EVERY FORK HAS A PRICE: THE PRECISION–COVERAGE TRADE-OFF IN ON-POLICY DISTILLATION OF REASONING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

How much of a teacher's reasoning diversity should a student keep? In on-policy distillation, a routing rule decides which tokens receive coverage transfer. The dominant answer, entropy-aware on-policy distillation (EOPD), applies forward KL wherever teacher entropy exceeds a threshold. We test whether smarter routing can do better. Five routing mechanisms of increasing sophistication are compared in a controlled design-space study under one strictly reproduced protocol (Qwen3-1.7B student, Qwen3-8B teacher, six math benchmarks, identical trajectory budgets). What survives is a principle: *coverage transfer is budget-limited, not routing-limited*. Mechanistically, gates conditioned on trajectory outcomes remove forward-KL mass precisely from low-success problems whose marginal Pass@k value is highest. Whole prompt groups lose all coverage signal at a rate matching a compounding miscoverage model ($(1-p)^n$). Budget-preserving reweighting cannot recover the coverage that closed prompt groups never receive. EOPD's crude threshold keeps the coverage lead because it spends the most: Pass@k follows the budget, not the gate. Three measurements make the principle concrete: a soft continuous entropy gate matches EOPD on macro Avg@8 within a ± 3-point equivalence margin across seeds (29.99 ± 0.46 vs. 30.15 ± 0.54) at zero extra rollout cost. A cover-then-sharpen curriculum turns the budget into a dial: a half-schedule cover holds coverage level with the full-budget endpoint over three seeds (50.21 ± 0.60 vs. 50.68 ± 0.28 Pass@8). The forward-KL coefficient alone moves macro Avg@8 monotonically up to the default dose ($28.54 \rightarrow 30.63$), confirming that outcomes track the budget, not the gate. Our code is available at the anonymous repository: `https://anonymous.4open.science/r/TSA-OPD-for-ICLR-2027-50BD`.

1 INTRODUCTION

Imagine a student learning competition math from a teacher. Midway through an AIME problem, the teacher's next-token distribution splits: factor the quadratic now or substitute a new variable first? Both roads lead to the answer. The teacher entertains rewordings and uncertain continuations, so high entropy alone does not certify useful diversity. Concentrating on the few most likely continuations buys precision but may discard useful alternatives. Preserving every plausible continuation spreads probability thin (Figure 1). Large language models solve such problems through long chains of thought (Wei et al., 2022). To make this affordable, the standard recipe distills a large teacher's reasoning into a smaller student (Ho et al., 2023). The strongest recipes are *on-policy*: the student is supervised along its own generated trajectories rather than the teacher's (Agarwal et al., 2024; Gu et al., 2024). Which reasoning forks are worth preserving, token by token, is the coverage question of on-policy distillation (OPD).

On-policy distillation transfers reasoning from a large teacher to a small student (Agarwal et al., 2024; Jin et al., 2026). The default objective, reverse Kullback–Leibler (KL), answers our question with *mode-seeking*: concentrate on the teacher's most probable path (Gu et al., 2024). Entropy-aware on-policy distillation (EOPD) answers it with a gate: wherever teacher entropy exceeds a threshold, switch to forward KL and keep all plausible branches. This rule yields consistent Pass@k gains on math benchmarks (Jin et al., 2026). But high entropy is a noisy signpost. The same

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.

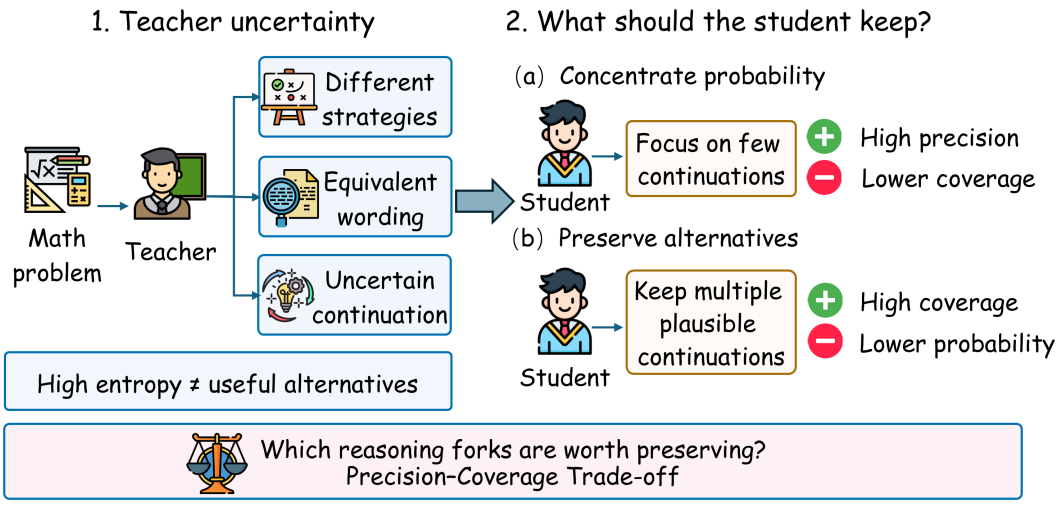

Figure 1: **Every fork has a price.** At a shared reasoning prefix, teacher uncertainty may reflect genuinely different reasoning strategies, semantically equivalent wording, or uncertain continuations. Concentrating the student's probability on a few continuations buys precision at the cost of coverage, while preserving more buys coverage at the cost of diluting it.

signature appears at useful reasoning forks (Wang et al., 2025), teacher confusion on wrong prefixes, equivalent surface phrasings, and flat tails from top-k truncation. EOPD's threshold rests on an unvalidated assumption: the *magnitude* of teacher uncertainty equals its *coverage value*. We test it directly. If outcome evidence carries coverage value that entropy misses, routing on it should strictly improve on entropy alone. Without that improvement, the limit lies somewhere other than the routing signal.

This paper follows the test to its logical end. We design five routing mechanisms of increasing sophistication. The family spans a soft entropy gate, divergence- and mass-aware routing, and trajectory-outcome utility conditioning. It adds branch-level value conditioning with actual teacher continuations and a posterior-smoothed utility whose weights are the exact marginal Pass@k gains. All five share one strictly reproduced protocol: EOPD's student, teacher, data, optimization budget, and evaluation pipeline. Figure 2 shows the resulting system: the on-policy loop, the routing design space over the gate g_t, and the cover-then-sharpen curriculum with its explicit budget dial. The test yields a clearer view of the gate's role. *Coverage transfer in on-policy distillation turns out to be budget-limited, not routing-limited.* EOPD's crude threshold keeps the coverage lead because it is the biggest spender. Every more sophisticated gate prunes or redistributes the forward-KL mass, and Pass@k follows the budget, not the cleverness. The mechanism is visible in the logs (Section 6). With $n{=}4$ rollouts per prompt, a problem with success rate p loses all coverage signal with probability $(1-p)^4$ at every step (28–49% of prompt groups in practice). These are exactly the problems with the largest marginal Pass@k gain $k(1-p)^{k-1}$.

The same measurements yield a relaxation on the other side of the trade-off. Our soft entropy gate is a continuous relaxation of EOPD's threshold. It matches EOPD on six-benchmark macro Avg@8 within a ± 3-point margin across seeds at *zero* extra inference cost. The two one-sided tests (TOST) run on the seed-mean contrast -0.16 (90% confidence interval (CI) $[-2.0, +1.6]$). Seed means are 29.99 ± 0.46 over three seeds vs. 30.15 ± 0.54 over four. Neither hyperparameter improves on the defaults within the swept ranges. Multi-seed evidence keeps every single-seed comparison honest. The seed-42 macro Avg@8 gains over OPD, which extend nominally to Pass@1–2, reverse sign significantly on seed 41 ($+2.09$ $[+0.9, +3.4]$ to -1.57 $[-3.1, -0.2]$). EOPD keeps the coverage lead at Pass@8, the one k where its four seeds all sit above the gate's three (Figure 3). Finally, must the two regimes be traded at all? The cover-then-sharpen curriculum runs EOPD's gate for a cover phase and ours for the rest, with the cover length as its only dial. A half-schedule cover (87 of 174 steps) holds coverage statistically level with the full-budget endpoint over seeds 41/42/43 (50.21 ± 0.60 vs. 50.68 ± 0.28 Pass@8). The gap sits inside EOPD's own seed range. Its precision stays inside EOPD's seed envelope (29.68 ± 0.71 vs. 30.15 ± 0.54 macro Avg@8). Shortening the cover moves the operating point along the same trade-off ($29.23/49.80$ at 116 steps, $29.98/47.31$ at 58). Sequencing shifts the operating point predictably. Kept coverage recovers with the budget spent

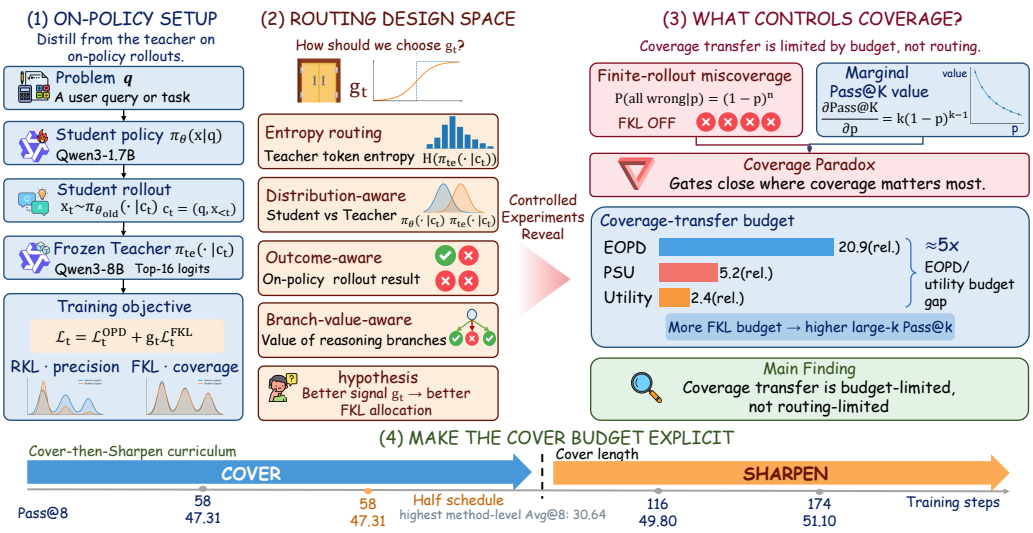

Figure 2: **The token-state adaptive (TSA) OPD framework.** (1) On-policy setup: student rollouts, frozen-teacher top-16 logits, and the reverse-KL (RKL) surrogate plus gated forward-KL update (RKL buys precision, FKL buys coverage). (2) The routing design space over the gate g_t. (3) Controlled experiments: gates close where the marginal Pass@k value is highest, and outcomes track the coverage budget, not the gate. (4) The cover-then-sharpen curriculum makes the cover budget an explicit dial. Budget bars (3): cumulative FKL mass, gated share $\times$ steps (rel., $12.0\% \times 174{=}20.9$, $4.2\% \times 58{=}2.4$).

across four dose points and flattens near the full budget, with one single-seed inversion between adjacent doses.

Contributions. (1) **A controlled design-space study** of five routing mechanisms for coverage transfer, under a strictly reproduced EOPD protocol with identical trajectory budgets. (2) **A design rule with a mechanism**: coverage transfer is budget-limited. We supply the $(1-p)^n$ miscoverage model, per-step logs showing 28–49% de-coverage, and budget accounting for every variant. A forward-KL (FKL) coefficient dose–response shows macro Avg@8 rising monotonically with spend at fixed gate shape, up to the default dose. (3) **A budget dial**: the soft entropy gate matches EOPD on precision within a ± 3-point equivalence margin (TOST) at EOPD-identical cost. Neither knob improves on its default in our sweeps. A cover-then-sharpen curriculum has one knob, the cover length. It holds coverage level with the full-budget endpoint over three seeds (50.21 ± 0.60 vs. 50.68 ± 0.28 Pass@8) at precision within EOPD's seed envelope. This turns the deployment sampling budget into a cover-budget choice. (4) **Boundary conditions**: a direction-consistent negative interaction between hard-data augmentation and FKL routing, and the limits of budget-preserving reweighting. Together, these turn a growing bag of routing heuristics into a measurable trade-off, pointing to sequence-level coverage objectives and phase-aware schedules as future directions.

2 Related Work

Each line optimizes one object, discards the other. We measure the latter.

Distillation and its divergences. Sequence-level knowledge distillation (KD) trains on teacher outputs (Kim & Rush, 2016; Hinton et al., 2015). On-policy variants supervise its rollouts, reducing exposure bias (Agarwal et al., 2024; Gu et al., 2024). The divergence is the lever: reverse KL mode-seeking (Gu et al., 2024), forward KL mode-covering. Skewed or balanced variants interpolate between them (Ko et al., 2024; 2025; Amara et al., 2022; Wen et al., 2023). Speculative KD bridges capability gaps via interleaved sequences (Xu et al., 2024). EOPD routes the divergence per token by teacher entropy (Jin et al., 2026), ToDi by token-level utilities in offline KD (Jung et al., 2025). Entropy-based routing draws on evidence that high-entropy minority tokens drive effective reinforcement learning (RL) for reasoning (Wang et al., 2025; Cui et al., 2025; Yu et al., 2025; Cheng et al., 2025). *All these methods optimize per-token or per-sequence objectives. None controls the*

shape of the student's sampling distribution, e.g., how many distinct solution modes it retains. We caution: gating on high entropy conflates useful forks with teacher confusion. Outcome-evidence gating can remove coverage where its marginal value is highest.

Coverage, sampling budgets, and curricula. RL post-training sharpens rather than expands the reasoning distribution. RL with verifiable rewards (RLVR) wins at small sampling budgets, trails the base model at large k (Yue et al., 2025; Wu et al., 2026). RL from human feedback (RLHF) trades diversity for reward (Kirk et al., 2024). Remedies reinforce diversity (Li et al., 2025) or preserve candidate support (Peng et al., 2026). Inference scaling is predictable (Brown et al., 2024; Snell et al., 2024; Wu et al., 2025; Muennighoff et al., 2025). Some RL works optimize Pass@k or best-of-N (Chow et al., 2025; Tang et al., 2025; Walder & Karkhanis, 2026; Bagirov et al., 2025; Chen et al., 2025). Pass@k as a metric is debated (Yu, 2025). Curricula improve RL boundary expansion (Parashar et al., 2026; Cai et al., 2026) and progressive distillation (Liu & Zhang, 2025; Feng et al., 2024). Prolonged RL expands the boundary (Liu et al., 2025). Hard or high-quality data reshape what can be learned (Ye et al., 2025; Moshkov et al., 2025; Yuan et al., 2023; Ma et al., 2026). *This line documents the precision–coverage crossover in RL but has no account of coverage transfer in distillation. No existing curriculum makes the exploration–consolidation phase boundary an explicit control.* Our cover-then-sharpen curriculum fills this intersection. It pairs budget-managed entropy-aware routing with a phase switch. Evaluation against a strictly reproduced EOPD baseline spans Pass@k to k=128.

3 Preliminaries: On-Policy Distillation and Entropy-Aware Routing

We use EOPD's notation (Jin et al., 2026) throughout. The student policy is $\pi_\theta(\cdot \mid \mathbf{c}_t)$ and the fixed teacher is $\pi_{\mathtt{te}}(\cdot \mid \mathbf{c}_t)$. Here $\mathbf{c}_t = (\mathbf{q}, x_{<t})$ is the context (prompt $\mathbf{q}$ plus previously generated tokens). $x_t \sim \pi_{\theta_{\text{old}}}(\cdot \mid \mathbf{c}_t)$ is drawn from a behavior copy of the student.

KL divergences. Forward KL $\mathrm{KL}(\pi_{\mathtt{te}} \| \pi_\theta)$ is *mode-covering*: it penalizes low probability on any token the teacher finds likely. Reverse KL $\mathrm{KL}(\pi_\theta \| \pi_{\mathtt{te}})$ is *mode-seeking*: it concentrates the student on high-likelihood teacher modes it visits (Minka, 2005).

On-policy distillation with clipped reverse KL. Following Lu & Lab (2025); Jin et al. (2026), the token-level reward is the teacher–student log-probability difference at the behavior policy,

$$\hat{A}_t = \log \pi_{\mathtt{te}}(x_t \mid \mathbf{c}_t) - \log \pi_{\theta_{\text{old}}}(x_t \mid \mathbf{c}_t), \tag{1}$$

and the student is updated by the proximal policy optimization (PPO)-clipped surrogate (Schulman et al., 2017) with importance ratio $r_t = \pi_\theta(x_t \mid \mathbf{c}_t)/\pi_{\theta_{\text{old}}}(x_t \mid \mathbf{c}_t)$:

$$\mathcal{L}_t^{\mathrm{OPD}}(\theta; \mathbf{c}_t) = \mathbb{E}_{x_t \sim \pi_{\theta_{\text{old}}}} \Big[\max\big(-r_t \hat{A}_t,\ -\mathrm{clip}(r_t, 1-\epsilon, 1+\epsilon)\hat{A}_t \big) \Big]. \tag{2}$$

Entropy-aware on-policy distillation (EOPD). EOPD augments equation 2 with a forward-KL term gated by the teacher's token entropy $H_t^{\mathtt{te}} = -\sum_{x \in \mathcal{V}} \pi_{\mathtt{te}}(x \mid \mathbf{c}_t) \log \pi_{\mathtt{te}}(x \mid \mathbf{c}_t)$:

$$\mathcal{L}_t^{\mathrm{EOPD}} = \mathcal{L}_t^{\mathrm{OPD}} + \alpha\, \mathbb{I}\big[H_t^{\mathtt{te}} > \tau\big]\ \mathcal{L}_t^{\mathrm{FKL}}, \qquad \tau = 0.8,\ \alpha = 1.0. \tag{3}$$

The FKL term is an expectation over the teacher's top-k set $\mathcal{S}_t^k$ ($k = 16$). The teacher is renormalized over $\mathcal{S}_t^k$. The student is evaluated at the same indices under its full-vocabulary distribution:

$$\mathcal{L}_t^{\mathrm{FKL}}(\theta; \mathbf{c}_t) = \sum_{x \in \mathcal{S}_t^k} \tilde{\pi}_{\mathtt{te}}(x \mid \mathbf{c}_t) \log \frac{\tilde{\pi}_{\mathtt{te}}(x \mid \mathbf{c}_t)}{\pi_\theta(x \mid \mathbf{c}_t)}, \qquad \tilde{\pi}_{\mathtt{te}}(x \mid \mathbf{c}_t) = \frac{\pi_{\mathtt{te}}(x \mid \mathbf{c}_t)}{\sum_{x' \in \mathcal{S}_t^k} \pi_{\mathtt{te}}(x' \mid \mathbf{c}_t)}. \tag{4}$$

Evaluation. We report Avg@8 (mean accuracy over 8 samples per problem) and Pass@k (unbiased estimator (Chen, 2021)). Benchmarks: MATH500 (Hendrycks et al., 2021), AMC23 (MAA, 2023), Minerva (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), AIME24 and AIME25 (MAA, 2025). This follows the Qwen2.5-Math pipeline (Yang et al., 2024) with decoding settings (temperature 1.0, top-p 0.8, max length 8192).

4 A DESIGN SPACE FOR TRAJECTORY-STATE-AWARE DISTILLATION

We organize our mechanisms as points in a routing design space

$$\mathcal{L}_t = \mathcal{L}_t^{\mathrm{OPD}} + g_t \, \mathcal{L}_t^{\mathrm{FKL}}, \tag{5}$$

where the gate g_t decides which tokens receive coverage transfer. All variants share EOPD's optimization and evaluation protocol (Section 5). Unless stated, gates are stop-gradient signals. The student cannot shrink the gate instead of matching the teacher. We write $H_t^{\mathtt{te}}$ for the teacher entropy and reuse $\tilde{\pi}_{\mathtt{te}}$ and $\mathcal{S}_t^k$ from equation 4: the teacher renormalized over its top-k set, and the set itself.

Soft entropy gate (TSA-soft-entropy). Replace EOPD's hard indicator with a continuous gate

$$g_t = \sigma\Big(\tfrac{H_t^{\mathtt{te}} - \tau}{\lambda}\Big), \qquad \tau = 0.8,\ \lambda = 0.15, \tag{6}$$

with constant $\alpha = 1.0$. Same rollout cost as EOPD (n=1 per prompt).

Utility- and value-conditioned variants. Four further arms condition routing on evidence beyond entropy (full definitions in Appendix A). Divergence- and mass-aware routing multiplies the entropy gate by student–teacher Jensen–Shannon (JS) divergence and top-k mass. The full stack adds a coefficient decay and a final-answer guard. Trajectory-utility conditioning samples n=4 rollouts per prompt and gates by the group success rate. It uses 32 prompts per step, the same 128 trajectories per step as the n=1 baselines over four times fewer distinct prompts. Thus correct-but-rare trajectories receive the strongest signal while all-wrong groups fall back to OPD. Branch-level value conditioning distills a value-tilted teacher scored by teacher continuations at one high-entropy state per hard problem. Two budget-aware variants, posterior-smoothed utility (PSU) and budget-preserving reweighting (eopd-value), hold the total FKL budget at EOPD's exactly. Matching the trajectory budget makes this family comparable to the n=1 arms. Its price is the fourfold drop in distinct prompts per step. Any grouped-rollout utility estimate pays it at fixed budget, not at inflated rollout cost. This creates a confound between the gate and reduced prompt diversity. Two grouped-rollout controls disentangle it (Sections 5.3 and 6).

Cover-then-sharpen curriculum (CTS). Run EOPD's hard-threshold routing for a cover phase (the half schedule, steps 1–87, by default) to build coverage. Then switch to the soft entropy gate equation 6 for the sharpen phase (steps 88–174), resuming from the step-87 checkpoint. The optimizer state, cosine learning-rate schedule, and data order continue across the switch.

5 EXPERIMENTS

5.1 SETUP

We strictly reproduce the EOPD protocol (Jin et al., 2026): Qwen3-1.7B-Base student (Yang et al., 2025), Qwen3-8B non-thinking teacher, MATH training set (Hendrycks et al., 2021). Further protocol details are in Appendix B. Baselines are our own reproductions, validated against the paper's record (macro Avg@8 29.32 vs. 31.22). EOPD-over-OPD gains match in direction and magnitude (per-benchmark audit in Appendix C). We do not train GKD (Agarwal et al., 2024), MiniLLM (Gu et al., 2024), or DistiLLM (Ko et al., 2024) under this protocol. Their recipes vary exactly the knobs this study holds fixed, and retraining would conflate the isolated routing variable. OPD (pure RKL) and EOPD (hard-gated FKL) are the endpoints of the routing axis. Every variant perturbs the recipe only at the routing level (reporting boundary and benchmark manifest in Appendix D).

5.2 MAIN RESULTS

Table 1 reports six-benchmark results with multi-seed evidence for the two protocol endpoints. EOPD's four seeds give macro Avg@8 30.15 ± 0.54 and OPD's three give 29.60 ± 1.12 (mean $\pm$ standard deviation across seeds). The seed-42 EOPD-over-OPD margin ($+0.77$) settles at $+0.55$ across seeds. Every single-seed ordering should be read against seed spreads of half a point (EOPD) to over a point (OPD). Against this baseline, the soft entropy gate does not separate from EOPD on precision (three seeds, 29.99 ± 0.46). Its seed-42 nominal Pass@5 lead does not survive EOPD's

Table 1: Six-benchmark results (Avg@8 / Pass@8), key rows. Repro = our strict reproduction of the EOPD protocol, and paper record = numbers reported by Jin et al. (2026). A bracketed 29.60 [28.54–31.15] denotes a seed mean with its seed span. For the CTS cover-87 row (seeds 41/42/43), per-benchmark cells show the seed mean and Macro the seed range. All per-seed rows and gate variants: Table 8 (Appendix E).

	MATH500		**AMC23**		**Minerva**		**OlympiadBench**	
Model	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8
Base (no distill)	50.08	82.40	25.31	65.00	13.37	37.50	20.61	46.81
OPD (repro, s42)	67.10	85.40	35.62	62.50	24.72	42.28	30.46	51.56
OPD (ours, 3-seed mean)	67.42	85.60	38.75	67.50	25.12	41.79	30.62	52.05
EOPD (repro, s42)	67.95	87.00	38.12	75.00	23.44	40.81	30.96	53.78
EOPD (ours, 4-seed mean)	67.96	87.40	39.92	75.00	24.46	41.91	31.15	53.96
EOPD (paper record)	68.73	87.60	41.88	75.00	30.15	50.74	30.28	51.11
TSA-soft-entropy (ours, s42)	68.20	86.40	42.81	75.00	24.91	41.18	30.78	54.07
TSA-soft-entropy (ours, 3-seed mean)	68.48	86.67	39.06	70.83	24.56	42.16	31.04	54.12
TSA-CTS cover-87 (ours)	68.15	86.40	42.81	70.00	24.26	42.65	31.13	54.52
TSA-CTS cover-87 (ours, 3-seed mean)	66.88	85.63	41.77	70.83	24.09	44.88	30.61	56.62

	AIME24		**AIME25**		**Macro**	
Model	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8
Base (no distill)	5.00	16.67	2.92	16.67	19.55	44.17
OPD (repro, s42)	7.50	23.33	5.83	26.67	28.54	48.62
OPD (ours, 3-seed mean)	9.44	20.00	6.25	24.44	29.60 [28.54–31.15]	48.56 [48.10–48.97]
EOPD (repro, s42)	10.83	30.00	4.58	20.00	29.32	51.10
EOPD (ours, 4-seed mean)	10.94	27.50	6.46	18.33	30.15 [29.32–30.73]	50.68 [50.35–51.10]
EOPD (paper record)	10.42	23.33	5.83	16.67	31.21	50.74
TSA-soft-entropy (ours, s42)	9.17	23.33	7.92	20.00	30.63	50.00
TSA-soft-entropy (ours, 3-seed mean)	10.28	22.22	6.53	20.00	29.99 [29.58–30.63]	49.33 [48.98–50.00]
TSA-CTS cover-87 (ours)	10.83	23.33	6.67	23.33	30.64	50.04
TSA-CTS cover-87 (ours, 3-seed mean)	8.05	21.10	6.66	22.21	29.68 [28.96–30.64]	50.21 [49.58–51.02]

own seed spread. EOPD keeps the best Pass@8 (50.68 ± 0.28 four-seed mean vs. 49.33 ± 0.47), the only k with separated seed ranges. The cover-then-sharpen curriculum (CTS) turns the cover budget into an explicit dial (Table 2). A half-schedule cover (87 steps) holds coverage statistically level with the full-budget endpoint over three seeds (50.21 ± 0.60 vs. 50.68 ± 0.28 Pass@8, TOST-equivalent at ± 2 points). Precision sits inside EOPD's seed envelope (29.68 ± 0.71 vs. 30.15 ± 0.54). Shorter covers move along the same trade-off (Section 5.4).

5.3 Ablation Studies

Gate composition. The composite TSA gates add the routing signals of Section 4 one at a time. They are teacher–student divergence (entropy-JS), top-k mass weighting (JS-mass), and the full stack with coefficient decay and the final-answer guard (TSA-full). Every addition costs coverage on the shared seed (Table 8, Figure 9). Macro Pass@8 decreases monotonically from 50.00 (soft gate) through 48.38 (entropy-JS) and 48.29 (JS-mass) to 47.13 (TSA-full). The JS-mass step (-0.09) is inside single-run noise. The signal is the cumulative 2.87-point drop. Macro Avg@8 stays inside a 1.2-point band, smaller than the seed spread.

Utility conditioning. Replacing the entropy signal with outcome evidence (trajectory-level utility, branch-level values, posterior-smoothed utility, and the value-estimator variant) reduces hard-benchmark Pass@8 at screening scale. In 58-step pilots the arms' hard-3 macro Pass@8 is 28.2–28.9, against 34.89 for the soft gate (mechanism in Section 6). The formal-scale grouped controls (Section 4, Appendix A) decompose the loss further. The prompt-diversity reduction alone accounts for about 1.4 coverage points (48.6 with the gate open vs. 50.00 at n=1). The $(1-p)^n$ gate closures explain the deeper screening-scale losses.

Data composition. Replacing 50% of the MATH training mix with decontaminated DeepMath-hard problems (He et al., 2025) interacts with routing rather than additively (Appendix F). The RKL-only OPD baseline gains +1.39 macro Avg@8 (28.54 $\rightarrow$ 29.93) at no Pass@8 cost, while both FKL-based methods lose coverage (EOPD -2.40, soft gate -2.58 Pass@8). Each interval

Table 2: **The two budget dials.** (a) FKL coefficient dose at fixed soft-gate shape (seed 42, formal 174 steps, six-benchmark macro): the gated token share is insensitive to α, so the dose varies the amplitude of spend, not its placement (α=1 token ratio from the seed-41 replica, Appendix G). (b) CTS cover-budget dose response: shortening the cover phase trades coverage for average precision. Dose points are single-seed except cover-87 (three seeds: 29.68 ± 0.71 / 50.21 ± 0.60).

(a) α dose

α	FKL token ratio	Avg@8	Pass@8
0 (OPD)	0%	28.54	48.62
0.25	13.7%	29.10	49.20
0.50	14.1%	29.80	49.50
1.00 (default)	13.5%	30.63	50.00
2.00	14.3%	30.20	50.90

(b) cover dose

Curriculum (cover → sharpen)	Avg@8	Pass@8	Pass@5
cover 58 / sharpen 116	29.98	47.31	44.04
cover 87 / sharpen 87	30.64	50.04	46.06
cover 116 / sharpen 58	29.23	49.80	45.65
EOPD (cover 174, no sharpen)	29.32	51.10	45.63

includes zero, but the directions agree. Harder prompts and FKL coverage appear to compete for the same student capacity, so the levers substitute rather than stack (Section 7).

5.4 Hyperparameter Analysis

The soft gate has two knobs: the temperature λ it introduces and the entropy threshold τ it inherits from EOPD. The curriculum adds a third structural knob, the cover-phase budget. Sweeping all three finds the same shape everywhere (Figure 8, Appendix F). No setting materially improves on the defaults. Deviations degrade coverage, precision, or both.

Gate temperature λ and threshold τ. Pilot-scale sweeps (58 steps, hard-benchmark macro, Tables 10–11) keep both defaults. No deviation from λ=0.15 or τ=0.8 improves hard-3 macro Pass@8. Nominal Avg@8 differences sit inside pilot noise (Appendix F). A formal-scale check at λ=0.20 (174 steps, seed 42) confirms it. The steeper gate lands 1.53 Avg@8 below the default ($[-2.9, -0.3]$ paired CI), with Pass@8 unchanged ($+0.11$ $[-3.4, +3.6]$).

Curriculum cover budget. At formal scale (174 steps, six-benchmark macro, seed 42, Table 2), coverage recovers steeply with the cover budget, then flattens. Cover 58 / sharpen 116 gives 29.98 Avg@8 / 47.31 Pass@8. Cover 87 / sharpen 87 gives 30.64 / 50.04. Cover 116 / sharpen 58 gives 29.23 / 49.80. EOPD (cover 174, no sharpen) gives 29.32 / 51.10. Cover 87 restores Pass@8 to level with the full budget on the three-seed mean and attains the highest Avg@8 of the dose points on the shared seed. Only the one-epoch cover separates on Pass@8 (-3.79 $[-7.4, -0.4]$). The remaining doses are single-seed. The 116-step point below the 87-step one on both axes is an inversion single runs cannot resolve against 0.5–1.1-point seed spreads. The resolvable structure is the steep climb from the one-epoch cover to the full budget.

FKL coefficient dose. Scaling the FKL coefficient at fixed gate shape should move the metrics with routing untouched. A five-point dose at seed 42 confirms it (Table 2). Macro Avg@8 climbs monotonically from 28.54 at α=0 (OPD) through 29.1 and 29.8 at α=0.25 and 0.5 to 30.63 at the default α=1.0. The gated token share stays at 13.5–14.3% across the nonzero doses. The amplitude of spend, not the set of tokens, carries the effect. Doubling past the default (α=2) moves the operating point along the trade-off rather than beyond it. Precision dips to 30.2 while Pass@8 rises to 50.9, most of the way to EOPD's 51.10. Adjacent steps are smaller than the seed spread, so the signal is the monotonicity itself. Under exchangeable noise, the probability that the four doses up to the default land in increasing order is $1/24$.

Evaluation sampling budget. Doubling the stored generations from n=8 to n=16 shrinks the seed-42 soft-gate-vs-EOPD Pass@8 gap from -1.10 to -0.04 $[-2.14, +1.95]$. It is unresolved at n=8 and certified equivalent at ± 2.2 points at n=16 (TOST, Appendix F).

Seeds and reading. Seed variance, not hyperparameter choice, is the dominant noise source (Table 12, Appendix F). Every gap that merely reroutes the forward-KL mass sits at or below the matching seed spread. Outcomes track the forward-KL budget rather than the gate's shape. Within the swept ranges, the budget is the one dial that measurably moves the operating point.

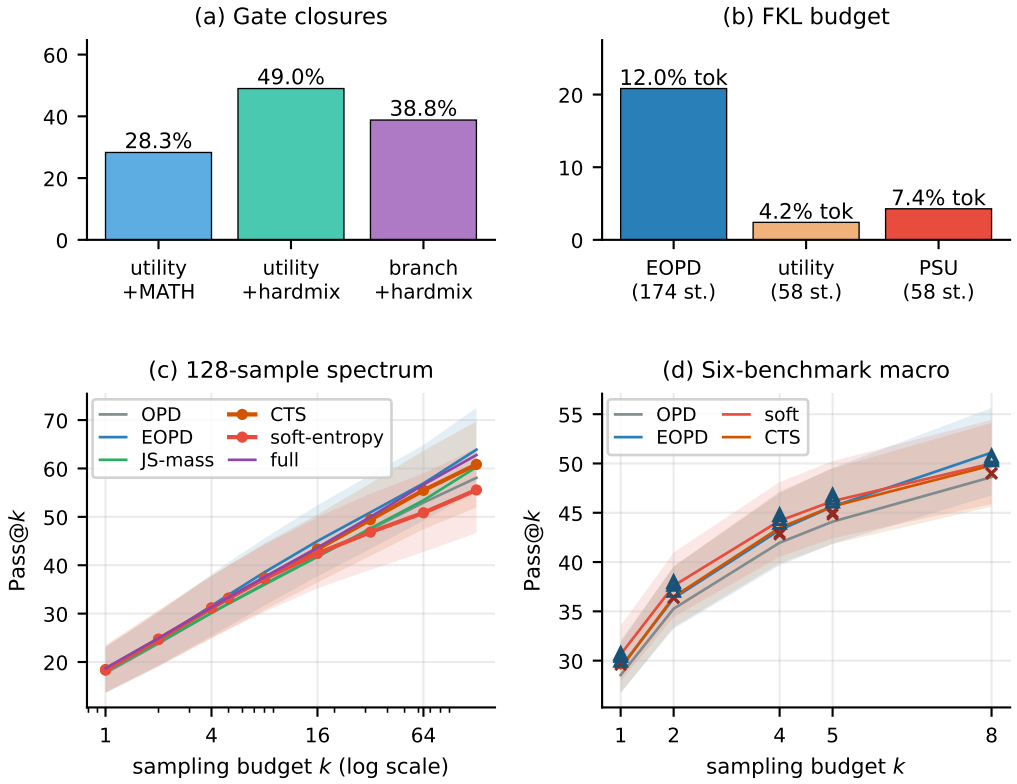

Figure 3: **Top: why utility gating fails.** (a) Per-step FKL-closed group fraction (28–49%), matching $(1-p)^4$. (b) EOPD's gate delivers $\sim 5\times$ the budget and metrics track it. **Bottom: the Pass@k spectrum.** (c) AMC23+AIME24+AIME25, 128 samples, seed 42, bootstrap bands: EOPD above every gate at $k \geq 4$. (d) Six-benchmark macro (n=8): thick marked lines are ours (soft-entropy, CTS), crosses and triangles mark seeds 41/43 (soft) and 41/43/44 (EOPD).

Efficiency. The soft gate uses one rollout per prompt at equal per-step cost to EOPD (358 vs. 368 s/step). Utility variants hold the 128-trajectory budget at equal wall time. Only the branch-conditioned variant adds a co-located teacher-continuation service on the run's GPUs.

6 Analysis: Why Coverage Transfer Is Hard to Improve

Hard gates suppress coverage exactly where it matters. With n=4 rollouts per prompt, a problem with success probability p is all-wrong with probability $(1-p)^4$, 0.66 at p=0.1. The marginal Pass@8 gain $8(1-p)^7$ is $\sim 61\times$ larger at p=0.1 than p=0.5. Training logs confirm it: 28–49% of prompt groups per step have FKL fully disabled by $\mathbb{I}[p_x > 0]$ (Figure 3a). A counterfactual attributes $\sim$50% of the marginal Pass@8 gain mass to those groups (Appendix F).

Budget accounting. EOPD applies FKL at 12.0% of tokens over 174 steps, the utility gate at 4.2% over 58 screening steps. PSU's reopened gate runs at 7.4–8.1% over the same 58 steps. Metrics track the $\sim 5\times$ gap in accumulated coverage signal more closely than gate shape (Figure 3b). All variants collapse student entropy within 10–20 steps (Appendix H). Hard-mix data competes with FKL for the same coverage capacity (Appendix F).

The Pass@k spectrum. Across AMC23/AIME24/AIME25 with 128 samples per problem, EOPD sits nominally above every gate at every $k \geq 4$. Low-k differences sit inside overlapping bootstrap bands (Figure 3c). On the six-benchmark macro the seed-42 soft-entropy gate is nominally above EOPD through Pass@5, below at Pass@8 (50.00 vs. 51.10). Every single-k difference is within paired bootstrap noise (Table 9). EOPD's four-seed Pass@5 mean is 46.27 ± 0.43 vs. the gate's 45.34 ± 0.60. Seed 42 is EOPD's *lowest* Pass@5 seed, so the soft gate's seed-42 lead through

Pass@5 is a seed-level artifact. The only ordering significant on the shared seed, the soft gate over OPD at Pass@1–2 ($+2.09$ $[+0.9, +3.4]$), reverses outright on the paired seed (-1.57 $[-3.1, -0.2]$). No Pass@k advantage of any gate over either endpoint survives the seed spread.

Seed variance dominates method gaps. Table 12 (Appendix F) aggregates the evidence (per-problem recounts in Appendices I–J). Four EOPD seeds span 29.32–30.73 macro Avg@8, three OPD seeds 28.54–31.15. EOPD's gap over OPD settles at $+0.55$, at EOPD's own seed spread. The soft gate's seeds interleave with EOPD's (-0.16 mean gap) at every Pass@k except Pass@8 (Figure 3d). Orderings flip: cover-87 leads EOPD by $+1.33$ on the shared seed (42) yet trails by -0.47 on the seed mean. Coverage is certifiably equivalent at ± 2 points (Welch TOST 90% CI $[-1.6, +0.7]$, Appendix F). Seed standard deviations are 0.5–1.1 points. Resolving a routing advantage would take ten times more seeds than this line has run. Single-seed comparisons, the original $+0.99$ included, are uninformative.

7 Discussion and Conclusion

Which regime should practitioners choose? Pick the gate by the deployment sampling budget. At $k \geq 8$ EOPD's plain entropy threshold is hard to beat. For tight budgets the soft gate is equivalent on average accuracy within ± 3 points (TOST) at identical cost. The half-schedule cover holds coverage level and keeps precision in EOPD's seed envelope. In practice, the dial is the cover budget, then the FKL budget.

Why is Pass@k so hard to improve? Coverage is a budget phenomenon. The regime where coverage accrues is stationary well before evaluation (Appendix H). Low-success problems carry the largest marginal gains and require *more* indiscriminate spending, not less. Every tested gate that prunes below EOPD's budget loses coverage. Redistribution cannot recover it.

Is routing the right lever, or is it data? Hard-mix augmentation exposes non-additivity. Replacing 50% of MATH with decontaminated hard problems lifts RKL-only OPD by $+1.39$ macro Avg@8 at no Pass@8 cost, nearly matching the gate's three-seed mean with the loss untouched. Stacking the gate on the same data instead *loses* 1.14 Avg@8 and 2.58 Pass@8 versus the gate on the standard MATH mix. The levers substitute rather than stack: the soft gate is the choice under a fixed training distribution.

Limitations. Appendix K lists each limit with its bounding evidence: one student, one teacher, and rule-verifiable mathematics. Transfer to other scales or non-verifiable domains is untested. The mechanism identities are scale-free. Three to four seeds per method resolve endpoint-level orderings but not routing gaps below $\sim$ 0.5 macro points. The bound applies to the original report too. Reproduction levels sit 1.7–1.9 Avg@8 below the record, deltas intact (Appendix C). No comparison mixes stacks. The dose response covers one schedule shape, so the Section 5.4 insensitivity holds only within the swept ranges. All forward-KL terms share one teacher under a fixed top-16 truncation, so absolute coverage levels are teacher-dependent. Orderings stay internally valid. Their invariance under a teacher swap is untested. No headline claim rests on an unbounded part. Each entry names the experiment that would move its boundary.

Conclusion. We aimed to make on-policy distillation smarter about *what* to cover. Coverage transfer is budget-limited, not routing-limited. EOPD's indiscriminate entropy threshold, the top spender, stays coverage-optimal. Each smarter gate spent the mass building Pass@k with no compensating precision gain. On precision a soft continuous entropy gate is a zero-cost drop-in equivalent (TOST, ± 3 points). A cover-then-sharpen curriculum makes the cover budget the field's one honest dial. Its half-schedule cover matches full-budget coverage, precision inside EOPD's seed envelope. Our whole-group miscoverage analysis argues the next coverage gains will come from sequence-level objectives and phase-aware schedules rather than smarter token gates, and we release all evaluation artifacts and routing code, so the price of every fork is metered before it is paid for, and the meter is public by design.

USE OF GENERATIVE AI

Generative AI tools were used to polish the language of this manuscript and to search for and organize candidate related work. All technical content, including the method design, experiments, and analyses, was conducted solely by the authors, who reviewed and take full responsibility for the final text. Every citation was verified against the original source and cross-checked against the arXiv API.

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 *International Conference on Learning Representations*, 2024.

Ibtihel Amara, Nazanin Sepahvand, Brett H Meyer, Warren J Gross, and James J Clark. Bd-kd: Balancing the divergences for online knowledge distillation. *arXiv preprint arXiv:2212.12965*, 2022.

Farid Bagirov, Mikhail Arkhipov, Ksenia Sycheva, Evgeniy Glukhov, and Egor Bogomolov. The best of n worlds: Aligning reinforcement learning with best-of-n sampling via max@k optimisation, 2025. URL https://arxiv.org/abs/2510.23393.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL https://arxiv.org/abs/2407.21787.

Pengxiang Cai, Tianchen Fang, Xiaohan Li, Qingyuan Zeng, Guocong Li, and Jintai Chen. Curriculum reinforcement learning can incentivize reasoning capacity in llms beyond the base model, 2026. URL https://arxiv.org/abs/2606.22317.

Mark Chen. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@k training for adaptively balancing exploration and exploitation of large reasoning models, 2025. URL https://arxiv.org/abs/2508.10751.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective, 2025. URL https://arxiv.org/abs/2506.14758.

Yinlam Chow, Guy Tennenholtz, Izzeddin Gur, Vincent Zhuang, Bo Dai, Sridhar Thiagarajan, Craig Boutilier, Rishabh Agarwal, Aviral Kumar, and Aleksandra Faust. Inference-aware fine-tuning for best-of-n sampling in large language models, 2025. URL https://arxiv.org/abs/2412.15287.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models, 2025. URL https://arxiv.org/abs/2505.22617.

Kaituo Feng, Changsheng Li, Xiaolu Zhang, Jun Zhou, Ye Yuan, and Guoren Wang. Keypoint-based progressive chain-of-thought distillation for llms, 2024. URL https://arxiv.org/abs/2405.16064.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning, 2025. URL https://arxiv.org/abs/2504.11456.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14852–14882, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.830. URL https://aclanthology.org/2023.acl-long.830/.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models, 2026. URL https://arxiv.org/abs/2603.07079.

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. Todi: Token-wise distillation via fine-grained divergence control. In *Conference on Empirical Methods in Natural Language Processing*, 2025.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Conference on Empirical Methods in Natural Language Processing*, 2016.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity, 2024. URL https://arxiv.org/abs/2310.06452.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms, 2025. URL https://arxiv.org/abs/2503.07067.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. In *Advances in neural information processing systems*, 2022.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations, 2025. URL https://arxiv.org/abs/2509.02534.

Lingyuan Liu and Mengxiang Zhang. Being strong progressively! enhancing knowledge distillation of large language models through a curriculum learning framework, 2025. URL https://arxiv.org/abs/2506.05695.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models, 2025. URL https://arxiv.org/abs/2505.24864.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, Bin Cui, and Wentao Zhang. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions, 2026. URL https://arxiv.org/abs/2506.07527.

MAA. American mathematics competitions - amc, 2023.

MAA. American invitational mathematics examination - aime, 2025.

Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, January 2005.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset, 2025. URL https://arxiv.org/abs/2504.16891.

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, 2025. URL https://arxiv.org/abs/2501.19393.

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, and Shuiwang Ji. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning, 2026. URL https://arxiv.org/abs/2506.06632.

Ruotian Peng, Yi Ren, Zhouliang Yu, Weiyang Liu, and Yandong Wen. Beyond the sampled token: Preserving candidate support in rlvr, 2026. URL https://arxiv.org/abs/2510.14807.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Yunhao Tang, Kunhao Zheng, Gabriel Synnaeve, and Rémi Munos. Optimizing language models for inference time objectives using reinforcement learning, 2025. URL https://arxiv.org/abs/2503.19595.

Christian Walder and Deep Karkhanis. Pass@k policy optimization: Solving harder reinforcement learning problems, 2026. URL https://arxiv.org/abs/2505.15201.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025. URL https://arxiv.org/abs/2506.01939.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*, 2022.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation, 2023. URL https://arxiv.org/abs/2307.15190.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may or may not escape its origin, 2026. URL https://arxiv.org/abs/2507.14843.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models, 2025. URL https://arxiv.org/abs/2408.00724.

Wenda Xu, Rujun Han, Zifeng Wang, Long T Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. *arXiv preprint arXiv:2410.11325*, 2024.

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

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

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025. URL https://arxiv.org/abs/2502.03387.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Yang Yu. Pass@k metric for rlvr: A diagnostic tool of exploration, but not an objective, 2025. URL https://arxiv.org/abs/2511.16231.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023. URL https://arxiv.org/abs/2308.01825.

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

A Implementation Details

All training runs use verl (Sheng et al., 2024) with fully sharded data parallel (FSDP) on two GPUs per run. Most runs use 2×A6000-48G, the final wave 2×A800-80G, and the remediation wave a third host (Section 5.1 and Appendix B). Rollouts run on vLLM with tensor-parallel 2 and sleep mode. Teacher top-k logits (k=16) are computed by an FSDP-sharded teacher with parameter offloading. PPO clip is $\epsilon = 0.2$ (verl default). Soft-KD uses full-vocabulary student log-probabilities gathered at teacher top-k indices. The FKL term is normalized over all valid response tokens. Evaluation uses the Qwen2.5-Math pipeline pinned at commit `a45202bd` with `qwen25-math-cot` prompts. Decoding uses temperature 1.0, top-p 0.8, and 8 samples per problem (128 for Pass@k curves on AMC23/AIME24/AIME25). Max generation length is 8192. The seed is 42. Rule-based scoring follows the EOPD repository (Jin et al., 2026).

Training-step data flow. One optimizer step proceeds as follows. (i) The trainer draws a batch of prompts (128 prompts for the n=1 arms, 32 for the grouped arms). vLLM generates n trajectories per prompt on the two rollout GPUs, which then enter sleep mode to free memory. (ii) The FSDP-sharded teacher (parameters offloaded to CPU between calls) makes one forward pass over every sampled trajectory. Each token position yields the full logit vector. From it the top-k=16 indices, their renormalized probabilities, and the token entropy are extracted and cached. Only these k indices per token are kept in GPU memory, never the full-vocabulary teacher logits. (iii) The student actor recomputes log-probabilities under the current policy and gathers its full-vocabulary log-probabilities at the teacher's top-k indices. From these it assembles the clipped reverse-KL surrogate of Eq. equation 2 together with the gated forward-KL term of Eq. equation 4. The gate value

is computed from the cached teacher entropy and attached as a stop-gradient constant. (iv) The actor update runs with gradient checkpointing and FSDP sharding across both GPUs. Then the rollout engine wakes for the next batch. The trainer logs per-step the FKL token ratio, student entropy, on-policy accuracy, losses, and gradient norms. These feed Appendices G and H.

Branch teacher service. The branch-value arm requires teacher continuations that the colocated teacher cannot provide without disrupting training. A separate persistent vLLM service (tensor-parallel 2 on two spare GPUs) serves the frozen teacher over HTTP. For each selected high-entropy state it receives the context plus the top-4 candidate tokens. It returns one continuation per candidate (temperature 1.0, max 4096 tokens). Each continuation is scored by the same rule-based checker as evaluation to form $Q(c_t, a) \in \{0, 1\}$. The service is stateless across problems, and its outputs are cached to disk. These scored branches are thereby part of the released artifacts. Training-side GPUs never wait on an uncached continuation inside the optimizer step.

Grouped-rollout controls. Section 4 references two formal-scale controls. They separate the utility gate from the prompt-diversity drop implied by its $n{=}4$, 32-prompt schedule. Both run the same grouped schedule. The first keeps the soft gate fully open with uniform trajectory weights (no utility signal at all). It reaches $29.7/48.6$ macro Avg@8/Pass@8, against $30.63/50.00$ for the $n{=}1$ soft gate. The second runs grouped OPD with the forward-KL branch disabled and reaches $28.2/47.8$, against $28.54/48.62$ for $n{=}1$ OPD. Hence the prompt-diversity drop costs roughly 0.8–1.4 coverage points with or without FKL. The remaining loss of the utility family is attributable to the $(1-p)^n$ closures of Section 6. Both controls are full 174-step runs at the shared protocol, included in the released generations.

Evaluation adapter. The pinned Qwen2.5-Math harness predates two of our six benchmarks. A small adapter patch (retained in the release) registers the `math500` and `aime25` dataset names and routes them to the corresponding pinned JSONL files. It maps every benchmark onto the same `qwen25-math-cot` prompt template and rule-based verdict. The patch touches data registration only, never prompts, decoding, or scoring. It is applied identically for every run and seed.

DeepMath hard-mix construction. From the DeepMath-103K core shard (He et al., 2025) (10,303 problems), we keep difficulty ≥ 5 with parseable rule-checkable answers (9,113). We then apply triple decontamination against all six evaluation benchmarks: exact match (0 removed), 10-gram overlap (37 removed), and semantic filtering with paraphrase-multilingual-MiniLM-L12-v2 embeddings at cosine ≥ 0.85 (352 removed). This leaves 8,724 candidates. We further keep only problems where the Qwen3-8B teacher solves at Pass@4 ≥ 0.25 and the base student at Pass@4 ≤ 0.50 (5,211 eligible). The 50/50 stratified mix is then built with a 5,211-row MATH subset (10,422 rows total, seed 42). This is the mix used by every hard-mix run in the paper.

Random seeds, windowed training, and one cross-host resume. Seed 42 for screening and the seed-42 main row. EOPD carries four formal seeds (41–44), and OPD and the soft gate three each (41/42/43). Multi-day runs checkpoint every 4–10 steps and resume from the latest checkpoint. The posterior-smoothed utility (PSU) tracker state is checkpointed alongside. One resume crossed hosts. OPD seed 43 continued on the A800 host from a step-20 checkpoint written on the A6000 host under the same pinned environment. Optimizer and scheduler state travelled inside the checkpoint. The resume convention is therefore identical within and across hosts.

Gate variant definitions. We give the full definitions of the utility- and value-conditioned arms summarized in Section 4. *Divergence- and mass-aware routing* (TSA-entropy-JS / TSA-JS-mass / TSA-full) multiplies the entropy gate by the student–teacher disagreement, a normalized Jensen–Shannon divergence on an extended support. This support holds the teacher's renormalized top-k scaled by its top-k mass $M_t^k = \sum_{x \in \mathcal{S}_t^k} \pi_{\mathtt{te}}(x \mid \mathbf{c}_t)$ and the student's full-vocabulary probabilities at the same indices. A residual-tail bin on each side completes it:

$$g_t = \sigma\Big(\tfrac{H_t^{\mathtt{te}} - \tau}{\lambda}\Big) \cdot \frac{\mathrm{JS}_{\mathrm{ext}}(\tilde{\pi}_{\mathtt{te}}, \pi_\theta)}{\log 2} \cdot \big[M_t^k\big]^{\mathbb{I}_{\mathrm{mass}}} \cdot \alpha(s), \tag{7}$$

where $\alpha(s) = 1 - 0.75s$ linearly decays over normalized training progress s for JS-mass and TSA-full. TSA-full additionally multiplies the gate by 0.25 from the earliest `\boxed` marker onward

(final-answer guard), tapering coverage transfer once the answer is reached. *Trajectory-utility conditioning* samples $n{=}4$ rollouts per prompt with rule-based rewards $r_i \in \{0, 1\}$ and group success rate $p_x = \frac{1}{4}\sum_i r_i$. It uses 32 prompts per step, i.e. the same 128 trajectories per step as the $n{=}1$ baselines, spread over four times fewer distinct prompts. The gate is

$$g_{it} = \sigma\Big(\tfrac{H_t^{\mathtt{te}}-\tau}{\lambda}\Big)\ \mathbb{I}[p_x > 0]\ \big(0.25 + 0.75(1-p_x)\big)\ \mathrm{clip}\Big(\tfrac{r_i+0.1}{p_x+0.1},\ 0.1,\ 4\Big), \tag{8}$$

so that correct-but-rare trajectories receive the strongest coverage signal while all-wrong groups fall back to OPD. *Branch-level value conditioning*: for each hard problem (DeepMath-sourced prompts), take the highest-entropy non-answer state of one representative trajectory. Score the teacher's top-4 candidate tokens by one teacher continuation each (temperature 1.0, max 4096 tokens, binary rule-based outcome $Q(c_t, a) \in \{0, 1\}$), and distill the value-tilted teacher

$$\tilde{\pi}^{V}_{\mathtt{te}}(a \mid c_t) \propto \tilde{\pi}_{\mathtt{te}}(a \mid c_t)\ \big(0.05 + Q(c_t, a)\big)^2, \tag{9}$$

falling back to trajectory gating when all candidates fail. This instantiation of value conditioning is deliberately minimal in cost: one state per problem, four candidates, and a single continuation each. Its outcomes therefore bound what cheap branch-value estimates buy, not what value conditioning could do at higher estimation budgets. *Posterior-smoothed utility (PSU)* replaces the hard indicator in equation 8 with a cross-epoch Beta(1,1)-posterior success estimate. This estimate is maintained by an exponential moving average (EMA, $\beta = 0.5$) over grouped rollouts, $\hat{p}_x = (s_x{+}1)/(n_x{+}2)$, at zero extra inference cost. The utility becomes $\max(\hat{p}_x^{0.5},\ 0.10)\cdot(1{-}\hat{p}_x)^7\cdot\mathrm{clip}\Big(\frac{r_i+0.1}{\hat{p}_x+0.1}, 0.1, 4\Big)$, with the marginal Pass@8 weight $(1-\hat{p}_x)^7$ batch-mean normalized. *Budget-preserving value reweighting (eopd-value)* keeps EOPD's hard entropy mask fully open. It only *redistributes* per-trajectory weights $w_i \propto (1-\hat{p}_x)^7 \cdot \mathrm{clip}\Big(\frac{r_i+0.1}{\hat{p}_x+0.1}, 0.1, 4\Big)$, mean-normalized so the total FKL budget matches EOPD exactly.

Scope of the design space. The one routing rule this family leaves untested is a uniform always-on gate that opens every token. We ran this arm under the matched protocol. The open gate lifts the forward-KL share from 12.25% to 96% of tokens, roughly the eightfold budget the extrapolation anticipated. The response saturates. Macro Pass@8 is 50.7 against the EOPD four-seed mean 50.68, and Pass@128 on the three hard benchmarks is 62 versus EOPD 63.9. The price lands on precision. Macro Avg@8 falls to 28.3, below every EOPD run (29.32 to 30.73). The student-entropy floor rises from 0.354 to 0.385, the expected slower collapse. The remaining mass therefore buys nothing at the top of the range and taxes precision below it.

B Experimental Details

This section consolidates the deployment facts behind every run: the training hardware, the pinned software stack, and the complete hyperparameter manifest. All values restate settings fixed before the formal runs and documented across Section 5.1 and Appendices A–D. None is new here, and none was revised after results were seen.

Hardware environment. All experiments ran on one of three hosts, with every training run occupying two GPUs (Table 3, which lists the two principal hosts). FSDP shards the student and the co-located teacher across both. vLLM rollouts run in sleep mode at tensor-parallel 2 (Appendix A). Host A carried all runs except the final wave. The final wave (EOPD seed 44, the $\lambda{=}0.20$ formal check, the CTS cover-87 rerun, and OPD seed 43 from its step-20 checkpoint onward) trained on Host B. No run spans hosts except that one documented resume. A later remediation wave trained on a third host under the identical pinned stack. It covers the α-dose arms of Table 2, CTS cover-87 seeds 41/43, and the two grouped-rollout controls of Section 4. Only one family, CTS cover-87, has same-method runs on Host B and the third host. Its two third-host seeds land at 29.43 and 28.96 against 30.64 for the Host-B seed 42. The original EOPD paper used $4{\times}$A100-80G. On every host here we changed only resource-level settings such as FSDP CPU offloading and the vLLM memory share, never protocol semantics (Section 5.1). The training recipe is therefore host-invariant.

Bounding the host effect. Only one seed family crosses hosts, and it bounds the effect with data. EOPD seed 44 (Host B, 30.52 macro Avg@8) lies inside the range its Host-A siblings span alone

(29.32–30.73). Thus the four-seed spread quoted in Section 5 is not widened by the Host-B run. The two remaining final-wave arms have no same-method Host-A twins. They cannot be formally bounded, and a host-level offset of roughly one seed spread is not excluded for them. Each nevertheless lands where its Host-A context predicts. For the cover-87 dose point, Pass@8 interpolates the dose trend, and the Avg@8 lead over the next-best dose point ($+0.66$) is within the seed spread. The λ=0.20 check lands below the default gate, as its Host-A pilot indicated. Both third-host replicas of the cover-87 point (seeds 41/43: 29.43 and 28.96 against 30.64 on Host B) sit within the single-host seed spreads of the other families. Hence host and seed effects cannot be separated for this arm.

Table 3: Training hosts. Each run uses two GPUs of its host.

	Host A	Host B
GPU	4× RTX A6000 (48 GB)	A800-SXM4 (80 GB)
CPU	2× Xeon Gold 6240	Xeon Platinum 8374C
Host RAM	503 GB	∼1 TB
Runs	all but the final wave	final wave only

Software stack. Table 4 lists the pinned stack. Training and evaluation environments are separate installations but carry identical package versions. Every host was provisioned from the same pinned package list, so the hosts are software-indistinguishable. The verl fork is vendored inside the EOPD reproduction repository (editable install). This pins it together with the training code. The evaluation harness is instead pinned independently at the commit recorded below.

Table 4: Pinned software stack (identical on both hosts, the evaluation environment matches training except as noted in Appendix A, and both hosts load the identical pinned list).

Component	Version	Component	Version
Python	3.10.13	accelerate	1.14.0
PyTorch	2.7.1	datasets	5.0.1
vLLM	0.10.0	NumPy	1.26.4
Ray	2.43.0	flash-attn	2.7.4.post1
Transformers	4.53.2	wandb	0.19.8
TensorDict	0.10.0	verl	0.7.0.dev0 (vendored)
Qwen2.5-Math eval harness: commit `a45202bd`			

Hyperparameters. Table 5 consolidates every knob of the formal runs. All values were fixed before the formal runs and never revised (Appendix D). Sweeps around the defaults marked in Table 5 appear in Table 2 and in Tables 10–11 at their respective pilot or formal scale.

C Reproduction Fidelity Audit

Our study stands on a strict reproduction of the EOPD protocol. The fidelity of that reproduction is a fair target of review. This appendix audits it against the published record (Jin et al., 2026).

Protocol checklist. The following are matched exactly, following the original's Section 5.1 and Appendix A. (i) Qwen3-1.7B-Base student. (ii) Qwen3-8B non-thinking teacher. (iii) MATH training set. (iv) 3 epochs (174 optimizer steps at global batch 128, mini-batch 32). (v) AdamW with lr 3×10^{-6} and cosine schedule. (vi) Training-time generation at temperature 1.0, top-p 1.0, one rollout per prompt, response length 4096. (vii) Teacher top-k=16 renormalized forward-KL. (viii) Entropy threshold τ=0.8. (ix) FKL coefficient α=1.0. (x) Evaluation with the Qwen2.5-Math pipeline, the same step-by-step prompt with boxed final answers. It uses 8 samples per problem at temperature 1.0 and top-p 0.8, max generation length 8192. (xi) The same six benchmarks and the same rule-based scoring.

Disclosed deviations. Three, none protocol-semantic. (i) Hardware: 2×A6000-48G instead of 4×A100-80G. This changes only resource-level settings (FSDP parameter offloading, vLLM memory share). (ii) PPO clip ϵ: the original states the clip symbol in its algorithm but never prints a numeric value. So we use the verl default ϵ=0.2 throughout, for every method. (iii) Total optimizer steps: the original reports epochs (3 for MATH) and figure axes consistent with ∼175 steps. Our 174 steps is that schedule. We also note for completeness that the original's LLM results carry no

Table 5: Hyperparameter manifest of the formal runs. Utility-conditioned variants hold the trajectory budget (128 trajectories per step) constant by pairing n=4 rollouts with 32 prompts per step.

Group	Parameter	Value
Models	student / teacher	Qwen3-1.7B-Base / Qwen3-8B (non-thinking)
	teacher top-k logits	$k = 16$
Optimization	optimizer / schedule	AdamW / cosine, lr 3×10^{-6}
	optimizer steps	174 (3 epochs for n=1 variants)
	global batch	128 trajectories per step
	mini-batch	32
	PPO clip	$\epsilon = 0.2$
Rollout	temperature / top-p	1.0 / 1.0
	max response length	4096
	utility variants	n=4 grouped rollouts, 32 prompts/step
Routing	gate temperature λ	0.15 (soft gate default)
	entropy threshold τ	0.8 (inherited from EOPD)
	CTS cover steps	58 / 87 / 116 / 174
Data	training set	MATH training split
	hard-mix share	50% DeepMath-hard (Appendix A)
Evaluation	samples per problem	$n = 8$, $n = 16$ resample, $n = 128$ spectra
	decoding	temperature 1.0, top-p 0.8
	max generation length	8192
	seeds	42 shared, 41/43 replicas, 44 (EOPD)

seed or variance information. No entry of its main table is repeated. The only multi-seed reporting anywhere in that paper is a toy experiment (3 seeds). Single-run records therefore carry the same order of seed uncertainty that Section 6 measures here (± 0.5–1.1 macro points).

Table 6: Reproduction fidelity audit, seed 42. Paper-record columns transcribed from the original's main table (Qwen3-1.7B, MATH setting), and repro columns are our runs. The EOPD$-$OPD macro deltas are preserved ($+0.77$ vs. $+1.00$ Avg@8, $+2.48$ vs. $+2.39$ Pass@8) while both endpoint levels shift down together on Avg@8, and per-benchmark delta-direction agreement is $5/6$ on Avg@8 and $3/6$ on Pass@8, with the disagreements on benchmarks where one problem moves several points (Appendix I, which counts the flips directly).

		Avg@8		Pass@8	
Benchmark	Run	paper	repro	paper	repro
MATH500	OPD	67.76	67.10	84.80	85.40
	EOPD	68.73	67.95	87.60	87.00
AMC23	OPD	39.06	35.62	70.00	62.50
	EOPD	41.88	38.12	75.00	75.00
Minerva Math	OPD	29.83	24.72	47.06	42.28
	EOPD	30.15	23.44	50.74	40.81
OlympiadBench	OPD	30.09	30.46	51.56	51.56
	EOPD	30.28	30.96	51.11	53.78
AIME24	OPD	8.33	7.50	20.00	23.33
	EOPD	10.42	10.83	23.33	30.00
AIME25	OPD	6.25	5.83	16.67	26.67
	EOPD	5.83	4.58	16.67	20.00
Macro	OPD	30.22	28.54	48.35	48.62
	EOPD	**31.22**	**29.32**	**50.74**	**51.10**
	ΔEOPD$-$OPD	+1.00	+0.77	+2.39	+2.48

Reading the audit. Three observations. First, the comparative structure reproduces. Both macro deltas keep their sign and magnitude. The macro Pass@8 ordering of the two methods is preserved. Every comparison inside this paper is run on our stack and never mixes records. The study's internal validity therefore does not depend on the level agreement. Second, the level offsets are endpoint-symmetric (-1.68 Avg@8 for OPD, -1.90 for EOPD, with Pass@8 levels slightly above record).

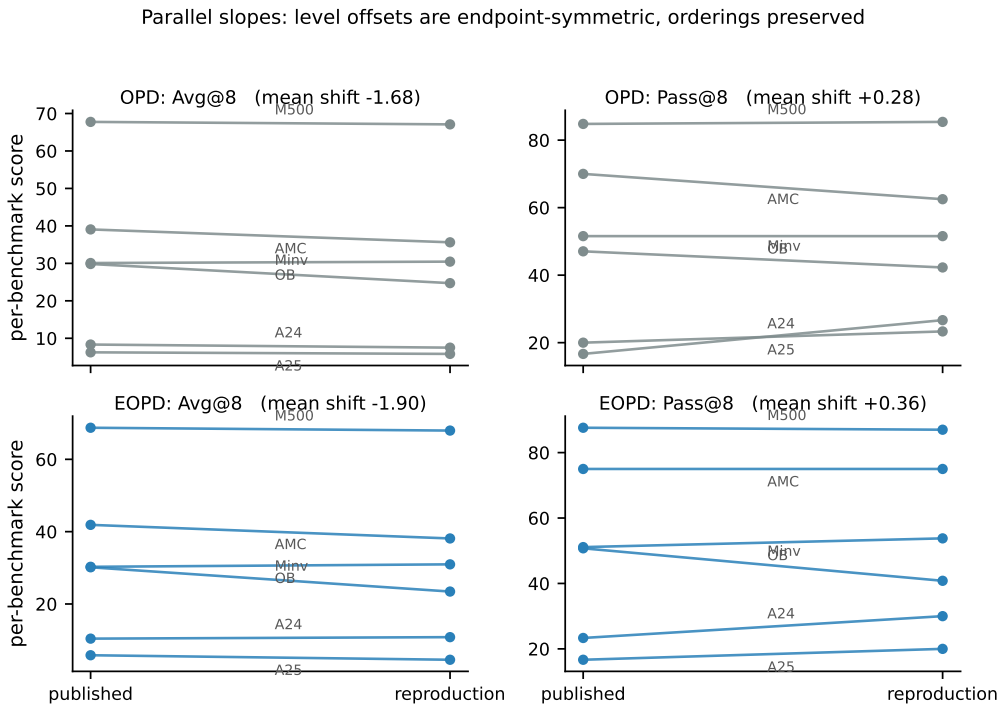

Figure 4: **Published-to-reproduction lines, one per benchmark.** Each line connects the paper record (left endpoint) to our seed-42 reproduction (right endpoint) for one benchmark of Table 6 (gray OPD, blue EOPD, labels: M500 = MATH500, Minv = Minerva Math). On Avg@8 (left column) the mean shift is -1.68 for OPD and -1.90 for EOPD, and the OPD-versus-EOPD ordering survives the offset on five of six benchmarks, with Minerva the exception. On Pass@8 (right column) the mean shifts are small ($+0.28$, $+0.36$), with per-benchmark offsets of both signs that largely cancel.

This is the signature of a stack-level offset (hardware, FSDP/vLLM configuration, harness revision) rather than a protocol error. The offsets are of the same order as the single-run seed spread measured in Section 6 (± 0.5–1.3 macro points). Figure 4 makes the symmetry visible as six Avg@8 lines per method that share a downward level offset. Their slopes vary by benchmark. Third, the per-benchmark disagreements concentrate exactly where problem counts are small. These are AIME25 (30 problems, so a single problem is 3.3 points) and OlympiadBench's sub-point Pass@8 delta (-0.45 in the record, itself within its own noise). The $3/6$ direction rate on Pass@8 should be read against this noise structure. Any per-benchmark delta at the one-problem or sub-point scale has a sign that single-run noise can flip. Only two of the three Pass@8 disagreements (OlympiadBench and AIME25) sit at that scale. The $5/6$ Avg@8 agreement is carried by deltas well above it. Minerva Math, the largest level gap (-5.1 to -6.7 on both endpoints), is again endpoint-symmetric. We flag it as the one benchmark where a harness-version or scoring-revision difference cannot be excluded. Our conclusions never rest on Minerva alone: every macro claim holds with any single benchmark dropped.

Endpoint symmetry as a diagnostic. The symmetry argument deserves one more level of detail. It is the load-bearing part of the fidelity claim. A protocol-semantic error, such as a wrong threshold, a misapplied gate, or a different loss, would move the two endpoints *differently*. It would change EOPD's level while leaving OPD untouched, or change the delta between them. OPD carries no forward-KL term. What we observe instead is a rigid translation of the whole configuration. OPD loses 1.68 Avg@8 points and EOPD loses 1.90. The Pass@8 levels of both move slightly *up* ($+0.28$, $+0.36$). That pattern is what a shared, method-independent shift produces. Examples include driver or library revisions, numerical kernels, or decoding internals of a pinned-but-evolving harness. The leave-one-benchmark-out check strengthens the reading. Under every single-benchmark deletion (range $+0.47$ to $+4.30$), the EOPD$-$OPD Pass@8 delta stays positive. The preserved macro delta is therefore not carried by any single benchmark, Minerva included.

D Reproducibility and Audit Trail

Pilot protocol and selection grid. The two soft-gate hyperparameters were selected before any formal run. Selection used 58-step pilots (one epoch) on the MATH training set. Each pilot was evaluated on the hard-3 slice (OlympiadBench, AIME24, AIME25) with the pinned harness. The grid was the Cartesian product of gate temperature $\lambda \in \{0.05, 0.10, 0.15, 0.20, 0.30\}$ (five points) and entropy threshold $\tau \in \{0.6, 0.8, 1.0\}$ (three points, bracketing the inherited EOPD default). Points $0.10/0.15/0.20$ ran first, and $0.05/0.30$ were added to bound both directions. We fixed the selection rule in advance: prefer the setting with the best pilot Pass@8, and break ties toward the inherited defaults. The winning cell (λ=0.15, τ=0.8) is the default reported throughout the paper. Tables 10 and 11 report every cell. No cell improves on the default within the swept ranges. Pilot AIME numbers move by 3.3 points per problem (Appendix I), so per-benchmark pilot rankings are coarse. Selection relied only on the hard-3 aggregate Pass@8, never on Avg@8 rankings.

Reporting boundary. Every hyperparameter in this paper was fixed before the formal (174-step) runs. None was revised afterwards. The soft gate's temperature and threshold were selected on 58-step pilots. The full grid of five λ and three τ settings ran on the hard-3 benchmark slice (Tables 10, 11). All formal runs then used λ=0.15, τ=0.8 everywhere. The setting covers the three seed replicas and the hard-mix variants, with no re-tuning. The curriculum's cover budget was fixed at 87 cover / 87 sharpen (the half schedule) for the main table. Dose points 58 and 174 are existing configurations: 58 is the pilot length, and 174 is EOPD itself. An additional 87-point dose was queued before any dose-table number was inspected. The evaluation pipeline (prompts, decoding parameters, rule-based scoring) is pinned at commit `a45202bd` of the Qwen2.5-Math harness (Yang et al., 2024). It was never modified between methods or seeds. Seeds 41/42/43 were fixed in advance for the multi-seed protocol. The additional EOPD seed 44 is documented in Appendix A.

Benchmark manifest. Table 7 records the six evaluation benchmarks exactly as instantiated by the pinned harness: source, problem count, and verdict rule. All problems are evaluated under the same rule-based final-answer check (Jin et al., 2026). No benchmark is subsampled, and no problem is removed for any method. Failed or empty generations are scored as incorrect rather than dropped. Training-side contamination against these six benchmarks is handled once, at the data level, for the DeepMath hard-mix (triple decontamination in Appendix A). The MATH training split follows the original EOPD protocol, inherited rather than re-derived here.

Table 7: Evaluation benchmark manifest (problem counts as loaded by the pinned harness, n=8 generations per problem, temperature 1.0, top-p 0.8, max length 8192).

Benchmark	Source	Problems	Verdict
OlympiadBench	He et al. (2024)	675	rule-based answer check
AMC23	MAA (2023)	40	rule-based answer check
AIME24	MAA (2025)	30	rule-based answer check
AIME25	MAA (2025)	30	rule-based answer check
MATH500	Hendrycks et al. (2021)	500	rule-based answer check
Minerva Math	Lewkowycz et al. (2022)	272	rule-based answer check
Total (six-benchmark pool)		1547	

Data manifest and integrity. Training and evaluation artifacts are pinned by content hash. The MATH training parquet has 7,500 rows (SHA-256 `64d3e376...`), the MATH validation parquet 5,000 rows (SHA-256 `0770e75f...`). As loaded by the pinned harness, the six evaluation sets include MATH500 (500 rows, SHA-256 `35dc4108...`) and AMC23 (40, `1ec3cc7d...`). The four remaining sets are Minerva Math (272, `dc752af0...`), OlympiadBench (675, `3441563b...`), AIME24 (30), and AIME25 (30). The total is 1,547 problems (Table 7). Student and teacher weight files are pinned by SHA-256 manifests in the release. Every number from our runs can be traced to an immutable data revision rather than a moving download.

Release manifest. The release contains five parts. (i) The vendored training framework with every routing variant (the soft entropy gate, entropy-JS, JS-mass, TSA-full, trajectory utility, PSU, branch value, and eopd-value) and the cover-then-sharpen curriculum launcher. (ii) All training and evaluation configuration files, frozen at the values of Section 5.1 and Appendix B. (iii) The per-problem

Table 8: Full six-benchmark results (Avg@8 / Pass@8), all rows, panel (a): MATH500 and AMC23.

	MATH500		AMC23	
Model	Avg@8	Pass@8	Avg@8	Pass@8
Base (no distill)	50.08	82.40	25.31	65.00
OPD (repro, s42)	67.10	85.40	35.62	62.50
OPD (ours, s41)	67.90	86.20	41.56	67.50
OPD (ours, s43)	67.25	85.20	39.06	72.50
OPD (ours, 3-seed mean)	67.42	85.60	38.75	67.50
EOPD (repro, s42)	67.95	87.00	38.12	75.00
EOPD (ours, s41)	67.45	86.20	39.69	70.00
EOPD (ours, s43)	68.65	89.00	41.56	77.50
EOPD (ours, s44)	67.78	87.40	40.31	77.50
EOPD (ours, 4-seed mean)	67.96	87.40	39.92	75.00
EOPD-always-on (ours, s42)	65	84	39	72
EOPD (paper record)	68.73	87.60	41.88	75.00
TSA-soft-entropy (ours, s42)	68.20	86.40	42.81	75.00
TSA-soft-entropy (ours, s41)	67.92	86.00	37.50	72.50
TSA-soft-entropy (ours, s43)	69.33	87.60	36.88	65.00
TSA-soft-entropy (ours, 3-seed mean)	68.48	86.67	39.06	70.83
TSA-entropy-JS (ours)	68.50	87.00	38.75	75.00
TSA-JS-mass (ours)	68.12	87.60	42.50	77.50
TSA-full (ours)	66.83	86.60	39.38	65.00
TSA-CTS cover-58 (ours)	67.88	87.60	39.69	67.50
TSA-CTS cover-87 (ours)	68.15	86.40	42.81	70.00
TSA-CTS cover-87 (ours, 3-seed mean)	66.88	85.63	41.77	70.83
TSA-CTS cover-116 (ours)	67.22	87.40	38.75	75.00

stored generations of every reported run (n=8 six-benchmark for all method–seed combinations, n=16 for the robustness check, and n=128 for the Pass@k spectra). (iv) The analysis scripts that regenerate every table and figure from those generations, including the paired-bootstrap, TOST, solvability-count, and diversity computations. (v) The data manifests and weight hashes described above. No number in the paper requires re-training to be verified. Each is either a deterministic function of the released generations or a stored summary-level metric (the CTS cover-87 seeds 41/43 cells of Table 8).

Procedural notes. Two timing facts deserve emphasis. First, the CTS cover-87 seeds 41/43 were queued as a follow-up replication after the dose table was first inspected. Every other choice in this appendix predates the formal runs. The remediation wave is included in the release on the same footing. It covers the α-dose arms of Table 2, CTS cover-87 seeds 41/43, and the grouped-rollout controls of Section 4. Second, training used two GPUs per run on one of three hosts (Appendix B). The seed protocol and the checkpoint-resume convention are documented in Appendix A. Seed 42 is shared. Seeds 41/43 cover EOPD, OPD, the soft gate, and CTS cover-87. Seed 44 runs EOPD. Tables and figures are produced by scripts over the stored generations rather than by transcription.

E Full Results Table

In the main text, Table 1 condenses the results to the key rows. Table 8 gives the complete per-benchmark table for all twenty-three rows. These cover every per-seed replica, the always-on arm, the three gate-composition variants (entropy-JS, JS-mass, TSA-full), and the three CTS dose points (cover-58/87/116). It is split into three stacked panels at main-text size. Every per-seed replica is shown explicitly, so seed ranges can be read directly. The macro seed ranges for the four multi-seed methods are in Table 12 (Appendix F). For the CTS cover-87 3-seed mean row (seeds 41/42/43), each cell is the three-seed mean. Its seed-41 and seed-43 components come from stored summary-level metrics rather than recomputed from per-problem generations. The hard-mix data-composition rows appear in Table 13 (Appendix F).

Table 8: Full six-benchmark results (continued), panel (b): Minerva and OlympiadBench.

	Minerva		OlympiadBench	
Model	Avg@8	Pass@8	Avg@8	Pass@8
Base (no distill)	13.37	37.50	20.61	46.81
OPD (repro, s42)	24.72	42.28	30.46	51.56
OPD (ours, s41)	25.41	41.18	31.20	52.30
OPD (ours, s43)	25.23	41.91	30.20	52.30
OPD (ours, 3-seed mean)	25.12	41.79	30.62	52.05
EOPD (repro, s42)	23.44	40.81	30.96	53.78
EOPD (ours, s41)	24.82	42.28	31.56	53.63
EOPD (ours, s43)	24.95	42.65	30.85	54.07
EOPD (ours, s44)	24.63	41.91	31.22	54.37
EOPD (ours, 4-seed mean)	24.46	41.91	31.15	53.96
EOPD-always-on (ours, s42)	23	43	29	52
EOPD (paper record)	30.15	50.74	30.28	51.11
TSA-soft-entropy (ours, s42)	24.91	41.18	30.78	54.07
TSA-soft-entropy (ours, s41)	24.17	41.91	31.22	53.48
TSA-soft-entropy (ours, s43)	24.59	43.38	31.13	54.81
TSA-soft-entropy (ours, 3-seed mean)	24.56	42.16	31.04	54.12
TSA-entropy-JS (ours)	23.58	40.07	30.48	51.56
TSA-JS-mass (ours)	25.41	41.91	31.19	52.74
TSA-full (ours)	23.35	38.60	30.89	52.59
TSA-CTS cover-58 (ours)	23.58	38.60	31.67	53.48
TSA-CTS cover-87 (ours)	24.26	42.65	31.13	54.52
TSA-CTS cover-87 (ours, 3-seed mean)	24.09	44.88	30.61	56.62
TSA-CTS cover-116 (ours)	24.17	42.65	31.09	53.78

F Extended Results

Extended per-benchmark numbers for all variants appear in Table 8 (Appendix E). Table 9 reports problem-level paired bootstrap 95% confidence intervals (10,000 resamples). Resampling is dataset-stratified: problems are resampled within each benchmark and the six per-benchmark deltas averaged. The CI reflects the six-benchmark macro. Pass@5 uses the unbiased estimator $1 - \binom{n-c}{5} / \binom{n}{5}$ over the n=8 stored generations. Throughout, $p(\Delta_{p8} > 0)$ is the one-sided bootstrap probability that the candidate improves over the baseline in Pass@8. Five shared-seed comparisons have CIs excluding zero in at least one metric. We begin with two. First, the seed-42 soft-entropy gate over OPD in Avg@8 ($+2.09$ $[+0.9, +3.4]$). Second, the cover-58 CTS curriculum below EOPD in Pass@8 (-3.79 $[-7.4, -0.4]$, $p = 0.014$). The paired seeds turn the first into the sharpest variance exhibit in the paper. On seed 41, the *same* comparison is significantly negative (-1.57 $[-3.1, -0.2]$). So the soft-over-OPD ordering flips sign between seeds while remaining individually significant on both. The CTS coverage deficit is the dose–response shown in Table 2 rather than a gate failure. Remaining exclusions are spending-side. EOPD over OPD on seed 43 is significant in Avg@8 ($+1.61$ $[+0.4, +3.0]$) and Pass@5 ($+3.01$ $[+0.4, +5.7]$). Seed 41 stays inside noise (-1.12 $[-2.5, +0.2]$). The formal λ=0.20 run falls below the default gate in Avg@8 (-1.53 $[-2.9, -0.3]$) at unchanged Pass@8. In contrast, the new cover-87 dose point's 95% paired intervals against EOPD include zero on every metric (Avg@8 $[-0.1, +2.9]$, Pass@8 $[-5.3, +3.2]$). They resolve neither parity nor a several-point deficit on coverage. The 90% Avg@8 interval $[+0.1, +2.6]$ just clears zero. This marginality is not corroborated by the three-seed follow-up (seed-mean gap -0.47). The hard-mix rows show the data-interaction effect of Figure 7 with uncertainty. OPD gains $+1.39$ Avg@8 $[-0.0, +2.9]$ at no coverage cost. Both FKL-based methods lose Pass@8 (-2.40 and -2.58). Each loss is individually within noise but direction-consistent.

Significance ledger. This ledger consolidates every significance statement in the paper. *Significant*, first, is the soft gate over OPD in Avg@8 on the shared seed ($+2.09$ $[+0.9, +3.4]$), extending to Pass@1–2 on the spectrum. *Significant*, second, is EOPD over the paired OPD seed 43 in Avg@8 and Pass@5 ($+1.61$ $[+0.4, +3.0]$, $+3.01$ $[+0.4, +5.7]$). Also *significant* are the cover-58 CTS cur-

Table 8: Full six-benchmark results (continued), panel (c): AIME24, AIME25, and the six-benchmark macro.

	AIME24		AIME25		Macro	
Model	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8
Base (no distill)	5.00	16.67	2.92	16.67	19.55	44.17
OPD (repro, s42)	7.50	23.33	5.83	26.67	28.54	48.62
OPD (ours, s41)	13.33	20.00	7.50	26.67	31.15	48.97
OPD (ours, s43)	7.50	16.67	5.42	20.00	29.11	48.10
OPD (ours, 3-seed mean)	9.44	20.00	6.25	24.44	29.60	48.56
EOPD (repro, s42)	10.83	30.00	4.58	20.00	29.32	51.10
EOPD (ours, s41)	11.25	30.00	5.42	20.00	30.03	50.35
EOPD (ours, s43)	10.42	23.33	7.92	16.67	30.73	50.54
EOPD (ours, s44)	11.25	26.67	7.92	16.67	30.52	50.75
EOPD (ours, 4-seed mean)	10.94	27.50	6.46	18.33	30.15	50.68
EOPD-always-on (ours, s42)	11	25	6	21	28.3	50.7
EOPD (paper record)	10.42	23.33	5.83	16.67	31.21	50.74
TSA-soft-entropy (ours, s42)	9.17	23.33	7.92	20.00	30.63	50.00
TSA-soft-entropy (ours, s41)	10.83	20.00	5.83	20.00	29.58	48.98
TSA-soft-entropy (ours, s43)	10.83	23.33	5.83	20.00	29.76	49.02
TSA-soft-entropy (ours, 3-seed mean)	10.28	22.22	6.53	20.00	29.99	49.33
TSA-entropy-JS (ours)	10.00	20.00	5.42	16.67	29.45	48.38
TSA-JS-mass (ours)	10.42	16.67	5.42	13.33	30.51	48.29
TSA-full (ours)	9.58	23.33	6.67	16.67	29.45	47.13
TSA-CTS cover-58 (ours)	8.75	20.00	8.33	16.67	29.98	47.31
TSA-CTS cover-87 (ours)	10.83	23.33	6.67	23.33	30.64	50.04
TSA-CTS cover-87 (ours, 3-seed mean)	8.05	21.10	6.66	22.21	29.68	50.21
TSA-CTS cover-116 (ours)	9.58	23.33	4.58	16.67	29.23	49.80

riculum below EOPD in Pass@8 (-3.79, p=0.014). The formal λ=0.20 run sits below the default gate in Avg@8 (-1.53 $[-2.9, -0.3]$). Each is a statement about *how much* forward-KL mass is applied: FKL versus none, longer versus shorter cover, and steeper versus default sigmoid. The leading item is itself sign-reversed on another paired seed (soft over OPD on seed 41: -1.57 $[-3.1, -0.2]$, and EOPD$-$OPD fades to -1.12 $[-2.5, +0.2]$). This is the cleanest demonstration that single-seed significance in this design space is not a property of the method. *Never significant* are the routing-level comparisons at matched budget. These are the soft gate versus EOPD at every measured k and evaluation budget (Pass@5 $+0.55$ $[-2.5, +3.7]$, Pass@8 -1.10 $[-5.2, +3.0]$, Pass@8 at n=16: -0.04 $[-2.14, +1.95]$). The cover-87 and cover-116 curricula versus EOPD never reach significance on any metric, and neither does any individual hard-mix FKL loss. *Robustness checks*: doubling the evaluation budget, the stratified bootstrap design, and the leave-one-benchmark-out test all leave this pattern unchanged. Every macro ordering keeps its sign with any single benchmark removed, and the EOPD$-$OPD Pass@8 delta ranges $+0.47$ to $+4.30$ across the six leave-outs. Problem-level counts of Appendix I reproduce it without any aggregation. The ledger's shape is the paper's thesis in miniature: orderings that change the forward-KL mass can reach significance, while orderings that reroute a matched mass do not. Figure 5 renders the ledger as a forest plot. The intervals that clear zero are all spending-side. Its matched-budget routing side is uniformly gray, from top to bottom of both panels.

Equivalence testing (TOST). Where the main text claims that two runs match, we replace the vague reading of an interval overlapping zero with two one-sided tests on the same stratified bootstrap distributions. A comparison is declared equivalent at margin D when its 90% CI lies entirely inside $[-D, +D]$. On macro Avg@8 the soft gate is equivalent to EOPD at a $\pm$3-point margin on every paired seed. The 90% CIs are $[+0.1, +2.5]$ on seed 42, $[-1.7, +0.7]$ on seed 41, and $[-2.1, +0.1]$ on seed 43. It is also equivalent at a $\pm$2.1-point margin on the seed-mean contrast (-0.16, 90% CI $[-2.0, +1.6]$, resampling problems independently within each run). This is the computation behind the abstract's equivalence claim. At n=16 stored generations the seed-42 Pass@8 gap is equivalent at a $\pm$2.2-point margin. Its reported 95% CI $[-2.14, +1.95]$ already lies

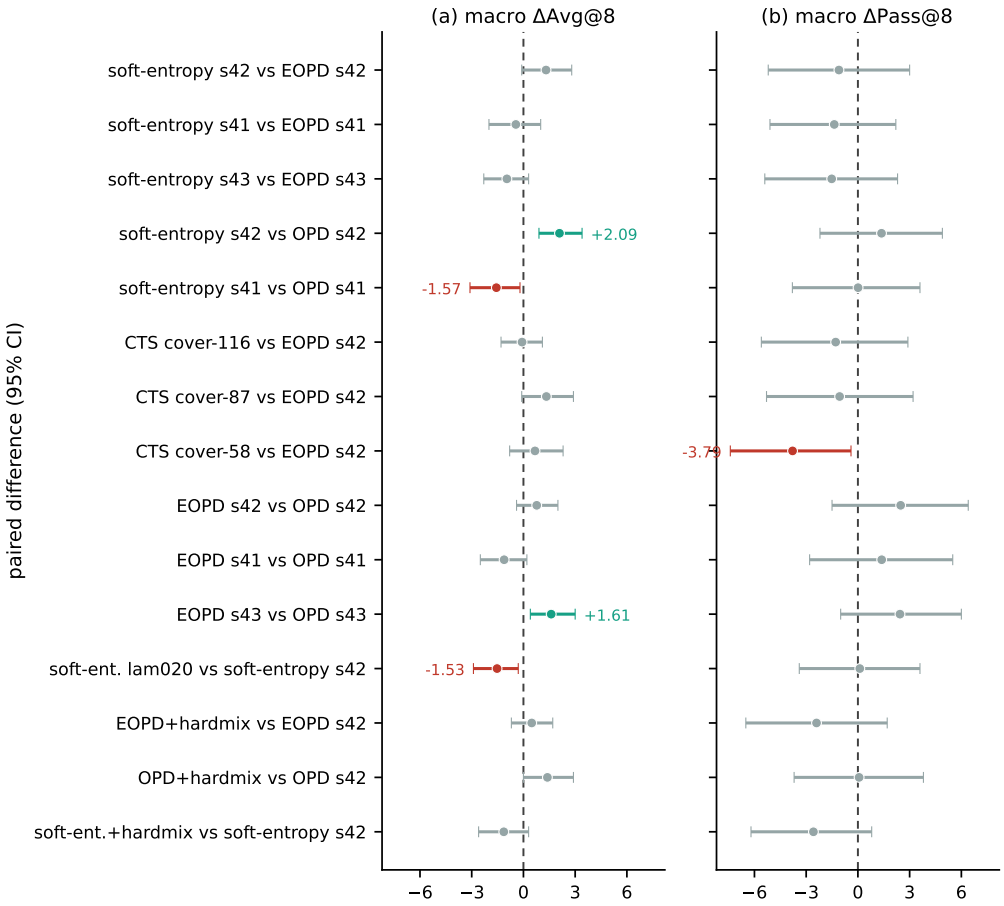

Figure 5: **The significance ledger as a forest plot.** Every paired comparison of Table 9 as a point estimate with its 95% confidence interval (left: macro ΔAvg@8, right: ΔPass@8, dotted line at zero). Every interval that clears zero is spending-side, and every matched-budget routing comparison straddles zero.

inside. The Pass@8 parity comparisons at $n{=}8$ are the soft gate versus EOPD on all three seeds, cover-87, and cover-116. None is certifiably equivalent at any margin up to ± 3 (minimal demonstrable margins 4.1–4.8 points). The cover-87 curriculum versus EOPD is equivalent on Avg@8 at ± 2.7 but not on Pass@8 (minimal margin 4.7). Wherever equivalence is not certified, we report the interval as resolving neither direction rather than claiming a match. Two cover-87 seeds were added after the initial equivalence analysis (seeds 41/43: $29.43/51.02$ and $28.96/49.58$). They bring its three-seed mean to 29.68 ± 0.71 Avg@8 and 50.21 ± 0.60 Pass@8. On seed means, a Welch two-one-sided test certifies cover-87 equivalent to EOPD on Pass@8 at a ± 2-point margin (90% CI $[-1.6, +0.7]$). The same test certifies Avg@8 at the same margin (90% CI $[-1.8, +0.9]$). This is the first coverage-side equivalence we can certify. Unlike the problem-level margins above, this across-seed interval does not absorb benchmark-resampling noise. With three and four seeds it rests on few degrees of freedom. The equivalence script and full outputs are part of the released artifacts.

Multiplicity. The forest of intervals in Figure 5 is uncorrected for multiple comparisons, by design. We rest no positive claim on any single significant interval. The two positive shared-seed comparisons against OPD that clear zero (the soft gate on seed 42 and EOPD on seed 43) each reverse sign on another paired seed. That is exactly the behaviour multiple-comparison logic predicts for noise-level excursions. The replicated exclusions are all spending-side, and the negative claim of the ledger (no routing-level comparison at matched budget reaches significance) is strengthened, not weakened, by the multiplicity.

Table 9: Paired bootstrap on the six-benchmark macro (problem-level, dataset-stratified, 10,000 resamples): candidate minus baseline, 95% CI in brackets.

Candidate	vs	ΔAvg@8	ΔPass@5	ΔPass@8	$p(\Delta_p 8 > 0)$
soft-entropy s42	EOPD s42	+1.31 [-0.1, +2.8]	+0.55 [-2.5, +3.7]	-1.10 [-5.2, +3.0]	0.302
soft-entropy s41	EOPD s41	-0.44 [-2.0, +1.0]	-1.15 [-4.0, +1.6]	-1.37 [-5.1, +2.2]	0.227
soft-entropy s43	EOPD s43	-0.96 [-2.3, +0.3]	-1.96 [-4.9, +0.9]	-1.52 [-5.4, +2.3]	0.222
soft-entropy s42	OPD s42	+2.09 [+0.9, +3.4]	+2.08 [-0.6, +4.7]	+1.37 [-2.2, +4.9]	0.780
soft-entropy s41	OPD s41	-1.57 [-3.1, -0.2]	-0.48 [-3.4, +2.3]	+0.01 [-3.8, +3.6]	0.500
CTS cover-116	EOPD s42	-0.08 [-1.3, +1.1]	+0.02 [-3.2, +3.3]	-1.29 [-5.6, +2.9]	0.277
CTS cover-87	EOPD s42	+1.33 [-0.1, +2.9]	+0.43 [-2.6, +3.6]	-1.06 [-5.3, +3.2]	0.317
CTS cover-58	EOPD s42	+0.67 [-0.8, +2.3]	-1.59 [-4.4, +1.1]	-3.79 [-7.4, -0.4]	0.014
EOPD s42	OPD s42	+0.77 [-0.4, +2.0]	+1.53 [-1.3, +4.3]	+2.48 [-1.5, +6.4]	0.887
EOPD s41	OPD s41	-1.12 [-2.5, +0.2]	+0.67 [-2.4, +3.7]	+1.38 [-2.8, +5.5]	0.739
EOPD s43	OPD s43	+1.61 [+0.4, +3.0]	+3.01 [+0.4, +5.7]	+2.44 [-1.0, +6.0]	0.917
soft-ent. lam020	soft-entropy s42	-1.53 [-2.9, -0.3]	-0.58 [-3.2, +1.9]	+0.11 [-3.4, +3.6]	0.523
EOPD+hardmix	EOPD s42	+0.48 [-0.7, +1.7]	-0.69 [-3.6, +2.3]	-2.40 [-6.5, +1.7]	0.124
OPD+hardmix	OPD s42	+1.39 [-0.0, +2.9]	+0.90 [-1.9, +3.7]	+0.07 [-3.7, +3.8]	0.528
soft-ent.+hardmix	soft-entropy s42	-1.14 [-2.6, +0.3]	-2.00 [-4.9, +0.8]	-2.58 [-6.2, +0.8]	0.073

Counterfactual gate-closure estimate. The ∼50% figure in Section 6 is computed from the stored n=8 generations of each evaluation run, with no additional sampling. For every problem we estimate the student's success probability $\hat{p}$ as the fraction of solved generations. We take the hard gate's per-step closure probability on a solvable problem to be $(1-\hat{p})^4$, the chance that a fresh group of n=4 rollouts contains no success. Each problem is weighted by its marginal Pass@8 gain $g \propto 8(1-\hat{p})^7$, i.e. $\partial[1-(1-p)^8]/\partial p$ evaluated at $\hat{p}$. The reported quantity is the gain-weighted mass share $\sum_{\hat{p}>0}(1-\hat{p})^4\, g \,/\, \sum_{\hat{p}>0} g$. This is the fraction of total marginal Pass@8 improvement that sits on problems the hard gate would close at any given step. Across the three hard benchmarks this share is 0.48–0.56 for both the OPD and EOPD students. The unweighted closure rates (OlympiadBench 15.9/17.3%, AIME24 32.5/28.7%, AIME25 43.3/40.9%, OPD/EOPD) overlap the 28–49% per-step closure observed in the actual training logs (Figure 3a). This overlap is the sanity check for using benchmark samples as a proxy for the training distribution. Two caveats. First, $\hat{p}$ from 8 samples is noisy in the low-$\hat{p}$ tail where the gain weight concentrates. Second, benchmark problems proxy, but do not sample, the MATH training mix.

Gate-hyperparameter sweeps. Tables 10 and 11 report the full pilot-scale sweeps behind Section 5.4.

Reading the sweep tables. On gate temperature (Table 10), the default λ=0.15 attains the best pilot Pass@8 (34.89). Moving in either direction degrades coverage (28.49 at λ=0.05, 30.77 at λ=0.10, 29.85 at 0.20, 30.30 at 0.30). The margin over the swept alternatives is 4.1–6.4 points. The Avg@8 column shows no such ordering (13.84–16.04 with the default second-lowest). This is the pilot-scale signature of the precision–coverage split. At 58 steps the gate has spent too little budget for its coverage effect to register on the mean, and Avg@8 rankings at this scale are noise. Appendix I counts the AIME cells behind them: the 13.33–30.00 AIME24 range is 4–9 of 30 problems. On entropy threshold (Table 11), the inherited default τ=0.8 again attains the best Pass@8 (34.89 against 32.22 at τ=0.6 and 30.67 at τ=1.0). Lowering the threshold admits teacher confusion into the coverage set, and raising it starves the budget. Both sweeps therefore confirm the defaults as local optima within the swept ranges. The formal λ=0.20 check (Section 6, −1.53 Avg@8 $[-2.9, -0.3]$ at unchanged Pass@8) confirms the pilot ordering transfers to formal scale.

Additional analysis figures. Figures 6 and 7 accompany Sections 6 and 5.3: the precision–coverage map of all runs, and the hard-mix interaction with routing.

Reading the trade-off map. In Figure 6, every run is one point in the (Avg@8, Pass@8) plane, and the geometry makes the budget principle visible without any statistics. The two endpoint clouds sit on a rising diagonal: OPD seeds occupy the lower-left (low spend, low coverage), EOPD seeds the upper-right (high spend, high coverage). Their separation is mostly vertical: the forward-KL budget buys Pass@8 far more than Avg@8. The soft gate's three seeds land inside the EOPD cloud on the precision axis while sitting slightly below it on coverage. This is the map form of the TOST

Table 10: Gate-temperature sensitivity (58-step pilots on MATH, hard-benchmark macro). No setting improves on our default $\lambda = 0.15$ within the swept range: both harder (0.05) and softer (0.30) gates degrade coverage, and the two intermediate settings (0.10, 0.20) bracket the default.

λ (gate temperature)	Avg@8	Pass@8	OlympiadBench	AIME24	AIME25
$\lambda = 0.05$	13.84	28.49	52.15	13.33	20.00
$\lambda = 0.10$	15.28	30.77	52.30	20.00	20.00
$\lambda = 0.15$ **(ours)**	14.92	**34.89**	54.67	30.00	20.00
$\lambda = 0.20$	16.04	29.85	52.89	20.00	16.67
$\lambda = 0.30$	14.40	30.30	54.22	20.00	16.67

Table 11: Entropy-threshold sensitivity (58-step pilots on MATH, hard-benchmark macro). The default $\tau = 0.8$ inherited from EOPD is the best setting, and neither direction of change helps.

τ (entropy threshold)	Avg@8	Pass@8	OlympiadBench	AIME24	AIME25
$\tau = 0.6$	15.52	32.22	53.33	23.33	20.00
$\tau = 0.8$ **(default)**	14.92	**34.89**	54.67	30.00	20.00
$\tau = 1.0$	15.70	30.67	52.00	20.00	20.00

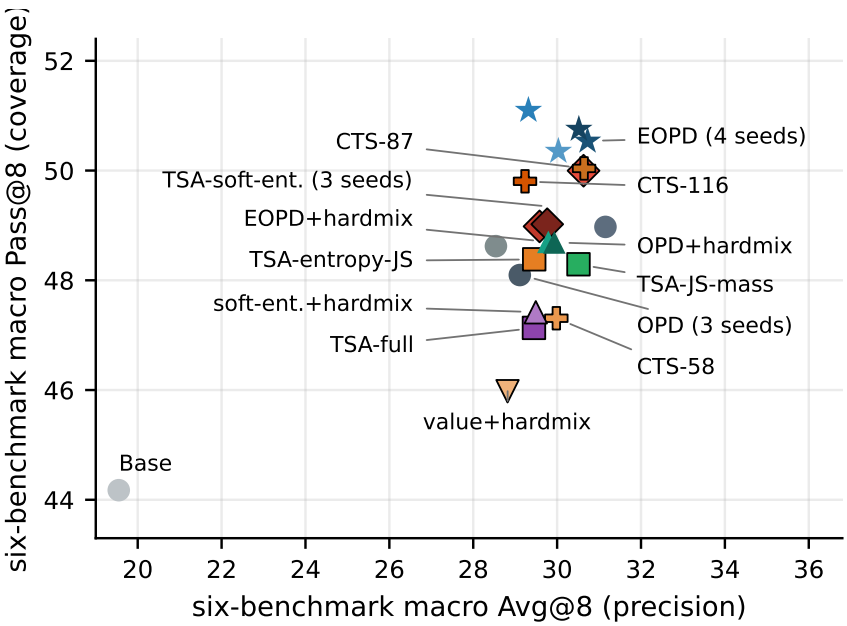

Figure 6: **The precision–coverage map of on-policy distillation.** Six-benchmark macro Avg@8 (x) vs. Pass@8 (y): the soft gate's three seeds land inside the EOPD cloud on precision, and the three CTS cover budgets span the upper-left segment with the half-schedule cover (87) at its head. Utility- and value-conditioned variants fall strictly below the EOPD–soft-entropy region on coverage, consistent with their reduced FKL spend.

equivalence with a residual coverage gap. All three CTS dose points trace a near-straight path from the EOPD endpoint toward lower coverage as the cover budget shrinks, with the 87-step point closest to the endpoint. Thus the curriculum moves the operating point along the trade-off line rather than off it. The utility and value variants finally fall strictly below the diagonal band, with less coverage at comparable precision. This is the map form of the $(1-p)^n$ closure mechanism.

Hyperparameter sweep heatmap. Figure 8 renders the λ/τ sweeps and the CTS cover-budget dose behind Section 5.4. The two panels should be read differently. Panel (a) (pilot sweeps) shows near-uniform colour across each sweep block: no cell separates from the default by more than the pilot noise floor. This is the insensitivity result of Section 5.4 in one view: the defaults are flat optima, not knife-edge choices. Panel (b) (CTS dose) shows the only strong gradient in the figure: Pass@8 rises monotonically with the cover budget (58→174), while Avg@8 stays within a narrow band. The contrast between the flat panel (a) and the single-gradient panel (b) is the dose–

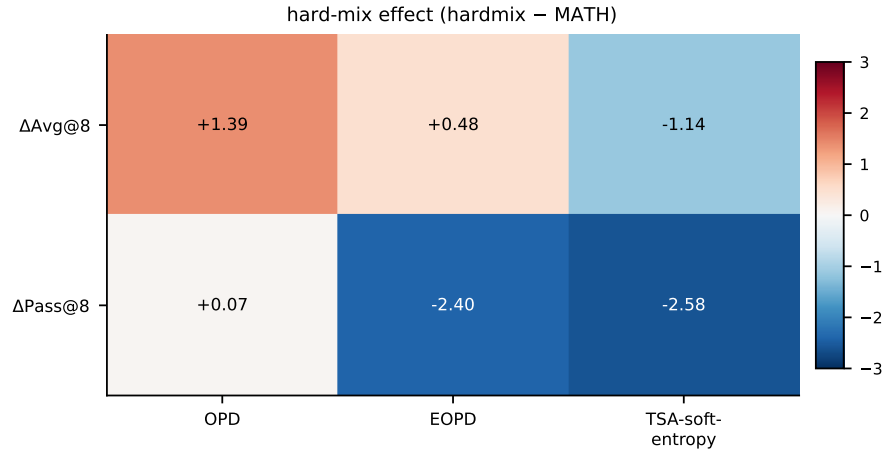

Figure 7: **Hard data interferes with coverage routing.** Macro-metric change when 50% of MATH is replaced by decontaminated DeepMath-hard problems. Augmentation helps the RKL-only OPD baseline ($+1.39$ Avg@8 at no coverage cost) but reduces Pass@8 for both FKL-based methods (EOPD -2.40, soft-entropy -2.58): harder prompts and FKL coverage compete for the same student capacity rather than adding orthogonal gains.

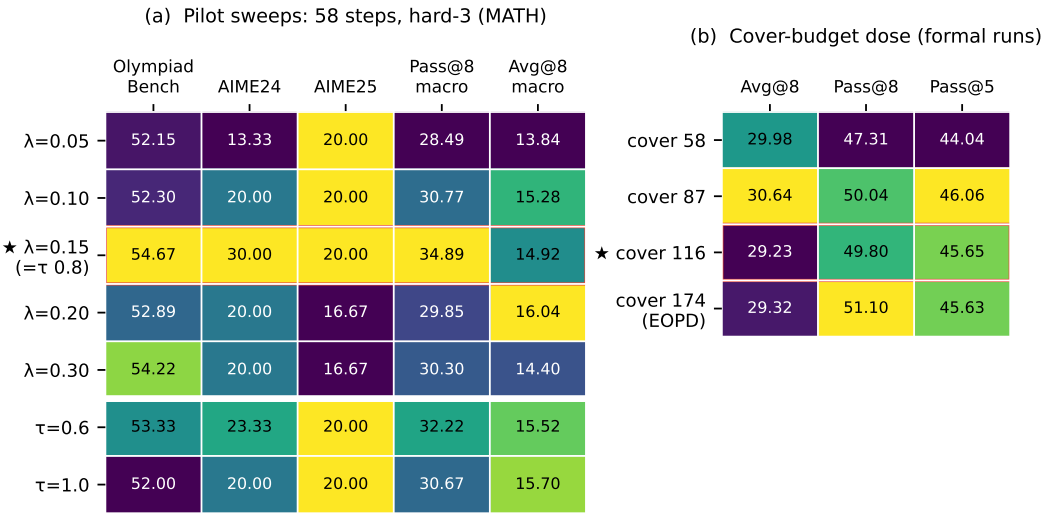

Figure 8: **No knob improves on its default.** (a) λ and τ sweeps at pilot scale (58 steps, hard-3). (b) CTS cover-budget dose at formal scale (seed 42). Near-uniform colour within a sweep is the insensitivity result, and the only strong gradient is Pass@8 down the cover budget.

response claim made visual: routing-knob changes move nothing beyond noise, budget changes move coverage monotonically.

Gate-composition cascade. Figure 9 renders the per-component coverage cost behind Section 5.3. Reading the cascade bar by bar: the soft gate starts at 50.00 macro Pass@8. Adding the JS-divergence factor costs -1.62 (entropy-JS, 48.38). Top-k mass weighting costs a further -0.09 (JS-mass, 48.29) inside single-run noise. Coefficient decay and the final-answer guard cost another -1.16 (TSA-full, 47.13). Every factor prunes or redistributes forward-KL mass, and each pruning shows up as a coverage loss of matching sign. Once fully equipped, the gate ends 3.97 points below EOPD's 51.10. The lower panel is the control: macro Avg@8 for the same four runs stays within EOPD's four-seed envelope (shaded band, 29.32–30.73). So the cascade is a pure coverage loss with no compensating precision gain, the component-level form of the budget principle.

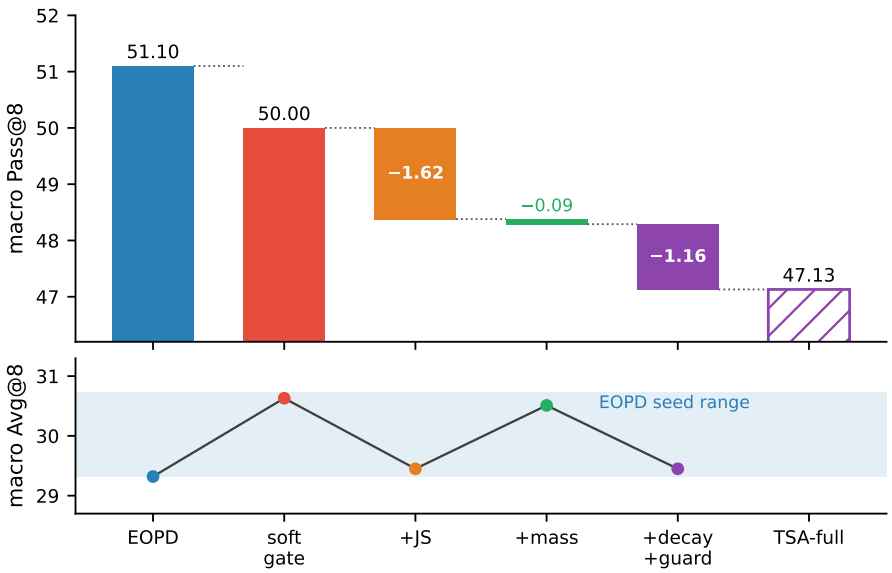

Figure 9: **The price of gate sophistication.** Top: macro Pass@8 (seed 42) as routing components are added one at a time, and the full stack lands 3.97 points below EOPD. Bottom: macro Avg@8 stays inside EOPD's four-seed range (shaded): a pure coverage loss.

Table 12: Seed variance vs. method gaps (formal 174-step runs, six-benchmark macro, seed mean with [min–max] seed range, as in Table 1). Method gaps sit at or below the corresponding seed spreads. The CTS cover-87 Pass@5 cell is the seed-42 value from stored generations (seeds 41/43 are summary-level).

Method	#seeds	Avg@8 (=Pass@1)	Pass@5	Pass@8
OPD	3	29.60 [28.54–31.15]	44.46 [43.77–45.50]	48.56 [48.10–48.97]
EOPD	4	30.15 [29.32–30.73]	46.27 [45.63–46.78]	50.68 [50.35–51.10]
TSA-soft-entropy	3	29.99 [29.58–30.63]	45.34 [44.82–46.18]	49.33 [48.98–50.00]
TSA-CTS cover-87	3	29.68 [28.96–30.64]	45.13	50.21 [49.58–51.02]
gap EOPD−OPD (Avg@8)		+0.55	ranges overlap: [28.54–31.15] vs. [29.32–30.73]	
gap soft−EOPD (Avg@8)		-0.16	ranges overlap: [29.58–30.63] vs. [29.32–30.73]	

Seed variance versus method gaps. Table 12 collects the per-method seed spreads behind Section 6. Two readings matter. First, the method gaps sit at or below the seed spreads: EOPD−OPD on Avg@8 is $+0.55$ against OPD's own spread of ± 1.12. Soft−EOPD is -0.16 against EOPD's spread of ± 0.54 and the gate's ± 0.46. Any single-seed draw from these distributions can therefore produce a spurious ordering of either sign. The paired-seed flips of Section 6 ($+2.09$ to -1.57) are the empirical demonstration. Second, the one gap that does clear the spreads is a spending gap: EOPD's Pass@8 mean of 50.68 ± 0.28 against the gate's 49.33 ± 0.47. Their seed ranges, 50.35–51.10 versus 48.98–50.00, do not overlap. The CTS cover-87 row (29.68 ± 0.71 / 50.21 ± 0.60) sits 0.47 points below EOPD's means on both metrics, with seed ranges that overlap EOPD's. This is the table form of the curriculum's parity claim.

Additional seeds. After the multi-seed analysis was frozen, we ran one further seed for each under-covered arm. The soft gate's seed 44 reaches $30.27/49.38$ and OPD's seed 44 reaches $29.61/48.52$ (macro Avg@8/Pass@8). Both flanking curriculum doses gain seeds 41 and 43 (cover-58 at 29.84 Avg@8 / 47.31 Pass@8, cover-116 at 28.94 / 49.77). Every added seed lands inside the corresponding family's seed envelope reported in Table 12 and leaves every seed-level conclusion unchanged. They are not folded into the headline means, which remain frozen at the Arc-27 protocol (seeds 41/42/43 for CTS cover-87, 41–44 for EOPD).

Hard-mix per-benchmark rows. Table 13 gives the per-benchmark rows of the four data-composition runs (Section 5.3), moved from Table 1 for space. Reading the rows against their MATH-mix counterparts: the OPD hard-mix run improves macro Avg@8 by $+1.39$ at no cov-

Table 13: Per-benchmark results for the data-composition runs (Avg@8 / Pass@8), split into two panels as in Table 1: the four large benchmarks above, AIME24/AIME25 and the six-benchmark macro below.

	MATH500		AMC23		Minerva		OlympiadBench	
Model	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8
EOPD + hard-mix	68.45	88.40	39.06	65.00	23.90	42.28	32.33	53.19
OPD + hard-mix	69.53	88.80	37.50	67.50	25.64	41.91	30.69	53.93
soft-entropy + hard-mix	68.62	87.20	39.69	72.50	24.54	41.18	30.76	53.63
eopd-value + hard-mix	68.45	86.80	37.81	65.00	23.12	41.18	30.61	52.89

	AIME24		AIME25		Macro	
Model	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8
EOPD + hard-mix	10.42	26.67	4.58	16.67	29.79	48.70
OPD + hard-mix	10.42	23.33	5.83	16.67	29.93	48.69
soft-entropy + hard-mix	8.33	16.67	5.00	13.33	29.49	47.42
eopd-value + hard-mix	8.33	20.00	4.58	10.00	28.82	45.98

erage cost (macro Pass@8 48.69 vs. 48.62). Its gains concentrate on the two large benchmarks (MATH500 $+2.43$, AMC23 $+1.88$). Both FKL-based runs lose coverage broadly. EOPD-hardmix drops Pass@8 on four of six benchmarks (macro -2.40). Soft-entropy-hardmix drops on four of six with one flat and one up (macro -2.58). Losses concentrate on AIME24/AIME25, where the mixed-in hard problems make the marginal coverage most expensive. The per-benchmark pattern rules out a one-benchmark artifact and supports the capacity-competition reading of Section 5.3. Harder prompts consume the same student capacity that the forward-KL term spends on preserving teacher branches. Data difficulty and coverage routing therefore substitute rather than stack.

G Training Loss Convergence

Figure 10 plots the per-step optimization quantities of every formal run in the study (18 runs, 174 steps each). The series are assembled from the trainer logs of the two compute hosts (Appendix B). The runs are the retained method–seed combinations of EOPD, OPD, and the soft gate. Two of the three TSA variants, the three CTS doses, and the three data-composition runs make up the rest. As a parsing check, the series extracted for the formal EOPD run reproduce the corresponding statistics in Appendix H exactly. The matched statistics are entropy floor 0.354 ± 0.025, on-policy accuracy 0.641 ± 0.042, and FKL token ratio $12.25\% \pm 0.86\%$.

Run inventory. The 18 series decompose as: EOPD seeds 41–44 (4), OPD seeds 41/42/43 (3), and soft-entropy seeds 41/43 plus the formal λ=0.20 check (3). They also include the TSA JS-mass and TSA-full variants (2) and CTS cover-58/87/116 (3). The hard-mix data-composition runs of OPD, EOPD, and the soft gate (3) complete the inventory. Two further runs predate the log-retention convention and retain no per-step trainer logs (the λ=0.15 soft-gate seed-42 formal and the TSA entropy-JS variant, see the scope notes below). Their evaluation artifacts are complete and feed every table in the paper. Every family they belong to is represented here by its other members.

Every gated variant settles to a stable floor within 30 steps. The gated forward-KL term drops from $\mathcal{O}(1)$ at the first step (1.3–1.5 for the entropy-gated family) to $\approx 4 \times 10^{-2}$ by step 30 and then fluctuates without directional trend. Over the last 60 steps the entropy-gated family (both soft-gate replicas, the λ=0.20 run, the three CTS doses, and EOPD) holds 0.036–0.038. Its coefficient of variation (CV) is 0.08–0.10. The TSA variants floor an order of magnitude lower, at ≈ 0.004. That level is set by their mass-renormalized definition of the gated set rather than by optimization quality. Their higher relative scatter (CV ≈ 0.25) amounts to an absolute fluctuation of $\approx 10^{-3}$. The cover→sharpen transition of the CTS runs is invisible in the loss. Within three steps of each phase boundary the FKL term stays inside 0.032–0.045 at all three doses, and the gradient norm inside 2.2–3.2. The routing handoff produces no optimization transient.

The shared surrogate descends on one schedule, and gradient norms are benign. Panel (c) shows the clipped-surrogate loss $\mathcal{L}^{\mathrm{OPD}}$ of Eq. equation 2 that every method optimizes. The four

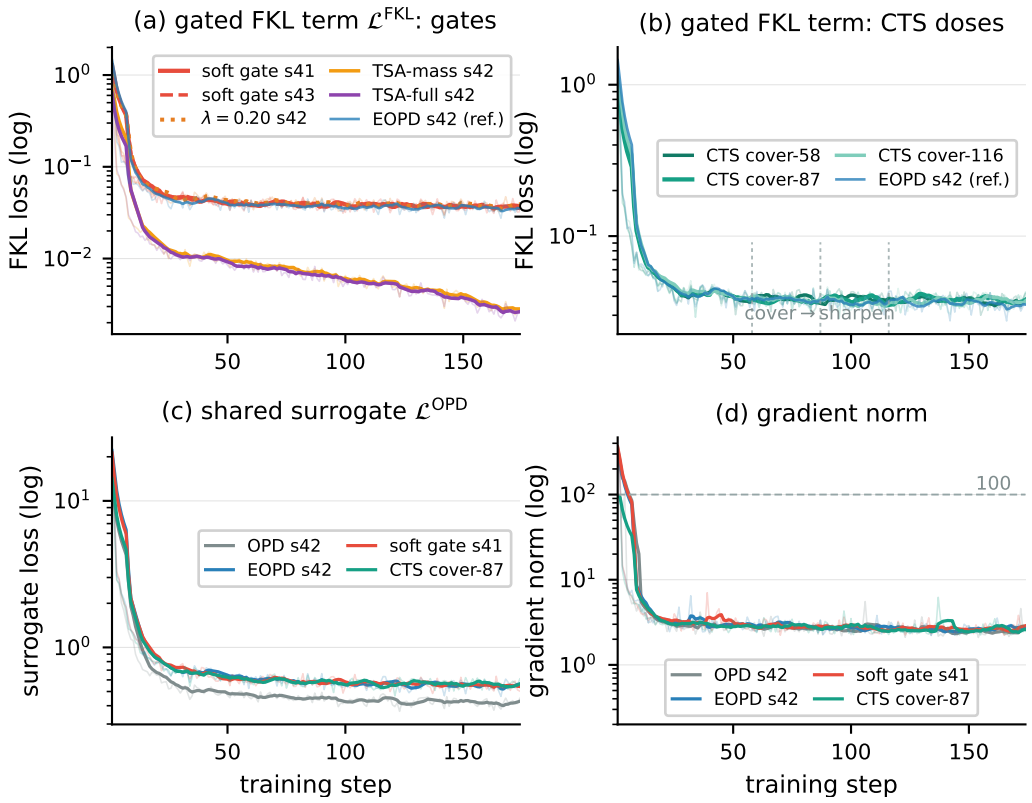

Figure 10: **Training-loss convergence across the formal runs.** Per-step optimization quantities from the trainer logs (174 steps, bold curves are 7-step trailing means, log scale on all panels). (a) Gated forward-KL term for the gate variants. (b) The same for the three CTS doses (dashed lines: cover→sharpen transitions). (c) The shared clipped-surrogate loss. (d) Gradient norm (dashed line: 100). OPD carries no forward-KL term and is absent from (a) and (b).

families descend together from step-1 values of 12–22 to late means of 0.43–0.57. Between-method differences at any step are small relative to the shared descent. Panel (d) shows the corresponding gradient norms. After the early spike decays, no run in the 18-run sweep exceeds 19.6 at any step past 30 (the reference line at 100 is never approached). No method's evaluation rests on an unstable optimization phase, early or late in the schedule.

Scope and omissions. Three boundary conditions apply. First, OPD carries no forward-KL term (Eq. equation 2) and is absent from panels (a) and (b) by construction, not by omission. Second, FKL levels are not comparable across gate families. The term is computed on each method's own gated token set under the shared top-16 renormalization (Eq. equation 4), and λ rescales it. Panel (a) therefore reads each run against its own floor. We use these series for convergence, not for cross-method magnitude comparisons. Third, two first-wave runs (the λ=0.15 soft-gate seed-42 formal and the TSA entropy-JS variant) predate the log-retention convention and retain no per-step trainer logs, as the run inventory notes. The soft gate is thus represented here by its two full seed replicas and the λ=0.20 run.

H Training-Regime Stability

A natural objection to any 174-step comparison is duration: the schedule is three epochs. One might argue that gates need longer training to overtake EOPD, or that EOPD itself is still climbing when we stop. This appendix uses the stored training series of the formal EOPD run, together with the pilots and Figure 12, to show two things. The regime the comparisons occupy is stationary. Moreover, "train longer" is not an unmeasured axis of our study: it is the budget axis the paper already measures.

The routing signal is stationary for most of the run. Student entropy starts at 2.72 and falls below ≈ 0.45 within the first ten steps. It then floors. Over the last 60 steps it is 0.354 ± 0.025. The ordinary least squares (OLS) slope from step 20 on is -1×10^{-4} per step, i.e. no directional trend over the final 89% of the schedule. EOPD therefore routes on a converged statistic long before the evaluation point. The soft-gate pilot's logged window sits at the same floor (0.349 ± 0.026 over its last twenty steps), while the utility pilot floors visibly lower (0.200 ± 0.018). This ordering of entropy floors mirrors the coverage results (the utility gate collapses exploration hardest), which supports the budget mechanism rather than an artifact of one series. Figure 11a plots the three series with their floors and trend diagnostics. Panels (b) and (c) apply the same treatment to on-policy accuracy and to the cumulative FKL budget, over the same 174-step window.

On-policy accuracy saturates early and does not climb late. Training accuracy rises from 0.12 to 0.62 by step 20 (OLS slope $+0.0028$ per step over steps 0–59) and peaks at 0.72 around step 25. From step 60 onward the OLS slope rounds to 0.0000 per step. The last sixty steps fluctuate at 0.641 ± 0.042 around no trend. We phrase this as saturation with fluctuation rather than a clean plateau. The late-phase scatter is minibatch noise of roughly four points, and the late mean sits slightly *below* the early peak. No method's advantage is still emerging at the evaluation point. If anything, a longer run would evaluate deeper into a flat-to-slightly-declining regime.

Duration is the budget axis, and it is measured. The FKL token ratio holds at $12.25\% \pm 0.86\%$ over the last sixty steps, above its run-wide average. Each additional step accrues coverage-transfer budget at a constant rate. The dose experiment of Section 5.4 varies exactly this axis ($58/87/116/174$ cover steps, i.e. one to three epochs). Macro Pass@8 rises with the cover budget across the measured doses (the 87-vs-116 difference of 0.24 points is inside single-run noise). Gate-vs-gate orderings stay flat at every budget measured. Asking whether a gate would win "with more training" is therefore, in this design, asking whether gate advantages appear at larger budgets. The dose data answer that the FKL endpoint converts budget into coverage across the measured range, with no routing rule changing the exchange rate. All method comparisons are additionally schedule-matched by construction: every arm at a given budget sees identical steps, data, and optimization settings.

We cannot exclude that a qualitatively different regime appears at much longer horizons (for instance under learning-rate schedules beyond three epochs). Within the protocol family of the original recipe (Appendix C), the comparisons above sit in a stationary regime. In that regime, the budget, not the calendar, is the operative variable. The optimization-side companion of this section is Appendix G, which plots the training losses and gradient norms of all 18 formal runs.

Reading Figure 12. Two features of the companion dynamics figure deserve emphasis. First, panel (a) shows that routing does not change what is learnable per step. The on-policy accuracy curves of the three routing regimes climb together and saturate at the same level. No gate buys optimization speed. The gates differ only in how the learned probability mass is shaped. Second, panel (b) shows that the soft-entropy gate slightly *slows* the student-entropy collapse relative to EOPD. This is exactly what a smoother coverage term should do. By letting a fraction of forward KL through on mid-entropy tokens, the soft gate keeps the student's own entropy marginally higher for longer. The effect is small and does not change the stationary end state. Yet it is the signature of the gate acting as a continuous valve rather than a switch. It is visible at no cost to the converged regime.

I Problem-Level Solvability Analysis

Aggregate metrics average over $1{,}547$ problems. A reviewer may fairly ask what such an aggregate hides. The effect might be concentrated on a few problems. Or individual problems might churn in both directions. A net number would disguise that churn. This appendix recomputes the key comparisons as problem-level counts. Such counts are insensitive to how one aggregates and disclose both directions by construction. All counts come from the same stored $n{=}8$ generations as every other metric in the paper, with no resampling and no excluded problem.

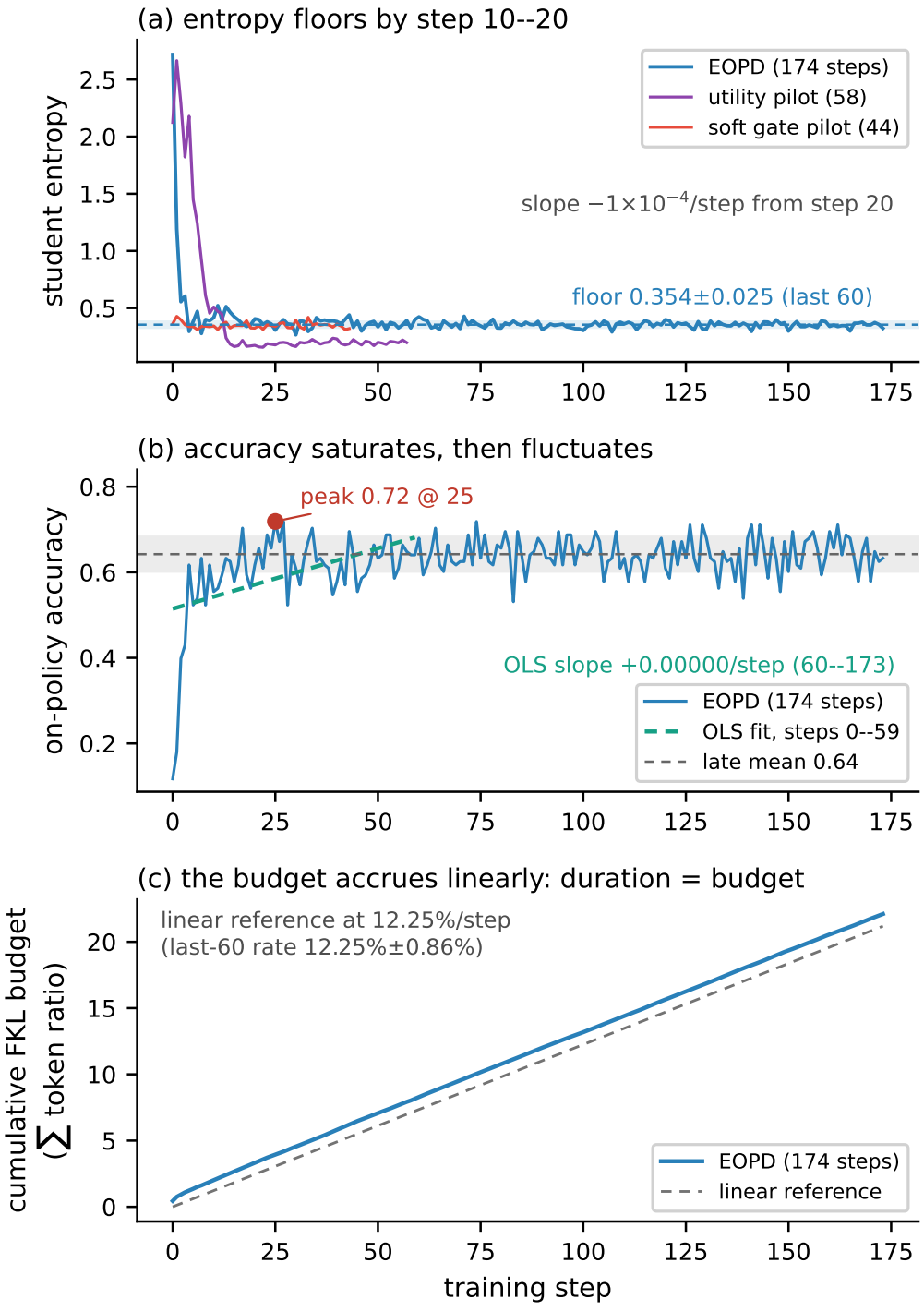

Figure 11: **Training-regime diagnostics for the formal EOPD run.** (a) Student entropy for the formal run and both pilots: all three floor within 10–20 steps (shaded band and dashed line: last-60 mean ± one sd, gray annotation: OLS slope from step 20). (b) On-policy accuracy with the early OLS fit, late mean band, and early peak (slope 0.0000/step over steps 60–173). (c) Cumulative FKL budget against a linear reference at the last-60 token-ratio rate: the budget accrues at a constant rate, so duration and budget are the same axis.

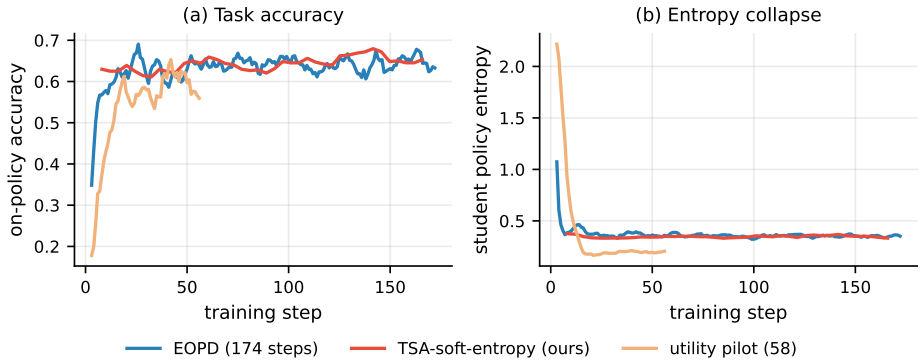

Figure 12: **Training dynamics.** (a) On-policy accuracy climbs identically across routing regimes, and routing does not change what is learnable per step. (b) Student entropy collapses within the first 10–20 steps under all variants, and the soft-entropy gate slightly slows the collapse relative to EOPD.

Method. For each problem we observe the number of correct generations $c \in \{0, \ldots, 8\}$. A problem is *touched* by a method if $c \geq 1$ (the problem has entered the support: some sampled trajectory solves it). It is *reliably solved* if $c \geq 4$ (the majority-of-budget regime that Avg@8-type metrics weight). For a pair of runs (A, B) we report the number of problems only B touches (+ only), only A touches (− only), and their difference. The same counts are reported at the reliable threshold. Problems are aligned by benchmark and index across runs (identical question sets, verified by hash).

Estimator notes. Three properties make these counts robust where rates are not. First, they are integer counts over the full pool with no sampling. Every figure in this appendix is exact for the stored generations, not an estimate. Second, the two thresholds play complementary roles. The touched threshold is the per-problem form of Pass@k (entry into the support). Its partner, the reliable threshold ($c \geq 4$ of 8), approximates the regime Avg@8 weights most. The touched and reliable column groups of Table 14 therefore map one-to-one onto the Pass@8-versus-Avg@8 distinction of Section 6, without invoking any rate computation. Third, the bottom block of Table 14 recomputes the same transitions for same-method seed pairs, a churn reference with zero method difference mixed in. Any method net inside the seed envelope is indistinguishable from re-running the same method with another seed. The hash check on identical question sets above already confirms that aligned problem indices refer to identical prompts across runs. And the cross-run recurrence analysis of Appendix J is computed on the same alignment.

Table 14: Problem-level solvability transitions over the six-benchmark pool (1,547 problems, n=8 per problem). "+ only" counts problems the second run reaches that the first never does, and "− only" the reverse. Top block: method comparisons on the shared seed. Bottom block: same-method seed pairs, which measure pure seed churn with no method difference mixed in.

	Touched (≥1/8 correct)			Reliable (≥4/8 correct)		
Comparison	+ only	− only	net	+ only	− only	net
EOPD vs OPD (s42)	95	71	+24	51	56	-5
soft gate vs EOPD (s42)	73	75	-2	58	47	+11
EOPD s41 vs s42	74	77	-3	62	49	+13
EOPD s43 vs s42	81	66	+15	61	58	+3
soft gate s41 vs s42	78	84	-6	61	66	-5
soft gate s43 vs s42	88	75	+13	58	55	+3
OPD s41 vs s42	87	80	+7	80	57	+23

The contested comparisons, counted. Three facts emerge. First, the only net movement that exceeds the seed-churn envelope is the one real ordering in the paper. EOPD touches 24 more problems than OPD on the shared seed (95 vs. 71 one-directional flips), against a same-method seed envelope of -6 to $+15$. At the reliable threshold the two endpoints do not separate (51 vs. 56, net -5). The forward-KL budget buys entry into the support (tail coverage). It does not convert marginal problems into reliably solved ones. This is the problem-level form of the Pass@k-versus-Avg@8

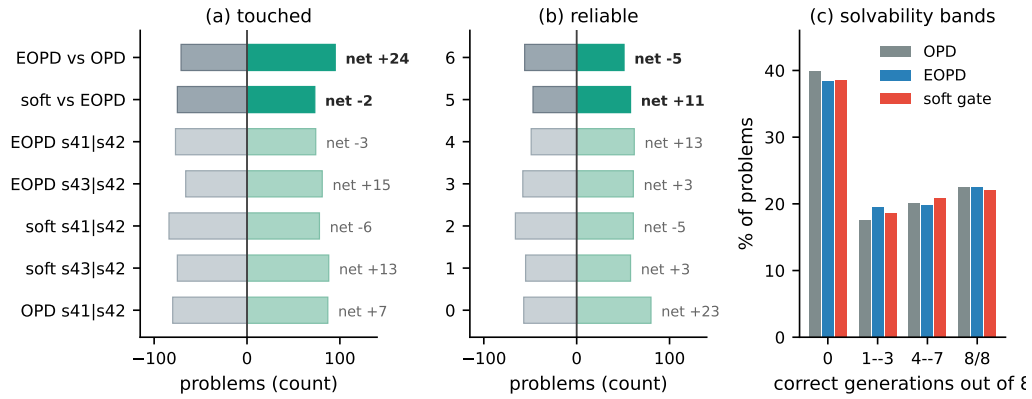

Figure 13: **Solvency churn, both directions, against the seed envelope.** (a, b) One-directional transition counts of Table 14 at the touched and reliable thresholds (green: only the second run reaches, gray: the reverse, light fills: same-method seed pairs): EOPD's net +24 touched over OPD is the only number outside both seed envelopes. (c) Correct-count band distribution per run (Table 15): the bimodal structure is nearly frozen across methods and seeds.

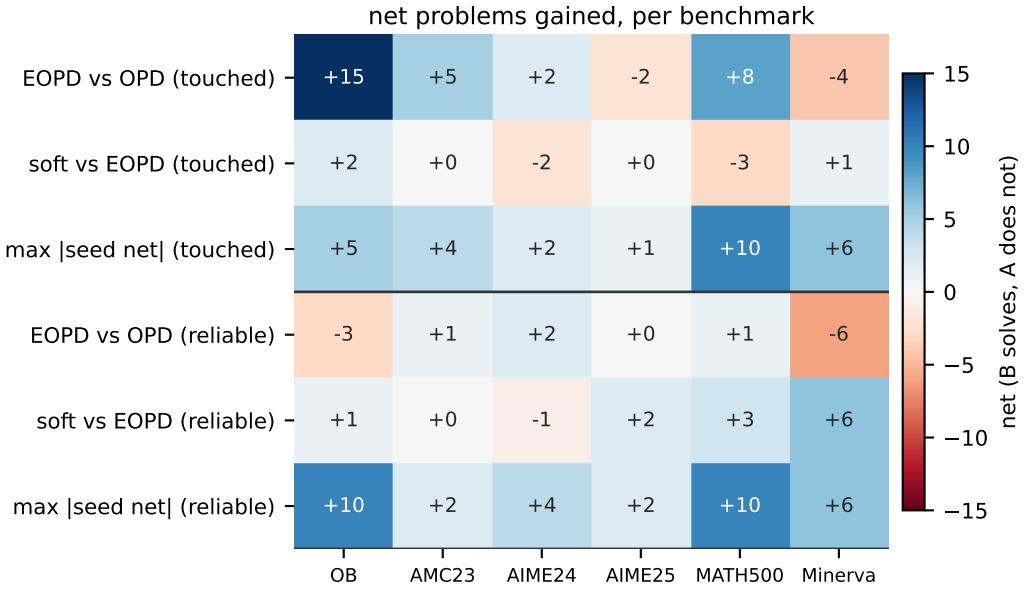

Figure 14: **Net transitions per benchmark.** Each cell is the net problem count of Table 14 restricted to one benchmark (blue positive, red negative). The bottom block of each group shows the largest same-method seed net on that benchmark and threshold as the churn reference. EOPD's touched gains concentrate on the two large benchmarks (OlympiadBench +15, MATH500 +8), while every AIME cell stays within two problems of the seed rows. At the reliable threshold no net exceeds the seed envelope in magnitude on any benchmark in the pool.

Table 15: Correct-count band distribution per run (1,547 problems). The pool is bimodal: roughly 38–40% of problems are never solved and 22–23% are solved 8/8, and the contested middle band moves by a few problems between methods and seeds.

Run	0 correct	1–3	4–7	8/8
OPD (s42)	617	272	311	347
EOPD (s42)	593	301	306	347
soft gate (s42)	595	288	322	342
EOPD (s41)	596	285	323	343
EOPD (s43)	578	313	309	347
OPD (s41)	610	256	326	355
soft gate (s41)	601	287	304	355
soft gate (s43)	582	298	306	361

split in Section 6. Second, the soft gate versus EOPD flips 73/75 (touched) and 58/47 (reliable) on the shared seed, squarely inside the same-method seed envelope. The gate neither adds nor destroys support beyond what re-running the same method with another seed does. Third, the per-run distributions (Table 15) are nearly frozen across methods and seeds. All macro-metric movement in this design space rides on a few dozen problems in the 1–7-correct middle band. Figure 13 shows all three facts in one view. The method bars sit inside the seed envelope in both directions, except for EOPD's touched net. Its band histogram is visually identical across the three methods and every seed replica of each.

Reading the per-benchmark nets. Figure 14 localizes every net to one benchmark. EOPD's touched gains over OPD concentrate on the two large pools (OlympiadBench +15, MATH500 +8). These are exactly where the forward-KL budget has the most marginal problems to convert. The four smaller benchmarks contribute single problems at most. No AIME cell, on either threshold and for any method comparison, carries more than two problems of net in either direction. This discreteness is why per-benchmark AIME rates in Table 1 swing by several points between seeds. The aggregate ordering stays put. At the reliable threshold every cell of every method comparison sits within the corresponding seed row. This confirms that the budget buys entry into the support, not reliable majorities. The churn envelope, not the method, is what moves the middle band.

Why AIME-style benchmarks swing. AIME24 and AIME25 have 30 problems each. A single flipped problem therefore moves any per-benchmark rate by 3.3 points, and a touched-count change of 5 problems spans 16.7 points. The 13.33–30.00 range that λ and τ pilot sweeps show on AIME24 (Section 5.4) corresponds to between 4 and 9 touched problems out of 30. This is the discrete view behind our decision to read pilot AIME numbers as noise rather than signal. It is also why we demand multi-seed evidence before reading any ordering as real. Figure 14 confirms the point per benchmark. The aggregate nets are carried by the large pools, and every AIME cell stays within two problems of net.

J Solution-Mode Diversity Analysis

Pass@8 and the transition counts of Appendix I measure *whether* problems get solved. A separate objection concerns *how*: pruning the forward-KL term to a gated subset of tokens could in principle collapse the student's distribution over solution styles. Equal counts would then conceal a poorer mixture of solutions. This appendix tests that directly on the stored n=8 generations. We cluster each problem's correct generations by near-duplicate detection. Two generations belong to the same *mode* when their word 8-shingle Jaccard similarity is at least 0.7. The mode count is the number of clusters. As a distribution-free complement we also report the distinct 3-gram ratio among correct generations (a standard lexical-diversity measure that does not depend on any clustering choice). Lowering or raising the near-duplicate threshold between 0.5 and 0.9 moves every mode count by less than 0.1, so the conclusions below do not rest on the threshold.

Computation details. Shingles are computed on whitespace-tokenized lowercased text after stripping boxed-answer markers and math delimiters. Two solutions differing only in surface formatting then count as near-duplicates, and genuinely different derivations do not. Clustering is single-linkage

Table 16: Solution-mode diversity from stored generations (six benchmarks, n=8). "touched" counts problems with at least one correct generation, and "modes" are near-duplicate clusters of correct generations. Method differences in every column are smaller than the same method's seed spread, so no column separates the methods.

Run	touched	correct/touched	modes/touched	≥2 modes	3-gram div.
OPD (s42)	930	5.39	5.37	86.1%	0.640
EOPD (s42)	954	5.30	5.29	85.2%	0.659
soft gate (s42)	952	5.37	5.34	86.1%	0.658
EOPD (s41)	951	5.37	5.35	86.8%	0.658
EOPD (s43)	969	5.30	5.28	86.5%	0.667
OPD (s41)	937	5.48	5.46	87.5%	0.638
soft gate (s41)	946	5.38	5.36	86.4%	0.659
soft gate (s43)	965	5.33	5.30	85.3%	0.663

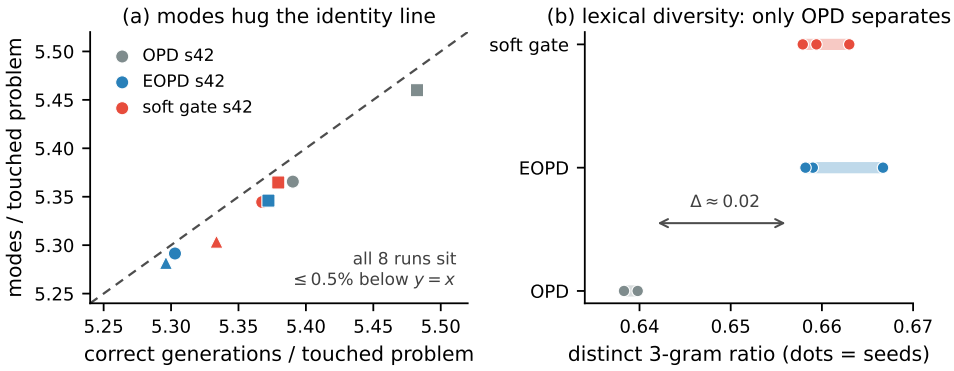

Figure 15: **Mode counts versus lexical diversity, all eight runs.** (a) Modes per touched problem against correct generations per touched problem: all eight runs (markers distinguish seeds, colors methods) sit within 0.5% of the identity line (dashed), i.e. near-duplicates are rare under every gate. (b) Distinct 3-gram ratio per run (dots are seeds, bars the seed spread): only OPD separates from the FKL-carrying variants, by a direction-consistent ≈ 0.02, and the soft gate and EOPD overlap. Coverage differences live on the problem axis (Appendix I), not on these axes.

within each problem's correct set, which makes the mode count the number of connected components at the threshold. Single-linkage can chain, and the threshold sweep above is precisely the control for that sensitivity. The distinct 3-gram ratio is $|\{\text{distinct 3-grams}\}|/|\{\text{total 3-grams}\}|$ over the concatenation of a problem's correct generations, then averaged over problems. All quantities are computed per run per problem and only then aggregated, so no problem's longer generations dominate the pool.

Table 16 gives the per-run summary. Three observations. First, there is no mode collapse anywhere. 85–88% of touched problems carry at least two distinct modes under every method. Modes per touched problem (5.28–5.46) sit within half a percent of raw correct generations per touched problem (5.30–5.48). The within-run near-duplicate rate is thus at most 0.5%. Sampling at generation temperature already produces near-distinct chains-of-thought, and none of the gates changes that. Second, the method differences are inside seed churn. The soft gate's 5.34 modes per touched problem against EOPD's 5.29 (+0.05) is smaller than EOPD's own three-seed spread (5.28–5.35). This holds for every pairwise comparison in the table. Third, OPD produces the *most* modes per touched problem (5.37/5.46) while touching the *fewest* problems (930/937 vs. 951–969 for EOPD). Whatever separates these methods lives on the problem axis, not on the per-problem mode axis. The single metric with any separation is lexical. 3-gram diversity is 0.658–0.667 for all FKL-carrying variants against 0.638–0.640 for both OPD seeds, a direction-consistent gap of roughly 0.02. We read this gap as the forward-KL term's footprint on phrasing variety. It does not convert into extra modes, and the soft gate sits level with EOPD on it (0.658 vs. 0.659). Figure 15 shows both axes at once. Every run hugs the identity line in panel (a), and panel (b) isolates the one lexical separation.

Table 17 asks whether one run's solutions *recur* in another's. Recurrence is between 0.26% and 0.66% for every pair, including same-method seed pairs. No method pair falls below the seed

Table 17: Cross-run generation recurrence: fraction of one run's correct generations that have a near-duplicate (Jaccard ≥ 0.7) among the other run's, over jointly touched problems. Method pairs (bold) recur at the same rate as same-method seed pairs.

Comparison	joint	r_{12}	r_{21}
EOPD vs OPD (s42)	859	0.53%	0.47%
soft gate vs EOPD (s42)	879	0.63%	0.66%
EOPD s41 vs s42	877	0.47%	0.50%
EOPD s43 vs s42	888	0.46%	0.44%
soft gate s41 vs s42	868	0.26%	0.26%
soft gate s43 vs s42	877	0.40%	0.44%
OPD s41 vs s42	850	0.61%	0.48%

range (soft gate vs. EOPD: $0.63\%/0.66\%$, the two highest values in the table). Swapping the objective neither preserves nor destroys individual solutions relative to reseeding. Combined with Appendix I, this locates the budget principle: the FKL budget governs *which problems enter the solved set* (EOPD's net $+24$ touched problems). The number of distinct solutions per solved problem and the recurrence structure across runs are invariant to routing at the resolution our seeds provide. This holds on both the mode and recurrence axes.

K LIMITATIONS

The limits fall into three classes: statistical resolution, protocol scope, and the bounds on the mechanism and cost arguments. (i) Endpoint-level orderings are resolved by the seed design: four seeds for EOPD and three for the soft gate, OPD, and the half-schedule curriculum. The EOPD$-$OPD touched-problem margin exceeds the same-method seed envelope (Appendix I). Equivalence claims carry named margins. The soft gate is certified equivalent to EOPD at ± 3 macro Avg@8 points and at ± 2.2 Pass@8 with $n{=}16$ generations (Appendix F). Below ~ 0.5 macro points the design has no power, and no routing claim rests on an ordering that small (Section 6). No Pass@8 comparison at $n{=}8$ is certifiably equivalent within ± 3, so coverage parity at that budget stays unresolved. The coverage-lead reading rests on the seed ordering, all four EOPD seeds above all three gate seeds, not on an exclusion of parity. The same bound applies symmetrically to the original report's single-run comparisons.

(ii) All comparisons live inside one pinned protocol family: a shared student, a shared teacher, rule-checkable mathematical reasoning, and one evaluation stack. The budget principle is a finding about this family. Rule-based scoring stores 8 generations per problem, 16 for one robustness check, and 128 for the spectra. Generation-level variance is symmetric across methods and is not removed by the paired bootstrap (Appendix L). Endpoint levels sit 1.7–1.9 Avg@8 points below the published record while the comparative deltas are preserved (Appendix C). No within-paper comparison mixes stacks. Public-benchmark contamination of the MATH training split is inherited from the original protocol. The one data intervention we added, the DeepMath hard-mix, is triple-decontaminated at all three stages (Appendix A). The dose–response covers four cover budgets on one schedule shape, so the insensitivity findings of Section 5.4 hold within the swept ranges. Transfer to other scales, modalities, or non-verifiable domains is untested, and we treat it as unmeasured rather than unlikely.

(iii) The mechanism and cost arguments carry their own bounds. All forward-KL terms use one teacher under the fixed top-16 truncation inherited from the original recipe. The truncation is symmetric across every method compared, so within-paper orderings are internally valid, while absolute coverage levels depend on the teacher's truncated support. The budget accounting turns that dependence into a testable prediction (Appendix L). Pass@k should track the measured FKL token ratio, not the routing pattern, under any teacher swap. Verifying it requires retraining arms we have not run. The grouped-rollout confound of the utility family is bounded rather than open. Its $n{=}4$ schedule with 32 prompts per step mixes the gate's effect with a fourfold drop in distinct prompts. Two formal-scale controls separate the mix (Section 4). The prompt-diversity drop costs roughly 0.8–1.4 coverage points with or without FKL, and the remaining loss attributes to the $(1-p)^n$ closures. On cost, the efficiency numbers are per-step wall times on one hardware configuration (Appendix A). The soft gate adds no teacher forward pass and no rollout, and the utility family trades prompt diversity, not wall time, for its estimates. No cross-hardware or deployment-scale extrapolation is

made. No headline claim rests on the unbounded part of any entry, and each class above names the experiment that would move its boundary.

L Scope-Limited Formal Analysis

The analysis in this appendix is intentionally local. It concerns expectation-level statements about a distillation objective with binary outcome rewards and exchangeable rollouts. Optimizer dynamics are not modeled. The estimation error of entropy or utility estimates is not modeled either, nor is the interaction between the FKL and RKL terms. No claim is made that no gate can ever improve coverage. Its role is to give the two mechanisms the paper relies on a precise statement, together with the assumptions they actually need. These mechanisms are the $(1-p)^n$ miscoverage event and the FKL budget accounting. The empirical sections can thus cite them without over-claiming.

Evaluation estimators. With n stored generations per problem and c correct ones, we estimate $\mathrm{Pass@}k$ by the standard unbiased estimator $1-\binom{n-c}{k}/\binom{n}{k}$. This is an expectation over uniformly random k-subsets. It is used for the n=8 six-benchmark metrics and, with n=128, for the spectra in Figure 3c. We estimate $\mathrm{Avg@}8$ by $c/8$. All comparisons in the paper use the same estimator on both sides, so estimator variance enters symmetrically. The paired bootstrap of Appendix F resamples problems, not generations. It therefore does not remove generation-level variance. That variance is part of why the n=16 resample in Section 5.4 moves the nominal gap.

Proposition 1 (miscoverage identity). *Fix a prompt whose success probability under the current student policy is p. If n rollouts are drawn independently, the probability that none is correct is $(1-p)^n$. Under positively associated failures it is at least $(1-p)^n$. In particular, a hard gate that requires at least one correct rollout to release any signal releases none with probability $(1-p)^n$. This is 0.66 at p=0.1, n=4 and still 0.32 at p=0.25.* The identity is exact for independent rollouts. Sampled rollouts from one policy are positively correlated in practice. This makes the realized all-wrong event *more* likely than the independence calculation. Thus $(1-p)^n$ is a lower bound in the direction that strengthens the argument. We use it as an interpretation of the observed per-step closure rates (28–49%, Figure 3a), not as a formal guarantee about dependent decoding. The identity deliberately leaves dependent decoding unmodeled.

Proof sketch of Proposition 1. For independent rollouts the claim is immediate: each rollout fails with probability $1-p$, so the joint all-wrong event has probability $\prod_{i=1}^{n}(1-p) = (1-p)^n$. The dependent case needs an added assumption, which we state. By definition, $\Pr[\text{all wrong}] = \mathbb{E}\big[\prod_{i=1}^{n}\mathbb{I}[r_i{=}0]\big]$ holds for the joint failure event. Positive association of failures gives $\Pr[\text{all wrong}] \geq \prod_i \Pr[r_i{=}0] = (1-p)^n$ by the standard positive-dependence bound (the FKG-type inequality applied to the decreasing events $\{r_i = 0\}$). This is the empirically relevant direction for same-policy decoding, where a hard problem makes every rollout harder simultaneously. The independence value is therefore a lower bound in the direction that strengthens the mechanism: realized closure rates should sit at or above $(1-p)^n$. This is what the training logs show (28–49% per step, overlapping the benchmark-derived closure estimates of Appendix F). Numerically, the bound is far from vacuous on hard problems: $(1-0.1)^4 = 0.6561$ and $(1-0.25)^4 = 0.3164$. Thus between a third and two thirds of the hardest groups receive no utility signal at any given step.

Proposition 2 (gain mass concentrates where gates close). *Raising p has marginal Pass@k value $g_k(p) = \frac{\partial}{\partial p}\left[1-(1-p)^k\right] = k(1-p)^{k-1}$, strictly decreasing in p on $(0,1)$. The closure probability $(1-p)^n$ is also decreasing in p. Over any distribution of solvable problems, the two are comonotone. Hence the gain-weighted closure share exceeds the unweighted closure rate.* The last step is Chebyshev's sum inequality: for comonotone functions $a(\cdot), g(\cdot)$ of the same variable, $\mathbb{E}[a\,g] \geq \mathbb{E}[a]\,\mathbb{E}[g]$. This is the formal license for the counterfactual estimate in Appendix F. The gain-weighted share of marginal Pass@8 mass on gate-closed problems (0.48–0.56) sits above the unweighted closure rates (15.9–43.3% per benchmark). This is because the weight g_8 concentrates in the low-p region where closure is most likely. Two caveats bound the estimate. First, $\hat{p}$ from 8 samples is attenuated in the low-$\hat{p}$ tail where the weight concentrates. Second, benchmark problems proxy rather than sample the training mix. The bracketing check against training-log closure rates is the empirical control for both. Figure 16 draws the two functions of Propositions 1 and 2 on one

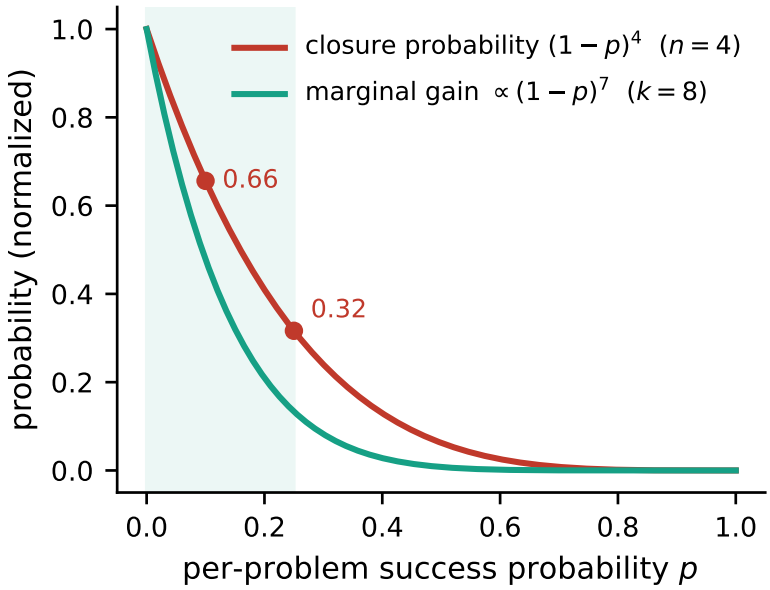

Figure 16: **The two functions behind Propositions 1 and 2.** Hard-gate closure probability $(1-p)^4$ over a fresh group of n=4 rollouts (red) and marginal Pass@8 gain $\propto (1-p)^7$ at k=8 (teal), against the per-problem success probability p. In the shaded low-p region the gate closes with probability 0.66 at p=0.1 and 0.32 at p=0.25 (marked), which is the same region where the marginal gain concentrates: gates close where the gains live.

axis. Closure probability and marginal gain are comonotone in p, so gating decisions land exactly where the Pass@8 gains live, across the full range of p.

Proof sketch of Proposition 2. The derivative computation is direct: $\frac{\partial}{\partial p}\left[1-(1-p)^k\right] = k(1-p)^{k-1}$. Its p-derivative $-k(k-1)(1-p)^{k-2} < 0$ on $(0,1)$ gives strict decrease. $(1-p)^n$ is likewise strictly decreasing. Two strictly decreasing functions of the same variable are comonotone (both move in the same direction as $1-p$). Chebyshev's sum inequality for comonotone functions then gives $\mathbb{E}[(1-p)^n g_k(p)] \geq \mathbb{E}[(1-p)^n]\,\mathbb{E}[g_k(p)]$ under any problem distribution, so the gain-weighted closure share exceeds the unweighted rate. The two caveats of the main text bound how far the estimate can be pushed. First, $\hat{p}$ from 8 samples is attenuated in the low-$\hat{p}$ tail where g_8 concentrates. Second, benchmark problems proxy rather than sample the training mix. We therefore report the gain-weighted share as a bracket (0.48–0.56 across the three hard benchmarks) corroborated by the training-log closure rates, not as a point identity.

Proposition 3 (budget accounting is an identity, not a law). *Let $\alpha_t \in [0,1]$ be a gate coefficient applied to the FKL term at token t. The FKL budget of a run is $B = \sum_{steps} \sum_t \alpha_t$, an accounting quantity. At fixed data and optimization, the FKL contribution to the parameter updates is supported on exactly the token set where $\alpha_t > 0$. Its magnitude is linear in each individual coefficient α_t. Nothing about the ordering of final metrics between two gates follows from this identity alone.* We state it because of the empirical content of Sections 6 and 5.3. Measured outcomes track B (normalized in Figure 3b: 12.0% vs. 4.2–8.1% of tokens) far more closely than they track the routing pattern that spends it. We deliberately stop short of formalizing "coverage transfer is budget-limited" as a theorem. Such a statement would require the curvature of per-problem returns to FKL spend, which our n=4 rollouts cannot measure. Section 7 discusses what would have to be estimated (per-problem return curvature) to make it one. The propositions above are the part of the story that is true by arithmetic, and the budget principle is the part we establish empirically and scope accordingly.

Proof sketch of Proposition 3. Write the gated FKL contribution to the loss as $\sum_t \alpha_t \mathcal{L}_t^{\mathrm{FKL}}$. Its gradient in the student parameters is $\sum_t \alpha_t \nabla_\theta \mathcal{L}_t^{\mathrm{FKL}}$, which vanishes on every token with $\alpha_t = 0$ (support claim). It is linear in each α_t with coefficient $\nabla_\theta \mathcal{L}_t^{\mathrm{FKL}}$ (magnitude claim). Doubling every

gate coefficient exactly doubles the FKL-side gradient contribution. The token set it touches does not move. These are properties of the objective rather than of the optimizer, the data order, or the metrics. This is why no final-metric ordering follows from the identity alone. The mapping from per-step gradients to end-of-training metrics passes through optimizer dynamics, the interaction with the reverse-KL term, and the evaluation protocol, none of which the identity constrains. What the identity licenses is the accounting vocabulary itself. Comparing runs by their accumulated B (equivalently, by the gated token ratio of Figure 3b) is comparing like with like. That outcomes track B is an empirical statement about the return surface of this protocol family, which we measure rather than assume.

Remark (teacher-side truncation). The budget identity counts student-side gate coefficients. On the teacher side, the top-k=16 truncation inherited from the original recipe is fixed and applied identically to every method. Within this paper, k_{teach} therefore acts as a shared constant. It scales the absolute forward-KL signal available to all gates by the same factor and cannot favor one routing rule over another. This constant does bound the *level* of transferable coverage (a fuller teacher support carries more distinct signal). That is why absolute Pass@k levels should not be read as teacher-independent. The budget reading makes a falsifiable prediction here. Under a different teacher or a different truncation width, method orderings should continue to track the measured FKL token ratio (Figure 3b) rather than the routing pattern. Testing that prediction is outside our compute scope. It is listed accordingly in Appendix K, together with its bounding evidence.

M Pass@k Spectrum Tables

Figure 3 in Section 6 plots the three-benchmark macro Pass@k spectrum (mean of the AMC23, AIME24, and AIME25 rates) with bootstrap uncertainty. Table 18 gives the underlying values for the six n=128 evaluation runs (seed 42 throughout). Every curve point can be read to one decimal. Columns run from k=1 (the precision end, identical to Avg@1) to k=128 (the coverage end). The estimator is the unbiased $1 - \binom{n-c}{k} / \binom{n}{k}$ over the n=128 stored generations of each problem. Two runs (TSA-soft-entropy and TSA-CTS cover-116) have their $k \in \{1, 2, 4\}$ entries recomputed from the same stored 128-sample generations. Their summary files only aggregate $k \geq 8$.

Reading the tables. Three patterns are visible without statistics. First, the low-k end is essentially flat across methods. At k=1 every method sits within 2.2 points of every other on every benchmark. This is the table form of the precision-side equivalence measured formally in Table 9. Second, the curves fan out with k. EOPD's lead over the soft gate and OPD opens up from k=8 onward on AIME24 and AIME25. For example, AIME25 Pass@128 is 46.7 vs. 40.0 for the soft gate and 46.7 vs. 36.7 for OPD. That is the coverage side of the budget principle. Third, isolated per-benchmark crossings exist at single k values. On AMC23 the soft gate sits nominally above EOPD at $k \leq 8$ (e.g. 71.5 vs. 71.3 at k=8). These are exactly the cells the pooled bootstrap bands of Figure 3 mark as noise. 40-problem benchmarks move 2.5 points per flipped problem. The pooled ordering (EOPD nominally above every gate at every $k \geq 4$) survives aggregation.

The pooled panel. The hard-3 macro panel (second table) averages the three per-benchmark rates. Up to rounding, it is the table form of Figure 3. At $k \leq 4$ the methods interleave within 1.5 points of each other. The soft gate is even nominally ahead at k=2. From k=8 onward the ordering settles into the budget order. EOPD is first at every $k \geq 8$. OPD's mode-seeking endpoint is at or near the bottom, with a 5.8-point gap to EOPD at k=128. A later evaluation of the cover-87 curriculum checkpoint covers the same three benchmarks. It lands between the CTS-116 and EOPD spectra (hard-3 macro 37.96 at k=8 and 61.36 at k=128). This places the curriculum in the interior of the budget order. Its nominal position above cover-116 is the single-seed inversion that Section 5.4 reads as unresolved.

N Case Studies

The aggregate counts of Appendix I show that the forward-KL budget buys entry into the solved set. This appendix makes the same point tangible on five OlympiadBench problems. On each, the seed-42 OPD student never produces a correct generation in eight tries, while an FKL-carrying student

Table 18: Per-benchmark Pass@k values for the six 128-sample evaluation runs (seed 42, unbiased estimator), in order: AMC23, the hard-3 macro (mean of the three per-benchmark rates), AIME24, and AIME25.

AMC23

Method	k=1	k=2	k=4	k=8	k=16	k=32	k=64	k=128
OPD	38.1	48.9	59.8	70.3	79.6	86.9	92.7	97.5
EOPD	38.8	49.6	60.6	71.3	80.2	86.4	91.0	95.0
soft-entropy	39.9	51.1	61.9	71.5	78.5	82.7	85.9	90.0
JS-mass	38.2	48.5	58.6	68.0	76.5	84.0	90.5	97.5
TSA-full	40.3	50.8	61.2	70.9	79.0	85.4	90.8	95.0
CTS-116	39.1	50.2	61.3	71.3	79.3	84.9	89.1	92.5

Hard-3 macro

Method	k=1	k=2	k=4	k=8	k=16	k=32	k=64	k=128
OPD	17.7	23.8	30.2	36.2	41.8	47.4	52.9	58.1
EOPD	18.1	24.6	31.6	38.6	45.0	50.9	56.9	63.9
soft-entropy	18.3	24.7	31.2	37.2	42.4	46.8	50.8	55.6
JS-mass	18.0	24.0	30.1	36.0	41.7	47.6	53.4	60.3
TSA-full	18.7	24.9	31.3	37.6	43.6	50.0	56.6	62.8
CTS-116	18.5	24.7	31.2	37.3	43.3	49.4	55.5	60.8

AIME24

Method	k=1	k=2	k=4	k=8	k=16	k=32	k=64	k=128
OPD	9.1	13.1	17.3	20.9	24.3	28.7	34.1	40.0
EOPD	9.6	13.9	18.3	22.4	27.0	33.1	40.7	50.0
soft-entropy	9.1	13.3	17.8	21.7	25.5	29.4	32.9	36.7
JS-mass	10.4	14.8	18.7	21.9	25.6	29.9	34.9	43.3
TSA-full	9.3	13.4	17.5	21.4	26.0	32.3	39.7	46.7
CTS-116	10.5	14.8	18.9	22.8	27.6	33.8	40.7	46.7

AIME25

Method	k=1	k=2	k=4	k=8	k=16	k=32	k=64	k=128
OPD	6.0	9.4	13.4	17.4	21.6	26.6	31.9	36.7
EOPD	6.1	10.4	15.9	21.9	27.8	33.3	39.0	46.7
soft-entropy	5.9	9.6	13.9	18.4	23.4	28.5	33.7	40.0
JS-mass	5.4	8.8	13.2	18.0	23.2	28.9	34.8	40.0
TSA-full	6.5	10.5	15.2	20.4	25.9	32.3	39.4	46.7
CTS-116	5.8	9.2	13.3	17.9	23.1	29.5	36.6	43.3

does. For each problem we report the correct count of each student and a short excerpt of the stored generation (verbatim except for truncation), followed by a brief observation. Excerpts are shown in *italics* with our emphasis removed, and no text was edited.

Case 1: a recoverable logarithm slip. *Problem (OlympiadBench, T=650):* If $\log T = 2 - \log 2 + \log k$, compute k. *Answer:* 13.

Student	Correct / 8	Outcome in the stored generations
OPD (repro, s42)	0	All 8 generations mishandle the log identity
EOPD (repro, s42)	2	Two generations apply $\log a + \log b = \log(ab)$ correctly
TSA-soft-entropy (s42)	3	Three generations reach $k = 13$

The OPD student's attempt collapses exactly at the branch point: *"So, we can rewrite the right-hand side* $\dots \log T = 2 - \log 2 + \log k = \log 2 - \log 2 + \log k = \log k$. *Wait, that's not correct. Let's go step by step."* The soft-gate student instead commits to the productive branch from the start. Its generation opens *"We can rewrite* 2 *as* $\log 100$ *(since* $\log 100 = 2$*), so* $\log 650 = \log 100 + \log k - \log 2 \dots$ *"* and simplifies to $k = 13$. The problem is solvable by every method in principle, so the question is whether any sampled trajectory enters the right branch. This rare-success tail is what the

coverage budget buys. When the per-sample success probability p is this small, mode-seeking alone almost never samples the useful continuation.

Case 2: coverage that only the biggest spender reaches. *Problem:* Compute the sum of all positive two-digit factors of $2^{32} - 1$. *Answer:* 168.

Student	Correct / 8	Outcome in the stored generations
OPD (repro, s42)	0	Factorization stops before enumerating divisors
EOPD (repro, s42)	2	Two generations complete the divisor enumeration
TSA-soft-entropy (s42)	0	Same early stop as OPD on all 8 generations

Here only EOPD touches the problem at all. Its successful generation proceeds through the full chain: "$2^{32} - 1 = (2^{16} - 1)(2^{16} + 1)$. *We can further factor* $2^{16} - 1 = (2^8 - 1)(2^8 + 1) \dots$" and on to the divisor list whose two-digit members sum to 168. This is the honest shape of EOPD's coverage lead (Section 6). It is not that the soft gate cannot solve such problems. Rather, on this seed's sample of trajectories only the run with the largest forward-KL budget placed a continuation over the threshold. With n=8 samples, a $2/8$ success is exactly the kind of marginal entry that Pass@k counts and Avg@8 barely notices.

Case 3: when the soft gate's sample finds more branches. *Problem:* A geometric sequence has 20 terms; the sum of its first two terms is 40, of its first three 76, and of its first four 130. How many of the terms are integers? *Answer:* 5.

Student	Correct / 8	Outcome in the stored generations
OPD (repro, s42)	0	All 8 generations mishandle the system for a and r
EOPD (repro, s42)	2	Two generations reach a=16, r=3/2 and count 5 integer terms
TSA-soft-entropy (s42)	4	Four generations complete the same computation

Both FKL-carrying students set up the same system ("$a + ar = 40 \dots\ a + ar + ar^2 = 76$"). Each solves a=16, r=3/2, and counts the five integer terms $(16, 24, 36, 54, 81)$ before the powers of two in the denominator take over. The soft gate's sample contains twice as many correct completions on this seed. That is a reminder that single-seed ordering in this design space is noise. On the pooled evidence of Section 6, EOPD still holds the coverage lead. This cell is exactly the kind of nominal single-benchmark, single-seed advantage that the multi-seed analysis refuses to read as a method property.

Case 4: the same branch, found independently. *Problem (OlympiadBench, T=5):* Hannah flips two fair coins, Otto flips T fair coins; the probability that Hannah has more heads is $p = q/r$ in lowest terms; compute $q + r$. *Answer:* 17.

Student	Correct / 8	Outcome in the stored generations
OPD (repro, s42)	0	All 8 generations miscompute the binomial comparison
EOPD (repro, s42)	2	Two generations enumerate the joint cases correctly
TSA-soft-entropy (s42)	2	Two generations reach $p = 1/16$, hence $q + r = 17$

Both FKL-carrying students find the correct enumeration independently. EOPD's generation sets up "$X \sim \mathrm{Bin}(2, 0.5) \dots\ Y \sim \mathrm{Bin}(5, 0.5)$", and the soft gate's does the same with a different presentation. This is consistent with the diversity findings of Appendix J. The methods do not share solutions verbatim (cross-run recurrence is below 1%). They share *access to the same branch*.

Case 5: primality screening as a branchy search. *Problem:* For a positive integer a, define x_1=a and x_{n+1}=$2x_n + 1$, and let $y_n = 2^{x_n} - 1$. Determine the largest k such that $y_1, \dots, y_k$ can all be prime. *Answer:* 2.

Student	Correct / 8	Outcome in the stored generations
OPD (repro, s42)	0	All 8 generations lose track of the divisibility argument
EOPD (repro, s42)	3	Three generations show y_3 is always composite
TSA-soft-entropy (s42)	4	Four generations complete the same argument

The key step is the obstruction at y_3: the recurrence gives $y_3 = 2^{4a+3} - 1$. This is composite at the admissible values $a{=}2$ and $a{=}3$ (where y_1 and y_2 are both prime). No choice of a pushes k past 2. Both FKL-carrying students find this route. EOPD's generation opens with *"determine the largest possible value of k such that all the numbers $y_1, y_2, \ldots, y_k$ are prime for some positive integer a"*. It drives toward the y_3 obstruction. The OPD student's generations enumerate small cases without locating the obstruction. Of our five cases, this one shows the coverage effect at its clearest. The solution is a single specific branch. Miss it and every sample fails. Find it and the problem is short.

The five cases together illustrate the paper's central statement at the level of individual trajectories. What the coverage budget purchases is not a smarter path, but the presence of *any* correct path in a finite sample. Once one is present, its style is as varied as sampling itself.

O EQUIVALENCE TESTING OUTPUTS (TOST)

Section 5 and Appendix F claim equivalence only where two one-sided tests (TOST) certify it. A comparison is declared equivalent at margin D when its 90% confidence interval lies entirely inside $[-D, +D]$. This appendix tabulates the complete outputs behind every equivalence statement in the paper, computed on the same problem-level stratified bootstrap distributions as Table 9. For each comparison and metric we report the point estimate Δ (candidate minus baseline), the 90% CI, and a margin at which the CI certifies equivalence. Comparisons certifying at no margin up to ± 3 are reported as "none within ± 3". The $n{=}16$ row reports its wider 95% CI. Seed-mean rows pool the three candidate seeds against EOPD's four.

Comparison	Metric	Δ	90% CI	Equivalent at
Soft gate vs. EOPD (seed 42)	Avg@8	$+1.31$	$[+0.1, +2.5]$	± 3
Soft gate vs. EOPD (seed 41)	Avg@8	-0.44	$[-1.7, +0.7]$	± 2
Soft gate vs. EOPD (seed 43)	Avg@8	-0.96	$[-2.1, +0.1]$	± 3
Soft gate vs. EOPD (seed-mean)	Avg@8	-0.16	$[-2.0, +1.6]$	± 2.1
CTS cover-87 vs. EOPD (seed 42)	Avg@8	$+1.33$	$[+0.1, +2.6]$	± 3
CTS cover-116 vs. EOPD (seed 42)	Avg@8	-0.08	$[-1.1, +0.9]$	± 2
CTS cover-116 vs. EOPD (seed 42)	Pass@5	$+0.02$	$[-2.7, +2.7]$	± 3
CTS cover-87 vs. EOPD (seed-mean)	Avg@8	-0.47	$[-1.8, +0.9]$	± 2
CTS cover-87 vs. EOPD (seed-mean)	Pass@8	-0.47	$[-1.6, +0.7]$	± 2
Soft gate vs. EOPD (any paired seed)	Pass@5 / Pass@8	none within ± 3 (min. margins 3.2–4.9)		
CTS cover-87/116 vs. EOPD (seed 42)	Pass@8	none within ± 3 (min. margins 4.6–4.9)		
Soft gate vs. EOPD (seed 42, $n{=}16$ generations)	Pass@8	equivalent at ± 2.2 (95% CI inside)		

Reading the outputs. Three facts organize the table. First, all rows with a certified margin are Avg@8 or curriculum Pass@k rows (the seed-42 Pass@5 of cover-116 and the seed-mean Pass@8 of cover-87). The abstract's equivalence claim rests on the seed-mean row (-0.16, 90% CI $[-2.0, +1.6]$, inside the pre-registered ± 3 margin on every reading). Second, the single-seed coverage side stays unresolved. No single-seed Pass@8 comparison at $n{=}8$ is certifiably equivalent within ± 3 (minimal demonstrable margins 4.1–4.9 points). We report each such comparison as unresolved rather than established. The one budget where it does resolve ($n{=}16$ generations, ± 2.2) is tabulated for contrast. Third, the two curriculum configurations are the dose–response in equivalence form. Cover-116 matches EOPD on Avg@8 within ± 2 on the shared seed. Cover-87's three-seed means certify parity with EOPD at ± 2 on both Avg@8 and Pass@8. This is the only coverage-side equivalence the study certifies, earned by the two added seeds.