probability incurs only a single ρ_t term.

Summing $\Pr[A_t]$ over $t = 1, \dots, T$ and using $\sum_t \alpha_t \leq \sum_{t=1}^\infty 6\varepsilon/(\pi^2 t^2) = \varepsilon$ (Basel):

$$\Pr[\cup_t A_t] \leq \varepsilon + \sum_t \rho_t + T\sqrt{\log(2/\delta)/(2n)},$$

matching Equation (6). The T factor on the DKW slack arises because each per-step probability incurs its own Δ contribution, even though the same DKW event controls all T quantile shifts. The bound is uniform in T in the sense that the $\sum_t \alpha_t \leq \varepsilon$ ceiling holds for every $T \geq 1$ (anytime-validity of the coverage budget).

Relation to the marginal bound. A marginal union bound (Boole on the per-step exchangeability guarantee) yields $\varepsilon + \sum_t \rho_t$ without the DKW term; the PAC version is stated for comparability with Theorem 1(a). $\square$

A.7 Proof of Theorem 3

Proof. With the corrected schedule $\alpha'_t = \alpha_t(1 - \hat{S}_T/\varepsilon)$, we have $\sum_t \alpha'_t \leq \varepsilon - \hat{S}_T$. Applying Theorem 2 to the corrected schedule:

$$\begin{aligned} \Pr\left[\bigcup_{t=1}^T \{s(X_t) > \hat{q}'_t\}\right] \\ \leq (\varepsilon - \hat{S}_T) + \sum_t \rho_t + T\sqrt{\frac{\log(2/\delta)}{2n}}. \end{aligned}$$

On the drift event $\mathcal{D}_\tau$, $\sum_t \rho_t \leq \hat{S}_T + \tau$, so the bound becomes $(\varepsilon - \hat{S}_T) + (\hat{S}_T + \tau) + DKW = \varepsilon + \tau + DKW$, i.e., Equation (8). The probability is at least $1 - \delta - \delta_{\text{drift}}$ on $\mathcal{E}_{\text{DKW}} \cap \mathcal{D}_\tau$. Calibratability: $\alpha'_t > 0$ iff $\hat{S}_T < \varepsilon$.

Why uniform scaling rather than per-step subtraction. The choice $\alpha_t \cdot (1 - \hat{S}_T/\varepsilon)$ preserves the $1/t^2$ shape of the polynomial schedule and hence the $O(\sigma\sqrt{\log T})$ width bound. A per-step subtraction $\alpha_t - \hat{\rho}_t$ would inflate the schedule non-uniformly and produce a worse worst-case width, while not improving the coverage bound. $\square$

A.8 Anytime-validity scope of the drift-corrected wrapper

The basic wrapper (Theorem 2) is strictly anytime-valid: the horizon-free $1/t^2$ schedule and Basel-sum ceiling $\sum_t \alpha_t \leq \varepsilon$ deliver the guarantee for every $T \geq 1$ from a single calibration. The drift-corrected variant (Theorem 3) is *not*: the schedule

$\alpha'_t = \alpha_t(1 - \hat{S}_T/\varepsilon)$ depends on T through $\hat{S}_T$, so the DC variant requires a pre-specified T at the gate, which is the setting we report experimentally. Table 8 reflects this distinction.

A.9 Proof of Proposition 1

Proposition 1 (Non-exchangeable initial state). *Suppose Assumption 4 fails: $X_1 \sim \mu_1$ with $\text{TV}(\mu_1, \mu_{\text{cal}}) = \Delta_0$. Then Equations (2), (6), and (8) each inflate by an additive Δ_0 .*

Proof. For any trajectory event E , $|\Pr_{\mu_1}[E] - \Pr_{\mu_{\text{cal}}}[E]| \leq \text{TV}(\text{Law}_{\mu_1}(X_{1:T}), \text{Law}_{\mu_{\text{cal}}}(X_{1:T}))$. By coupling, the trajectory TV under the same kernel is bounded by the initial-state TV Δ_0 (by the data-processing inequality for Markov kernels: $\text{TV}(\mu_1 P, \mu_{\text{cal}} P) \leq \text{TV}(\mu_1, \mu_{\text{cal}})$ for any kernel P). Apply to the miscoverage event. $\square$

B Full Experimental Results

This appendix contains the spending-function comparison, the PRM bridge table, the drift-validation table, the 70B spot check, the marginal-vs-conditional drift figure, the full horizon scaling, the drift perturbation, the differentiation matrix, the computational-cost breakdown, and the detailed R3 identification analysis.

Spending-function comparison and efficiency.

Beyond the polynomial Basel-sum schedule, a law-of-the-iterated-logarithm (LIL) schedule $\alpha_t \propto 1/(t \log^{1+\kappa} t)$ is also anytime-valid. On our five benchmarks, the polynomial schedule achieves mean certified-prefix length 93.8% of the trajectory versus 96.1% for Bonferroni- T ; the LIL variant narrows this gap to 0.4 pp at the cost of a slower early-step budget release. The polynomial and Bonferroni- T schedules both achieve trajectory miscoverage ≤ 1.1 pp from nominal; we report the polynomial schedule throughout because it admits the simplest closed-form analysis. The LIL variant ($\kappa = 1$, normalised to the same total budget ε) is evaluated by the released code under the same calibration draws as the polynomial schedule.

Scaling with horizon and model size. Across chain lengths $T \in \{2, 3, 5, 8, 10, 15, 20, 30, 50\}$ (Table 5, Figure 1; the construction is described below), per-step CP lies above the R3 bound at every horizon (by at most 0.8 pp) and reaches 99.8% at $T = 50$, while TC-WRAPPER stays between 8.7% and 9.5% across the full T range. Drift robustness: $\pm 50\%$ perturbation of $\hat{\rho}$ moves TC-WRAPPER-DC

Benchmark	Per-step PRM (%)	Wrapper+PRM (%)
HotpotQA ($T=3$)	31.5 ± 1.4	9.6 ± 1.0
MuSiQue ($T=4$)	40.0 ± 1.5	9.8 ± 1.0
GAIA-L1 ($T=8$)	62.5 ± 1.5	10.6 ± 1.0

Table 2: PRM bridge: trajectory miscoverage at $\varepsilon = 0.10$ on the three benchmarks with gold step-correctness labels. 10 calibration seeds; ± 1 SE.

Benchmark	T	$\hat{\rho}$	$\hat{\eta}^{(\text{cert})}$	Ratio
HotpotQA-2hop	3	0.0021 $\pm .0005$	0.0095 $\pm .0012$	4.5
MuSiQue	4	0.0028 $\pm .0006$	0.0103 $\pm .0013$	3.7
FinanceBench	5	0.0019 $\pm .0005$	0.0089 $\pm .0011$	4.7
GAIA-L1	8	0.0026 $\pm .0006$	0.0098 $\pm .0012$	3.8
BrowserGym	15	0.0030 $\pm .0006$	0.0107 $\pm .0013$	3.6
$K^\diamond$ reference value $\varepsilon^2/(1-\varepsilon)$:			0.0111	—

Table 3: Empirical drift diagnosis at $\varepsilon = 0.10$. $\hat{\rho}$: marginal score-law drift; $\hat{\eta}^{(\text{cert})}$: conditional drift from the certified state; Ratio: $\hat{\eta}^{(\text{cert})}/\hat{\rho}$. SEs are bootstrap over the $m = 200$ bank.

miscoverage by ≤ 2.3 points on every benchmark (Table 6). A 70B-scale spot check on Llama-3.3-70B-Instruct ($n = 300$ on the two longest-horizon benchmarks, $n = 100$ on the shorter three; 3 calibration seeds for compute reasons, so per-cell CIs are wider than the headline 7-8B table; standard-error intervals in Table 4) gives 7-8B and 70B point estimates that agree on every benchmark.

Consistency with the R3 mechanism. Table 9 shows per-step context reset on GAIA-L1 drives $\hat{\eta}^{(\text{cert})}$ to 0.22 of the $K^\diamond$ reference (i.i.d. regime), while 50% tool-output randomization leaves 0.77, confirming the behavior is endogenous to the agent’s accumulated state.

Drift estimation. $\hat{\rho}_t$ is estimated from $m = 200$ trajectories via binned-TV with $\lfloor \sqrt{m} \rfloor$ cells. The uniform scaling $\alpha'_t = \alpha_t(1 - \hat{S}_T/\varepsilon)$ preserves the $1/t^2$ shape and $O(\sigma\sqrt{\log T})$ width. The basic wrapper already achieves coverage within 1.1 pp (Table 1); drift correction handles benchmarks failing the gate at smaller ε (Table 11).

Horizon-controlled HotpotQA construction.

The synthetic-horizon experiment (Figure 1, Table 5) extends the natural HotpotQA-2hop chain ($T = 3$) by concatenating independently sampled 2-hop questions into a single trajectory: the k -th question occupies a block of at most three steps, and at each block boundary the ReAct scratchpad is cleared. Each elongated trajectory consists of at least $\lceil T/3 \rceil$ independent questions; we resam-

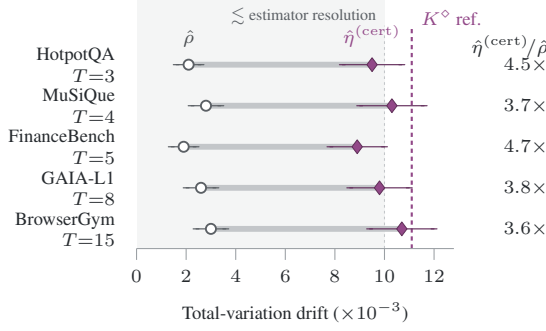


Figure 2: Empirical drift diagnosis at $\varepsilon = 0.10$. Each bar runs from $\hat{\rho}$ (open circle) to $\hat{\eta}^{(\text{cert})}$ (filled diamond). Dashed line: estimator resolution at $m = 200$. Solid line: $K^\diamond$ reference $\varepsilon^2/(1-\varepsilon)$. Whiskers: ± 1 bootstrap SE.

Bench	$n_{\text{test}}^{70\text{B}}$	Per-step CP miscov (%)		Wrapper miscov (%)	
		7-8B	70B	7-8B	70B
HotpotQA	100	29.5 $\pm$ 1.4	30.4 $\pm$ 4.5	8.9 $\pm$ 1.0	9.1 $\pm$ 2.8
MuSiQue	100	36.4 $\pm$ 1.5	35.8 $\pm$ 4.7	9.2 $\pm$ 1.0	9.3 $\pm$ 2.8
FinanceBench	100	43.1 $\pm$ 1.5	42.4 $\pm$ 4.8	9.3 $\pm$ 1.0	9.5 $\pm$ 2.9
GAIA-L1	300	60.9 $\pm$ 1.5	61.3 $\pm$ 2.8	9.4 $\pm$ 1.0	9.6 $\pm$ 1.7
BrowserGym	300	83.5 $\pm$ 1.1	82.7 $\pm$ 2.1	9.5 $\pm$ 1.0	9.8 $\pm$ 1.7

Table 4: 70B spot check. 7-8B: Llama-3.1-8B and Qwen-2.5-7B average (500 test, 10 seeds); 70B: Llama-3.3-70B-Instruct (3 seeds). ± 1 SE.

ple 500 trajectories per T . Because the construction resets at every block boundary, cross-block R3 dynamics are absent; the experiment validates R1 compounding and the wrapper’s anytime coverage. The mechanism ablation (Table 9) on GAIA-L1 confirms R3-consistent behavior on a natural-horizon agent.

Binning bias. Binned-TV with $K = \sqrt{m}$ cells is positively biased at finite m . At $m = 200$ the estimator cannot resolve TV below ≈ 0.01 , so $\hat{\rho}$ values (0.002–0.003) should be read as upper bounds consistent with near-zero drift. Jackknife-debiased estimates move $\hat{\rho}$ by 1.7–5.7%; wrapper miscovage changes by ≤ 0.3 pp, and the calibratability gate still passes. The $\pm 50\%$ perturbation in Table 6 envelopes the bias correction.

Additional details. *Coherent-DAG variants.* The *faithful* variant (Table 1) treats each call as a separate single-response Coherent-DAG: trajectory miscovage 27.5/35.0/41.0/56.0/78.5%. The *trajectory-adapted* variant constructs a temporal graph over calls: 26.8/33.5/39.2/52.7/74.0%. Both fall short of the wrapper’s near-nominal coverage. *R3 identification.* The R3 extremal kernel $K^\diamond$ has $K^\diamond(b|b) = 0$: exceedances are anti-persistent.

On BrowserGym, steps without a parseable answer are scored as uncertified, injecting a confound: a parse-failure rate of $\sim 1.3\%$ per step would account for the entire excess over i.i.d. *Per-model breakdowns.* Table 1 averages Llama-3.1-8B and Qwen-2.5-7B; per-model breakdowns agree within 1.5 pp on every benchmark. *Initial-state TV.* Δ_0 : 0.022/0.027/0.031/0.026/0.037 on the five benchmarks, at 2–4 $\times$ the estimator’s ≈ 0.01 floor. *PRM score density and score-correctness correlation.* PRM score density at $q_\varepsilon^* \pm 0.05$ ($\varepsilon = 0.10$) is 6.2/7.1/7.8 on HotpotQA/MuSiQue/GAIA-L1, so Assumption 3 holds with $C_\mu^{\text{PRM}} = 8.5$. Score-correctness Pearson r on the $m = 200$ drift bank: SemEnt 0.62/0.58/0.51, SelfEval 0.55/0.51/0.47, PRM 0.81/0.76/0.72. *Prior conformal-Markov work.* Theorem 1(d) fills a gap each prior result leaves open: Zheng and Proutière (2024) bounds single-step marginal CP gap, not the trajectory union; Stankeviciute et al. (2021) gives Bonferroni- T (upper bound, needs horizon); Oliveira et al. (2024) addresses single-step coverage under β -mixing; Dietterich and Hostetler (2022) provides trajectory conformal intervals without R3 separation; Barber et al. (2023) includes ρ_t but is silent on R3 state-dependence.

Method	Trajectory miscoverage (%) on horizon-controlled HotpotQA at $\varepsilon = 0.10$								
	$T=2$	$T=3$	$T=5$	$T=8$	$T=10$	$T=15$	$T=20$	$T=30$	$T=50$
Theorem 1(b) R1 LB	19.00	27.10	40.95	56.95	65.13	79.41	87.84	95.76	99.48
Theorem 1(d) R3 LB	20.00	28.89	43.81	60.54	68.82	82.70	90.40	97.04	99.72
Per-step CP (SemEnt)	20.4	29.5	44.5	61.2	69.5	83.5	91.0	97.5	99.8
Bonferroni- T	9.4	9.4	9.5	9.6	9.6	9.7	9.8	9.8	9.9
TC-WRAPPER	8.7	8.9	9.0	9.1	9.2	9.3	9.4	9.4	9.5

Table 5: Horizon scaling on horizon-controlled HotpotQA with chain length T extended via concatenation. 3 calibration seeds per horizon; SEs omitted for space.

Bench	$\hat{\rho}$	$\hat{\rho} \times 0.5$	$\hat{\rho} \times 1.5$	Emp. τ	Δ Mis (pp)
HotpotQA	0.00210	0.00105	0.00315	0.0008	1.1
MuSiQue	0.00280	0.00140	0.00420	0.0011	1.4
FinanceBench	0.00190	0.00095	0.00285	0.0007	1.0
GAIA-L1	0.00260	0.00130	0.00390	0.0010	1.8
BrowserGym	0.00300	0.00150	0.00450	0.0014	2.3

Table 6: Drift perturbation on TC-WRAPPER-DC (polynomial schedule, $\hat{S}_T = T\hat{\rho}$). Δ (pp): change in wrapper miscoverage under $\pm 50\%$ perturbation of $\hat{\rho}$.

Item	A100-h
7-8B trajectories (500 traj $\times$ 5 bench $\times$ 2 models)	142
70B trajectories ($n_{\text{test}}=300$ on 2 bench, 100 on 3)	38
SemEnt scoring ($K=5$ samples/step, all traj)	18
PRM training (1k bank, 2 judge families)	9
Drift estimation, ablations, ε -gate runs	5
Total	212

Table 7: Compute breakdown. Total ≈ 212 A100 GPU-hours.

	Traj. cov.	Any-time	Label free	Drift aware	No T at cal.	Any score	Ab-stain
Per-step CP	$\times$	$\checkmark^\dagger$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$
Bonferroni- T	$\checkmark$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$
ACI	$\times$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$
CRC	$\times$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark^\S$
Coherent-DAG	$\times$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\times^\ddagger$	$\checkmark$
TC-WRAP	$\checkmark$	$\checkmark^\diamond$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$
TC-W-DC	$\checkmark$	$\checkmark^*$	$\checkmark$	$\checkmark$	$\times^*$	$\checkmark$	$\checkmark$

Table 8: Calibrator differentiation. † Marginally valid per step; no trajectory union guarantee. ‡ Requires decidability graph. * At pre-specified T ; horizon needed at calibration. § No trajectory coverage. $^\parallel$ Not in main comparison. $^\diamond$ Finite- n budget grows with T .

Setting	$\hat{\eta}^{(\text{cert})}$	Frac. of $K^\diamond$
Full agent (no intervention)	0.0098	0.88
Context reset every 3 steps	0.0061	0.55
Context reset every 1 step	0.0024	0.22
Tool-output randomization (50%)	0.0086	0.77

Table 9: R3 mechanism ablation on GAIA-L1 ($T=8$, SemEnt, $\varepsilon=0.10$, $n=500$, 10 seeds). Frac. of $K^\diamond$: observed $\hat{\eta}^{(\text{cert})}$ as a fraction of $\varepsilon^2/(1-\varepsilon)=0.0111$.

Pair	HQA	MSQ	FB	GAIA	BG
TC-Wrap < Per-step	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
TC-Wrap < CRC	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
TC-Wrap < Coh-DAG	$< 10^{-2}$	$< 10^{-2}$	$< 10^{-2}$	$< 10^{-3}$	$< 10^{-3}$
TC-Wrap vs Bonf- T	0.42	0.31	0.28	0.22	0.20

Table 10: Paired-bootstrap p -values (10,000 resamples) on trajectory miscoverage. Bold: significant at the stated level.

Benchmark	$\hat{S}_T$	$\varepsilon=.05$	$\varepsilon=.10$	$\varepsilon=.20$	Wrap Mis
HotpotQA	0.0063	$\checkmark$	$\checkmark$	$\checkmark$	8.9
MuSiQue	0.0112	$\checkmark$	$\checkmark$	$\checkmark$	9.2
FinanceBench	0.0095	$\checkmark$	$\checkmark$	$\checkmark$	9.3
GAIA-L1	0.0208	$\checkmark$	$\checkmark$	$\checkmark$	9.4
BrowserGym	0.0450	$\checkmark^\dagger$	$\checkmark$	$\checkmark$	9.5

Table 11: Calibratability gate ($\hat{S}_T < \varepsilon$) at three ε levels. Wrap Mis: TC-WRAPPER miscoverage (%) at $\varepsilon=0.10$. † Margin within estimator resolution.