VERIFIER-INDUCED STRATEGY COLLAPSE: FAMILY-CORRELATED FALSE NEGATIVES ELIMINATE VALID STRATEGIES WHILE CAPABILITY METRICS IMPROVE

Anonymous authors
Paper under double-blind review

ABSTRACT

When a terminal verifier in agentic reinforcement learning falsely rejects one strategy family, group-relative policy optimization (GRPO) does not merely mis-rank that family: it deletes it from the policy — in our experiments within about 30 steps — while every standard capability metric improves. The cause is the correlation structure of the checker's false negatives (FN), not their rate: recent theory models verifier noise as independent across solutions, an abstraction under which the deletion cannot occur. For GRPO with a terminal checker that rejects a fraction r_f of valid rollouts of strategy family f, we derive — in the tabular family-mixture abstraction — the exact finite-group dynamics: the expected one-step drift of a family's probability mass is $m_f(q_f - \bar{q})$ up to a positive factor, where $q_f = s_f(1 - r_f)$ is the family's verifier acceptance rate and $\bar{q}$ the group-mean acceptance, so selection reduces to a single threshold equal to mean verifier quality. The aggregate FN rate is provably insufficient to preserve strategy ordering, and a family goes extinct once $r_f > 1 - \bar{q}/s_f$. In a controlled tool-chain environment with an exact success oracle, a checker with family-correlated false negatives drove the targeted valid family from a third of the population (banking to ~ 0.5) to extinction within about 30 GRPO steps while true success rose from 0.39 to ≈ 1.0 and the checker's measured reliability converged to perfect. Matched-iid controls at the identical aggregate error rate, five-seed and 7B replications, a second model family, stress tests showing real capability loss on route-blocked and task-shifted environments, family-resolved audits of published GSM8K, HumanEval, and MATH-500 checker configurations support the theory and mark its scope.

1 INTRODUCTION

Agentic reinforcement learning with verifiable rewards reduces supervision to a terminal checker's verdict, whose error profile is thus the training signal's error profile. Audits show real verifiers — unit-test oracles, execution harnesses, model-based judges — carry substantial false-negative (FN) rates on valid solutions (Xin, 2026; Egashira et al., 2026; Xu et al., 2025; Cai et al., 2025); recent theory analyzes how such noise deforms GRPO (Guo et al., 2025). These analyses share one abstraction: checker errors are *independent* across solutions, indifferent to *how* a solution works — a falsely rejected valid solution is treated exactly like an invalid one.

Real checkers develop structured blind spots: they reject solutions using an unusual tool, an unfamiliar route, or a minority style — errors attaching to a *strategy family* rather than to individual rollouts; the family is the unit the checker fails on. Under family-conditioned false negatives, group-relative advantage stops being fair competition. Falsely rejected members of a hostile family are punished precisely when they co-occur with accepted members of favored families, the penalty growing with the accepted majority; the pressure compounds into directed elimination.

Our contributions are threefold. **(1) Exact dynamics.** For GRPO with family-conditioned FN rates r_f, we derive the expected one-step family-mass drift exactly for every finite group size (Theorem 1): selection reduces to a single moving threshold, the group-mean verifier acceptance $\bar{q}(m)$,

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.

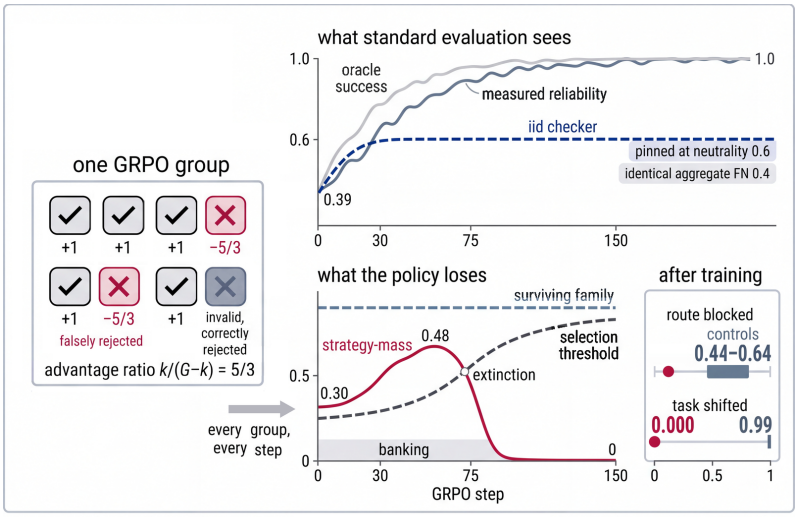

Figure 1: On one training timeline, every standard metric improves while a valid strategy family is deleted. Left: the mechanism — one GRPO group (G=8): the checker falsely rejects valid rollouts of the targeted family, punished $k/(G-k) = 5/3$ times harder than any accepted reward. Top band: what standard evaluation sees — oracle success and the corr checker's measured reliability rise toward 1.0, while the matched-iid checker stays pinned at neutrality 0.6. Bottom band: what the policy loses — the targeted family's mass banks, then collapses once group-mean acceptance $\bar{q}(t)$ crosses $q_{f_1} = s(1-r)$ (extinction), absorbed by the surviving family. Right: after training — standard evaluation passes every check, while route-blocked (0.125 vs. 0.44–0.64) or task-shifted (0.000 vs. 0.99) evaluation exposes the loss. Schematic; exact trajectories in Figure 2.

with extinction boundary $r_f > r_f^* = 1 - \bar{q}/s_f$. The threshold is endogenous — it rises as training drains invalid modes and strengthens survivors — producing *banking-then-collapse* trajectories. **(2) Insufficiency of the aggregate error rate.** No bound on the aggregate FN rate preserves the true strategy ordering (Proposition 2): a family-correlated profile at any small aggregate rate can invert the ordering that a family-uniform profile at the same rate preserves. **(3) Controlled evidence and an audit protocol.** In a categorical tool-chain environment with an exact success oracle, a 1.5B model trained with GRPO under a family-correlated checker loses a valid family holding $\sim$30% of the initial population while oracle success rises $0.39 \to \approx 1.0$ and the checker's *measured* reliability converges to perfect — the verifier masks itself as its victims disappear (Corollary 2). A matched-iid checker at the identical aggregate FN rate preserves the family, isolating family structure as the cause. A family-resolved audit protocol — built on a harm metric exact in a categorical simulator — extends the diagnosis to stress tests, checker repair, and published benchmark checkers.

Figure 1 previews the argument as one paradox on a single training timeline.

2 Related Work

Verifier noise and RLVR dynamics. Verifier-guided post-training has moved from scalar reward models to checkable signals — unit tests, execution harnesses, rubric and LLM judges (Zhao et al., 2026; Cai et al., 2025; Kwok et al., 2026) — and audits document substantial FN mass on valid solutions: category-level heterogeneity and convergence-delaying systematic errors (Xin, 2026; Egashira et al., 2026), lightweight-judge FN rates (Xu et al., 2025), cross-lingual selection bottlenecks (Zhou et al., 2026). On the training side, GRPO (Guo et al., 2025) — whose group standard-deviation normalization its main variants share (Bay & Yearick, 2026) — popularized group-relative advantages with verifiable rewards, and a growing body of work analyzes its robustness to noisy rewards and reward gaming (Radhakrishna Palandye et al., 2026; Zhang, 2026; Helff et al., 2026; Ichihara et al., 2025), gradient corrections that presuppose noise independence (Mansouri et al., 2025), and variance or group-composition effects (Wang et al., 2026a; Yu et al., 2026; Zhang et al., 2025; Gao et al.,

2026; Jiang et al., 2026; Dash et al., 2026; Vo et al., 2026). Closest to us, RLVεR (Rad et al., 2026) derives GRPO replicator dynamics over reasoning modes with mode-independent FN/FP rates, governed by a single scalar (Youden's index): with $J > 0$ noise is "rate, not fate." Our Proposition 2 identifies the scope of that criterion: once FN rates are family-structured, the aggregate rate is no longer a sufficient statistic, and a checker with $J > 0$ and zero FPR can eliminate a valid family (§5); the mode-uniform law of Rad et al. (2026) is the special case of Theorem 1 in which all families share one r_f.

Support contraction and diversity collapse. Wei et al. (2026) document the cross-task analogue under a correct verifier: later-objective support depletes while pass@1 improves. Our setting differs in the source and observability of the loss: under a fallible checker the deletion is within-task, driven by the checker's own error profile, masked by self-masking (Corollary 2), and invisible to best-k readouts until a shift demands the family; the boundary (Theorem 1), audit protocol, and stress tests are new. Under correct rewards the same contraction appears: RLVR raises small-k performance while narrowing the large-k reasoning boundary (Yue et al., 2025; Zhou, 2026), traced to GRPO's distribution sharpening (He et al., 2025); UCPO (Lochab et al., 2026) attributes diversity collapse to objective indifference over correct solutions plus stochastic concentration; ours is directional selection by the checker's error profile; Wang et al. (2026b) rectify irreversible recursive support contraction with support-constrained updates, the repair-side complement of our reopening analysis (Appendix F.2); mode or advantage collapse under group-relative objectives (He et al., 2026; Li et al., 2026) is the aggregate face of the same hazard.

Reward misspecification, monitoring, and hidden failures. That headline signals look healthy while failures hide is an active concern: implicit reward hacking measured through reasoning effort (Wang et al., 2025), red-teamed weak-to-strong monitoring (Kale et al., 2025), and Bayesian non-negative reward modeling against spurious reward bias (Duan et al., 2026). Our stress tests instantiate it: aggregate monitors improve while the audited capability disappears (§6).

3 Exact Dynamics of Group-Relative Selection under Family-Correlated Verifier Errors

This section derives, exactly and for every finite group size, how a family-conditioned verifier error profile reshapes group-relative policy optimization (GRPO; Guo et al., 2025). Exact statements are proven for all finite G from the objective; additional assumptions are marked as hypotheses and assigned to dedicated experiments (§5, §6). The law is deliberately minimal: exactness, not generality, is what fixes the sign, the boundary, and the clock that §4 and §5 must match.

3.1 Setting and Abstraction Conditions

The trajectory space of a task instance x is partitioned into K valid strategy families plus an invalid mode $\perp$; families are defined by the strategy implemented, never by verifier outcome. **(A1)** the G rollouts of a group are i.i.d. draws from $\pi(\cdot \mid x)$; **(A2)** given family f, outcomes are i.i.d. with true success s_f and family-conditioned false-negative rate r_f on valid rollouts (FPR zero); **(A3)** binary rewards $R_i = V(\tau_i)$; **(A4)** tabular mixture parameterization $\pi(f \mid x) = m_f(x)$, token-level sharing deferred to §7.

Family f's verifier-visible acceptance rate is $q_f = s_f(1 - r_f)$; the group-mean acceptance is $\bar{q} = \sum_j m_j q_j$ over all modes ($q_\perp = 0$).

3.2 Event Structure: Punishment Requires a Mixed Group

Throughout, $a_i = (R_i - k/G)/\hat{\sigma}_G$ with $\hat{\sigma}_G = \sqrt{k(G-k)/(G(G-1))}$ the unbiased group std (advantages zero in degenerate groups); any positive normalization preserves signs. The implementation uses TRL's convention, $\hat{\sigma}_{\mathrm{MLE}} = \sqrt{k(G-k)/G}$ with an inert 10^{-4} floor — a constant $\approx 1.07\times$ advantage rescaling that rescales drift speeds only; every sign condition, threshold, and boundary below is invariant.

Lemma 1 (Event structure)**.** *For any realized group with $\hat{\sigma}_G > 0$: $a_i < 0$ iff $R_i = 0$ and the group contains at least one verifier-accepted rollout.*

A valid-but-falsely-rejected rollout is punished iff its group is mixed.

3.3 One-Step Family Drift: A Replicator Law with a Moving Threshold

Define the group-dependent magnitude factors

$$\alpha_+(G,\bar{q}) = \mathbb{E}_{K\sim\mathrm{Bin}(G-1,\,\bar{q})}\left[\sqrt{\tfrac{(G-1-K)(G-1)}{(1+K)G}}\right], \tag{1}$$

$$\alpha_-(G,\bar{q}) = \mathbb{E}_{K\sim\mathrm{Bin}(G-1,\,\bar{q})}\left[\sqrt{\tfrac{K(G-1)}{(G-K)G}}\right], \tag{2}$$

the mean $|a|$ of an accepted / rejected member ($+1$ success absorbed in the binomial index accounts for the focal member).

Theorem 1 (Exact finite-G family drift)**.** *Under A1–A4, one GRPO step changes the expected mass of family f by*

$$\mathbb{E}[\Delta m_f] \;=\; \eta\, m_f\big(\alpha_+ + \alpha_-\big)\big(q_f - \bar{q}\big), \tag{3}$$

exactly, for every finite G.

Proof sketch. Conditioning on $i \in f$ leaves the other $G-1$ members i.i.d. mixture draws (A1), giving $\mathbb{E}[a \mid i \in f] = q_f\alpha_+ - (1-q_f)\alpha_-$; the within-group zero-sum identity $\sum_i (R_i - k/G)/\hat{\sigma}_G = 0$, averaged over a uniformly random member, yields $\bar{q}\,\alpha_+ = (1-\bar{q})\,\alpha_-$. Full proof in Appendix A. □

Corollary 1 (Single selection threshold)**.** *Family f gains mass iff $q_f > \bar{q}$ and loses mass iff $q_f < \bar{q}$, for every finite G. Group-relative selection is governed by a single moving threshold, the group-mean verifier quality $\bar{q}(m)$, itself a function of the population mixture.*

Proposition 1 (Softmax parameterization; irreversibility)**.** *(i) Under A1–A3 with the softmax parameterization $m_f = e^{z_f}/\sum_g e^{z_g}$ and per-sample policy-gradient updates on the family logits, $\mathbb{E}[\Delta z_g] = \eta\, m_g \mu_g$ with $\mu_g = (\alpha_+ + \alpha_-)(q_g - \bar{q})$ — exact at the logit level for every η (exact one-step mass map: $m'_f = m_f e^{\Delta z_f}/\sum_g m_g e^{\Delta z_g}$); to first order in η,*

$$\mathbb{E}[\Delta m_f] = \eta\, m_f(\alpha_+ + \alpha_-)\Big(m_f(q_f - \bar{q}) - C(m)\Big), \quad C(m) = \sum_g m_g^2(q_g - \bar{q}), \tag{4}$$

with sign condition $q_f > \bar{q} + C(m)/m_f$: the threshold shift is family-size-normalized, and interior fixed points coincide with the tabular ones. (ii) If $q_f < \bar{q}$ on the survivor-supported mixtures, extinction is absorbing under both laws — exponential decay at rate $\eta(\alpha_+ + \alpha_-)(\bar{q} - q_f)$ tabularly, algebraic $\sim 1/t$ decay under equation 4, whose drift on a rare family scales as m_f^2: punishment is throttled by the family's own frequency. (iii) If the checker is later repaired to a common acceptance rate across valid families, valid-family mass shares are invariant (tabular) or shrink for the smaller family while invalid mass drains (softmax): the deleted family never regains its pre-collapse standing under the repaired reward signal alone; restoring it requires non-reward mechanisms (exploration resets, replay, distillation).

Proof sketch. (i) follows from $\nabla_{z_g} \log m_f = \mathbf{1}[g = f] - m_g$ and the zero-sum identity; the threshold form divides the bracket by m_f. (ii) evaluates both laws as $m_f \to 0$ (the extinct family's contribution to $C(m)$ is $O(m_f^2)$). (iii) compares relative drifts under equal acceptance rates. Full proofs in Appendix A. □

Monte-Carlo evaluation of the exact group loop confirms equation 4 to four decimals at the experimental calibration point; on the logged token-level trajectories all three laws track the observed f_1 drift comparably ($R^2 \in [0.19, 0.51]$, five corr runs) — the qualitative predictions are parameterization-robust.

Equation 3 is a replicator equation with fitness q_f, except that the threshold is endogenous, rising as the population improves. Three structural consequences follow.

Extinction boundary and the race dynamics. Rewriting the sign condition for family f gives the fate transition in FN space:

$$r_f^*(m) = 1 - \frac{\bar{q}(m)}{s_f}: \quad r_f > r_f^*(m) \;\Rightarrow\; \text{negative drift.} \tag{5}$$

Because $\bar{q}(m)$ co-evolves with the mixture, the boundary is a moving obstacle: $\bar{q}$ starts low ($m_\perp \approx 0.6$) and rises as the invalid mode drains, so a family with positive initial drift banks mass early and collapses once the rising threshold overtakes its fixed acceptance rate. A fixed-point argument sharpens this: since clean families have $q = 1$ and the hostile family $q_1 = 1 - r_1 < 1$, any mixture containing a clean family has $\bar{q} > q_1$ — coexistence is impossible; extinction is a question of when (§4.2).

Amplification through co-occurrence. The per-event punishment magnitude satisfies $A_-(k)/A_+(k) = k/(G-k)$: in majority-accepted groups the falsely-rejected member is punished more than accepted members are rewarded. A hostile family co-occurs with accepted families systematically, so its FN events concentrate in high-k groups, exceeding i.i.d. harm at identical aggregate rate. The accumulating quantity is the cumulative harmful-selection exposure

$$\mathrm{HSE}_f(T) = \sum_{t \le T} \eta\, m_f(t)\, s_f\, r_f\, \alpha_-\big(G, \bar{q}(t)\big), \tag{6}$$

and the exact three-channel decomposition of $\mathbb{E}[\Delta \log m_f]$ is validated to $R^2 > 0.99$ in the exact simulator (§4).

Group size. Expected dynamics depend on G only through $\alpha_\pm(G, \bar{q})$, weakly beyond $G = 4$; extinction time is approximately G-independent within A1–A4; G controls variance, degeneracy, and rare-family coverage (H-G; Appendix F.5).

3.4 The Aggregate FN Rate Is Insufficient

Proposition 2 (Insufficiency of the aggregate FN rate)**.** *(i) If $r_f = r$ for all f, then $q_f = (1-r)s_f$ and the q-ranking of families coincides with the true s-ranking for every G. (ii) There exist configurations with identical aggregate FN rate in which one profile preserves the true ordering and a family-correlated profile inverts it; the inversion can be driven by an arbitrarily small aggregate FN rate when two families are near-tied in true success.*

Proof sketch. (i) is immediate. For (ii), $s = (0.6, 0.4)$ with aggregate FN 0.24: iid FN ($r = (0.24, 0.24)$) gives $q = (0.456, 0.304)$ — ordering preserved; correlated FN ($r = (0.40, 0)$) gives $q = (0.360, 0.400)$ — inverted. As $s_g/s_f \to 1$ the inverting $r_f = 1 - s_g/s_f \to 0^+$. Full details in Appendix A. □

The corr checker is the live counterexample: self-masking drives its realized aggregate FN to zero, so Youden's J tends to its learning-regime maximum on the day a valid family is gone — no aggregate FN bound secures strategy ordering, and published "capability boundaries" measured under fallible checkers may be checker boundaries.

3.5 Self-Masking as a Corollary

Corollary 2 (Self-masking verifier)**.** *Under the drift law, the FN exposure channel of family f carries the factor $m_f(t)$: as f is driven to extinction its FN events vanish with its mass, so the checker's observed error rate on the trained policy falls toward zero even though the checker itself is unchanged. Formally, $\mathrm{HSE}_f(t) \to 0$ and the measured population FN rate $\bar{r}(t) = \sum_f m_f(t) s_f r_f / \sum_f m_f(t) s_f \to 0$ as $m_f(t) \to 0$.*

This predicts the signature we exploit in §5: the checker appears most reliable exactly when its victim is gone. Fixed-distribution audits on the frozen pre-training checkpoint separate this policy-induced masking from genuine checker improvement, which cannot occur — the checker is never updated. Proposition 1(ii)–(iii): the deletion is irreversible under continued training and inert under repair — the reward signal that caused the loss cannot restore it (§6).

4 EXACT CATEGORICAL SIMULATOR

We implement A1–A4 literally as a tabular categorical simulator over the four modes of the tool-chain environment (§5.1): the policy is a mixture vector, groups of G=8 rollouts are drawn i.i.d., outcomes and verdicts follow the exact (s_f, r_f) profile, and the mass vector updates by the group-relative gradient.

4.1 VERIFICATION OF THE DRIFT LAW

Two checks validate the implementation: regressing realized one-step mass changes on the three exact channels yields $R^2 > 0.99$ (mapping A) and 0.80–0.85 (mapping B) — the simulator-side validation of the HSE accounting — and the realized drift's sign agrees with $q_f \lessgtr \bar{q}$ at every step of every run.

4.2 THE MOVING THRESHOLD AND THE RACE TO COLLAPSE

We sweep $r_{f_1} \in [0.30, 0.75]$ under two mappings. Under the heterogeneous-success mapping ($s = (0.7, 0.5, 0.3)$, uniform valid masses, 10% invalid), the static boundary $r^* = 1 - \bar{q}(m_0)/s_{f_1} = 0.10$ is exceeded by the whole range: every point decays monotonically from step zero. Under the LLM-calibrated mapping ($s_f = 1$; initial masses $(0.017, 0.304, 0.074, 0.605)$ from the base-model calibration), the boundary is $r^*(m_0) = 0.605$, above the experimental operating point 0.5198: every point banks mass first and collapses only when the rising $\bar{q}(t)$ overtakes $q_{f_1} = 1 - r_{f_1}$. Flip time is strongly monotone in r_{f_1} and banked peak mass falls as r_{f_1} rises (Appendix G; Figures 5 and 6). The fixed-point argument above explains both mappings: coexistence is impossible, and extinction timing is the race between the hostile family's banking rate and the invalid-mode drain that inflates the early threshold.

4.3 MATCHED-AGGREGATE COUNTEREXAMPLE REALIZED

The experimental configuration itself is the paired construction of Proposition 2 realized in the simulator: identical aggregate FN rate on valid rollouts, with errors either spread uniformly (iid) or concentrated on one family (correlated). In the tabular abstraction the paired arms reproduce the inversion: under the family-uniform arm the strongest family wins as it should (final mass 0.987 vs. 0.009 for the inferior family; ordering preserved), while under the matched family-correlated arm the strongest family goes extinct ($m = 0.003$) and the inferior family takes over (0.992) — the true ordering is inverted at identical aggregate error. The inversion also costs capability: oracle success at horizon is 0.498 under the correlated arm versus 0.697 under clean — O-pass improves in every arm even as diversity collapses. Checker-measured success tracks the oracle under the correlated arm (0.497) while staying depressed under iid (0.417): self-masking is visible already in the abstraction. The LLM experiments of §5 confirm the prediction at token level.

5 CONTROLLED EXPERIMENTS

5.1 ENVIRONMENT AND CHECKER ARMS

Our testbed is a categorical tool-chain environment with an exact success oracle: a module needs dependency P (alpha on mirror-1 or gamma on mirror-2), then Q (beta), then publication (delta); success iff the chain is realized without precondition violations. The action multiset defines the exact, verifier-independent family classifier: f_0 = alpha-route, f_1 = gamma-route, f_2 = both, else invalid $f_\perp$; a stylistic header serves the family-orthogonal surface control. All arms share the frozen calibration of the base model (Qwen2.5-1.5B-Instruct, LoRA r=16): per-family masses $(m_{f_0}, m_{f_1}, m_{f_2}, m_{f_\perp}) = (0.017, 0.304, 0.074, 0.605)$ and oracle success rate 0.395, measured on n=4096 generation probes before training. Four checker arms ran 150 GRPO steps each ($B = 32$ groups, $G = 8$ rollouts, binary rewards): clean ($V = O$); iid, rejecting a uniformly random 40% of all valid rollouts; corr, rejecting a random 51.98% of valid rollouts of family f_1 only and never rejecting any other family; and surface, rejecting valid rollouts whose header is `SHORTCUT` (prevalence ≈0.45). Under the calibration masses, corr and iid are matched at an aggregate FN rate of 0.4 on valid rollouts; they differ only in whether errors attach to a strategy family or to rollout identity —

Arm	O-pass $0 \to 150$	$m_{f_1}(150)$	V-pass@150	$\text{HSE}_{\text{benign}}$	HSE_{FN}
corr	0.387–0.461 $\to$ 0.986–1.000	0.000 (5/5)	0.995±0.005	3074±49	−960±37
iid	0.387–0.461 $\to$ 0.988–1.000	0.387–1.000	0.595±0.021	17329±39	−16177±20
iidm	0.387 $\to$ 0.998–1.000	1.000	0.607±0.010	2375±16	−1022±21
surface	0.387–0.461 $\to$ 0.992–1.000	0.988–1.000	0.996±0.003	1956±52	−88±8
clean	0.387–0.461 $\to$ 0.996–1.000	0.992–0.996	0.998±0.001	1851±34	0

Table 1: Only the corr arm ends the run with the targeted family eliminated while every standard metric reads healthy. Arm-level outcomes of the controlled experiment (Qwen2.5-1.5B, 150 GRPO steps, G=8): corr, iid, iidm, surface, and clean arms — five seeds per arm except iidm, which uses its two registered seeds (ranges span seeds; ±: seed sd; the iid m_{f_1} range is survivor-share noise, every seed at or above the initial 0.30). O-pass: oracle-measured success; $m_{f_1}(150)$: the targeted family's mass at step 150; V-pass: checker-measured success on valid rollouts; $\text{HSE}_{\text{benign}}$ / HSE_{FN}: cumulative benign-positive / harmful-FN selection exposure of the logged channels. The clean arm records no harmful-FN exposure by construction; the surface arm's small HSE_{FN} reflects occasional surface-trigger false rejections, not family-constitutive harm. Full per-run states are in Table 3.

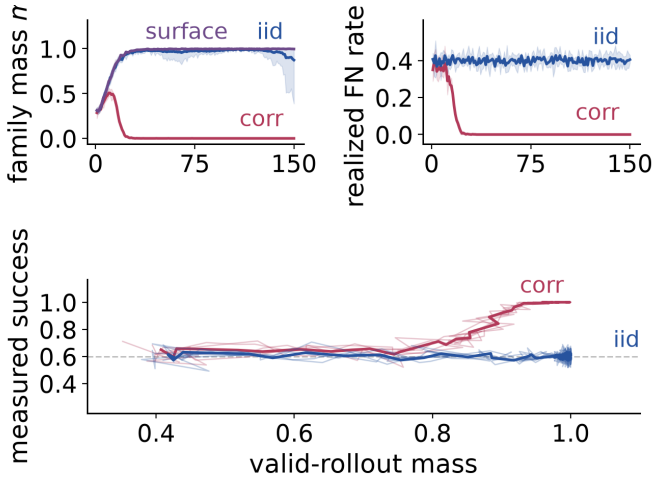

Figure 2: Family-correlated noise deletes the targeted family; matched uniform noise at the identical aggregate rate preserves it. Per-step dynamics, five seeds per arm (solid: mean; band: min–max envelope). Top left, targeted-family mass m_{f_1}: extinction under corr, persistence under iid and the surface-trigger control. Top right, realized population FN rate: the corr checker's FN rate self-masks to zero as its victims disappear, while iid's persists at the planted 0.4. Bottom, valid-rollout mass vs. checker-measured success: corr converges to perfect measured reliability while iid is pinned at the neutrality threshold $q_f = 0.6$.

the experimental realization of §3, with corr setting $q_{f_1} = 0.4802$, $q_{f_0} = q_{f_2} = 1$, and iid the neutrality threshold $q_f = 0.6$ uniformly. Arm definitions, calibration hashes, and registered acceptance criteria are pre-registered in Appendix C; all runs completed 150 logged steps under byte-identical frozen constants.

5.2 Main Contrast: Same Aggregate Error, Opposite Fate

Across five seeds, corr and iid converge to opposite strategy fates: $\min_s m_{f_1}(\text{iid}, s) - \max_s m_{f_1}(\text{corr}, s) = +0.387 \gg 0.15$, and every corr seed replicates the extinction endpoint exactly ($m_{f_1} = 0.000$; Table 1). Under corr the targeted family goes from its calibrated $\sim$0.30 population share to zero observed valid rollouts while oracle success rises $0.39 \to \approx 1.0$; under matched-iid it retains a large mass throughout (0.99, 0.98, 0.39, 1.00, 1.00 — the third from slow neutral drift pinned at the neutrality threshold, as the q_f accounting predicts).

The clean arm certifies ordinary learnability (O-pass $0.39 \to 0.996$), ruling out training artifacts.

Two features of Table 1 merit emphasis. O-pass improves to ≥ 0.986 in every arm: aggregate capability metrics are blind to the collapse. And measured reliability moves in opposite directions: under corr, V-pass converges to the oracle rate (1.000 to 0.986), while under iid it stalls near 0.6 — the corr checker's victims take its false negatives with them.

The iidm row makes the contrast controlled: its per-step injected FN count matches the corr arm's realized schedule of the same seed, but each rejection is drawn uniformly over families. The outcome is the iid one ($m_{f_1} \to 1.000$), the schedule decaying to zero by step $\sim$90 as the corr twin self-masks: identical FN volume, opposite fate — the optimizer amplifies family-correlated error structure into extinction.

5.3 Self-Masking of the Verifier

Corollary 2 predicts a family-conditioned checker looks increasingly reliable as it destroys its target; the V-pass trajectories confirm it (Figure 2 bottom; per-seed anatomy: Appendix B). Under corr, checker-measured success rises $0.683 \to 1.000$ and the population FN rate $\bar{r}(t)$ improves $0.4 \to 0$ with the checker unchanged: any post-hoc audit certifies a near-perfect verifier. Under iid the same error rate pins success at the neutrality threshold ($0.644 \to 0.600$): its errors attach to rollouts that keep being produced.

5.4 Family-Resolved Harmful-Selection Exposure

The three-channel decomposition of §3 separates the profiles mechanistically: in corr (seed 0), family f_1 saw 916 accepted valid rollouts (benign advantage $+0.086$) against 991 falsely rejected (-0.94) — an $\approx 11:1$ penalty on one family; $f_\perp$ and f_2 saw zero FN events. In matched-iid, rejected valid rollouts share one mean advantage (-1.098), f_1 carrying 97.9% of the (much larger) FN mass, $\approx 17\times$ corr's, yet fate reverses (Proposition 2). The surface-trigger arm accumulates comparable FN volume (-86 to -100) on a family-orthogonal trigger and leaves f_1 intact ($m_{f_1} \geq 0.988$) — the collapse is family-constitutive (Appendix D).

5.5 How Far Does the Phenomenon Generalize?

The contrast is robust across seeds, checker arms, scales, and model families: the grid brings corr to five of five extinctions against seventeen intact controls; target permutation drives any family extinct regardless of initial mass, and the FN-rate sweep shows extinction is threshold-free in the registered range (Appendix E). SmolLM2-1.7B reproduces the corr-vs-iid contrast under degenerate calibration, OLMo-2-1B fails the competence gate, and the 7B replication reproduces the causal signature (Appendix F.3); the stress tests (Figure 3), fixed-distribution audits, and protocol transfer are in Appendix F (§§F.1–F.4); G-dose in §F.5.

6 Hidden Capability Loss

Extinction under a friendly checker matters only if the lost family is relearnable on demand; the stress tests hold the checker fixed and alter the environment (Figure 3), and the deletion is invisible in standard logs.

6.1 Route Blocking

Masking the surviving routes forces a switch to the eliminated gamma-route family or fail (pre-registered loss band $\Delta \geq 0.15$). Controls reroute and succeed; corr collapses ($\Delta \geq 0.32$): the punished behavior is gone as usable ability, with nothing latent to reroute.

6.2 Cost Shift and Task Shift

Task shift admits only the eliminated route (mirror-1 outage: any alpha use invalidates the rollout), so recovery must come from existing distributions; readouts mirror §6.1. Registered cost shift needs graded rewards the binary environment lacks.

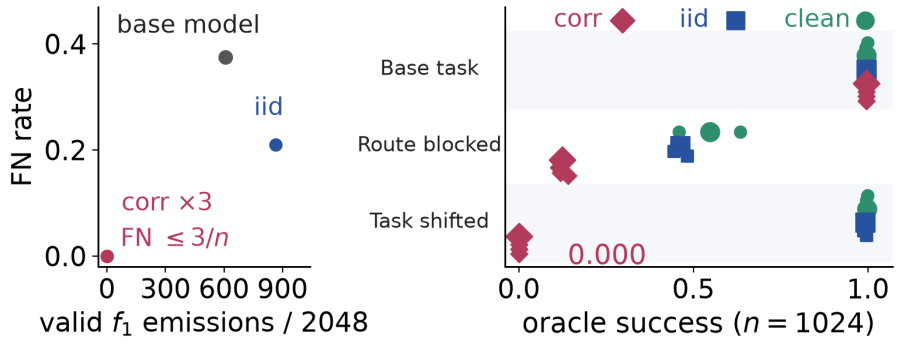

Figure 3: Self-masking at its endpoint, and the loss is real. Left: family-resolved audits (n=2048 per checkpoint): the base model emits 606 valid f_1 rollouts at FN rate 0.49, every corr final zero. Right: per-seed readouts (dots: seeds; larger: means): corr collapses under route blocking (0.125 vs. 0.44–0.64) and scores 0.000 under task shift; controls retain the family (0.44–0.64, 0.99^{+}).

6.3 The Deletion Is Invisible in Standard Logs

Extinction is support-level (no valid f_1 rollouts from step ≈30; survivor entropy $0.024 \to 0.008$) while every aggregate metric improves — standard logs show nothing to act on. Figure 3's task-shift row makes the deletion legible: the removed family was the only competent response.

7 Discussion

Checker vs. capability boundaries. Under family-correlated checker errors the repertoire is an artifact of the reward oracle: it contracts invisibly while every aggregate metric improves — three families became two as O-pass rose $0.39 \to 1.0$. Families below the population threshold die at any error budget (Proposition 2); self-masking hides the loss — FN rates fall as victims die. Family-resolved oracle measurement with HSE attribution (§5.4) locates it.

Mitigation directions. Harm enters through $q_f = s_f(1 - r_f)$: audit r_f per family and re-allocate verification budget to equalize it. Policy-side fixes are group-composition control or within-stratum advantage re-normalization (Appendix F.1). Both diagnoses run on frozen rollouts at initialization, before any training spend; and token-level temperature is not a family-addressable knob, transiently re-opening the deleted family's support before the repaired checker's common-acceptance dynamics drive it back out (Appendix F.2).

Limitations. Our categorical environment has an exact family classifier; the published-benchmark audit (Appendix F.4) finds standard configurations safe. Token-level sharing (A4) is excluded by theory and bounded by leakage variants and the LLM runs. False positives are excluded by construction, with FPR $= 0$; the symmetric channel is separate (Zhang, 2026).

Operational auditing implication. On-policy aggregate accuracy is not evidence of verifier quality: measured success rises simply because the policy stops emitting the family on which the checker fails. Audits should therefore run on a fixed reference distribution with family-resolved reporting of r_f, read jointly with the HSE attribution to separate genuine checker improvement from policy-induced self-masking before a distribution shift surfaces the hidden loss.

8 Conclusion

Family-attached false negatives drive a family across the extinction boundary $r_f^* = 1 - \bar{q}/s_f$: family structure, not error volume, decides fate. A family holding a third of the population vanishes in thirty steps while oracle success reaches ≈1.0, and the deletion survives repair and re-opening (Appendix F.2). These results motivate family-resolved verifier audits — measuring what a checker fails on, not merely how often — before aggregate metrics are read as evidence that nothing was lost.

REPRODUCIBILITY STATEMENT

The controlled experiments use a frozen training stack: the categorical tool-chain environment and GRPO trainer at code freeze 8200536 (P2 extensions are a documented copy-then-patch of the same trainer, covered by a unit suite of 22 trainer tests and 31 audit-library tests), and one constants file (`constants.json`, `md5 862b20c97364778da6c6c916d91ad744`) that fixes the calibration, the checker arms, and the evaluation protocol. The training objective is vanilla policy gradient on token log-probabilities with group-relative advantages ($B = 32$ groups, $G = 8$ rollouts, advantages normalized by the MLE group standard deviation with a 10^{-4} floor); there is no importance ratio, no objective clipping, and no KL penalty in the objective — KL and entropy against the base policy are logged as probes every 25 steps. Optimization is AdamW (learning rate 10^{-5}, weight decay 0, gradient-norm clip 1.0) on LoRA r=16, α=32 adapters over Qwen2.5-1.5B-Instruct (bf16 weights, float32 log-probabilities), with generation at temperature 1.0, top-p 1.0, $\leq$160 new tokens, a pool of 512 tasks reshuffled per epoch, and 150 steps with frozen 512-rollout evaluations at steps 0, 50, 100, and 150. Base-model calibration used 4096 generation probes. The environment is deterministic given seeds, and the checker arms are deterministic functions of the constants and per-call RNG streams. Arm definitions, calibration, registered pass/fail criteria, and analysis choices were pre-registered before the corresponding training runs; per-step metric streams, configuration hashes, the frozen family classifiers, the vendored published checkers used in the benchmark audits, and the exact categorical simulator will be released. Model coverage: Qwen2.5-1.5B (main experiments, five seeds per checker arm), Qwen2.5-7B (30-step replication, two seeds), SmolLM2-1.7B (cross-family replication in a degenerate calibration regime), and OLMo-2-0425-1B (stopped at a pre-registered competence gate; no training run).

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure: assisting in the development of the theoretical model (the family-mixture abstraction and its exact drift laws) and in formulating, checking, and writing up the mathematical claims and proofs — every derivation in Appendix A was re-verified step by step by the authors; proposing and refining hypotheses and providing structured feedback on the experimental methodology and the registered pass/fail criteria; implementing the research software (the training, audit, and analysis pipelines and their unit and audit-library test suites) and operating the experiment pipeline — calibration, training, audit, and stress runs on our compute allocation — with AI-driven tooling under the authors' supervision, with all runs, logs, and artifacts inspected and validated by the authors; supporting the interpretation of results; aiding the writing of the manuscript through language editing, restructuring of passages, and wording revisions, with all final text, claims, and framing reviewed and approved by the authors; and retrieving, discovering, and verifying the related work (every bibliographic entry was cross-checked against authoritative sources). We have not used generative AI tools to generate synthetic datasets: the categorical environment is deterministic software, the training data are rollouts sampled from the open-weight models we train, and the published-benchmark audits operate on public benchmark slices with independent ground truth.

We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI. See the ICLR 2027 AI Policy for Authors for more details.

REFERENCES

Yong Yi Bay and Kathleen A. Yearick. Grpo, dr. grpo, and dapo are three operations on one number: The group-standard-deviation identity, 2026. URL http://arxiv.org/abs/2607.00152v1.

Xin-Qiang Cai, Wei Wang, Feng Liu, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Reinforcement learning with verifiable yet noisy rewards under imperfect verifiers, 2025. URL http://arxiv.org/abs/2510.00915v4.

Saurabh Dash, Pierre Clavier, John Dang, Matthias Galle, Marzieh Fadaee, Ahmet Üstün, and Beyza Ermis. Soft-sverl: Self-verified reinforcement learning with soft rewards, 2026. URL http://arxiv.org/abs/2605.28561v1.

Zhibin Duan, Guowei Rong, Zhuo Li, Bo Chen, Mingyuan Zhou, and Dandan Guo. Mitigating reward hacking in rlhf via bayesian non-negative reward modeling, 2026. URL http://arxiv.org/abs/2602.10623v2.

Kazuki Egashira, Mark Vero, Jasper Dekoninck, Florian E. Dorner, Robin Staab, and Martin Vechev. Delay, plateau, or collapse: Evaluating the impact of systematic verification error on rlvr, 2026. URL http://arxiv.org/abs/2605.02909v2.

Haowen Gao, Zhenyu Zhang, Liang Pang, Fangda Guo, Hongjian Dou, Guannan Lv, Shaoguo Liu, Tingting Gao, Huawei Shen, and Xueqi Cheng. Diva-grpo: Enhancing multimodal reasoning through difficulty-adaptive variant advantage, 2026. URL http://arxiv.org/abs/2603.01106v1.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Zhenhua Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bowen Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guan-Ting Chen, Guowei Li, Hongjun Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jiagang Tu, Junjie Qiu, Junlong Li, Jiali Cai, Jiaqi Ni, Jian Liang, Jing Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, L. Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Runji Wang, R. J. Chen, Rong Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Yu, Shunfeng Zhou, Shuting Pan, Sansan Li, Shuang Zhou, Shaoqing Wu, and Tao Yun. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Aishan He, Daniel B. Fried, and Sean Welleck. Rewarding the unlikely: Lifting grpo beyond distribution sharpening. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25559–25571. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1298.

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Advantage collapse in group relative policy optimization: Diagnosis and mitigation, 2026. URL http://arxiv.org/abs/2605.21125v2.

Lukas Helff, Quentin Delfosse, David Steinmann, Ruben Härle, Hikaru Shindo, Patrick Schramowski, Wolfgang Stammer, Kristian Kersting, and Felix Friedrich. Llms gaming verifiers: Rlvr can lead to reward hacking, 2026. URL http://arxiv.org/abs/2604.15149v1.

Yuki Ichihara, Yuu Jinnai, Tetsuro Morimura, Mitsuki Sakamoto, Ryota Mitsuhashi, and Eiji Uchibe. Mo-grpo: Mitigating reward hacking of group relative policy optimization on multi-objective problems. arXiv:2509.22047, 2025. URL http://arxiv.org/abs/2509.22047v3.

Yuxin Jiang, Yufei Wang, Qiyuan Zhang, Xingshan Zeng, Liangyou Li, Jierun Chen, Chaofan Tao, Haoli Bai, and Lifeng Shang. From verifiable dot to reward chain: Harnessing verifiable reference-based rewards for reinforcement learning of open-ended generation, 2026. URL http://arxiv.org/abs/2601.18533v1.

Neil Kale, Chen Bo Calvin Zhang, Kevin Zhu, Ankit Aich, Paula Rodriguez, Scale Red Team, Christina Q. Knight, and Zifan Wang. Reliable weak-to-strong monitoring of llm agents, 2025. URL http://arxiv.org/abs/2508.19461v1.

Jacky Kwok, Shulu Li, Pranav Atreya, Yuejiang Liu, Yixing Jiang, Chelsea Finn, Marco Pavone, Ion Stoica, and Azalia Mirhoseini. Llm-as-a-verifier: A general-purpose verification framework, 2026. URL http://arxiv.org/abs/2607.05391v2.

Xiaozhe Li, Yang Li, Xinyu Fang, Shengyuan Ding, Peiji Li, Yongkang Chen, Yichuan Ma, Tianyi Lyu, Linyang Li, Dahua Lin, Qipeng Guo, Qingwen Liu, and Kai Chen. Beyond mode collapse: Distribution matching for diverse reasoning, 2026. URL http://arxiv.org/abs/2605.19461v1.

Anamika Lochab, Bolian Li, and Ruqi Zhang. Uniform-correct policy optimization: Breaking rlvr's indifference to diversity, 2026. URL http://arxiv.org/abs/2605.00365v1.

Omar El Mansouri, Fathinah Asma Izzati, Mohamed El Amine Seddik, and Salem Lahlou. Noise-corrected grpo: From noisy rewards to unbiased gradients. arXiv:2510.18924, 2025. URL http://arxiv.org/abs/2510.18924v3.

Ali Rad, Khashayar Filom, Darioush Keivan, Peyman Mohajerin Esfahani, and Ehsan Kamalinejad. Rate or fate? rlvεr: Reinforcement learning with verifiable noisy rewards. arXiv:2601.04411, 2026. URL http://arxiv.org/abs/2601.04411v1.

Anagha Radhakrishna Palandye, Rebecca Glick, and Osheen Kaul. Reward granularity in rlvr: Comparing process and outcome reward structures for mathematical reasoning in small language models, 2026. URL http://arxiv.org/abs/2607.02869v1.

Thi Kim Trang Vo, Nam Tien Le, Thi Kim Nguyet Vo, Minh Khang Tran, and Duy Phuong Tran. A verifier-guided explainable reasoning framework with gold-anchored qlora, task-aware mixture-of-experts, and group-relative rlvr, 2026. URL http://arxiv.org/abs/2609.05221v1.

Jingyi Wang, Lei Zhu, Tengjin Weng, Song-Li Wu, Haochen Tan, Jierun Chen, Chaofan Tao, Haoli Bai, Lu Hou, Lifeng Shang, and Xiao-Ping Zhang. Grpo-vps: Enhancing group relative policy optimization with verifiable process supervision for effective reasoning, 2026a. URL http://arxiv.org/abs/2604.20659v1.

Tianyi Wang, Long Li, Hongcan Guo, Yibiao Chen, Yixia Li, Yong Wang, Yun Chen, and Guanhua Chen. Anchored policy optimization: Mitigating exploration collapse via support-constrained rectification, 2026b. URL http://arxiv.org/abs/2602.05717v1.

Xinpeng Wang, Nitish Joshi, Barbara Plank, Rico Angell, and He He. Is it thinking or cheating? detecting implicit reward hacking by measuring reasoning effort, 2025. URL http://arxiv.org/abs/2510.01367v4.

Shaohang Wei, Zikun Su, Feifan Song, Wen Luo, Wei Li, Guangyue Peng, and Houfeng Wang. Verifier-induced support reshaping in on-policy optimization, 2026. URL https://arxiv.org/abs/2608.00220.

Esther Xin. Where the verifier fails: A category-level audit of reward signals in rlvr, 2026. URL http://arxiv.org/abs/2609.01354v1.

Zhangchen Xu, Yuetai Li, Fengqing Jiang, Bhaskar Ramasubramanian, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. Tinyv: Reducing false negatives in verification improves rl for llm reasoning, 2025. URL http://arxiv.org/abs/2505.14625v2.

Song Yu, Li Li, Wenwen Zhao, and Zhisheng Yang. Ep-grpo: Entropy-progress aligned group relative policy optimization with implicit process guidance, 2026. URL http://arxiv.org/abs/2605.04960v1.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? NeurIPS 2025 Oral; arXiv:2504.13837, 2025. URL http://arxiv.org/abs/2504.13837v5.

Chuyifei Zhang. When the reward suite is leaky: A preregistered causal contrast of natural verifier false positives in rlvr, 2026. URL http://arxiv.org/abs/2607.11022v1.

Xichen Zhang, Sitong Wu, Yinghao Zhu, Haoru Tan, Shaozuo Yu, Ziyi He, and Jiaya Jia. Scaf-grpo: Scaffolded group relative policy optimization for enhancing llm reasoning, 2025. URL http://arxiv.org/abs/2510.19807v2.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, Xiaolei Wang, Yifan Du, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Peiyu Liu, Yiwen Hu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *Frontiers of Computer Science*, 20(12), 2026. doi: 10.1007/s11704-026-60308-3.

Chenyu Zhou, Qiliang Jiang, and Xu Zhou. Multilingual verifier bias in rlvr: Benchmark, rollout diagnosis, and the cross-lingual selection bottleneck, 2026. URL http://arxiv.org/abs/2608.20362v1.

Tao Zhou. When rlvr shrinks the reasoning boundary: Diagnosing pass@k inversion. arXiv:2607.20543, 2026. URL https://arxiv.org/abs/2607.20543.

A Proofs

A.1 Event Structure (Lemma 1)

With $k = \sum_i R_i$ accepted rollouts and $\hat{\sigma}_G > 0$, member i has advantage sign $\operatorname{sign}(R_i - k/G)$. If $R_i = 1$ then $k \geq 1$ so $R_i - k/G \geq 0$ (strictly when $k < G$; equality when $k = G$ is the degenerate case excluded by $\hat{\sigma}_G > 0$). If $R_i = 0$ and $k \geq 1$ the advantage is negative; if $k = 0$ the group is degenerate. Hence $a_i < 0$ iff $R_i = 0$ and $k \geq 1$. □

A.2 Exact Finite-G Drift (Theorem 1)

Under A1–A4 the score-function estimator of the mixture parameter m_f with learning rate η gives, for a group of size G, $\mathbb{E}[\Delta z_f] = \eta\,\mathbb{E}[a(\mathbf{1}\{i \in f\} - m_f)]$, so $\mathbb{E}[\Delta m_f] = \eta\, m_f\, \mathbb{E}[a \mid i \in f]$ after subtracting the population mean (the identity $\sum_f m_f \mathbb{E}[a \mid i \in f] = \mathbb{E}[a] = 0$ holds because advantages are zero-mean within each group). Condition on $i \in f$: by A1 the remaining $G - 1$ members are i.i.d. mixture draws, so the number K of accepted co-members is $\mathrm{Bin}(G - 1, \bar{q})$; the focal member is accepted with probability q_f (contribution $q_f \alpha_+$) and otherwise rejected while $K \geq 1$ members were accepted (contribution $-(1 - q_f)\alpha_-$; when $K = 0$ the group is degenerate and the advantage is zero, which the binomial averaging of α_- handles through the $K = 0$ term vanishing). The within-group zero-sum identity $\sum_i (R_i - k/G)/\hat{\sigma}_G = 0$, averaged over a uniformly random member, gives $\bar{q}\,\alpha_+ = (1-\bar{q})\,\alpha_-$: the expected positive advantage mass equals the expected negative one. Substituting, $\mathbb{E}[a \mid i \in f] = q_f \alpha_+ - (1 - q_f)\alpha_- = (\alpha_+ + \alpha_-)(q_f - \bar{q})$, since $q_f \alpha_+ - (1 - q_f)\alpha_- = q_f(\alpha_+ + \alpha_-) - \alpha_-$ and $\alpha_- = \bar{q}(\alpha_+ + \alpha_-)$. □

A.3 Insufficiency of the Aggregate Rate (Proposition 2)

(i) With $r_f = r$ for all f, $q_f = s_f(1 - r)$ is an affine, strictly order-preserving transform of s_f, so the drift signs of Theorem 1 induce the same ordering as the true successes for every G and every mixture m. (ii) The two-family example in the statement preserves the aggregate rate $\sum_f m_f s_f r_f / \sum_f m_f s_f$ by construction: $0.6 \times 0.4 = 0.24 = (0.6 + 0.4)/2 \times 0.24 \cdot 2/2$ under uniform masses, and the induced q-vectors $(0.456, 0.304)$ and $(0.360, 0.400)$ have opposite orderings. For the arbitrarily-small-rate version, take any $\varepsilon > 1 - s_g/s_f > 0$ and set $r_f = \varepsilon$, $r_g = 0$: then $q_f < q_g$ while the aggregate FN rate $m_f s_f \varepsilon / \sum_j m_j s_j \to 0$ as $\varepsilon \to 0$. □

A.4 Self-Masking (Corollary 2)

The FN exposure channel of family f is $\eta\, m_f(t)\, s_f\, r_f\, \alpha_-(G, \bar{q}(t))$, which is $O(m_f(t))$; as $m_f(t) \to 0$ the channel's contribution to both the drift of other families and the population FN rate $\bar{r}(t) = \sum_f m_f s_f r_f / \sum_f m_f s_f$ vanishes. The checker is unchanged by hypothesis, so the improvement in measured reliability is entirely attributable to the support change. □

A.5 Softmax Parameterization and Irreversibility (Proposition 1)

Part (i). Treat the family as a single action with $m_f = e^{z_f}/\sum_g e^{z_g}$, so $\nabla_{z_g} \log m_f = \mathbf{1}\{g = f\} - m_g$. The GRPO update applies $z_g \mathrel{+}= (\eta/G)\sum_i a_i(\mathbf{1}\{g = f_i\} - m_g)$, whose expectation is $\mathbb{E}[\Delta z_g] = \eta(m_g\mu_g - m_g\mathbb{E}[a])$; the second term vanishes by the within-group zero-sum of the advantages, giving $\mathbb{E}[\Delta z_g] = \eta\, m_g\mu_g$ exactly, with $\mu_g = q_g\alpha_+ - (1-q_g)\alpha_- = (\alpha_+ + \alpha_-)(q_g - \bar{q})$ by the identity $\bar{q}\alpha_+ = (1-\bar{q})\alpha_-$. For small η, first-order expansion of $m'_f = e^{z_f+\Delta z_f}/\sum_g e^{z_g+\Delta z_g}$ gives $\Delta m_f = m_f(\Delta z_f - \sum_g m_g\Delta z_g) = \eta\, m_f(m_f\mu_f - \sum_g m_g^2\mu_g) = \eta\, m_f(\alpha_+ + \alpha_-)\big(m_f(q_f - \bar{q}) - C(m)\big)$, since $\sum_g m_g^2\mu_g = (\alpha_+ + \alpha_-)\, C(m)$. The sign condition is $m_f(q_f - \bar{q}) > C(m)$, i.e. $q_f > \bar{q} + C(m)/m_f$. At an interior fixed point every family satisfies $m_f(q_f - \bar{q}) = C(m)$; multiplying by $1/m_f$, averaging with weights m_f, and using $\sum_f m_f(q_f - \bar{q}) = 0$ forces $C(m) = 0$, hence $q_f = \bar{q}$ for every active family — the tabular fixed points. Monte-Carlo verification: 4×10^5 simulated groups at the frozen calibration mixture reproduce the four family drifts to within Monte-Carlo error ($\leq 1\%$ relative) against equation 4.

Part (ii). Suppose $q_g \geq \bar{q}(m)$ for every surviving g with $m_g > 0$ and $q_f < \bar{q}$ persists on all mixtures supported on the survivors plus f. Tabular: the drift is $\eta\, m_f(\alpha_+ + \alpha_-)(q_f - \bar{q}) < 0$ for every $m_f > 0$, so $m_f(t)$ decreases monotonically and exponentially at rate $\eta(\alpha_+ + \alpha_-)(\bar{q} - q_f)$. Softmax: the bracket is $m_f(q_f - \bar{q}) - C(m)$; the extinct family contributes $O(m_f^2)$ to C, so near $m_f = 0$ the bracket tends to $-C(\hat{m})$ evaluated on the extinct-neighborhood mixture $\hat{m}$. If $C(\hat{m}) \geq 0$ the drift is strictly negative at every $m_f > 0$; if $C(\hat{m}) < 0$ the drift vanishes at the floor $m^* = |C(\hat{m})|/(\bar{q} - q_f)$ and changes sign only below it, so the family can at most stabilize at a floor of order C — never regrow. In both cases decay is monotone toward extinction; under the softmax law the per-family drift scales as m_f^2 (self-throttling), giving $m_f(t) \approx m_f(0)/(1 + \eta(\alpha_+ + \alpha_-)(\bar{q} - q_f)m_f(0)\, t)$, i.e. algebraic $\sim 1/t$ decay.

Part (iii). Repair the checker so that all valid families share a common acceptance q (in the controlled environment, $q = s_f = 1$). Tabular: $\mathbb{E}[\Delta m_f]/m_f = \eta(\alpha_+ + \alpha_-)(q - \bar{q})$ is the same for every valid family, so $\mathrm{d}(m_f/m_V)/\mathrm{d}t = 0$ exactly — shares among valid families are invariant while the invalid mode drains. Softmax: for two valid families f, g with $m_f < m_g$, $\mathrm{d}\log(m_f/m_g)/\mathrm{d}t = \eta\,\mu\,(m_f - m_g) < 0$ while $\mu = q - \bar{q} > 0$ (invalid mass still present), so the smaller family's share strictly shrinks. In neither parameterization does the repaired reward provide a differential gradient toward the damaged family; regaining the pre-collapse share would require the family's acceptance rate to exceed the surviving families', which the repaired checker by construction denies. Hence restoration requires non-reward mechanisms: exploration resets or raised temperature (re-opening the support), replay of stored family examples, or distillation from a teacher retaining the strategy. □

Remark. The proposition fixes what the repaired reward can and cannot do; it does not certify that the listed non-reward mechanisms suffice. A simulator quantification (§F.2) shows their sufficiency is parameterization-dependent: reseeding restores and holds the family under the tabular law, while under the softmax law every reopening mechanism leaves the family below its pre-collapse share.

B Trajectory Anatomy and Frozen Evaluation

B.1 Extinction Invariant, Survivor Random; Trajectory Anatomy

Across the corr seeds the extinct family is invariant (always the FN-targeted f_1) while the absorbing survivor is seed-dependent: f_2 absorbs the displaced mass in two seeds (0.805, 0.957) and f_0 in one (0.953) — the drift law fixes which family dies, while above-threshold survivors' relative shares are settled by early sampling noise. The trajectory follows the race predicted by §3.3 and plotted in Figure 2 top left: m_{f_1} banks $0.281 \to 0.477$ over the first ten steps (threshold $\bar{q}(t)$ still below q_{f_1}), collapses sharply between steps 10 and 20 ($0.477 \to 0.086$), and is zero-observed from step ≈ 30 onward in probes and frozen evaluation alike; route-token entropy tightens ($0.024 \to 0.008$) and the KL probe against the base policy stays ordinary (0.059–0.075 nats per response token over steps 25–150), so the collapse is not a drift-away artifact. Because extinction removes the family from the support, zero observed mass in finite samples is the operational reading; the high-sample frozen evaluation of §B.2 reports binomial zero-count bounds rather than claiming a mathematical zero.

B.2 Frozen Evaluation

At registered checkpoints (steps 0, 50, 100, 150) we evaluate every run on fresh held-out tasks with the frozen family classifier and the frozen checker. The corr-arm checkpoints show $m_{f_1} = 0$ in all post-collapse evaluations for all seeds, while O-pass and V-pass coincide to within sampling error, matching the self-masking prediction; the iid-arm checkpoints keep generating the targeted family at $m_{f_1} \approx 0.99$ under an unchanged aggregate error rate. The high-sample fixed-distribution audits ($n = 2048$ generation rollouts per checkpoint, with rule-of-three zero-count bounds for the "zero observed" claims) are reported in Appendix F (Table 4, Figure 3).

C Pre-Registration and Verification Artifacts

Frozen configuration. All controlled runs share one frozen constants file (`code/s1/constants.json`, `md5 862b20c97364778da6c6c916d91ad744`) recording the calibration (masses $(0.0173, 0.3040, 0.0737, 0.6050)$, oracle success 0.3950, hostile family 1, $r_{f_1}^{\text{corr}} = 0.5198$) and the arm definitions of §5.1. The training code is frozen at commit 8200536; the extension runs use a documented copy-then-patch snapshot whose unit suite (22 tests, including exact-count FN injection, schedule matching, family-resolved HSE bookkeeping, mixed-group detection, shifted completion masks, and a memory-safe probe equivalence check) plus 31 audit-library tests pass on the HPC environment. Configuration hashes are recorded in each run's info stream.

Hyperparameters. Qwen2.5-1.5B-Instruct; LoRA r=16, α=32, dropout 0 on all attention and MLP projections; bf16 base weights, float32 log-probabilities; group-relative advantages with $B = 32$ groups, $G = 8$ rollouts ($a_i = (R_i - k/G)/(\hat{\sigma}_{\text{MLE}} + 10^{-4})$, $\hat{\sigma}_{\text{MLE}} = \sqrt{k(G-k)/G}$, zero advantages when $\hat{\sigma}_{\text{MLE}} < 10^{-8}$); vanilla policy-gradient update on token log-probabilities, normalized by the batch's total response-token count — no importance ratio, no objective clipping, and no KL penalty in the objective (a KL/entropy probe against the base policy is logged every 25 steps); AdamW, lr 10^{-5}, weight decay 0, gradient-norm clip 1.0; generation at temperature 1.0, top-p 1.0, $\leq$160 new tokens; pool of 512 tasks reshuffled per epoch; 150 steps; frozen evaluation of 512 rollouts (128×4 tasks) at steps 0, 50, 100, 150. Each 1.5B run completes in roughly 50 minutes on one A100.

Pre-registered criteria. (i) contrast $m_{f_1}(\text{iid}) - m_{f_1}(\text{corr}) \geq 0.15$ at step 150, both seeds, same direction — passed with $+0.992$ / $+0.984$; (ii) clean-arm O-pass improvement ≥ 0.10 with $m_{f_\perp} \leq 0.10$ — passed ($+0.609$ / $+0.598$); (iii) no realized-rate drift beyond ± 0.03 of calibration — passed. The extension arms of §5.5 carry their own pre-registered criteria (multi-seed reversal of the contrast, the matched-noise control eliminating it, systematic phase-boundary failure, target-permutation failure).

Reproducibility. The environment is deterministic given seeds; the checker arms are deterministic functions of the frozen constants and per-call RNG streams seeded by (run, step, group, member). Per-step metric streams (family masses, oracle- and checker-visible pass rates, realized FN rates, advantage magnitudes, cumulative HSE channels), the frozen evaluation protocols, and all configuration hashes will be released; the family classifier is a pure function of the action multiset and is included.

D Regime Map

The theory locates the published failure regimes on a single axis: the family-constitutiveness of the FN trigger. Surface-trigger FNs tied to a learnable-away style feature produce degeneracy inflation and delay, then recovery, because the punished behavior is not the family itself (Egashira et al., 2026); item-correlated FNs tied to task instances break A1 and reproduce the published delay regimes rather than family suppression; family-constitutive FNs — errors that attach to what the solution is rather than how it looks — activate the within-group sign-flip channel analyzed above and produce strategy collapse. The aggregate error rate sets the time scale, not the fate.

r_{f_1}	Seed	Bank	Bank step	Onset step	$m_{f_1}(150)$
0.40	0	0.621	12	19	0.004
0.40	1	0.590	10	19	0.000
0.45	0	0.566	10	18	0.000
0.45	1	0.598	10	18	0.000
0.50	0	0.523	12	17	0.000
0.50	1	0.594	10	18	0.000
0.55	0	0.492	10	16	0.000
0.55	1	0.547	9	16	0.000
0.65	0	0.469	10	15	0.000
0.65	1	0.488	8	16	0.000

Table 2: FN-rate sweep: hostile-family FN rate r_{f_1} vs. flip dynamics (150 GRPO steps, per seed). Bank = peak targeted-family mass; onset = first step with $m_{f_1} < 0.5$ bank after the peak; $m_{f_1}(150)$ the terminal mass.

Arm	Seed	m_{f_0}	m_{f_1}	m_{f_2}	$m_{f_\perp}$	agg. FN
clean	0	0.000	0.996	0.004	0.000	0.000
clean	1	0.000	0.996	0.004	0.000	0.000
clean	2	0.000	0.996	0.000	0.004	0.000
clean	3	0.000	0.996	0.004	0.000	0.000
clean	4	0.000	0.992	0.000	0.008	0.000
corr	0	0.188	0.000	0.805	0.008	0.000
corr	1	0.953	0.000	0.047	0.000	0.000
corr	2	0.031	0.000	0.969	0.000	0.000
corr	3	0.055	0.000	0.934	0.012	0.000
corr	4	0.039	0.000	0.953	0.008	0.000
iid	0	0.000	0.992	0.004	0.004	0.400
iid	1	0.000	0.984	0.012	0.004	0.400
iid	2	0.000	0.387	0.594	0.020	0.386
iid	3	0.000	1.000	0.000	0.000	0.410
iid	4	0.000	1.000	0.000	0.000	0.430
iidm	0	0.000	1.000	0.000	0.000	0.000
iidm	1	0.000	1.000	0.000	0.000	0.000
surface	0	0.000	0.988	0.008	0.004	0.000
surface	1	0.000	0.996	0.000	0.004	0.000
surface	2	0.000	1.000	0.000	0.000	0.000
surface	3	0.000	0.996	0.000	0.004	0.000
surface	4	0.000	1.000	0.000	0.000	0.000

Table 3: Final-state family masses and realized aggregate FN rate for every completed controlled run (22 runs: five corr seeds, five iid, two iidm, five surface, five clean). Every corr final shows $m_{f_1} = 0$ and realized FN $= 0$ (self-masking); every control arm retains $m_{f_1} \in [0.387, 1.0]$ with realized FN ≈ 0.4 under iid and ≈ 0 under the non-rejecting arms.

Table 2 tabulates the sweep behind this map: for each trigger rate and seed, the peak bank and its step, the onset step, and the terminal family mass.

E Complete Final-State Grid

The extension arms complete the grid. Target permutation (hostile $= f_0$ and $= f_2$, two seeds each) drives the target extinct regardless of initial mass — the sixty-percent family f_2 is erased by step $\sim$15 (bank 0.11), the trivial f_0 within five steps, and the population collapses onto the one family both solvent and checker-clean: any family-constitutive trigger channel is lethal to its target. The FN-rate sweep (ten grid runs) collapses with monotonically ordered flip times (median onset 19, 18, 17.5, 16, 16 for $r_{f_1} = 0.40$–0.65) and banking mass $0.62 \to 0.47$: the static boundary gates only the initial growth phase; extinction is threshold-free in the registered range. Table 3 lists the final state of every completed run.

Audit	Ckpt	O-pass	V-pass	agg. FN	f_1 count	r_{f_1}
base	base	0.388	0.243	0.3748	606	0.492
corr s0	final	0.997	0.997	0.0000	0	—
corr s1	final	0.997	0.997	0.0000	0	—
corr s2	final	0.996	0.996	0.0000	0	—
iid s2 recover	final	0.996	0.787	0.2099	866	0.494

Table 4: Family-resolved checker audits ($n = 2048$ generation rollouts; f_1 = hostile family). "—": family absent, FN rate undefined.

Policy	O (base)	f_1 (base)	O (block)	f_1 (block)	O (shift)
clean s0	0.994	1017	0.635	650	0.998
clean s1	0.998	1020	0.459	470	0.995
corr s0	0.996	0	0.119	52	0.000
corr s1	0.996	0	0.114	15	0.000
corr s2	0.995	0	0.141	74	0.000
iid s0	0.995	1013	0.482	494	0.996
iid s1	0.997	1015	0.443	454	0.987

Table 5: Stress readouts ($n = 1024$ per condition): oracle success and f_1 emission counts on the base task, under route blocking, and under task shift.

F Audit and Stress Readouts

Tables 4 and 5 report the frozen family-resolved audits and the stress readouts (§6.1–§6.2); Figure 3 (main text) visualizes both. The audit rows quantify self-masking at its endpoint: the same checker whose f_1 FN rate is 0.49 ($n = 606$) at initialization measures as error-free in aggregate (0.0000, $n = 2048$) at every corr final, with the audited family absent from emissions (count 0; a rule-of-three upper bound on its emission rate of ≤ 0.0018 at $n = 2048$).

F.1 Stratified Advantage Re-Centering

The registered mitigation arm tests the minimal family-level intervention: after computing the raw group-relative advantages of a step, each stratum's (family's) mean advantage is subtracted within the stratum, preserving within-family ordering exactly and removing the cross-family comparison channel through which family-conditioned FN enters the drift law (§3, Theorem 1). Three arms run under the corr checker on the S1 policy, 150 steps (Table 6).

The all-strata variant preserves the targeted family ($m_{f_1} = 0.281$ vs. 0.000; realized aggregate FN rate stays near 0.3 rather than self-masking to 0) but, by re-centering the invalid stratum too, removes the systematic negative selection on invalid modes and caps oracle success near 0.5 — the same protection applied to a stratum that should be punished. The valid-only variant leaves the invalid stratum under raw advantages (invalid mass drains normally, final $o = 0.99$, clean-level) and restores the targeted family to dominance ($m_{f_1} \to 0.977$); the family-conditioned FN now falls on the single unprotected valid stratum, which goes extinct ($m_{f_0} \to 0.000$). The net hazard-sum-of-expectations, ≈ 0 in both variants, quantifies the compensation: the re-centering exactly offsets the checker's FN-driven drift wherever it is applied, and the boundary moves to wherever it is not. Repairing the checker (§F.2) or conditioning selection fully within families are the interventions in this environment that remove the boundary outright.

Arm	m_{f_1}	m_{f_0}	o
corr checker, no mitigation	0.000	0.996	1.000
stratum (all families re-centered)	0.281	0.027	0.482
stratum-valid (valid families only)	0.977	0.000	0.992

Table 6: Family-level mitigation under the corr checker (1.5B, seed 0).

F.2 POST-HOC REPAIR ARM

The mitigation arm above alters the selection channel before any damage; the question a practitioner actually faces is whether repairing the checker after collapse restores the lost family. Two 150-step runs continue the frozen corr checkpoints under the repaired (clean, $V = O$) checker with a fresh optimizer: repair-early from step 25 ($m_{f_1} = 0.0078$) and repair-late from step 150 ($m_{f_1} = 0$). Pre-registered criteria: f_1's valid share stays $\leq 2\%$ and never approaches its pre-collapse $\sim$77%; oracle pass returns to $\sim$1; any resurgence is transient blips, not sustained regrowth. All three are met. In repair-early, f_1 never exceeds its inherited 0.0078 and is extinct from step 50 at every evaluation, although every f_1 rollout the policy emits passes ($o_{f_1} = 1.0$ at all 14 nonzero appearances) — a fully favorable reward signal that the deleted family does not convert into recovery, the empirical face of Proposition 1(ii)'s strictly negative drift at $C(m) > 0$. In repair-late, f_1 emits nothing in 150 steps — the deletion is support-level. Oracle pass returns to 1.0 in both runs. A further observation is registered as post-hoc: once the invalid mode drains, the deterministic differential between the surviving valid families vanishes ($\mu \to 0$) and their composition performs a nearly neutral random walk with absorbing support loss — repair-late flips from f_2-dominated (0.758) to f_0-absorbed (1.0) within 50 steps, echoing the seed-dependent survivor of §B.1. Post-collapse repair therefore preserves the aggregate reliability the checker reports while the population it certifies keeps drifting; restoration of a deleted family requires non-reward mechanisms (Proposition 1(iii)).

Simulator quantification of the reopening question. A pre-registered simulator study (100 seeds per arm) probes whether the listed non-reward mechanisms restore the deleted family under each parameterization: phase 1 runs the corr checker from the calibration mixture until m_{f_1} decays to 0.029 (the algebraic law's slow tail), and phase 2 switches to the genuinely repaired checker for 300 steps under five arms — repair only; repair + replay (re-seeding f_1 to its phase-1 peak 0.58); repair + sampling temperature $\tau = 1.5$; repair + support floor $\epsilon = 0.02$; and exploration ($\tau = 1.5$) without repair. The two laws agree that repair alone never restores the family, and disagree on whether a reseeding can be held. Under the tabular law the repaired checker freezes valid-family shares exactly as Proposition 1(iii) states: f_1 holds its inherited 0.029 under repair, τ, and the floor, and a replay reset is maintained to the horizon (0.585 at step 300; restored in 100/100 seeds). Under the softmax law no arm restores: repair alone continues the (decelerating) relative decay to 0.020; the replay reset re-decays to the same endpoint 0.020 — the common-acceptance drag attracts the restored family back regardless of its reseeded share; temperature lifting produces only a transient peak of 0.082 and an endpoint of 0.047 ($1.6\times$ the inherited base, restoration in 0/100 seeds); the support floor lifts to ϵ-scale (0.021); and exploration without repair is the worst arm (0.016). The empirical repair-early arm above — decay to extinction although every emitted f_1 rollout passed — sits on the softmax side of this disagreement, consistent with real policies being logit-parameterized rather than tabular.

LLM reopening arm. The temperature arm of the simulator has a direct LLM-level test: continuing the frozen corr checkpoint at step 25 ($m_{f_1} = 0.0078$) under the repaired checker with training-time sampling temperature $\tau = 1.5$, evaluation protocol unchanged ($\tau = 1.0$). The pre-registered criterion was two-sided: f_1 sustained at ≥ 0.05 would demonstrate LLM-level recovery; continued decay confirms the softmax-law prediction. The outcome is the second branch, and more decisive than the simulator's transient lift (0.082): flattened training sampling surfaces the latent family immediately — f_1 appears in 13 of the first 18 training probes, peaking at 0.035 ($4.5\times$ its inherited share) — and then re-decays to zero by step 18, never to return; the frozen evaluation ($\tau = 1.0$) never observes it at any checkpoint. Token-level temperature is thus not a family-addressable knob: it re-opens the family's sampling support transiently, but the repaired checker's common-acceptance dynamics drive the reopened mass back out. In the same window the repair-late absorption replicates at a different initialization: the dominant f_2 (0.818 at init) is absorbed by the minority survivor f_0 (0.141) within $\sim$35 steps ($f_0 \to 0.99$, $f_2 \to 0$), and oracle pass returns to 1.0 — the aggregate reliability the checker certifies is preserved while the population it certifies narrows to a single family.

F.3 CROSS-MODEL-FAMILY REPLICATION

The corr-vs-iid contrast is replicated on a second model family (SmolLM2-1.7B-Instruct) at a qualitatively different calibration regime. Fresh calibration follows the frozen rule (hostile = argmax

Arm	Seed	m_{f_1} (train)	m_{f_1} (eval)	O-pass (eval)	agg. FN
corr	0	0.000	0.000	1.000	0.0000
corr	1	0.000	0.000	1.000	0.0000
iid	0	0.164	0.176	0.998	0.4039
iid	1	0.035	0.031	1.000	0.3945

Table 7: 7B minimal replication (30-step horizon, two seeds per arm): train-final and frozen-eval m_{f_1}, oracle success, and realized aggregate FN rate. corr extinguishes the targeted family and masks the checker's aggregate FN to 0; matched-iid preserves it $(0.035, 0.164)$ with persistent ≈ 0.4 FN.

valid share; r chosen for aggregate FN $= 0.4$): here the model can barely solve the environment at initialization ($o = 0.015$; 98.5% invalid rollouts; hostile family mass 0.009), so the "third of the population" precondition does not hold and the contrast is read at pool-quantum resolution. The corr checker first grows the hostile family through the invalid-draining learning phase (mass $0.004 \to 0.37$ by step 60 — the initial growth phase under a low threshold), then drives it extinct once oracle success approaches 1: eval mass $0.008 \to 0.201 \to 0.006 \to 0.000$, realized aggregate FN self-masking $0.58 \to 0.0000$. The matched-iid twin preserves and elevates the same family to dominance ($m_{f_0} = 0.973$ at step 150) with aggregate FN pinned at 0.41 — no self-masking. The corr-vs-iid contrast is $+0.97$, well beyond the registered margin; a second model family, a second calibration regime, and the threshold-crossing structure of Theorem 1 are all reproduced.

The 7B minimal replication (30-step horizon, two seeds per arm; Table 7) reproduces the causal signature in a different calibration regime: the targeted family starts rare ($m_{f_1}(0) \approx 0.03$), corr extinguishes it by step ~ 19 with realized aggregate FN $0.052 \to 0.0000$, and matched-iid preserves it with persistent ≈ 0.40 FN.

A third family failed the pre-registered competence gate: OLMo-2-0425-1B-Instruct, calibrated with the frozen rule (4096 rollouts), reached $o = 0.032$ with hostile mass 0.018 against the required ≥ 0.10 and ≥ 0.05, and was not trained. Two non-Qwen open 1B-class instruct models thus sit in the degenerate regime at this environment's difficulty, so the corr-vs-iid contrast at non-degenerate base competence is demonstrated within the Qwen family; the calibration-only gate fixes this boundary before any training outcome exists.

F.4 Naturalistic Audit on Published Benchmark Checkers

The audit protocol of §F transfers to published benchmark settings where no family labels are given: family discovery is itself a measurement problem. We audit two published-style checker configurations with an independent ground truth and a static presentation-family classifier, pre-registering the trigger threshold ($n_{\text{correct}} \geq 40$ in a family, family-conditioned FN ≥ 0.20, Fisher $p < 0.001$) before generation.

Math (GSM8K). Qwen2.5-1.5B-Instruct, test split $[0, 250)$, $k = 8$ samples per problem ($n = 2000$, $T = 0.8$). Ground truth: multi-pattern extraction plus numeric normalization against the dataset reference. Audited checkers: a narrow `\boxed`-only extractor (the dominant published pattern) and a lenient multi-pattern variant. Of 1393 gt-correct solutions (69.7%), 1380 use the `\boxed` presentation family, on which the narrow checker's FN is 0.0036 (exact binomial CI $[0.0012, 0.0084]$); the twelve bare-final presentations are FN $12/12 = 1.0$ (CI $[0.7354, 1.0]$) against a Fisher enrichment $p \approx 1.7 \times 10^{-25}$, while the lenient checker passes 11/12 of them. The structural signature is therefore present and extreme — but the misaligned family's mass ($12/2000 \approx 0.6\%$) falls far below the pre-registered trigger, a negative control: standard GSM8K RLVR pipelines align the checker with the dominant presentation family and sit in the safe regime; the mechanism requires checker–family misalignment rather than checker narrowness alone.

Code (HumanEval + EvalPlus ground truth). Sandbox-executed completions ($n = 1312$, 8 per problem) with the OFFICIAL HumanEval base unit tests as the audited published checker (the sparse reference-derived suite an RLVR pipeline actually trains against) and the EvalPlus v0.1.0 PLUS suites as independent ground truth (10 s budget both sides); families by AST classifier. Of 580 PLUS-correct completions (44.2% of emissions), the official suite rejects zero in every family (direct

$n = 498$, CI $[0, 0.0074]$; brute-force $n = 33$; recursion $n = 49$; Fisher $p = 1.0$ throughout): the PLUS suites are a superset of the official tests, so this configuration hosts no FN on valid solutions at this calibration. A pre-registered efficiency-channel check re-executes the official checker at the tight settings published RLVR pipelines deploy (1 s timeout, recursion 1000, 512 MB) against the same honest ground truth (10 s budget): FN is again zero in every family (direct $n = 498$, brute $n = 33$, recursion $n = 49$; Fisher $p = 1.0$ throughout), so the timeout channel carries no family structure here either; the audit re-executes the stored completions against extracted test segments.

Math (MATH-500). The third audit asks whether a stricter published checker develops form-linked false negatives when the answer space itself carries presentation diversity, which GSM8K's numeric references cannot exhibit. Same model and prompt protocol (published-style boxed coaching), $k = 8$, first 250 problems ($n = 2000$, $T = 0.8$, max 1024 tokens); ground truth is the published math-verify checker (0.9.0); the audited checker is the strict Minerva configuration vendored verbatim from the lm-eval minerva math task (last-`\boxed` extraction, the appendix-D normalization, sympy latex-parse equivalence with a 5 s timeout; archived upstream copy and vendored module both in the released code). Families are syntactic form classes of the final answer (integer, fraction, radical, decimal, text, other, bare-final). Of 910 gt-correct solutions (45.5%), the strict checker rejects 2.4% overall: integers ($n = 621$) and radicals ($n = 48$) are clean (FN $= 0.000$), fractions run 6.7% ($n = 119$, Fisher $p = 0.004$ vs. rest), decimals 21% ($n = 19$, $p = 8 \times 10^{-4}$) — no family meets the pre-registered trigger (mass ≥ 40, FN ≥ 0.20, $p < 0.001$), a negative control at the family level. The residual rejections are nonetheless form-linked and mechanistically specific, and each was reproduced in isolation under the vendored module: tuple answers such as $(3, \pi/2)$, which the latex parser cannot equate (a documented limitation of the checker), leading-dot reference decimals ($.0000672$ vs. 0.0000672), and the normalizer's article-substitution rule corrupting variable-a expressions written with spaces ($(a + 5)(b + 2) \mapsto (+5)(b + 2)$, rejected against $(a + 5)(b + 2)$). The presentation form of a correct solution thus modulates a published checker's verdict at the $\sim 2\%$ level — invisible in aggregates, absent from the integer-form families, and below our per-family trigger because the form axis is coarser than the failure modes it localizes.

The GSM8K and HumanEval configurations therefore sit in the safe regime of the mechanism — checker–family alignment (math) or uniform FN (code): the aggregate accuracy of a checker cannot distinguish these safe configurations from a misaligned one at initialization; the family-resolved audit can. The MATH-500 audit adds the complementary observation: even in a safe-regime configuration the published checker's residual FN is form-linked at the $\sim 2\%$ level, structured more finely than any presentation-family axis we pre-registered — the aggregate FN rate cannot locate that structure either.

All three audits run on the base policy (no RL): they measure the checker's family structure on the same distribution a training run would see at initialization, which is the quantity the drift law consumes.

F.5 Group-Size Dose–Response

Exact simulation of the hostile-family race (100 seeds per point, mapping B, $r_{f_1} = 0.5198$) across group sizes $G \in \{2, 4, 8, 16, 32\}$ shows the dose–response shape predicted by Theorem 1: extinction onset contracts monotonically with G (mean extinction step $922.6 \to 463.8 \to 368.1 \to 336.9 \to 325.2$, seed-sd ≤ 21; Figure 4); the contraction flattens for $G \gtrsim 8$ as the per-step number of sampled hostile rollouts $m_{f_1} G$ grows and the relative group-sampling variance in the exact drift (Theorem 1) shrinks. Extinction always occurs; G sets only the clock.

G Simulator Race Trajectories

Figure 5 summarizes the race to collapse over the registered sweep: the flip step contracts monotonically in r_{f_1} while the banked peak mass falls. Figure 6 shows the corresponding trajectories — family mass $m_{f_1}(t)$ (solid) against the rising threshold $\bar{q}(t)$ (dashed), with collapse beginning where the threshold overtakes $q_{f_1} = 1 - r_{f_1}$; higher r_{f_1} shortens the banking phase and lowers the banked peak.

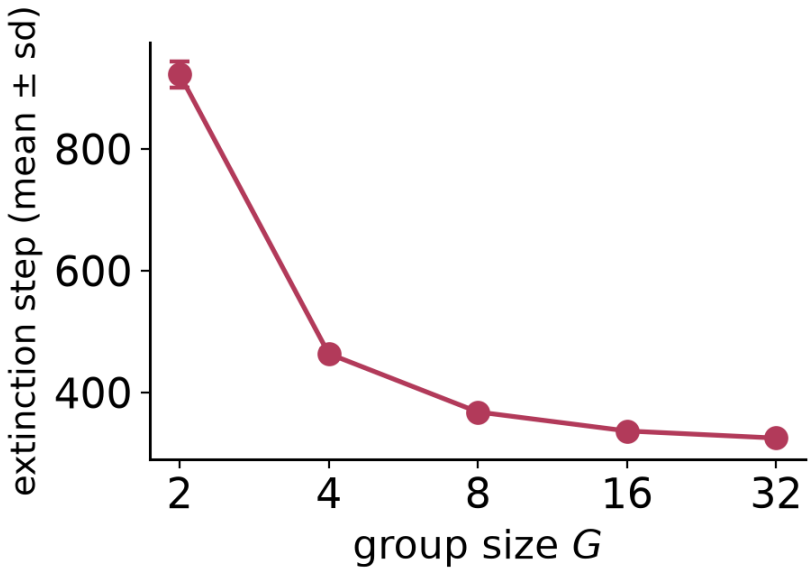

Figure 4: Extinction always occurs; G sets only the clock. Mean extinction step of the hostile family in the exact simulator (100 seeds per point) contracts monotonically with group size and flattens beyond $G \approx 8$; bars are seed standard deviations, on a log-2 axis.

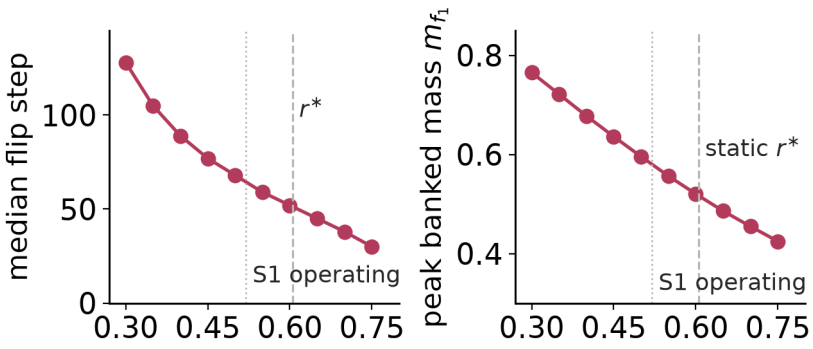

Figure 5: The race to collapse, summarized. Left: flip step — the step at which the hostile family's mass first drops below half its banked peak — contracts monotonically with the hostile-family FN rate r_{f_1} (128 at 0.30 down to 30 at 0.75; 20 seeds per point). Right: banked peak mass falls from 0.765 to 0.425 as r_{f_1} rises. The static initialization boundary $r^*_{f_1}(m_0) = 0.605$ (gray dashed) sits above the S1 operating point 0.5198 (red dotted): every registered point banks first and collapses only when the rising group-mean acceptance $\bar{q}(t)$ overtakes $q_{f_1} = 1 - r_{f_1}$.

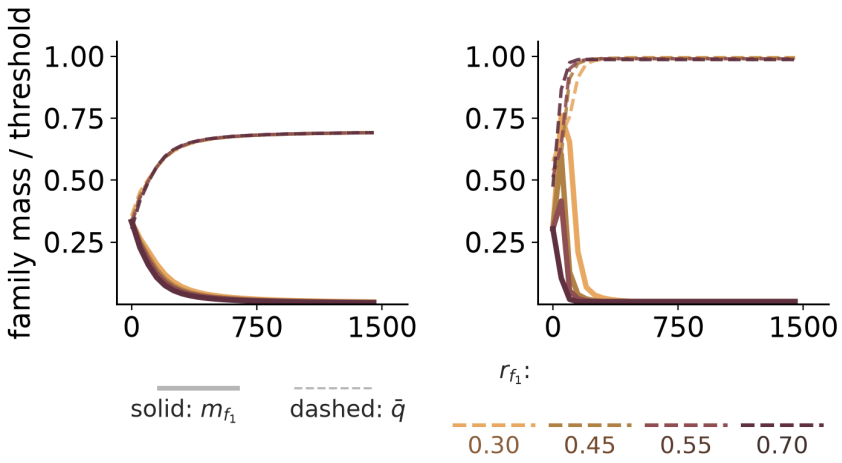

Figure 6: Family-mass trajectories $m_{f_1}(t)$ (solid) against the moving selection threshold $\bar{q}(t)$ (dashed), per sweep point. The shared legend strip below the panels maps line styles to family mass and threshold and swatch colors to the hostile-family FN rate r_{f_1} (0.30, 0.45, 0.55, 0.70; light to dark). Collapse begins where the threshold curve overtakes q_{f_1}; higher r_{f_1} both shortens the banking phase and lowers the banked peak mass.