STYLEOPD: SELECTING THE GENERATION PROMPT FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) supervises a student on its own rollouts. Recent work improves that supervision by changing what the teacher is conditioned on or which tokens receive loss, while the prompt under which the student generates its rollouts is left at the task default, even though those rollouts determine what supervision exists at all. We treat the generation prompt as a selection variable and choose it by measurement. Reusing the criterion behind teachability-based token selection — corrective disagreement the student can still act on — we turn a per-token gate into a per-prompt score, computed once on the untrained student with no labels, no training and no change to the rest of the OPD recipe. In a controlled study distilling Qwen3.5-27B into Qwen3.5-9B, the choice of prompt moves accuracy by more than the prompt-free OPD gain over the initial student, every added prompt improves on vanilla OPD averaged over training, and the selected prompt beats vanilla OPD on every seed. Disagreement is what decides: every disagreement-based signal we measured, including our score, selects the best prompt, while every rule built on agreement or accuracy selects a worse one, and the best training prompt is the one under which the untrained student is least accurate. The selected prompt outperforms a sequence-level distillation baseline that needs extra teacher data, and it composes with token selection: the prompt helps on every seed with or without token selection, and more with it over training.

1 INTRODUCTION

On-policy distillation (OPD) is an increasingly common way to transfer reasoning ability from a strong teacher into a smaller student. Rather than training the student on teacher-written text, the student generates its own trajectories and the teacher scores them token by token, so the supervision is dense and is delivered exactly at the states the student visits. This removes the exposure mismatch of sequence-level distillation (Kim & Rush, 2016) and is markedly more sample efficient than off-policy imitation in recent reasoning-distillation studies (Li et al., 2026).

Recent work improves OPD by reshaping that supervision, and there are three variables to set. The first is what the teacher is conditioned on, such as a verified answer in on-policy self-distillation (Zhao et al., 2026). The second is which tokens receive loss, by estimated importance (Xu et al., 2026) or by *teachability*, disagreement that still lies inside the student's plausible continuations (Wang et al., 2026b). The third is the prompt the student samples under, almost always left at the task default.

This paper is about the third variable. Token selection reweights a sample already drawn and cannot create a state the student never visited: the generation prompt fixes the supervision available, and the scoring side only decides how to spend it. That this matters is known: aligning the student's format with the teacher's template improves OPD (Li et al., 2026), and output-style bias is treated as a first-order problem (Wu et al., 2026).

What is missing is a way to act on this. Template alignment says the generation format matters, but it identifies a format rather than a reasoning instruction: our student and teacher come from the same family and are already aligned in that sense, and six candidates (five reasoning instructions and no added prompt) still spread 1.33 points of final accuracy between them — more than the prompt-free OPD gain over the untrained student. The obvious heuristic, training under the prompt the model handles best, points the wrong way. The question is therefore concrete: *given a pool*

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.

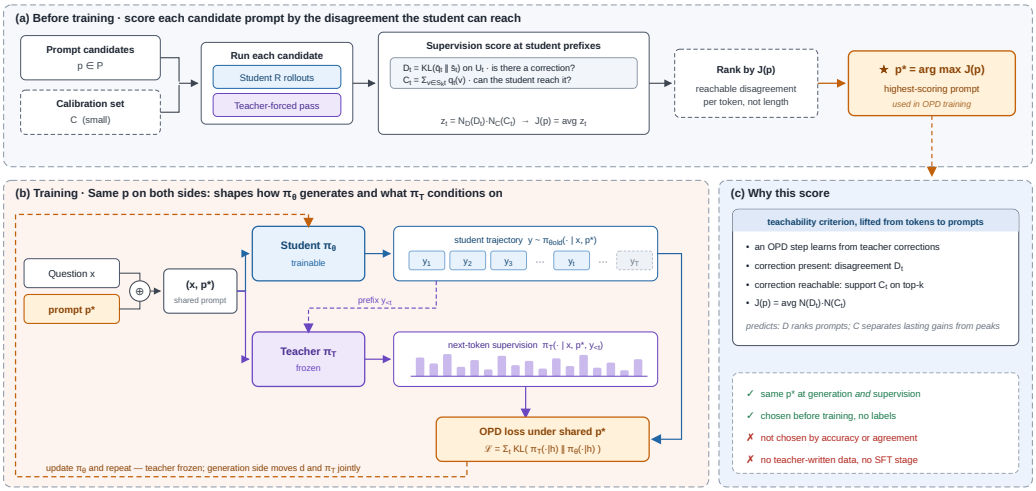


Figure 1: StyleOPD. (a) Before training, each candidate prompt is run on a small calibration set: the student samples rollouts, the teacher scores them in one teacher-forced pass, and $J(p)$ averages over student prefixes whether a correction is present (disagreement D_t) times whether the student can reach it (teacher mass C_t on the student's top-k). The highest-scoring prompt is kept. (b) Ordinary OPD runs unchanged with that prompt shared by student and teacher. (c) Together the two factors lift the teachability condition from tokens to prompts, and each makes a prediction tested in §4.3.

of task-preserving reasoning instructions, can a measurable quantity tell us which one to generate under?

Our answer reuses the criterion that already governs token selection. Teachability-based token selection (TA-OPD) asks, for each position, whether the teacher's disagreement with the student is corrective while remaining inside the set of tokens the student is willing to produce; we ask the same of a prompt. The score is computed once on the untrained student from one rollout pass and one teacher-forced pass — no labels, no gradient step, $1{,}024$ responses per candidate, under 6% of the rollouts of one training run, so that scoring our whole pool costs less than training one extra arm — and ordinary OPD then runs unchanged with the selected prompt shared by student and teacher (Figure 1). We study one teacher–student pair in depth: every arm is trained to completion and evaluated at every checkpoint, and the central comparisons are repeated over three seeds.

Findings. **(i) The generation prompt is a first-order variable.** The selected prompt beats vanilla OPD on every seed, by 1.92 points on average, and every one of the nine prompts we added trains better than no prompt over the course of training (§4.2). **(ii) Disagreement decides, accuracy misleads.** Every disagreement statistic we measured finds the winning prompt, while agreement and accuracy heuristics miss it; the winner is the prompt under which the untrained student is *least* accurate. **(iii) Reachability helps rank.** With the winner removed, the support-weighted score still picks the best remaining prompt whereas bare disagreement picks one that peaks early, and no signal we measured ranks any of our pools better (§4.3). **(iv) Composable and cheap.** The selected prompt helps with or without token selection on every seed, gaining more with it over training, and the two together reach 72.09 average accuracy over five benchmarks (AVG5) at the best checkpoint (§4.4); our sequence-level distillation baseline spends $89{,}836$ teacher responses and a supervised fine-tuning stage and still trails (§5.1).

Contributions. We frame teacher conditioning, token weighting and the generation prompt as three interventions on one supervision signal (§3); derive a label-free, training-free prompt score from the teachability criterion with predictions stated before training (§3.2–§3.3); and test each on fully trained runs (§4.3–§4.4).

2 Related work

From sequence-level to on-policy distillation. Knowledge distillation matches a student to a teacher's output distribution (Hinton et al., 2015); its sequence-level form trains on teacher-generated outputs (Kim & Rush, 2016) and is how reasoning is commonly distilled into small models (Ho et al., 2023; Hsieh et al., 2023; Guo et al., 2025), supervising the student only at states the teacher would visit. In the imitation-learning view, training on the learner's own states removes this mismatch (Ross et al., 2011; Lin et al., 2020), and on-policy distillation samples from the student and matches the teacher there (Agarwal et al., 2024; Gu et al., 2024), with further work on the divergence and the sampling schedule (Wen et al., 2023; Ko et al., 2024); it is now part of production post-training (Yang et al., 2025), and Li et al. (2026) give a systematic account for large reasoning models. We adopt this standard setting — student rollouts, teacher-forced scoring, no task reward, no auxiliary supervised fine-tuning — and use a sequence-level adaptation as a baseline.

Selective OPD: choosing tokens. Since positions differ in how much they teach, several methods weight them within a trajectory. TIP estimates token importance and concentrates the objective on high-importance positions (Xu et al., 2026), as token selection does for pretraining (Lin et al., 2024). TA-OPD formalizes *teachability*: disagreement is useful when the teacher's preferred continuation is still among the tokens the student itself puts mass on, and disagreement outside that set is hard to absorb (Wang et al., 2026b), a token-level form of the capacity gap that makes larger teachers less useful to small students (Cho & Hariharan, 2019; Mirzadeh et al., 2020). Our score descends directly from this criterion; what we contribute is its use as a generation-side decision rule, and in §4.4 we adopt an existing token selector unchanged.

Privileged self-distillation: choosing the teacher's context. A parallel line changes what the teacher conditions on. On-policy self-distillation (OPSD) instantiates teacher and student from one model and gives the teacher privileged information — a verified answer or a reference trace — while the student sees only the question (Zhao et al., 2026). Its documented failure mode is that an oracle-conditioned teacher moves outside the student's support, so the student imitates a hindsight shortcut (Zhang, 2026; Shen et al., 2026); keeping supervision close to what the student can produce is the criterion TA-OPD applies to positions and we apply to prompts (§3.2).

Thinking patterns and teacher-aligned prompting. Li et al. (2026) show that compatibility between student and teacher thinking patterns governs how much ability transfers, and report gains from prompting the student in a format aligned with the teacher's post-training template. This is the closest prior work: it identifies a format, but not which reasoning instruction to use once the format is fixed. Wu et al. (2026) approach the same problem from the teacher side by correcting output-style bias, and their method serves as a comparator here. Cross-tokenizer distillation (Wang et al., 2026a) is orthogonal to prompt choice. The candidate pool itself draws on the zero-shot reasoning-prompt literature: step-numbered chain of thought (Kojima et al., 2022), plan-then-solve decomposition (Wang et al., 2023), and tabular chain of thought (Jin & Lu, 2023).

Prompt optimization and disagreement-based selection. Chain-of-thought prompting (Wei et al., 2022) and its zero-shot variants choose instructions for the model's own inference, and prompt optimizers search for them by downstream accuracy (Zhou et al., 2023; Yang et al., 2024). We choose an instruction for what it does to training, and find the accuracy heuristic points the wrong way. Selecting inputs by model disagreement is the premise of query-by-committee active learning (Seung et al., 1992; Settles, 2009); here nothing is labeled, and the prompt changes the states at which the teacher is queried rather than which examples are. We are not aware of a prompt selector built for on-policy distillation, so §4.3 constructs the comparison set explicitly.

3 Method

3.1 Three interventions on one supervision signal

Let π_θ be the student and π_T the frozen teacher. A prompt p is a task-preserving instruction appended to the input x: it constrains how the model organizes its reasoning and leaves the question and the answer contract unchanged. In on-policy distillation the student samples $y \sim \pi_{\theta_{\text{old}}}(\cdot \mid x, p)$ and

is trained to match the teacher at every visited prefix $h = (x, p, y_{<t})$, giving the per-position loss $\ell_{\mathrm{OPD}}(\theta; h)$, in our implementation the sampled-token estimator of $\mathrm{KL}(\pi_\theta(\cdot \mid h) \,\|\, \pi_T(\cdot \mid h))$.

Selective variants attach a non-negative weight $w(h)$ to each visited position and optimize the normalized objective

$$\mathcal{L}(\theta; p, w) = \frac{\mathbb{E}_{h \sim d_{\theta_{\mathrm{old}}, p}}[w(h)\, \ell_{\mathrm{OPD}}(\theta; h)]}{\mathbb{E}_{h \sim d_{\theta_{\mathrm{old}}, p}}[w(h)]}, \qquad \mathbb{E}[w] > 0, \tag{1}$$

where $d_{\theta_{\mathrm{old}}, p}$ is the occupancy measure over prefixes induced by sampling from the current student under prompt p. Three choices are open. The teacher's conditioning enters through π_T, which privileged self-distillation replaces with a view conditioned on information the student does not have, such as a verified answer (Zhao et al., 2026). The scoring side chooses w, redistributing loss within whatever trajectories were drawn (Xu et al., 2026; Wang et al., 2026b). The generation side chooses p, which moves the measure $d_{\theta_{\mathrm{old}}, p}$ itself and, because we prompt student and teacher identically, moves $\pi_T(\cdot \mid h)$ with it. Sharing the prompt is deliberate: the teacher then corrects how the student carries out the instruction, not whether it follows one the teacher never saw.

We study the third. For a fixed p, every objective reachable by choosing w is an average of ℓ_{OPD} over prefixes the student already visits under p, since positions outside the support of $d_{\theta_{\mathrm{old}}, p}$ receive no loss for any admissible w: the generation side determines the supervision the scoring side is able to spend. Two consequences matter. The value of p can be measured before training, because $d_{\theta_{\mathrm{old}}, p}$ and π_T are both available at θ_{old}; and because student and teacher share the prompt, changing p moves the visited states and the teacher conditional together, so we identify that joint intervention rather than a generation-only pathway.

3.2 Which prompt supplies more usable signal?

Two degenerate regimes bound the problem. If student and teacher already agree at the visited prefixes, matching the teacher changes nothing; if they disagree in a way that places the teacher's preferred continuation far outside anything the student currently entertains, the target is hard to absorb. TA-OPD makes this trade-off at the token level (Wang et al., 2026b); we make its prompt-level counterpart:

> *Prefer the generation prompt whose trajectories carry more corrective disagreement while the teacher's preferred continuations remain among the tokens the student itself puts mass on.*

The criterion has two parts, and they can fail separately. A correction must be present: where the teacher agrees with the student there is nothing to learn. And it must be absorbable: the KL gradient reaches every logit, but a small update mainly reshuffles probability among continuations the student already entertains, so a correction toward tokens far outside that set takes hold slowly. We call such a correction reachable and use the student's top-k as a proxy for the set; it is a heuristic, not a hard constraint. Stated before any training, this makes three predictions, which §4 tests in turn. **(P1)** A prompt's disagreement at the student's own prefixes predicts how well it trains; statistics of agreement, or of the models' accuracy under the prompt, do not. **(P2)** Between prompts with similar disagreement, the one whose disagreement falls more inside the student's support trains better throughout; weighting disagreement by support should therefore rank a pool better than disagreement alone, most visibly when no candidate stands out. **(P3)** A token selector can keep only what the prompt produced, so under a higher-scoring prompt token selection should cost less and the prompt should gain more.

3.3 The prompt score

Fix a calibration set $\mathcal{C}$ of questions from the training domain and a candidate pool $\mathcal{P}$. For each $p \in \mathcal{P}$ we draw R rollouts per question from the *initial* student, then run one teacher-forced pass so that both models are evaluated at the identical prefixes the student produced; sizes are in §4.1. At position t let s_t and q_t be the full-vocabulary, temperature-one student and teacher distributions, let S_t^k and T_t^k be their top-k sets, and let $U_t = S_t^k \cup T_t^k$. We measure

$$D_t = \mathrm{KL}\Big(\bar{q}_t^{\,U_t} \,\Big\|\, \bar{s}_t^{\,U_t}\Big), \qquad C_t = \sum_{v \in S_t^k} q_t(v), \tag{2}$$

where $\bar{q}_t^{U_t}$ and $\bar{s}_t^{U_t}$ are renormalized on U_t. Both terms follow the teachability criterion of Wang et al. (2026b) without reproducing its deployed functional, which is why we call the score TA-inspired; the two are written out side by side in Appendix H. The disagreement term D_t is restricted to tokens either model considers plausible, so it is not dominated by tail mass; the support term C_t is the teacher probability mass falling on tokens the student is already willing to produce, a probability mass rather than a set-overlap ratio.

The two terms have different units and dynamic ranges, so we place them on a common scale with calibration quantiles: with $Q_{.05}(a)$ and $Q_{.95}(a)$ denoting the 5th and 95th percentiles of quantity a over the calibration bank, $N_a(u) = \mathrm{clip}_{[0,1]}\left((u - Q_{.05}(a))/(Q_{.95}(a) - Q_{.05}(a))\right)$ and $z_t = N_D(D_t) \cdot N_C(C_t)$. Quantiles are estimated from a fixed number of deterministic, stratified positions per response, pooled over all candidates, so that every response contributes equal calibration mass and no candidate is normalized against itself. Pooling makes the scale depend on the pool, so adding a candidate can shift it; ranking by the unnormalized product selects the same prompt here (Appendix B). The score itself uses all positions and aggregates with uniform weights at every level,

$$J(p) = \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \frac{1}{R} \sum_{r=1}^{R} \frac{1}{|y_{x,p,r}|} \sum_{t=1}^{|y_{x,p,r}|} z_{x,p,r,t}, \qquad p^\star = \arg\max_{p \in \mathcal{P}} J(p), \tag{3}$$

so J is a density of usable signal rather than a per-sequence total and a candidate cannot win by being verbose. Aggregation covers thinking and post-thinking text, excludes prompt and padding positions, and retains incorrect responses, since those trajectories are part of what OPD trains on.

The two factors are what the predictions test, so they are also what we ablate. Removing C leaves bare disagreement — either KL direction, either model's entropy, or combinations of them — which P1 says should still pick well; removing D leaves support or top-k overlap, which measure agreement and P1 says should not. P2 is the contribution of C on top of D. Disagreement varies across prompts far more than support does, so J is led by D and C acts as a tie-breaker among prompts of similar disagreement, as the teachability gate does among tokens. The two predictions therefore separate what we claim for the criterion from what we claim for the score: any disagreement statistic should find a prompt that stands out, and only the support-weighted score should keep choosing well when none does.

3.4 Using the selected prompt, and how the score is judged

Computing J costs one rollout pass over $|\mathcal{C}| \cdot R$ responses per candidate and one teacher-forced pass, with no labels, no reward model, no gradient step and no evaluation: $1{,}024$ responses per candidate at the sizes of §4.1, and $6{,}144$ for the whole six-candidate pool, about a third of one run's $17{,}920$ rollouts. Training is then ordinary OPD with $p^\star$ appended to every training question and shared by student and teacher, preserving the original answer contract; same data, budget and optimizer. Evaluation always uses the plain task prompt, so a candidate cannot be credited for making the test easier.

Each prediction is checked as a decision: a high-score group should outperform a low-score group, and the chosen prompt should have low regret. The pool, score direction and group boundaries are fixed for all audited signals, and every candidate is reported.

4 Experiments

4.1 Setup

Models, data and training. Student and teacher are Qwen3.5-9B and Qwen3.5-27B, both in thinking mode; OPD trains on the $17{,}917$ mathematical-reasoning questions of DAPO-Math-17K (Yu et al., 2025). Every arm runs the identical recipe: prompt batch 128, one rollout per question, $32{,}768$-token response budget, learning rate 5×10^{-6}, 140 updates, checkpoints every ten updates, 32 GPUs, no task reward and no auxiliary supervised fine-tuning. Only the prompt — and, in §4.4, the token selector — differs between arms; results use seed 23 unless stated. The score is calibrated on 256 training-domain questions with four initial-student rollouts each, $k = 16$, and quantiles from 128 stratified positions per response.

Training prompt	J	Best checkpoint AIME24	AIME25	AIME26	MATH500	HMMT25	AVG5	Δ	Curve average AVG5	Δ
Initial student (9B)	–	70.83	57.50	66.67	96.16	52.50	68.732	–	–	–
Teacher (27B)	–	81.87	74.58	81.25	97.82	70.83	81.270	–	–	–
PSV	0.1211	72.29	61.67	68.54	96.59	55.42	70.902	1.28	69.303	1.68
Canonical Freeform	0.1178	71.25	61.46	66.67	96.69	55.00	70.214	0.59	68.558	0.94
Tabular-CoT	0.1138	70.42	64.58	68.75	96.81	52.71	70.654	1.03	68.046	0.43
Plan	0.1106	72.29	60.42	69.79	96.34	53.54	70.476	0.86	68.535	0.92
Step-by-Step	0.1087	70.83	61.04	67.92	96.39	51.67	69.570	-0.05	67.789	0.17
Vanilla OPD	0.1076	71.67	60.42	67.71	95.80	52.50	69.620	0.00	67.620	0.00
Formal PSV	0.1146	72.08	59.17	71.46	96.66	52.08	70.290	0.67	68.560	0.94
Post PSV	0.1126	71.88	62.08	65.00	96.16	52.92	69.608	-0.01	68.061	0.44
Labels Only	0.1093	71.25	59.79	67.29	96.59	52.50	69.484	-0.14	68.368	0.75
Semantic Freeform	0.1091	73.33	60.00	69.58	96.11	51.25	70.054	0.43	67.999	0.38
SeqKD SFT	–	68.33	61.46	66.04	96.31	51.88	68.804	-0.82	–	–
SeqKD + vanilla OPD	–	70.42	61.04	65.83	96.09	49.17	68.510	-1.11	66.852	-0.77
SeqKD + PSV OPD	–	70.42	61.67	69.17	96.45	55.21	70.584	0.96	68.607	0.99

Table 1: Ten candidate-prompt OPD runs, ordered by J within the first two blocks, with reference rows and baselines. First block: the six ranked candidates fixed in advance. Second block: variants of the selected prompt, built afterwards (§5.2). Third block: baselines; SeqKD is sequence-level distillation, SFT supervised fine-tuning. PSV: plan–solve–verify. Δ is against vanilla OPD under the same rule; curve averages span updates 10–140. J is measured on the untrained student.

Candidates and terminology. The ranked pool is six candidates fixed in advance: five reasoning styles drawn from the prompting literature — step-by-step reasoning (Kojima et al., 2022), plan-then-solve (Wang et al., 2023), tabular chain of thought (Jin & Lu, 2023), a free-form reasoning instruction and a plan–solve–verify structure — together with the no-added-prompt arm, which is both the training baseline and a pool member, since a selection rule must be allowed to decide that adding nothing is best. The winner among them, PSV, asks the model to make a plan, execute it and verify the result, in that order, inside its thinking trace. Four variants of PSV, built afterwards and trained under the same recipe, are ablations (§5.2) outside the ranked pool. All ten strings appear in Appendix K. Throughout, *vanilla OPD* means training with no added prompt and every evaluation uses the *plain* task prompt.

Evaluation and reporting. All numbers use the plain task prompt, a 32,768-token output cap and a 65,536-token context. AVG5 is the equally weighted mean of AIME24, AIME25, AIME26, MATH500 (Hendrycks et al., 2021; Lightman et al., 2024) and HMMT25, each scored as mean correctness over 16 sampled responses per question, which estimates pass@1; an AVG5 exists only where all five components are verified, and incomplete cells are never imputed. Each run is reported twice. The *best checkpoint* gives deployment numbers comparable with prior work. The *curve average*, the mean AVG5 over the checkpoints at updates 10–140, measures what the prompt does over the whole run; since a prompt is chosen before training and acts on every update, it is what we test the prompt score against. Seed-to-seed variation is in Appendix G.

4.2 How much is the generation prompt worth?

Table 1 gives the result. Among the six ranked candidates the selected prompt reaches 70.902 AVG5 at its best checkpoint against 69.620 for vanilla OPD; the initial student is at 68.732 and the teacher remains 10.368 points above the best full-token OPD student. The six choices span 1.332 points, more than the 0.888 points vanilla OPD gains over the initial student. Averaged over training the intervention is positive: every one of the nine added prompts trains better than vanilla OPD, by 0.17 to 1.68 points, although three of them do not beat it at the best checkpoint. The selected prompt beats vanilla OPD on all three seeds, by 1.28, 2.35 and 2.13 points (Table 8); the distillation baselines of the third block are discussed in §5.1.

4.3 Testing the criterion

P1: disagreement predicts, agreement and accuracy do not. Table 2 lists each rule with what it computes. Every rule that measures disagreement selects PSV, the best of the six, and orders

		Six ranked candidates					Without PSV	
Selection rule	Per-token quantity	Top two	ρ	Gap	Pick's student acc.	Regret	Pick	Regret
J (ours)	$N_D(D_t)\, N_C(C_t)$	PSV, CanonFF	0.94	1.23	49.99	0.00	CanonFF	0.00
Union KL	D_t, Eq. 2	PSV, Tabular	0.83	0.97	49.99	0.00	Tabular	0.51
Forward KL	$\mathrm{KL}(q_t \,\|\, s_t)$	PSV, Tabular	0.83	0.97	49.99	0.00	Tabular	0.51
Reverse KL	$\mathrm{KL}(s_t \,\|\, q_t)$	PSV, Tabular	0.83	0.97	49.99	0.00	Tabular	0.51
Student entropy	$H(s_t)$	PSV, Tabular	0.83	0.97	49.99	0.00	Tabular	0.51
Teacher entropy	$H(q_t)$	PSV, Tabular	0.83	0.97	49.99	0.00	Tabular	0.51
Support	C_t, Eq. 2	Vanilla, Step	-0.83	-0.97	68.73	1.68	Vanilla	0.94
Normalized support	$N_C(C_t)$	Vanilla, Step	-0.83	-0.97	68.73	1.68	Vanilla	0.94
Top-k overlap	$\lvert S_t^k \cap T_t^k \rvert / k$	CanonFF, Vanilla	0.14	-0.20	69.73	0.74	CanonFF	0.00
Label-using diagnostics, not deployable								
Student accuracy	AVG5 of initial student under p	CanonFF, Vanilla	-0.37	-0.83	69.73	0.74	CanonFF	0.00
Teacher accuracy	AVG5 of teacher under p	CanonFF, Plan	-0.09	-0.13	69.73	0.74	CanonFF	0.00
No added prompt	–	Vanilla	–	–	68.73	1.68	Vanilla	0.94

Table 2: Every selection rule on the six ranked candidates, scored by the curve average of the completed runs. Quantities are averaged per token as in Eq. 3; s_t, q_t are the student and teacher distributions, S_t^k, T_t^k their top-k sets, and entropy is an uncertainty proxy for disagreement. The first of the top two is the pick; "Gap" is the top-two minus bottom-two mean, ρ the rank correlation, "Pick's student acc." the untrained student's AVG5 under the pick, regret the distance to the best prompt; the last two columns repeat the decision without PSV. CanonFF: Canonical Freeform. More signals: Table 6.

all four high×low pairs. These statistics are correlated, so they are not independent confirmations; what they show is that the decision does not hinge on our functional, and the strongest simple alternative, taking the prompt with the largest KL, makes the same choice. Every rule that measures agreement or accuracy selects a worse prompt, and the selected prompt is the one under which the untrained student is least accurate (49.99 AVG5 against 68.73). The advantage holds along the whole rollout, not only at the opening tokens: PSV's local score exceeds the default's in all 20 position bins (Figure 2a; raw support in Figure 3).

P2: reachability ranks the rest. A selector must also rank, not only spot an outlier, so we remove the winner and ask again. The five remaining candidates lie within 0.94 points. On this pool the score, and every signal weighting disagreement by normalized support, picks the best of them, Canonical Freeform; 18 of the 19 bare disagreement signals, largest KL included, pick Tabular-CoT, 0.51 points lower over training (Table 2, last columns); top-k overlap and the label-using accuracy heuristics also find Canonical Freeform here, having missed PSV on the full pool. Over the whole ranking, no signal we measured orders the six candidates, all ten arms or the nine without PSV better than the score ($\rho = 0.94, 0.88, 0.83$); bare disagreement at best ties it on the six candidates and trails on both larger pools (Appendix B). The decisive margin is single-seed and within seed-to-seed variation, so this is evidence consistent with P2, not a proof of it. Tabular-CoT shows what bare disagreement misses: its best checkpoint, 70.65, is an isolated peak at update 30 (68.29 and 69.17 either side), and it averages 68.05 over training against 68.56. Raw support is saturated on this model pair (99.57% to 99.73% for every candidate), but after normalization Tabular-CoT has the lower support ($N_C = 0.906$ against 0.916): more of its disagreement lies outside what the student entertains.

4.4 Prompt selection composes with token selection

P3 concerns the token side, so we train the two together, adopting the selector unchanged. The two compose: the selected prompt helps under both token regimes on every seed, relative to the default prompt in the same regime, by 1.92 and 2.37 points on average at the best checkpoint and by 1.91 and 2.84 over training, and the combination reaches 72.087 AVG5 at its best checkpoint (Table 3). Under token selection the selected prompt is ahead of the default at most checkpoints (Figure 2c). Over training the prompt gains more under token selection on every seed, the direction P3 predicts: token selection lowers the default prompt's curve on every seed (-0.98, -0.98, -0.74) and leaves the selected prompt's essentially unchanged ($+0.52$, -0.54, $+0.12$). At the best checkpoint the effect of token selection varies across seeds by more than the interaction, so we read P3 as consistent with the data rather than established.

Prompt	Tokens	Best checkpoint AIME24	AIME25	AIME26	MATH500	HMMT25	AVG5	Δ	Curve AVG5	Δ
Mean of three seeds										
Vanilla	All	71.88	61.88	67.29	95.91	50.63	69.517	0.00	67.787	0.00
	TA 5%	70.97	60.42	67.08	96.16	53.96	69.719	0.20	66.889	-0.90
PSV	All	72.98	63.68	69.30	96.57	54.65	71.439	1.92	69.694	1.91
	TA 5%	72.50	64.93	70.62	96.69	55.69	72.087	2.57	69.726	1.94
Seed 23										
Tabular-CoT	All	70.42	64.58	68.75	96.81	52.71	70.654	1.03	68.046	0.43
	TA 5%	71.25	61.46	67.08	96.04	53.54	69.874	0.25	66.468	-1.15
Step-by-Step	All	70.83	61.04	67.92	96.39	51.67	69.570	-0.05	67.789	0.17
	TA 5%	70.42	61.46	67.71	96.44	53.12	69.830	0.21	67.432	-0.19

Table 3: Prompt selection crossed with TA token selection (Wang et al., 2026b) (TA 5%: loss on the 5% most teachable tokens), at each run's best checkpoint and averaged over training. Upper block: the default and selected prompts, averaged over three seeds. Lower block: two further prompts, seed 23 only. "All" is full-token OPD. Δ is against full-token vanilla OPD on the same seeds. Per-seed values: Table 8.

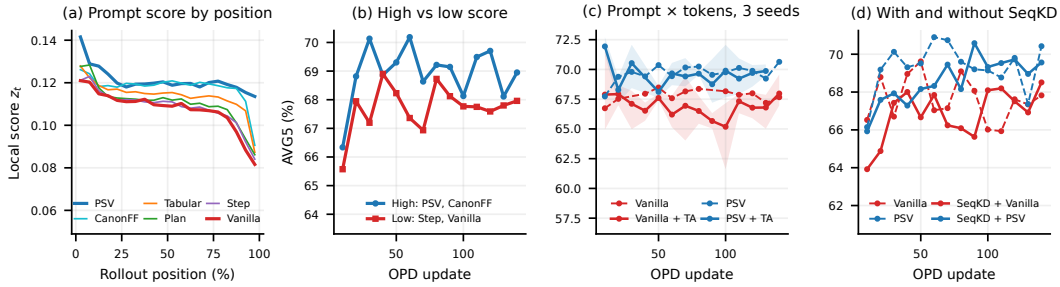


Figure 2: Prompt score and training curves. (a) Local score $z_t = N_D(D_t)\, N_C(C_t)$ of the six ranked candidates along the untrained student's calibration rollouts, in 20 relative-position bins; J is its mean. (b) The two highest- and two lowest-scoring of the six ranked candidates under full-token OPD, seed 23. (c) The four arms of Table 3's upper block, mean of three seeds at each update where all three seeds are evaluated; bands show the seed range of the full-token arms. (d) Vanilla and PSV OPD with and without SeqKD initialization (§5.1), seed 23.

5 Analysis

5.1 Three ways to close a style gap

A student and a teacher that reason in different styles can be reconciled by changing the student, by changing the teacher's signal, or by changing the context both of them see. Sequence-level distillation (SeqKD) does the first, Lightning OPD 2.0 (Wu et al., 2026) the second and prompt selection the third; Table 4 lists what each requires and Figure 2d shows the SeqKD and prompt-selection curves.

Fine-tuning the student on teacher-written responses leaves it at 68.80 AVG5, level with the initial 68.73: imitating the teacher's text does not by itself make the student better, echoing the limits of imitating a stronger model's text (Gudibande et al., 2024). OPD started from it opens lower (63.92 after ten updates, against 66.53 for vanilla OPD) and stays lower, averaging 66.85 over training against 67.62; the same holds with the selected prompt (68.61 against 69.30). The prompt gain survives the change of initialization, 1.75 points on the curve from the SeqKD start against 1.68 from the initial student: the prompt supplies something the teacher-written corpus does not, and the corpus itself costs training quality. Our Lightning OPD 2.0 reproduction never rises above the initial student in 84 updates (peak 68.73); we have not reconciled it with the published results and draw no conclusion from it.

Prompt selection needs no extra stage or data and combines with SeqKD initialization and token selection (§4.4).

Method	What it changes	Needs beyond OPD	Extra cost
SeqKD → OPD (Kim & Rush, 2016)	student initialization	teacher corpus, SFT stage	89,836 teacher responses at 64K
Lightning OPD 2.0 (Wu et al., 2026)	teacher signal	SFT reference, replay bank	style-bias fitting
TA-OPD (Wang et al., 2026b)	token weights	–	scoring at every update
StyleOPD (ours)	shared prompt	–	1,024 student responses and one teacher pass per candidate, once

Table 4: What each method requires beyond vanilla OPD.

5.2 What part of the instruction carries the effect?

At the best checkpoint, the lower block of Table 1 separates what the selected prompt asks for from where it applies.

Organizing the reasoning is what matters, not marking it. Labels Only requests the same three headings without saying what each section should do, and is the weakest of the ten arms: 1.42 points below the full instruction and below adding no prompt at all. Semantic Freeform asks for the same three activities without imposing headings, and lands 0.85 below the full instruction but 0.43 above the default. The effect tracks the behavioral content, not the markers.

The instruction has to reach the tokens that OPD trains on. The three variants requesting the same plan, solve and verify behaviour differ only in where it must hold. Inside the thinking trace is best; dropping the mandatory paragraph that pins the structure to the text inside `<think>` (Formal PSV) costs 0.61 points, and applying it only to the final visible answer (Post PSV) costs 1.29 and is no better than adding no prompt. OPD supervises the long reasoning trace, and an instruction shaping only the short final presentation leaves that supervision unchanged. This is a consistent ordering rather than a position-controlled experiment: the three strings are matched on content only.

Is PSV just a harder prompt? PSV leaves the untrained student far weaker (native AVG5 49.99 against 68.73, most of it on MATH500), so it might help only by making training rollouts harder or longer. Difficulty alone does not explain the ranking: after PSV, Semantic Freeform and Post PSV depress native accuracy most (66.76, 67.13), yet over training they trail Canonical Freeform (68.00, 68.06 against 68.56 curve AVG5), which raises native accuracy. Nor are PSV's training rollouts much longer ($22{,}997$ against $22{,}659$ mean tokens), and a hard-first ordering would be a curriculum (Bengio et al., 2009), which fixed-prompt arms do not implement. Training under PSV raises accuracy under PSV by 15.7 points, the largest matched gain in the pool, yet the model evaluated without it is better still, so the instruction is a training signal to drop at deployment (Appendix E).

5.3 Scope and limitations

We traded breadth for depth: one teacher–student pair (9B←27B, one family), competition mathematics with verifiable answers, thinking mode and a 32K budget. Transfer to other pairs and domains is untested, and the score is calibrated per pair. The pool is five common reasoning styles and no prompt; zero regret on it does not imply that the score would rank an arbitrary new instruction correctly. Only the default and selected prompts were run under three seeds, so the other arms, and the 0.51-point margin behind P2, carry single-seed uncertainty of about that size. Student and teacher share the prompt, so we do not isolate a generation-only effect; arms are matched on updates, prompt batch and rollout count but not on generated tokens; and the SeqKD baseline uses one recipe.

6 Conclusion

Work on OPD has tuned the teacher's conditioning and the token weights but left the generation prompt at its default. A score computed once on the untrained student, without labels or training, chooses a prompt that beats vanilla OPD on every seed, outperforms our SeqKD baseline, which needs teacher-written data, and composes with token selection. In the pair and pool we studied, what decides is how much the teacher has to correct on the student's own rollouts, not how well either model already answers, and weighting that correction by whether the student can absorb it helps rank the rest. Prompt choice belongs in the OPD recipe.

Reproducibility statement

The score is specified in §3.3 to the level a reimplementation needs: the calibration set size, rollout count, top-k, the two measured quantities, the quantile normalization and its estimation protocol, and the aggregation. Training and evaluation protocols, including the checkpoint reporting convention, are in §4.1. The exact text of all ten candidate prompts is reproduced in Appendix K, and the difference between the offline prompt score and the deployed token selector is written out in Appendix H so that neither is mistaken for the other. Appendix A gives the full selector audit and the main per-candidate signals, Appendix G gives the seed-to-seed variation of the four arms run on all three seeds, and Appendix I gives the per-seed and per-candidate curves. The supplementary material contains a standalone implementation of the score, reference implementations of the OPD loss and token selector, the training and evaluation configurations, and anonymized per-checkpoint results with a script that recomputes the main tables (Appendix L).

Ethics statement

This work distills mathematical reasoning between open-weight models on public competition-mathematics benchmarks; it involves no human subjects, no personal data and no newly collected dataset. The main foreseeable risk is indirect: a cheap selection rule that improves distillation efficiency also lowers the cost of copying capabilities out of a stronger model, including in cases where the model owner has not authorized it. The method requires full output distributions from the teacher, so it applies to models whose logits the practitioner can already access, which limits but does not eliminate that concern. Training and evaluation consumed substantial GPU compute; we report the per-arm budget in §4.1 so that the cost of reproduction is visible, and note that the selection procedure itself is a small fraction of one training run and is amortized across runs on a model pair.

AI use statement

We used generative AI tools only to polish the writing of this paper (grammar, wording and copy-editing). They were not used for research ideation, experiment design, generating or analyzing results, or producing any reported number. The authors reviewed all edited text and take full responsibility for the content of this paper.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, 2009.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

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

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary language models. In *International Conference on Learning Representations*, 2024.

Daya Guo et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645:633–638, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Ziqi Jin and Wei Lu. Tab-CoT: Zero-shot tabular chain of thought. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10259–10277, 2023. doi: 10.18653/v1/2023.findings-acl.651. URL https://aclanthology.org/2023.findings-acl.651/.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of EMNLP*, 2016. URL https://arxiv.org/abs/1606.07947.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *International Conference on Machine Learning*, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.11916.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026. URL https://arxiv.org/abs/2604.13016.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Not all tokens are what you need for pretraining. In *Advances in Neural Information Processing Systems*, 2024.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 1992.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, Gang Chen, and Jieping Ye. Purified OPSD: On-policy self-distillation without losing how to think. *arXiv preprint arXiv:2607.02234*, 2026. URL https://arxiv.org/abs/2607.02234.

Hao Wang, Kun Yuan, Wenlin Zhong, Minglei Zhang, Han Xiao, Ming Sun, and Honggang Qi. Cross-tokenizer on-policy distillation via byte-prefix marginalization. *arXiv preprint arXiv:2607.22334*, 2026a. URL https://arxiv.org/abs/2607.22334.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2609–2634, 2023. doi: 10.18653/v1/2023.acl-long.147. URL https://aclanthology.org/2023.acl-long.147/.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026b. URL https://arxiv.org/abs/2605.26844.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Yecheng Wu, Song Han, and Han Cai. Lightning OPD 2.0: Mitigating style bias in cross-teacher on-policy distillation for large reasoning models. *arXiv preprint arXiv:2607.28449*, 2026. URL https://arxiv.org/abs/2607.28449.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026. URL https://arxiv.org/abs/2604.14084.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024.

Qiying Yu et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Wenhao Zhang. Beyond absolute imitation: Anchored residual guidance for privileged on-policy distillation. *arXiv preprint arXiv:2606.10385*, 2026. URL https://arxiv.org/abs/2606.10385.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026. URL https://arxiv.org/abs/2601.18734.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *International Conference on Learning Representations*, 2023.

A FULL CANDIDATE AND SELECTOR AUDIT

Table 5 gives every candidate under the main audited signals, together with the downstream result. It is the source of the signal rankings in Table 2, and it shows the support saturation discussed in §F: the spread of C_{16} across the entire pool is about 0.16 percentage points, against a factor of 1.4 in the disagreement terms.

Candidate	J	Fwd. KL	Rev. KL	Student entropy	C_{16} (%)	AVG5
PSV	0.1211	0.0678	0.0732	0.4843	99.570	70.902
Canonical Freeform	0.1178	0.0560	0.0620	0.4325	99.693	70.214
Formal PSV	0.1146	0.0546	0.0606	0.4320	99.690	70.290
Tabular-CoT	0.1138	0.0565	0.0629	0.4493	99.645	70.654
Post PSV	0.1126	0.0526	0.0584	0.4205	99.706	69.608
Plan	0.1106	0.0521	0.0586	0.3977	99.716	70.476
Labels Only	0.1093	0.0514	0.0572	0.3996	99.715	69.484
Semantic Freeform	0.1091	0.0513	0.0576	0.3990	99.714	70.054
Step-by-Step	0.1087	0.0505	0.0563	0.3922	99.727	69.570
Vanilla OPD	0.1076	0.0499	0.0558	0.3876	99.731	69.620

Table 5: All ten candidates under the main audited signals, ordered by J. AVG5 is the best-checkpoint value from that candidate's own OPD run.

Signal	Family	High pair	Pick	ρ	Gap
$J = N_D(D) \cdot N_C(C)$ (ours)	dis. × norm. support	PSV, CanonFF	PSV	0.94	1.23
$N(D_{16}) \cdot N(C_{16})$	dis. × norm. support	PSV, CanonFF	PSV	0.94	1.23
$N(H) \cdot N(C_{16})$	dis. × norm. support	PSV, CanonFF	PSV	0.94	1.23
Soft-OR$(H, D) \cdot N(C)$	dis. × norm. support	PSV, CanonFF	PSV	0.94	1.23
Probability ℓ_2 distance	disagreement	PSV, CanonFF	PSV	0.94	1.23
Union KL $N_D(D)$	disagreement	PSV, Tabular	PSV	0.83	0.97
Top-16 KL D_{16}	disagreement	PSV, Tabular	PSV	0.83	0.97
$N(D_{16})$	disagreement	PSV, Tabular	PSV	0.83	0.97
Forward KL	disagreement	PSV, Tabular	PSV	0.83	0.97
Reverse KL	disagreement	PSV, Tabular	PSV	0.83	0.97
Student entropy	disagreement	PSV, Tabular	PSV	0.83	0.97
Teacher entropy	disagreement	PSV, Tabular	PSV	0.83	0.97
Entropy sum	disagreement	PSV, Tabular	PSV	0.83	0.97
Entropy product	disagreement	PSV, Tabular	PSV	0.83	0.97
$\exp((H_s + H_t)/2)$	disagreement	PSV, Tabular	PSV	0.83	0.97
$N(H)$	disagreement	PSV, Tabular	PSV	0.83	0.97
Soft-OR(H, D)	disagreement	PSV, Tabular	PSV	0.83	0.97
$N(D_{16}) \cdot C_{16}$	dis. × raw support	PSV, Tabular	PSV	0.83	0.97
$N(H) \cdot C_{16}$	dis. × raw support	PSV, Tabular	PSV	0.83	0.97
Entropy $\cdot C_{16}$	dis. × raw support	PSV, Tabular	PSV	0.83	0.97
Forward KL $\cdot C_{16}$	dis. × raw support	PSV, Tabular	PSV	0.83	0.97
Reverse KL $\cdot C_{16}$	dis. × raw support	PSV, Tabular	PSV	0.83	0.97
Soft-OR$(H, D) \cdot C_{16}$	dis. × raw support	PSV, Tabular	PSV	0.83	0.97
Raw support C_{16}	agreement	Vanilla, Step	Vanilla	-0.83	-0.97
Normalized support $N(C_{16})$	agreement	Vanilla, Step	Vanilla	-0.83	-0.97
Top-8 overlap	agreement	CanonFF, Vanilla	CanonFF	-0.14	-0.20
Top-16 overlap	agreement	CanonFF, Vanilla	CanonFF	0.14	-0.20
Top-32 overlap	agreement	CanonFF, PSV	CanonFF	0.43	0.64

Table 6: The high/low audit of §4.3 for every measured signal on the six ranked candidates, scored by the curve average over updates 10–140. Every distributional disagreement signal listed selects PSV; every agreement signal listed selects a different prompt, and no signal ranks the pool better than J. The signals that disagree differ only on the runner-up, Canonical Freeform against Tabular-CoT. Multiplying by *raw* support changes no ranking, because raw support spans only 0.16 percentage points across the pool. Here H denotes H_s; H_s and H_t are the student and teacher entropies, D_{16} is the KL on the top-16 union, and C_{16} is the teacher mass on the student's top-16. Soft-OR$(H, D) = N(H) + N(D) - N(H)N(D)$ with D the forward KL; CanonFF is Canonical Freeform.

B THE SELECTOR AUDIT ON ALL TEN ARMS

The main audit uses the six candidates fixed in advance. Table 7 repeats it, top three against bottom three, on all ten trained arms and on the nine that remain after removing PSV. On all ten arms the score orders all nine pairs correctly and separates the groups by 1.00 points, against 8/9 pairs and 0.71 for reverse KL. Removing PSV, so that no conclusion rests on the one prompt the method selects, keeps the high group ahead for every disagreement signal, at a smaller margin, and the score still separates the groups at least as well as any other signal (0.59 points, $\rho = 0.83$). The four PSV variants exist only because PSV was selected, which is why they are not part of the main pool.

Normalization. The quantile normalization maps each raw quantity onto a common scale but does not create the ordering: ranking the pool by the raw, unnormalized product $D_{16} \cdot C_{16}$ instead agrees with the normalized score at $\rho = 0.964$ and selects the same candidate, so the quantile normalization does not drive the decision.

Signal	Pick	Pairs	Gap	ρ
All ten arms				
J (ours)	PSV	9/9	1.00	0.88
Reverse KL	PSV	8/9	0.71	0.76
Forward KL	PSV	9/9	0.83	0.77
Student entropy	PSV	8/9	0.65	0.71
Raw support C_{16}	Vanilla	1/9	-0.66	-0.68
Nine without PSV				
J (ours)	CanonFF	9/9	0.59	0.83
Reverse KL	Tabular	8/9	0.46	0.67
Forward KL	Tabular	9/9	0.59	0.68
Student entropy	Tabular	8/9	0.41	0.60
Raw support C_{16}	Vanilla	1/9	-0.41	-0.57

Table 7: High/low audit with three prompts per group, scored by the curve average. Gap and ρ as in Table 2; abbreviations as in Table 6. "Pairs" counts the nine high×low pairs ordered correctly.

C THE SCORE ALONG THE ROLLOUT

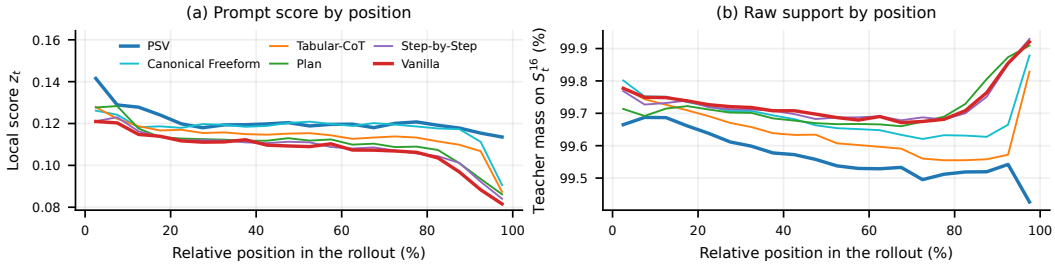


Figure 3: Local score and raw teacher support of the six ranked candidates along the untrained student's calibration rollouts, in 20 relative-position bins.

D THE CROSSED DESIGN IN FULL

Table 8 gives the four cells of the 2 × 2 design of §4.4 at their best checkpoints, per seed. Table 9 extends the crossing to the four prompts for which both token regimes were trained, which lets the high/low prompt audit be repeated inside each regime. On these four prompts the high-score pair beats the low-score pair by 1.183 points under full-token OPD and 1.983 points under token selection. Only the vanilla and selected prompts were trained on all three seeds, so this is a single seed and we do not read its margin closely.

Configuration	23	37	53	Mean	Δ
Vanilla prompt, full-token OPD	69.62	69.54	69.39	69.517	0.000
Selected prompt, full-token OPD	70.90	71.89	71.52	71.439	1.922
Vanilla prompt, TA token selection	68.12	69.40	71.64	69.719	0.202
Selected prompt, TA token selection	72.04	71.51	72.71	72.087	2.571

Table 8: The four cells of the 2 × 2 design at their best checkpoints, per seed and averaged. Δ is against selecting neither.

Prompt	J	Full	TA	Δ
PSV	0.1211	70.90	72.04	1.14
Tabular-CoT	0.1138	70.65	69.87	-0.78
Step-by-Step	0.1087	69.57	69.83	0.26
Vanilla OPD	0.1076	69.62	68.12	-1.50

Table 9: Four prompts under both token regimes at their best checkpoints, seed 23 only, ordered by J. Δ is TA minus Full for the same prompt.

E NATIVE AND MATCHED-PROMPT ACCURACY

The distilled models are evaluated under the plain task prompt throughout, because the prompt is a training-time intervention and the deployed model should not need it. Table 10 also applies each

prompt at inference. The untrained student is far weaker under PSV than without a prompt (49.994 against 68.732), and the drop is largest on MATH500 (96.16 to 53.93), the easiest benchmark, which points to responses that do not reach an answer within the budget rather than to wrong reasoning. Training under PSV raises accuracy under PSV by 15.69 points, to 65.688, by far the largest matched gain in the pool (the next is Tabular-CoT's 2.69): the student learns to carry out the instruction, which is the correction the score anticipated. The plain-prompt model is better still, so the instruction is best used as a training signal and dropped at deployment. Matched values use the checkpoint selected under the plain prompt.

Prompt	Native 9B	Native 27B	Trained, matched	Trained, plain
Vanilla OPD	68.732	81.270	69.620	69.620
PSV	49.994	79.034	65.688	70.902
Tabular-CoT	68.450	79.698	71.144	70.654
Plan	68.002	81.496	70.200	70.476
Formal PSV	69.672	81.030	70.100	70.290
Canonical Freeform	69.728	81.610	70.106	70.214
Semantic Freeform	66.764	80.718	68.244	70.054
Post PSV	67.126	79.900	68.108	69.608
Step-by-Step	68.128	80.620	69.406	69.570
Labels Only	67.338	81.708	68.548	69.484

Table 10: AVG5 with the candidate prompt applied at inference, supporting §5.2. "Native" columns apply the prompt to the untrained models; "trained, matched" applies it at test time to the distilled model, at the checkpoint selected under the plain prompt; "trained, plain" evaluates that checkpoint under the plain prompt.

F SUPERVISION DIAGNOSTICS

Two properties of the pool constrain how the score should be read. Support is saturated: every candidate places at least 99.5% of teacher mass inside the student's top-16 set (Appendix A), a spread of 0.16 percentage points, so the composite is dominated by the disagreement term and the product form is not shown to be necessary here. And the direction is not "minimize divergence": the candidates that help are those with *more* teacher–student divergence at the visited prefixes. That should not be extrapolated, since driving divergence arbitrarily high — with an instruction the student cannot follow — would leave the regime in which support stays high. The pool spans a narrow band of divergence at uniformly high support, and that is the band we measured. Position-resolved diagnostics (Figure 3) show the separation persisting across relative positions rather than decaying after the opening tokens, which argues against the instruction acting only as a prefix perturbation.

G SEED-TO-SEED VARIATION

Table 11 gives how far the best checkpoint of an identical arm moves across three seeds, for the four arms of the crossed design, the only arms trained on all three.

Arm	SD	Range
Full-token OPD, default prompt	0.095	0.230
Full-token OPD, selected prompt	0.409	0.992
TA token selection, default prompt	1.453	3.516
TA token selection, selected prompt	0.493	1.204

Table 11: Seed-to-seed standard deviation and range of each arm's best-checkpoint AVG5 across three seeds. This is the seed-to-seed variation referred to in §4.1.

H Offline prompt score versus the deployed token selector

These two quantities share a motivation and are not the same function; conflating them would misstate what the composition experiment tests. The offline prompt score is Eqs. 2–3: a forward KL between student and teacher distributions *renormalized on the union* of their top-16 sets, multiplied by a support term, each normalized by calibration-wide 5th/95th percentiles pooled over all candidates, and averaged with uniform weights per question, per rollout and per token.

The deployed token selector instead evaluates raw full-vocabulary probabilities on the student's top-16 ids,

$$d_t^{\text{train}} = \max\left(0, \sum_{v \in S_t^{16}} s_t(v) \log \frac{s_t(v)}{q_t(v)}\right), \qquad c_t^{\text{train}} = \sum_{v \in S_t^{16}} q_t(v), \tag{4}$$

so the disagreement term is a clipped *reverse*-direction contribution restricted to the student's own top-k set rather than a renormalized-distribution KL on the union. The two terms are combined by the vendored selection routine with `clip_top_frac=0.02`, and the adapter keeps exactly $\lceil 0.05N \rceil$ valid positions per selection batch — a global budget with no per-response floor — with the loss normalized by the number of selected tokens.

The differences are therefore the KL direction, the token set the disagreement is measured on, the normalization, and batch-global versus per-response budgeting. We call the prompt score TA-inspired for this reason, and we describe the joint runs of §4.4 as two related mechanisms applied together rather than as one calibrated functional acting on both axes.

I Training curves

I.1 Seeds

Each seed is plotted separately. Averaging curves whose maxima occur at different updates would misrepresent both the level and the location of the peak, and missing benchmark cells are never interpolated: lines connect complete AVG5 points only.

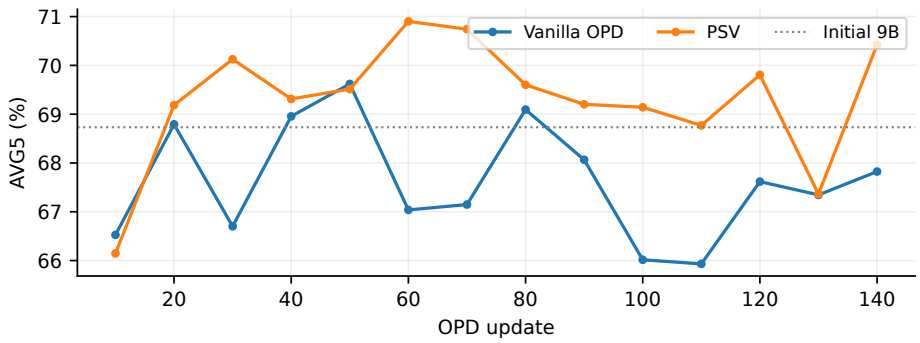


Figure 4: Seed 23. Missing benchmark cells are not interpolated into AVG5; lines only connect complete points.

I.2 Every candidate against the common reference

Each panel shows one candidate against the same prompt-free OPD curve on seed 23, so that the candidate's advantage or deficit can be read as a sustained offset rather than as a difference of maxima.

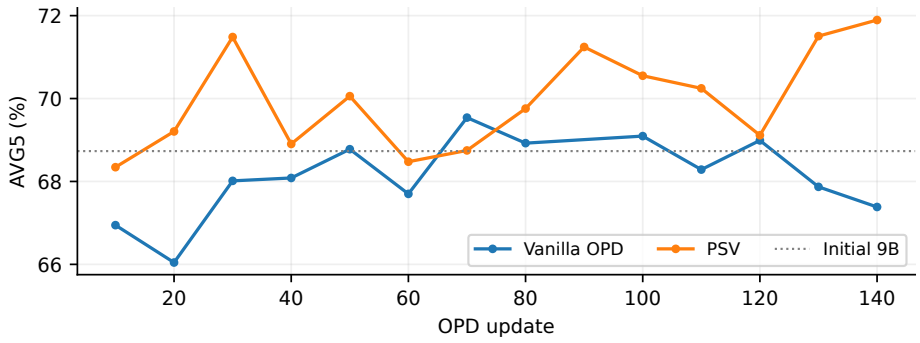


Figure 5: Seed 37. Missing benchmark cells are not interpolated into AVG5; lines only connect complete points.

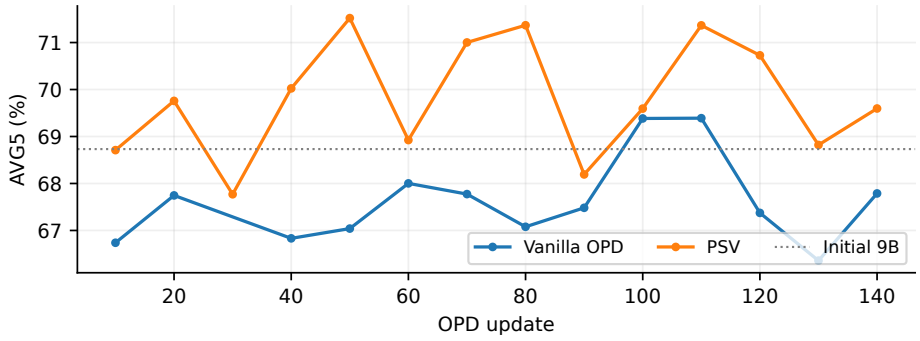


Figure 6: Seed 53. Missing benchmark cells are not interpolated into AVG5; lines only connect complete points.

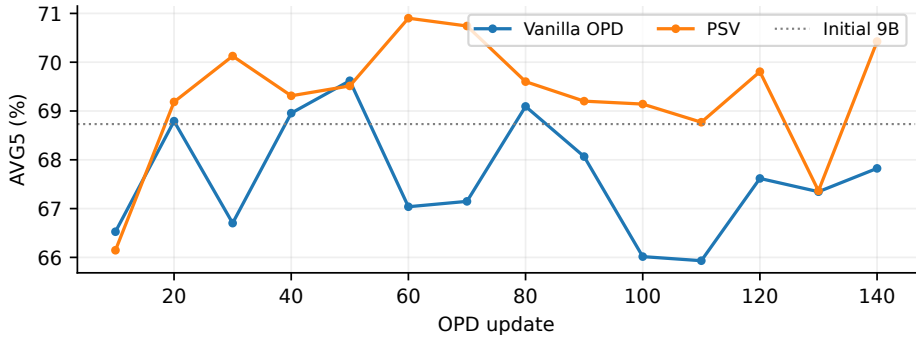


Figure 7: PSV and the common vanilla reference, seed 23. Plain task prompt, 32K output cap, average of 16 samples.

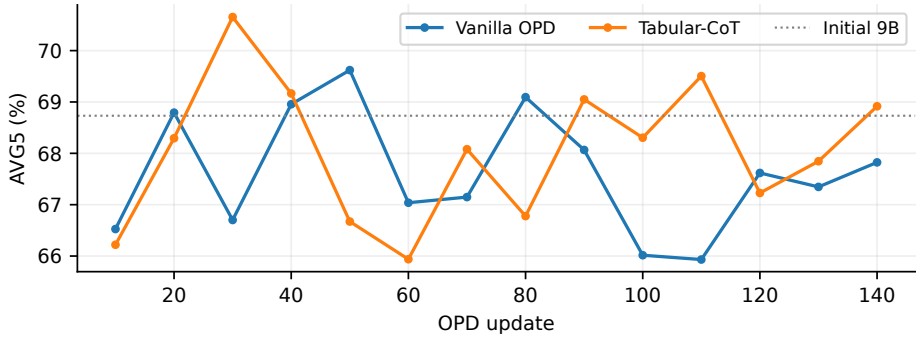


Figure 8: Tabular-CoT and the common vanilla reference, seed 23. Plain task prompt, 32K output cap, average of 16 samples.

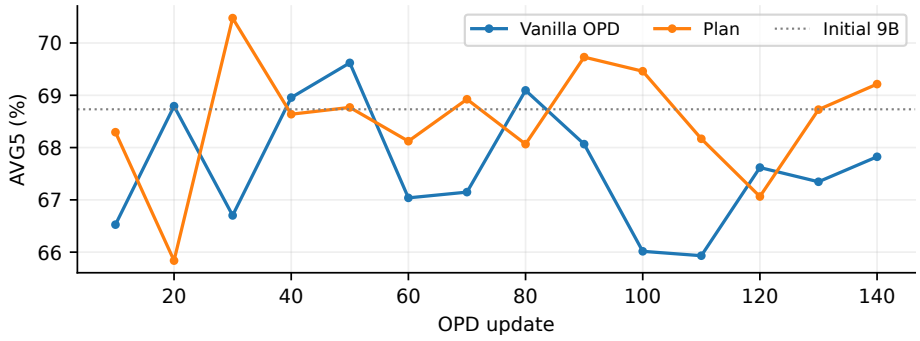


Figure 9: Plan and the common vanilla reference, seed 23. Plain task prompt, 32K output cap, average of 16 samples.

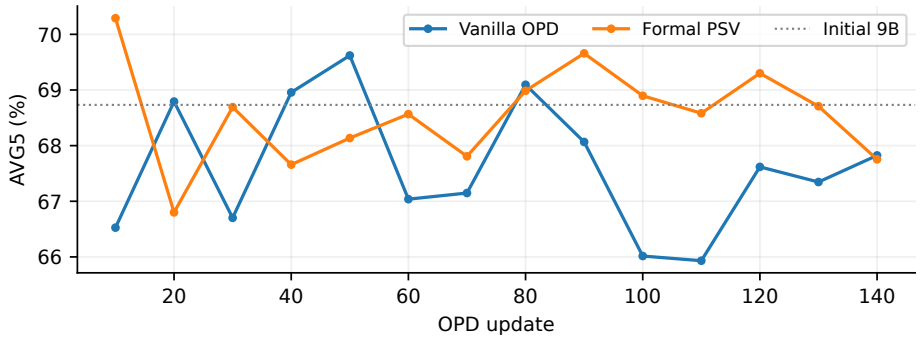


Figure 10: Formal PSV and the common vanilla reference, seed 23. Plain task prompt, 32K output cap, average of 16 samples.

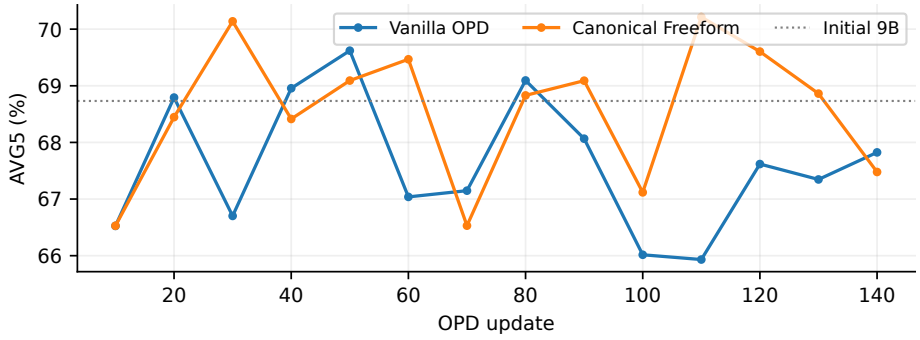


Figure 11: Canonical Freeform and the common vanilla reference, seed 23. Plain task prompt, 32K output cap, average of 16 samples.

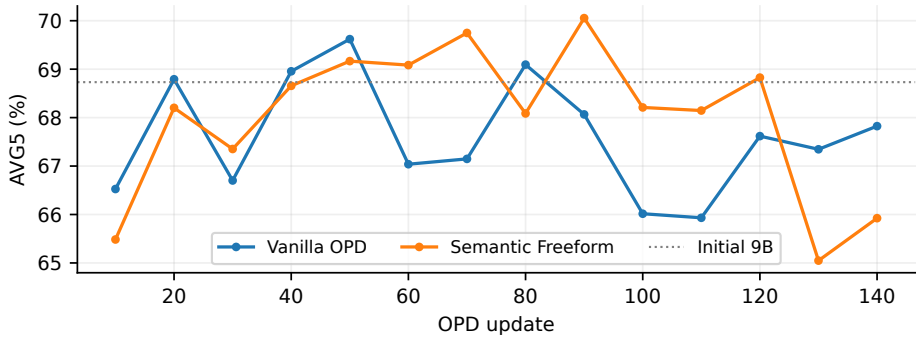


Figure 12: Semantic Freeform and the common vanilla reference, seed 23. Plain task prompt, 32K output cap, average of 16 samples.

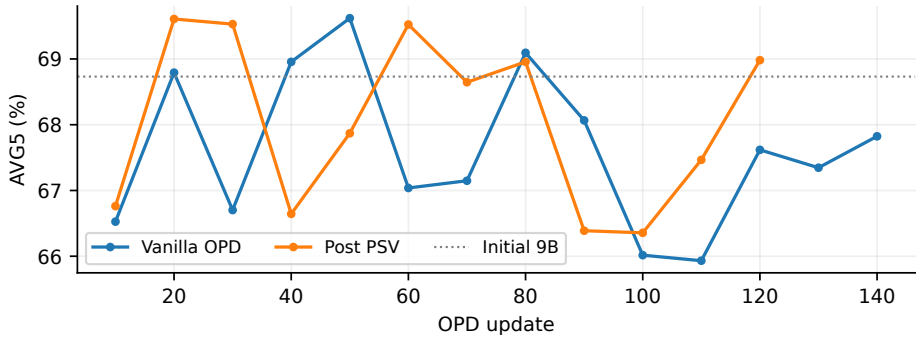


Figure 13: Post PSV and the common vanilla reference, seed 23. Plain task prompt, 32K output cap, average of 16 samples.

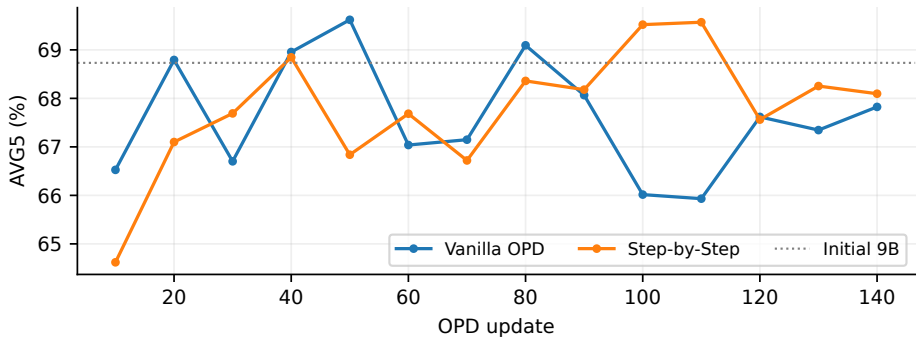


Figure 14: Step-by-Step and the common vanilla reference, seed 23. Plain task prompt, 32K output cap, average of 16 samples.

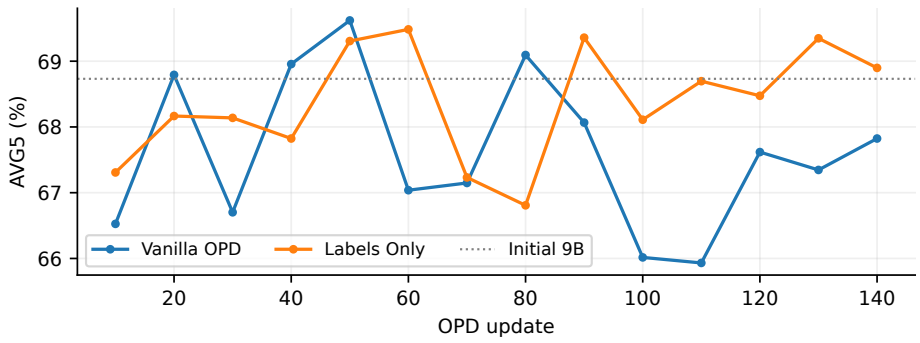


Figure 15: Labels Only and the common vanilla reference, seed 23. Plain task prompt, 32K output cap, average of 16 samples.

J BASELINE RECIPES

SeqKD fine-tunes the initial student for one epoch on $89{,}836$ correct, complete teacher responses generated at a 64K budget, with full thinking supervision, global batch 64, learning rate 10^{-5} with cosine decay and 5% warm-up, and a $98{,}304$-token sequence limit; OPD then starts from the fine-tuned weights under the recipe of §4.1. We used this single recipe and did not tune it. The Lightning OPD 2.0 reproduction follows the published description as far as we could and is reported only descriptively.

K EXACT CANDIDATE PROMPTS

The following strings are appended to the task while preserving its original final-answer contract. Naming is stable throughout the paper: PSV is the variant with the explicit in-thinking clause, Formal PSV is the same instruction without that clause, and Post PSV constrains only the post-thinking answer. The behavioral constraints are adaptations inspired by the cited reasoning-prompt literature and are not verbatim reproductions of those papers' prompts.

K.1 VANILLA OPD

```
[No added style instruction.]
```

K.2 PSV

```
Reasoning-style constraint for this response. Apply it from the first reasoning token and
    inside the reasoning trace. It does not change the required final-answer format:
Start the reasoning with `Plan:` and organize it in exactly three ordered sections: `Plan:`,
    `Solve:`, and `Verify:`. Plan states the route, Solve executes it, and Verify checks the
    decisive condition once. Do not restart a section.

Thinking-channel placement (mandatory): apply the organization above to the reasoning text
    inside <think>...</think>, meaning the generated text before the first </think> close
    marker. The first non-whitespace text inside <think> must be `Plan:`, and the requested
    organization must continue there until </think>. Do not postpone the requested
    organization until after </think> or apply it only to the visible final answer.
```

K.3 TABULAR-COT

```
Reasoning-style constraint for this response. Apply it from the first reasoning token and
    inside the reasoning trace. It does not change the required final-answer format:
Start the reasoning with this exact Markdown header: `| Step | Subgoal | Derivation | State
    |`. Follow it with the separator row and one concise reasoning transition per table row.
    Carry symbols through the State column and do not restart the table.
```

K.4 PLAN

```
First state an overall route and subgoals under Plan:. Then work through the solution in your
    usual freeform presentation. Keep the original final-answer requirements unchanged.
```

K.5 FORMAL PSV

```
Reasoning-style constraint for this response. Apply it from the first reasoning token and
    inside the reasoning trace. It does not change the required final-answer format:
Start the reasoning with `Plan:` and organize it in exactly three ordered sections: `Plan:`,
    `Solve:`, and `Verify:`. Plan states the route, Solve executes it, and Verify checks the
    decisive condition once. Do not restart a section.
```

K.6 CANONICAL FREEFORM

```
Reasoning-style constraint for this response. Apply it from the first reasoning token and
    inside the reasoning trace. It does not change the required final-answer format:
Start the reasoning with `Approach:`. Then use concise continuous freeform reasoning without
    numbered steps, additional section headings, bullet lists, or tables. Reuse symbols
    consistently and correct errors inline.
```

K.7 SEMANTIC FREEFORM

```
Identify an overall route and subgoals, work through the solution, and independently check
    the final result. Use your usual freeform presentation; no special headings are
    required. Keep the original final-answer requirements unchanged.
```

K.8 POST PSV

```
Format the final user-visible solution into exactly three ordered sections: `Plan:`, `Solve:`,
    and `Verify:`. Plan states the route, Solve executes it, and Verify checks the decisive
    condition once. Do not restart a section. Apply this organization only to the final
    user-visible response; your private reasoning may use its usual freeform style. Preserve
    the original required final-answer format.
```

K.9 STEP-BY-STEP

```
Organize your reasoning into consecutively numbered steps labeled Step 1:, Step 2:, and so
    on. Each step should advance the solution. Keep the original final-answer requirements
    unchanged.
```

K.10 LABELS ONLY

```
Organize your response with exactly one ordered sequence of headings: Plan:, Solve:, Verify:.
    Keep the original final-answer requirements unchanged.
```

L EVIDENCE PACKAGE AND PROTOCOL DETAILS

Supplementary code and data. The supplementary material contains: a standalone implementation of the prompt score (per-token statistics of Eq. 2, pooled quantile normalization and the aggregation of Eq. 3) with a calibration driver; reference implementations of the per-token OPD loss and of the token selector of Appendix H; the ten prompts; the training and evaluation configurations; and anonymized per-checkpoint results for every arm reported here (per-benchmark accuracies and AVG5), the per-candidate calibration signals and the position-binned statistics behind Figure 3. A script recomputes Tables 1–3 and the headline numbers from these files with no network, GPU or API access, checks them against the values printed here, and fails rather than emitting a partial AVG5. Raw model outputs and checkpoints are not included. Training used an in-house distributed trainer that is not released; the loss, selector and every non-default setting are given in the reference code and configuration.

Training configuration retained. 140 updates, prompt batch 128, one rollout per question, 32K response budget, learning rate 5×10^{-6}, no task reward, 32 GPUs per job. Seeds are 23/37/53 and control the order of training data. The token-selection arms use the TA scoring mode with retention 0.05, an exact token budget, top-16 support and selected-token loss normalization. Pipeline and sequence-parallel layouts are recorded per arm rather than inferred from the GPU count, since equal GPU counts do not imply equal realized token throughput.

Evaluation configuration retained. Thinking mode, 32K output cap, 64K context, 16 sampled responses per question including MATH500, local answer verification. Sampler settings, tokenizer and post-training template, dataset revisions and prompt suffixes are preserved in the manifests. No score obtained under one budget is relabeled as another.

Known gaps. Two historical benchmark cells in the baseline curves (one HMMT25 and one MATH500 point on prompt-free arms at particular checkpoints) were not recovered by the current harvester because of a schema change in an older job configuration; the affected AVG5 points are omitted rather than imputed. They do not alter the reported maxima but could alter the shape of those curves. The four-arm design of §4.4 is complete for 38 of the 42 possible seed–checkpoint cells, counting a cell only when all four arms are evaluated. Online generation length and truncation were logged for the two head-to-head arms only: $22{,}997$ against $22{,}659$ mean tokens and 30.7% against 27.9% truncation for the selected and default prompts respectively.