EMIT: MEASURING THE RELIABILITY OF PER-PROMPT DECISIONS IN RLVR CURRICULA

Anonymous authors
Paper under double-blind review

ABSTRACT

Per-prompt controllers in reinforcement learning with verifiable rewards transform small groups of verifier outcomes into consequential training actions, yet the reproducibility and semantic fidelity of those decisions are rarely characterized. We introduce EMIT, a reliability-aware evaluation framework that treats adaptive curriculum control as a measurable decision process. EMIT formalizes exact-action reproducibility, connects difficulty concentration and decision boundaries to rollout requirements, and combines these predictions with source-traced implementation and verifier audits. Applied to frozen real-server rollouts, the framework reveals how sampling resolution and answer-format semantics govern controller behavior. A completion-token-matched multi-run study further shows substantial observed gains for both the released adaptive controller and a carefully tuned static curriculum over the no-curriculum reference. The adaptive controller attains the highest observed primary-task mean while closely tracking the tuned static curriculum across the broader suite. By making intermediate curriculum decisions auditable, EMIT provides a principled basis for reporting, sampling, pooling, and reliability-aware controller design.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) uses verifier feedback to strengthen mathematical reasoning (Shao et al., 2024; DeepSeek-AI et al., 2025). Recent curricula additionally map a small outcome group to a *per-prompt* action: how much assistance to reveal, whether to retain an item, or whether a rollout group should contribute to an update (Li et al., 2025; Lee et al., 2026; Yu et al., 2025; Xu & Fang, 2026; Mao et al., 2026). These controllers share a common interface: estimate a prompt state, map it to an action, and allow that action to shape training. This intermediate decision is scientifically important because it determines which experience the learner receives, yet it is usually evaluated only through the final model.

Near a decision boundary, two valid rollout groups can support different actions. Additional sampling improves resolution, whereas verifier semantics determine what the resolved quantity means—a concern complementary to the benefits of outcome and process verification (Cobbe et al., 2021; Lightman et al., 2024). A complete evaluation therefore needs three forms of evidence: repeatability of the executed action, validity of the verifier signal, and incremental training utility relative to a tuned pooled policy.

EMIT (Evaluating Measurement Integrity in Training) treats the controller as a measurement instrument. Its workflow (Figure 1) first defines reliability at the action level, then checks whether a simple sampling account predicts independent repeats, traces the exact decision unit through published and released implementations, and probes sensitivity to verifier semantics. These analyses let EMIT translate the diagnosis into reporting requirements and controller-design guidance. This ordering connects mathematical analysis, fixed-policy server measurements, released-code replays, and matched training evidence without conflating their estimands.

Closed-form predictions accurately track frozen measurements across decision families and checkpoints (Section 4). The implementation audit identifies small decision groups as a recurring design choice (Section 5), while the released-selector study demonstrates that verifier semantics can reshape the curriculum action distribution (Section 6). Finally, the protocol-frozen server comparison places this diagnostic evidence beside a matched multi-run training result (Section 7).

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.

Contributions.

1. A **decision-reliability instrument** that connects a controller's rule to repeatability, discard risk, and sample-budget design, with plug-in calibration on frozen server rollouts (Sections 3 and 4).
2. A **source-traced implementation audit** that links published specifications to an executed released selector and distinguishes sampling-driven variability from verifier sensitivity (Sections 5 and 6).
3. A **protocol-frozen paired comparison**: five paired real-server runs compare the released adaptive controller against the development-selected $(4 \rightarrow 4)$ static schedule under a common completion-token cap (Section 7), providing an uncertainty-aware view of incremental utility.

Evidence synthesis. Frozen-policy evidence and the protocol-frozen real-server study answer complementary questions and together connect mechanism-level diagnosis to downstream utility. The complete experimental scope and reproduction details appear in Appendix M, and Section 3.5 translates the reliability analysis into a concrete controller-design template.

2 FORMALIZATION

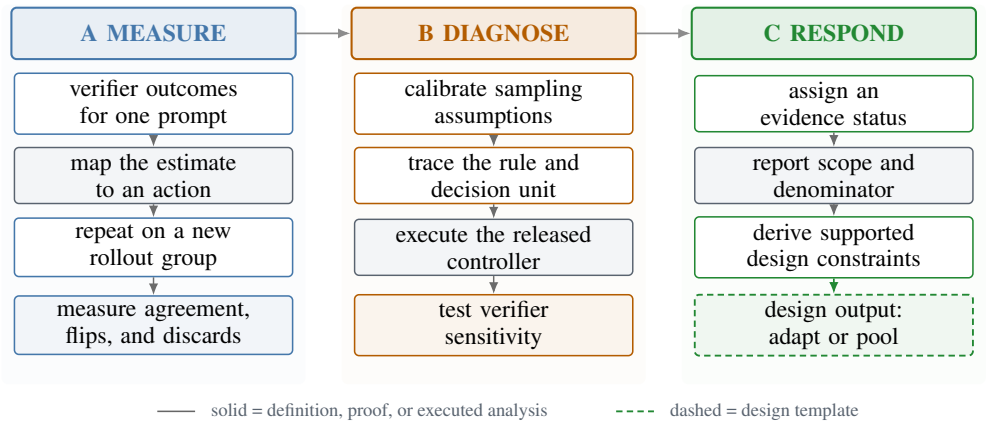

Figure 1: EMIT separates measurement, diagnosis, and response. **A:** measure two-repeat exact-action consistency. **B:** audit the sampling account, implemented rule, and verifier. **C:** record evidence status and supported design constraints. Solid paths denote executed analyses of frozen server outputs; the dashed path denotes the resulting controller-design template.

Each rule maps a small-group estimate to a per-prompt action: **estimate** $\rightarrow$ **threshold/invert** $\rightarrow$ **act** (Figure 1). We model verifier-positive counts as $X \sim \mathrm{Bin}(G, p_x)$ and set $\hat{p} = X/G$, where p_x denotes the fixed-verifier positive probability under the specified policy, decoder, and prompt/hint. Semantic validity is audited as a separate property. Frozen-sweep decision functionals are consistent with this sampling account at the measured resolution (Section 4). The common estimate–action structure yields the exact-action consistency criterion below.

EMIT evaluation procedure. Given a controller R, frozen rollout groups, and a source-provenance record, EMIT returns a four-part evidence profile: action repeatability, sampling-model calibration, verifier sensitivity, and matched downstream utility. It first re-executes the released decision path on independent groups, evaluates the corresponding finite-sample prediction, and perturbs only audited verifier semantics. A protocol-frozen multi-run comparison then evaluates whether the measured controller behavior translates into incremental training value. Keeping these outputs separate makes the procedure applicable to threshold, filtering, and multi-action curriculum rules.

Exact-action consistency criterion. Let θ_x collect every latent quantity used by a controller (for SEELE, the verifier-positive probabilities across hint levels), and let $\mathcal{D}_G$ denote one finite rollout group. For a finite action set $\mathcal{A}$, define $q_a(\theta_x) = \Pr\{R(\mathcal{D}_G) = a \mid \theta_x\}$. Its two-repeat *decision consistency* is $C_R(\theta_x) = \sum_{a \in \mathcal{A}} q_a(\theta_x)^2$: the probability that independent groups produce the

same action. It is at least $1/|\mathcal{A}|$ and, for a binary scalar threshold, reduces to Equation (1). Consistency quantifies action reproducibility. EMIT evaluates semantic validity, informativeness, and downstream utility as complementary axes; a constant controller, whose consistency is one, illustrates why these axes are reported jointly. Its population average depends on the selected difficulty distribution.

Finite action spaces are covered directly. Binary scalar thresholds yield the closed forms below, while complete-rule execution supplies the corresponding evaluation for multi-action controllers.

3 MEASUREMENT THEORY

3.1 A CONCENTRATION–SAMPLE-SIZE LAW

Let a rule act iff $\hat{p} \geq \tau$, i.e. iff $X \geq m$ with $m = \lceil \tau G \rceil$, and let $F(p) = \mathbb{P}(X \geq m)$ be its acceptance probability. Two independent replications agree with probability

$$C(p) = F(p)^2 + (1 - F(p))^2. \tag{1}$$

Proposition 1 (local-Gaussian sample law)**.** *Let difficulty near the threshold be $p = \tau + \sigma Z$ with $Z \sim \mathcal{N}(0,1)$ and negligible boundary truncation. Under the local Gaussian approximation $F(p) = \Phi((p-\tau)/s)$, $s^2 = \tau(1-\tau)/G$, define $\rho^2 = G\sigma^2/[\tau(1-\tau)]$. Then average consistency is*

$$\bar{C}_G = \frac{1}{2} + \frac{1}{\pi} \arcsin\left(\frac{\rho^2}{1+\rho^2}\right), \qquad G_\kappa = \frac{\tau(1-\tau)}{\sigma^2} \frac{r_\kappa}{1 - r_\kappa}, \quad r_\kappa = \sin\big(\pi(\kappa - \tfrac{1}{2})\big). \tag{2}$$

Thus $G_\kappa \propto \tau(1-\tau)\sigma^{-2}$. The inverse-square scaling extends to other location–scale densities under the same local approximation, with a distribution-specific factor; the exact finite-binomial behavior is evaluated separately below.

Sketch (full argument in Appendix A). With $F = \Phi(\rho Z)$, $\mathbb{E}[F] = 1/2$ and a standard bivariate-normal orthant identity gives $\mathbb{E}[F^2] = 1/4 + (2\pi)^{-1} \arcsin(\rho^2/(1+\rho^2))$. Substitute into $\mathbb{E}[F^2 + (1-F)^2]$ and invert. □

Interpretation. Equation (2) is a signal-to-noise relation: resolving a distribution of width σ requires standard error proportional to that width, whereas binomial standard error decreases as $G^{-1/2}$. The frozen comparisons in Section 4 test the associated fixed-G binomial functionals without outcome-fitted calibration constants. Under concentrated difficulty, resolving small differences near τ calls for a larger reliability budget; this is precisely the regime in which adaptive curricula focus their decisions. The law therefore turns an observed difficulty distribution into an interpretable sampling requirement.

3.2 CONSISTENCY IS PINNED NEAR $\frac{1}{2}$ AT THE TARGET

Corollary 1. *At a prompt sitting exactly on the target ($p = \tau = \frac{1}{2}$) with G even, symmetry of $\mathrm{Bin}(G, \frac{1}{2})$ gives $F = \frac{1}{2} + P_0/2$ with $P_0 = \binom{G}{G/2}/2^G$, hence $C(\delta_\tau, G) = \frac{1}{2} + P_0^2/2 \approx \frac{1}{2} + 1/(\pi G)$.*

This pointwise result complements Proposition 1, which averages over a positive-width distribution whose mass leaves the near-chance band as G grows. At the target itself, consistency instead approaches $\frac{1}{2}$. The even-G decrease is a second-order tie effect: for odd G there is no central atom and $F = C = \frac{1}{2}$ exactly (Appendix J). The robust implication for a non-degenerate binomial threshold is that continuity of $F(p)$ from 0 to 1 guarantees some p^* with two-run consistency $\frac{1}{2}$. Increasing G narrows the near-chance band without raising this worst-case floor (Figure 2).

3.3 EXCLUSIONS UNDER ZERO-VARIANCE FILTERING

Theorem 1 (single-draw zero-variance exclusion)**.** *Let a rule discard a group iff $X \in \{0, G\}$. A prompt with $0 < p < 1$ has nonzero expected Bernoulli reward variance, yet is observed all-same, and thus excluded, with probability*

$$D(p) = p^G + (1-p)^G, \tag{3}$$

which $\to 1$ *as* $p \to 0$ *or* 1 *and is minimized at* $p = \frac{1}{2}$ *with* $D = 2^{1-G}$. *The population single-draw exclusion rate under difficulty* μ *is* $\int D\, d\mu$. *This theorem characterizes observation-driven exclusion, while the matched server study evaluates downstream training utility.*

Because latent p is not directly observed, our 16-rollout reference defines a matched observable using the original group A and an independent group B, each of size $G = 8$. Writing $d_g(p) = p^g + (1-p)^g$, this conditional *reveal rate* is

$$R_G(\mu) = \Pr(A \cup B \text{ mixed} \mid A \text{ all-same}) = \frac{\int [d_G(p) - d_{2G}(p)]\, d\mu(p)}{\int d_G(p)\, d\mu(p)}. \tag{4}$$

It tests an observable implication of Equation (3); it is not $\int D\, d\mu$ itself. DAPO's dynamic sampling applies exactly this rule, as does GP-POPO (Mao et al., 2026), which defines an ineffective group by zero reward variance; its authors abbreviate it POPO too, so we write GP-POPO. POPO discards only the $X{=}0$ tail—the same analysis with one term. Unlike Proposition 1 this result is irreducibly discrete—about the atoms at $X{=}0$ and $X{=}G$—so the binomial structure is essential; Section 4 matches Equation (4) to within one point on three models.

3.4 Design tool: required G

The first-crossing scan in Table 3 prices target consistency under several difficulty distributions. The required rollout budget grows sharply as difficulty concentrates near the boundary. Consequently, a group size adequate for a broad prompt population may provide limited resolution for a curriculum intentionally targeting a narrow difficulty band. Short non-monotone ranges induced by binomial rounding and parity further motivate evaluating the exact discrete rule rather than relying only on asymptotics.

3.5 Decision reliability under fixed shrinkage

Shrinkage is a standard response to noisy per-prompt estimates: it borrows strength across prompts through a population or difficulty-bucket prior, and recent RLVR work applies this principle to the gradient baseline (Han et al., 2026). For a thresholded action, reliability depends on both the estimator and the map applied before the decision.

Lemma 1 (fixed-map threshold equivalence)**.** *Fix* $w \in (0, 1]$ *and* $\mu \in [0, 1]$ *independently of the evaluated* $\hat{p}$, *and let* $T(\hat{p}) = w\hat{p} + (1-w)\mu$. *Then* $\mathbf{1}[T(\hat{p}) \geq \tau] = \mathbf{1}[\hat{p} \geq \tau']$ *with* $\tau' = (\tau - (1-w)\mu)/w$. *For* $\hat{p} = X/G$ *and* $0 < \tau' \leq 1$, *this rule has* $\min_{p\in[0,1]} C_{\tau'}(p) = \frac{1}{2}$. *If* $\tau' \leq 0$ *or* $\tau' > 1$, *the decision is constant and its consistency is* 1; *at* $w = 0$ *it is likewise constant, determined by* $\mu \geq \tau$.

Our same-sample diagnostic is data-dependent: it fits rank buckets, moments, and prompt weights from the noisy vector being evaluated. This EMIT analysis construction is distinct from EBPO's released training estimator and is more expressive than a relocated threshold. The exact counterexample in Appendix A and cross-fit controls in Appendix B and Table 4 show that its consistency depends on the population and fitting protocol. EMIT therefore evaluates the executed map directly and treats shrinkage, allocation, and abstention as distinct, independently auditable design choices.

Reliability-aware action gate. EMIT turns the reliability analysis into an actionable controller interface. For a target two-run consistency $\kappa \in (\frac{1}{2}, 1)$, define

$$q_\kappa = \frac{1 + \sqrt{2\kappa - 1}}{2}, \qquad s = \sqrt{\frac{\hat{p}(1-\hat{p})}{G}}, \qquad \text{act iff } |\hat{p} - \tau| \geq \lambda \Phi^{-1}(q_\kappa) s, \tag{5}$$

where the pre-specified multiplier λ controls the margin beyond the Gaussian base rule; our instantiation uses $\lambda = 1.5$. When the gate is closed, the controller returns a pre-specified pooled action rather than committing an unresolved per-prompt choice. The same margin supplies an allocation score: cells with the largest estimated additional sample requirement receive new rollouts first. The resulting design composes measurement, abstention, and targeted allocation in one transparent rule, directly connecting EMIT's diagnosis to curriculum control.

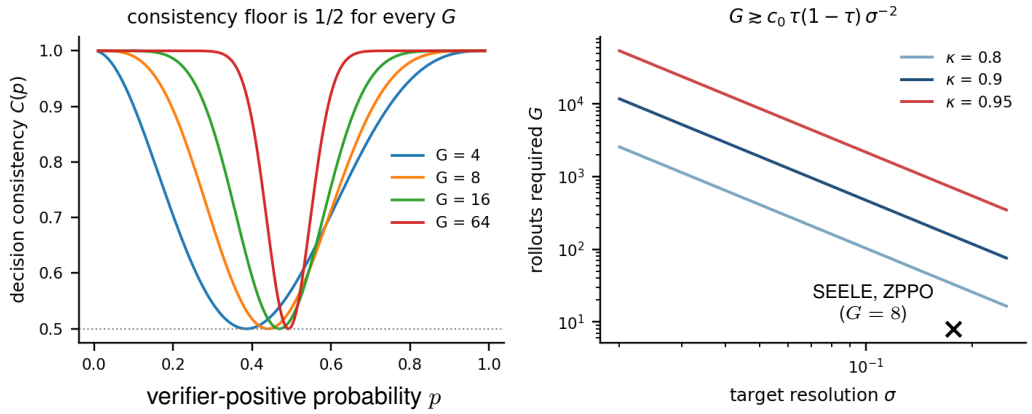

Figure 2: Decision-consistency geometry and its sampling implication. **Left:** every $C(p)$ curve reaches its chance-agreement floor at the decision boundary; larger groups narrow the low-consistency region. **Right:** the inverse-square relation converts local difficulty concentration into a required rollout budget. The marker locates G=8 for SEELE and ZPPO; DAPO uses G=16 (Table 1).

4 Empirical validation

Here "frozen" means analysis of stored real-server rollouts rather than simulated model responses; the theory-side Monte-Carlo design scan remains separately identified in Section 3.4.

Finite-group sampling account. The binomial law provides a direct account of decision-level behavior in the controlled frozen-policy regime and supports the calibrated comparisons below.

Direct plug-in calibration. Figure 3 compares plug-in predictions with frozen measurements across consistency, conditional-reveal, and selector-flip families. Difficulty distributions and measurements share a rollout source, creating a direct calibration test without an outcome-fitted constant. Predictions recover both the ordering and scale of measured behavior across models, with sub-percentage-point mean absolute error. The multi-action selector is the most demanding calibration case because its fit-and-snap path compounds several sampling events.

Disjoint-prompt calibration. We additionally separate distribution estimation from agreement measurement across prompt subsets. For each of 200 random splits, one half of the 400 prompts estimates the difficulty distribution and the other measures independent A/B action agreement at the original group size. Agreement remains strong across these disjoint populations (r=0.997, MAE 0.0026); using only eight prediction-side rollouts preserves the pattern (r=0.993, MAE 0.0048). This establishes that the sampling account transports across disjoint prompt subsets while retaining its decision-level interpretation.

Additional decision rules. The zero-variance and selector-flip families execute DAPO's rule and SEELE's selector, respectively. Replaying ZPPO's admission threshold and POPO's one-tailed filter on the same rollouts yields the cross-model ordering predicted by the sampling account. The five-action fit/invert/snap selector poses a stronger repeatability test than a single threshold at the same group size, and a threshold sweep preserves this qualitative ordering. Thus the diagnostic transfers across several published rules while remaining tied to their executed semantics.

Rule fidelity and model variation. The rule-specific calibrations replay their exact executed rules within the corresponding source cells (Appendix G). Across models, mass near the target explains the observed ordering, and the measured difficulty distributions recover the model-specific magnitude of selector variability.

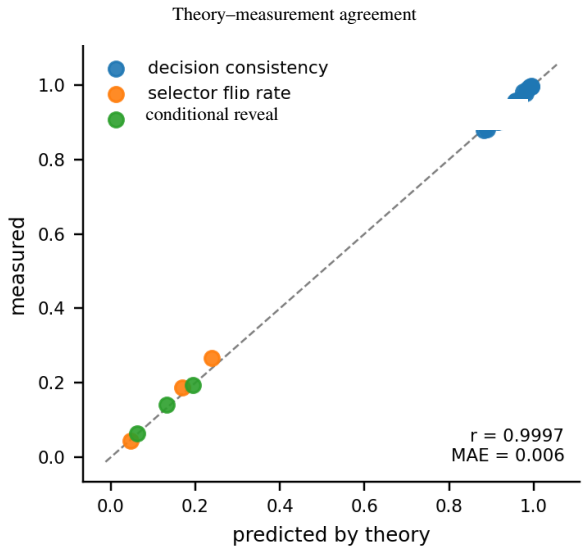

Figure 3: Frozen-rollout calibration of the sampling account. Points compare predictions with measurements for exact-action consistency, conditional reveal, and selector flips; the diagonal is exact agreement. No outcome-fitted constant is used, and the comparison executes the measured fixed-policy regime directly. The dense upper-right cluster contains the consistency cases.

5 Retrospective Audit of Recent Configurations

Eligibility required a discrete per-prompt action computed from a small verifier-outcome group and a recoverable primary paper or release specification. The resulting pre-specified five-method case series spans threshold, filtering, and multi-action rules and is source-traced from paper text, configuration, or launch scripts (Appendix F).

Table 1: Rule-specific diagnostics for recent per-prompt RLVR configurations. Here $p = 0.15$ is a common illustrative operating point. Flip and exclusion incidence form distinct rule-specific profiles. C_D is the two-repeat consistency of the binary keep/exclude action, and the SEELE row isolates one binary threshold component of its five-action selector.

controller/rule	G	reference state	EMIT diagnostic	provenance
SEELE threshold	8	binary component, $p{=}\tau$	flip .463	**verified**
ZPPO	8	$p{=}0.5$	flip .463	**verified**
DAPO	16	$p{=}0.15$	exclude .074; $C_D{=}.863$	**verified**
POPO	8	$p{=}0.15$	exclude .272; $C_D{=}.604$	**verified**
GP-POPO	8	$p{=}0.15$	exclude .272; $C_D{=}.604$	rule verified

At the displayed operating point, the POPO and GP-POPO rates agree to reported precision because the all-correct tail is negligible, despite their distinct one- and two-tailed rules. The fixed (G, p) choices and row-level provenance are documented in Appendix F.

ZPPO makes the tension explicit: it places its strict admission boundary at $\bar{r}{=}0.5$ on the stated rationale that reward standard deviation is largest there. At latent $p = 0.5$, Bernoulli estimator variance is maximal and the $G = 8$ admission action flips with probability 0.463; the discrete consistency minimum lies nearby. Thus a point selected for informativeness also benefits from an explicit repeatability report, allowing information content and decision stability to be assessed together.

6 CASE STUDY: SEELE

Experimental setup. We draw a pre-specified uniform sample of 400 prompts from SEELE's released DeepMath training set. Its five-level action reveals progressively longer prefixes of a segmented hint. For every prompt and level, real-server generation produces 16 rollouts that are partitioned into two independent decision groups, matching the released selector's group size. The principal sweeps use Qwen2.5-Math-7B (Yang et al., 2024), Qwen2.5-3B (Qwen Team, 2024), and Llama-3.2-3B (Meta AI, 2024); a matched base/instruction-tuned pair supports the format decomposition (Table 5). Generation uses vLLM (Kwon et al., 2023) under the frozen decoding configuration specified in Appendix H.

Strict grading executes SEELE's shipped answer-tag reward. A branch-audited tolerant extractor additionally recognizes boxed or final numeric expressions and serves as a controlled semantic sensitivity condition. We define $P(\text{format})$ as the fraction of rollouts emitting a strict-parseable answer, irrespective of content. Every action is produced by the upstream `hint_manager` path (Appendix K), thereby isolating released selector behavior on the frozen outputs.

6.1 VERIFIER SEMANTICS CONCENTRATE RELEASED-SELECTOR ACTIONS

Under the shipped grader, the fitted selector assigns the maximum hint to every audited Math and Llama prompt and to a large majority of Qwen prompts (Figure 4). For the first two models, the adaptive selector and a fixed maximum-hint schedule therefore induce the same frozen action vector, while Qwen retains a broader action distribution. This pattern establishes action adaptivity as a directly measurable decision-level property alongside terminal metrics.

Mechanistically, the released fit-and-cap path appends a top-right anchor to complete the observed curve. Every audited prompt then traverses the fit-and-inverse stage, while the final concentration emerges during discretization. This branch trace localizes the effect to the released selector and complements the cross-method sampling audit in Section 5.

6.2 MECHANISTIC INTERPRETATIONS

On the same Math-7B outputs, tolerant extraction spreads actions across every hint level and raises action entropy (Figure 4), locating a verifier-to-action pathway. Across checkpoints, maximum-level selection declines as strict parseability rises, connecting initial controller state to the completion-token-matched evaluation in Section 7.

6.3 VERIFIER SEMANTICS STRONGLY SHAPE ACTIONS

The expected strict-positive rate factors as $P(\text{format})P(\text{math} \mid \text{format})$. Across checkpoints, strict parseability varies by more than an order of magnitude. In the matched base/instruction pair, the plotted format and conditional-content factors separate sharply (Figure 6). A separate sensitivity analysis finds a $0.99\times$ instruction-tuned-to-base extractor-score ratio when tolerant credit is taken at face value (Appendix I). This contrast specifically identifies answer-format sensitivity under the released reward semantics, while larger rollout groups sharpen the complementary sampling analysis.

6.4 SELECTOR BEHAVIOR UNDER RELEASED-STATE REPLAY

State-faithful replay preserves the strong maximum-hint concentration under strict grading, while tolerant grading reproduces the paired tolerant replay (Appendix L and Table 8). This localizes the action change to verifier semantics within the released frozen-policy lifecycle.

7 COMPLETION-TOKEN-MATCHED TRAINING EVALUATION

The protocol-frozen study pairs five runs of no-curriculum GRPO, the development-selected $(4 \to 4)$ static schedule, and released SEELE on Qwen2.5-3B. MATH500 is primary (Lightman et al., 2024; Hendrycks et al., 2021), with a broader benchmark suite reported below (Cobbe et al., 2021; He et al., 2024; Lewkowycz et al., 2022). All arms share data, batch construction, token cap, and

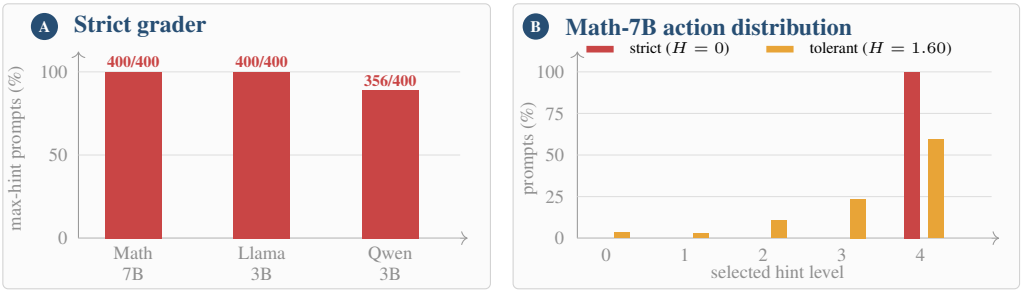

Figure 4: Released-selector actions on frozen real-server rollouts. **A:** the shipped grader strongly concentrates maximum-hint actions. **B:** changing only the audited extractor recovers a distribution across all hint levels. Both panels isolate the verifier-to-action pathway on the same outputs.

checkpoint rule (Appendix H and Table 6), adding a tuned-static, common-budget comparison to the upstream equal-response evaluation (Li et al., 2025).

Completed real-server runs use the development-selected static schedule and a common calibrated cap of 1.024 billion completion tokens. Under this matched budget, both curriculum arms have higher observed mean accuracy than the no-curriculum reference on every benchmark. Released SEELE attains the highest primary MATH500 mean and closely tracks the development-selected static schedule across the suite. The paired intervals in Figure 5 quantify the precision of that incremental comparison, while the aggregate table represents every run. The pre-specified MATH500 contrast favors SEELE by $+0.3$ percentage points, with a paired 95% interval of $[-0.7, +1.3]$ and paired Cohen's d_z=0.40. Together, the results show that the released adaptive interface preserves the substantial curriculum gains under the matched budget while achieving the strongest observed mean on the primary benchmark.

Table 2: Observed real-server exact match at the protocol-frozen Qwen2.5-3B endpoint (mean $\pm$ sample SD in percentage points across five paired runs); static is development-selected.

arm	M500	AMC23	AIME24	OlyBench	GSM8K	Minerva	Δ M500
GRPO	45.8 ± 1.4	24.0 ± 2.9	1.3 ± 1.8	15.0 ± 0.8	78.5 ± 0.8	17.3 ± 1.5	–
selected static $(4 \to 4)$	64.7 ± 0.9	39.0 ± 2.2	5.3 ± 1.8	28.3 ± 0.5	85.9 ± 0.6	25.7 ± 0.9	0.0
released SEELE	65.0 ± 0.7	39.5 ± 2.1	6.0 ± 1.5	28.3 ± 0.6	85.8 ± 0.3	26.0 ± 0.8	$+0.3$

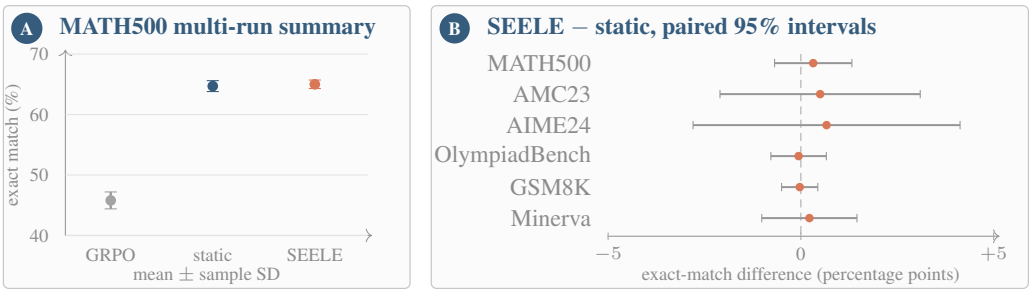

Figure 5: Aggregate real-server terminal results. **A:** multi-run MATH500 means with sample-standard-deviation whiskers. **B:** paired SEELE-minus-static means with paired 95% intervals across the benchmark suite.

8 INTERPRETING RELIABILITY EVIDENCE

EMIT combines decision consistency, verifier validity, and incremental utility. Calibration, semantic contrasts, and matched comparisons measure these axes; Section 3.5 instantiates the corresponding reliability-aware gate.

9 RELATED WORK

Adaptive RLVR curricula. Group-relative optimization constructs learning signals from within-group rewards (Shao et al., 2024), while recent controllers convert small rollout groups into prompt-level assistance, replay, filtering, or prioritization decisions (Li et al., 2025; Lee et al., 2026; Yu et al., 2025; Xu & Fang, 2026; Mao et al., 2026). Related methods study online difficulty filtering (Bae et al., 2026) or combine reinforcement learning with hint-based supervision (Liu et al., 2025a; Huang et al., 2025). EMIT complements these algorithms by treating the executed action as the object of measurement and asking whether finite-sample evidence supports a repeatable decision across threshold, filtering, and multi-action rules.

Curriculum learning and measurement. Classical curriculum learning orders examples by difficulty or learning progress (Bengio et al., 2009; Graves et al., 2017; Portelas et al., 2020; Narvekar et al., 2020). EMIT adds a distinct question: how reliably does the available evidence support the selected action? This action-consistency view connects training-time control to item response theory and adaptive testing (Lord, 1980; van der Linden & Glas, 2000), extending measurement ideas used to compare or compress evaluation sets (Vania et al., 2021; Maia Polo et al., 2024).

Shrinkage and rollout allocation. Empirical-Bayes shrinkage stabilizes noisy item estimates (Robbins, 1956; Efron & Morris, 1975), and EBPO applies this principle to a group-relative gradient baseline (Han et al., 2026). Rollout-allocation methods target coverage or training value (Wang et al., 2026; Yao et al., 2026), continuing the broader idea of prioritized experience (Schaul et al., 2016). EMIT targets decision reliability: it characterizes how shrinkage interacts with thresholds and allocates additional evidence against an explicit consistency target.

Verifier semantics and reproducibility. Outcome and process verifiers provide scalable supervision (Cobbe et al., 2021; Lightman et al., 2024), while imperfect or exploitable rewards can alter the optimized behavior (Amodei et al., 2016; Skalse et al., 2022; Gao et al., 2023; Shao et al., 2025; Cai et al., 2025). EMIT traces such semantic changes through the resulting curriculum-action distribution. This complements multi-run reporting for final performance (Henderson et al., 2018; Agarwal et al., 2021; Bouthillier et al., 2021) with paired analysis of the intermediate decisions that shape the training trajectory. Detailed comparisons appear in Appendix D.

10 CONCLUSION

EMIT establishes per-prompt curriculum decisions as measurable objects in RLVR, connecting sampling, executed actions, verifier semantics, and downstream utility. Its theory and real-server analyses recover decision behavior and yield concrete design guidance. Under a matched token budget, the released adaptive controller attains the highest observed primary-task mean and closely tracks the tuned static curriculum across the broader suite; both deliver substantial observed gains over the no-curriculum reference. Together, EMIT provides a rigorous foundation for reliability-aware design.

AI USE STATEMENT

The authors used AI assistance only for manuscript writing and language polishing. The authors conducted the scientific analyses and reported server experiments, verified every AI-assisted change and all scientific content, and take full responsibility for the manuscript.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *arXiv preprint arXiv:2108.13264*, 2021.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, Jeongyeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoning oriented reinforcement learning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 700–719, 2026. doi: 10.18653/v1/2026.eacl-long.30. URL https://aclanthology.org/2026.eacl-long.30/.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning (ICML)*, 2009. https://doi.org/10.1145/1553374.1553380.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769, 2021.

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

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. https://doi.org/10.18653/v1/2020.emnlp-main.745.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? statistical power analysis in deep reinforcement learning experiments. *arXiv preprint arXiv:1806.08295*, 2018.

DeepSeek-AI, 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, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong

Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Bradley Efron and Carl Morris. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975. doi: 10.1080/01621459.1975.10479864.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. https://doi.org/10.18653/v1/p19-1267.

Alex Graves, Marc G. Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. *arXiv preprint arXiv:1704.03003*, 2017.

Kevin Han, Yuhang Zhou, Mingze Gao, Gedi Zhou, Serena Li, Abhishek Kumar, Xiangjun Fan, Weiwei Li, and Lizhu Zhang. EBPO: Empirical bayes shrinkage for stabilizing group-relative policy optimization. *arXiv preprint arXiv:2602.05165*, 2026.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 3828–3850, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2103.03874.

Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo M. Ponti, and Ivan Titov. Blending supervised and reinforcement fine-tuning with prefix sampling. *arXiv preprint arXiv:2507.01679*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180.

Byung-Kwan Lee, Ximing Lu, Shizhe Diao, Minki Kang, Saurav Muralidharan, Karan Sapra, Andrew Tao, Pavlo Molchanov, Yejin Choi, Yu-Chiang Frank Wang, and Ryo Hachiuma. Zone of proximal policy optimization: Teacher in prompts, not gradients. *arXiv preprint arXiv:2606.18216*, 2026. URL https://arxiv.org/abs/2606.18216v3. Version 3.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ziheng Li, Zexu Sun, Jinman Zhao, Erxue Min, Yongcheng Zeng, Hui Wu, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Xu Chen, and Zhi-Hong Deng. Staying in the sweet spot: Responsive reasoning evolution via capability-adaptive hint scaffolding. *arXiv preprint arXiv:2509.06923*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. UFT: Unifying supervised and reinforcement fine-tuning. *arXiv preprint arXiv:2505.16984*, 2025a.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, 1980.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Yixiu Mao, Yun Qu, Qi Wang, Heming Zou, and Xiangyang Ji. RLVR without ineffective samples: Group prioritized off-policy optimization for LLM reasoning. *arXiv preprint arXiv:2606.01281*, 2026.

Meta AI. Llama 3.2 model card. GitHub repository, 2024. URL https://github.com/meta-llama/llama-models.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50, 2020. URL https://jmlr.org/papers/v21/20-212.html.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. Automatic curriculum learning for deep RL: A short survey. *arXiv preprint arXiv:2003.04664*, 2020.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Herbert Robbins. An empirical bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 157–163. University of California Press, 1956. URL https://projecteuclid.org/euclid.bsmsp/1200501653.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *International Conference on Learning Representations*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in RLVR. *arXiv preprint arXiv:2506.10947*, 2025. URL https://arxiv.org/abs/2506.10947.

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.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *arXiv preprint arXiv:2209.13085*, 2022.

Wim J. van der Linden and Cees A. W. Glas (eds.). *Computerized Adaptive Testing: Theory and Practice*. Kluwer Academic Publishers, 2000.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. Comparing test sets with item response theory. In *Annual Meeting of the Association for Computational Linguistics (ACL/IJCNLP)*, 2021. https://doi.org/10.18653/v1/2021.acl-long.92.

Tao Wang, Shuo Li, Yan Sun, Dongsheng Ding, and Edgar Dobriban. Where to spend rollouts: Hit-utility optimal rollout allocation for group-based RLVR. *arXiv preprint arXiv:2605.07114*, 2026.

Mingwei Xu and Hao Fang. Beyond negative rollouts: Positive-only policy optimization with implicit negative gradients. *arXiv preprint arXiv:2605.06650*, 2026.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024. URL `https://arxiv.org/abs/2409.12122`.

Zhiyuan Yao, Yi-Kai Zhang, Yuxin Chen, Yueqing Sun, Zishan Xu, Yu Yang, Tianhao Hu, Qi Gu, Hui Su, and Xunliang Cai. CoBA-RL: Capability-oriented budget allocation for reinforcement learning in LLMs. *arXiv preprint arXiv:2602.03048*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan 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, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

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? *arXiv preprint arXiv:2504.13837*, 2025.

A Proofs and derivations

Equation (1). Let the two independent replications accept with the same probability $F(p)$. They agree iff both accept or both reject: $C(p) = F^2 + (1-F)^2$. Writing $f = F(p), C = 2f^2 - 2f + 1$ is strictly convex with minimizer $f = \frac{1}{2}$ giving $C = \frac{1}{2}$; since F is continuous and increasing from 0 to 1 (for $1 \le m \le G$) there is a unique p^* with $F(p^*) = \frac{1}{2}$, so $\min_p C(p) = \frac{1}{2}$. This uses only continuity and monotonicity of F—not the detailed binomial form—so the same floor holds whenever a continuous acceptance curve crosses $\frac{1}{2}$; it is the near-boundary *width* that carries the sampling law.

Proposition 1. Under the local model, $F = \Phi(\rho Z)$ with $\rho = \sigma\sqrt{G/[\tau(1-\tau)]}$ and Z, U, V independent standard normals. Then $\mathbb{E}F^2 = \Pr(U \le \rho Z, V \le \rho Z)$. The pair $(\rho Z - U, \rho Z - V)$ is centered bivariate normal with correlation $c = \rho^2/(1+\rho^2)$, so its positive-orthant probability is $1/4 + (2\pi)^{-1}\arcsin(c)$. Symmetry gives $\mathbb{E}F = 1/2$, hence $\bar{C}_G = 2\mathbb{E}F^2 - 2\mathbb{E}F + 1 = 1/2 + \pi^{-1}\arcsin(c)$. Setting $\bar{C}_G = \kappa$ gives $c = r_\kappa = \sin(\pi(\kappa - 1/2))$ and $\rho^2 = r_\kappa/(1-r_\kappa)$, which yields Equation (2). For a fixed non-Gaussian location–scale density, the same change of variables shows dependence on ρ and thus the σ^{-2} scaling, but not this closed-form factor.

Table 3: First-crossing rollout budgets for target consistency at $\tau = 1/2$, estimated with $3 \cdot 10^4$ Monte-Carlo draws (integers to approximately $\pm 10\%$). U, B_{22}, and $B_{.5}$ denote Uniform$(0,1)$, Beta$(2,2)$, and Beta$(1/2,1/2)$; N_σ is a truncated $\mathcal{N}(1/2,\sigma^2)$ on $[0,1]$. "Outside scan" means that no first crossing was obtained within the evaluated finite grid; it is not an extrapolated budget.

prior	κ=0.8	κ=0.9	κ=0.95	more-G-is-worse over
U	7	30	122	—
B_{22}	15	69	291	—
$B_{.5}$	3	12	49	—
$N_{0.10}$	105	484	1943	$G \in [2, 24]$
$N_{0.05}$	423	1939	*outside scan*	$G \in [2, \geq 40]$
$N_{0.02}$	2648	*outside scan*	*outside scan*	$G \in [2, \geq 40]$

Corollary 1. At $p = \tau = \frac{1}{2}$ with G even, by symmetry $\mathbb{P}(X > G/2) = \mathbb{P}(X < G/2) = (1 - P_0)/2$ with $P_0 = \binom{G}{G/2} 2^{-G}$, so $F = \mathbb{P}(X \geq G/2) = (1 - P_0)/2 + P_0 = \frac{1}{2} + P_0/2$ and $C = \frac{1}{2} + P_0^2/2$. Stirling gives $P_0 \sim \sqrt{2/(\pi G)}$, hence $C \approx \frac{1}{2} + 1/(\pi G)$. For odd G there is no central atom, $F = \frac{1}{2}$ and $C = \frac{1}{2}$ exactly. We verified the closed form against exact enumeration: $C(8)$=0.53738, $C(16)$=0.51928, $C(32)$=0.50979, $C(64)$=0.50493.

Theorem 1. $\mathbb{P}(X \in \{0, G\}) = (1-p)^G + p^G$ for $X \sim \mathrm{Bin}(G, p)$; the extrema follow from $\frac{d}{dp}[p^G + (1-p)^G] = G[p^{G-1} - (1-p)^{G-1}]$, zero at $p = \frac{1}{2}$ (minimum 2^{1-G}) and $\to 1$ at the endpoints.

Lemma 1. For fixed (w, μ) with $w > 0$, multiplying an inequality by w does not reverse it, hence $T(\hat{p}) \geq \tau \iff \hat{p} \geq w^{-1}(\tau - (1-w)\mu) =: \tau'$, including equality under the paper's "$\geq$" tie rule. If $\hat{p} = X/G$ and $0 < \tau' \leq 1$, then $m = \lceil G\tau' \rceil \in \{1, \ldots, G\}$ and $F_{\tau'}(p) = \Pr[\mathrm{Bin}(G, p) \geq m]$ is continuous and strictly increasing from 0 to 1. Therefore $C_{\tau'}(p) = 2(F_{\tau'}(p) - \frac{1}{2})^2 + \frac{1}{2}$ attains exactly $\frac{1}{2}$ at the unique p for which $F_{\tau'}(p) = \frac{1}{2}$. For $\tau' \leq 0$ the rule always accepts, while for $\tau' > 1$ it always rejects, so in either case $C \equiv 1$. At $w = 0$, $T \equiv \mu$ and the same constant-decision conclusion applies. □

Our same-sample diagnostic uses $w(x) = v/[v + x(1-x)/G]$ (apart from its endpoint clip), not a bucket-constant w. For fixed $v > 0$ and $\mu \in [0, 1]$,

$$T'(x) = \frac{v\{v + [x^2 + \mu(1-2x)]/G\}}{[v + x(1-x)/G]^2} = \frac{v\{v + [(x-\mu)^2 + \mu(1-\mu)]/G\}}{[v + x(1-x)/G]^2} > 0,$$

and the clipped endpoint pieces are affine and increasing. This conditional monotonicity still does not cover parameters and rank buckets estimated from the same noisy vector.

An exact implementation-realizable counterexample rules out universal data-dependent non-improvement. Take G=8, τ=1/2, K=10, and 20 prompts with latent probabilities $(1/8, 0, \ldots, 0)$. A positive focal count is rank-bucketed with one deterministic zero. At X=4, $(\hat{p}, \mu_b, \sigma_b^2) = (1/2, 1/4, 1/16)$, the within variance is $1/32$, $w = 2/3$, and $T(\hat{p}) = 5/12 < 1/2$; at X=5, $T(\hat{p}) = 115/208 > 1/2$. Hence raw accepts for $X \geq 4$ and shrinkage for $X \geq 5$. With $X \sim \mathrm{Bin}(8, 1/8)$ their acceptance probabilities are $188707/16777216$ and $20637/16777216$, respectively, and shrinkage raises the focal two-replication consistency by $21754392555/1099511627776 \approx 0.019786$. Exact arithmetic and the diagnostic floating-point implementation agree on all nine counts.

B GROUPED HELD-OUT SHRINKAGE CONTROL

Before running this control, we fixed a deterministic five-fold split by underlying prompt: all three models and five hint levels for a DeepMath `orig_index` share one fold. For each replicated measurement pair under fixed randomization, raw and same-sample decisions follow the G=8, K=10, τ=0.5 path. The cross-fit estimator fits rank-bucket means and population variances on the other four folds. Because $\hat{p}$ has only nine values, ties are unavoidable; held-out values use the pre-specified empirical mid-rank bucket. Empty means default to 0.5 and empty/singleton variances to 10^{-4}, though no such fitted bucket occurred in the full run.

Table 4: Exact paired-decision consistency for raw, same-sample diagnostic shrinkage, and grouped cross-fit shrinkage. Synthetic denominators are 160,000; frozen real is 240,000. Cross-fit improves on same-sample in every row and attains the highest observed consistency on the tolerant-extractor frozen real distribution; differences from raw are reported for every control. Entries are finite-count consistency diagnostics.

difficulty	raw	same-sample	cross-fit	cross-fit − raw
uniform	.819219	.809475	.813981	-.005238
Beta$(2, 2)$	.751963	.737763	.739319	-.012644
concentrated	.581556	.574813	.581556	.000000
bimodal	.840138	.831625	.837350	-.002788
frozen real, tolerant	.946925	.952525	.954263	+.007338

The real fold contains 400 prompt groups and 6,000 model–prompt–hint cells; no group crosses folds, and missing, invalid, and duplicate counts are zero. Two complete runs are byte-identical. The author-side audit records exact action changes, acceptance counts, bucket assignments, source and input hashes, and the pre-specified protocol. MSE is reported for synthetic settings with known latent truth; the frozen-real evaluation uses exact-action consistency.

C Format competence and selector behavior

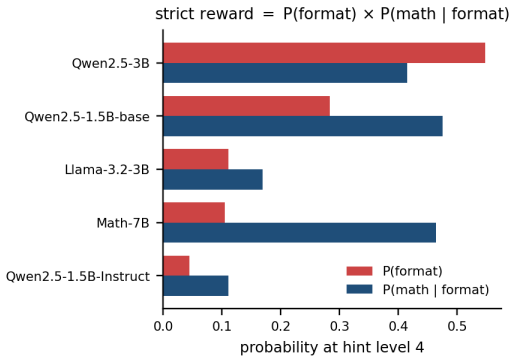

Figure 6: Frozen-policy strict-reward decomposition. Bars show $P(\text{format})$ and $P(\text{math} \mid \text{format})$ at hint level 4; their product is the expected strict-positive rate. Tolerant-extractor sensitivity is reported separately in Appendix I.

D Extended related work

This section expands the comparison in Section 9 and delineates the contribution relative to each neighboring literature.

Per-prompt adaptive RLVR. SEELE (Li et al., 2025), ZPPO (Lee et al., 2026), DAPO (Yu et al., 2025), POPO (Xu & Fang, 2026) and GP-POPO (Mao et al., 2026) threshold a small-group accuracy estimate into a per-prompt action (Figure 1); we found no action-repeatability estimate in these five traced specifications. UFT (Liu et al., 2025a) and Prefix-RFT (Huang et al., 2025) scaffold with hints but do not threshold per prompt. The five controllers above use small-group success statistics to choose prompt actions. GRPO instead uses rewards within a group to construct relative advantages for optimization (Shao et al., 2024); R1-style recipes and analyses (DeepSeek-AI et al., 2025; Liu et al., 2025b; Yue et al., 2025) build on that value-model-free update. This is distinct from prompt selection and from PPO-based post-training (Schulman et al., 2017; Ouyang et al., 2022).

Curriculum learning. Ordering or adapting training items by difficulty is long-standing (Bengio et al., 2009); reinforcement-learning curricula often use learning progress (Graves et al., 2017). See

Portelas et al. (2020) and Narvekar et al. (2020) for surveys. That literature asks *which item to show next*. We ask a prior question—whether the per-item measurement the choice is made from can support the choice—and our answer directly covers schemes that threshold a small-sample estimate per item. Curricula driven by aggregate or long-run statistics form a complementary design class.

Shrinkage and allocation. Shrinking a noisy per-item estimate toward a population mean is the classical empirical-Bayes remedy (Robbins, 1956; Efron & Morris, 1975), and EBPO (Han et al., 2026) brings it to RLVR. Its target is the *gradient baseline*, an estimation problem, where shrinkage is the right tool. Lemma 1 concerns the other target—the *decision*—and shows that a fixed affine map exactly reparameterises its threshold. Our executed same-sample map is richer and population-dependent, so Table 4 evaluates its decision behavior directly. Budget allocation is the complementary lever: HORA (Wang et al., 2026) allocates for coverage and CoBA-RL (Yao et al., 2026) by training value, both in the tradition of prioritized replay (Schaul et al., 2016). Neither allocates for *decision reliability*, which is what Table 3 prices.

Verifier noise and reward hacking. Outcome and process supervision can improve mathematical reasoning (Cobbe et al., 2021; Lightman et al., 2024), but an imperfect reward can mis-specify or be exploited (Amodei et al., 2016; Skalse et al., 2022), including through reward-model over-optimization (Gao et al., 2023). Spurious-reward RLVR results further show why a training gain alone need not validate reward semantics (Shao et al., 2025). Section 6.3 contributes one measured instance in the RLVR setting: the shipped verifier assigns strongly format-sensitive rewards. Since every rule in Section 5 consumes a binary verifier, this sensitivity changes the quantity controlled by the curriculum. The resulting claim preserves the exact verifier semantics consumed by each audited curriculum.

IRT in machine learning. Our framing draws on adaptive testing (Lord, 1980; van der Linden & Glas, 2000). Item response theory has been imported into ML mainly to analyse *evaluation*: to compare test sets (Vania et al., 2021) and to shrink benchmarks while preserving their measurement properties (Maia Polo et al., 2024). SEELE imports the item-response *curve* into training. We import the other half of the same theory—the reliability of the resulting score—and apply it to training decisions, which is where the small-sample regime bites hardest and where, unlike evaluation, no reliability figure is customarily reported.

E MODELS

Table 5: Models used as public, frozen checkpoints in the frozen-policy analyses. Section 6.3's approximately $27\times$ base-to-instruction-tuned strict-score contrast is the last two rows—same lineage and parameter count, differing only in instruction tuning.

checkpoint	params	role
`Qwen/Qwen2.5-Math-7B`	7B	main sweeps
`Qwen/Qwen2.5-3B`	3B	main sweeps
`meta-llama/Llama-3.2-3B`	3B	main sweeps
`Qwen/Qwen2.5-1.5B`	1.5B	selector sweep + format decomposition
`Qwen/Qwen2.5-1.5B-Instruct`	1.5B	format decomposition (instruct arm)

F AUDIT PROVENANCE, ROW BY ROW

Every row of Table 1 is traced to a primary paper or released implementation. Rule semantics are verified directly; where a displayed group size is reconstructed or an ablation arm is selected, the table and the notes below say so explicitly.

- SEELE (Li et al., 2025): $G{=}8$ and $\tau{=}0.5$ read from the released launch script; the selector itself is the upstream `hint_manager` we execute unchanged (Appendix K).
- ZPPO (Lee et al., 2026): in version 3, $G{=}8$ appears in Table 11, while the $\bar{r}{<}0.5$ admission rule and its standard-deviation rationale appear in the accompanying text.
- DAPO (Yu et al., 2025): group size 16 and the $X \in \{0, G\}$ discard rule stated in the paper text.

- POPO (Xu & Fang, 2026): $G{=}8$ is stated in the training hyperparameters, and the one-tailed $X{=}0$ discard rule is specified in Algorithm 1.
- GP-POPO (Mao et al., 2026): the two-tailed rule verified from its method section; G is ablated over $\{4, 8, 16, 32\}$ rather than fixed, and we quote the $G{=}8$ arm.

G RULE-SPECIFIC CALIBRATION IDENTITY

Section 4 reports the disjoint-prompt calibration for the consistency family. The six rule-specific points retain their executed same-source calibration, with an exact conditional identity at fixed p: $\Pr(A \text{ all-same}) = d_G(p)$ and $\Pr(A \text{ all-same}, A \cup B \text{ mixed}) = d_G(p) - d_{2G}(p)$; integrating and conditioning gives Equation (4). The selector family instead forward-resamples through the upstream rule.

H TRAINING PROTOCOL AND RESOURCE ACCOUNTING

The upstream and confirmatory studies provide complementary comparator and resource views; Section 7 summarizes their connection. This appendix records the protocol and aggregate resource accounting for the completion-token-matched server study in Section 7.

H.1 QUESTION, MODEL, AND ARMS

The primary question is whether released SEELE differs from the development-selected $(4 \to 4)$ static schedule under a matched completion-token budget, with GRPO as a no-curriculum reference. The study uses Qwen2.5-3B revision `3aab1f19...08227a9b`, five paired training initializations, and a group-disjoint subset of the released DeepMath training set. The model revision, prompt groups, schedule family, selection rule, whole-batch stopping rule, and evaluator were byte-identified before execution.

All three confirmatory arms have executable semantics: GRPO uses no prompt-level curriculum; the selected $(4 \to 4)$ static arm applies the chosen monotone global schedule with an explicit per-prompt feasibility clamp; and the released-selector arm executes SEELE unchanged. Static+EBPO, static+HORA, and RGC define modular analytic extensions alongside the three protocol-frozen arms.

H.2 SPLITS, TUNING, BUDGET, AND EVALUATION

The released DeepMath file contains 21,908 rows. Exact normalized-question groups, including known duplicates, stay within one split: 200 development groups occupy 201 source rows and 100 probe groups occupy 100 rows. Removing all 301 rows leaves the identical 21,607-row training pool for every arm. The frozen audit found no normalized overlap between that source and the six test assets. The development split uses the lowest source-row index as the representative of its one duplicated group; neither the probe split nor any benchmark test item participates in tuning. The normalization rule, stable group hash, development and probe membership, candidate schedule family, selection metric, and lexicographic tie rule are frozen in the study manifests. Every run binds the model, run identifier, arm, checkpoint, dataset, grader, and configuration bytes. The completed non-confirmatory pilot established correctness, runtime, memory, and token accounting; benchmark evaluation was reserved for the confirmatory runs.

The static comparator is selected from all 15 monotone pairs (a, b) with $0 \le a \le b \le 4$. It uses level a through committed update 250 and level b thereafter, with the effective per-prompt level clamped to the number of available hint segments minus one. Each candidate is trained independently for 500 updates from the same frozen development initialization and ranked by greedy strict accuracy on the development groups; ties choose the lexicographically smallest pair. The selected schedule is then retrained from the base checkpoint for every confirmatory run initialization. No test benchmark participates in this choice.

Table 6: Protocol-frozen experiment design. Panel A distinguishes the three executed Qwen-only arms; a check indicates a prompt-level action. Panel B records the frozen study constants. These are design parameters, not results.

A. Arm decomposition

arm	hints	per-item action	estimator	allocation	purpose
GRPO	none	—	group-relative	uniform	no-curriculum reference
selected static $(4 \to 4)$	fixed	—	group-relative	uniform	adaptation control
released SEELE	adaptive	✓	group-relative	uniform	audited method

B. Frozen study constants

component	setting	component	setting
data	21,607 training rows; 200 development groups; 100 disjoint probe groups	paired runs	three arms $\times$ five paired initializations
training	batch 256; scheduled checkpoints 0/100/250/500; terminal checkpoint selected by budget	prompts	maximum 2,048 tokens; overlength prompts filtered
rollouts	maximum 2,048 tokens; temperature 1; top-p 1; 32 responses per prompt	selector	target τ=0.5; four rounds; G=8 per round
optimization	AdamW, 10^{-6} learning rate, 0.01 weight decay; one PPO epoch; clip 0.2	regularization	KL coefficient 0.001 (low-variance KL); SFT coefficient 0.001; entropy coefficient 0
static tuning	15 monotone level pairs in $\{0, \ldots, 4\}^2$; switch at update 250; fixed development initialization	selection	500 updates/path; 200-group strict-accuracy development set; lexicographic tie break
token calibration	separate fixed initialization; 50 updates/arm; no performance evaluation	terminal rule	1.024B completion-token cap; last complete whole-batch checkpoint; no partial update
evaluation	greedy, one sample, 2,048-token maximum; no chat template; strict exact match	inference sets	MATH500 500; AMC23 40; AIME24 30; OlympiadBench 675; GSM8K 1,319; Minerva 272
statistics	training run is the independent unit; mean, sample SD, two-sided 95% t_4 interval	primary effect	paired SEELE–static MATH500 delta, interval, and Cohen's d_z

The primary resource axis is cumulative committed completion tokens. Calibration uses a separate frozen initialization for 50 updates per arm and performs no accuracy evaluation. If $r_{\min}$ is the smallest of the three measured completion-token rates, the cap is fixed to $B = 10^6 \lfloor 400 r_{\min} / 10^6 \rfloor = 1.024 \times 10^9$ before confirmatory evaluation. Committed tokens include every sampled response in a batch that produces a complete optimizer update, even when the released all-correct/all-wrong filter later removes a response. Training generates and grades a whole candidate batch; if committing it would exceed B, the pipeline rejects the entire batch before the actor update and atomically selects the preceding complete checkpoint. It never truncates a response or group. The rejected batch and every failed attempt remain charged as physical cost. A run that reaches 500 committed updates without triggering the cap is invalid under this protocol. Generated responses, committed and physical completion tokens, retained actor examples, optimizer updates, GPU-hours, wall time, and peak memory are recorded separately; evaluation generations use a separate ledger.

The confirmatory endpoint is greedy, single-sample MATH500 exact-match accuracy for Qwen2.5-3B at the terminal matched-token checkpoint. Evaluation uses the Qwen base boxed prompt, no chat template, temperature zero, top-p one, one sample, and at most 2,048 generated tokens. The exact test-set denominators are 500 (MATH500), 40 (AMC23), 30 (AIME24), 675 (OlympiadBench), 1,319 (GSM8K), and 272 (Minerva Math). The latter five are pre-specified secondary evaluations. The frozen evaluation suite comprises these six exact, hash-verified assets, with no post-hoc benchmark substitution. All checkpoints are selected by budget, never by test performance. Token-fraction interpolation is descriptive only.

H.3 MEASURED SERVER RESOURCE USE

Table 7 reports observed server telemetry for the confirmatory runs. Shared static selection and token calibration are accounted for separately from the multi-run summaries.

H.4 ESTIMANDS AND STATISTICAL RULES

Training initializations are paired across arms. The primary contrast is released SEELE minus the selected $(4 \to 4)$ static schedule on Qwen2.5-3B, with GRPO reported as a reference. Benchmark outcomes are reported as pre-specified multirun means and sample standard deviations. Physical costs use the same aggregate convention, and failed train/evaluation attempts enter the arm-level attempt totals. No best-run or best-checkpoint selection is permitted. Before any confirmatory result, we fixed an untruncated two-sided Student-t interval over the five paired run-level values (four degrees of freedom). The primary contrast additionally reports its raw paired percentage-point effect and paired Cohen's d_z=0.40. This interval estimates the conditional paired effect for the frozen split and development-selected static schedule. Prompt-level observations are not treated as independent training runs, following established cautions about run variation, power, and split-induced instability (Colas et al., 2018; Card et al., 2020; Gorman & Bedrick, 2019). The pre-specified secondary benchmarks provide descriptive generalization checks.

The manuscript reports the pre-specified mean, sample standard deviation, and paired interval across all confirmatory runs; no individual run is selected or excluded.

I THE TOLERANT EXTRACTOR AND ITS BRANCH-EXPOSURE AUDIT

The **strict** grader is SEELE's shipped reward: extract the final `<answer>...</answer>` line and compare to ground truth after normalization. The **tolerant** grader additionally accepts (i) a `\boxed{}` expression and (ii) the last numeric expression in the response, under the same normalization. Tolerant is *our construct*; strict is what the released reward optimizes. The frozen audit file samples 50 tolerant-positive responses and 30 strict-wrong/tolerant-right responses. In the disputed stratum, $25/30$ carry an explicit `\boxed{}` answer and $5/30$ use no boxed form, exposing the loosest final-numeric branch. The audit's estimand is branch exposure: it records extractor and format fields, and uses $5/30$ exclusively as a branch-exposure statistic rather than an error-rate estimate.

At hint level 4, 60.7% of the base checkpoint's tolerant credit is disputed against 98.5% for the instruction-tuned one. To parameterize the full sensitivity range, let f denote the false-positive fraction among disputed responses and propagate it through $a_m(f) = t_m(1 - f d_m)$; the table reports $a_{\text{inst}}(f)/a_{\text{base}}(f)$. At f=0 tolerant is taken at face value; at f=1 we have identically $a_m(1) = t_m(1 - d_m) = s_m$—tolerant degenerates to strict—so the instruction-tuned-to-base extractor-score ratio there is *exactly* the strict-reward ratio $1/27$.

f	instruct/base score ratio	reading
0	$0.99\times$	tolerant taken at face value
$5/30$	$0.92\times$	illustrative stress point equal to the loose-branch share
1	$1/27$	every disputed response treated as false positive

The endpoints are identifiable and the sensitivity analysis leaves f parametric between them; the $0.92\times$ row is an illustrative sensitivity point. Accordingly, Section 6.3 establishes strict-grader sensitivity to answer form as its targeted estimand, while the strict-selector 400/400 headline remains independent of tolerant extraction.

J ODD/EVEN AND ROUNDING TECHNICAL NOTE

Corollary 1 is a statement at $p = \tau$ with τG integral. For $\tau = \frac{1}{2}$: even G has a central atom giving $C = \frac{1}{2} + P_0^2/2 > \frac{1}{2}$; odd G has none, giving $C = \frac{1}{2}$ exactly. Hence the sequence $\{C(\delta_\tau, G)\}_G$ is not monotone—it is strictly decreasing within even G and constant on the floor for odd G. The convention "$>$" vs "$\geq$" at the threshold leaves C unchanged. The global minimizer of $C(p)$ is

Table 7: Observed real-server resource accounting. Panel A reports the five confirmatory runs per arm (mean ± sample SD); completion-token, response, and retained actor-example counts are in millions, and T/E denotes train/evaluation attempts. Panel B keeps shared static selection and token calibration outside those run means.

A. Confirmatory runs only

arm	committed M	physical M	updates	responses M	retained M
GRPO	1021.8 ± 0.7	1024.7 ± 0.6	400.6 ± 1.1	3.28 ± 0.01	2.13 ± 0.01
selected static $(4 \to 4)$	1021.5 ± 0.7	1024.7 ± 0.5	388.6 ± 1.1	3.18 ± 0.01	1.46 ± 0.01
released SEELE	1021.5 ± 0.6	1024.6 ± 0.5	384.4 ± 1.1	3.15 ± 0.01	2.02 ± 0.01

arm	train GPU-h	eval GPU-h	wall h	peak GB	attempts T/E
GRPO	782.6 ± 11.6	5.84 ± 0.11	99.5 ± 1.5	73.0 ± 0.2	5/30
selected static $(4 \to 4)$	818.8 ± 11.5	6.14 ± 0.11	103.9 ± 1.4	74.5 ± 0.2	5/30
released SEELE	843.2 ± 14.3	6.24 ± 0.11	107.2 ± 1.8	75.4 ± 0.3	5/30

B. Shared pre-confirmatory overhead

item	runs × updates	tokens	GPU-h	accounting scope
static selection	15×500	19.73 B	15.8k	shared comparator selection; accounted separately from run means
token calibration	3×50	0.393 B	0.31k	shared cap selection; no performance evaluation

$p^* \neq \tau$ in general (G=8: $p^* \approx 0.44$), with $p^* \to \tau$ as G grows. These atoms are a measure-zero technicality for the distribution-level law in Proposition 1.

K UPSTREAM PARITY

Every selector-action number in Section 6 uses the upstream `hint_manager` code, executed with `ray`/`verl` stubbed (`ray.remote`→synchronous, `ray.get`→ identity), so the real `fit_model`, `regularize_y` and `avg_y_for_same_x` run byte-for-byte. An independent reimplementation used as a cross-check omitted `regularize_y` and agreed with upstream on 98.6% of decisions; all $49/3600$ disagreements were single-level ($|\Delta L|$=1) at floor/ceil snapping boundaries and none changed the 400/400 · 400/400 · 356/400 headline. The strict-grader branch counts under the upstream code are $\{\text{inverse} : 400\}$ for all three models (no early exit fires), confirming that every prompt executes the fit-and-inverse path; the resulting maximum-level counts remain model-dependent as reported in Section 6.

L RELEASED-CODE-FAITHFUL OFFLINE REPLAY

We freeze the policy and reuse the stored, server-generated 16-outcome cells; the replay requires no additional rollout generation. Each logical prompt encounter imports the released `SeeleHintManager`, starts a new in-memory manager, and executes four sequential decisions. Each decision draws G=8 current outcomes without replacement from the corresponding stored cell. The fitter receives the current round observations and, only for a sparse fit, at most one aggregate point from the preceding dataset write-back. At the fourth round the four current aggregates replace the preceding prompt state. Stored cells may be revisited in later rounds or logical passes. Released-lifecycle replay is the primary analysis, and a cumulative-state variant provides a lifecycle-sensitivity comparison.

Table 8: Released-lifecycle frozen-policy replay after ten logical passes (n=400 prompts per model). Entropy H is in bits; "max" is the number selecting hint level 4.

Model	strict max	strict H	tolerant max	tolerant H
Llama-3.2-3B	$400/400$	0.000	$397/400$	0.064
Math-7B	$400/400$	0.000	$319/400$	0.955
Qwen2.5-3B	$395/400$	0.097	$330/400$	0.790

Across $48{,}000$ comparable batch-end decisions, the released-lifecycle and cumulative-state variants agree on $45{,}493$ (94.777%); the $2{,}507$ differences show that lifecycle semantics are material. Update-to-update exact-action consistency over passes 2–10 is respectively $1.000/1.000/0.998$ under strict and $0.998/0.969/0.977$ under tolerant for Llama/Math/Qwen. The strict-grader's strong concentration therefore remains visible throughout the frozen replay, with small residual action diversity for Qwen. After three strict passes, switching to tolerant grading produces the same recorded batch-end trajectory as a clean tolerant replay. These controlled fixed-policy observations isolate lifecycle semantics and connect directly to the downstream training evidence in Appendix H.

M EXPERIMENTAL COVERAGE AND REPRODUCIBILITY

EMIT combines source-faithful replay across multiple models with five paired, completion-token-matched Qwen2.5-3B training runs, linking mechanism-level diagnosis to downstream utility. Source-traced decisions, semantic perturbations, disjoint-prompt calibration, and multidimensional resource accounting make the complete evidence chain auditable (Table 7).

End-to-end execution contract. A reproducible EMIT evaluation makes the controller, measurement, and downstream comparison separately auditable:

1. **Freeze the instrument:** record the verifier, rollout group size, estimate-to-action map, tie convention, randomization, and every state carried between decisions.
2. **Measure exact actions:** re-execute the frozen rule on independent rollout groups and report agreement, flips, or conditional reveals at the same decision unit used by training.
3. **Calibrate the sampling account:** estimate the difficulty distribution on a disjoint prompt subset and compare its finite-group prediction with measured action behavior.
4. **Trace implementation and semantics:** connect each reported rule to its primary source and released path, then vary only audited verifier semantics to isolate their action-level effect.
5. **Convert evidence into control:** apply Equation (5) to act, pool, or request additional rollouts according to a pre-specified consistency target.
6. **Test incremental utility:** compare the resulting controller with a tuned pooled policy under a common resource axis, paired runs, and a frozen checkpoint-selection rule.

The reported frozen-policy analyses use stored real-server rollout sweeps containing 400 prompts $\times$ 5 hint levels $\times$ 16 outcomes per model, graded under both extractors. Analyses of these fixed outcomes require no new model generation. The completion-token-matched training study followed protocol `EMIT-MINIMUM-PRIMARY-V1` (SHA-256 `d8aac78c...1ef56142`); its manifests bind the model, data, evaluator, configuration, and benchmark hashes. The reported aggregate covers every paired confirmatory run and benchmark. Upon acceptance, the authors will release the algorithmic implementation of EMIT.