BEFORE THE BREAKTHROUGH: BOTTLENECK CROSSINGS AS PRECURSORS IN LONG-HORIZON AGENT RL

Anonymous authors
Paper under double-blind review

ABSTRACT

When validation success plateaus, it is difficult to tell whether a long-horizon LLM agent is still making hidden progress toward a breakthrough or is genuinely stalled. We introduce the *Crack-Front Monitor*, which tracks clusters of verifier-confirmed bottleneck crossings as candidate precursors of breakthroughs, without additional model inference or environment interaction. We characterized the signal on 15 discovery runs of 8–14-turn executable tool tasks. We then froze the monitor, its thresholds, and its predictions before any crack or breakthrough event in a ten-run prospective cohort spanning independent seeds and a second task distribution in the same environment family. In the completed cohort, crack-fronts preceded all nine observed sustained breakthroughs across both task distributions and both training arms, with no false alarms. The one run that never fronted never broke through. In an exploratory comparison with thresholds frozen before application, the monitor's median online warning lead was 20 steps versus 5 for a near-miss baseline; a post-hoc crossing-rate rule achieved a similar lead, consistent with the signal residing in the crossing events themselves. The frozen timing window and restart budgets transferred poorly: only four of nine lead times fell within the preregistered 30–60-step window, and the restart budgets would have stopped three runs that later broke through. These results provide initial prospective evidence that bottleneck crossings can reveal progress before validation improves, but they do not yet support reliable breakthrough timing or stopping rules. Code and data are available at `https://anonymous.4open.science/r/crack-front-monitor-8997`.

1 INTRODUCTION

Reinforcement learning for multi-turn tool-use agents has a waiting problem before it has a credit problem. On hard task distributions, training sits at a success-rate floor for hundreds of steps: validation curves are flat, mean training reward is flat, and two runs that look identical at step 150 can have opposite futures: one is about to break through, while the other stays flat for its entire budget. Practitioners face this regime with learning-curve extrapolation and multi-fidelity early stopping (Li et al., 2018; Klein et al., 2017), which read *aggregate* progress curves; at the success floor those curves carry little signal, and stopping rules built on them stopped several discovery runs that later ignited (§3).

We instead track a different statistic: the time–task clustering of *rare bottleneck-passage events*. Across seventeen training runs under matched settings on executable tool tasks (8–14 user turns; fifteen of the seventeen monitor-applicable, App. E), every breakthrough followed the same four-stage anatomy: a hard per-turn *wall* (conditional pass rate $\sim 10^{-4}$ in every arm); a stochastically initiated *crack-front* (a concentrated burst of rollouts passing *through* the wall across many tasks); *sparks*; and a self-reinforcing *cascade*. In our data the crack-front behaves as a precursor: it precedes ignition in every run that ignites (median lead 45 steps, range 25–160) and appears in none of the runs that stay flat, while validation remains at the floor throughout the lead window.

We turn this observation into an online monitor and evaluate it prospectively. The monitor's full specification, from the wall-identification protocol to its two predictions, was frozen under a git-

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.

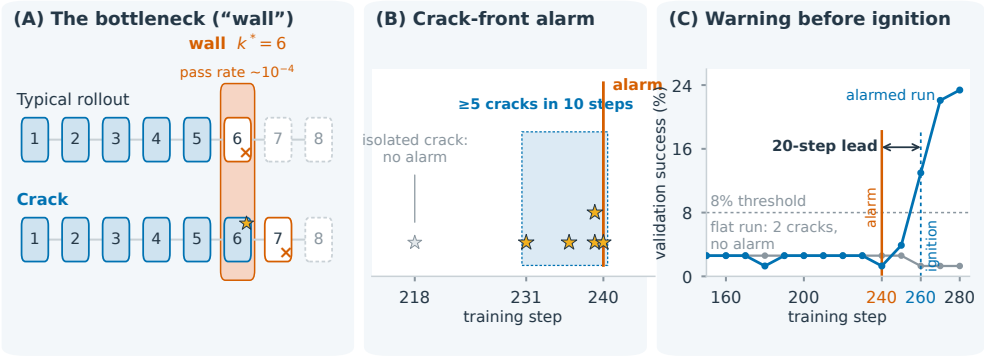


Figure 1: How the Crack-Front Monitor works. (A) Schematic: the existing checker records how many leading turns each rollout passes; almost all rollouts that reach the wall turn k^* (identified from training steps 1–40, not hand-chosen) fail to pass it, and a rollout that does counts as one crack, which is not full-task success. (B) Observed crossings in prospective decomposition run s148: an isolated crack never alarms; five cracks inside the closed window $[231, 240]$ raise the alarm at step 240 (231 is the retrospective front onset, not the online alarm time). (C) Validation in the same run (blue) is still at the floor at the alarm; sustained ignition begins at step 260 (confirmed at 280), an online warning lead of 20 steps. Gray: a discovery run on the same task distribution with two cracks in 280 steps, no alarm, and no ignition. The example illustrates the alarm mechanism; stopping-policy validation is evaluated separately (§4).

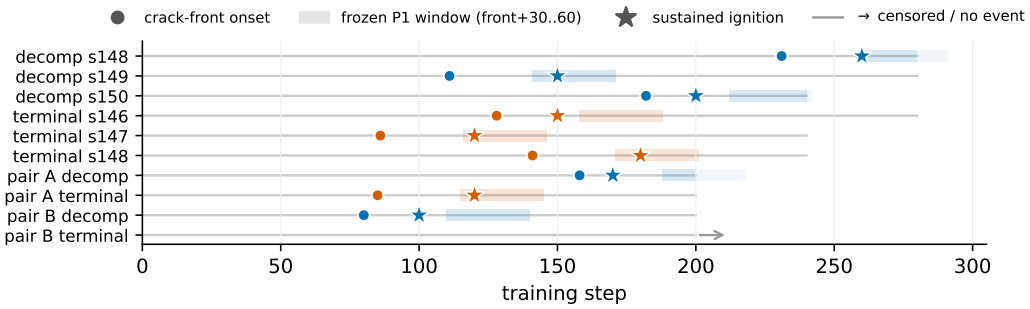


Figure 2: The frozen prediction window versus observed events, per prospective run (final cutoff 2026-09-23, cohort complete). Dot = retrospective crack-front onset (the online alarm trails by 9 steps); gray lines end at the observation cutoff; the P1 band (front $+30$ to $+60$) is lighter beyond it, prediction only; star = sustained-ignition onset. Every ignition is front-preceded (9/9), the front-free run never ignites, no false alarms; most ignitions arrive before the window opens (leads 12–39, median 29; four inside). The specification predates every event shown (App. E).

timestamped protocol freeze, its registration document and amendment history released verbatim (App. E), before any crack or breakthrough event in a cohort of held-out runs, on both the original task distribution (independent seeds) and a newly generated second distribution (true pairs, arms differing only in the credit signal). Section 4 reports all outcomes of that test, including censorings and failures.

Two further sets of results address when the signal is useful and why it appears. First, the monitor's information is *incremental* against three frozen comparator rules on identical budgets: validation-plateau stopping commits three false kills, mean-progress does not discriminate, and the near-miss baseline matches the monitor's decision accuracy with the opposite error type; in an exploratory prospective-cohort comparison the median online warning lead is 20 steps for the monitor versus 5 for near-miss (§3, §4). Second, we examine *why* bottleneck passage may carry useful optimization signal (§5): the verifiable outcome already contains an exact per-turn decomposition, available dur-

ing training at zero cost, and arms trained with it produced earlier and wider crack-fronts among the fronting discovery runs; in corruption arms that preserve every rollout's *total* reward, two frontier-attenuated records front later than every random-attenuation record while a third does not (§5.2); the probes are consistent with the placement hypothesis but do not establish it. We identify a group-normalization failure mode of reward-space credit injection, 47% of final-turn advantages flipping sign as groups homogenize; injecting credit after normalization removes it and preserves optimal deterministic policies under stated conditions (Theorem 1).

Contributions.

(1) **Breakthroughs are preceded by a recurring bottleneck-crossing pattern** (§2, §3). Across 17 runs under matched settings, every breakthrough followed the same wall $\rightarrow$ crack-front $\rightarrow$ spark $\rightarrow$ cascade sequence. The crack-front, read from verifier outputs already computed for the reward, retrospectively preceded every ignition (7/7) with zero false alarms at all nine settings of a 3×3 threshold grid.

(2) **Prospective validation supports precursor detection but not timing or restart calibration** (§4, Table 1). We froze the monitor, its comparison rules, and its predictions under a git-timestamped protocol freeze before any cohort event, and scored them on completion across two task distributions. Baseline and restart-policy comparisons are discovery-cohort analyses; the prospective near-miss lead comparison is exploratory. The misses, in lead-window calibration and restart-budget transfer, show that front-to-ignition timing did not transfer reliably from the discovery runs to the prospective cohort.

(3) **Where credit sits is associated with front timing, and reward-space injection has a normalization failure mode** (§5). Sum-preserving placement probes (identical per-rollout totals, credit moved off vs. onto the frontier turn) yield later fronts in two of three frontier-starved records, not the third. We also show why reward-space redistribution flips advantage signs as groups homogenize, and derive an advantage-space alternative with an optimality-preservation result for the idealized estimator under stated assumptions.

We use the per-turn decomposition as an instrument (§5); the replay-based credit audit is App. C, and the short-horizon study delimiting the regime is App. D.

2 Setting and the Anatomy of Sparse-Regime Breakthroughs

Experimental setting. A trajectory consists of user turns $t = 1, \ldots, T$; turn t comprises the t-th request, the policy's tool calls, and the environment's responses. Turns are the clock: the harness injects exactly T requests, so every trajectory has T turns while the number of calls within a turn is free. A deterministic checker executes the policy's calls and the reference calls in parallel environments and marks turn t passed if the environment states (and read-op responses) agree after it. The verifiable outcome is $Y = \frac{1}{T}\#\{\text{leading passed turns}\} \in [0, 1]$; the training reward R adds a dense call-overlap term and a full-completion bonus (App. A). Qwen3-4B (Yang et al., 2025) trained from the instruction-tuned checkpoint with DAPO/GRPO (Shao et al., 2024; Yu et al., 2025), 16 rollouts per prompt, matched data order, steps, and token budgets across arms; environments are the executable multi-turn backends of the Berkeley Function-Calling Leaderboard (BFCL v4) (Patil et al., 2025), with procedurally generated, checker-verified, decontaminated training tasks (App. A). The **long regime**, this paper's setting, uses 8–14-turn tasks (1,543 checker-verified; starting success 2.6%); a second, newly generated task distribution for the prospective cohort is described in §4. Arms differ *only* in the credit signal delivered to the group-relative estimator: **terminal-only** GRPO, the exact **decomposition** (Eq. 1), a **judge**-credit pipeline (RWTP; App. C), a **masked-judge** variant, and structured **corruption** arms (§5). Held-out validation is scored every 10 steps.

Every run sits at the 2.6% floor for at least 90 steps (validation curves and per-run crossing rates: Figure 4, App. E). Rollout forensics across these runs reveal the same four-stage transition.

(1) The wall. One turn depth at which the conditional pass rate collapses: $q_k = P(\text{pass turn } k \mid \text{passed } 1..k{-}1)$ is 0.77–0.82 for turns 1–5 and $\sim 10^{-4}$ at turn 6, identically in every arm and on both task distributions (re-identified per corpus from steps 1–40 before any readout; §3). Below the wall, per-turn conditionals, group contrast supply, and the sub-wall tail are statistically indistinguishable across arms.

(2) Crack-front initiation. Within a few tens of steps, partial wall-crossings (*cracks*: rollouts on $\geq$8-turn tasks passing $\geq$6 turns) appear across many tasks at once, hundreds of events over 20–30 task types in decomposition arms; tens over 10–18 types in controls. The front precedes ignition in every igniting discovery run (median lead 45, range 25–160); no flat run exceeds single-digit crossings.

(3) Sparks. Full near-completions; a lone spark without a front dies out (observed once), sparks backed by a front convert.

(4) Cascade. Self-reinforcing amplification to a converged endpoint band (13–21% by step 280) regardless of arm: among the igniting discovery runs, endpoints fell within a similar range despite different onset times (Rad et al., 2026; Khatri et al., 2026).

Initiation is stochastic: matched arms diverge (two decomposition seeds never front in 280 steps) and we find no observable pre-spark differences between arms beyond the front itself. The formal timing comparison uses first-passage survival analysis with right-censoring (front initiation: KM median 130 vs. not reached, permutation log-rank p=0.17; ignition: median 175 vs. not reached, p=0.17; n=6 vs. 9 runs): the timing difference is not statistically significant, one motivation for the prospective cohort of §4.

3 The Crack-Front Monitor

We convert the anatomy into an operational monitoring procedure, frozen verbatim under a git-timestamped protocol freeze before any prospective-cohort crack or breakthrough event (App. E reproduces the frozen text and commit hashes).

Signal construction and wall identification. The monitor reads only $(n_{\text{turns}}, \text{turns_passed})$ pairs the checker already computes for the training reward: no judge, learned model, extra rollouts, or additional forward passes. For a new task distribution, the wall is identified from steps 1–40 statistics, before any monitor readout: the turn k^* at which the conditional pass-rate curve q_k collapses by orders of magnitude. On both our corpora k^*=6 ($q_6 = 8 \times 10^{-5}$ pooled on the second corpus, vs. $q_{1..5} \in [0.77, 0.82]$). If no wall is identifiable, the monitor abstains, and the abstention is counted in applicability coverage.

Decision rule (frozen). A *crack* is a rollout on an $\geq$8-turn task with turns_passed $\geq k^*$. The *front* is the first step s such that window $[s, s{+}9]$ contains $\geq$5 cracks (the rule imposes no task-diversity requirement; the observed fronts were nonetheless cross-task, §2). Predictions: **P1**, front $\Rightarrow$ ignition (3 consecutive validation readings >8%) between 30 and 60 steps after the front; **P2**, no front by budget R (162 for decomposition arms, 256 for controls) $\Rightarrow$ no ignition by $R{+}40$, and the run is restarted at R under the policy.

Robustness and incremental value. Across the tested grid ($w \in \{5, 10, 20\} \times c \in \{3, 5, 10\}$), every setting yields the same classification (7/7 fronts precede ignition, 0/7 false alarms, one censored-after-front), only the median lead moving (22–62 steps; App. E). R' denotes the restart budget R used as a comparison checkpoint. Table 7: validation-plateau stopping kills three runs that later ignite, whose validation sat at the floor through R'; mean-progress continues everything; near-miss (a rollout passing $\geq T{-}1$ leading turns without full completion; rate over long-horizon rollouts in a trailing 20-step window) reaches 14/15 with one false kill, the costly error type; the monitor's single error is one unsuccessful continuation under the fixed-horizon endpoint, on a run censored before sustained ignition was confirmed. Accuracy ties at R', but *online warning lead* does not: the crack-front preceded all seven discovery ignitions (median lead 36, range 16–151), near-miss only four (median 6; at or after onset in three), and the paired difference favors the crack-front in every run (crack-front alarm minus near-miss alarm, negative = crack-front earlier: median -18, bootstrap CI $[-21, -15]$); a near-miss restart policy is also dominated on decomposition arms (362 vs. 274). Nor is a *single* crossing the signal (discovery-set, post hoc): the seven never-igniting runs contain 0–8 isolated cracks each, so a first-crossing alarm false-alarms on six of seven flat runs where the front rule fires on none. A cumulative variant ($\geq$5 lifetime cracks, no window) alarms within a few steps of the windowed rule and false-alarms on one flat run: repetition does most of the filtering, the window only removes that alarm, and clustering's independent value beyond repetition is not estab-

Table 1: What was frozen, and what happened. Each row quotes a preregistered commitment (git-timestamped before any cohort event) and its outcome at the final cutoff: one held outright, one held in direction but not calibration, one mixed, one failed prospectively.

frozen commitment	outcome (10/10 runs, final cutoff)	reading
Fronts precede ignition; no ignition without a front	9/9 front-preceded, two distributions, both arms; the front-free run never ignited; zero false alarms	**held**
P1: ignition 30–60 steps after the front	leads 12–39 (median 29); 4/9 in window	direction held; **window miscalibrated**
P2: no front by $R \Rightarrow$ no ignition by $R{+}40$	one SUPPORTED, one REFUTED (by 2 steps), one censored	**mixed**; R transferred poorly (prospective decomp fronts 80–231 vs. discovery 90–142, two of five beyond R)
Restart at R beats fixed deadlines (discovery estimate)	the online protocol would have stopped three eventual breakthroughs (two under onset accounting); monitor and near-miss decide identically	**failed prospectively**; budget unresolved

lished here. Discovery-set analyses; §4's registered layer tests the monitor only, and its near-miss lead comparison is exploratory.

Budget implications and scope. Policies are evaluated offline on the collected discovery-run trajectories, never as measured improvements from an executed allocation policy. With censored-after-front attempts counted as failures at full cost and the fixed deadline D tuned on the same runs *in the comparator's favor*: expected steps to ignition $\mathbb{E}[L]/q$, under *onset-based offline accounting*, is **274** for monitor+restart on decomposition arms vs. 318 for the best fixed deadline and 332 for no restart; on control arms the conservative treatment favors the tuned deadline (680 vs. 740). The online-cost variant preserves every ordering (decomposition 291 vs. 345 vs. 350; controls 760 vs. 740 vs. 800). Bootstrap intervals and $q{=}0$ resample fractions in App. E. The monitor is not learning-curve prediction, since aggregate curves carry little signal at the floor; it does not guarantee ignition, since initiation is stochastic; and it is not validated beyond one environment family (§4 is within-family).

4 Prospective Validation

All rules, thresholds, and pairings below were frozen before any cohort event; rows updated only by accruing steps through the final cutoff (2026-09-23).

Cohort and pairing. Two tiers (s148 had trained 22 steps at registration, metrics at floor and zero cracks, disclosed; the other nine had not started). **Original distribution** (independent seeds, stated plainly as unpaired): decomposition seeds 148/149/150 vs. terminal seeds 146/147/148, 240–280 steps. **New task distribution** (true pairs): a second corpus generated under a distinct recipe, single-domain compositions of 596 tasks gated only on checker validity, zero benchmark leakage, structural difference, and pipeline stability, wall identifiability being an outcome rather than a gate. Pair A = decomposition seed 4242 + terminal seed 4242 and pair B = seeds 4243, identical model seed and data order within a pair, arms differing only in credit; 200 steps each. The wall for the new corpus was identified from steps 1–40 of both arms and recorded before any monitor readout ($k^*{=}6$, $q_6 = 8 \times 10^{-5}$); the record was written at run-step $\sim$78, past the preregistered step-60 deadline, a deviation disclosed in App. E.

Scoring (frozen). Each run is scored SUPPORTED / REFUTED / CENSORED against P1 and P2 as frozen; censoring at the run's last observed step; no post-hoc window adjustment. We distinguish interval accuracy from the verdicts: a window MISS records an ignition outside the 30–60-step interval and is not automatically REFUTED under the original criteria. Runs polluted by infrastructure retry-storms are excluded under a rule fixed before readout (App. E).

Table 2: Prospective scorecard at the final cutoff (2026-09-23, cohort complete; rules unchanged since freezing). P1: SUPP = onset lead in the frozen 30–60 window; MISS = front preceded ignition, lead (shown) outside it. P2: SUPP / REF / CENS; n/a = not applicable (front by R). Left: original distribution (independent seeds; budgets 240–280). Right: second distribution (true pairs; budgets 200).

run	front	ign.	P1	P2
decomp s148	231	260	MISS (29)	SUPP
decomp s149	111	150	SUPP (39)	n/a
decomp s150	182	200	MISS (18)	REF
terminal s146	128	150	MISS (22)	n/a
terminal s147	86	120	SUPP (34)	n/a
terminal s148	141	180	SUPP (39)	n/a

run	front	ign.	P1	P2
pair A decomp	158	170	MISS (12)	n/a
pair A terminal	85	120	SUPP (35)	n/a
pair B decomp	80	100	MISS (20)	n/a
pair B terminal	none	none	n/a	CENS

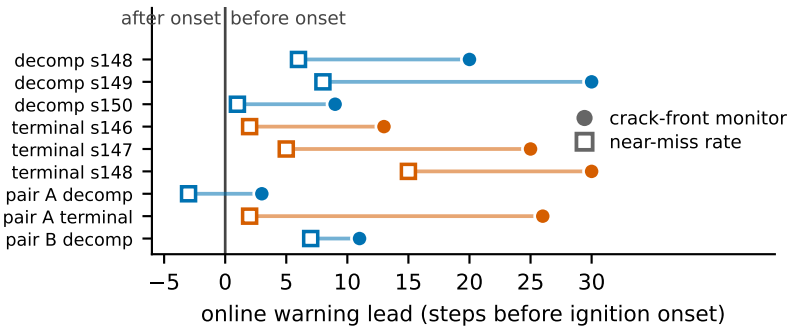


Figure 3: Per-run online warning leads (steps from alarm to ignition onset), prospective cohort; exploratory, both rules' thresholds frozen before application; decomposition blue, terminal vermillion. The monitor (filled) leads near-miss (open) in all nine igniting runs; the tenth run alarms under neither rule and never ignites.

Registered analyses and results. Five analyses were registered; each is reported with its outcome at the final cutoff. (a) P1/P2 scorecard per tier: under the registered primary endpoint (original-distribution comparisons truncated at 240, later steps follow-up only; Amendment 4), eight observed ignitions, all front-preceded, with decomp s148 alarmed at 240 and igniting only in follow-up (260); over the full follow-up, precedence 9/9, zero false alarms; the frozen window captured 4/9 (onset leads 12–39 vs. discovery 25–160; alarms trail fronts by 9 steps). (b) Applicability coverage including abstentions: 10/10 at full budget (terminal s147's late start after a disclosed queue fault, App. F). (c) The decomposition timing effect re-estimated on the true pairs only, by within-pair front and ignition ordering: the pairs split (pair A terminal-first, pair B decomposition-first; Table 2), so decomposition acceleration is not confirmed on paired data.

(d) Conditional discrimination, whether front presence predicts subsequent ignition among run-step points matched on step and group mean-progress, computed with online information only (alarm times, not retrospective fronts) on the pooled discovery+prospective runs: among points matched on step and group mean progress ($|\Delta| \leq 0.15$), with exactly one run alarmed and neither ignited, the alarmed run ignites first in 191/216 decidable comparisons (censoring-guarded; concordance 0.88). These are matched run-step comparisons that reuse the 25 runs across grid times, not independent samples; a cluster bootstrap resampling runs (each comparison weighted by its runs' multiplicities) gives CI $[0.73, 0.99]$, supportive but low-powered, as the small run count dictates, and not the product of one run or one tier (leave-one-run-out 0.86–0.95; discovery-only 0.88, cohort-only 0.91; grid and tolerance inherited, untuned); (e) The budget-value tables of §3 recomputed with the prospective runs appended at freeze: pooled onset-based expected-cost estimates move to 263 (decomposition) and 427 (control), a composition effect of the new cohort's higher breakthrough rate, not a same-cohort policy gain; the frozen R remains miscalibrated (three eventual breakthroughs stopped under

the online protocol, two under onset accounting), so the budget layer is unresolved exactly as registered.

An exploratory, post-hoc comparison of *online warning time* (near-miss at its calibrated 0.5% trailing rate, not Table 7's 1% variant) separates what decision accuracy could not (Figure 3; the monitor leads in all nine igniting runs, medians 20 vs. 5). These are frozen operational rules, not optimal tradeoffs; still, eight near-miss thresholds (0.0003–0.05, zero observed false alarms on the seven flat discovery runs) top out at a cohort median lead of 10, while a crossing-*rate* rule (trailing 10-step cracks per rollout, threshold $1.01\times$ the flat runs' maximum) matches the monitor (21 vs. 20, zero flat false alarms): the information is the bottleneck-crossing event at the identified depth, not the aggregation.

4.1 Structural transfer: moving the wall

Does the bottleneck identification procedure transfer when the wall is shifted by construction? We constructed a third corpus that moves the bottleneck by design, prefixing every second-distribution task with three trivial checker-verified read-only turns (displacing each depth by exactly three; fixed before training, no task selected using monitor outcomes). The first attempt produced an *exactly* zero conditional pass rate at absolute turn 6 (26,118 attempts), which we declined to treat as a semantic wall (grounds to inspect infrastructure limits; no exact-zero abstention criterion was preregistered) and traced, via execution records, to the harness's assistant-turn budget (documented and excluded, App. F). With the budget corrected, the frozen procedure applied to steps 1–40 recovers the wall at turn 9, the constructed displacement, at 7.1×10^{-5} (vs. the second corpus's 8×10^{-5}; prefix turns pass at ~ 0.999, displaced turns at 0.76–0.84). With the displaced crack rule, the front fired at step 163 of the decomposition arm's full 200-step run and not in the terminal arm (an asymmetry also seen in pair B); the fronting arm's validation then rose to 17.2–20.7% by budget end, but two missing readings leave the frozen sustained rule unscorable (censored), and a further pair stayed flat in both arms with no front (App. E). The identification procedure transfers to the displaced wall, and the front statistic detects concentrated crossings there; the front's *predictive* value at the new depth is not yet established.

5 Mechanism: Why Placement Carries the Signal

5.1 Free exact credit

The training reward computes $k = \#\{\text{leading passed turns}\}$ on its way to $Y = k/T$; the exact decomposition

$$c_t^{\mathrm{dec}} = \tfrac{1}{T}\,\mathbf{1}[t \le k], \qquad \textstyle\sum_t c_t^{\mathrm{dec}} = Y, \tag{1}$$

is therefore available at every training step at zero marginal cost (§3). The formula is potential-based shaping (Ng et al., 1999) with $\phi(s_t) = k_t/T$, the exact limit of stepwise progress attribution (Wang et al., 2025), and an instance of verifier-derived supervision (Li et al., 2026c; Yuan et al., 2026). The outcome implies two constraints, the *sum* constraint $\mathbf{1}^\top c = Y$ and a *support* constraint ($c_t = 0$ for $t > k$): only leading passed turns can have earned credit. Judge-estimated credit violates it badly (17.6% of its mass lands after the first failure; App. C), one reason approximate credit routes poorly even when it helps.

Effect of exact decomposition. In the discovery cohort, decomposition arms produced the earliest and widest crack-fronts (the prospective pairs split, §4; fronts 90–142 across four runs and both injection mechanisms vs. 149–236 for every fronting control, an order of magnitude more crack events), without guaranteeing initiation: in the discovery runs, exact decomposition is associated with earlier front emergence, and the prospective paired runs did not establish consistent acceleration. At the 2.6% floor, 80–83% of 16-rollout groups have identical k across all rollouts; in the remaining groups, the decomposition concentrates the within-group contrast *at the position of the frontier turn*, while terminal reward spreads the same contrast over the whole trajectory.

Table 3: Sum-preserving placement probes. The removed credit is redistributed within the same rollout (random-starved: to the other turns; frontier-starved: to the earlier prefix); front times listed per seed.

arm	what moves	front
exact decomposition	nothing	90–142 (4 runs)
random-turn starved (frontier spared)	one random non-frontier turn pays $0.3\times$	84, 183, 192
frontier starved	the frontier turn pays $0.3\times$	150, 248, none by 240

5.2 Placement probes: identical totals, different outcomes

The frontier-payment hypothesis admits a by-construction test. Both corruption arms of Table 3 preserve *every rollout's total reward exactly* (the removed credit is redistributed within the same trajectory), so per-rollout reward-total explanations are excluded by construction; the arms differ in *where* the credit sits (token-level advantage magnitudes are not matched).

The released records do not fully separate: two frontier-starved fronts (248; none observed by step 240) are later than every random-starved front (84, 183, 192; one igniting at 175), but a third frontier-starved seed fronted at 150 and ignited at 170, inside the random-starved range (an earlier internal note listing this seed's front at 216 is superseded by the released series; App. F). We therefore do not claim a replicated ordering of front-emergence times from these probes. The design still excludes per-rollout reward-total accounts of any individual front, and does not exclude token-level advantage-magnitude accounts (matching totals does not match gradient magnitudes).

5.3 The σ_R pathology and its repair

Reward-space credit injection places $r_t = \beta c_t$ at each turn's final token and $r_{\text{term}} = R - \beta Y$ at the terminal token, so every rollout's total equals R: group statistics are exactly GRPO's, and $\beta{=}0$ is GRPO; this natural form fails at scale (Prop. 1).

Proposition 1 (Sign flip under reward-space redistribution)**.** *Let $P_{<t} = \sum_{k<t} c_k$ be the credit already paid before turn t, μ_R, σ_R the group mean and spread of totals, and $A^{\text{GRPO}} = (R - \mu_R)/\sigma_R$. Under reward-space redistribution the advantage of turn-t tokens is $A_t = A^{\text{GRPO}} - \beta P_{<t}/\sigma_R$. Hence the final turn's advantage has the opposite sign from A^{GRPO} whenever $\beta P_{<T} > R - \mu_R > 0$; as groups homogenize ($\sigma_R \to 0$ with $R - \mu_R = O(\sigma_R)$), for any fixed $\beta > 0$ this holds for every above-average rollout whose paid prefix credit $P_{<T}$ is bounded away from zero.*

Measured on 6,116 late-training rollouts: the final-turn advantage flips sign in **47%**, and successful rollouts' completion turns shift by -9.7σ: perfect completions are punished. An offline recompute over 1.15M discovery rollouts under the *real* composite return (App. E; an earlier 92.7% broadcast-zeroing claim was a binary-stand-in artifact, withdrawn) shows the terminal advantage tracking the exact-credit scalar ($+0.99$): the schemes are near-equivalent *per rollout* and differ in *where* the payment lands. Reward-space exact credit pays the wall spread-dependently (frontier-contrast/spread correlation -0.49 under the real return; -0.75 under the simplified probe). The repair is to inject *after* normalization: $A_j = A^{\text{GRPO}} + \beta_a z_j$ with $z_j = (C_j - \mu_C)/\sigma_C$ the standardized credit return-to-go; the credit term carries its own scale, spread-decoupled.

Theorem 1 (informal; formal statement and assumptions in App. B). For a fixed prompt with normalization constants held fixed with respect to the policy parameters, the idealized un-clipped advantage-space injection is a policy-gradient estimator for $\mathbb{E}[R + \kappa Y]$ with $\kappa = \beta_a \sigma_R/\sigma_C$ (this does not by itself establish unbiasedness of the finite-group training update or a single global coefficient for the multi-prompt objective, App. B); it preserves optimal deterministic policies under the reward-alignment and deterministic-dynamics assumptions of App. B (bounded by a stochastic counterexample), prefix-measurable credits yield an unbiased estimator under the stated assumptions, and post-hoc credits can introduce bias, linear in β_a in the case we exhibit. The statement is optimality-preserving rather than policy-invariant; preserving the optimum does not pin down where finite training ends; assumptions are verified, and their necessity shown by counterexample, in enumerable chain MDPs (App. B).

Regime delimitation. Short-horizon tasks are the opposite regime: group comparison already carries failure-location information (Sullivan & Koller, 2026), and every credit arm scores below terminal-only GRPO there (-4.1 to -7.4 points; descriptive, 1–2 seeds per arm; App. D). The mechanism here operates where group signal is starved, not everywhere.

6 Related Work

Early stopping and curve prediction. Multi-fidelity methods (Li et al., 2018) and learning-curve extrapolation (Klein et al., 2017) allocate budget from aggregate progress curves. In contrast, we study rare bottleneck-crossing events while aggregate validation remains near the floor; the information lives in their clustering at matched time and progress. Abrupt post-plateau transitions are documented, with continuous progress beneath them (Sun et al., 2026; Reddy, 2024); our contribution is an *online, zero-cost* verifier-derived precursor, its preregistered prospective test, and the budget accounting.

Credit assignment for agents. Verifier-native credit is concurrent: VICT traces verifier atoms to actions (Li et al., 2026c); VPR converts intermediate oracles into dense turn rewards (Yuan et al., 2026); the decomposition we use as an instrument is a simple instance of these (shaping Ng et al., 1999; SPA-RL's exact limit Wang et al., 2025). Step-credit mechanisms (TRACE Tao et al., 2026; IGPO Wang et al., 2026a; GiGPO Feng et al., 2025; SALT Li et al., 2026a; RTMC Wang et al., 2026c; CARL Kumar et al., 2026; AgentPRM Xi et al., 2026; Flow-GRPO Li et al., 2026e; CW-GRPO Wang et al., 2026b; SCAR Cao et al., 2025; SSVPO Li et al., 2026d; φ-PPO Li et al., 2026b) *produce* credit and validate by RL score; our replay-based audit (App. C) inverts the direction, as does a concurrent resampling audit (Zhang, 2026). Coverage-vs-targeting (Zhou et al., 2026) predicts targeting pays where verifier information density is high; our probes examine how frontier-credit placement affects bottleneck-crossing emergence. Concurrent work reports a sibling normalization defect for dense auxiliary rewards (Wang, 2026); ours concerns outcome-derived redistribution, with a different failure and repair.

7 Limitations and Conclusion

Limitations. Our evidence comprises 15 monitor-applicable discovery runs and a 10-run prospective cohort; survival contrasts remain trends ($p\approx0.17$), and the corruption study uses three seeds per rung, with one record superseding an internal note (App. F). The monitor requires an identifiable wall (it abstains otherwise, and abstentions are counted), and the free decomposition needs leading-prefix outcomes. Our main experiments use a single 4B policy, with a 2B study in App. D. Budget claims are offline discovery estimates, and the frozen budgets did not transfer prospectively (§4). Total compute was approximately 5,200 GPU-hours (App. F).

Crack-fronts preceded every observed breakthrough: we characterized the anatomy, froze a zero-cost monitor before any cohort event, and scored every prediction. Lead windows and stopping budgets did not transfer; a persistently absent front warrants scrutiny, not a validated stop. Identifying the wall early and counting its crossings requires no additional inference or environment interaction, and in these cohorts the count preceded every breakthrough.

AI Use Statement

We used large language models to assist with language polishing, drafting limited portions of the manuscript, identifying some related work, implementing parts of the experimental code, and generating synthetic training tasks. Generated tasks were checked using deterministic verifiers and audited for benchmark contamination. The authors reviewed the AI-assisted contributions and take responsibility for the final manuscript and results.

Ethics Statement

This work studies training dynamics of tool-use agents on procedurally generated, checker-verified synthetic tasks; it involves no human subjects, personal data, or sensitive content. Released corpora are audited for contamination against the public benchmark.

Reproducibility Statement

Code, data, and the preregistration history are anonymously available at `https://anonymous.4open.science/r/crack-front-monitor-8997`. The repository contains the complete analysis code (monitor, readouts, baselines, survival and budget analyses, mechanism recomputes, theorem checks), the replay-audit engine, the training patches and launchers, the corpus generators with the decontamination audit, the cleaned corpora with SHA-256 hashes, the preregistration document with its git timestamps, an incident summary, and derived per-run event series (per-step crack counts and validation curves) from which every monitor-related number in the paper is regenerated by an included script; full raw rollout logs exceed the anonymous repository's size limit and are released with the public repository upon publication.

References

Meng Cao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. SCAR: Shapley credit assignment for more efficient RLHF. *arXiv preprint arXiv:2505.20417*, 2025. URL https://arxiv.org/abs/2505.20417.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems*, volume 38, pp. 46375–46408. Curran Associates, Inc., 2025. doi: 10.52202/085713-1544. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/75ea88aca02640cca54c1231db8cacdd-Abstract-Conference.html.

Aaron Klein, Stefan Falkner, Jost Tobias Springenberg, and Frank Hutter. Learning curve prediction with Bayesian neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=S11KBYclx.

Abhijit Kumar, Zoey Wu, and Mohit Suley. Knowing when to ask: Segment-level credit assignment for LLM tool use. *arXiv preprint arXiv:2605.27788*, 2026. URL https://arxiv.org/abs/2605.27788.

Jiazheng Li, Yawei Wang, Qiaojing Yan, Yijun Tian, Zhichao Xu, Huan Song, Panpan Xu, and Lin Lee Cheong. SALT: Step-level advantage assignment for long-horizon agents via trajectory graph. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 4709–4725. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.findings-eacl.247. URL https://aclanthology.org/2026.findings-eacl.247/.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018. URL https://jmlr.org/papers/v18/16-558.html.

Mingxuan Li, Kai-Zhan Lee, and Elias Bareinboim. Counterfactual Shapley credit assignment. *Reinforcement Learning Journal*, 7, 2026b. URL https://rlj.cs.umass.edu/2026/papers/Paper45.html.

Pengcheng Li, Zhengyang Zhang, Dongxu Zhang, Sui Huang, and Shaohua Ma. VICT: Verifier-instrumented credit tracing for long-horizon LLM agent reinforcement learning. *arXiv preprint arXiv:2608.28128*, 2026c. URL https://arxiv.org/abs/2608.28128.

Yugu Li, Zehong Cao, Jianglin Qiao, and Siyi Hu. SSVPO: Effective step-level credit assignment for RL training of language models. In *International Conference on Learning Representations*, 2026d. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4da4dd613cc9e63932d643b9c6e8f86f-Abstract-Conference.html.

Zhuofeng Li, Haoxiang Zhang, Seungju Han, Sheng Liu, Jianwen Xie, Yu Zhang, Yejin Choi, James Zou, and Pan Lu. In-the-flow agentic system optimization for effective planning and tool use. In *International Conference on Learning Representations*, pp. 50524–50570, 2026e. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/52c4a736cd63b3c7bf52011a1f66f3f6-Abstract-Conference.html.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pp. 278–287, 1999. URL https://people.eecs.berkeley.edu/~russell/papers/icml99-shaping.pdf.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392. PMLR, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Ali Rad, Khashayar Filom, Darioush Keivan, Peyman Mohajerin Esfahani, and Ehsan Kamalinejad. Rate or fate? RLV$^{\varepsilon}$R: Reinforcement learning with verifiable noisy rewards. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=LwB2EacVT6.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/540a3bf4863672cdaebfed69fa2ce335-Abstract-Conference.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Michael Sullivan and Alexander Koller. GRPO is secretly a process reward model. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=nMGaOCVlDW.

Yiyou Sun, Yuhan Cao, Pohao Huang, Haoyue Bai, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. RL grokking recipe: How does RL unlock and transfer new algorithms in LLMs? In *International Conference on Learning Representations*, pp. 93190–93220, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/96842011407c2691ab4eefff48fc864d-Abstract-Conference.html.

Leitian Tao, Baolin Peng, Wenlin Yao, Tao Ge, Hao Cheng, Mike Hang Wang, Jianfeng Gao, and Sharon Li. TRACE: Turn-level reward assignment via credit estimation for long-horizon agents. *arXiv preprint arXiv:2607.13988*, 2026. URL https://arxiv.org/abs/2607.13988.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/aadaedf7fc275bf68ad217bf1b38626e-Abstract-Conference.html.

Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. SPA-RL: Reinforcing LLM agents via stepwise progress attribution. *arXiv preprint arXiv:2505.20732*, 2025. URL https://arxiv.org/abs/2505.20732.

Junzhe Wang, Zhiheng Xi, Yajie Yang, Hao Luo, Shihan Dou, Tao Gui, and Qi Zhang. Enhancing LLM-based search agents via contribution weighted group relative policy optimization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31704–31718. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.1462. URL https://aclanthology.org/2026.acl-long.1462/.

Tao Wang, Suhang Zheng, and Xiaoxiao Xu. RTMC: Step-level credit assignment via rollout trees. *arXiv preprint arXiv:2604.11037*, 2026c. URL https://arxiv.org/abs/2604.11037.

Yu Wang. The dark room in the reward channel: Dense prediction rewards collapse GRPO-trained LLM agents—and the channel, not the content, decides what works. *arXiv preprint arXiv:2607.21273*, 2026. URL https://arxiv.org/abs/2607.21273.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Yajie Yang, Wenxiang Chen, Zhihao Zhang, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. AgentPRM: Process reward models for LLM agents via step-wise promise and progress. In *Proceedings of the ACM Web Conference 2026*, pp. 4184–4195. Association for Computing Machinery, 2026. doi: 10.1145/3774904.3792551. URL https://doi.org/10.1145/3774904.3792551.

Minheng Xiao and Xian Yu. Reward redistribution via Gaussian process likelihood estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(32):27073–27080, 2026. doi: 10.1609/aaai.v40i32.39921. URL https://ojs.aaai.org/index.php/AAAI/article/view/39921.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244. Curran Associates, Inc., 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Huining Yuan, Zelai Xu, Huaijie Wang, Xiangmin Yi, Jiaxuan Gao, Xiao-Ping Zhang, Yu Wang, Chao Yu, and Yi Wu. Verifiable process rewards for agentic reasoning. *arXiv preprint arXiv:2605.10325*, 2026. URL https://arxiv.org/abs/2605.10325.

Haiyue Zhang. Credit without ground truth: Auditing step-level credit assignment in LLM agents against executed replay. *arXiv preprint arXiv:2608.19760*, 2026. URL https://arxiv.org/abs/2608.19760.

Chenyu Zhou, Qiliang Jiang, Shuning Wu, and Xu Zhou. Coverage, not targeting: A structural regime in multi-turn agent credit assignment. *arXiv preprint arXiv:2609.02417*, 2026. URL https://arxiv.org/abs/2609.02417.

A Reward and Environment Details

Reward composition. All arms share the same scalar reward

$$R = \underbrace{k/T}_{Y,\ \text{leading-turns progress}} + 0.5 \cdot \mathbf{1}[k{=}T] + 0.4 \cdot \text{overlap},$$

where k is the largest number of leading turns that all pass the checker (turn j is checked in the cumulative state of turns $1..j$), the completion bonus restores weight on the benchmark's own full-completion metric which the linear term under-prices, and overlap is the fraction of each turn's reference calls the policy also emitted (name and arguments matched), averaged over turns; this dense term is verifiable and involves no LLM judge. Assistant-turn and parallel-call caps 10/10 per rollout (16 assistant turns in the wall-shift and 8B reruns); response cap 10,000 tokens (6,000 in the long regime from step 175, a documented protocol change). Judge (judge arms only): a 35B open-weight model, one hindsight call per rollout.

Corpora and decontamination. The benchmark ships no RL training split, so training tasks are procedurally generated and checker-verified. The contamination audit (user-turn text, reference call sequences, initial states) removed 51 leaked tasks from a predecessor corpus and, for the prospective variant corpus, 3 of 599 gated tasks (2 initial-configuration collisions, 1 near-duplicate); both audits are released. Because arms differ only in the credit signal at fixed corpus, reward, and hyperparameters, corpus choice affects external validity, not internal comparisons.

Effect of the auxiliary reward terms (measured). Over the first 60 steps of both long-regime arms, **80.1% / 82.9%** of 16-rollout groups have identical k; under $R = Y$ these carry zero within-group advantage and dynamic filtering would discard them. The overlap term restores within-group variance in $\sim$10–11% of those groups; the completion bonus prices the near-miss-to-completion conversion at $0.5 + 1/T$, exactly the event that turns the crack-front into a cascade.

B Theory

Proof of Proposition 1. With $r_t = \beta c_t$ at turn ends and $r_{\text{term}} = R - \beta Y$, the return-to-go at turn t is $G_t = R - \beta P_{<t}$; group statistics are computed on totals $G_1 = R$, so $A_t = (G_t - \mu_R)/\sigma_R = A^{\text{GRPO}} - \beta P_{<t}/\sigma_R$. The sign condition follows. □

Theorem 1 (Objective of advantage-space injection)**.** *Let credits sum to the anchor, $\sum_t c_t = Y$, and let z_j use a single pooled group scale σ_C (the convention throughout; per-position standardization yields a position-reweighted objective instead, App. B). Then (i) treating the group statistics μ_C, σ_C as fixed (the idealized, infinite-group limit; in practice they are sample-dependent, and the finite-group estimator is self-normalized, like GRPO's own advantage), $\beta_a z_j$ is the return-to-go of the shaped per-turn reward $\beta_a c_t$ up to baseline and scale, so the idealized estimator is, up to the positive scalar $1/\sigma_R$, a policy-gradient estimator for $J_{\beta_a}(\pi) = \mathbb{E}_\pi[R + \kappa Y]$ with $\kappa = \beta_a \sigma_R/\sigma_C$; (ii) if every component of R is maximized only by complete success ($Y{=}1$) and the turn-pass dynamics are deterministic given the actions, the sets of optimal deterministic policies of J_0 and J_{β_a} coincide for every $\beta_a \geq 0$; under the weaker condition that some J_0-optimal policy attains $Y{=}1$ almost surely, for every $\beta_a > 0$, $\operatorname{argmax} J_{\beta_a} = \operatorname{argmax} J_0 \cap \{\pi : \mathbb{E}_\pi[Y]{=}1\}$, so J_{β_a}-maximizers remain J_0-optimal but the full argmax sets need not coincide; (iii) prefix-measurability of the credits (functions of the trajectory up to turn t, as the exact checker decomposition is) is a* sufficient *condition for the estimator to be unbiased for ∇J_{β_a}; without it, unbiasedness is not guaranteed, and for post-hoc credits (such as RWTP's, which depend on Y, hence on future actions) we exhibit a bias term proportional to β_a in an enumerable counterexample; we do not claim every non-prefix-measurable credit is biased.*

Proof sketch of Theorem 1. (i) $C_j = \sum_{k \geq j} c_k$ is the undiscounted return-to-go of the reward c_t placed at turn ends; standardization over the group subtracts a baseline (unbiased for the gradient of a fixed objective) and rescales by $1/\sigma_C$, while the GRPO term carries its own scale $1/\sigma_R$; factoring out $1/\sigma_R$ leaves the objective $R + \kappa Y$ with $\kappa = \beta_a \sigma_R/\sigma_C$ (verified numerically to 10^{-9}). (ii) If each component of R attains its maximum only at $Y{=}1$, then $R + \kappa Y$ is maximized exactly by the *trajectories* that maximize R; under deterministic turn-pass dynamics a deterministic policy

selects a single trajectory, so the argmax sets over deterministic policies coincide. When instead some J_0-optimal policy attains $Y{=}1$ a.s. and $\beta_a > 0$ (so $\kappa > 0$), $\max J_{\beta_a} = \max J_0 + \kappa$ and it is attained exactly by the J_0-optimal policies with $\mathbb{E}[Y]{=}1$ (using $Y \leq 1$), giving the stated intersection identity, containment rather than equality of the full J_0 argmax. Without either condition the trajectory-wise argument does not transfer to policy expectations; a stochastic commit-to-risk chain where preservation fails is below. (iii) Writing $C_j = Y - \sum_{i<j} c_i$, the estimator's credit term satisfies $\mathbb{E}[\sum_j s_j C_j] = \nabla\mathbb{E}[Y] - \mathbb{E}[\sum_j s_j \sum_{i<j} c_i]$ with $s_j = \nabla \log \pi(a_j \mid h_j)$, so $\mathbb{E}[\hat{g}] - \frac{1}{\sigma_R}\nabla\mathbb{E}[R + \kappa Y] = -\frac{\beta_a}{\sigma_C}\,\mathbb{E}\big[\sum_j s_j \sum_{i<j} c_i\big]$. If each credit is measurable when paid, the omitted past-credit sum is h_j-measurable and $\mathbb{E}[s_j \mid h_j] = 0$ makes the term vanish; the checker decomposition satisfies this. Bias arises when past-labelled credits, omitted from the return-to-go, depend on current or future actions, as post-hoc credits do (dependence of future returns on future actions is not itself a source of bias); the enumerable counterexample below exhibits this term, exactly linear in β_a. □

Toy-MDP verification and counterexample. All statements were checked numerically in exactly enumerable chain MDPs ($T{=}4, 6$ turns, two actions per turn, deterministic checker pass tables, the paper's own reward family $R = Y + 0.5 \cdot \mathbf{1}[k{=}T] + 0.4 \cdot \mathrm{overlap}$; script released, fixed seed). **(i)** For a softmax policy with exact score-function gradients over all 2^T trajectories, the idealized pooled-σ_C estimator matches the policy gradient of $\frac{1}{\sigma_R}\mathbb{E}[R + \kappa Y]$ to residual $\sim 10^{-16}$, and the implied constant equals $\beta_a\sigma_R/\sigma_C$ to 10^{-9} (the earlier draft's $\kappa = \beta_a/\sigma_C$ was off by the factor σ_R). **(ii) counterexample.** Instance B is a $T{=}4$ chain whose first action commits to a mode: *risky* (each turn passes w.p. 0.8; $\mathbb{E}[R] = 0.795$, $\mathbb{E}[Y] = 0.590$) or *safe* (deterministic $k{=}3$; $\mathbb{E}[R] = 0.750$, $\mathbb{E}[Y] = 0.750$). Every R component is maximized only on $Y{=}1$ trajectories and $Y{=}1$ is reachable, yet exhaustive policy enumeration shows the J_κ-argmax switches from risky to safe at $\kappa > 0.283$: component-wise reward alignment and reachability alone are insufficient without the additional conditions in the theorem, which the repaired statement therefore includes. A per-turn-choice stochastic variant does *not* break preservation (the optimal policy plays safe early and risky late and dominates both objectives); the commitment structure is essential. **(iii)** Prefix-measurable credits are unbiased to machine precision; uniform post-hoc credits ($c_t = Y/T$) give a relative gradient bias of 2.7–7.1% over $\beta_a \in [0.25, 2]$ that is *exactly* linear in β_a (residual 4×10^{-15}). Instance C violates the component assumption instead (overlap maximized on a $Y{=}0.5$ trajectory, $Y{=}1$ infeasible) and breaks preservation at $\kappa \geq 0.2$, showing that assumption is also necessary. **Standardization convention.** The finite-N self-normalized estimator converges (cosine 0.957 at $N{=}2$ to 0.9996 at $N{=}64$) to the *per-turn*-standardized limit; with $\sigma_C(t)$ varying $6\times$ across positions in Instance A, that limit is a position-reweighted credit objective (cosine 0.994 to the pooled form) rather than $\mathbb{E}[R{+}\kappa Y]$ for any κ, which is why the theorem fixes the pooled-σ_C convention explicitly; the finite-N toy is therefore a diagnostic of the per-turn-standardized idealized estimator, and none of these checks establishes unbiasedness of the implemented finite-group pooled update.

Gauss–Markov form, anchor sensitivity, never-lose selection (RWTP). For positive-definite $\widehat{\Sigma}$, the projection is the conditional mean of $\mathcal{N}(d, \widehat{\Sigma})$ given $\mathbf{1}^\top c = Y$; under correctly specified Σ it is unbiased for $c^\dagger$ with covariance $\Sigma - \Sigma\mathbf{1}(\mathbf{1}^\top\Sigma\mathbf{1})^{-1}\mathbf{1}^\top\Sigma \preceq \Sigma$. Anchor sensitivity: $\partial c^*/\partial Y = \Sigma\mathbf{1}/(\mathbf{1}^\top\Sigma\mathbf{1})$, so anchor errors flow to *uncertain* turns (graceful degradation, verified to 30% anchor label error). Special cases: $T{=}1$ gives $c{=}Y$; constant σ recovers uniform projection; $\sigma_t \to 0$ pins turn t. Never-lose shrinkage: $\widehat{\sigma}^2_\gamma = (1-\gamma)\widehat{\sigma}^2 + \gamma\,\overline{\widehat{\sigma^2}}$ with γ grid-selected on the calibration split; $\gamma{=}1$ *is* uniform projection, so the calibrated estimator's expected credit error is at most uniform's up to grid selection error. Verified in 25/25 synthetic noise×reliability regimes (unshrunk noisy reliabilities lose up to 12.8% credit MSE under weak heterogeneity; calibration selects $\gamma{=}1$ under shuffled or inverted reliabilities; calibrated gains $+23$–26% under genuine heterogeneity, decaying with horizon from $+55\%$ at $T{=}2$ to $+12\%$ at $T{=}16$).

C The Judge-Credit Pipeline and the Replay Audit

This appendix retains, as supporting methodology, the auditable judge-credit pipeline (RWTP: reliability-weighted terminal projection of judge increments, with calibrated never-lose shrinkage) and the executable counterfactual audit instrument (deletion credit, gold-splice deviation cost, exact Shapley by 2^T replays; replay engine agrees with the benchmark checker on 120/120 probes).

Findings used in the main text: judge credit violates the support constraint (17.6% of mass after the first failure, 63% of it positive; marginal-turn payment 44% lower than the decomposition's (0.056 vs. 0.100); $+0.53$ total over-claim); support masking raises rank correlation with the decomposition from 0.64 to 0.92 and helps when it fires (onset 195 vs. 225) without repairing reliability (1/3 seeds). Audit-methodology findings: accuracy-against-Shapley claims are dominated by target front-loading (a judge-free positional prior reaches 0.84 and beats every judge estimator on absorbing-outcome targets; each target's position-correlation must be reported); the judge's genuine information is *where damage occurs* (best on the gold-splice target, 0.62, where the positional prior is -0.40); published mechanisms audited: outcome broadcasting carries zero localization by construction, multiplicative advantage rescaling anti-localizes on front-loaded targets (-0.12).

Calibration data and pipeline. 850 trajectories with exact replay labels (600 perturbed-reference across six perturbation types; 250 on-policy rollouts), split 70/15/15 by task. RWTP projects judge increments onto the outcome (Gauss–Markov step with calibrated never-lose shrinkage; the conditioning identity is shared with GP-based return redistribution Xiao & Yu, 2026 and we claim no novelty for it), debiases with a ridge head on runtime transcript features, and self-disables under shuffled or inverted reliabilities.

Table 4: Rank agreement (Spearman) of credit estimators with five replay-computed targets on the held-out test split (96 trajectories). Bottom row: each target's own rank correlation with $-t$ (front-loading). Judge-free rows in italics.

estimator	$\text{Shapley}_{\text{absorb}}$	$\text{Shapley}_{\text{any-turn}}$	checker	gold-splice	deletion
judge, uniform projection	0.19	0.41	0.60	**0.62**	0.19
RWTP (all features)	0.76	0.61	0.60	0.14	0.76
RWTP w/o position features	0.31	0.48	**0.73**	0.36	0.31
positional prior, no judge	**0.84**	0.62	0.57	-0.40	**0.84**
features-only ridge, no judge	0.84	**0.62**	0.57	-0.36	0.84
checker vector$/T$ (uses checker)	0.54	0.60	0.70	0.67	0.54
target's correlation with $-t$	0.84	0.62	0.57	-0.40	0.84

Table 5: Published credit mechanisms audited against replay targets (same 96 trajectories).

mechanism (source)	$\text{Shapley}_{\text{absorb}}$	checker	gold-splice
outcome broadcast (Flow-GRPO)	n/a (constant)	n/a	n/a
multiplicative advantage rescale (CW-GRPO)	-0.12	0.19	0.32
raw judge (uniform projection)	0.19	0.60	0.62
RWTP w/o position features	0.31	0.73	0.36

Reading Table 4 row-wise: reported alone, "0.17→0.77 against Shapley" would have measured target front-loading, not judge correction, and we retract that reading of our own earlier results. Every credit-accuracy claim should report several targets, each target's position-correlation, and judge-free prior rows. Reading Table 5: outcome broadcasting is constant within a trajectory (zero localization by construction); multiplicative rescaling anti-localizes on front-loaded targets because the advantage's sign inverts the within-trajectory ranking on below-baseline trajectories; neither defect is visible to validation by RL score.

D THE SHORT REGIME: WHERE DENSE CREDIT IS A TAX

Setup as in §2 with 3–5-turn tasks (two domains held out; starting success 29%); final checkpoints additionally scored on a powered set of 1,242 fresh tasks (± 2.8 points at 95%).

In-distribution: dense credit is a tax. Every credit arm, including exact credit through the scale-safe advantage-space injection, scores below terminal-only GRPO at matched steps (-4.1 to -7.4

Table 6: Short regime (3–5 turns), full-task success at step 160. Left: frozen validation (n=143, $\pm$8.2 at 95%; the three 43.4 entries are three independent runs at 62/143). Right: powered validation of final checkpoints (n=1242; task-level 95% intervals $\pm$2.8; descriptive comparisons, 1–2 training seeds per arm). Per-run sources ship with the code.

arm	val@160	arm	powered (n=1242)
GRPO (terminal-only)	50.3	GRPO	**47.8**
RWTP (s42 / s43)	47.6 / 45.5	RWTP (s43 / s42)	43.7 / 40.4
uniform projection	43.4	uniform	41.5
decomposition (reward-space)	43.4	decomposition, advantage-space	41.5
decomposition, advantage-space	43.4		

Table 7: Decision-rule comparison at the frozen budget R' on discovery runs. CONTINUE/STOP per rule; truth = sustained ignition by budget end; false kill = stopping a run that ignites. The near-miss rate is the strongest decision-rule comparator.

rule	correct	false kills	miss kills
crack-front monitor (frozen)	**14/15**	**0**	1
validation-plateau stop	10/15	3	2
mean-progress stop	7/15	0	8
near-miss-rate stop ($<$1%)	14/15	1	0

points; descriptive, 1–2 seeds per arm). Rollout forensics show the ordering already on the training distribution (train full-pass 50.1 / 44.4 / 43.6 / 41.7 for GRPO / RWTP / decomposition / uniform), suggesting the deficit is not solely a generalization effect; these checkpoint-level comparisons do not establish robustness across training seeds. With 16-rollout groups on 3–5-turn tasks the within-group comparison already carries the failure-location information (GRPO as implicit process reward Sullivan & Koller, 2026); additional credit perturbs the advantage and costs rate. A 2B model on this regime's development corpus shows the pathology's parameter-dependence: there uniform projection was catastrophic (train full-pass 2.2% vs. 14.8%) while judge credit was best in-distribution yet tied on the public benchmark. Seed caveat (frozen in the preregistration): the powered comparison rests on 1–2 seeds per arm.

The public benchmark disagrees, with large seed variance. On the locked public benchmark (n=800, single pass, zero ground-truth-error rate), the base model scores 35.1; after 160 short-regime RL steps, terminal-only GRPO scores 32.6, RWTP 33.0 and 39.8 in its two seeds, uniform 33.9, decomposition arms 30.1–33.3. Two honest readings: most RL arms land at or below base (short-horizon RL on procedural data mostly *costs* public-benchmark performance), and the judge-credit arms cost the least, with one seed clearly above base. In-distribution, the evaluated credit-trained checkpoints score lower; out-of-distribution, only a judge-credit checkpoint exceeds the base model; both are reported, and the 6.8-point spread between seeds of the same method bounds what any single-seed public-benchmark comparison can show.

E Long-Regime Forensics and Preregistration

Frozen monitor text and amendment log. The preregistration document (released verbatim, with its git history; timestamps are author-maintained git records, not an independent timestamping service) was registered at 2026-09-16 15:00 UTC (`1174835`) and amended six times the same afternoon, every amendment before any prospective-cohort event: A1 15:13 (`51b014e`; comparator baselines, five-metric scoring, disclosure of already-seen data, separation of precursor and restart conclusions); A2 15:19 (`7978bbc`; short-regime replication lock, later withdrawn pre-launch); A3 15:30 (`1b1a54d`; long-horizon refocus, the withdrawal disclosed, final frozen run list including the variant pairs); A4 15:35 (`385cb03`; explicit pairing map, original distribution = independent seeds, variant = true pairs, 240-step primary horizon, corpus-gate criteria that exclude wall identifiability, offline policy-evaluation wording); A5 15:40 (`ff1b9d2`; signal-vs-policy separation,

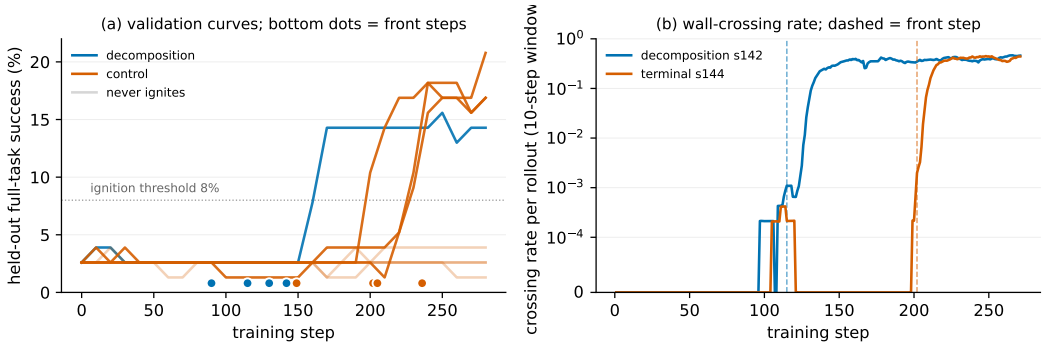


Figure 4: The anatomy on the discovery cohort. (a) Held-out validation for the eight monitor-applicable runs with complete logged histories (decomposition blue, controls vermillion; faded = never ignites); seven further runs lost early validation history to a disclosed logging incident and are omitted, their full event series shipping with the release. Bottom dots mark *every* fronting run's crack-front step, curve drawn or not (front times come from the complete rollout logs): wherever the surviving validation history covers a front step, validation is still at the floor there, and fronts precede ignition in all seven fronting runs with an observed sustained ignition (median lead 45 steps, range 25–160); the eighth fronting run, a control, fronted at 236 and was censored at budget end mid-ignition. (b) Wall-*crossing rate* (cracks per long-horizon rollout, 10-step windows) for one igniting run per group: the rate erupts from $\sim 4 \times 10^{-4}$ before the front to a ~ 0.4 post-front plateau. Dashed lines mark the retrospective front onsets; the corresponding online alarms occur nine steps later. Raw counts are exposure-dependent (an incident replay burst near step 175 inflates one run's counts $\sim 19\times$ while leaving the rate unchanged, App. F), so the rate is displayed.

Table 8: Run-level event table (discovery runs; referenced from §2). Front (frozen rule $w{=}10, c{=}5$) and sustained ignition (3 consecutive validation readings $>8\%$; the onset is the first of the three) reported separately; 'cens@s' = right-censored at step s. The threshold-sensitivity grid ships with the released event series.

run	group	front	sustained ignition	last step
decomp s142	decomposition	115	175	280
decomp s143	decomposition	130	175	280
decomp s145	decomposition	none (cens@280)	none	280
decomp s146	decomposition	none (cens@280)	none	280
decomp s147	decomposition	90	250	280
decomp s342 (AS)	decomposition	142	170	280
terminal s142	control	236	none (cens@280)	280
terminal s143	control	none (cens@280)	none	280
terminal s144	control	202	230	280
judge s242 / s243	control	none / none	none / none	280
judge s244	control	205	230	280
masked s442	control	149	200	280
masked s443 / s444	control	none / none	none / none	280

matched-time/matched-progress conditional test, fixed-deadline comparator); A6 15:59 (`82bf6db`; run-count and wording precision, budget accounting rules). The variant-corpus wall identification (turn 6, $q_6 = 8 \times 10^{-5}$ pooled, both arms independently) was recorded 2026-09-17 14:25 UTC (`0352b1f`) from steps 1–40 only; the record notes one protocol deviation, disclosed there: the spec asked for the record before run-step 60 and it was written at run-step $\sim$78, before any monitor read-out of those logs (the readout instrument had never been run on them; the earliest historical front is step 90).

Full event table and threshold grid. Table 8 (App. E) lists the discovery cohort. Additional rows carried with flags rather than dropped: a shuffle-corruption arm, s544, and terminal s145 are

STORM-POLLUTED (a 2026-09-15 infrastructure incident re-rolled early steps 100+ times, making raw-log front times artifacts; the exclusion rule was fixed and disclosed before any prospective readout); *decomposition N4 s942* (group size 4) is monitor-N/A, its corpus contains no $\geq$8-turn tasks, so cracks are undefined and the monitor abstains by rule. The 3×3 threshold grid (window $\{5, 10, 20\}\times$count $\{3, 5, 10\}$) yields, at every setting: 7/7 igniting runs with a preceding front, 0/7 flat runs with a false alarm, 1 censored-after-front run (terminal s142: front 236, first crossing 270, budget end 280 before three consecutive readings), median front→ignition lead 22 (w5,c10) to 62 (w20,c3) steps, 45 at the frozen (w10,c5).

Decision-rule detail. At R'=162/256: the monitor's CONTINUE set is exactly the front set; validation-plateau (any reading above the 2.6% floor by R') stops decomposition s142/s143/s147, all of which ignite (fronts 115/130/90, first crossings 175/175/250, all after R'=162), and continues judge s243 and masked s444, which never ignite; mean-progress (group mean turns-passed at $[R'-9, R']$ vs. $[31, 40]$) continues all 15 runs (mean k drifts up everywhere, e.g. $3.59 \rightarrow 3.87$ in a flat run); near-miss ($\geq$1% rate in $[R'-19, R']$, the statistic's own 20-step window) errs once, stopping decomposition s342 (rate 0.6%) which ignites at 170. Operationalization choices not fixed by the frozen text (window widths for mean-progress and near-miss; literal "above floor" > 0.026, with a $\times$2-margin variant reported as sensitivity) were made once, before any prospective readout, and are documented in the released script.

Budget tables. $\mathbb{E}[\text{total}] = \mathbb{E}[L]/q$ per the frozen Metric-4 formula; censored-after-front attempts count as failures at full cost B=280; the fixed deadline D is grid-searched on the same runs in the comparator's favor; bootstrap is run-level (20k resamples) with q=0 resamples reported, not dropped. Decomposition arms (R=162, 6 runs): monitor **274** [186, 537] (q=0: 0.1%) vs. best fixed D=250: 318 [186, 712] vs. no restart: 332 [186, 772]. Control arms (R=256, 9 runs): monitor 740 [351, 2326] (q=0: 2.8%) vs. best fixed D=230: **680** [335, 2070] vs. no restart: 780 [360, 2470]; the conservative censoring treatment favors the tuned deadline on controls, and we report it so. Online-cost variant (alarm $\leq R$ decides; the confirmation step, the third consecutive above-threshold validation reading, is the success cost): decomposition 291 [203, 554] vs. best fixed D=270: 345 vs. no restart 350; controls 760 vs. best fixed D=250: 740 vs. 800; every ordering is preserved.

Survival analysis. Run is the statistical unit; runs that never reach an event are right-censored at their last step. Kaplan–Meier medians and 10k-permutation log-rank on the discovery cohort (storm-polluted runs excluded by the pre-fixed rule): front initiation, decomposition median 130 vs. control median not reached, p=0.17; ignition, median 175 vs. not reached, p=0.17 (n=6 vs. 9; an earlier draft cited p=0.12 from a pre-completion snapshot that predated one control's late ignition, corrected here). Both are trends at this run count and are reported as such; the prospective cohort is the designed follow-up.

Extension cohort readout (2026-09-26, post-cutoff, same frozen rules; supersedes the 2026-09-24 readout). Additional runs outside the preregistered ten-run cohort, launched under the same frozen rules, are labeled extension and analyzed separately; none is merged into the registered cohort. *Second distribution.* The seed-4244 pair is right-censored at its last recorded steps (192 and 200) with no front and no sustained ignition, two front-free non-igniting observations. Pair D (seeds 4245): the *terminal* arm fronted at step 129 (online alarm 138) with validation at the floor through the alarm, and sustained ignition began at step 150 (17.2% at steps 150–180; online lead 12, onset lead 21), the run still training at step 184 at this readout; the decomposition arm shows one isolated crack, no front, and flat validation through step 173 (censored). Pair D therefore splits terminal-first, as pair A did, and contributes a tenth front-preceded ignition outside the registered cohort, with a lead again below the frozen 30–60 window. Across the prospective and extension cohorts the tally is: ten ignitions, all front-preceded; four front-free runs, none igniting by their cutoffs; zero false alarms (the discovery set separately contributes seven front-preceded ignitions and seven front-free flat runs, Table 8). *Displaced wall* (third corpus; crack = turns_passed $\geq$ 9 on tasks of $\geq$ 11 displaced turns, the original $\geq$ 8-turn filter shifted by the prefix): the attempt-2 pair completed 200 steps; the decomposition arm fronted at 163 (alarm 172) and its validation reached 17.2% at step 180 and 20.7% at step 200, but the 170 and 190 readings are missing from the released series, so the sustained rule cannot be scored (censored, rising); the terminal arm logged one displaced crack and stayed flat. A further pair (seeds 5244): the decomposition arm completed 200 steps with zero dis-

placed cracks and flat validation; the terminal arm is flat through step 139 (two displaced cracks at that step, no front; censored). The derived files for the displaced-corpus runs carry `cracks9_lh11` (displaced-wall cracks under the ≥ 11-turn filter), so these statements regenerate from the package.

Second-policy extension (Qwen3-8B): abstention. An exploratory 8B pair (only the wall-identification procedure carries over frozen; no other 8B rule was preregistered) reached 143–145 steps by the readout. The identification procedure *abstained*: the turn-6 conditional pass count is exactly zero across 16,474 attempts while turns 1–5 decline smoothly (0.89 to 0.70), and the full-run maximum of turns-passed is 5 over 72,520 rollouts with tool-call counts piled at 5–7, the same infrastructure signature (an exhausted assistant-turn budget under a more verbose model) that invalidated the first wall-shift attempt. The abstention was applied on inspection: an exactly zero conditional pass count is grounds to investigate infrastructure limits (no numerical abstention threshold was preregistered), and execution records confirmed the same assistant-turn budget boundary as the first wall-shift attempt; a repeat with the corrected 16-turn budget did not produce training steps (repeated worker deaths at step 0; the run keeper stopped under its escalation rule), so the 8B extension remains at abstention and is not interpreted here.

Pre-registered predictions scored to date. Before extending the first long-regime runs from 160 to 280 steps we registered: (1) the ignited seed keeps rising (confirmed); (2) the second decomposition seed converts (confirmed, at step 150); (3) terminal-only stays at floor (refuted: it ignited late in 2 of 3 seeds), and this refutation converted the claim to its rate-and-reliability form. The frontier-attenuation preregistration predicted *no* ignition and observed *late* ignition, weakening the binary reading to the graded one. Two early single-seed verdicts (in both directions) were reversed by replication.

Real-return mechanism recompute (correction record). The offline scheme comparison was first computed with the outcome component (binary completion flag) standing in for the training return. Re-verifying against the logged composite return $R = k/T + 0.5 \cdot \mathbf{1}[k{=}T] + 0.4 \cdot \text{overlap}$ (identity holds on 100% of 1.31M raw log records, 1.15M after lineage dedup, within 2.5×10^{-4}, the logging precision): (i) the earlier statement that terminal broadcasting zeroes the direction signal for 92.7% of rollouts with nonzero exact advantage is an artifact of the binary stand-in and is withdrawn; under the real return the terminal whole-return advantage correlates $+0.99$ with the exact-credit scalar, near-zero cases are 0.07%, and sign disagreements 1.5%. The corrected reading strengthens the placement account: the schemes are near-equivalent per rollout and differ in intra-trajectory placement. (ii) The spread-coupling of reward-space exact credit's frontier contrast is -0.49 under the real return (-0.75 under the simplified probe); direction unchanged, magnitude weaker. Both original numbers replicate exactly under the original definitions; the discrepancy is definitional and disclosed here.

Judge support-violation statistics. 120 real long-regime rollouts, both pipelines replayed offline: 17.6% of judge credit mass beyond the passed prefix, 63% positive; marginal-leading-turn payment 0.056 vs. 0.100; total over-claim $+0.53$; on $Y{=}0$ trajectories the judge injects $|c| = 0.28$ of pure noise.

F Compute and Incidents

Training totals at the data freeze: $\approx$5,200 GPU-hours across the full campaign (10,363 logged training steps over all runs on both hosts, at measured per-step times of 15.0 min on H200 and 14.5 min on B200 lanes of 2 GPUs; this counts costs as experienced, including incident downtime absorbed into step times). Judge calls: 7.4×10^5, one hindsight call per rollout in judge arms only; the monitor and the decomposition add zero calls, zero extra inference, and zero extra environment interaction. Calibration data collection was one-time (850 labeled trajectories; accuracy saturates by $\sim$50 tasks / 300 trajectories). Protocol changes and incidents are recorded in the released ledger: the response-cap change at step 175 (uniform across long-regime runs; in the 40 released runs spanning it, per-rollout passed-turn depth, crack rate, and near-miss rate show no discontinuity between steps 155–174 and 176–195: flat runs stay at zero cracks with depth shifts under ± 0.2 turns of mixed sign, and active runs continue smooth trends); epoch extensions; two judge-relocation windows degrading $\sim$500 rollouts each of two judge-arm seeds; retry-storm pollution rules with the two excluded runs; the

late start of terminal s147 (a stale queue process holding a pre-amendment stage list); the wall-shift transfer's first attempt, invalidated by an assistant-turn budget artifact that the wall-identification protocol itself exposed (exact-zero conditional pass rate at the harness boundary), documented with the fix; and one corruption-probe record (attenuate seed 546) whose released series (front 150, sustained ignition 170, last released step 229; a cross-host continuation is not merged) supersedes an earlier internal note (front 216, sustained 225) that the released data do not reproduce.