# LET INTUITION TEACH DELIBERATION: REWARDING PREFIX SUCCESS POTENTIAL IN REASONING MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Two reasoning trajectories can reach the same answer while differing in whether a correct short answer was already available halfway through. We use this distinction to let verified short-answer probes supervise long-form deliberation. Prefix Success Potential (PSP) measures the probability of producing a correct short answer from a prefix under a fixed protocol. The same policy samples auxiliary answers, a verifier scores them, and their success rates reward the complete original trajectory. Our variant $\mathrm{PSP}_{\mathrm{ME}}$ (ME: midpoint and endpoint) favors higher midpoint potential at equal endpoint potential. Probes serve only for scoring; no external teacher or process annotations are required. We train with an 8,192-token budget and evaluate at 32,768/38,912 tokens. Across Qwen3-1.7B, Qwen3-8B, Gemma4-E4B-it, and Gemma4-12B-it, $\mathrm{PSP}_{\mathrm{ME}}$ achieves Math-7 scores of 51.06%, 68.70%, 65.56%, and 79.93% respectively. These represent gains of 3.89, 4.45, 5.55, and 6.56 percentage points over matched outcome-reward GRPO. Fixed-prefix evaluation reveals two key findings. First, RL training dissociates answer potential from natural completion: Gemma4-12B-it improves potential by 14.32 pp while completion drops by approximately 29 pp, suggesting that PSP teaches models to maintain extractable reasoning states rather than forcing earlier generation. Second, gains transfer heterogeneously through different mechanisms: Gemma models improve short-answer potential while Qwen3-8B does not, yet both achieve substantial long-budget gains, demonstrating that midpoint supervision enhances reasoning quality through multiple pathways rather than a single uniform mechanism.

## 1 INTRODUCTION

Consider two reasoning trajectories that eventually give the same correct answer. Halfway through the first, the model can already produce a correct short answer when asked; halfway through the second, it cannot. A final-answer reward treats these trajectories alike, although their intermediate states differ in answer availability. We investigate whether this difference provides useful supervision for long-form reasoning.

We approach this question through verified short-answer probes: the same model produces answers from existing reasoning prefixes within a short budget, and a task verifier checks each answer. We formalize this probe as *Prefix Success Potential* (PSP): the probability of passing the verifier under a fixed short-answer protocol. After sampling the original trajectories, we estimate PSP at selected prefixes and average the estimates into one reward for each complete response (Figure 1). Our main configuration, $\mathrm{PSP}_{\mathrm{ME}}$, uses two probes of at most 256 tokens at each of the midpoint and endpoint. We apply policy gradients only to original responses. The method requires no external teacher, process annotations, or learned value network.

**Probes versus continuations.** VinePPO evaluates intermediate states by sampling continuations to estimate future success (Kazemnejad et al., 2025), but its auxiliary generation grows with the number of evaluated states, samples per state, and remaining continuation lengths. PSP instead asks whether the model can answer *now*, within 256 tokens, and averages these readouts to reward the original trajectory. This probe-based approach changes both the measured quantity and credit assignment.



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.

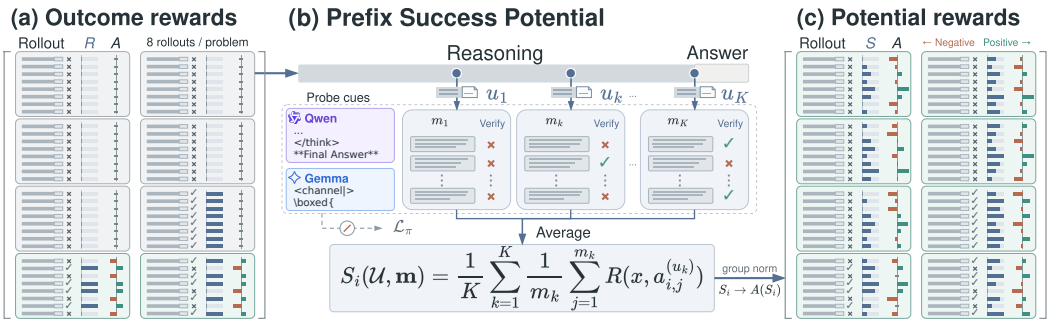


Figure 1: **Verified short answers supervise complete reasoning trajectories.** (a) Each problem has eight original rollouts; binary rewards give zero task advantages in tied groups. (b) The same policy samples short answers from selected prefixes using model-family cues. Verified success rates are averaged within and then across positions to give $S_i$; gold answers enter only verification. (c) GRPO updates the complete original responses. The crossed connection excludes cues and probe tokens from the policy loss. Prefix rules are model-specific (Appendix B.2.1); the recovered cards illustrate the observed predominance of all-wrong groups.

At equal endpoint potential, $\mathrm{PSP}_{\mathrm{ME}}$ favors higher midpoint potential. The method can distinguish trajectories with identical binary outcomes, including tied groups where group relative policy optimization (GRPO; Shao et al., 2024) has zero task advantages.

We evaluate Qwen3-1.7B/8B and Gemma4-E4B-it/12B-it with an 8,192-token training cap and 32,768/38,912-token mathematical test caps. $\mathrm{PSP}_{\mathrm{ME}}$ improves Math-7 over matched outcome-reward GRPO by 3.89, 4.45, 5.55, and 6.56 percentage points, respectively. We report initial-model benchmarks and pre-RL comparisons in Appendix C.1.

We also measure natural-answer correctness and short-answer potential on shared prefixes, revealing that these dimensions are dissociable. Both Gemma models improve 8K potential over GRPO despite fewer naturally correct completions; Qwen shows different patterns.

Our contributions are: (1) PSP rewards that use verified short-answer probes to supervise complete reasoning trajectories; (2) empirical evidence that RL training dissociates answer potential from natural completion: PSP improves short-answer potential by 14.32 pp while natural completion drops by approximately 29 pp; (3) heterogeneous patterns across model families; and (4) ablations of re-answering, probe count, and prefix position, together with measurements of transfer, costs, and tradeoffs across four models.

## 2 Related Work

**Outcome rewards and additional sampling.** GRPO compares responses within a question group (Shao et al., 2024). DAPO replaces tied groups with additional complete responses (Yu et al., 2025); learnability-based selection and Prompt Replay alter question exposure (Foster et al., 2025; Baroian & Berger, 2026). PSP preserves the question sampler and rescores original trajectories using short-answer probes.

**Continuation values and process feedback.** Process supervision and verified continuations provide intermediate feedback (Lightman et al., 2024; Wang et al., 2024). VinePPO estimates continuation values for step-level credit (Kazemnejad et al., 2025). Rewarding Progress uses changes in future success (Setlur et al., 2025), PPPO updates retained prefixes (Sun et al., 2026), and PACR and GRPO-VPS use changes in correct-answer confidence (Yoon et al., 2025; Wang et al., 2026a). SPAE also probes intermediate answers for potential but combines confidence-based measures with gold-answer token probabilities and shapes step-level advantages using potential differences and saturation penalties (Wu et al., 2026). PSP uses verified success rates rather than confidence scores, averages absolute potential values rather than differences, and assigns one trajectory reward rather than step-level advantages.

**Prefix probes and answer extractability.** Answer-convergence analysis probes truncated reasoning for stable answers (Liu & Wang, 2025); TRACE aggregates verifier rewards across forced-answer cutoffs (Wang et al., 2026b). The detection–extraction gap identifies that success under free continuations can differ from forced answer extraction (Wang & Zhu, 2026). Our work differs in two ways: we use probe success as RL training reward rather than only for analysis, and we measure how potential and completion change after training.

**Budgets, fast answers, and early exits.** e3 studies short-budget RL and task/budget curricula (Setlur et al., 2026), while L1 uses length-sensitive rewards (Aggarwal & Welleck, 2025). PSP operates at a fixed 8K training cap, focusing on trajectory supervision. Budget forcing and S-GRPO control or reward early exits (Muennighoff et al., 2025; Dai et al., 2025), as does JET (Han et al., 2026); DEER uses trial-answer confidence for inference-time stopping (Yang et al., 2026). PSP instead rewards trajectory quality without a stopping objective.

**Teacher feedback.** GKD and OPD provide teacher feedback on student sequences (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025). PSP uses verified outcomes from the same model rather than teacher distribution matching. We include a separately configured OPD reference (Appendix C.7).

## 3 PRELIMINARIES

For a problem $x \sim \mathcal{D}$, a frozen rollout policy $\pi_{\text{old}}$ independently samples $n \geq 2$ original responses $y_i = c_i \oplus a_i$, where $c_i$ is reasoning and $a_i$ the final-answer segment. A verifier returns $R_i = R(x, a_i) \in \{0, 1\}$.

Given trajectory scores $r = (r_1, \ldots, r_n)$, GRPO (Shao et al., 2024) forms normalized advantages

$$A_i(r) = \frac{r_i - \bar{r}}{\sigma_r + \epsilon}, \qquad \bar{r} = \frac{1}{n}\sum_i r_i, \qquad \sigma_r^2 = \frac{1}{n-1}\sum_i (r_i - \bar{r})^2, \tag{1}$$

with $\epsilon > 0$, and maximizes

$$J(\theta; r) = \mathbb{E}_{x, y_{1:n}}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T_i}\sum_{t=1}^{T_i}\min\{\rho_{i,t}A_i(r),\ \text{clip}(\rho_{i,t}, 1-\delta, 1+\delta)A_i(r)\} - \beta\widehat{d}_{i,t}\right], \tag{2}$$

where $T_i = |y_i|$, $\rho_{i,t} = \pi_\theta(y_{i,t} \mid x, y_{i,<t})/\pi_{\text{old}}(y_{i,t} \mid x, y_{i,<t})$, and $\widehat{d}_{i,t}$ is the sampled KL divergence to a fixed reference. Equal scores give $A_i = 0$ for all $i$. We replace the trajectory scores $r$ in Eq. equation 1 while retaining this clipped objective and KL regularization.

## 4 PREFIX SUCCESS POTENTIAL REWARDS

### 4.1 SUCCESS POTENTIAL UNDER A SHORT-ANSWER PROTOCOL

For a state $s = (x, c_{\leq t})$, fix a protocol $\mathcal{P}$ specifying the answer cue, sampling parameters, and short output budget. The same rollout snapshot induces a distribution $\pi_{\text{old}}^{\mathcal{P}}$ and the Prefix Success Potential (PSP)

$$V_{\mathcal{P}}^{\pi_{\text{old}}}(s) = \mathbb{E}_{a \sim \pi_{\text{old}}^{\mathcal{P}}(\cdot|s)}[R(x, a)]. \tag{3}$$

This is the probability of answering correctly under protocol $\mathcal{P}$, not a step-correctness score or the value of unrestricted continuation. Probe contexts contain the problem and a model-specific contiguous prefix. Qwen excludes content after the first reasoning close. Gemma4-E4B-it retains continuous history up to the last reasoning close, or the full available history when reasoning reopens. These boundary rules define protocol $\mathcal{P}$. We use gold answers only for verification. Appendix B.3 examines a small offline search for answer cues.

At a fixed state, $m$ independent probes estimate the value by

$$\widehat{V}_{\mathcal{P},m}(s) = \frac{1}{m}\sum_{j=1}^{m} R(x, a_j), \qquad \mathbb{E}[\widehat{V}_{\mathcal{P},m}(s) \mid s] = q, \quad \text{Var}(\widehat{V}_{\mathcal{P},m}(s) \mid s) = \frac{q(1-q)}{m}, \tag{4}$$

where $q = V_{\mathcal{P}}^{\pi_{\text{old}}}(s)$. More independent probes reduce conditional estimation variance for the same target.

### 4.2 FROM PREFIX VALUES TO A TRAJECTORY REWARD

Choose positions $\mathcal{U} = \{u_1, \ldots, u_K\}$ with $0 \le u_1 < \cdots < u_K \le 1$, and probe counts $\mathbf{m} = (m_1, \ldots, m_K)$. For reasoning length $L_i$, define

$$t_{i,k} = \lfloor u_k L_i \rfloor, \qquad s_i^{(u_k)} = (x, c_{i,\le t_{i,k}}), \qquad \Phi_i^{(u_k)} = \frac{1}{m_k} \sum_{j=1}^{m_k} R(x, a_{i,j}^{(u_k)}). \tag{5}$$

Positions use each trajectory's actual prefix-window length, not its generation cap. The endpoint follows the model-specific boundary rule above. For Gemma, $L_i$ counts the continuous history window. If no closing marker occurs, the available generated prefix determines $L_i$.

We average position-wise estimates to obtain

$$S_i(\mathcal{U}, \mathbf{m}) = \frac{1}{K} \sum_{k=1}^{K} \Phi_i^{(u_k)} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{m_k} \sum_{j=1}^{m_k} R(x, a_{i,j}^{(u_k)}). \tag{6}$$

Positions remain equally weighted even if their probe counts differ. The score is in $[0, 1]$. Conditional on the original trajectory, the score estimates $\mu_i(\mathcal{U}) = K^{-1} \sum_k V_{\mathcal{P}}^{\pi_{\text{old}}}(s_i^{(u_k)})$. Increasing $m_k$ changes estimation precision; changing $\mathcal{U}$ also changes the target. The score excludes the original reward $R_i$ and uses absolute values rather than differences between adjacent positions.

We denote endpoint probing by $\text{PSP}_{\text{E}}$ (E: endpoint) and midpoint-and-endpoint probing by $\text{PSP}_{\text{ME}}$, with $m = 2$ probes per position by default. Our main instance, $\text{PSP}_{\text{ME}}$, uses $\mathcal{U} = \{0.5, 1.0\}$ and $\mathbf{m} = (2, 2)$, yielding $S_i = (\Phi_i^{(0.5)} + \Phi_i^{(1.0)})/2$. $\text{PSP}_{\text{E}}$ ($m$ probes) uses only $u = 1$ with $m$ probes. Within each model, all positions share the same protocol: Qwen uses an eight-token answer cue and Gemma a six-token native-close/boxed cue, followed by at most 256 generated tokens at temperature 1 and top-$p$ 1.

**Absolute potential versus step advantages.** VinePPO uses Monte Carlo continuation values for step-level credit assignment (Kazemnejad et al., 2025). PSP instead averages absolute short-answer values into one trajectory reward. Both the measured value and credit assignment differ; differencing our probes would not reproduce VinePPO.

**What the midpoint adds.** Write $q_{i,u} = V_{\mathcal{P}}^{\pi_{\text{old}}}(s_i^{(u)})$. $\text{PSP}_{\text{ME}}$ estimates $\mu_i = (q_{i,0.5} + q_{i,1.0})/2$. For two trajectories with equal endpoint potential,

$$\mu_i - \mu_j = \tfrac{1}{2}(q_{i,0.5} - q_{j,0.5}). \tag{7}$$

Higher midpoint potential therefore gives higher expected reward at equal endpoint potential. The midpoint is relative to each trajectory's length, so this preference does not imply answer availability at a shorter absolute budget. Section 5.3 tests availability and natural completion separately.

### 4.3 UPDATING THE ORIGINAL TRAJECTORIES

We substitute $r = S$ into Eqs. equation 1 and equation 2. One scalar advantage weights all original reasoning and answer tokens. We exclude probe cues and outputs from the policy loss. Original responses and probes share a frozen rollout snapshot; scoring uses stopped gradients, with no segment-specific advantage. This provides an alternative trajectory reward, not an unbiased replacement for the complete policy gradient of original-answer accuracy (Appendix A.4).

### 4.4 ESTIMATION AND REWARD DISCRIMINATION

Conditional on fixed trajectories and independent probes, $S_i$ is unbiased for $\mu_i$, with mean-squared error at most $1/(4N_{\text{eff}})$, where $N_{\text{eff}} = K^2/(\sum_k m_k^{-1})$. For a target gap $\Delta_{ij} > 0$, the probability of reversing or tying the ordering is at most $\exp(-N_{\text{eff}}\Delta_{ij}^2)$ (Proposition 1, Appendix A). More probes improve precision, but these bounds can be loose with four probes.

Group advantages are nonzero exactly when scores are not all tied. On shared trajectories, the gain in nonzero-advantage groups equals recovered original ties minus newly introduced ties (Proposition 2). Sampling noise can also break ties, so reward discrimination alone does not establish useful learning; Section 5.2 tests both discrimination and final performance.

Table 1: Matched GRPO and PSP comparisons across four models, with a separate Qwen3-1.7B OPD reference (percent). GRPO and PSP train with an 8,192-token cap; OPD uses 4,096. Mathematical testing allows 32,768 tokens for GSM8K/MATH500 and 38,912 for competition mathematics.

| | Mathematical reasoning | | | | | | | | Other capabilities | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Math-7 | GSM8K | MATH500 | AIME24 | AIME25 | AIME26 | HMMT25 | HMMT26 | GPQA$^\diamond$ | IFEval | BFCL | HumanEval |
| *Qwen3-1.7B* | | | | | | | | | | | | |
| GRPO | 47.17 | 89.69 | 90.80 | 41.82 | 30.83 | 30.73 | 20.83 | 25.47 | 36.77 | 68.76 | 94.00 | **81.71** |
| $\mathrm{PSP_E}$ | 49.80 | 89.31 | 89.80 | 45.73 | 34.17 | 36.98 | **24.38** | 28.22 | 38.38 | **71.90** | 92.00 | 79.88 |
| $\mathrm{PSP_{ME}}$ | **51.06** | **90.90** | **92.00** | **49.17** | **35.47** | **37.50** | 23.96 | **28.41** | **39.29** | 69.69 | **95.00** | **81.71** |
| OPD-4K | 51.98 | 90.14 | 91.20 | 49.48 | 36.88 | 40.52 | 25.73 | 29.92 | 39.60 | 70.06 | 93.00 | 81.10 |
| *Qwen3-8B* | | | | | | | | | | | | |
| GRPO | 64.24 | 95.75 | 95.20 | 70.52 | 53.75 | 60.62 | 34.17 | 39.68 | 58.79 | 85.21 | 93.00 | **91.46** |
| $\mathrm{PSP_E}$ | **68.89** | **95.91** | **96.20** | 73.54 | **65.99** | **64.58** | 42.08 | **43.94** | **61.46** | 84.84 | 96.00 | 86.59 |
| $\mathrm{PSP_{ME}}$ | 68.70 | 95.60 | 96.00 | **75.36** | 65.21 | 64.06 | **42.50** | 42.14 | 61.41 | **85.95** | **97.00** | 90.85 |
| *Gemma4-E4B-it* | | | | | | | | | | | | |
| GRPO | 60.01 | **93.40** | 94.80 | 60.62 | 45.42 | 51.88 | 37.08 | 36.84 | 54.80 | **87.99** | 90.00 | **85.37** |
| $\mathrm{PSP_E}$ | 63.17 | 92.42 | **95.00** | 63.59 | 47.71 | **62.71** | 40.31 | 40.44 | 57.58 | **87.99** | 88.00 | 84.15 |
| $\mathrm{PSP_{ME}}$ | **65.56** | 92.65 | 93.00 | **65.26** | **54.48** | 61.98 | **45.31** | **46.21** | **59.39** | 86.69 | **93.00** | **85.37** |
| *Gemma4-12B-it* | | | | | | | | | | | | |
| GRPO | 73.37 | 96.74 | 97.40 | 77.29 | 63.23 | 68.85 | 55.94 | 54.17 | 66.57 | **93.35** | 96.00 | 96.34 |
| $\mathrm{PSP_E}$ | 78.91 | **96.82** | **98.60** | **82.45** | 72.92 | 75.94 | 65.62 | 60.04 | **70.96** | 92.79 | 96.00 | 96.34 |
| $\mathrm{PSP_{ME}}$ | **79.93** | 96.06 | 98.40 | 82.24 | **73.91** | **78.33** | **67.92** | **62.69** | 66.57 | 90.57 | **98.00** | **98.17** |

Bold marks maxima among GRPO, $\mathrm{PSP_E}$, and $\mathrm{PSP_{ME}}$ within each model, including ties; blue highlights $\mathrm{PSP_{ME}}$. The separated OPD-4K row uses a Qwen3-8B teacher and a different training recipe, is excluded from bold comparisons, and does not match training resources (Appendix C.7). Initial-model scores are in Appendix C.1, with official-report checks in Appendix B.8. Math-7 weights seven datasets equally. HumanEval uses a 16,384-token output cap shared by thinking and final code. IFEval uses strict prompt-level accuracy; BFCL is a local 100-example subset. Full likelihood-based benchmarks appear in Appendix C.3. Sampling is fixed within each model; differences use unrounded scores.

# 5 Experiments

## 5.1 Transfer beyond the training rollout budget

**Shared setup.** We train Qwen3-1.7B, Qwen3-8B, Gemma4-E4B-it, and Gemma4-12B-it on DAPO-Math-17k. Unless stated otherwise, training uses an 8,192-token output cap, $n = 8$ responses per problem, 32 problem groups per step, and 200 steps. Benchmark results use the final checkpoints (training seed 42). All training runs start from shared Initial checkpoints with Math-7 scores of 50.26% (Qwen3-1.7B), 69.57% (Qwen3-8B), 56.39% (Gemma4-E4B-it), and 81.83% (Gemma4-12B-it); full Initial results appear in Appendix C.1. These defaults apply throughout the paper and appendix. Within each model, data exposure, response count, optimizer settings and intended execution settings are matched; probes add computation. Gemma4-12B-it $\mathrm{PSP_E}$ requires checkpoint recovery and a runtime repair (Appendix D.2). Appendix B details model-dependent protocols and systems.

Table 1 shows that $\mathrm{PSP_{ME}}$ consistently outperforms GRPO across all four models on Math-7, the equal-weight mean over seven mathematical datasets. Test caps are 32,768 tokens on GSM8K/MATH500 and 38,912 on competition mathematics; repeated answers estimate single-response correctness without voting. These caps exceed our RL budget, not an established limit of the starting models' prior training. Sampling and evaluation are fixed within each model (Appendix B).

**Long-budget mathematical performance.** $\mathrm{PSP_{ME}}$ reaches 51.06%, 68.70%, 65.56%, and 79.93% Math-7 on the four models, improving over GRPO by 3.89, 4.45, 5.55, and 6.56 pp. Additional midpoint benefit is model-dependent: $\mathrm{PSP_{ME}}$ exceeds $\mathrm{PSP_E}$ by 1.26 pp on Qwen3-1.7B and 2.39 pp (95% CI $[1.19, 3.61]$) on Gemma4-E4B-it. The larger models show smaller or statistically insignificant midpoint benefits: Qwen3-8B shows $-0.20$ pp (95% CI $[-1.04, 0.65]$) and Gemma4-12B-it shows $+1.02$ pp (95% CI $[-0.05, 2.11]$), with both CIs including zero. Both larger models perform below their initial checkpoints after RL training (Appendix C.1), so PSP's gains represent partial mitigation of GRPO's performance degradation rather than absolute improvements. Full comparisons and intervals appear in Appendix C.

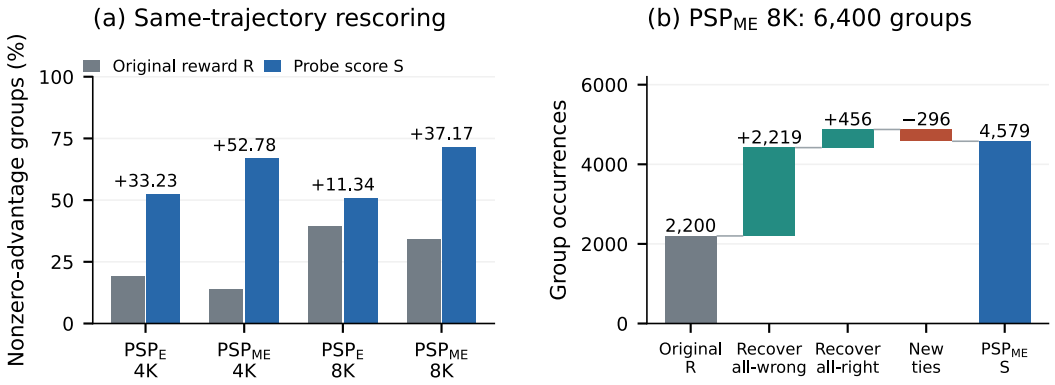


Figure 2: **Reward discrimination and recovery.** These diagnostics rescore fixed Qwen3-1.7B trajectories. (a) Original rewards $R$ and actual scores $S$ rescore the same trajectories within each run; annotations give the increase in percentage points. Here $R$ verifies original responses generated within the indicated 4K/8K training output cap. Runs have different trajectory distributions. (b) $\mathrm{PSP}_{\mathrm{ME}}$ 8K has $2{,}200 + 2{,}219 + 456 - 296 = 4{,}579$ groups with nonzero task advantages. Each run contains 6,400 group occurrences; questions may recur. The statistic excludes KL and does not guarantee useful gradients. Counts aggregate the group occurrences observed during training.

**Other capabilities and teacher reference.** Mathematics-only RL yields task-dependent transfer: $\mathrm{PSP}_{\mathrm{ME}}$ improves GPQA$^{\diamondsuit}$ over GRPO on three of four models and achieves the highest BFCL scores on all four models, but both Gemma models lose IFEval instruction-following accuracy. HumanEval and likelihood-based results also vary by task (Appendices C.2 and C.3). The separate OPD-4K reference reaches 51.98% Math-7 on Qwen3-1.7B, versus 51.06% for $\mathrm{PSP}_{\mathrm{ME}}$, using a frozen Qwen3-8B teacher and a different recipe; it is not a matched-resource comparison (Appendix C.7).

### 5.2 Why verified short-answer rewards help

Table 2: Qwen3-1.7B reward-construction ablations under the shared setup and long-cap mathematical test protocol.

| Reward | Positions $\mathcal{U}$ | Probes/response | Math-7 |
|---|---|---|---|
| GRPO | Original answer | 0 | 47.17 |
| $\mathrm{PSP}_{\mathrm{E}}$ | $\{1.0\}$ | 1 | 49.52 |
| $\mathrm{PSP}_{\mathrm{E}}$ | $\{1.0\}$ | 2 | 49.80 |
| $\mathrm{PSP}_{\mathrm{E}}$ | $\{1.0\}$ | 4 | 48.95 |
| $\mathrm{PSP}_{\mathrm{ME}}$ | $\{0.5, 1.0\}$ | $2 + 2$ | 51.06 |

**Re-answering and position allocation.** $\mathrm{PSP}_{\mathrm{E}}$ ($m = 1$) improves on GRPO by 2.36 pp (95% CI $[1.23, 3.50]$), exceeding binary reward alphabet effects. $\mathrm{PSP}_{\mathrm{E}}$ ($m = 4$) scores 0.57 pp below $\mathrm{PSP}_{\mathrm{E}}$ ($m = 1$) (95% CI $[-1.69, 0.52]$), indicating non-monotonic performance with probe count.

$\mathrm{PSP}_{\mathrm{ME}}$ and $\mathrm{PSP}_{\mathrm{E}}$ ($m = 4$) use four probes and the same five score levels. $\mathrm{PSP}_{\mathrm{ME}}$ gains 2.11 pp (95% CI $[1.20, 3.03]$), supporting position allocation beyond probe count on Qwen3-1.7B. Equal probe counts do not match actual computation or establish step-level credit. Other models compare two endpoint probes with four distributed probes, so they do not isolate position allocation.

**Reward discrimination on shared trajectories.** We define a trainable group narrowly as one containing at least one nonzero *task* advantage. For each run, we retain original rewards and actual training scores on its own trajectories. Figure 2 shows that on Qwen3-1.7B $\mathrm{PSP}_{\mathrm{ME}}$ 8K trajectories, the fraction of trainable groups rises from 34.38% under $R$ to 71.55% under $S$, a gain of 37.17 pp. All four audited Qwen3-1.7B probe configurations increase same-trajectory discrimination. This is not a counterfactual GRPO training run: trajectories rescored with $R$ were still generated by the probe-trained policy.

Table 3: Natural completion and endpoint potential at an 8K prefix budget (percent).

| *Qwen3-1.7B* | | | | | *Qwen3-8B* | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Math-7 | | Clean | | Method | Math-7 | | Clean | |
| | $R$ | $\widehat{\phi}_{100}$ | $R$ | $\widehat{\phi}_{100}$ | | $R$ | $\widehat{\phi}_{100}$ | $R$ | $\widehat{\phi}_{100}$ |
| Initial | 33.11 | 42.89 | 25.29 | 46.73 | Initial | 37.22 | 54.82 | 36.72 | 71.46 |
| GRPO | 39.48 | 41.57 | 42.48 | 48.41 | GRPO | 54.28 | 58.17 | 64.94 | 72.29 |
| $\mathrm{PSP_E}$ | 32.11 | 38.23 | 30.47 | 47.51 | $\mathrm{PSP_E}$ | 42.79 | 57.18 | 45.02 | 72.09 |
| $\mathrm{PSP_{ME}}$ | 27.83 | 42.49 | 20.02 | 48.29 | $\mathrm{PSP_{ME}}$ | 32.93 | 56.95 | 32.32 | 72.14 |
| *Gemma4-E4B-it* | | | | | *Gemma4-12B-it* | | | | |
| Method | Math-7 | | Clean | | Method | Math-7 | | Clean | |
| | $R$ | $\widehat{\phi}_{100}$ | $R$ | $\widehat{\phi}_{100}$ | | $R$ | $\widehat{\phi}_{100}$ | $R$ | $\widehat{\phi}_{100}$ |
| Initial | 52.97 | 54.43 | 58.69 | 57.10 | Initial | 36.79 | 58.26 | 32.32 | 61.99 |
| GRPO | 55.61 | 57.00 | 61.82 | 60.72 | GRPO | 70.93 | 49.38 | 73.54 | 55.96 |
| $\mathrm{PSP_E}$ | 53.51 | 57.95 | 59.77 | 62.04 | $\mathrm{PSP_E}$ | 50.72 | 63.16 | 53.12 | 69.31 |
| $\mathrm{PSP_{ME}}$ | 30.74 | 59.33 | 37.70 | 63.99 | $\mathrm{PSP_{ME}}$ | 41.94 | 63.69 | 42.68 | 66.50 |

Within each model, $R$ and $\widehat{\phi}_{100}$ use the same original trajectories; Clean is a separate 256-question set. Potential averages four probes, each with up to 256 extra tokens. Math-7 uses MathEqual for $R$ and DeepScaler for probes; Clean uses DeepScaler for both. All cells are measured. See Appendix B.7 for provenance and boundary rules.

Gemma4-E4B-it shows a different pattern: for $\mathrm{PSP_{ME}}$, discrimination rises from 42.73% to 67.98%, but $\mathrm{PSP_E}$ *decreases* discrimination from 44.75% to 40.73% while still improving long-cap mathematics (Table 1). This demonstrates that increased reward discrimination is not necessary for performance gains, suggesting multiple mechanisms may contribute to PSP's effectiveness.

### 5.3 Natural completion and fixed-prefix answer potential

Natural completion and answer potential measure different behavior. We score the original output visible within $B$ tokens as $R(B)$, and measure endpoint potential $\widehat{\phi}_{100}(B)$ with four forced short answers after model-specific prefix processing. Table 3 compares both on the same original trajectories. Math-7 uses different verifiers for the two metrics (Appendix B.7).

Table 3 reveals a striking dissociation between answer potential and natural completion for Gemma models. At 8K, $\mathrm{PSP_{ME}}$ improves potential over GRPO: by 2.33 pp (95% CI $[1.26, 3.45]$) on Math-7 and 3.27 pp (95% CI $[0.61, 5.96]$) on Clean for Gemma4-E4B-it, and by 14.32 pp (95% CI $[12.36, 16.31]$) on Math-7 and 10.55 pp (95% CI $[7.30, 13.87]$) on Clean for Gemma4-12B-it. However, Gemma4-12B-it's natural completion drops substantially: from 70.93% (GRPO) to 41.94% ($\mathrm{PSP_{ME}}$) on Math-7, and from 73.54% to 42.68% on Clean. This pattern—higher extractable answer potential despite lower natural completion rates—suggests that $\mathrm{PSP_{ME}}$ training teaches the model to maintain accessible reasoning states throughout the trajectory without necessarily generating natural final answers within the 8K budget.

$\mathrm{PSP_{ME}}$ is not uniformly strongest across all comparisons: on Gemma4-12B-it, Clean potential trails $\mathrm{PSP_E}$ by 2.81 pp (95% CI $[-5.08, -0.59]$), while the 0.53 pp Math-7 advantage (95% CI $[-0.52, 1.58]$) is not statistically significant.

Qwen models show a contrasting pattern, revealing that PSP's benefits arise through different mechanisms across model families. Neither Qwen3-1.7B nor Qwen3-8B improves short-answer potential over GRPO on Clean; Qwen3-8B $\mathrm{PSP_{ME}}$ trails GRPO in Math-7 potential by 1.61 pp at 2K (95% CI $[-2.52, -0.79]$), 1.66 pp at 4K (95% CI $[-2.62, -0.78]$), and 1.22 pp at 8K (95% CI $[-2.40, -0.05]$). In contrast, Gemma4-12B-it gains 9.99 pp at 2K, 7.95 pp at 4K, and 14.32 pp at 8K.

This model-family divergence is significant: Gemma models improve short-answer extractability (higher $\widehat{\phi}_{100}$) while Qwen3-8B does not, yet both families achieve substantial long-budget performance gains over GRPO (Table 1, +4.45 pp for Qwen3-8B, +6.56 pp for Gemma4-12B-it). The dissociation demonstrates that midpoint supervision can enhance reasoning quality through multiple pathways—earlier answer availability for Gemma versus other trajectory improvements for Qwen—rather than a single uniform mechanism.

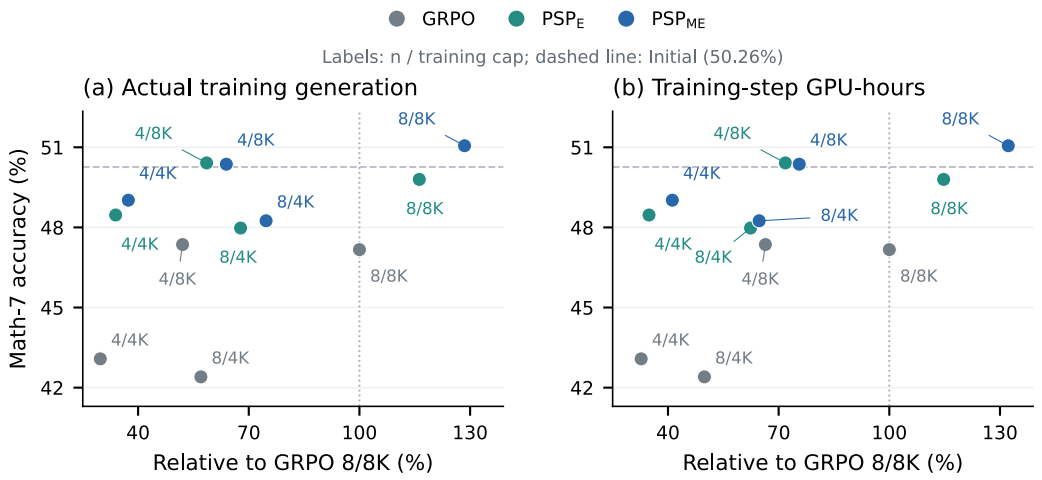


Figure 3: **Qwen3-1.7B performance under reduced training budgets.** Training uses four H100 GPUs. Labels give responses/problem $n$ and training output cap; color denotes reward, and each point represents one configuration. Generation includes all probe outputs. Each cost axis is normalized to GRPO $n = 8$, 8K (320.804M generated tokens; 33.163 training-step GPU-hours). Scores and costs refer to the same final checkpoints. Timers exclude startup and work outside training steps. All scores use the shared long-cap Math-7 protocol. Points compare observed configurations, not matched-compute learning curves.

Appendix C.5 provides complete model-wise comparisons and intervals. These intervals reflect question-level variability only; training-seed variability and multiple-comparison adjustments are not included.

Appendix C.6 tracks Clean performance over training (Figure 6). For Gemma4-E4B-it, training midpoint success rises from 12.77% to 60.76% while original trajectories lengthen and natural correctness falls (Appendix D.1). Because trajectories and relative positions change, this does not establish earlier insight at fixed token budgets. Appendix F illustrates how midpoint probes distinguish two unfinished responses with tied endpoint scores; its broader token diagnostic does not establish selective reinforcement of local reasoning patterns.

### 5.4 Training cost and test-generation length

Under the shared Qwen3-1.7B setup, $\mathrm{PSP}_{\mathrm{ME}}$ adds 28.50% actual training generation and 32.24% timed training-step duration relative to GRPO. The nominal probe reservation, $4 \times 256/8192 = 12.5\%$, does not capture changes in original length or prefix processing. Figure 3 shows that $\mathrm{PSP}_{\mathrm{ME}}$ achieves substantially higher Math-7 performance (51.06% versus 47.17% for GRPO) while adding moderate training cost. VinePPO and full DAPO are valid comparators, but we have not evaluated them at matched cumulative compute. Such a comparison must match models, data, and evaluation, and count auxiliary continuations, discarded samples, and prefix processing within the total training budget (Appendix D).

Reducing responses per problem to four retains 50.37% Math-7 for $\mathrm{PSP}_{\mathrm{ME}}$, with 63.92% of the reference's actual training generation and about 76% of its step time. At $n = 4$, 4K, $\mathrm{PSP}_{\mathrm{ME}}$ reaches 49.02%, using 37.36% of generation and 41.18% of timed GPU-hours. Table 19 and Appendix E provide complete configurations and accounting. Test-time generation increases substantially: $\mathrm{PSP}_{\mathrm{ME}}$ responses average 79% to 174% more tokens than GRPO across the four models (Table 17).

## 6 Discussion and Conclusion

PSP uses verified short-answer probes to supervise complete reasoning trajectories. Re-answering and position-allocation ablations support this reward design; tie-recovery analysis explains one effect on GRPO feedback. Across four models, gains over GRPO transfer from 8K training to

longer mathematical test budgets. Initial-model comparisons (Appendix C.1) and non-mathematical tradeoffs delimit these gains.

**Baseline, cost, and robustness.** Training starts from Initial checkpoints (Math-7: 50.26%, 69.57%, 56.39%, 81.83%). Three GRPO configurations underperform Initial; two $\mathrm{PSP}_{\mathrm{ME}}$ models remain below Initial despite gains over GRPO, indicating suboptimal GRPO setup. PSP consistently improves over GRPO under matched conditions (mean +3.89 pp), with Gemma4-E4B-it achieving +9.17 pp net gain over Initial. However, $\mathrm{PSP}_{\mathrm{ME}}$ generates 1.79–2.74$\times$ longer responses than GRPO (Table 17), increasing inference cost; under restricted 8K budgets, $\mathrm{PSP}_{\mathrm{ME}}$ underperforms GRPO across all models (Table 3). Single training seeds limit variance quantification, though consistent gains across four models from two independent families provide complementary robustness evidence; formal replicates remain necessary for statistical significance.

**Dissociable dimensions and heterogeneous mechanisms.** Fixed-prefix evaluation reveals that answer potential and natural completion are dissociable. For Gemma4-12B-it, $\mathrm{PSP}_{\mathrm{ME}}$ improves 8K potential by 14.32 pp over GRPO while natural completion drops by approximately 29 pp (Table 3). This suggests that midpoint supervision teaches models to maintain extractable knowledge throughout reasoning rather than forcing earlier answer generation. PSP's benefits arise through different mechanisms across model families: Gemma models improve short-answer extractability while Qwen3-8B does not, yet both achieve substantial long-budget gains. Jointly studying answer extraction and stopping is a natural next step.

## 7 LIMITATIONS

PSP depends on the cue, sampler, budget, and verifier; it measures protocol-specific answer availability, which can reflect extraction or adaptation as well as reasoning progress. Short probes measure extraction success under a fixed protocol, not reasoning completeness; answer extraction may succeed or fail independently of free continuation. The method requires verifiable outcomes and is untested on open-ended tasks. Finite probes add noise; our theory establishes neither complete-gradient unbiasedness nor improved convergence (Appendix A.4).

We compare methods under the same maximum output cap, not under matched actual generation cost. PSP models generate 1.79–2.74$\times$ longer responses than GRPO (Table 17); under restricted 8K budgets, $\mathrm{PSP}_{\mathrm{ME}}$ underperforms GRPO across all models (Table 3). Accuracy-versus-cost curves remain untested. Combining PSP with early-exit or length-control methods remains untested.

Results use one training seed per configuration (seed 42); question-bootstrap intervals capture evaluation variance but not training-run variation. Multi-seed replicates are necessary for statistical significance. Equal-count position ablations and teacher comparisons cover only Qwen3-1.7B. No matched-compute VinePPO/full-DAPO comparison or complete natural-answer budget sweep is available. A complete train–test overlap audit and full-project resource accounting remain outstanding (Appendix B.10).

### AI USE STATEMENT

We used generative AI tools to assist with refining the methodological and theoretical formulation, drafting and checking mathematical arguments, designing diagnostic experiments, implementing and debugging experimental and analysis code, and interpreting results. We also used these tools for literature search, translation, manuscript drafting and editing, and preparing figures and tables. In the shared-trajectory token-signal diagnostic (Appendix F), two AI agents reviewed candidate text spans with outcomes, advantages, and method signals hidden; these were AI-assisted annotations, not human annotations. We have reviewed all AI-assisted work. Validation included numerical and consistency checks against archived experimental outputs, checks of figure and table values, and finite-enumeration checks for theoretical identities. The scope of the checks and remaining reproducibility limitations are documented in Appendices A and B.10. The authors take responsibility for the final text, claims, code, analyses, and other artifacts, including those prepared with AI assistance.

## REFERENCES


Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Pranjal Aggarwal and Sean Welleck. L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=4jdIxXBNve.

Andrei Baroian and Rutger Berger. Prompt replay: speeding up grpo with on-policy reuse of high-signal prompts, 2026. URL https://arxiv.org/abs/2603.21177.

Muzhi Dai, Chenxu Yang, and Qingyi Si. S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 48178–48204. Curran Associates, Inc., 2025. doi: 10.52202/085713-1608. NeurIPS proceedings.

Thomas Foster, Anya Sims, Johannes Forkel, and Jakob Foerster. LILO: Learning to Reason at the Frontier of Learnability. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 48941–48974. Curran Associates, Inc., 2025. doi: 10.52202/085713-1634. NeurIPS proceedings.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, December 2023. URL https://zenodo.org/records/10256836.

Gemma Team. Gemma 4 Technical Report, 2026. URL https://arxiv.org/abs/2607.02770v1. Version 1, Table 5.

Jinyi Han, Ying Huang, Ying Liao, Haiquan Zhao, Zishang Jiang, Xinyi Wang, Xikun Lu, Guanghao Zhou, Sihang Jiang, Jiaqing Liang, Weikang Zhou, Zeye Sun, Fei Yu, and Yanghua Xiao. Your Models Have Thought Enough: Training Large Reasoning Models to Stop Overthinking. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2u5ZRzDyS0.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. URL http://www.jstor.org/stable/2282952.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining Credit Assignment in RL Training of LLMs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29557–29590. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/kazemnejad25a.html.

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

Xin Liu and Lu Wang. Answer Convergence as a Signal for Early Stopping in Reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17896–17907,

Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.904. URL https://aclanthology.org/2025.emnlp-main.904/.

Kevin Lu and Thinking Machines Lab. On-Policy Distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. URL https://thinkingmachines.ai/blog/on-policy-distillation/. Technical blog.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Qwen Team. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388v1. Version 1, Table 19.

Ben Rank, Hardik Bhatnagar, Ameya Prabhu, Shira Eisenberg, Karina Nguyen, Matthias Bethge, and Maksym Andriushchenko. PostTrainBench: Can LLM agents automate LLM post-training? *arXiv preprint arXiv:2603.08640*, 2026. URL https://arxiv.org/abs/2603.08640.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=A6Y7AqlzLW.

Amrith Setlur, Matthew Yang, Charlie Snell, Jeremiah Greer, Ian Wu, Virginia Smith, Max Simchowitz, and Aviral Kumar. e3: Learning to Explore Enables Extrapolation of Test-Time Compute for LLMs. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 127323–127361, 2026. ICLR proceedings.

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, 2024. URL https://arxiv.org/abs/2402.03300.

Yiliu Sun, Zicheng Zhao, Yang Wei, Yanfang Zhang, and Chen Gong. Well Begun, Half Done: Reinforcement Learning with Prefix Optimization for LLM Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):33144–33152, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i39.40598. URL https://doi.org/10.1609/aaai.v40i39.40598.

Hanyang Wang and Mingxuan Zhu. The Detection-Extraction Gap: Models Know the Answer Before They Can Say It, 2026. URL https://arxiv.org/abs/2604.06613.

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 https://arxiv.org/abs/2604.20659.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Xinpeng Wang, Nitish Joshi, Barbara Plank, Rico Angell, and He He. Is it Thinking or Cheating? Detecting Implicit Reward Hacking by Measuring Reasoning Effort. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 96956–96981,

2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/9d0ee47d732bd526c4a03d4aecc39bc1-Paper-Conference.pdf.

Fei Wu, Zhenrong Zhang, Qikai Chang, Jianshu Zhang, Quan Liu, and Jun Du. Step Potential Advantage Estimation: Harnessing Intermediate Confidence and Correctness for Efficient Mathematical Reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6361–6379, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.317. URL https://aclanthology.org/2026.findings-acl.317/.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=NpU7ZXafRi.

Eunseop Yoon, Hee Suk Yoon, Jaehyun Jang, SooHwan Eom, Qi Dai, Chong Luo, Mark A. Hasegawa-Johnson, and Chang D. Yoo. PACR: Progressively Ascending Confidence Reward for LLM Reasoning, 2025. URL https://arxiv.org/abs/2510.22255.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, juncai liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113222–113244. Curran Associates, Inc., 2025. doi: 10.52202/085713-3775. NeurIPS proceedings.

# A THEORY AND PROOF DETAILS

## A.1 ESTIMATION AND RECOVERY STATEMENTS

Let $\mathcal{H} = (x, y_{1:n}, \pi_{\text{old}}, \mathcal{P}, \mathcal{U}, \mathbf{m})$ fix the conditions before scoring. Write $q_{i,k} = V_{\mathcal{P}}^{\pi_{\text{old}}}(s_i^{(u_k)})$ and $\mu_i = K^{-1}\sum_k q_{i,k}$.

**Proposition 1** (Estimation and ranking)**.** *Suppose probe verification outcomes are mutually independent conditional on $\mathcal{H}$. For $N_{\text{eff}} = K^2/(\sum_k m_k^{-1})$, the score in Eq. equation 6 satisfies*

$$\mathbb{E}[S_i \mid \mathcal{H}] = \mu_i, \qquad \mathbb{E}[(S_i - \mu_i)^2 \mid \mathcal{H}] = \frac{1}{K^2}\sum_k \frac{q_{i,k}(1-q_{i,k})}{m_k} \le \frac{1}{4N_{\text{eff}}}. \tag{8}$$

*If $\mu_i - \mu_j = \Delta_{ij} > 0$, then $\mathbb{P}(S_i \le S_j \mid \mathcal{H}) \le \exp(-N_{\text{eff}}\Delta_{ij}^2)$.*

The result follows from weighted Bernoulli variances and Hoeffding's inequality (Hoeffding, 1963); Appendix A.2 gives the proof. Equal counts give $N_{\text{eff}} = Km$. These bounds concern the protocol-defined target and may be loose with few probes; they do not guarantee correct ranking in an individual group.

**Proposition 2** (Exact recovery decomposition)**.** *Let $\mathcal{E}_r = \{r_1 = \cdots = r_n\}$ and $I_r = \mathbf{1}\{\exists i : A_i(r) \neq 0\}$. For original rewards $R$ and probe scores $S$ on the same trajectories,*

$$\mathbb{P}(I_S = 1) - \mathbb{P}(I_R = 1) = \mathbb{P}(\mathcal{E}_R \cap \mathcal{E}_S^c) - \mathbb{P}(\mathcal{E}_R^c \cap \mathcal{E}_S). \tag{9}$$

Because the advantage denominator is positive, $I_r = \mathbf{1}_{\mathcal{E}_r^c}$; subtracting these indicators proves the identity, including its finite-count version. No independence assumption is required. The gain is recovered ties minus new ties. Nonzero advantages can reflect sampling noise even when all $\mu_i$ are identical; neither proposition guarantees a beneficial update.

Define $Y_{i,k,j} = R(x, a_{i,j}^{(u_k)})$ and $w_{k,j} = 1/(Km_k)$, so $S_i = \sum_{k,j} w_{k,j}Y_{i,k,j}$ and $C_{\mathbf{m}} = \sum_{k,j} w_{k,j}^2 = N_{\text{eff}}^{-1}$. Conditional independence refers to these Bernoulli variables given $\mathcal{H}$; it is not unconditional independence. Original rewards are fixed by $\mathcal{H}$ and need not follow the probe distribution.

## A.2 PROOF OF PROPOSITION 1

Linearity of expectation gives $\mathbb{E}[S_i \mid \mathcal{H}] = \sum_{k,j} w_{k,j}q_{i,k} = \mu_i$. Independence removes cross-covariances, giving

$$v_i := \text{Var}(S_i \mid \mathcal{H}) = \sum_{k,j} w_{k,j}^2 q_{i,k}(1-q_{i,k}) = \frac{1}{K^2}\sum_k \frac{q_{i,k}(1-q_{i,k})}{m_k} \le \frac{C_{\mathbf{m}}}{4}. \tag{10}$$

Unbiasedness identifies this variance with the mean-squared error. Unbiasedness itself requires only the stated conditional means, whereas the variance formula requires zero covariances. With correlated probes, the general expression includes $2\sum_{(k,j)<(\ell,r)} w_{k,j}w_{\ell,r}\,\text{Cov}(Y_{i,k,j}, Y_{i,\ell,r} \mid \mathcal{H})$; the independent bound cannot simply be retained.

For independent bounded variables $Z_\ell \in [a_\ell, b_\ell]$, Hoeffding's lower-tail bound (Hoeffding, 1963) is

$$\mathbb{P}\left(\sum_\ell Z_\ell - \mathbb{E}\sum_\ell Z_\ell \le -t\right) \le \exp\left(-\frac{2t^2}{\sum_\ell (b_\ell - a_\ell)^2}\right), \quad t > 0. \tag{11}$$

Each summand of $S_i$ has range width $w_{k,j}$. Applying both tails gives

$$\mathbb{P}(|S_i - \mu_i| \ge t \mid \mathcal{H}) \le \min\{1, 2e^{-2N_{\text{eff}}t^2}\}. \tag{12}$$

For $S_i - S_j$, the two independent sets of signed summands have total squared range width $2C_{\mathbf{m}}$. Its conditional mean is $\Delta_{ij}$, so

$$\mathbb{P}(S_i \le S_j \mid \mathcal{H}) \le e^{-2\Delta_{ij}^2/(2C_{\mathbf{m}})} = e^{-N_{\text{eff}}\Delta_{ij}^2}. \tag{13}$$

A whole-group tie implies a tie for any selected pair, and therefore has the same upper bound whenever a pair has a positive target gap. This does not give a nontrivial guarantee when the gap is zero. For four equally weighted probes and $\Delta = 0.1$, the ranking bound is $e^{-0.04} \approx 0.961$; it does not certify reliable ranking with the main sampling budget.

By Cauchy–Schwarz, $K^2 \le (\sum_k m_k)(\sum_k m_k^{-1})$, hence $N_{\text{eff}} \le \sum_k m_k$, with equality precisely for equal counts. This is a worst-case estimator bound under equal position weights. It does not show that equal allocation is optimal for every set of success probabilities or that effective sample size measures GPU cost.

The recovery identity holds pointwise and hence also for finite counts. Figure 2 gives its empirical instance. Recovering a fraction of original tied groups differs from the net reduction after new ties; group activity also differs from the fraction of individually nonzero advantages. Numerical masks and actual advantages must be checked.

### A.3 Separating target variation from measurement noise

Under conditional independence, write $S_i = \mu_i + e_i$ with $\mathbb{E}[e_i \mid \mathcal{H}] = 0$ and $\mathbb{E}[e_i^2 \mid \mathcal{H}] = v_i$. For sample variances across trajectories,

$$\mathbb{E}[\sigma_S^2 \mid \mathcal{H}] = \sigma_\mu^2 + \frac{1}{n}\sum_i v_i. \tag{14}$$

To prove this, expand $\sum_i (S_i - \bar{S})^2$. The cross term has zero expectation, while $\mathbb{E}\sum_i (e_i - \bar{e})^2 = \sum_i v_i - n\,\mathrm{Var}(\bar{e}) = (1 - 1/n)\sum_i v_i$. Division by $n-1$ gives Eq. equation 14. Across random trajectories with other conditions fixed, total variance similarly gives $\mathrm{Var}(S_i \mid x) = \mathrm{Var}(\mu_i \mid x) + \mathbb{E}[v_i \mid x]$. Thus greater score variance or fewer exact ties can include measurement noise. These formulas concern scores before normalization; the random denominator prevents replacing $\mathbb{E}[A_i]$ by normalized target means.

### A.4 Counterexamples and optimization boundaries

**Averaging positions can eliminate distinctions.** The ideal pairs $(q_{0.5}, q_{1.0}) = (1,1),(0,1),(1,0),(0,0)$ give expected rewards $1, 1/2, 1/2, 0$. Take two trajectories with stage probabilities $(0,1)$ and $(1,0)$. Endpoint scores are deterministically 1 and 0, whereas equal-weight two-position scores are both $1/2$, for any number of probes. More positions therefore need not reduce ties.

**Fewer ties need not mean more useful signal.** If all $q_{i,k} = p \in (0,1)$, all target values are identical. Under this homogeneous model, adding independent trials can reduce exact tie probability: noise shrinks while the score grid becomes finer. At a fixed total probe count and equal weights, endpoint-only and distributed probes have the same binomial score distribution under this homogeneous model. It gives no intrinsic advantage to position diversity.

**Protocol changes and one-response groups.** If a short-answer protocol always fails while natural answers have nondegenerate binary rewards, the new score always ties even though original rewards can distinguish trajectories. For $n = 1$, the sample standard deviation is undefined and group centering gives a zero numerator regardless of any chosen numerical convention. Single-response experiments require another baseline or advantage estimator; the group-recovery analysis does not extend to them.

**Value unbiasedness is not complete-gradient unbiasedness.** For a prefix at the natural answer boundary, define $V^{\pi}_{\text{ans}}(x,c) = \mathbb{E}_{a\sim\pi(\cdot|x,c)} R(x,a)$. This natural-answer value need not equal the short-answer value under $\mathcal{P}$. Consider the ideal natural-answer objective for fixed $x$, without group normalization or clipping. Factor $\pi_\theta(c, a \mid x) = \pi_\theta(c \mid x)\pi_\theta(a \mid x, c)$ and assume regularity for exchanging derivatives and expectation. Then

$$\begin{aligned}\nabla_\theta J_x = {} & \mathbb{E}_c\left[V^{\pi_\theta}_{\text{ans}}(x,c)\nabla_\theta \log \pi_\theta(c \mid x)\right] \\ & + \mathbb{E}_{c,a}\left[R(x,a)\nabla_\theta \log \pi_\theta(a \mid x, c)\right].\end{aligned} \tag{15}$$

The first term can conditionally average over the answer because its score depends only on $c$. In contrast, weighting answer tokens by a stopped-gradient conditional mean gives

$$\mathbb{E}_{a|x,c}\left[V_{\text{ans}}^{\pi_\theta}(x,c)\nabla_\theta \log \pi_\theta(a \mid x,c)\right] = 0, \tag{16}$$

by normalization of the answer distribution. The original answer term is generally nonzero. For example, a fixed prefix and $a \sim \text{Bernoulli}(\theta)$ with reward $a$ give $\nabla_\theta J_x = 1$, whereas replacing the reward by the stopped-gradient mean $\theta$ gives zero expected score-weighted gradient. A conditionally independent auxiliary estimate has the same issue. This counterexample rules out complete-gradient equivalence; it does not assert that answer-token gradients vanish in our actual multi-position, normalized, clipped surrogate.

# B Experimental details and reproducibility

## B.1 Scope, data, and model identity

All runs start from the released post-trained checkpoints `Qwen/Qwen3-1.7B`, `Qwen/Qwen3-8B`, `google/gemma-4-E4B-it` and `google/gemma-4-12B-it`, called Initial. Shared defaults are in Section 5.1; Tables 4–5 specify training and model differences. Qwen3-1.7B additionally varies response count, cap and probe count. Steps 85/170 support diagnostics; final results use step 200, without test-based checkpoint selection.

Training uses a schema-only conversion of the archived DAPO-Math-17k snapshot, preserving source row order: 17,398 rows and 17,187 distinct prompt identifiers. Data are shuffled with zero loader workers, a 2,048-token prompt limit, long-prompt filtering and errors for remaining overlength inputs. Training-time validation is disabled. The 6,400 core group occurrences can repeat questions. A complete exact/semantic train–test overlap audit remains outstanding; the Clean name does not establish decontamination. Gemma4-12B-it has neither E4B's per-layer embeddings nor shared cross-layer KV, so E4B-specific missing-parameter exemptions do not apply. Its endpoint run's recovery is recorded in Appendix D.2.

## B.2 Prompts, reasoning boundaries, and verification

Training and mathematical evaluation use one user message with the following text, preserving the two newlines:

```
{question}

Please reason step by step, and put your final answer within \boxed{}.
```

The frozen native chat template supplies the generation prompt. Mathematical evaluation enables native thinking. Qwen training uses its template default; Gemma training additionally appends a prompt-only thought opening, specified below, which is omitted at evaluation. No teacher solution or gold answer is appended.

The exact Qwen answer marker, using `\n` to denote a newline, is

```
"\n...\n</think>\n\n**Final Answer**\n"
```

It encodes to eight tokens under the frozen tokenizer. Probe input concatenates the already-templated prompt tokens, the reasoning-prefix tokens, and this marker, without wrapping them in another chat message. Up to 256 additional tokens are generated. The first `</think>` marks the reasoning endpoint; that marker and the original answer after it are excluded. If it is absent, all available original-response tokens are treated as visible reasoning, including capped responses. For length $L_c$, $\text{PSP}_{\text{ME}}$ uses cuts $\lfloor L_c/2 \rfloor$ and $L_c$, without sentence or step alignment. Empty prefixes are permitted.

Original responses and probes use the same parameter snapshot, temperature 1, top-$p$ 1, and disabled top-$k$ truncation. The implementation checks policy versions. Asynchronous inference does not guarantee identical trajectories across trained methods or bitwise deterministic replay.

Training uses the frozen DeepScaler-style verifier: extract the answer after the thinking or response boundary, normalize the boxed answer, and apply mathd or SymPy equivalence checks. Missing

scoreable segments or boxed answers receive zero. Probe verification concatenates the marker and probe text before scoring. Verifier exceptions are recorded and fail the run rather than silently counting as successful execution. Original rewards are retained for diagnostics but not added to probe scores. Probe tokens and markers never enter the policy loss.

#### B.2.1 GEMMA TRAINING PREFIX AND VERIFICATION PROTOCOL

Gemma uses the same mathematical user instruction but appends the artificial opening `<|channel>thought\n` to the prompt (three tokens: 100, 45518, 107). It is excluded from PPO and KL action losses; all actually generated original-response tokens remain trainable. The scoring procedure preserves repeated or reopened thought channels rather than dropping malformed-format responses.

Probe inputs retain a contiguous generated history. The endpoint is immediately before the last native thought close, unless a complete thought opening follows that close, in which case the entire available history is used; absent a close, the entire history is also used. An initial complete thought header is excluded from the length used for the 50% cut but retained in context. Earlier self-generated answer segments are not removed. Consequently, this window measures re-answering potential of continuous history, not exclusively pure reasoning content. The exact appended probe cue is

```
"\n<channel|>\n\boxed{"
```

It contains six tokens (107, 101, 107, 236785, 57296, 236782). $\mathrm{PSP_E}$ samples two 256-token continuations at the endpoint; $\mathrm{PSP_{ME}}$ samples two each at midpoint and endpoint. Both use temperature 1, top-$p = 1$, and disabled top-$k$, matching the original training rollouts. Artificial cues and probe outputs are excluded from the policy loss.

For original-response scoring, the last-answer parser takes the content after the final `<channel|>` or `</think>` marker, or the entire output if neither exists, then extracts its last boxed answer. A malformed final box does not fall back to an earlier answer. Reopened thoughts and missing boundaries are recorded without dropping the response. Probe scoring applies the same rule to the appended cue concatenated with the generated continuation. Gold answers are never inserted into either context. This differs from Qwen's first-close boundary rule; the same formula for $S$ does not imply identical measurement protocols.

### B.3 PROBE SELECTION ACROSS MODEL FAMILIES

**Offline selection protocol.** We compare four cues on frozen Qwen3-1.7B, Qwen3-8B, Gemma4-12B-it, and Gemma4-E4B-it. Each uses the same 32 development and 32 disjoint confirmation questions from the archived training data, eight 8K responses per question, and four 256-token probes at each reasoning midpoint and endpoint. Candidates share responses within each model. With Full fixed as control, development endpoint correctness on complete, nonempty reasoning selects among the three alternatives; ties prefer shorter prefill, then fixed candidate order. All models select Boxed before confirmation. Temperature is 1; Qwen uses top-$p = 1$ with top-$k$ disabled, Gemma top-$p = 0.95$, top-$k = 64$. Sampling/bootstrap seed 43 is distinct from the training seed. Exact appended strings follow: `\n` denotes a newline; replace `C` with `</think>` (Qwen) or `<channel|>` (Gemma). Quotes are excluded; retain the Answer cue's trailing space.

```
Full:         "\n...\nC\n\n**Final Answer**\n"
Native close: "\nC"
Answer cue:   "\nC\nThe final answer is "
Boxed:        "\nC\n\boxed{"
```

Boxed uses six prefill tokens outside the generation budget. The model generates the answer and closing brace; gold and the original final-answer segment are excluded. Verification includes the actual prefill.

**Results and budget check.** Figure 4 shows confirmation gains of 62.94 pp for Gemma4-12B-it (95% CI $[52.90, 72.85]$) and 55.18 pp for Gemma4-E4B-it ($[44.78, 65.13]$), versus 0.00 pp ($[0.00, 0.00]$) and 0.21 pp ($[0.00, 0.72]$) for Qwen3-1.7B/8B. Native close's zero scores need not mean semantically incorrect answers. On 16 development questions, extending the tail to 1,024 tokens improves Full but leaves both Gemma models' Boxed endpoint correctness unchanged while increasing generation cost, indicating a cue–format–budget interaction.

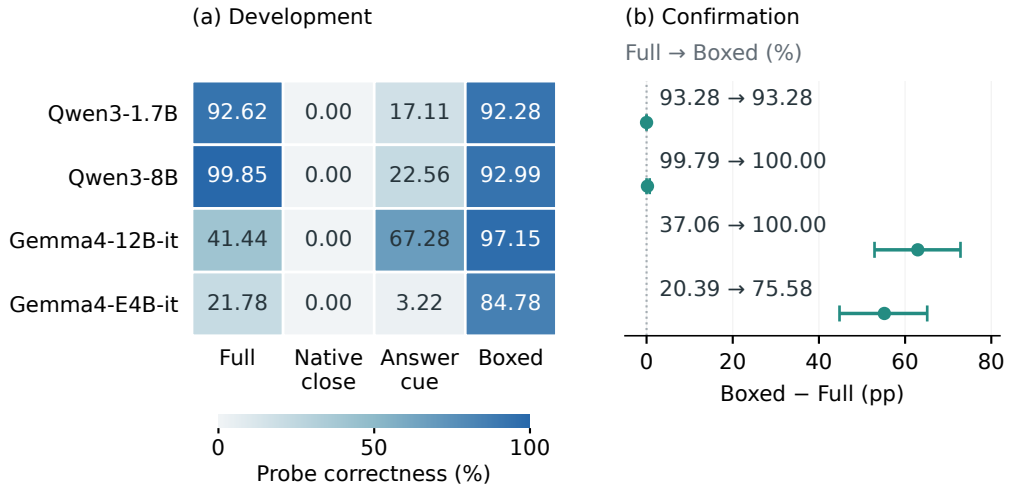


Figure 4: **Answer-cue selection across models.** Endpoint per-probe correctness under strict boxed verification, conditional on complete, nonempty reasoning. (a) Development: 149, 164, 149, 225 eligible responses in row order, out of 256 per model. (b) Confirmation: Full→Boxed percentages and paired differences with 95% intervals from 100,000 question-cluster bootstrap resamples, pooling eligible responses within resampled clusters. Eligible responses: 119, 120, 114, 217 across 22, 19, 19, 32 questions. Both narrow Qwen intervals include zero. These are extraction rates, not benchmark accuracies or cross-model capability rankings.

**Protocol choice and limits.** Qwen training retains Full because Boxed has no strict confirmation gain and slightly lower all-response rates; Gemma uses Boxed with the continuous-history rule above. E4B's midpoint gain is inconclusive ($+4.72$ pp, $[-1.20, 11.74]$), and all 16 truncated confirmation responses fail Boxed endpoint verification. Extraction gains can saturate rewards: with two endpoint probes, Gemma4-12B-it active groups fall from 87.50% to 18.75%. This frozen cue audit does not establish RL gains.

### B.4 Optimization and systems

The basic objective is Eq. equation 2. The sampled KL term is $\widehat{d} = \pi_{\text{ref}}/\pi_\theta - \log(\pi_{\text{ref}}/\pi_\theta) - 1$, evaluated on the original token and history. Scores and advantages are held fixed during updates. Tables 4–5 give the settings; the equations below specify implementation corrections shared by the compared rewards.

The configured prompt batch of 32 expands to 256 responses for $n = 8$. Capped completions and groups with zero task advantages remain in the batch. There is no extra SFT, replay, learned critic, dynamic reference, or reward-driven data filtering.

For completeness, the shared implementation differs from the basic surrogate in Eq. equation 2 by numerical clipping, dual clipping, and rollout-backend correction. Define

$$\begin{aligned}\widetilde{\rho}_{i,t} &= \exp \operatorname{clip}(\log \pi_\theta - \log \pi_{\text{old}}, -20, 20),\\ \widetilde{g}_{i,t} &= \min\{\widetilde{\rho}_{i,t} A_i, \operatorname{clip}(\widetilde{\rho}_{i,t}, 0.8, 1.2) A_i\},\\ g^{\text{dc}}_{i,t} &= \begin{cases} \max(\widetilde{g}_{i,t}, 3A_i), & A_i < 0,\\ \widetilde{g}_{i,t}, & A_i \geq 0. \end{cases}\end{aligned} \tag{17}$$

All probabilities here are evaluated on $(y_{i,t}, h_{i,t})$. With recorded rollout probabilities $\pi_{\text{roll}}$,

$$u_{i,t} = \exp \operatorname{clip}(\log \pi_{\text{old}} - \log \pi_{\text{roll}}, -20, 20), \qquad w_{i,t} = \operatorname{sg}\left[u_{i,t} \mathbf{1}\{10^{-6} \leq u_{i,t} \leq 3\}\right]. \tag{18}$$

The correction zeroes out-of-range weights rather than clipping them to the thresholds. It multiplies only the policy term; zero-weight tokens remain in the sequence-length denominator. Let $z_{i,t} = \operatorname{clip}(\log \pi_{\text{ref}} - \log \pi_\theta, -20, 20)$ and $\widetilde{d}_{i,t} = \operatorname{clip}(e^{z_{i,t}} - z_{i,t} - 1, -10, 10)$. For a minibatch of $B$

Table 4: Qwen3-1.7B training settings. Qwen3-8B and both Gemma variants' differences are listed in Table 5. The 8K runs form the main comparison; 4K Qwen3-1.7B runs appear in the supplementary results.

| Item | Setting |
|---|---|
| Training budget | 200 steps; 32 prompts/step; 8 responses/prompt; 51,200 original responses |
| Token limits | Prompt 2,048; original response 4,096 or 8,192; probe output 256 |
| AdamW | Learning rate $10^{-6}$; betas $(0.9, 0.999)$; weight decay 0 |
| Schedule | 20 warmup steps, then constant; gradient norm clip 1.0 |
| PPO | One epoch per rollout; configured prompt minibatch 32; dynamic batching; per-GPU microbatch 1 |
| Advantages | Within-question sample standard deviation, $\epsilon = 10^{-6}$ |
| Policy loss | Sequence mean of token means; clip 0.2; negative-advantage dual clip 3.0; entropy coefficient 0 |
| KL | Fixed reference; coefficient 0.001; separate actor-loss term, not part of reward |
| Rollout correction | Token-level weights; outside $[10^{-6}, 3]$ set to zero; no batch normalization or rejection sampling |
| Training | Full-parameter fully sharded data parallelism (FSDP); gradient checkpointing; padding removal; no low-rank adapters |
| Precision/ offload | FSDP engine bfloat16, model dtype fp32; actor parameters/optimizer and reference parameters offloaded |
| Generation | vLLM bfloat16, tensor parallelism 1; prefix caching and chunked prefill |
| Engine capacity | Context 24,576; batched tokens 8,192; sequences 1,024; memory utilization 0.9 |
| Hardware | One node with four H100 SXM 80GB GPUs |
| Checkpoint rule | Save steps 85/170/200; main results fixed to step 200 |

Table 5: Model-specific training and mathematical evaluation settings. Core methods follow Section 5.1 and share learning rate $10^{-6}$, 20 warmup steps, and KL coefficient 0.001.

| Setting | Qwen3-1.7B | Qwen3-8B | Gemma4-E4B-it | Gemma4-12B-it |
|---|---|---|---|---|
| Probe cue | Full; 8 tokens | Full; 8 tokens | Close + boxed; 6 tokens | Close + boxed; 6 tokens |
| Prefix boundary | First close | First close | Last close[†] | Last close[†] |
| Added opening | None | None | 3 thought tokens | 3 thought tokens |
| Actor execution | FSDP[‡] | Static FSDP | Static FSDP | Static FSDP |
| Rollout | TP=1 | TP=2, four replicas | TP=2, four replicas | TP=2, four replicas |
| Hardware | H100 $\times 4$ | H200 $\times 8$ | H100 $\times 8$ | H200 $\times 8$ |
| Train context | 24,576 | 30,720 | 24,576 | 30,720 |
| Train sampler | $1/1/\text{off}$ | $1/1/\text{off}$ | $1/1/\text{off}$ | $1/1/\text{off}$ |
| Math sampler | $0.6/0.95/20$ | $0.6/0.95/20$ | $1/0.95/64$ | $1/0.95/64$ |
| Eval prompt | Native thinking | Native thinking | Native; no added opening | Native; no added opening |

[†]Both Gemma models use the full available history if reasoning reopens after the last close. [‡]Qwen3-1.7B uses FSDP with padding removal and dynamic batching. Static FSDP uses DP=8/SP=1, SDPA attention, padding, and microbatch 1. DP, SP and TP denote data, sequence and tensor parallelism. Sampler entries give temperature/top-$p$/top-$k$. Comparisons control settings within each model; actual computation is not matched across models or rewards. Qwen3-8B and Gemma4-12B-it evaluation uses H100 GPUs, distinct from their H200 training allocations. The 12B $\mathrm{PSP}_{\mathrm{E}}$ run includes the recovery patch described in Appendix D.2.

original responses, response masks $M_{i,t}$, and $T_i = \sum_t M_{i,t}$, the actor loss is

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^{B} \frac{\sum_t M_{i,t}[-w_{i,t} g^{\mathrm{dc}}_{i,t} + 10^{-3} \widetilde{d}_{i,t}]}{T_i + 10^{-8}}, \tag{19}$$

omitting distributed scaling used to recover the global mean. Scores, advantages, old/reference probabilities, and correction weights are fixed during differentiation. These are shared baseline details, not new components of the reward construction. The configuration-to-code correspondence does not establish a byte-for-byte reconstruction of historical shared dependencies.

Table 6: Frozen mathematical evaluation protocol. Repeats estimate average single-response correctness, not pass@$k$.

| Dataset | Questions | Replies/question | Total replies | Output cap |
|---|---|---|---|---|
| GSM8K | 1,319 | 1 | 1,319 | 32,768 |
| MATH500 | 500 | 1 | 500 | 32,768 |
| AIME24 | 30 | 64 | 1,920 | 38,912 |
| AIME25 | 30 | 64 | 1,920 | 38,912 |
| AIME26 | 30 | 32 | 960 | 38,912 |
| HMMT25 | 30 | 32 | 960 | 38,912 |
| HMMT26 | 33 | 32 | 1,056 | 38,912 |
| Total | 1,972 | — | 8,635 | Dataset-specific |

### B.5 Mathematical evaluation

Both Qwen models use mathematical sampling temperature 0.6, top-$p$ 0.95, top-$k$ 20, and min-$p$ 0 with context 40,960. Qwen3-1.7B uses tensor parallelism 1, memory utilization 0.9 and chunked prefill, with eight shards and at most 16 repeats per batch. Qwen3-8B uses four H100 GPUs as two TP=2 replicas, 16 concurrent requests per replica, and the frozen package-specific repeat partitions. HMMT26 is the local 33-question version, not an assumed 30-question set.

Gemma mathematical evaluation keeps the same question counts, repeats, and caps, but uses temperature 1, top-$p = 0.95$, top-$k = 64$, context 40,960, and TP=2 with two external replicas. Its native evaluation template omits the training-only artificial thought opening. Within each model these settings are held fixed across methods.

Six datasets use frozen MathEqual scores; GSM8K uses the DeepScaler-style verifier because of its reference-answer format. Unsuccessful verification, including missing verifier results, counts as incorrect and remains in the denominator; grading errors and truncations are recorded. For dataset $d$ with $N_d$ questions and $J_d$ replies per question,

$$\widehat{a}_d = \frac{1}{N_d}\sum_{q=1}^{N_d}\frac{1}{J_d}\sum_{j=1}^{J_d} R_{d,q,j}, \qquad \text{Math-7} = \frac{1}{7}\sum_{d=1}^{7}\widehat{a}_d. \tag{20}$$

There is no best-answer selection or majority vote. Length and truncation summaries also average within each dataset before the seven-dataset macro average. Total evaluation output instead sums every generated response. Records retain prompt IDs, trial indices, random-stream identifiers, verifier outcomes, errors, finish reasons, and token usage. Coverage, repeat counts, record uniqueness, and agreement between raw scores and aggregates are checked before accepting an evaluation result.

### B.6 Non-mathematical and short-budget evaluation

Generative GPQA$^\diamond$ uses 198 questions with ten replies each, a 32,768-token output cap, and the mathematical sampling settings above; answer choices are shuffled according to the frozen evaluation protocol. Its user template is

```
{question}

A: {a}
B: {b}
C: {c}
D: {d}

Please reason step by step, and put your final answer within \boxed{}.
Please only provide the letter of the answer in the box.
```

IFEval contains 541 prompts and reports strict prompt-level accuracy. For Qwen it uses the original task prompts, frozen chat template, a 32,768-token output cap, temperature 0.6, top-$p$ 0.95, top-$k$ 20, sampling enabled; the evaluator is configured with `</think>` as the thinking boundary.

Three additional generation protocols—BFCL, GPQA-Main (denoted GPQA$^{\mathrm{M}}$), and HumanEval—are adapted from PostTrainBench (Rank et al., 2026). We report each task independently. BFCL (a local 100-example subset) and GPQA$^{\mathrm{M}}$ (448 examples) have 16,000-token output caps; HumanEval (164 examples, one generation each) has a 16,384-token cap shared by thinking and final code for every reported model. Earlier 4,000-token results are retained only as budget-sensitivity evidence, not mixed into these scores. The new-budget answers are independently generated, not continuations of earlier outputs. API-judge tasks are excluded. GPQA$^{\mathrm{M}}$ uses a generation protocol distinct from GPQA$^{\diamondsuit}$. For the historical Qwen3-1.7B runs, these three task runners did not explicitly override all sampler parameters; reproduction requires the frozen Inspect/backend defaults and logs, not the mathematical sampler settings.

Likelihood-based evaluation uses lm-eval (Gao et al., 2023) for HellaSwag, MMLU, TruthfulQA, WinoGrande, and GPQA$^{\diamondsuit}$. These tasks score candidate answers without generating completions. TruthfulQA reports MC2, the question-averaged normalized probability mass assigned to all correct candidate answers. Appendix C.3 reports the complete five-benchmark suite to assess general-capability retention. Likelihood-based GPQA$^{\diamondsuit}$ remains separate from the generative reasoning protocol above.

Qwen3-8B reuses its frozen starting-model task protocols, including the GPQA-Main inputs and answer-option ordering; training sampling settings are not propagated to these suites.

For both Gemma models, non-mathematical tasks retain their frozen suite-specific samplers and native channel parsing; mathematical settings are not propagated to every suite. For Gemma4-E4B-it, the starting IFEval headline is retained: deterministic rescoring yields the same aggregate but changes some item-level outcomes, so these are not used as strictly paired evidence across scoring versions. Gemma4-E4B-it $\mathrm{PSP}_{\mathrm{ME}}$ GPQA-Main combines 350 previously completed samples with 98 recovered samples under the same task rules; the original task did not fix option-shuffle seeds, so unseen option orders cannot be reconstructed. All 448 questions are covered once. HumanEval uses the same 16,384-token cap as Qwen.

Gemma4-12B-it uses frozen starting-model questions/options for GPQA-Main. $\mathrm{PSP}_{\mathrm{E}}$ and $\mathrm{PSP}_{\mathrm{ME}}$ IFEval each use four disjoint partitions, aggregated by question/instruction rather than an unweighted mean of partition scores; all 541 questions are covered. Initial IFEval retains its historical scoring version and is descriptive only. Native evaluation templates omit the training thought opening.

### B.7 Fixed-prefix natural completion and endpoint potential

We use a fixed prefix-evaluation protocol. Clean uses 256 questions and four original replies each at steps 85/170/200, plus all four starting models. Its question set differs from Math-7. All four models' Math-7 prefix panels retain the original 1,972 questions and 8,635 replies per model from Table 6. Within each model, one original trajectory supplies all three budgets $B \in \{2048, 4096, 8192\}$; trajectories are not shared across models. Original IDs are reused for Gemma4-E4B-it trained Clean, Gemma4-12B-it GRPO Clean, and all states of both Gemma models and Qwen3-8B on Math. Gemma4-12B-it $\mathrm{PSP}_{\mathrm{E}}$/$\mathrm{PSP}_{\mathrm{ME}}$ Clean roots were generated for their own checkpoints in the accepted earlier batch. Missing IDs require new rollout from the matching checkpoint, used for all Qwen3-1.7B roots and Gemma/Qwen3-8B starting Clean; historical text is never re-tokenized as an original trajectory. Clean original sampling is temperature 1, with Qwen top-$p = 1$ and no top-$k$ restriction, and Gemma top-$p = 0.95$, top-$k = 64$. Math originals follow the model-specific mathematical sampling settings in Appendix B.5: new rollouts are capped at 8K, while reused long trajectories retain only their original first 8K IDs. New paired $R$ values are distinct from the archived historical diagnostics.

For original output IDs $y$, let $c_B$ be the model-specific contiguous prefix obtained from $y_{\leq \min(B,|y|)}$. We apply boundary detection only after truncation: Qwen ends before the first reasoning close; Gemma ends before the last close unless reasoning reopens. No tokens beyond $B$ enter boundary selection or the probe. Gemma may retain an earlier generated answer. We evaluate

$$\widehat{\phi}_{100}(B) = \frac{1}{4}\sum_{j=1}^{4} R_{\mathrm{probe}}(x, a_j), \qquad a_j \sim \pi_{\mathrm{ckpt}}^{\mathcal{P}_{\mathrm{eval}}}(\cdot \mid x, c_B). \tag{21}$$

Here 100 denotes the end of the retained window, not the probe count or output cap; this is the endpoint estimator of Equation 4. Four independent probes use the same evaluated checkpoint, Qwen

Full cue or Gemma native close plus boxed cue, temperature 1, top-$p = 1$, no top-$k$ restriction, and a maximum of 256 tokens each. The original answer score $R(B)$ verifies the original output visible within $B$, without a probe. The two contexts consequently share an original trajectory but need not retain identical lengths. Clean uses DeepScaler for both scores; Math-7 uses MathEqual for $R(B)$ and DeepScaler for probes. The latter difference prevents treating $\widehat{\phi}_{100} - R$ as a pure completion or truncation effect.

We first average probes, then original trials within each question, then questions; Math-7 equally weights seven datasets. Question-paired bootstrap uses 100,000 draws, seed 43, within datasets, preserving all within-question dependence. We report nominal intervals without multiple-comparison adjustment. Seeds identify model, dataset, question, trial, budget and probe independently of sharding or submission order. Frozen data preserve source hashes, IDs, cutpoints, probe seeds, scores and costs. Reused trajectories exclude their historical generation cost from the new evaluation allocation. The four probes estimate a single-sample success probability, not deployment-time best-of-four accuracy.

Archived records preserve source hashes, original token IDs, prefix cuts, probe seeds and scores. The extended evidence records model-specific task inventories and execution receipts; these do not constitute independent experimental repetitions.

### B.8 Initial-model evaluation against official reports

Qwen3-1.7B is within 1.8 pp of official reports on five overlapping tasks; model-specific evaluation settings and the Gemma4-E4B-it differences are documented here. We compare the Qwen3-1.7B and Gemma4-E4B-it Initial rows in Tables 8 and 9 with the corresponding thinking-mode checkpoints in the Qwen3 and Gemma 4 technical reports (Qwen Team, 2025; Gemma Team, 2026). These are the released post-trained models before our additional RL, rather than separately pretrained Base variants. Table 7 includes overlapping benchmarks with reported official scores, retaining discrepancies and distinguishing unresolved metric definitions.

Table 7: Initial-model scores against official thinking-mode reports (percent).

| Benchmark | Official | Ours | Δ pp | Our evaluation |
|---|---|---|---|---|
| *Qwen3-1.7B (thinking)* | | | | |
| MATH500 | 93.4 | 91.60 | −1.80 | 1 reply/question |
| AIME24 | 48.3 | 47.40 | −0.90 | 64 replies/question |
| AIME25 | 36.8 | 35.73 | −1.07 | 64 replies/question |
| GPQA-Diamond | 40.1 | 38.38 | −1.72 | 10 replies/question |
| IFEval | 72.5 | 71.35 | −1.15 | Strict prompt-level |
| *Gemma4-E4B-it (thinking)* | | | | |
| AIME26 | 42.5 | 49.58 | +7.08 | 32 replies/question |
| GPQA-Diamond | 58.6 | 56.16 | −2.44 | 10 replies/question |
| IFEval | 96.7 | 87.25 | — | Strict prompt-level |

Official values: Qwen3 report v1, Table 19 (Qwen Team, 2025); Gemma 4 report v1, Table 5 (Gemma Team, 2026). Ours reuses the Initial rows in Tables 8 and 9. Differences are ours minus official, computed before display rounding; official scores are available to one decimal place. For mathematics and GPQA, Qwen uses temperature 0.6, top-$p = 0.95$, top-$k = 20$; Gemma uses temperature 1, top-$p = 0.95$, top-$k = 64$. Both use 32,768-token caps for MATH500/GPQA and 38,912 for AIME, with average single-response accuracy and native thinking templates. Our IFEval scores use strict prompt-level accuracy. The Gemma report labels its entry only as IFEval without specifying scoring granularity, so its difference is left blank.

**Interpretation.** Qwen3-1.7B's five overlapping scores are 0.90–1.80 pp below the official reports. Samplers, caps, AIME repeats and GPQA repeats correspond, but equality of dataset revisions, templates, scorers and all repeat counts is not established. Gemma4-E4B-it is 2.44 pp below official GPQA and 7.08 pp above AIME26; the latter discrepancy remains unexplained. The official Gemma report does not establish matching task-specific protocols or IFEval granularity, so no IFEval difference is computed. These are descriptive starting-model checks, not equivalence tests. Local BFCL and likelihood MMLU are not equated with full BFCL or other MMLU variants.

### B.9 Statistical comparisons and the OPD reference

Scores, costs, lengths, and truncation rates refer to the corresponding trained checkpoints; training-history diagnostics use their recorded trajectories. OPD uses the training seed specified in Section 5.1, while evaluation and diagnostic sampling seeds are specified separately. We compute differences before display rounding.

Mathematical intervals use 100,000 paired question-bootstrap resamples, with bootstrap seed 20260909 for Qwen3-1.7B and 42 for Gemma. These bootstrap seeds are separate from training seeds. Repeated replies are first averaged within each question at each checkpoint. We subtract the two checkpoints' question means and resample these paired differences within each dataset, retaining equal weights across the seven datasets. This estimates uncertainty over test questions while preserving dependence among repeated replies.

OPD uses a frozen Qwen3-8B teacher, full-vocabulary generalized Jensen–Shannon divergence with $\alpha = 0.5$, 271 steps, effective batch size 64, one original response per problem, a 4K output cap, learning rate $5 \times 10^{-6}$, cosine scheduling, and 0.1 warmup fraction. This reference is neither a performance upper bound nor a matched-resource baseline.

### B.10 Implementation records and remaining reproducibility gaps

Archived configurations, custom source identities, checkpoint-transfer records and per-step diagnostics support the reported experiments. Historical Qwen runs used shared environments and a shared verl path, so declared revisions do not freeze every loaded dependency. Outstanding items include immutable training-image and dependency locks, complete exported-weight and benchmark-source identities, historical task-sampler receipts, a complete train–test overlap audit, peak memory and cache-adjusted processing, and independent evaluation/teacher resource totals. Strict numerical replay, a full-project resource bill and a same-question independently generated test-budget sweep remain unverified. Suite-specific environment declarations and detailed receipts are retained in the extended evidence.

## C Supplementary evaluation results

Results are organized by measurement, with models ordered Qwen3-1.7B, Qwen3-8B, Gemma4-E4B-it and Gemma4-12B-it. Table 1 is the common reference for trained-model long-cap mathematics; Appendix E retains all additional Qwen configurations.

### C.1 Initial-model benchmarks and pre-RL comparisons

Table 8 reports pre-RL mathematics; Tables 9 and 10 include Initial alongside the trained methods. Initial denotes the checkpoint before the RL training studied here, which may already have undergone substantial post-training. These comparisons contextualize the matched GRPO–PSP comparison in the main table.

Initial Math-7 scores are 50.26%, 69.57%, 56.39%, and 81.83% for Qwen3-1.7B, Qwen3-8B, Gemma4-E4B-it, and Gemma4-12B-it, respectively. $\mathrm{PSP}_{\mathrm{ME}}$ exceeds initialization by 0.80 pp on Qwen3-1.7B and 9.17 pp on Gemma4-E4B-it, but remains 0.87 pp below it on Qwen3-8B and 1.90 pp below it on Gemma4-12B-it. The two larger models therefore mainly recover long-budget performance lost after outcome-reward training. This recovery is consistent with the budget-related risks discussed in Section 2, but the fixed-cap reward comparison does not isolate training length as the cause of degradation. Paired comparisons appear in Table 11; official-report checks are in Appendix B.8.

### C.2 Other capabilities across the four models

The right-hand columns of Table 1 report the non-mathematical main comparison; this appendix provides additional detail. Generation tasks cover scientific knowledge reasoning (GPQA$^{\diamond}$), strict prompt-level instruction following (IFEval), code generation (HumanEval), and a local tool-use

Table 8: Initial-model mathematical results under long output caps (percent).

| Model | GSM8K | MATH500 | AIME24 | AIME25 | AIME26 | HMMT25 | HMMT26 | Math-7 |
|---|---|---|---|---|---|---|---|---|
| Qwen3-1.7B | 90.60 | 91.60 | 47.40 | 35.73 | 36.25 | 22.71 | 27.56 | 50.26 |
| Qwen3-8B | 95.91 | 94.80 | 75.57 | 67.34 | 65.73 | 42.29 | 45.36 | 69.57 |
| Gemma4-E4B-it | 93.48 | 93.60 | 55.26 | 39.53 | 49.58 | 31.25 | 32.01 | 56.39 |
| Gemma4-12B-it | 96.21 | 98.60 | 85.26 | 78.44 | 79.90 | 72.40 | 62.03 | 81.83 |

Initial denotes the pre-RL checkpoint. Evaluation uses the same model-specific protocols as Table 1; Math-7 weights seven datasets equally.

subset (BFCL). Table 9 collects the four models, including GPQA-Main; likelihood evaluations follow separately.

Table 9 shows task-dependent transfer rather than broad capability gains. In particular, Gemma4-E4B-it ME loses IFEval accuracy despite mathematical gains; Qwen3-8B E loses eight of 164 HumanEval problems relative to GRPO. The local BFCL subset has only 100 examples. All point estimates, including regressions, remain reported.

Table 9: Generative capabilities across four models (percent).

| Method | GPQA$^{\diamond}$ | IFEval | HumanEval | BFCL | GPQA$^{\mathrm{M}}$ |
|---|---|---|---|---|---|
| *Qwen3-1.7B* | | | | | |
| Initial | 38.38 | 71.35 | 80.49 | 93.00 | 34.82 |
| GRPO | 36.77 | 68.76 | 81.71 | 94.00 | 34.82 |
| $\mathrm{PSP}_{\mathrm{E}}$ | 38.38 | 71.90 | 79.88 | 92.00 | 35.71 |
| $\mathrm{PSP}_{\mathrm{ME}}$ | 39.29 | 69.69 | 81.71 | 95.00 | 33.93 |
| *Qwen3-8B* | | | | | |
| Initial | 60.40 | 85.21 | 89.63 | 96.00 | 52.01 |
| GRPO | 58.79 | 85.21 | 91.46 | 93.00 | 54.46 |
| $\mathrm{PSP}_{\mathrm{E}}$ | 61.46 | 84.84 | 86.59 | 96.00 | 55.80 |
| $\mathrm{PSP}_{\mathrm{ME}}$ | 61.41 | 85.95 | 90.85 | 97.00 | 53.57 |
| *Gemma4-E4B-it* | | | | | |
| Initial | 56.16 | 87.25 | 84.76 | 90.00 | 52.46 |
| GRPO | 54.80 | 87.99 | 85.37 | 90.00 | 56.03 |
| $\mathrm{PSP}_{\mathrm{E}}$ | 57.58 | 87.99 | 84.15 | 88.00 | 57.37 |
| $\mathrm{PSP}_{\mathrm{ME}}$ | 59.39 | 86.69 | 85.37 | 93.00 | 56.70 |
| *Gemma4-12B-it* | | | | | |
| Initial | 63.64 | 90.76 | 98.17 | 95.00 | 61.61 |
| GRPO | 66.57 | 93.35 | 96.34 | 96.00 | 68.75 |
| $\mathrm{PSP}_{\mathrm{E}}$ | 70.96 | 92.79 | 96.34 | 96.00 | 66.96 |
| $\mathrm{PSP}_{\mathrm{ME}}$ | 66.57 | 90.57 | 98.17 | 98.00 | 65.40 |

Initial is pre-RL. HumanEval uses 16,384 tokens; IFEval is strict prompt-level accuracy; BFCL is the local 100-example subset. Gemma4-12B-it Initial IFEval uses a historical scorer, so comparisons with initialization are descriptive (Appendix B.6). OPD is reported in Table 1.

### C.3 General-capability retention on likelihood benchmarks

Table 10 reports all five likelihood-based benchmarks for Initial, GRPO, $\mathrm{PSP}_{\mathrm{E}}$, and $\mathrm{PSP}_{\mathrm{ME}}$ across the four models, plus the separate Qwen3-1.7B OPD reference. These evaluations assess general knowledge, commonsense, truthfulness, and candidate-answer discrimination after mathematics-only training. They do not measure generated reasoning or instruction following.

ME differs from Initial by less than 0.40 pp on MMLU, HellaSwag, TruthfulQA and WinoGrande across all four models; likelihood-GPQA differences range from $-1.01$ to $+2.02$ pp. These descriptive results do not establish equivalence or non-inferiority, and do not erase the generative-task regressions.

Table 10: Complete likelihood-based evaluation of general capabilities (percent).

| Method | MMLU | HellaSwag | TruthfulQA | WinoGrande | GPQA$^{\diamond}$ |
|---|---|---|---|---|---|
| *Qwen3-1.7B* | | | | | |
| Initial | 55.44 | 60.40 | 45.84 | 61.48 | 20.20 |
| GRPO | 55.68 | 60.45 | 45.67 | 60.46 | 21.21 |
| $\mathrm{PSP_E}$ | 55.53 | 60.44 | 46.01 | 60.85 | 19.70 |
| $\mathrm{PSP_{ME}}$ | 55.41 | 60.45 | 45.92 | 61.17 | 20.20 |
| OPD-4K | 55.95 | 60.90 | 47.03 | 61.72 | 21.21 |
| *Qwen3-8B* | | | | | |
| Initial | 73.02 | 74.90 | 54.46 | 68.19 | 26.77 |
| GRPO | 72.81 | 74.91 | 54.63 | 67.88 | 28.28 |
| $\mathrm{PSP_E}$ | 72.95 | 74.87 | 54.37 | 67.96 | 28.28 |
| $\mathrm{PSP_{ME}}$ | 72.97 | 75.01 | 54.29 | 67.80 | 28.79 |
| *Gemma4-E4B-it* | | | | | |
| Initial | 69.90 | 68.14 | 59.03 | 66.46 | 23.74 |
| GRPO | 69.85 | 68.01 | 59.11 | 66.93 | 22.22 |
| $\mathrm{PSP_E}$ | 69.88 | 67.95 | 59.14 | 66.69 | 23.74 |
| $\mathrm{PSP_{ME}}$ | 69.99 | 67.85 | 59.05 | 66.54 | 22.73 |
| *Gemma4-12B-it* | | | | | |
| Initial | 69.21 | 56.16 | 54.59 | 56.59 | 24.75 |
| GRPO | 68.80 | 54.92 | 55.15 | 56.27 | 24.24 |
| $\mathrm{PSP_E}$ | 69.15 | 55.25 | 56.30 | 57.30 | 22.73 |
| $\mathrm{PSP_{ME}}$ | 69.26 | 56.03 | 54.26 | 56.51 | 24.24 |

All five likelihood-based benchmarks in the frozen evaluation suite are reported. Initial is the pre-RL checkpoint; blue highlights $\mathrm{PSP_{ME}}$. TruthfulQA reports MC2. GPQA$^{\diamond}$ here scores answer candidates by likelihood, without generated reasoning; it is distinct from generative GPQA in Table 1. OPD is the separately configured Qwen3-1.7B reference. Values are descriptive point estimates, not an equivalence test.

### C.4 Paired mathematical comparisons

Table 11 collects the within-model comparisons and Qwen3-1.7B probe-count ablations. Confidence intervals quantify question uncertainty, not training-run variation. Larger-model improvements over GRPO coexist with scores below Initial; neither larger model establishes a benefit of ME over E.

### C.5 Fixed-prefix budgets and answer potential

The boundary rules, verifiers, sampling and aggregation are specified in Appendix B.7. Figure 5 compares all four models across budgets; Table 12 gives the complete per-dataset 8K values. The three budgets reuse each original trajectory, not independent generations at each cap.

At 8K, the potential of Gemma4-E4B-it $\mathrm{PSP_{ME}}$ exceeds GRPO by 2.33 pp on Math-7 (paired question-bootstrap 95% CI $[1.26, 3.45]$) and 3.27 pp on Clean ($[0.61, 5.96]$), despite much lower natural completion. Qwen3-1.7B's corresponding differences are $+0.92$ pp ($[-0.14, 2.00]$) and $-0.12$ pp ($[-2.66, 2.39]$), with neither interval excluding zero. Clean uses different questions from Math-7 and is not subtracted from it.

On Clean, Qwen3-8B GRPO/$\mathrm{PSP_E}$/$\mathrm{PSP_{ME}}$ reach 72.29/72.09/72.14% potential (Figure 6), versus 71.46% at initialization. $\mathrm{PSP_{ME}}$ minus GRPO is $-0.15$ pp (95% CI $[-2.81, +2.51]$). On Math-7, $\mathrm{PSP_{ME}}$ reaches 56.95% versus 58.17% for GRPO: $-1.22$ pp (95% CI $[-2.40, -0.05]$). It differs from $\mathrm{PSP_E}$ by $-0.23$ pp (95% CI $[-1.02, +0.58]$) and exceeds Initial potential by $+2.13$ pp (95% CI $[+1.12, +3.18]$), despite remaining below initialization on long-cap accuracy. Thus a larger potential-minus-completion gap does not imply stronger prefixes than GRPO. These nominal intervals are not adjusted for multiple comparisons; the 8K Math-7 upper endpoint is close to zero.

The multi-budget Math-7 curves in Appendix C.5 (Figure 5) do not show a general earlier-answer advantage. Qwen3-8B $\mathrm{PSP_{ME}}$ minus GRPO is $-1.61$ pp (95% CI $[-2.52, -0.79]$) at 2K and $-1.66$ pp (95% CI $[-2.62, -0.78]$) at 4K, and remains negative at 8K. Qwen3-1.7B $\mathrm{PSP_{ME}}$ trails GRPO by 2.39 pp at 2K ($[-3.37, -1.48]$) and 1.12 pp at 4K ($[-1.93, -0.30]$). Gemma4-E4B-it is close at

Table 11: Question-paired long-cap Math-7 comparisons (percentage points).

| Comparison | Difference | 95% CI |
|---|---|---|
| *Qwen3-1.7B* | | |
| $\mathrm{PSP_E}$ ($m=1$) − GRPO | $+2.36$ | $[+1.23, +3.50]$ |
| $\mathrm{PSP_E}$ ($m=4$) − GRPO | $+1.78$ | $[+0.76, +2.83]$ |
| $\mathrm{PSP_E}$ ($m=1$) − $\mathrm{PSP_E}$ | $-0.27$ | $[-1.20, +0.64]$ |
| $\mathrm{PSP_E}$ ($m=4$) − $\mathrm{PSP_E}$ | $-0.85$ | $[-1.78, +0.06]$ |
| $\mathrm{PSP_E}$ ($m=4$) − $\mathrm{PSP_E}$ ($m=1$) | $-0.57$ | $[-1.69, +0.52]$ |
| $\mathrm{PSP_{ME}}$ − GRPO | $+3.89$ | $[+2.84, +4.99]$ |
| $\mathrm{PSP_{ME}}$ − $\mathrm{PSP_E}$ | $+1.26$ | $[+0.47, +2.06]$ |
| $\mathrm{PSP_{ME}}$ − $\mathrm{PSP_E}$ ($m=4$) | $+2.11$ | $[+1.20, +3.03]$ |
| *Qwen3-8B* | | |
| $\mathrm{PSP_E}$ − GRPO | $+4.65$ | $[+3.37, +5.97]$ |
| $\mathrm{PSP_{ME}}$ − GRPO | $+4.45$ | $[+3.18, +5.80]$ |
| $\mathrm{PSP_{ME}}$ − $\mathrm{PSP_E}$ | $-0.20$ | $[-1.04, +0.65]$ |
| GRPO − Initial | $-5.33$ | $[-6.76, -3.97]$ |
| $\mathrm{PSP_E}$ − Initial | $-0.68$ | $[-1.42, +0.04]$ |
| $\mathrm{PSP_{ME}}$ − Initial | $-0.88$ | $[-1.75, -0.04]$ |
| *Gemma4-E4B-it* | | |
| $\mathrm{PSP_E}$ − GRPO | $+3.16$ | $[+1.86, +4.51]$ |
| GRPO − Initial | $+3.62$ | $[+2.46, +4.85]$ |
| $\mathrm{PSP_E}$ − Initial | $+6.78$ | $[+5.36, +8.27]$ |
| $\mathrm{PSP_{ME}}$ − GRPO | $+5.55$ | $[+3.81, +7.33]$ |
| $\mathrm{PSP_{ME}}$ − $\mathrm{PSP_E}$ | $+2.39$ | $[+1.19, +3.61]$ |
| $\mathrm{PSP_{ME}}$ − Initial | $+9.17$ | $[+7.29, +11.12]$ |
| *Gemma4-12B-it* | | |
| $\mathrm{PSP_{ME}}$ − GRPO | $+6.56$ | $[+5.03, +8.15]$ |
| $\mathrm{PSP_E}$ − GRPO | $+5.54$ | $[+4.11, +6.99]$ |
| $\mathrm{PSP_{ME}}$ − $\mathrm{PSP_E}$ | $+1.02$ | $[-0.05, +2.11]$ |
| GRPO − Initial | $-8.46$ | $[-10.40, -6.54]$ |
| $\mathrm{PSP_E}$ − Initial | $-2.92$ | $[-4.07, -1.83]$ |
| $\mathrm{PSP_{ME}}$ − Initial | $-1.90$ | $[-3.12, -0.71]$ |

All intervals use 100,000 question-bootstrap draws after averaging replies within question, with equal dataset weights; they do not measure training-seed variation. Qwen3-1.7B follows Appendix B; Qwen3-8B uses bootstrap seeds 20260923 for trained-method comparisons and 20260924 against Initial; both Gemma models use 42.

2K ($+0.09$ pp, $[-0.80, 0.98]$), lower at 4K ($-1.30$ pp, $[-2.25, -0.37]$), and higher at 8K. Gains in forced-answer availability are therefore model- and budget-dependent.

### C.6 Clean checkpoint trajectories

Figure 6 shows the three core checkpoints for all four models at a fixed 8K evaluation cap. Complete numerical values, including all 21 Qwen3-1.7B configurations, are retained in the extended evidence. Single-trajectory configurations use the separate fixed-baseline estimator in Appendix E.3.

### C.7 Teacher-based OPD reference

PSP assigns a trajectory-level reward using same-policy short answers; OPD supplies token-position feedback from a frozen Qwen3-8B teacher. Its complete mathematical and generative scores are in Table 1, and likelihood results in Table 10. OPD reaches 51.98% Math-7, 0.92 pp above $\mathrm{PSP_{ME}}$, but uses a 4K cap, one response per problem, 271 steps and a different optimizer schedule (Appendix B). It is neither a matched-compute baseline nor an upper bound. The frozen-student diagnostic in Appendix F uses a different reduction/batch setup and does not reconstruct this training run.

Table 12: Per-dataset paired 8K natural completion and endpoint potential (percent).

| Dataset | Initial | | GRPO | | $\mathrm{PSP}_{\mathrm{E}}$ | | $\mathrm{PSP}_{\mathrm{ME}}$ | |
|---|---|---|---|---|---|---|---|---|
| | $R$ | $\widehat{\phi}_{100}$ | $R$ | $\widehat{\phi}_{100}$ | $R$ | $\widehat{\phi}_{100}$ | $R$ | $\widehat{\phi}_{100}$ |
| *Qwen3-1.7B* | | | | | | | | |
| GSM8K | 89.54 | 90.62 | 90.22 | 90.13 | 68.46 | 68.80 | 88.25 | 90.24 |
| MATH500 | 78.40 | 86.95 | 86.40 | 87.20 | 80.60 | 86.20 | 71.80 | 85.50 |
| AIME24 | 15.42 | 31.38 | 28.91 | 32.66 | 21.41 | 32.88 | 7.24 | 33.42 |
| AIME25 | 17.08 | 26.35 | 24.06 | 25.79 | 20.31 | 26.15 | 10.62 | 25.51 |
| AIME26 | 12.40 | 28.07 | 20.94 | 22.71 | 15.62 | 24.95 | 6.25 | 27.16 |
| HMMT25 | 8.12 | 17.66 | 13.02 | 15.13 | 8.65 | 13.26 | 5.52 | 16.46 |
| HMMT26 | 10.80 | 19.20 | 12.78 | 17.35 | 9.75 | 15.39 | 5.11 | 19.11 |
| Math-7 | 33.11 | 42.89 | 39.48 | 41.57 | 32.11 | 38.23 | 27.83 | 42.49 |
| *Qwen3-8B* | | | | | | | | |
| GSM8K | 95.00 | 93.97 | 95.75 | 94.29 | 95.22 | 93.92 | 95.00 | 93.97 |
| MATH500 | 80.00 | 91.55 | 93.00 | 91.35 | 86.20 | 93.25 | 76.20 | 92.00 |
| AIME24 | 25.10 | 54.27 | 51.15 | 58.01 | 36.15 | 56.54 | 18.59 | 56.74 |
| AIME25 | 17.24 | 39.00 | 39.22 | 44.11 | 23.28 | 42.77 | 12.03 | 43.31 |
| AIME26 | 16.98 | 47.14 | 49.27 | 54.92 | 26.98 | 50.81 | 12.81 | 49.79 |
| HMMT25 | 9.06 | 24.24 | 21.77 | 28.15 | 11.46 | 27.06 | 6.15 | 29.19 |
| HMMT26 | 17.14 | 33.59 | 29.83 | 36.32 | 20.27 | 35.94 | 9.75 | 33.64 |
| Math-7 | 37.22 | 54.82 | 54.28 | 58.17 | 42.79 | 57.18 | 32.93 | 56.95 |
| *Gemma4-E4B-it* | | | | | | | | |
| GSM8K | 93.78 | 90.20 | 93.86 | 92.36 | 92.80 | 89.33 | 93.48 | 92.49 |
| MATH500 | 92.40 | 91.25 | 94.00 | 91.75 | 92.80 | 90.80 | 77.60 | 92.35 |
| AIME24 | 50.36 | 55.17 | 54.01 | 58.18 | 51.77 | 59.28 | 14.48 | 58.45 |
| AIME25 | 36.67 | 38.79 | 41.56 | 42.99 | 40.10 | 44.15 | 13.23 | 46.86 |
| AIME26 | 46.46 | 49.97 | 47.50 | 49.90 | 47.71 | 56.69 | 6.46 | 55.47 |
| HMMT25 | 26.04 | 26.15 | 29.58 | 30.99 | 24.58 | 32.32 | 4.06 | 33.26 |
| HMMT26 | 25.09 | 29.47 | 28.79 | 32.86 | 24.81 | 33.05 | 5.87 | 36.46 |
| Math-7 | 52.97 | 54.43 | 55.61 | 57.00 | 53.51 | 57.95 | 30.74 | 59.33 |
| *Gemma4-12B-it* | | | | | | | | |
| GSM8K | 90.83 | 89.48 | 96.89 | 64.40 | 96.21 | 84.65 | 92.80 | 92.27 |
| MATH500 | 82.40 | 77.25 | 97.20 | 48.80 | 89.80 | 82.05 | 87.20 | 85.30 |
| AIME24 | 24.90 | 57.08 | 75.05 | 55.35 | 46.04 | 67.30 | 34.17 | 66.34 |
| AIME25 | 23.44 | 46.98 | 60.57 | 44.09 | 38.80 | 55.44 | 26.15 | 52.76 |
| AIME26 | 13.96 | 54.11 | 66.35 | 52.34 | 37.60 | 61.74 | 23.75 | 59.82 |
| HMMT25 | 10.00 | 44.56 | 51.88 | 40.42 | 21.98 | 46.85 | 11.88 | 44.56 |
| HMMT26 | 12.03 | 38.35 | 48.58 | 40.22 | 24.62 | 44.08 | 17.61 | 44.79 |
| Math-7 | 36.79 | 58.26 | 70.93 | 49.38 | 50.72 | 63.16 | 41.94 | 63.69 |

$R$ scores the original output visible within 8K; $\widehat{\phi}_{100}$ adds four short-answer probes after model-specific prefix processing. Math-7 weights the seven datasets equally. R uses MathEqual and probes use DeepScaler; their difference is not a pure truncation effect. Qwen3-1.7B roots were regenerated; Qwen3-8B and Gemma Math roots reused saved token IDs. These replace historical first-8K rescoring values; they are not separate generation runs at every budget.

Table 13: Gemma4-12B-it: paired ME contrasts in endpoint potential (percentage points).

| Panel | Comparison | Difference | 95% CI |
|---|---|---|---|
| Clean 8K | $\mathrm{PSP}_{\mathrm{ME}} - \mathrm{GRPO}$ | $+10.55$ | $[+7.30, +13.87]$ |
| Clean 8K | $\mathrm{PSP}_{\mathrm{ME}} - \mathrm{PSP}_{\mathrm{E}}$ | $-2.81$ | $[-5.08, -0.59]$ |
| Math-7 2K | $\mathrm{PSP}_{\mathrm{ME}} - \mathrm{GRPO}$ | $+9.99$ | $[+8.64, +11.40]$ |
| Math-7 2K | $\mathrm{PSP}_{\mathrm{ME}} - \mathrm{PSP}_{\mathrm{E}}$ | $+5.72$ | $[+4.58, +6.92]$ |
| Math-7 4K | $\mathrm{PSP}_{\mathrm{ME}} - \mathrm{GRPO}$ | $+7.95$ | $[+6.10, +9.84]$ |
| Math-7 4K | $\mathrm{PSP}_{\mathrm{ME}} - \mathrm{PSP}_{\mathrm{E}}$ | $+3.59$ | $[+2.56, +4.65]$ |
| Math-7 8K | $\mathrm{PSP}_{\mathrm{ME}} - \mathrm{GRPO}$ | $+14.32$ | $[+12.36, +16.31]$ |
| Math-7 8K | $\mathrm{PSP}_{\mathrm{ME}} - \mathrm{PSP}_{\mathrm{E}}$ | $+0.53$ | $[-0.52, +1.58]$ |

100,000 paired question-bootstrap draws, seed 43; Math-7 equally weights datasets. Both ME comparisons are retained at every reported budget, irrespective of sign or significance. Intervals exclude training-seed uncertainty and are not multiplicity-adjusted.

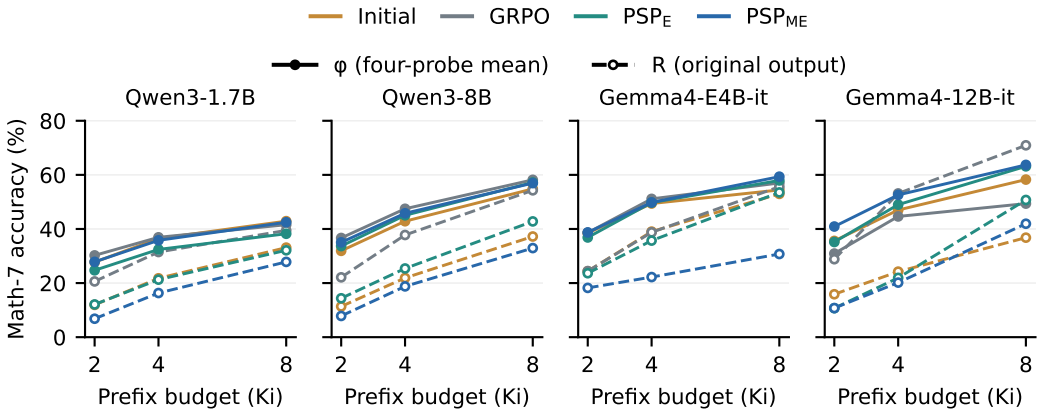


Figure 5: **Natural completion and forced-answer potential on the same trajectories.** Colors identify methods; solid curves show four-probe mean potential, dashed curves score original output. Budgets are 2,048/4,096/8,192 original-output tokens, before model-specific prefix processing; each probe adds up to 256 answer tokens. These are not independent rollouts at each budget or equal-total-compute comparisons. Math-7 weights seven datasets equally; $R$ uses MathEqual and probes use DeepScaler, so vertical gaps include verifier differences.

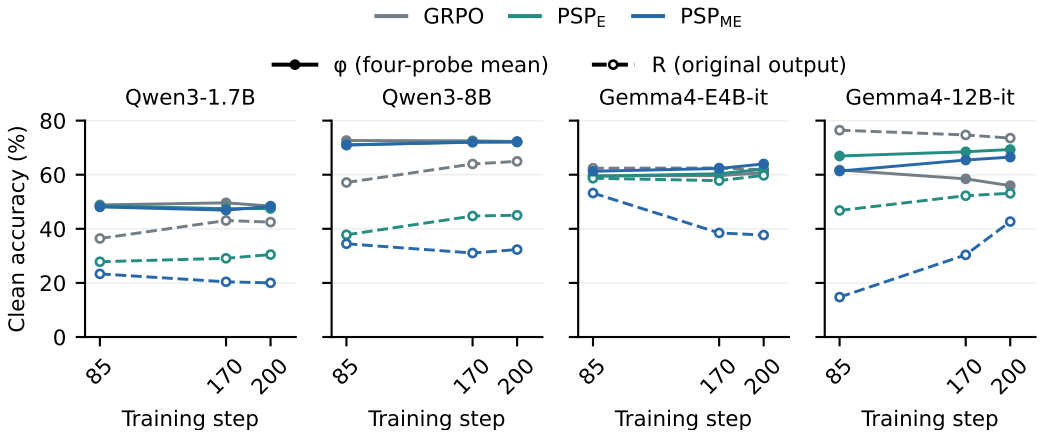


Figure 6: **Natural completion and endpoint potential across training checkpoints.** Each panel uses the separate Clean set of 256 questions with four native 8K responses per checkpoint. Colors identify methods; solid curves show four-probe mean potential, dashed curves show natural-answer $R$, both scored with DeepScaler on paired original trajectories. Each probe adds up to 256 tokens. Points are measured at steps 85/170/200; connecting lines show descriptive checkpoint trends, not intermediate measurements or uncertainty bands. Initial models are omitted because the horizontal axis denotes training checkpoints; their measured 8K baselines are reported in Table 3.

# D MECHANISM, TRAINING BEHAVIOR, AND COST

## D.1 TRAINING BEHAVIOR AND GROUP RECOVERY

Table 14 compares the first and last 20 steps on each run's own trajectories; Figures 7 and 8 show the Qwen3-1.7B and Gemma4-E4B-it histories. ME generally lengthens responses and increases cap hits while GRPO improves natural-answer correctness within the training cap. These are changing on-policy distributions, not held-out or causal comparisons. Cap hit means valid length equals 8,192, not an independently classified termination reason.

For Gemma4-E4B-it ME, training score rises from 40.81% to 66.28%, midpoint success from 12.77% to 60.76%, and endpoint success from 68.85% to 71.80%, while natural correctness falls. Table 15

also shows that E can improve long-cap performance while reducing group activity: its 453 recovered groups are outweighed by 710 new ties. Reward discrimination is therefore not necessary for a held-out gain. Recorded stability guards pass; absent fields and unavailable online extra-model audits are not treated as measured successes. The extended evidence retains detailed execution checks.

Table 14: Training behavior: first 20 → last 20 steps.

| Method | $R$ (%) | Mean root tokens | Cap hit (%) |
|---|---|---|---|
| *Qwen3-1.7B* | | | |
| GRPO | 33.93 → 52.32 | 7,154 → 5,691 | 62.54 → 33.67 |
| $\mathrm{PSP_E}$ | 33.75 → 38.03 | 7,131 → 7,014 | 62.19 → 58.26 |
| $\mathrm{PSP_{ME}}$ | 33.89 → 19.65 | 7,145 → 7,796 | 62.52 → 79.18 |
| *Qwen3-8B* | | | |
| GRPO | 37.87 → 71.70 | 7,189 → 5,157 | 60.80 → 21.09 |
| $\mathrm{PSP_E}$ | 38.81 → 47.54 | 7,189 → 6,854 | 59.92 → 51.13 |
| $\mathrm{PSP_{ME}}$ | 38.71 → 26.48 | 7,194 → 7,658 | 59.86 → 73.32 |
| *Gemma4-E4B-it* | | | |
| GRPO | 65.64 → 69.34 | 4,372 → 4,540 | 10.92 → 10.59 |
| $\mathrm{PSP_E}$ | 64.88 → 64.90 | 4,370 → 5,241 | 10.41 → 23.65 |
| $\mathrm{PSP_{ME}}$ | 64.77 → 27.46 | 4,369 → 7,526 | 10.70 → 70.76 |
| *Gemma4-12B-it* | | | |
| GRPO | 39.88 → 83.30 | 6,862 → 3,639 | 58.34 → 3.81 |
| $\mathrm{PSP_E}$ | 38.28 → 63.95 | 6,920 → 6,402 | 60.02 → 33.16 |
| $\mathrm{PSP_{ME}}$ | 37.66 → 45.20 | 6,958 → 7,020 | 60.84 → 53.83 |

R verifies original responses; cap hit is the fraction reaching the 8,192-token training output limit. Each entry compares steps 1–20 with 181–200 on the run's own trajectories, not a fixed held-out sample. Gemma4-12B-it follows final-checkpoint ancestry and excludes the superseded branch (Appendix D.2).

Table 15: Same-trajectory reward discrimination across models.

| Method | Active R | Active S | Recover 0 | Recover 1 | New ties |
|---|---|---|---|---|---|
| *Qwen3-8B* | | | | | |
| GRPO | 2124 | 2124 | 0 | 0 | 0 |
| $\mathrm{PSP_E}$ | 2149 | 2275 | 1481 | 22 | 1377 |
| $\mathrm{PSP_{ME}}$ | 2019 | 3873 | 2278 | 375 | 799 |
| *Gemma4-E4B-it* | | | | | |
| GRPO | 2592 | 2592 | 0 | 0 | 0 |
| $\mathrm{PSP_E}$ | 2864 | 2607 | 334 | 119 | 710 |
| $\mathrm{PSP_{ME}}$ | 2735 | 4351 | 1306 | 937 | 627 |
| *Gemma4-12B-it* | | | | | |
| GRPO | 1938 | 1938 | 0 | 0 | 0 |
| $\mathrm{PSP_E}$ | 2781 | 3776 | 975 | 982 | 962 |
| $\mathrm{PSP_{ME}}$ | 2530 | 5780 | 2559 | 795 | 104 |

Each run has 6,400 group occurrences; Active S = Active R + Recover 0 + Recover 1 − New ties. Rows rescore their own trajectories, so across-method differences are not shared-trajectory causal effects. Qwen3-1.7B appears in Figure 2.

### D.2 Training ancestry and execution checks

GRPO and $\mathrm{PSP_{ME}}$ complete 200 steps from the starting model. $\mathrm{PSP_E}$ completes 169 steps before a malformed rollout at step 170: eight valid tokens are followed by padding and non-finite log probabilities. Strict guards reject the response. Training resumes from the complete step-145 model, optimizer, scheduler, RNG and data state, using a defensive KV-cache initialization patch on wake-up. This patch does not change the intended sampler, reward or PPO objective, but introduces an implementation difference; it is not a demonstrated root-cause repair. The final ancestry is original steps 1–145 followed by recovered steps 146–200. Original steps 146–169 are a discarded branch, retained only for historical diagnostics and consumed-cost accounting. Clean step 85 comes from the ancestor, whereas steps 170/200 come from recovery. The final processes have zero exit codes and complete exports; this does not erase the earlier failure.

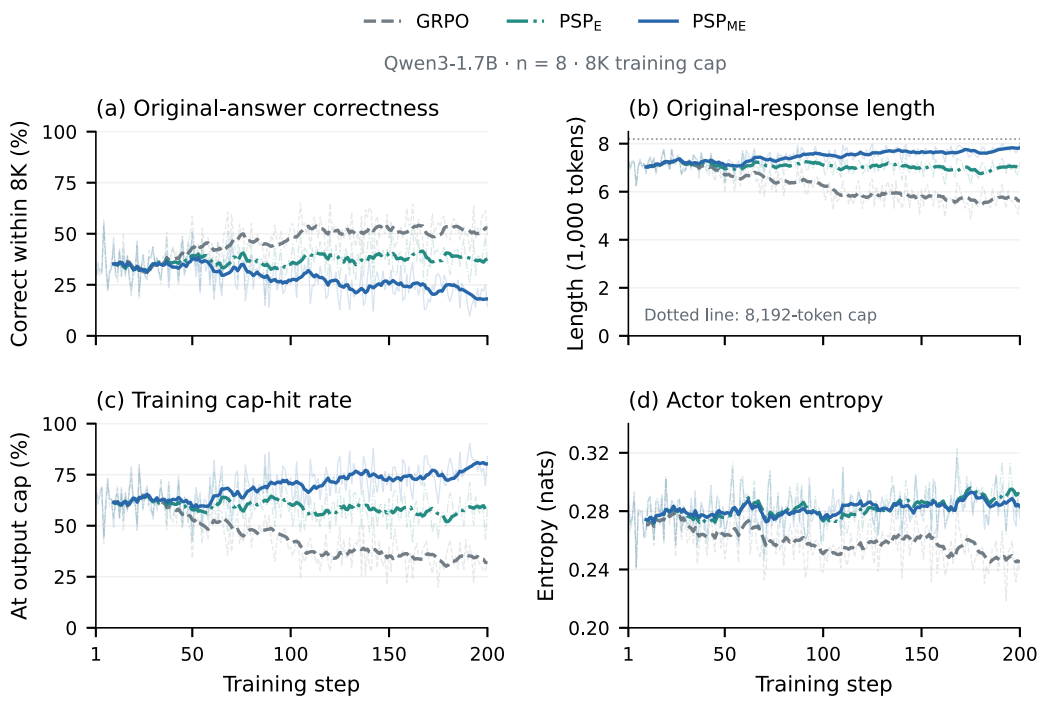


Figure 7: **Qwen3-1.7B original-response behavior during training.** The three methods follow the shared setup. Panels show (a) verifier accuracy of original training responses generated within the 8,192-token output budget (reasoning plus answer), including capped responses in the denominator, (b) mean original-response length including reasoning and answer, (c) the fraction reaching the 8,192-token training cap, and (d) pre-update actor vocabulary entropy in nats, averaged over valid response tokens within each sequence and then across sequences. Panel (a) measures correctness within this training budget, not probe success or long-cap Math-7 test accuracy. Prompt and probe tokens are excluded from panels (b)–(d). Faint lines show the full training history; dark lines are trailing ten-step means beginning at step 10, not confidence bands. Each method is measured on its own trajectories; entropy is neither a shared-context comparison nor a measure of reasoning quality.

The 600 retained Gemma4-12B-it records have finite metrics and zero recorded mask, verifier, leakage, policy-version, artificial-token and probe-training violations. The online extra-model three-way raw audit is unavailable. These checks neither establish numerical equivalence nor eliminate the observed generation fault. GRPO and $\mathrm{PSP}_{\mathrm{ME}}$ complete the original 200-step path; no checkpoint is selected by test performance.

### D.3 Training and evaluation accounting

**Qwen3-1.7B cost accounting.** Training-step GPU-hours equal the sum of step durations multiplied by four GPUs. They exclude startup, checkpoint export, separate evaluation, and other work outside the recorded timers. The Qwen3-1.7B cost audit covers eight core runs, seven $n = 4$ extension/control runs, and six single-trajectory runs: 21 configurations and 4,200 training steps. It excludes other ablations, failed/restarted jobs, engineering tests, OPD, and the rest of the project. Different submission batches can introduce system effects in timing comparisons. Logical probe input lengths count context presented to probes before cache reuse and are not prefill FLOPs. H200 hours are reported separately and are not normalized to H100 hours.

The same-trajectory $R/\widehat{\phi}_{100}$ curves in Figure 5 are now available. Independently generated natural-answer budget sweeps would instead hold questions, repeats, prompts, sampler, tokenizer and verifier fixed while changing the output cap. Existing 38,912-token competition results are not 32,768-token points. Neither prefix probing nor observed lower-cost final training configurations establishes matched actual test computation or learning curves at prespecified cumulative training budgets.

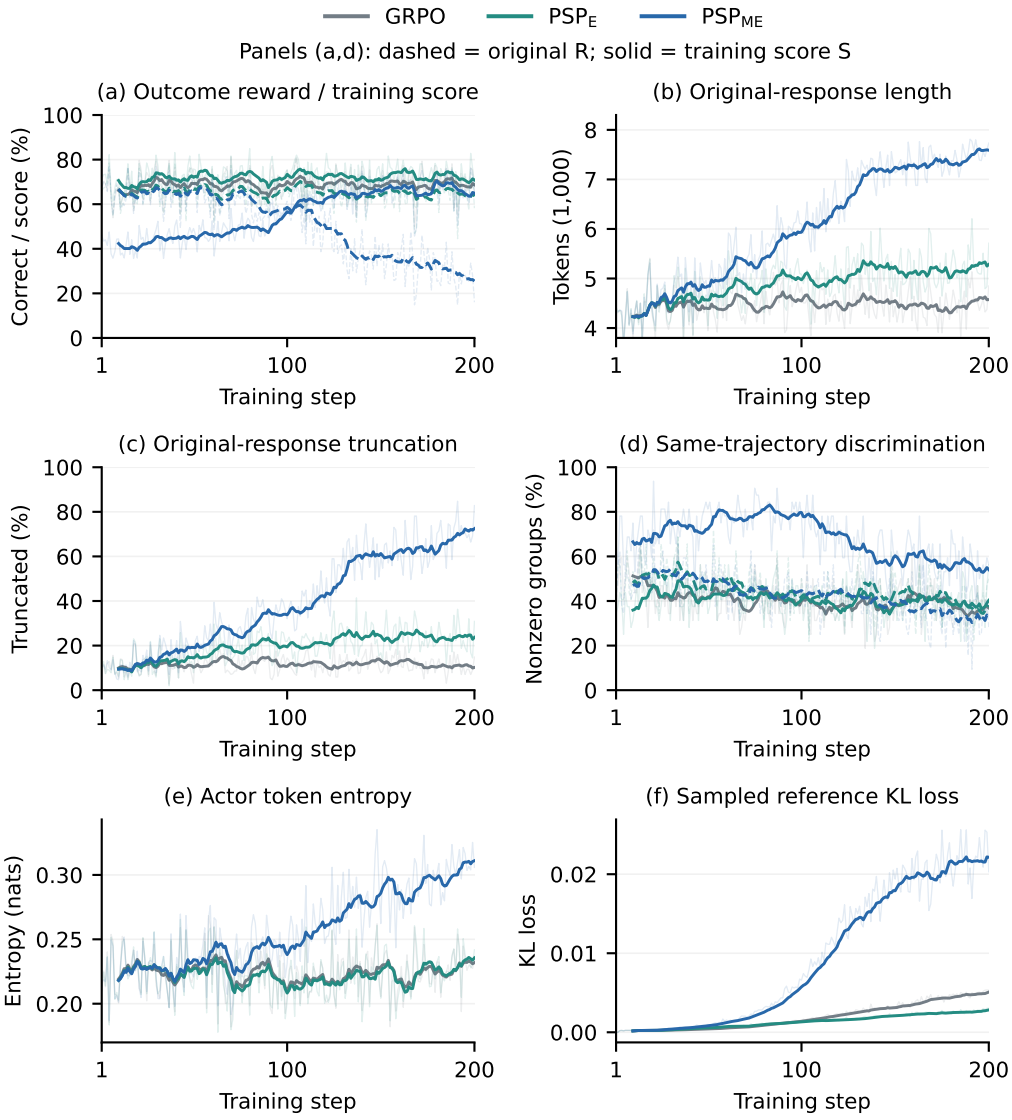


Figure 8: **Gemma4-E4B-it behavior throughout training.** Original-answer correctness and the actual training score are distinct measurements on each run's own responses. Other panels show original-response length, truncation, same-trajectory R/S group discrimination, actor entropy, and sampled reference KL loss. Faint lines are raw steps; dark lines are trailing ten-step means, not confidence bands. All methods retain padding and use the same static batching settings. These observations do not identify the cause of held-out gains.

Evaluation allocations are separate from training-step costs: Qwen3-8B's original/follow-up jobs consume 184.94/31.38 H100 GPU-hours, and Gemma4-12B-it's prefix follow-up consumes 46.196 H100 GPU-hours. These scopes exclude reused generation and other historical or failed work as detailed in the extended evidence; they are not full-project totals.

### D.4 Test-generation lengths and truncation

Table 17 summarizes mean output length and truncation across all four models. Initial values have no training cap. Longer test outputs are a cost and do not establish more efficient inference.

Table 16: Observed training costs under the available accounting scopes.

| Method / branch | Cap | Root M | Probe M | Input B | GPU-h |
|---|---|---|---|---|---|
| *Qwen3-1.7B (H100 ×4)* | | | | | |
| GRPO | 4K | 182.949 | 0.000 | 0.000 | 16.539 |
| $\mathrm{PSP_E}$ | 4K | 203.769 | 13.698 | 0.409 | 20.688 |
| $\mathrm{PSP_{ME}}$ | 4K | 206.423 | 33.199 | 0.634 | 21.462 |
| GRPO | 8K | 320.804 | 0.000 | 0.000 | 33.163 |
| $\mathrm{PSP_E}$ | 8K | 361.482 | 11.376 | 0.698 | 38.048 |
| $\mathrm{PSP_{ME}}$ | 8K | 382.739 | 29.491 | 1.132 | 43.855 |
| $\mathrm{PSP_E}$ ($m = 1$) | 8K | 361.770 | 6.015 | 0.349 | 39.630 |
| $\mathrm{PSP_E}$ ($m = 4$) | 8K | 357.817 | 20.857 | 1.379 | 40.833 |
| *Qwen3-8B (H200 ×8)* | | | | | |
| GRPO | 8K | 304.409 | 0.000 | 0.000 | 65.260 |
| $\mathrm{PSP_E}$ | 8K | 361.443 | 22.842 | 0.694 | 78.098 |
| $\mathrm{PSP_{ME}}$ | 8K | 382.124 | 44.118 | 1.125 | 86.123 |
| *Gemma4-12B-it (H200 ×8)* | | | | | |
| GRPO | 8K | 205.957 | 0.000 | 0.000 | 118.539 |
| $\mathrm{PSP_E}$ | 8K | 328.086 | 4.011 | 0.620 | 177.025 |
| $\mathrm{PSP_E}$ discarded | 8K | 38.823 | 0.449 | 0.073 | 20.760 |
| $\mathrm{PSP_{ME}}$ | 8K | 378.686 | 13.499 | 1.124 | 207.838 |

M/B denote millions/billions of tokens. Root and Probe count generated outputs; Input is logical probe input before cache reuse, not FLOPs. GPU-hours cover timed training steps and exclude startup, export and evaluation. H200 and H100 hours are not equivalent. The 24 discarded Gemma4-12B-it steps 146–169 are additional consumed compute, excluded from the final-ancestry rows; partial failures and complete job overhead remain unaccounted. No comparable Gemma4-E4B-it training-cost table is available.

Table 17: Long-cap generation length and truncation across models.

| Method | Training cap | Mean tokens | Test trunc. (%) |
|---|---|---|---|
| *Qwen3-1.7B* | | | |
| GRPO | 4K | 6,255 | 0.41 |
| $\mathrm{PSP_E}$ | 4K | 11,554 | 1.28 |
| $\mathrm{PSP_{ME}}$ | 4K | 14,531 | 2.70 |
| GRPO | 8K | 9,011 | 0.89 |
| $\mathrm{PSP_E}$ | 8K | 13,034 | 1.56 |
| $\mathrm{PSP_{ME}}$ | 8K | 16,103 | 3.02 |
| $\mathrm{PSP_E}$ ($m = 1$) | 8K | 13,158 | 1.73 |
| $\mathrm{PSP_E}$ ($m = 4$) | 8K | 11,629 | 1.12 |
| *Qwen3-8B* | | | |
| Initial | n/a | 14,353 | 4.34 |
| GRPO | 8K | 7,650 | 0.31 |
| $\mathrm{PSP_E}$ | 8K | 12,717 | 3.04 |
| $\mathrm{PSP_{ME}}$ | 8K | 14,727 | 2.95 |
| *Gemma4-E4B-it* | | | |
| Initial | n/a | 4,981 | 0.00 |
| GRPO | 8K | 5,164 | 0.00 |
| $\mathrm{PSP_E}$ | 8K | 7,005 | 0.00 |
| $\mathrm{PSP_{ME}}$ | 8K | 11,991 | 0.01 |
| *Gemma4-12B-it* | | | |
| Initial | n/a | 14,120 | 7.01 |
| GRPO | 8K | 4,559 | 1.01 |
| $\mathrm{PSP_E}$ | 8K | 9,352 | 2.00 |
| $\mathrm{PSP_{ME}}$ | 8K | 12,480 | 3.42 |

Means and truncation rates macro-average seven datasets and include incorrect and truncated responses. Initial is pre-RL, so its training cap is not applicable. These means differ from response-weighted totals across all 8,635 test replies. Qwen3-8B paired 8K R values are in Table 3.

### D.5 FIXED-PREFIX EVALUATION COST

Table 18 reports response-weighted lengths, which differ from dataset-macro accuracies. All generated probe tokens count, including text after a correct box. Gemma4-E4B-it ME probes hit the 256-token cap in 93.65% of Clean and 82.12% of Math-7 cases; cap hit is not equivalent to an incorrect answer. Logical input lengths do not measure uncached prefill, FLOPs or latency. Original generation is shared across budgets; larger prefixes and extra outputs do not establish lower actual test cost. Full method/budget and cap-hit records are retained in the extended evidence.

Table 18: Actual retained prefix and four-probe output at the 8K evaluation budget.

| Model | Reward | Clean prefix | Clean output | Math prefix | Math output |
|---|---|---|---|---|---|
| Qwen3-1.7B | GRPO | 4,817 | 240 | 5,612 | 296 |
| Qwen3-1.7B | $\mathrm{PSP}_{\mathrm{ME}}$ | 7,345 | 598 | 7,034 | 462 |
| Qwen3-8B | GRPO | 4,076 | 824 | 5,040 | 727 |
| Qwen3-8B | $\mathrm{PSP}_{\mathrm{ME}}$ | 6,953 | 863 | 6,885 | 831 |
| Gemma4-E4B-it | GRPO | 2,696 | 855 | 3,988 | 563 |
| Gemma4-E4B-it | $\mathrm{PSP}_{\mathrm{ME}}$ | 6,301 | 961 | 6,541 | 858 |
| Gemma4-12B-it | GRPO | 2693.5 | 293.7 | 3357.6 | 330.3 |
| Gemma4-12B-it | $\mathrm{PSP}_{\mathrm{ME}}$ | 6571.5 | 164.5 | 6432.2 | 222.1 |

Token means weight original responses, unlike question/dataset-weighted accuracy. Output sums four probes. Prefix lengths exclude boundary-removed content; repeated prefill and original generation are not included in output.

### D.6 FULL CONTINUATIONS, SAMPLING, AND RESCORING

Continuation-based estimation generates $\sum_k \sum_{j=1}^{m_k} \ell_{k,j}$ auxiliary tokens, where $\ell_{k,j}$ is continuation length; default ME caps auxiliary output at $4 \times 256 = 1{,}024$ tokens per original response. This excludes prefix processing and is not a total-compute or speedup claim. DAPO replaces tied groups, Prompt Replay changes question exposure, and PSP rescores original trajectories. Neither bounded probes nor nominal caps establishes superiority: matched comparisons must count original and auxiliary outputs, discarded candidates, prefix processing and optimization on matched hardware, models, data and evaluation. Both VinePPO and DAPO can improve efficiency despite additional sampling (Kazemnejad et al., 2025; Yu et al., 2025).

## E QWEN3-1.7B ABLATIONS AND BUDGET EXTENSIONS

### E.1 OBSERVED TRAINING-BUDGET COMPARISONS

Table 19 reports actual costs for all standard $n = 8$/$n = 4$, 4K/8K configurations, including outcome controls and the matched four-probe endpoint control. The reference is GRPO $n = 8$, 8K. ME $n = 4$, 8K retains 50.37% Math-7 at approximately 64% of reference generation and 76% of timed GPU-hours; $n = 4$, 4K reaches 49.02% at 37.36% and 41.18%, respectively. These compare final configurations, not learning curves at matched cumulative computation. The nominal default probe reservation of 12.5% omits changes in original-response lengths and prefix processing; actual core ME generation/time overhead is 28.50%/32.24%.

These cost ratios apply only to Qwen3-1.7B on four H100 GPUs. Other models and hardware are not pooled into the normalization; test costs are separate.

### E.2 ADDITIONAL CONFIGURATIONS AND COMPLETE RESULTS

Table 20 retains additional $n = 8$ configurations and all $n = 4$ controls; single-trajectory results are summarized separately. The $n = 4$ runs retain the core Qwen3-1.7B model and DAPO17K data, problem batch, training duration, optimizer schedule, and whole-response loss (Appendix B.4). Thus $n = 4$ uses 25,600 original responses and 6,400 problem occurrences, versus 51,200 responses for $n = 8$. Reducing $n$ changes the advantage normalization and response exposure; these are configuration comparisons, not reward-only ablations at equal actual computation. The 4K $n = 4$ $\mathrm{PSP}_{\mathrm{E}}$ row uses the whole-response loss, not a reasoning-only mask.

Table 19: Qwen3-1.7B training-budget configurations and actual costs.

| Reward | $n$ | Cap | Math-7 | Generated M | GPU-h |
|---|---|---|---|---|---|
| GRPO | 8 | 4K | 42.40 | 182.949 | 16.539 |
| $\mathrm{PSP_E}$ | 8 | 4K | 47.98 | 217.467 | 20.688 |
| $\mathrm{PSP_{ME}}$ | 8 | 4K | 48.25 | 239.622 | 21.462 |
| GRPO | 8 | 8K | 47.17 | 320.804 | 33.163 |
| $\mathrm{PSP_E}$ | 8 | 8K | 49.80 | 372.858 | 38.048 |
| $\mathrm{PSP_{ME}}$ | 8 | 8K | 51.06 | 412.230 | 43.855 |
| GRPO | 4 | 8K | 47.36 | 167.001 | 22.023 |
| $\mathrm{PSP_E}$ | 4 | 8K | 50.41 | 187.911 | 23.819 |
| GRPO | 4 | 4K | 43.08 | 95.446 | 10.856 |
| $\mathrm{PSP_E}$ | 4 | 4K | 48.46 | 108.802 | 11.559 |
| $\mathrm{PSP_{ME}}$ | 4 | 8K | 50.37 | 205.048 | 25.065 |
| $\mathrm{PSP_E}$ control | 4 | 8K | 48.86 | 194.668 | 24.232 |
| $\mathrm{PSP_{ME}}$ | 4 | 4K | 49.02 | 119.848 | 13.655 |

All rows use H100×4. Generated M counts all original and probe outputs in millions, including unused control probes. GPU-hours cover timed training steps only. Rows include corresponding outcome controls; these are final configurations, not matched-compute learning curves.

At $n = 4$, $\mathrm{PSP_E}$ uses two endpoint probes per response. $\mathrm{PSP_{ME}}$ uses two probes at each of the midpoint and endpoint. Its matched endpoint control also executes all four probes, but only the endpoint pair enters its reward; we label it $\mathrm{PSP_E}$ control and exclude it from the main low-budget scatter. Its unused midpoint probes remain in all cost totals. Hardware assignments and training configurations are recorded for each run; all executed probes are included in the reported costs.

$\mathrm{PSP_E}$ $n = 4$ exceeds R $n = 4$ by 3.06 pp on Math-7, with a 100,000-resample paired question-bootstrap interval $[2.08, 4.07]$ pp. Repeated replies are averaged within question at each checkpoint before resampling. This interval compares the two $n = 4$ configurations. At $n = 4$ and an 8K cap, $\mathrm{PSP_{ME}}$ versus its endpoint control scores 50.37% versus 48.86%, but IFEval falls by 2.03 pp, exceeding the prespecified maximum acceptable decrease of 1.0 pp. HumanEval at 16,384 tokens is also lower (75.61% versus 81.10%). The mathematical gain is therefore not an overall-capability success.

Table 20: Qwen3-1.7B additional mathematical results (percent).

| Reward | $n$ | Cap | GSM8K | MATH500 | AIME24 | AIME25 | AIME26 | HMMT25 | HMMT26 | Math-7 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Additional core configurations ($n = 8$)* | | | | | | | | | | |
| GRPO | 8 | 4K | 88.48 | 88.80 | 32.50 | 25.16 | 25.52 | 15.73 | 20.64 | 42.40 |
| $\mathrm{PSP_E}$ | 8 | 4K | 89.39 | 89.60 | 43.96 | 32.24 | 32.19 | 21.67 | 26.80 | 47.98 |
| $\mathrm{PSP_{ME}}$ | 8 | 4K | 90.45 | 90.60 | 43.65 | 33.33 | 33.44 | 20.42 | 25.85 | 48.25 |
| $\mathrm{PSP_E}$ ($m = 1$) | 8 | 8K | 89.23 | 89.80 | 44.38 | 36.04 | 35.73 | 22.60 | 28.88 | 49.52 |
| $\mathrm{PSP_E}$ ($m = 4$) | 8 | 8K | 90.45 | 90.00 | 44.48 | 34.06 | 34.90 | 21.88 | 26.89 | 48.95 |
| *Reduced response count ($n = 4$)* | | | | | | | | | | |
| GRPO | 4 | 8K | 89.92 | 89.60 | 41.56 | 31.30 | 31.56 | 20.10 | 27.46 | 47.36 |
| $\mathrm{PSP_E}$ | 4 | 8K | 89.54 | 91.60 | 48.28 | 33.59 | 36.88 | 23.75 | 29.26 | 50.41 |
| GRPO | 4 | 4K | 89.23 | 88.80 | 32.92 | 26.30 | 26.25 | 17.50 | 20.55 | 43.08 |
| $\mathrm{PSP_E}$ | 4 | 4K | 90.22 | 90.00 | 43.12 | 31.77 | 33.65 | 22.08 | 28.41 | 48.46 |
| $\mathrm{PSP_{ME}}$ | 4 | 8K | 89.99 | 90.80 | 46.88 | 34.11 | 39.17 | 23.02 | 28.60 | 50.37 |
| $\mathrm{PSP_E}$ control | 4 | 8K | 90.30 | 90.00 | 44.27 | 34.17 | 34.38 | 22.40 | 26.52 | 48.86 |
| $\mathrm{PSP_{ME}}$ | 4 | 4K | 90.30 | 89.80 | 45.57 | 33.59 | 35.83 | 22.19 | 25.85 | 49.02 |

Caps are training output limits; the standard $n = 8$, 8K results are in Table 1.

At $n = 4$, 4K, $\mathrm{PSP_{ME}}$ minus GRPO is +5.94 pp (question-bootstrap 95% CI $[+4.54, +7.40]$ pp).

At $n = 4$, 4K, $\mathrm{PSP_{ME}}$ minus $\mathrm{PSP_E}$ is +0.55 pp (question-bootstrap 95% CI $[-0.36, +1.46]$ pp).

Table 21: Capability tradeoffs in the matched four-probe control (percent).

| Reward | $n$ | Cap | GPQA$^{\diamond}$ | IFEval | BFCL | GPQA$^{\mathrm{M}}$ | HumanEval |
|---|---|---|---|---|---|---|---|
| $\mathrm{PSP}_{\mathrm{ME}}$ | 4 | 8K | 38.54 | 69.13 | 97.00 | 35.94 | 75.61 |
| $\mathrm{PSP}_{\mathrm{E}}$ control | 4 | 8K | 37.88 | 71.16 | 93.00 | 36.61 | 81.10 |

Both configurations execute four probes; the control uses only its endpoint pair for reward. HumanEval uses 16,384 output tokens. The IFEval decrease exceeds the prespecified 1.0 pp limit.

### E.3 Single-trajectory extension with a fixed baseline

Standard within-group centering gives no task advantage with one response and does not supply a single-trajectory instance of Equation 1. These historical runs instead assign the fixed-baseline scalar advantage $A = R - 0.5$ or $A = \widehat{\Phi}_4^{(1.0)} - 0.5$ to every original response token, with the same clipped PPO objective and separate KL regularization. The endpoint score averages four short-answer probes at the end of the actual reasoning. This tests endpoint potential, not the two-position $\mathrm{PSP}_{\mathrm{ME}}$ reward; the group-tie analysis is not applied to this estimator.

Each run uses 204 problems per step, one original response per problem, and 200 steps: 40,800 problem occurrences and original responses. Both the outcome control and endpoint-potential run execute four probes at each of the midpoint and endpoint (eight probes per response, each capped at 256 tokens), even though the midpoint probes do not enter either scalar reward. These diagnostic probes are retained in the cost accounting. Data, optimizer schedule, whole-response loss, and the fixed baseline are matched within each length pair; problem exposure differs from the core GRPO experiments.

Table 22: Fixed-baseline single-trajectory results and costs.

| Reward | Cap | Math-7 | $\Delta$ pp | Root M | Probe M | GPU-h | GPU |
|---|---|---|---|---|---|---|---|
| R scalar | 8K | 45.74 | — | 262.543 | 30.770 | 31.581 | H100 |
| Endpoint | 8K | 48.51 | +2.76 | 274.158 | 34.563 | 32.925 | H100 |
| R scalar | 16K | 46.98 | — | 382.203 | 33.168 | 60.876 | H100 |
| Endpoint | 16K | 48.71 | +1.73 | 390.499 | 33.717 | 62.383 | H100 |
| R scalar | 32K | 48.32 | — | 458.883 | 37.497 | 81.412 | H200 |
| Endpoint | 32K | 50.10 | +1.79 | 457.325 | 36.475 | 78.506 | H200 |

Deltas use unrounded scores and same-cap outcome controls. Both arms execute eight probes per response. GPU-hours cover training steps; H200 and H100 hours are not equivalent.

At 8K, endpoint potential gains 2.76 pp over its scalar control, while actual generation rises 5.25% and step time 4.25%; the generation increase exceeds the prespecified 2% limit. Because both arms execute probes, these increments do not measure the full overhead over outcome-only training. Its generation/time are 96.2%/99.3% of the core $n = 8$, 8K GRPO reference, not an eightfold cost reduction. The 16K/32K runs are length extensions, with actual generation 132.2%/153.9% of that reference; H200 time is not treated as H100-equivalent. Complete per-task, length, truncation and Clean results are retained in the extended evidence.

## F Shared-trajectory token-signal diagnostic

### F.1 Frozen sample and reward recovery

We compare GRPO, $\mathrm{PSP}_{\mathrm{ME}}$, and OPD on the same frozen Qwen3-1.7B starting student, using a frozen Qwen3-8B teacher for OPD. No parameters are updated. We sample 64 questions (seed 20260920) from 17,183 whitespace-deduplicated DAPO questions after excluding five conflicting-answer groups. Each question has eight shared 8K-capped responses and two 256-token probes at each midpoint and endpoint, following Appendix B. All 512 original responses, including 333 capped responses, remain in the analysis. Probe tokens are excluded from losses. The sample contains 3,706,880 original and 274,178 probe tokens. The archived seed mapping reuses 68 seeds between roots and probes of different questions; full-panel independent regeneration is not claimed.

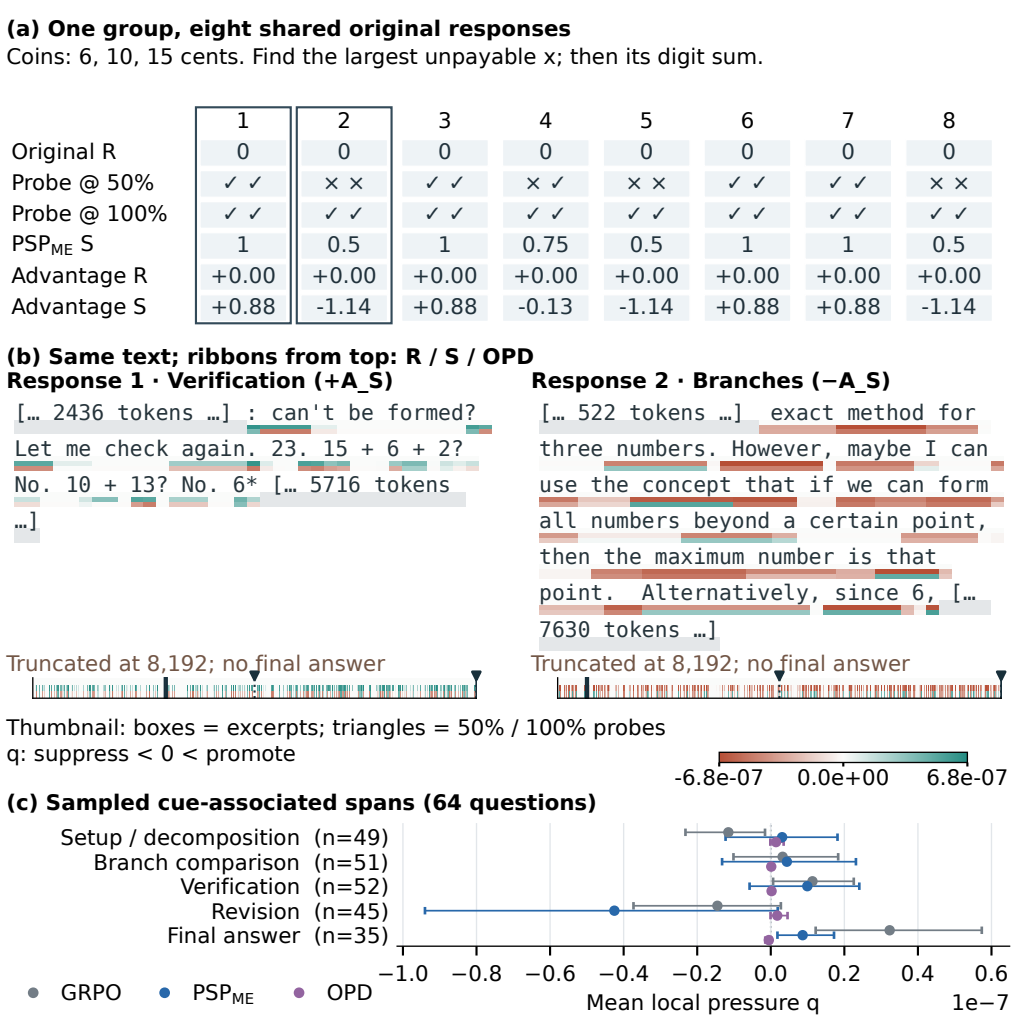


Figure 9: **From trajectory rewards to local token signals.** (a) A fixed-rule zero-reward recovery case. (b) Original text with R/$\mathrm{PSP}_{\mathrm{ME}}$/OPD ribbons: teal promotes and orange suppresses the sampled token. One shared asinh scale saturates 0.614% of values. (c) Blindly reviewed spans: question-averaged means, bootstrap 95% intervals, and coverage $n$. Zero signals remain; absent categories are excluded. Selection, reduction, and paired comparisons are specified in Appendix F.

Original rewards distinguish 23/64 groups, versus 48/64 for $\mathrm{PSP}_{\mathrm{ME}}$: 18 all-wrong and eight all-correct groups recover, one mixed group becomes tied, and twelve all-wrong groups remain tied. These are frozen-sample counts, not training repetitions. In Figure 9, all eight originals have zero reward and every endpoint probe succeeds, but midpoint success distinguishes the displayed responses. Scores 1 and 0.5 give advantages $+0.883$ and $-1.135$ around a group mean of 0.78125. Both originals are truncated without final answers. Two probes per position give a noisy distinction, not a certificate of correct reasoning.

### F.2 Signal Definition and Interpretation

For vocabulary logits $z_{i,t}$, the displayed task signal is

$$q_{i,t}^{\mathrm{local}} = -\langle \nabla_{z_{i,t}} \log p(y_{i,t} \mid h_{i,t}), \nabla_{z_{i,t}} \mathcal{L}_{\mathrm{signal}} \rangle. \tag{22}$$

It is the local first-order log-probability change along the negative loss gradient, not an observed parameter update. GRPO and PSP use the policy term in Appendix B.4; each response shares one advantage, with probabilities and importance correction determining the within-response texture.

PSP can restore or reverse trajectory support but does not assign independent step credit. OPD differentiates full-vocabulary Jensen–Shannon divergence with mixture weight $1/2$ and a differentiable student mixture; teacher probabilities and sequence importance weights are fixed. All objectives use sequence-mean/token-mean reduction in logical batches of 256 responses. This OPD diagnostic does not reconstruct the historical training recipe. Reference and student coincide, so the independent KL signal is zero. The shared signed-asinh scale saturates $0.614\%$ of method-token values; its limit is $6.82 \times 10^{-7}$.

### F.3 Selection, controls, and category findings

The illustrated group is selected among all-wrong recovered groups by proximity to median response length, with a stable question-hash tie-break; all 18 eligible groups average 8,192 tokens, so the hash determines the case. Positive/negative examples are nearest their within-group median advantages. Excerpts start at the earliest accepted sampled reasoning span. Observed unrecovered, newly tied and originally active controls and complete texts are preserved in the extended evidence.

Fixed lexical rules propose setup, branching, verification and revision spans; one candidate per response/category is selected by stable span hash without outcomes or signals. Two AI reviewers inspect disjoint responses with outcomes, advantages and signals hidden. Rejected candidates are not replaced. Review and boundary checks retain 987 of 1,795 candidates, covering 49/51/52/45 questions for the four reasoning categories and 35 for final answers. Labels describe explicit expressions, not mathematical correctness.

Signals are averaged within spans, responses and covered questions, retaining actual zeros but excluding absent categories. Paired question bootstrap uses 10,000 draws (seed 20260920), conditional on the observed sample and annotations. All four reasoning-category $\mathrm{PSP}_{\mathrm{ME}}$–GRPO intervals include zero: the case does not establish preferential reinforcement of any category or changes in subsequent pattern frequency. On final-answer spans the difference is $-2.37 \times 10^{-8}$ (95% CI $[-4.46, -0.766] \times 10^{-8}$); both mean signals remain positive, so this does not establish answer suppression. Intervals exclude regeneration and annotation-choice uncertainty and are not multiplicity-adjusted. Finite differences and closed-form checks validate the signal; complete derivations, controls and audit records accompany the extended evidence.