LADDEROPD: LOGIT-FREE ON-POLICY DISTILLATION FROM THE TEACHER'S TOP-k ORDER

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains a small model on its own samples while a teacher scores every token with its full next-token distribution. Many strong teachers are served through interfaces that return only sampled text or a short list of top candidates, so this distribution is often unavailable. We study how much of it OPD actually needs. We first introduce GroundKL, a measurement that takes an open teacher's logits as ground truth and tests how well each signal a closed interface could return recovers the exact per-token OPD signal. The signal is concentrated at high-entropy forking tokens, and there the order of the teacher's top-k candidates recovers it nearly as well as sixteen samples from the teacher. Guided by this finding, we propose LadderOPD, which trains on the teacher's ranked top-k token IDs alone. Across five teacher–student settings in two model families, LadderOPD is statistically equivalent to OPD with top-k probabilities or full logits within three GSM8K points in all ten comparisons, and it matches them on code as well. Shuffling the order at a growing share of forking tokens degrades the student step by step. For verifiable reasoning with a shared tokenizer, a teacher's ranked top-k list is therefore enough for on-policy distillation.

1 INTRODUCTION

Most capable small language models are distilled from larger ones (Hinton et al., 2015; DeepSeek-AI et al., 2025; Yang et al., 2025). For reasoning, on-policy distillation (OPD) works especially well. The student samples its own solutions and the teacher scores every token, so supervision falls on the states the student actually visits (Agarwal et al., 2024; Gu et al., 2024; Lu & Thinking Machines Lab, 2025). This avoids the exposure bias of imitating teacher text (Bengio et al., 2015; Ranzato et al., 2016; Ross et al., 2011), and because every token receives its own training signal, OPD needs far less compute than reinforcement learning (Lu & Thinking Machines Lab, 2025; Yang et al., 2025).



Figure 1: How much of the teacher does on-policy distillation need? (a) The access ladder at one forking token. Each rung adds information, and real interfaces stop at different rungs. LadderOPD uses L1, the ranked top-k IDs without their probabilities. (b) Mean GSM8K accuracy over the three Qwen2.5 pairs of Table 2. The order matches probabilities and logits, and a shuffled order falls below the untrained student.

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.

OPD, however, assumes that the teacher reveals its next-token distribution at every visited state, and many of the strongest teachers do not. Reasoning interfaces return sampled text, and chat interfaces that report log-probabilities cap the list at five to twenty candidates, partly because full distributions leak a model's internals (Carlini et al., 2024; Finlayson et al., 2024; Morris et al., 2024). Without logits, practitioners fall back on imitating teacher text (Kim & Rush, 2016; Hsieh et al., 2023), which reintroduces exposure bias, or on sequence-level and chunk-level rewards (Ye et al., 2025; Zhou et al., 2026), which lose the per-token signal of OPD.

This paper asks which part of the teacher's distribution OPD actually uses. We order teacher access as a ladder of four levels (Figure 1a). L0 is the teacher's top-1 token, L1 the top-k token IDs in the teacher's order, L2 adds their probabilities and L3 is the full logit vector. Any interface that offers L2 can offer L1 by dropping the probabilities, yet L1 has not been studied as a teacher signal. To see which level is needed, we introduce GroundKL, which uses an open teacher as ground truth and measures how much of the exact per-token OPD signal each level recovers (Section 4). The signal sits mostly at the high-entropy tokens where the student is uncertain, in line with the forking-token view of reasoning (Wang et al., 2025; Xu et al., 2026), and at these tokens the teacher's ranking of its top candidates recovers it nearly as well as sixteen samples from the teacher. Based on this finding we propose LadderOPD, an OPD method that trains on the ranked top-k IDs alone (Section 5).

Our experiments support three conclusions (Section 6). First, the ranking is enough. Across five teacher–student settings and two model families, LadderOPD is statistically equivalent to OPD with top-k probabilities or full logits within three GSM8K points in all ten comparisons (Figure 4a). Second, the ranking is causal. Shuffling it at a growing share of forking tokens degrades the student step by step, and a fully shuffled ranking leaves every student below its starting accuracy. Third, less information costs accuracy. Training on the top-1 token alone gives slightly weaker students, and training on teacher text gives clearly weaker ones. Our contributions are

- GroundKL, which measures how much of the exact OPD signal each logit-free signal recovers;
- LadderOPD, an on-policy distiller that needs only the teacher's ranked top-k token IDs and the same teacher compute as value-based OPD;
- a controlled comparison across the access ladder with equivalence tests and shuffled-order controls, covering code, a second model family, contamination-controlled evaluation and a 32B teacher.

2 Related Work

On-policy distillation. GKD (Agarwal et al., 2024) and MiniLLM (Gu et al., 2024) train the student on its own samples against the teacher's distribution. Later work tunes the divergence (Wen et al., 2023; Ko et al., 2024; 2025), interleaves teacher and student sampling (Xu et al., 2025), or restricts matching to the teacher's renormalized top-k (Fu et al., 2026). Token selection keeps the most informative positions (Xu et al., 2026), and OPRD moves the objective into hidden states (Yang et al., 2026). Offline distillation for pre-training subsamples cached teacher logits (Anshumann et al., 2025) or uses teacher signals to select data (Gu et al., 2025), and production pipelines use OPD to train small models from flagship ones (Yang et al., 2025; Lu & Thinking Machines Lab, 2025). All of them read the teacher's probabilities or internals, as does most LLM distillation (Xu et al., 2024).

Distillation without logits. Sequence-level distillation imitates teacher outputs (Kim & Rush, 2016), optionally with rationales (Hsieh et al., 2023). On-policy logit-free methods reward whole sequences with a discriminator (GAD, Ye et al., 2025), score the student's chunks at forks against sampled teacher continuations (OmniOPD, Zhou et al., 2026), or use the teacher's top-1 token as a per-token target (TGPO, Liu et al., 2026). PLaD instead learns offline from pseudo-preference pairs of teacher and student outputs (Zhang et al., 2024a). Cross-tokenizer methods still need the teacher's distribution, matching sorted probability vectors (Boizard et al., 2025) or aligning tokens through cross-model attention (Zhang et al., 2024b). None of these treats the teacher's ranking as the unit of supervision. OmniOPD checks its chunk scores against an open teacher's exact chunk probabilities, but none measures token by token how much of the exact OPD signal each logit-free observable recovers.

Ranks and relations in distillation. Matching relations instead of raw values has a history in classification and retrieval. DIST matches correlations between teacher and student predictions instead of their values (Huang et al., 2022), and ranking distillation transfers a recommender's top-ranked items (Tang & Wang, 2018), while decoupled distillation reweights the target and non-target

parts of the KL (Zhao et al., 2022). Work on the capacity gap shows that a stronger teacher does not always produce a better student (Cho & Hariharan, 2019; Mirzadeh et al., 2020). We bring the rank view to on-policy distillation, where the ranking is what a closed interface can offer.

3 BACKGROUND

On-policy distillation. Let π_S be the student and p_T the teacher. The student samples $y \sim \pi_S(\cdot \mid x)$, and at each position t with state $s_t = (x, y_{<t})$ it minimizes the per-token reverse KL

$$\mathcal{L}_t^{\mathrm{OPD}} = \mathrm{KL}\big(\pi_S(\cdot \mid s_t) \,\|\, p_T(\cdot \mid s_t)\big). \tag{1}$$

Its single-sample gradient weights the log-probability of the sampled token y_t by the log-ratio

$$g_t^\star = \log p_T(y_t \mid s_t) - \log \pi_S(y_t \mid s_t), \tag{2}$$

which we call the per-token OPD signal. Positive $g_t^\star$ raises the sampled token and negative $g_t^\star$ lowers it. The reverse direction makes the objective mode-seeking, so the student concentrates on continuations the teacher rates highly instead of spreading mass over everything the teacher allows (Gu et al., 2024; Agarwal et al., 2024). Computing $g_t^\star$ thus needs the teacher's logits.

Teacher access. At a state s_t we compare four levels of access. L0 returns the teacher's top-1 token. L1 returns the top-k set $S_t = (v_1, \ldots, v_k)$ ordered by the teacher's preference. L2 adds the probabilities $p_T(v_i \mid s_t)$ of that set, and L3 is the full distribution. Log-probability chat interfaces provide L2 with $k \leq 20$, reasoning interfaces return only sampled text, below L0, and standard OPD needs L3. Throughout, the teacher scores the student's sequence and returns its access level at every position, as prompt log-probability endpoints do. We seek the lowest rung that supports OPD.

4 GROUNDKL: WHAT DOES EACH SIGNAL RECOVER?

We sample two solutions to each of 200 GSM8K training problems from Qwen2.5-1.5B-Instruct, score all 81,212 tokens under Qwen2.5-7B-Instruct as an open teacher, and compute $g_t^\star$ exactly. We then ask how well each quantity that a closed interface could return predicts $g_t^\star$. The *order* signal is the negative rank of y_t in the teacher's top-k with $k = 20$, the *membership* signal records whether y_t is in that set at all, and *top-1 agreement* records whether y_t is the teacher's argmax. As a reference we draw 16 teacher samples at the same state and score y_t by its empirical frequency. This sampling ceiling, at sixteen queries per token, marks what repeated black-box access recovers at that budget. We split tokens by student entropy, calling the most uncertain 34% forking tokens, and by whether y_t lies in the teacher's top-k. For comparison we score the teacher's top-k log-probability of y_t, which the L2 rung returns, in the same way. Minus the student's log-probability it reproduces $g^\star$ up to the teacher's top-k mass, a Spearman correlation of 0.999.

Table 1: Recovery of $g^\star$ at forks inside the top-k (Spearman), from what each interface returns. The 3B teacher scores 0.5B rollouts.

Signal	3B	7B	14B
Order (L1)	0.635	0.625	0.608
Top-1 agreement (L0)	0.582	0.587	0.566
Sampling, 16 queries	0.667	0.645	0.640
Top-k values (L2)	0.701	0.669	0.665
Off-support share of $\lvert g^\star \rvert$	13%	6%	16%

The signal is concentrated. Sorted by student entropy, the top 20%, 34% and 50% of tokens carry 65%, 90% and 99% of the total $|g^\star|$ (Figure 2, left). Tokens outside the teacher's top-k are rare, 0.3% of all tokens, but each has a negative $g^\star$, averaging -6.9 against -0.23 overall, and together they hold 6% of the mass. The teacher's advice there is to avoid them, which no in-set signal can express.

The order recovers it. At forking tokens inside the set, the order signal reaches a Spearman correlation of 0.63 with $g^\star$, against 0.64 for the sampling ceiling and 0.67 for the teacher's values (Figure 2, right). Top-1 agreement reaches 0.59 but takes only two values, so it cannot tell the second-best candidate from the tenth. Membership is constant inside the set and recovers nothing by construction. The pattern holds for teachers of 3B, 7B and 14B parameters, where the order recovers 0.61 to 0.64 against 0.64 to 0.67 for sampling and 0.66 to 0.70 for the values (Table 1). On low-entropy tokens the order recovers only 0.16, since the sampled token is then nearly always the teacher's first choice. Two design rules follow. Off-support tokens need a term that acts on the set itself, and the ranking must be supervised at the tokens where the student forks.

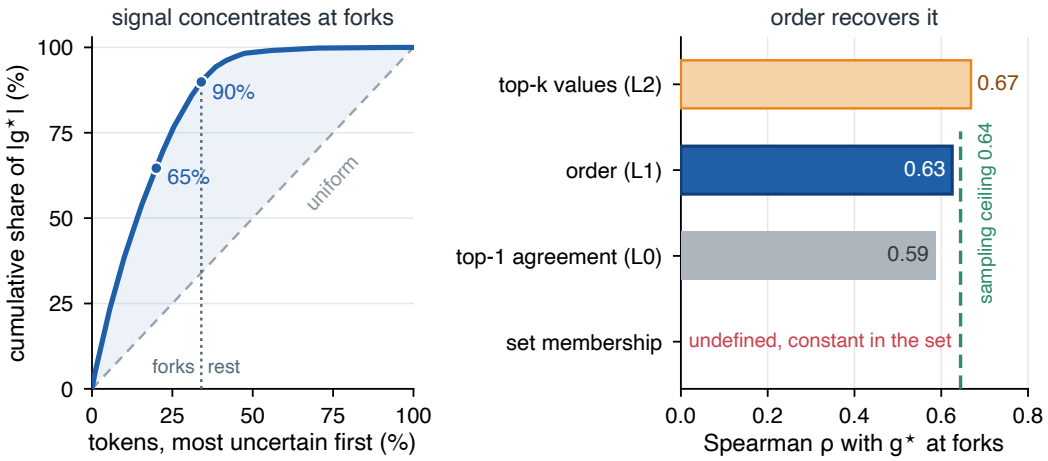


Figure 2: GroundKL. Left, the cumulative share of $|g^\star|$ held by the most uncertain tokens, 7B teacher. Right, how well each signal recovers $g^\star$ at forks, the most uncertain 34%, inside the top-k, with L2 as reference.

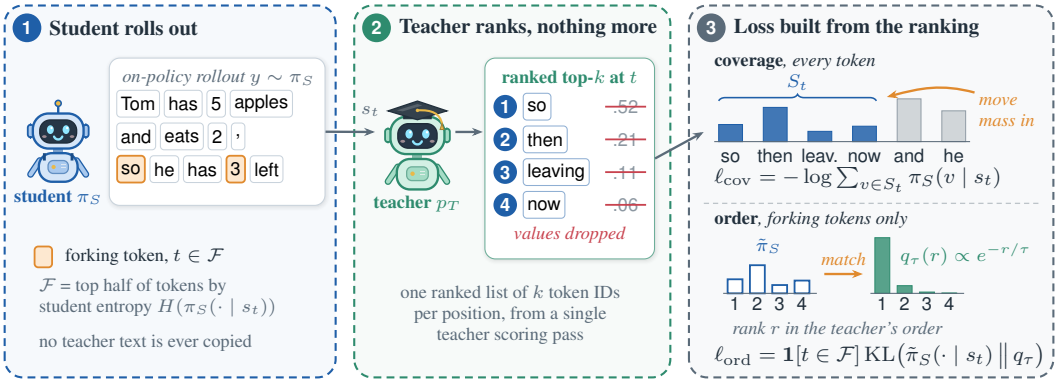


Figure 3: One step of LadderOPD. The student writes a solution and marks forking tokens by entropy. The teacher returns its ranked top-k IDs at each position, without probabilities. Coverage pulls the student's mass into that set at every token, and at forking tokens the student's distribution inside the set is matched to the rank-tempered target of Equation 3.

5 LadderOPD

On the student's rollout, at position t LadderOPD reads the teacher's ordered set $S_t = (v_1, \ldots, v_k)$ and nothing else. Let $\tilde{\pi}_S(r \mid s_t) = \pi_S(v_r \mid s_t) / \sum_{j=1}^{k} \pi_S(v_j \mid s_t)$ be the student's distribution renormalized on S_t and indexed by teacher rank r, and let q_τ be a fixed target over ranks,

$$q_\tau(r) = \frac{\exp(-r/\tau)}{\sum_{j=1}^{k} \exp(-j/\tau)}, \qquad r = 1, \ldots, k. \tag{3}$$

The per-token loss combines a coverage term and an order term,

$$\mathcal{L}_t = \underbrace{-\log \sum_{v \in S_t} \pi_S(v \mid s_t)}_{\ell_{\text{cov}}} + \lambda \underbrace{\mathbf{1}[t \in \mathcal{F}]\, \mathrm{KL}\big(\tilde{\pi}_S(\cdot \mid s_t) \,\|\, q_\tau\big)}_{\ell_{\text{ord}}}, \tag{4}$$

where $\mathcal{F}$ contains the completion tokens whose student entropy is at least the median over the batch, a wider set than the 34% of Section 4, and $\lambda = 1$. The batch loss averages ℓ_{cov} over all completion tokens and ℓ_{ord} over forking tokens. Figure 3 shows one step and Algorithm 1 the full procedure.

Each term answers one GroundKL finding. The coverage term ℓ_{cov} handles the rare off-support tokens, whose large negative $g^\star$ no in-set signal can express, and needs only the unordered set. The order term ℓ_{ord} supervises the ranking where GroundKL located it. Because q_τ depends on ranks

Algorithm 1 LadderOPD training

Require: student π_S, ranked interface $\textsc{TopK}(s)$ that returns the teacher's top-k IDs in order, prompts $\mathcal{D}$, temper τ, weight λ

$q_\tau(r) \leftarrow \exp(-r/\tau) / \sum_{j=1}^{k} \exp(-j/\tau)$ for $r = 1, \dots, k$
for step $= 1, \dots, N$ **do**
  Sample prompts $x \sim \mathcal{D}$ and G rollouts $y \sim \pi_S(\cdot \mid x)$ per prompt
  Query $S_t \leftarrow \textsc{TopK}(s_t)$ for every completion position, in one teacher pass per rollout
  $H_t \leftarrow$ entropy of $\pi_S(\cdot \mid s_t)$, without gradient; $\mathcal{F} \leftarrow \{t : H_t \geq \mathrm{median}(H)\}$
  $\ell_{\mathrm{cov}} \leftarrow \mathrm{mean}_t \left[-\log \sum_{v \in S_t} \pi_S(v \mid s_t) \right]$
  $\ell_{\mathrm{ord}} \leftarrow \mathrm{mean}_{t \in \mathcal{F}} \ \mathrm{KL}\big(\tilde{\pi}_S(\cdot \mid s_t) \,\|\, q_\tau\big)$ ▷ $\tilde{\pi}_S$ indexed by teacher rank
  Update π_S with AdamW on $\ell_{\mathrm{cov}} + \lambda\, \ell_{\mathrm{ord}}$
end for

alone, no probability value ever crosses the interface. The temper τ sets how much the target trusts the top candidate. As $\tau \to 0$ the target approaches the top-1 label, and a large τ flattens it toward uniform inside the set. We use $\tau = 0.5$, which gives rank one 86% of the target and spreads the other 14% over a short ordered tail.

A worked example. Take the forking token in Figure 3, where the teacher ranks *so*, *then*, *leaving* and *now* as its first four candidates. The coverage term raises the total probability that the student gives to the teacher's candidates and lowers every other token, whatever their order. The order term renormalizes the student's distribution over the candidates and compares it with q_τ, which at $\tau = 0.5$ gives 86% to *so*, 12% to *then* and less than 2% to each later rank. If the student prefers *then* to *so*, Equation 6 lowers the logit of *then* and raises that of *so* until the two rankings agree. At no point does a teacher probability enter the update.

What the teacher provides. Per completion token the teacher returns one list of k token IDs. Producing it takes one forward pass over the student's rollout, the same teacher compute as value-based OPD, so the two differ only in how much of that pass leaves the teacher. The loss uses the order only at forking tokens and the unordered set elsewhere, so an interface needs to reveal the ranking only at forks.

Relation to value-based OPD. Our top-k value-based baseline (L2) keeps the same coverage term and replaces the order term with $\mathrm{KL}(\tilde{\pi}_S \,\|\, \tilde{p}_T)$, where $\tilde{p}_T$ is the teacher renormalized on S_t, applied at every token in the main pairs and at forks with the 32B teacher. The order term has this form with teacher log-probabilities replaced by a linear function of rank,

$$\log \tilde{p}_T(v_r \mid s_t) \ \longrightarrow \ \log q_\tau(r) = -\frac{r}{\tau} - \log Z_\tau, \tag{5}$$

with Z_τ the normalizer of Equation 3. Writing z_r for the student's logit of v_r, the gradient of ℓ_{ord} is

$$\frac{\partial \ell_{\mathrm{ord}}}{\partial z_r} = \tilde{\pi}_S(r \mid s_t) \Big[\log \tilde{\pi}_S(r \mid s_t) - \log q_\tau(r) - \mathrm{KL}\big(\tilde{\pi}_S \,\|\, q_\tau\big) \Big], \tag{6}$$

so a candidate whose log-ratio to the rank target exceeds the student's average log-ratio is pushed down and one below that average is pulled up. The teacher enters every update only through q_τ and hence through its ranks, the quantity GroundKL found to carry the signal.

6 Experiments

Models and data. We use three Qwen2.5-Instruct teacher–student pairs (Qwen et al., 2024), 7B→1.5B, 3B→0.5B and 14B→1.5B, and train on a 4,000-problem pool from the GSM8K training set (Cobbe et al., 2021). We evaluate on the 1,319 problems of the GSM8K test set and on MATH500 (Hendrycks et al., 2021; Lightman et al., 2024) with greedy decoding, a 512-token budget and exact-match scoring. Section 6.5 adds a Llama pair (Grattafiori et al., 2024), code generation on MBPP (Austin et al., 2021), GSM-Symbolic (Mirzadeh et al., 2025) and a 32B teacher.

Training. All access levels share one trainer and one set of hyperparameters and differ only in the loss target. Each step samples 2 prompts with 4 rollouts each at temperature 0.8 and up to 256 new tokens. We train for 150 AdamW steps (Loshchilov & Hutter, 2019) at a learning rate of 10^{-5} with

Table 2: The access ladder on three Qwen2.5 pairs. Accuracy in %, GSM8K as mean±std over seeds and MATH500 as a mean. Untrained rows are the public checkpoints and bold marks the best trained method. †Scored on a fixed 200-problem subset. ‡Simplified reimplementation.

		7B→1.5B		3B→0.5B		14B→1.5B	
Access	Method	GSM8K	MATH	GSM8K	MATH	GSM8K	MATH
none	Untrained student	66.2	34.0	40.9	17.6	66.2	34.0
text	SeqKD	70.7±0.8	36.2	42.9±0.8	22.5†	69.3±0.6	34.7
text	OmniOPD‡	61.8±8.3	35.7	37.5±5.1	23.7	64.2±3.5	36.9
text	GAD	68.5±3.5	34.6	40.9±1.1	22.5	68.1±3.6	36.4
L0	Top-1 CE	72.6±0.7	36.3	48.0±0.8	22.7	73.1±0.5	35.8
L1	**LadderOPD**	**74.4**±1.3	36.8	**51.9**±1.1	23.7	**74.2**±0.6	36.2
L2	Top-k values	73.5±1.2	**37.3**	50.8±0.7	**24.0**	73.6±0.6	**37.7**
L3	Full logits	73.9±0.9	37.1	50.8±0.8	23.7	73.6±0.5	37.4
ctrl	Shuffled order	56.7±2.4	27.5	19.3±2.8	12.0	57.3±2.8	27.5

10 warmup steps, and LadderOPD uses $k = 20$ and $\tau = 0.5$. The 32B→7B pair takes 40 updates over 160 micro-steps at 7×10^{-6}. A run takes about half an hour, and Appendix C lists every setting.

Baselines. The ladder is completed by cross-entropy on the teacher's top-1 token (L0, the TGPO target), the coverage term plus reverse KL to the teacher's renormalized top-k probabilities (L2, following Revisiting-OPD), and full-vocabulary reverse KL (L3). We also run SeqKD on correctness-filtered teacher solutions, GAD ported from its official release, and a simplified chunk-verification baseline in the style of OmniOPD, which has no public code (Appendix D). A matched control replaces the teacher's order with a random permutation of the same top-k set, all else fixed.

Statistics. LadderOPD uses six seeds per pair on GSM8K and three on MBPP and with the 32B teacher, baselines three to six, all on shared seeds so that every comparison is paired. We test equivalence with two one-sided tests (TOST) at margins of ± 3 and ± 5 GSM8K points (Schuirmann, 1987; Lakens, 2017).

6.1 The order is enough

Table 2 reports every rung at matched seeds. Order-only LadderOPD reaches 74.4, 51.9 and 74.2 on GSM8K across the three pairs, against 73.5, 50.8 and 73.6 with top-k probabilities and 73.9, 50.8 and 73.6 with full logits. The order is significantly better than probabilities on the 3B pair ($p = .002$) and indistinguishable from both value-based rungs in every other comparison (Appendix E).

A non-significant difference is not evidence of equivalence, so Figure 4a places the 90% confidence interval of each paired difference against the two margins, for all five settings. All ten intervals lie inside ± 3, which is TOST equivalence at $\alpha = 0.05$ with $p \leq .041$, and no paired difference exceeds 1.1 points in either direction. Within three points, the ranking reproduces what OPD extracts from the teacher's probabilities and from its logits.

Figure 5 shows the gain over the untrained student in all five settings. Order, probabilities and logits add 9.9 to 15.6 GSM8K points to the 0.5B and Llama students and 7 to 8 points to the 1.5B student under either teacher, where they also raise MATH500 by 2 to 4 points. The 7B student of the 32B teacher starts at 91.7, and all three rungs keep it within 1.3 points of that start on GSM8K while adding 2.4 to 4.0 points on MATH500. On GSM8K the order tracks what OPD does to each student, and a shuffled order leaves every student 5 to 22 points below where it started.

6.2 The order is causal

If the ranking carries the signal, a wrong ranking should hurt. The matched control keeps the teacher's top-k set, the forking positions and every hyperparameter, and only permutes the order. It drops the student below its untrained accuracy on every pair, to 56.7, 19.3 and 57.3, while coverage alone, which sees the same set without its order, reaches 72.6 on the 7B pair (Table 3). A wrong order is thus far worse than no order, and removing the order costs 2.3 points on the three shared seeds ($p = .03$). Figure 4b shuffles a controlled fraction of forking tokens on that pair. Accuracy falls monotonically

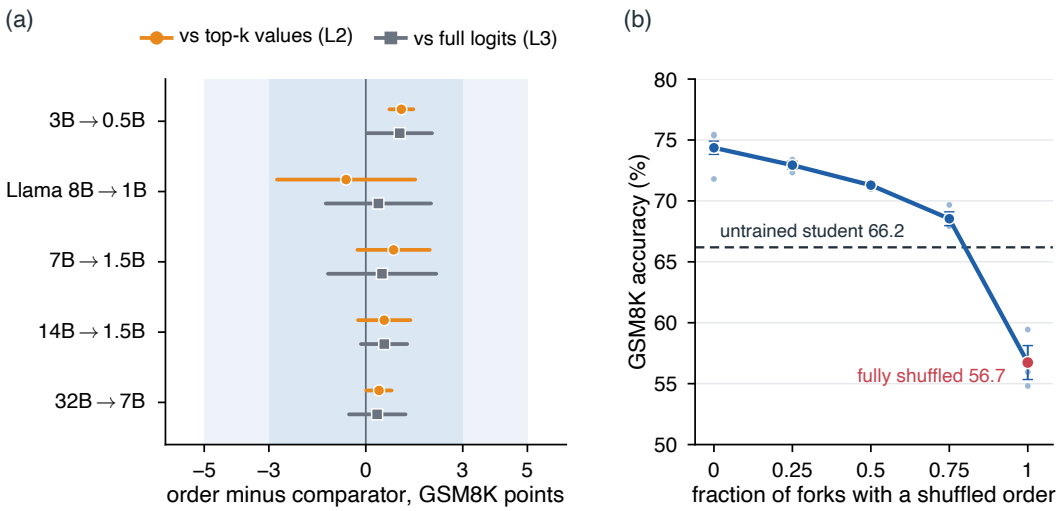


Figure 4: The order is enough and it is causal. (a) Paired difference to value-based OPD with 90% confidence intervals, against equivalence margins of ±3 (dark) and ±5 (light) GSM8K points. (b) Shuffling the order at a growing fraction of forking tokens on the 7B→1.5B pair, with seeds as light dots and mean±SEM, six seeds without shuffling and three otherwise.

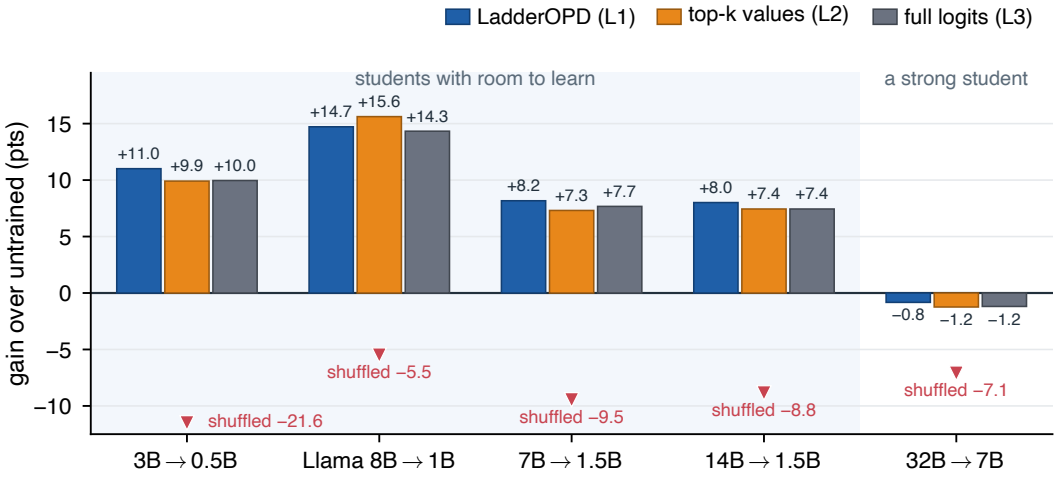


Figure 5: Gain over the untrained student on GSM8K in all five settings. The order tracks probabilities and logits whether OPD adds accuracy or not. Red triangles mark the shuffled order, drawn above the lower edge with its value in the label when off the scale.

from 74.4 through 72.9, 71.3 and 68.5 to 56.7 as the fraction grows from 0 to 1, so the damage grows with the share of the ranking that is wrong.

6.3 The ranking beats sparser signals

The top-1 token is the teacher's order truncated to one entry. It recovers most of the gain, at 72.6, 48.0 and 73.1 on the three pairs, but trails LadderOPD by 1.7, 3.9 and 1.1 points, significantly so on the 3B and 14B pairs ($p < .001$ and $p = .004$). Distillers that see only teacher text do worse. SeqKD on verified teacher solutions trails LadderOPD by 3.7 to 8.9 points and OmniOPD by 9.7 to 14.9 points on shared seeds. GAD stays below SeqKD on all three pairs, at 68.5, 40.9 and 68.1 against 70.7, 42.9 and 69.3.

Table 3: Ablations on 7B→1.5B, GSM8K. Variants use seeds 42, 123 and 456, on which LadderOPD averages 74.9, and $\tau = 1$ uses six seeds and $\tau = 2$ one.

Variant	GSM8K
LadderOPD	74.4±1.3
With GRPO reward	74.9±0.8
$\tau = 0.25$	75.5±0.7
$\tau = 1$	71.1±0.7
$\tau = 2$	65.0
$k = 10$	74.0±0.4
$k = 5$	73.6±1.5
Coverage only	72.6±0.7
Margin loss	72.0±0.7

Table 4: The fork gate across teachers, GSM8K on seeds 42, 123 and 456, six seeds on 3B.

Teacher	3B	7B	14B	32B
Gated (default)	**51.9**	**74.9**	**73.9**	**90.8**
No gate	50.9	74.5	73.7	90.4
Full logits (L3)	50.8	73.7	73.6	90.5

Table 5: Second family and code. Llama-3.1-8B→3.2-1B on GSM8K and Qwen 7B→1.5B on MBPP, three seeds per cell and six for Llama L1 and L3.

Method	Llama	MBPP
Untrained student	38.0	41.6
Top-1 CE (L0)	51.8±1.6	45.1±1.7
LadderOPD (L1)	52.7±0.9	46.9±0.3
Top-k values (L2)	**53.6**±0.6	**47.3**±1.1
Full logits (L3)	52.3±1.2	47.1±1.1
Shuffled order	32.5±4.1	26.4±2.4

Table 6: A 32B teacher. Qwen2.5-32B→7B, three seeds per row (Appendix H).

Method	GSM8K	MATH
Untrained student	91.7	47.4
LadderOPD (L1)	**90.8**±0.3	51.4±0.5
LadderOPD, no gate	90.4±0.6	**52.1**±0.8
Top-k values (L2)	90.4±0.2	50.2±0.4
Full logits (L3)	90.5±0.2	49.8±0.9
Shuffled order	84.6±1.3	46.1±2.0

6.4 What the recipe needs

Table 3 isolates each part on the 7B pair. The rank-tempered KL adds what coverage cannot, since coverage alone reaches 72.6 and a margin loss on adjacent ranks 72.0, against 74.9 for LadderOPD on the same seeds. Accuracy is highest at $\tau = 0.25$, where rank one holds 98% of the target, and falls steadily to 71.1 and 65.0 as τ grows to 1 and 2 (Appendix F). The support can be small. On shared seeds $k = 10$ and $k = 5$ lose 0.9 and 1.3 points, and $k = 10$ fits the caps of commercial interfaces. A verifiable-reward GRPO term (Shao et al., 2024) composes with the loss at 74.9.

Gating the order term to forks halves the tokens at which the teacher's ranking is revealed, and it costs no accuracy. Applying the order at every token lowers the 3B pair by 1.0 point, with $p = .02$ over six paired seeds, and moves the 7B, 14B and 32B pairs by at most half a point (Table 4).

6.5 Beyond the three pairs

A second family. Llama-3.2-1B-Instruct starts at 38.0 and gains 13.8 to 15.6 points under every access level from top-1 to full logits (Table 5). The order sits within a point of probabilities and logits and is TOST-equivalent to both at ±3, while the shuffled order drops the student to 32.5.

Code. On MBPP the untrained 1.5B student solves 41.6% of the test problems. Order, probabilities and logits lift it to 46.9, 47.3 and 47.1, and the order is TOST-equivalent to both at ±3. Top-1 reaches 45.1, and the shuffled order drops the student to 26.4, 15 points below where it started.

Contamination and transfer. GSM-Symbolic rewrites GSM8K items from templates, which controls for memorized problems (Mirzadeh et al., 2025). On 1,000 of its items the untrained 1.5B student scores 59.5, and order, probabilities and logits lift it to 68.9, 69.2 and 69.9, gains at least as large as on GSM8K, while a matched shuffled order falls to 51.5 (Table 19). On held-out SVAMP, ASDiv and GSM-Hard splits (Patel et al., 2021; Miao et al., 2020; Gao et al., 2023), the three evaluated rungs match or beat the untrained student for one seed, and top-1 scores highest on two sets (Table 20).

A 32B teacher. With a 32B teacher, more than twice the largest teacher of the main pairs, and a 7B student, the order stays within half a point of probabilities and logits, 90.8 against 90.4 and 90.5 and

TOST-equivalent to both at ± 3, while a shuffled control trained on the same schedule falls to 84.6 (Table 6). The untrained student already scores 91.7 on GSM8K and no rung improves on that within our 40 updates, but every rung adds 2.4 to 4.0 points on MATH500.

7 Discussion

Why the rank can replace the values. At forks inside the teacher's top-k the teacher's values track $g^\star$ with a Spearman correlation of 0.66 to 0.70 and its order with 0.61 to 0.64. The students trained on the two signals nonetheless end within 1.1 GSM8K points of each other in every paired comparison, and within 1.6 points on MATH500, where values lead with the 14B teacher ($p = .04$) and the order leads with the 32B teacher ($p \leq .02$). The extra fidelity of the values does not turn into accuracy, much as students match a teacher's full predictive distribution only loosely in general (Stanton et al., 2021), which suggests that what OPD uses at a fork is which candidates to prefer and in what order, far more than the exact margins between them. The rank-tempered target encodes exactly this preference, and Equation 6 shows that the teacher enters each update only through the position of a candidate in its list. The same equation explains why a wrong order is so damaging, since a shuffled list pushes probability toward candidates the teacher ranks low.

What in the ranking matters. With $\tau = 0.5$ the rank target puts 86% of its mass on the teacher's first candidate, and a sharper target with $\tau = 0.25$, 98% on rank one, does at least as well. The value of the ranked list thus lies mostly in the teacher's first choice and in the set it comes from, which the coverage term uses at every token, and less in the order of the remaining candidates. Top-1 cross-entropy, which lacks the set, trails LadderOPD by 1.1 to 3.9 points on four pairs.

Richer access does not pay off on GSM8K. Full logits never beat the order on GSM8K in our settings. Their mean accuracy is 0.4 to 1.1 points lower with every Qwen teacher from 3B to 32B and 0.4 points lower with the Llama teacher, and applying the order at every token instead of at forks does not help either. On MATH500 and MBPP the value-based rungs lead in seven of ten comparisons, by at most 1.6 points, and trail mainly with the 32B teacher. Up to a 32B teacher, whose 7B student is near its GSM8K ceiling, the ranking at forks carries what OPD uses from the teacher's distribution.

Which rung to ask for. With only a ranked list, LadderOPD matches value-based OPD at $k = 20$, and $k = 10$ costs under a point while fitting the caps of commercial interfaces. Black-box sampling tracks $g^\star$ slightly better than the order, at 0.64 to 0.67 against 0.61 to 0.64, but needs sixteen teacher queries per token where the ranked list needs one scoring pass. The fork gate halves the tokens at which the ranking must be revealed at no cost in accuracy, and ranked IDs reveal strictly less than log-probabilities while keeping the per-token density of OPD. For the designer of an interface, the same result says that an endpoint which scores a supplied sequence and returns a ranked list with k of at least ten already supports on-policy distillation without exposing a single probability.

Limitations. All students have 7B parameters or fewer and every training task has a verifiable answer. On open-ended instructions the order trails value-based OPD (Appendix J), and our attempts to carry it across tokenizers fell below the untrained student (Appendix I). The interface we assume scores a sequence supplied by the student, which not every commercial endpoint offers. With the 32B teacher the 7B student starts at 91.7 on GSM8K, so that setting tests equivalence more than gain. Absolute accuracies are specific to our answer format and 512-token budget (Appendix K). OmniOPD has no public code, so we compare against a simplified reimplementation.

8 Conclusion

On-policy distillation does not need the teacher's probabilities. Measured against an open teacher's exact signal, the part that matters sits at the tokens where the student hesitates and is carried by the order of the teacher's top candidates. A distiller built on that order matches distillation with probabilities or logits within three points across five settings, and corrupting the order degrades the student monotonically as the wrong share grows. For verifiable reasoning with a shared tokenizer, a teacher that scores the student's own sequence and returns a ranked list of ten or more candidates is therefore already sufficient for on-policy distillation. GroundKL needs only an open reference teacher and a few hundred rollouts, so it can be rerun for a new task before any training. Carrying the order signal to teachers whose tokenizer splits text differently from the student's is the natural next step.

Reproducibility Statement

All training and evaluation code, configurations and the per-run result file behind every number in this paper are included in the supplementary material, together with a table that maps each reported cell to its command and result file. Section 6 and Appendix C give every hyperparameter, Appendix E reports the paired tests behind every comparison in the text, and Appendix K describes the evaluation harness. All models and datasets we use are publicly available.

Ethics Statement

This work studies how much information a teacher model must reveal for distillation. A method that learns from ranked token lists lowers the barrier to distilling from closed models, which can matter for the terms of service of commercial interfaces. We use only open-weight teachers in all experiments, and users of the method should respect the licenses and usage policies of the models they distill from. The ranked interface reveals strictly less than the log-probabilities that such interfaces already expose.

AI Use Statement

During the experiments, we used large language models (LLMs) as an auxiliary tool for experiment monitoring and management. Specifically, the LLMs were used to monitor experiment logs and runtime information, identify potential execution issues or anomalies, and assist in reporting the status of ongoing experiments. The LLMs did not determine the research questions, experimental methodology, hyperparameter settings, or final experimental conclusions. All experimental configurations, result verification, analysis, and scientific conclusions were determined and validated by the authors. We take full responsibility for the final content and results of this work.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, et al. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *ICLR*, 2024.

Anshumann, Mohd Abbas Zaidi, Akhil Kedia, et al. Sparse Logit Sampling: Accelerating Knowledge Distillation in LLMs. In *ACL*, 2025.

Jacob Austin, Augustus Odena, Maxwell Nye, et al. Program Synthesis with Large Language Models. *arXiv preprint arXiv:2108.07732*, 2021.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In *NeurIPS*, 2015.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards Cross-Tokenizer Distillation: the Universal Logit Distillation Loss for LLMs. *TMLR*, 2025.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, et al. Stealing Part of a Production Language Model. In *ICML*, 2024.

Jang Hyun Cho and Bharath Hariharan. On the Efficacy of Knowledge Distillation. In *ICCV*, 2019.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025.

Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. Logits of API-Protected LLMs Leak Proprietary Information. In *COLM*, 2024.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, et al. Revisiting On-Policy Distillation: Empirical Failure Modes and Simple Fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Luyu Gao, Aman Madaan, Shuyan Zhou, et al. PAL: Program-aided Language Models. In *ICML*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *ICLR*, 2024.

Yuxian Gu, Hao Zhou, Fandong Meng, et al. MiniPLM: Knowledge Distillation for Pre-Training Language Models. In *ICLR*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. Measuring Mathematical Problem Solving With the MATH Dataset. In *NeurIPS Datasets and Benchmarks*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, et al. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In *Findings of ACL*, 2023.

Tao Huang, Shan You, Fei Wang, et al. Knowledge Distillation from A Stronger Teacher. In *NeurIPS*, 2022.

Yoon Kim and Alexander M. Rush. Sequence-Level Knowledge Distillation. In *EMNLP*, 2016.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards Streamlined Distillation for Large Language Models. In *ICML*, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, et al. DistiLLM-2: A Contrastive Approach Boosts the Distillation of LLMs. In *ICML*, 2025.

Daniël Lakens. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.

Hunter Lightman, Vineet Kosaraju, Yura Burda, et al. Let's Verify Step by Step. In *ICLR*, 2024.

Xinyu Liu, Kechen Jiao, Chunyang Xiao, et al. Teacher-Guided Policy Optimization for On-Policy Reasoning Distillation under Large Policy Divergence. *arXiv preprint arXiv:2605.13230*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

Kevin Lu and Thinking Machines Lab. On-Policy Distillation. Thinking Machines Lab: Connectionism, doi:10.64434/tml.20251026, 2025.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers. In *ACL*, 2020.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, et al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. In *ICLR*, 2025.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, et al. Improved Knowledge Distillation via Teacher Assistant. In *AAAI*, 2020.

John X. Morris, Wenting Zhao, Justin T. Chiu, et al. Language Model Inversion. In *ICLR*, 2024.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP Models Really Able to Solve Simple Math Word Problems? In *NAACL*, 2021.

Qwen, An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence Level Training with Recurrent Neural Networks. In *ICLR*, 2016.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *AISTATS*, 2011.

Donald J. Schuirmann. A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, et al. Does Knowledge Distillation Really Work? In *NeurIPS*, 2021.

Jiaxi Tang and Ke Wang. Ranking Distillation: Learning Compact Ranking Models With High Performance for Recommender System. In *KDD*, 2018.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, et al. Stanford Alpaca: An Instruction-following LLaMA Model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Shenzhi Wang, Le Yu, Chang Gao, et al. Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning. In *NeurIPS*, 2025.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-Divergence Minimization for Sequence-Level Knowledge Distillation. In *ACL*, 2023.

Wenda Xu, Rujun Han, Zifeng Wang, et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. In *ICLR*, 2025.

Xiaohan Xu, Ming Li, Chongyang Tao, et al. A Survey on Knowledge Distillation of Large Language Models. *arXiv preprint arXiv:2402.13116*, 2024.

Yuanda Xu, Hejian Sang, Zhengze Zhou, et al. TIP: Token Importance in On-Policy Distillation. *arXiv preprint arXiv:2604.14084*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, et al. OPRD: On-Policy Representation Distillation. *arXiv preprint arXiv:2606.06021*, 2026.

Tianzhu Ye, Li Dong, Zewen Chi, et al. Black-Box On-Policy Distillation of Large Language Models. *arXiv preprint arXiv:2511.10643*, 2025.

Rongzhi Zhang, Jiaming Shen, Tianqi Liu, et al. PLaD: Preference-based Large Language Model Distillation with Pseudo-Preference Pairs. In *Findings of ACL*, 2024a.

Songming Zhang, Xue Zhang, Zengkui Sun, et al. Dual-Space Knowledge Distillation for Large Language Models. In *EMNLP*, 2024b.

Borui Zhao, Quan Cui, Renjie Song, et al. Decoupled Knowledge Distillation. In *CVPR*, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS*, 2023.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, et al. OmniOPD: Logit-Free On-Policy Distillation via Speculative Verification. *arXiv preprint arXiv:2606.01476*, 2026.

APPENDIX

The appendix follows the order of the main text. Appendix A gives the full GroundKL measurements, Appendix B derives the properties of the loss, Appendix C covers the implementation and its cost, Appendix D the baselines and Appendix E every paired test. Appendices F to J report the remaining experiments, Appendix K describes the evaluation harness, Appendix L lists every model and dataset and Appendix M every seed. Every number traces to a per-run result file in the supplementary material, which also holds the scripts that compute each table.

A GROUNDKL IN FULL

Protocol. For each teacher we sample 200 GSM8K training problems, draw two completions per problem from the student at temperature 0.8, and score every completion token under the teacher with a single forward pass. Each completion is cut at its first end-of-sequence token. This gives 81,212 tokens for the 1.5B student, scored by the 7B and the 14B teacher, and 87,264 tokens for the 0.5B student, scored by the 3B teacher. For every token we record the exact $g_t^\star$ of Equation 2, the rank of y_t in the teacher's top-20, the teacher's argmax, the teacher's renormalized top-20 log-probabilities, and the empirical frequency of y_t among 16 teacher samples drawn at the same state. Forking tokens are the 34% of tokens with the highest student entropy, which corresponds to a threshold of 0.247 nats for the 1.5B student and 0.508 nats for the 0.5B student.

Where the signal lives. Table 7 splits $|g^\star|$ by stratum. Forking tokens are a third of all tokens and hold between 84% and 90% of the signal. Tokens outside the teacher's top-20 make up less than one percent of all tokens but hold 6% to 16% of it, and not one of them has a positive $g^\star$. Most of them are forking tokens, 97% for the 3B, 98% for the 7B and 73% for the 14B teacher. Concentration is robust to the entropy threshold. For the 7B teacher the most uncertain 10%, 20%, 34% and 50% of tokens hold 38.4%, 64.6%, 89.9% and 98.6% of $|g^\star|$.

Table 7: Where $|g^\star|$ sits, by stratum and teacher.

	3B→0.5B	7B→1.5B	14B→1.5B
Tokens scored	87,264	81,212	81,212
Forking share of tokens	34.0%	34.0%	34.0%
Forking share of $\|g^\star\|$	83.9%	89.9%	83.6%
Off-support share of tokens	0.82%	0.28%	0.53%
Off-support share of $\|g^\star\|$	13.0%	5.8%	15.6%
Mean $g^\star$ on off-support tokens	−6.81	−6.86	−11.48
Mean $g^\star$ on all tokens	−0.30	−0.23	−0.28

What each signal recovers. Table 8 reports the Spearman correlation between each observable and $g^\star$ in every stratum for the 7B teacher. The L2 signal, the teacher's renormalized log-probability minus the student's, equals $g^\star$ up to the log of the teacher's top-20 mass, so its correlation of 0.999 checks the bookkeeping rather than the information in the values. The like-for-like comparison with the other rows is the teacher's value alone. At forks it reaches 0.669 against 0.625 for the order, and at low-entropy tokens, where $g^\star$ varies mostly through the student's own probability, it falls to -0.153 while the order keeps 0.156 on the same tokens. Inside the support the order is strongest at the GroundKL forks, at 0.625 against 0.428 over all in-support tokens. The training gate at the median entropy also admits the 50th to 66th percentile band, where the order recovers only 0.23 to 0.31 (Table 10). Outside the support every token has the same clipped rank up to ties, and only the coverage term acts. Membership is constant inside the support and therefore has no correlation there by construction.

How certain the correlations are. Table 9 resamples whole rollouts. For every teacher the order leads top-1 agreement by 0.04 to 0.05 and trails the sampling ceiling by 0.02 to 0.03, and none of the six intervals for these differences includes zero. The order thus recovers significantly more than the top-1 token and slightly less than sixteen black-box samples.

Where in the entropy range the order works. Table 10 and Figure 6a split tokens into bands of student entropy. Among the 10% most uncertain tokens, which hold 34% to 38% of $|g^\star|$, the order reaches 0.72 to 0.73, close to the sampling ceiling of 0.72 to 0.76 and well above top-1 agreement at

Table 8: Spearman correlation with $g^\star$, 7B→1.5B. In-support columns exclude tokens outside the teacher's top-20, where membership is constant. The two value rows exist only inside the support, so all their entries are computed on in-support tokens.

Signal	All	Forking	Low-entropy	In support	Forking, in support
Order (L1)	0.435	0.635	0.157	0.428	0.625
Top-1 agreement (L0)	0.425	0.593	0.158	0.419	0.587
Membership	0.090	0.148	0.015	constant	constant
Sampling, 16 queries	0.351	0.652	0.296	0.345	0.645
Top-k value alone	0.160	0.669	−0.153	0.160	0.669
Value minus student (L2 signal)	0.999	1.000	1.000	0.999	1.000

Table 9: Bootstrap intervals for the recovery of $g^\star$ at forking tokens inside the support, with 95% intervals from 1,000 resamples of the 400 rollouts.

Quantity	3B teacher	7B teacher	14B teacher
Order (L1)	0.635 [0.625, 0.645]	0.625 [0.616, 0.635]	0.608 [0.597, 0.618]
Top-1 agreement (L0)	0.582 [0.572, 0.592]	0.587 [0.577, 0.597]	0.566 [0.556, 0.577]
Sampling, 16 queries	0.667 [0.658, 0.676]	0.645 [0.635, 0.654]	0.640 [0.630, 0.649]
Sampling minus order	0.031 [0.026, 0.036]	0.019 [0.014, 0.025]	0.032 [0.026, 0.037]
Order minus top-1	0.053 [0.049, 0.057]	0.038 [0.034, 0.042]	0.042 [0.038, 0.045]

0.59 to 0.63. Toward low entropy the order and top-1 agreement converge, since the sampled token is then almost always the teacher's first choice, and in the lower half of the entropy range both fall to 0.09 to 0.12. That half holds only 1.4% to 3.5% of the signal.

Table 10: Recovery by entropy band, Spearman correlation with $g^\star$ inside the support, with bands given as percentiles of student entropy.

Teacher	Entropy band	Share of $\lvert g^\star \rvert$	Order (L1)	Top-1 (L0)	Sampling
3B	90–100	34.2%	0.721	0.590	0.723
	80–90	24.6%	0.653	0.594	0.704
	66–80	25.1%	0.568	0.553	0.671
	50–66	12.6%	0.310	0.311	0.648
	0–50	3.5%	0.117	0.119	0.339
7B	90–100	38.4%	0.733	0.629	0.762
	80–90	26.2%	0.656	0.637	0.723
	66–80	25.3%	0.460	0.450	0.615
	50–66	8.7%	0.275	0.275	0.548
	0–50	1.4%	0.090	0.091	0.182
14B	90–100	35.3%	0.724	0.619	0.759
	80–90	23.5%	0.630	0.605	0.703
	66–80	24.8%	0.428	0.419	0.605
	50–66	12.9%	0.228	0.230	0.588
	0–50	3.5%	0.102	0.102	0.229

Support size. Figure 6b repeats the measurement with the teacher's top 5 and top 10 in place of the top 20. The order's recovery at forks falls from 0.61 to 0.64 at $k = 20$ to 0.59 to 0.61 at $k = 10$ and 0.55 to 0.58 at $k = 5$, and the share of $|g^\star|$ outside the support grows from 6% to 16% to 12% to 23% and 24% to 35%. Both trends agree with the small accuracy cost of $k = 10$ and $k = 5$ in Table 3.

Position in the completion. The first fifth of each completion holds 23% to 27% of $|g^\star|$ and has the highest rate of forking tokens, 46% (Figure 6c). With the 14B teacher the last fifth holds another 26%, and 1.4% of its tokens fall outside the teacher's top-20, about four times the rate in the middle of the completion.

Correct and incorrect rollouts. Incorrect rollouts carry more signal per token, with a mean $|g^\star|$ higher by 0.08 for every teacher, and 1.6 to 2.9 times the share of off-support tokens (Table 11). The order recovers the signal about equally well in both, with gaps of 0.001 to 0.032, and only the largest, for the 3B teacher, is significant under a rollout bootstrap, with a 95% interval from 0.009 to 0.054.

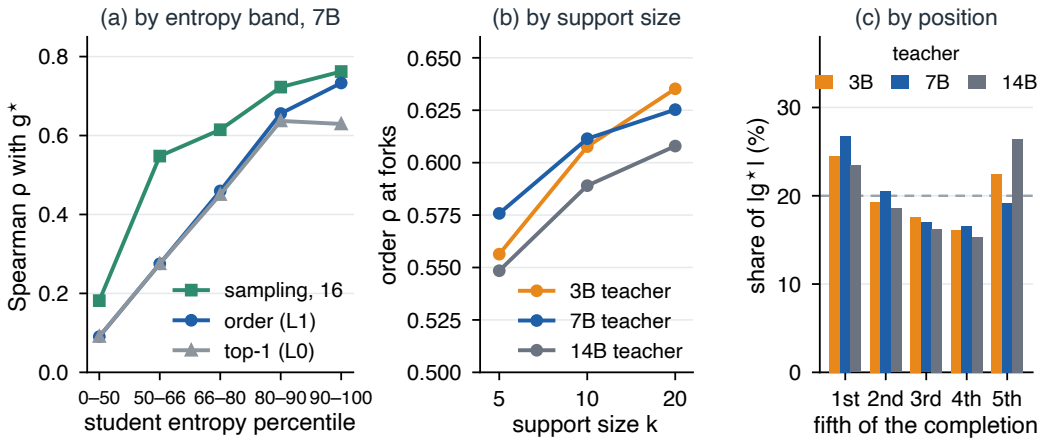


Figure 6: GroundKL in more detail. (a) Recovery of $g^\star$ inside the support by band of student entropy, 7B teacher. (b) Recovery by the order at forking tokens when the teacher returns its top 5, 10 or 20. (c) Share of $|g^\star|$ in each fifth of the completion, where the dashed line marks an even spread.

Table 11: Correct and incorrect rollouts. Mean $|g^\star|$ per token, the forking share of $|g^\star|$, the share of off-support tokens and the recovery at forks inside the support.

Teacher	Rollouts	Number	Mean $\lvert g^\star\rvert$	Forking share	Off-support	Order (L1)	Sampling
3B	Correct	110	0.369	81.4%	0.48%	0.661	0.685
	Incorrect	290	0.446	84.6%	0.92%	0.629	0.661
7B	Correct	232	0.299	88.1%	0.15%	0.633	0.648
	Incorrect	168	0.375	91.4%	0.43%	0.618	0.641
14B	Correct	232	0.354	79.7%	0.41%	0.609	0.637
	Incorrect	168	0.432	87.0%	0.66%	0.607	0.642

Tokens outside the support. Table 12 describes the rare tokens outside the teacher's top-20. Their median teacher rank is 30 to 49, and 9% to 27% of them lie beyond rank 100. The student gives them a median probability of 0.035 to 0.205, while the teacher's median is at most 1.5×10^{-4}, which explains their large negative $g^\star$. They sit at uncertain positions, with a mean student entropy of 1.20 to 2.13 nats against 0.33 to 0.50 over all tokens.

Table 12: Tokens outside the teacher's top-20.

	3B teacher	7B teacher	14B teacher
Tokens	713	231	433
Median teacher rank	34	30	49
Share beyond rank 100	14.3%	8.7%	27.3%
Median student probability	0.042	0.035	0.205
Median teacher probability	1.5×10^{-4}	8.8×10^{-5}	1.9×10^{-6}
Share at forking tokens	97.1%	98.3%	72.7%
Mean student entropy (nats)	2.13	1.82	1.20
Mean student entropy, all tokens	0.50	0.33	0.33

Forks chosen by teacher entropy. Taking the 34% of tokens with the highest teacher entropy instead captures 68% to 78% of $|g^\star|$ rather than 84% to 90%, and the order recovers 0.55 to 0.59 there rather than 0.61 to 0.64. The two selections share 83% to 87% of their tokens. Student entropy is the better selector, and it needs no access to the teacher.

B Properties of the Loss

Write z_u for the student's logit of token u at a fixed position, z_r for the logit of the teacher's r-th candidate v_r, and $\tilde{\pi}_r = e^{z_r} / \sum_{j=1}^{k} e^{z_j}$ for the student's distribution renormalized on S_t.

Gradient of the order term. The order term is $\ell_{\text{ord}} = \sum_j \tilde{\pi}_j\big(\log \tilde{\pi}_j - \log q_\tau(j)\big)$, and the renormalized softmax has the Jacobian $\partial\tilde{\pi}_j/\partial z_r = \tilde{\pi}_j\big(\mathbf{1}[j=r] - \tilde{\pi}_r\big)$. Hence

$$\frac{\partial \ell_{\text{ord}}}{\partial z_r} = \sum_{j=1}^{k} \tilde{\pi}_j\big(\mathbf{1}[j=r] - \tilde{\pi}_r\big)\big(\log \tilde{\pi}_j - \log q_\tau(j) + 1\big) = \tilde{\pi}_r\Big[\log \tilde{\pi}_r - \log q_\tau(r) - \mathrm{KL}\big(\tilde{\pi} \,\|\, q_\tau\big)\Big],$$

where the constant 1 drops out because $\sum_j \tilde{\pi}_j(\mathbf{1}[j=r] - \tilde{\pi}_r) = 0$. This is Equation 6. The order term does not depend on logits outside S_t, since $\tilde{\pi}$ is unchanged when they move.

Expected-rank form. Inserting $\log q_\tau(r) = -r/\tau - \log Z_\tau$ gives

$$\ell_{\text{ord}} = \frac{1}{\tau}\,\mathbb{E}_{r\sim\tilde{\pi}}[r] \;-\; H(\tilde{\pi}) \;+\; \log Z_\tau.$$

The order term therefore penalizes the expected teacher rank of the student's renormalized choice, scaled by $1/\tau$, minus an entropy bonus, and it vanishes exactly when $\tilde{\pi} = q_\tau$. The teacher's ranks enter linearly and nothing else about the teacher enters at all. As $\tau \to 0$, $\tau \log Z_\tau \to -1$ and $\tau H(\tilde{\pi}) \to 0$, so $\tau\,\ell_{\text{ord}} \to \mathbb{E}_{\tilde{\pi}}[r] - 1$. A sharp target thus still rewards moving mass from rank ten to rank two while the mass on rank one stays fixed, and cross-entropy on the top-1 token does not see such a change.

Gradient of the coverage term. With $M_t = \sum_{v\in S_t} \pi_S(v \mid s_t)$ and $\ell_{\text{cov}} = -\log \sum_{v\in S_t} e^{z_v} + \log \sum_u e^{z_u}$,

$$\frac{\partial \ell_{\text{cov}}}{\partial z_u} = \pi_S(u \mid s_t) - \mathbf{1}[u \in S_t]\,\frac{\pi_S(u \mid s_t)}{M_t} = \begin{cases} -\pi_S(u \mid s_t)\,(1 - M_t)/M_t, & u \in S_t, \\ \pi_S(u \mid s_t), & u \notin S_t. \end{cases}$$

Coverage raises every candidate in the set in proportion to its current probability, lowers every token outside it, and vanishes once $M_t = 1$. It is invariant to any permutation of S_t and so uses the unordered set only, so the ranking reaches the student only through ℓ_{ord} at forking tokens.

The temper. Table 13 lists the shape of q_τ at $k = 20$ for every temper we trained with, together with GSM8K accuracy on 7B→1.5B. Accuracy falls steadily as the target spreads over more ranks. On seed 42, the one seed trained at every temper, it drops from 76.2 with 1.1 effective ranks to 65.0 with 5.5, and the means over more seeds follow the same order.

Table 13: Shape of the rank target at $k = 20$. The effective number of ranks is $\exp H(q_\tau)$. GSM8K on 7B→1.5B is given for seed 42 and as the mean over the listed seeds.

τ	$q_\tau(1)$	Mass on ranks 1 to 3	Effective ranks	Seed 42	Mean	Seeds
0.25	0.982	1.000	1.10	76.2	75.5	3
0.35	0.943	1.000	1.26	75.1	75.1	1
0.5	0.865	0.998	1.58	74.6	74.4	6
0.75	0.736	0.982	2.19	73.8	73.8	1
1	0.632	0.950	2.83	71.0	71.1	6
1.5	0.487	0.865	4.15	67.7	67.7	1
2	0.393	0.777	5.49	65.0	65.0	1

C Implementation

Table 14 lists every training setting. All access levels in a column share these settings and differ only in the loss target. The loss follows Algorithm 1. The coverage term is averaged over all completion tokens in the batch, the order term over forking tokens, and the two are added with unit weights. The fork threshold is the median of the student's detached entropy over the completion tokens of the current batch, recomputed at every step. The optimizer is AdamW with no weight decay, a linear warmup followed by a constant learning rate, and gradient clipping at norm 1.0. The Llama pair uses the settings of the main pairs with its own models. The MBPP runs sample from the 374 problems of the MBPP training split, 187 to 206 distinct problems per run, with a coding system prompt, keep the settings of the main pairs, and score the 500 test problems for three seeds. The 3B→0.5B and Llama runs use two NVIDIA H100 80GB GPUs, one for the student and one for the teacher, and PyTorch

Table 14: Training settings. A micro-step processes one batch of rollouts.

Setting	Main pairs	32B→7B
Training pool per seed	4,000 GSM8K problems	4,000 GSM8K problems
Distinct problems drawn per run	285 to 294	157 to 158
Prompts per micro-step	2	1
Rollouts per prompt	4	4
Gradient accumulation	1	4
Micro-steps	150	160
Learning rate	1×10^{-5}	7×10^{-6}
Warmup steps	10	5
Rollout temperature, top-p and length	0.8, 0.95, 256	0.8, 0.95, 256
Top-k and temper τ	20 and 0.5	20 and 0.5
Fork quantile	0.5	0.5
Seeds for LadderOPD	6	3
GPUs	one A800 or two H100	two A800 80GB

2.6. All other runs use NVIDIA A800 80GB GPUs and PyTorch 2.12, with student and teacher on one GPU for the main pairs and on two GPUs for the 32B pair. All runs use Transformers 5.12.1 with bfloat16 weights.

Cost. Table 15 compares what each rung sends per completion token and the measured training time on 7B→1.5B, where a rollout has about 220 completion tokens. All four rungs take one teacher forward pass per rollout, so their training times agree within a minute, and they differ only in what leaves the teacher, from 80 bytes per token for the ranked list to 304 kilobytes for the full logits in bfloat16. The ranked list carries at most 344 bits per token, of which 61 bits encode the order inside the set and the rest its membership. The text-level baselines take longer. Generating and filtering the SeqKD demonstrations takes 50 minutes before 3 minutes of fine-tuning, GAD takes 25 minutes and our OmniOPD, which generates teacher continuations for every chunk, 56 minutes. Two runs shared each GPU, so these times compare methods only roughly.

Table 15: Interface payload and training time on 7B→1.5B. Bytes assume 32-bit IDs and values for L0 to L2 and bfloat16 values for L3, with 223 completion tokens per rollout. Time is the mean and range over six seeds for 150 steps on an A800.

Rung	Returned per token	Bytes per token	Bytes per rollout	Training time (min)
Top-1 CE (L0)	one token ID	4	0.9 kB	28.7 (26.7 to 33.3)
LadderOPD (L1)	20 ranked token IDs	80	17.8 kB	28.7 (26.6 to 31.0)
Top-k values (L2)	20 IDs and their values	160	35.7 kB	28.0 (24.6 to 29.8)
Full logits (L3)	151,936 values	303,872	67.8 MB	27.8 (25.7 to 28.8)

D BASELINES

Access-ladder rungs. All rungs run in the same trainer. L0 applies cross-entropy on the teacher's argmax at every completion token, the per-token target of TGPO (Liu et al., 2026). L2 keeps the coverage term of LadderOPD and replaces the order term with the reverse KL between the student and the teacher, both renormalized on the teacher's top-20, following Revisiting-OPD (Fu et al., 2026). This KL is applied at every completion token in the main pairs and on MBPP, and only at forking tokens with the 32B teacher and on the Alpaca instructions of Appendix J. L3 is the full-vocabulary reverse KL of Equation 1.

Shuffled-order control. The control uses the LadderOPD loss with the rank target q_τ permuted at random, independently at each forking token, while the teacher's top-20 set, the coverage term, the fork gate and every hyperparameter stay unchanged. The graded variant of Figure 4b applies the permutation to a random fraction of forking tokens and keeps the true order elsewhere.

SeqKD. For 3,000 GSM8K training problems we sample two teacher solutions at temperature 0.7 with top-p 0.95 and keep the first one with a correct final answer. This keeps 2,731 demonstrations from the 7B teacher and 2,801 from the 14B teacher. The student is fine-tuned on these demonstrations for 150 steps with batch size 8 and two-step gradient accumulation at a learning rate of 10^{-5}.

GAD. We port the official release of GAD (Ye et al., 2025) to our trainer. The discriminator is trained with the loss $-\log\sigma(s_T - s_S)$ on teacher and student responses, the student is warmed up with SFT on teacher text, and GRPO (Shao et al., 2024) then optimizes the discriminator score as a reward.

OmniOPD. No code is public, so we reimplemented the chunk-level verification idea of Zhou et al. (2026). Each student rollout is split into two equal halves. For each half the teacher greedily continues the preceding prefix for 48 tokens, and the token-level ROUGE-L F1 between the half and that continuation, cut to the half's length when the half is shorter, is its score. The two scores are averaged into one reward per rollout, which GRPO optimizes with advantages normalized over the four rollouts of a prompt. The published method differs in several ways. It weights the log-likelihood of each audited chunk by a smoothed similarity to ten sampled teacher rollouts, measures similarity by edit distance, audits ten 50-token chunks chosen at high-entropy forks, and adds a Dirichlet prior and a KL anchor.

E Statistical Tests

Table 16 reports every paired comparison between LadderOPD and the value-based rungs, computed over shared seeds from the per-run result files. Table 17 gives the corresponding TOST results that Figure 4a displays. Every comparison uses all shared seeds, six for the main pairs, three or six for Llama and three for MBPP and the 32B teacher.

Table 16: Paired differences, LadderOPD minus comparator in GSM8K points, MBPP points in the last row, with 95% confidence intervals and two-sided paired t-test p-values.

	vs top-k values (L2)			vs full logits (L3)		
Setting	n	Difference	p	n	Difference	p
7B→1.5B	6	$+0.86$ [-0.57, 2.29]	.183	6	$+0.50$ [-1.63, 2.64]	.570
3B→0.5B	6	$+1.10$ [0.63, 1.57]	.002	6	$+1.05$ [-0.22, 2.32]	.087
14B→1.5B	6	$+0.57$ [-0.46, 1.60]	.215	6	$+0.57$ [-0.34, 1.48]	.170
Llama 8B→1B	3	-0.61 [-3.77, 2.55]	.493	6	$+0.39$ [-1.69, 2.47]	.649
32B→7B	3	$+0.41$ [-0.18, 0.99]	.095	3	$+0.35$ [-0.93, 1.64]	.359
MBPP, 7B→1.5B	3	-0.47 [-3.34, 2.40]	.556	3	-0.20 [-3.68, 3.28]	.828

Table 17: Equivalence tests. 90% confidence intervals of the paired difference and TOST p-values at margins of ±3 and ±5 GSM8K points. Equivalence holds when $p < .05$.

	vs top-k values (L2)			vs full logits (L3)		
Setting	90% CI	$p_{\pm3}$	$p_{\pm5}$	90% CI	$p_{\pm3}$	$p_{\pm5}$
7B→1.5B	[-0.26, 1.98]	**.006**	**<.001**	[-1.17, 2.18]	**.015**	**.001**
3B→0.5B	[0.73, 1.47]	**<.001**	**<.001**	[0.05, 2.05]	**.005**	**<.001**
14B→1.5B	[-0.24, 1.38]	**<.001**	**<.001**	[-0.15, 1.28]	**<.001**	**<.001**
Llama 8B→1B	[-2.75, 1.53]	**.041**	**.013**	[-1.24, 2.02]	**.012**	**.001**
32B→7B	[0.01, 0.80]	**.001**	**<.001**	[-0.52, 1.23]	**.006**	**.002**
MBPP, 7B→1.5B	[-2.41, 1.48]	**.031**	**.010**	[-2.56, 2.16]	**.037**	**.014**

Table 18 gives the remaining paired comparisons cited in the text. The fork gate is tested the same way. Over six shared seeds on the 3B pair the gated recipe beats the ungated one by $+0.99$ points, 95% CI [0.23, 1.75], $p = .020$, and over three shared seeds on the 7B, 14B and 32B pairs the gated recipe leads by 0.43, 0.20 and 0.41 points with $p > .05$.

F Sensitivity to the Temper and the Support Size

Figure 7 sweeps τ and k on the 7B→1.5B pair. On seed 42, the one seed trained at every temper, accuracy falls steadily as the target flattens, from 76.2 at $\tau = 0.25$ through 74.6 at $\tau = 0.5$ to 71.0 at $\tau = 1$ and 65.0 at $\tau = 2$. On the shared seeds 42, 123 and 456, $k = 10$ and $k = 5$ lie 0.9 and 1.3 points below $k = 20$, and both fall inside the caps of five to twenty candidates that commercial interfaces allow. On seed 42 accuracy rises steadily with the support, from 71.9 at $k = 5$ through 73.6 and 74.6 to 75.6 at $k = 40$, the one k run with a single seed.

Table 18: Further paired comparisons, LadderOPD minus comparator in points, GSM8K unless marked, with 95% confidence intervals and two-sided paired t-test p-values.

Comparison	Setting	n	Difference	p
Order vs top-1	7B→1.5B	6	+1.74 [−0.16, 3.65]	.065
	3B→0.5B	6	+3.86 [3.35, 4.36]	<.001
	14B→1.5B	6	+1.06 [0.51, 1.62]	.004
	Llama 8B→1B	3	+1.16 [−3.36, 5.68]	.384
Order vs SeqKD	7B→1.5B	6	+3.69 [2.03, 5.34]	.002
	3B→0.5B	6	+8.92 [8.04, 9.81]	<.001
	14B→1.5B	6	+4.90 [4.56, 5.24]	<.001
Order vs coverage only	7B→1.5B	3	+2.27 [0.59, 3.95]	.028
Gated vs ungated	3B→0.5B	6	+0.99 [0.23, 1.75]	.020
	7B→1.5B	3	+0.43 [−0.04, 0.90]	.060
	14B→1.5B	3	+0.20 [−2.07, 2.48]	.737
	32B→7B	3	+0.41 [−1.88, 2.69]	.523
MATH500, order vs values	7B→1.5B	6	−0.47 [−1.57, 0.63]	.325
	3B→0.5B	6	−0.27 [−2.06, 1.53]	.718
	14B→1.5B	6	−1.57 [−3.00, −0.13]	.037
	32B→7B	3	+1.20 [0.70, 1.70]	.009
MATH500, order vs logits	7B→1.5B	6	−0.30 [−1.76, 1.16]	.620
	3B→0.5B	6	+0.03 [−0.73, 0.80]	.915
	14B→1.5B	6	−1.23 [−3.33, 0.87]	.192
	32B→7B	3	+1.60 [0.61, 2.59]	.020

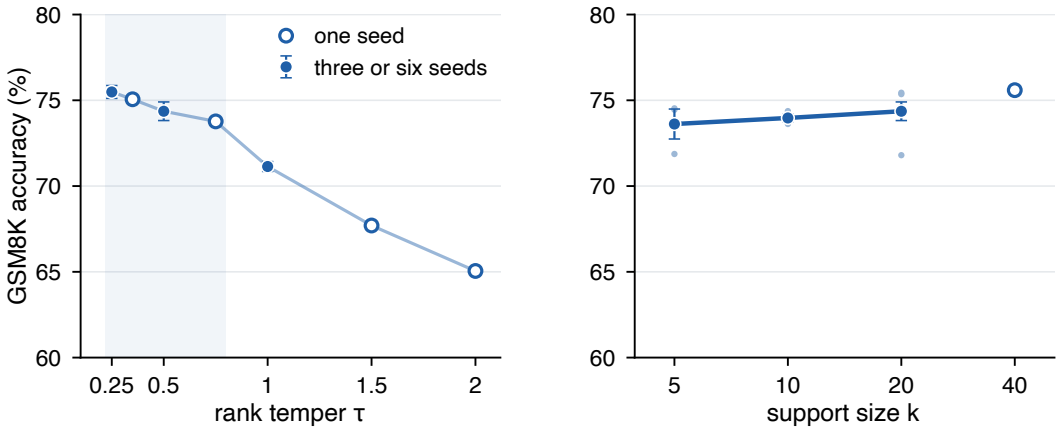


Figure 7: Sensitivity on 7B→1.5B. Filled points average three or six seeds with SEM bars, hollow points are single runs with seed 42, light dots in the right panel are individual seeds, and the shaded band marks τ from 0.25 to 0.75.

G Transfer and Contamination Control

Table 19 gives the full access ladder on GSM-Symbolic (Mirzadeh et al., 2025), evaluated on a fixed random subset of 1,000 items from its main configuration with the GSM8K harness. Every trained rung lifts the untrained student by 9 to 10 points and SeqKD by 5 points, while the matched shuffled order falls 8 points below the untrained student.

SVAMP, ASDiv and GSM-Hard (Patel et al., 2021; Miao et al., 2020; Gao et al., 2023) are scored on held-out splits of 60, 61 and 264 problems, the last fifth of each set after a shuffle with seed 0, using the seed-42 models of the L0, L1 and L2 rungs. At these sizes one problem moves accuracy by 1.7, 1.6 and 0.4 points, so gaps of a few points in Table 20 lie within noise.

Table 19: GSM-Symbolic, 7B→1.5B, 1,000 items, mean±std over three seeds and one seed for SeqKD.

Method	Accuracy
Untrained student	59.5
SeqKD	64.4
Top-1 CE (L0)	68.4±1.2
LadderOPD (L1)	68.9±1.1
Top-k values (L2)	69.2±0.5
Full logits (L3)	**69.9**±1.5
Shuffled order	51.5±1.3

Table 20: Transfer to held-out math sets, 7B→1.5B trained on GSM8K, seed 42.

Method	SVAMP	ASDiv	GSM-Hard
Untrained student	75.0	85.2	30.3
Top-1 CE (L0)	80.0	88.5	37.9
LadderOPD (L1)	75.0	88.5	34.8
Top-k values (L2)	78.3	90.2	35.6

H THE 32B TEACHER

The 7B student and the 32B teacher together need about 124GB of GPU memory, so they run on two GPUs, each micro-step processes a single prompt, and gradient accumulation over four micro-steps yields 40 optimizer updates over 160 micro-steps at a learning rate of 7×10^{-6} with 5 warmup steps. Every row of Table 6, the shuffled control included, uses this schedule and the seeds 42, 123 and 456, and no run diverged. Under our harness the untrained 7B student scores 91.7 on GSM8K and 47.4 on MATH500, so there is little GSM8K headroom. Every rung ends within 1.3 points below that start on GSM8K and 2.4 to 4.0 points above it on MATH500, 4.7 without the fork gate, and the shuffled order costs 7.1 points on GSM8K and 1.3 on MATH500.

I DISTILLATION ACROSS TOKENIZERS

We also tried to carry the order across tokenizers with OLMo-2-7B (**?**) as the teacher and Qwen2.5-1.5B-Instruct as the student. Their token IDs are not aligned, and OLMo-2 groups up to three digits per token where Qwen2.5 splits every digit, although 98.9% of OLMo-2's 100,278 token strings also occur in the Qwen2.5 vocabulary. Table 21 lists the results. The two mappings of the ranking that we tried, to first bytes and to four-character prefixes, do not match whole token strings. Mapping the teacher's ranking to first bytes drops the student to 17.7, since the first byte is usually a space, and four-character prefixes fare worse at 6.1. Text-level on-policy continuation, a DAgger variant (Ross et al., 2011), reaches 52.8, and 56.5 when only continuations with a correct answer are kept. All of these stay below the untrained student, while SeqKD on curated OLMo demonstrations lifts it to 70.1. An exact string match of the teacher's candidates, which the overlap would allow for most tokens, was not tried and is the direction we would pursue next.

Table 21: OLMo-2-7B→Qwen2.5-1.5B with non-aligned token IDs, GSM8K. Continuation rows average three seeds and the others use one.

Method	GSM8K
Untrained student	66.2
Order through first bytes	17.7
Order through four-character prefixes	6.1
On-policy continuation	52.8±4.4
On-policy continuation, correct answers only	56.5±1.5
SeqKD on curated demonstrations	70.1

J Open-Ended Instructions

We distill Qwen2.5-7B into Qwen2.5-1.5B on Alpaca instructions (Taori et al., 2023), where no verifiable reward exists, and judge pairs of responses with Qwen2.5-14B-Instruct (Zheng et al., 2023) on 120 instruction-only prompts drawn with a separate seed. Depending on the training seed, 8 to 15 of them lie in the 4,000-prompt pool a run samples from, and a reconstruction of the sampler finds 0 to 2 among the roughly 290 prompts a run actually draws. In this setting value-based OPD applies its KL at forking tokens only, and SeqKD fine-tunes with the schedule of Appendix D on greedy Qwen2.5-7B responses, kept without filtering, to one pool of 2,500 instruction-only prompts drawn with seed 42. That pool contains 4 of the judged prompts, and each SeqKD run draws 1 to 4 of them among roughly 1,550 distinct prompts. Each pair is judged in both orders to cancel position bias, a prompt is a win when the two verdicts together favor the candidate, and the win rate counts a tie as half a win. Table 22 lists the results. Here the order does not match the values. Against value-based OPD it wins 122, loses 197 and ties 281 of 600 comparisons over five seeds, a win rate of 43.8% with a 95% confidence interval over seeds from 39.9% to 47.6%, and the mirrored comparison over three seeds gives value-based OPD 57.1%. The order also trails SeqKD at 44.2%, while value-based OPD ties SeqKD at 50.0%. Both on-policy variants beat the untrained student, at 69.7% and 72.8%. We therefore restrict the claim that the order suffices to tasks with verifiable answers.

Table 22: Open-ended instructions, Qwen2.5-7B→Qwen2.5-1.5B on Alpaca, judged by Qwen2.5-14B-Instruct on 120 prompts per seed. Win rates are means over seeds with their standard deviation. The row of values against LadderOPD restates the first row's judgments on seeds 42, 123 and 456 from the other side.

Candidate	Opponent	Seeds	Win / loss / tie	Win rate (%)
LadderOPD (L1)	Values (L2)	5	122 / 197 / 281	43.8±3.1
LadderOPD (L1)	SeqKD	5	141 / 211 / 248	44.2±4.1
LadderOPD (L1)	Untrained	5	315 / 79 / 206	69.7±1.9
Values (L2)	LadderOPD (L1)	3	123 / 72 / 165	57.1±3.6
Values (L2)	SeqKD	3	90 / 90 / 180	50.0±3.6
Values (L2)	Untrained	3	199 / 35 / 126	72.8±3.1

K Evaluation Harness

All math benchmarks use the student's chat template with the system prompt below, greedy decoding and a budget of 512 new tokens.

```
You are a careful math problem solver.  Reason step by step, then
give the final answer on its own line in the form '#### <answer>'.
```

For GSM8K and its variants the answer is the last number in the text after the first #### marker, falling back in turn to an answer phrase, to the first `\boxed{}` expression and to the last number in the response, and it counts as correct when it equals the reference within 10^{-4}. MATH500 answers are taken from the last `\boxed{}` expression and compared numerically within 10^{-4}, or as strings otherwise, with the math-verify package used first when it is installed. MBPP responses are executed against the reference tests of each task.

Attention implementation. Every run that involves the 1.5B or 7B model trains and evaluates with PyTorch's scaled-dot-product attention (SDPA). With the eager implementation, bfloat16 inference corrupts Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct in the Transformers version we use. On a fixed random subset of 200 GSM8K test problems the 1.5B model scores 30.5 with eager attention and 64.0 with SDPA, and on the full test set the 7B model falls from 91.7 to 29.8. The 0.5B, 3B and 14B models change by at most one point on the same 200 problems, and Llama-3.2-1B scores 38.0 with eager attention in bfloat16 and 38.8 with SDPA in float16 on V100 GPUs. We therefore keep the original 3B→0.5B and Llama runs, which used eager attention, although the Llama-3.1-8B teacher was not re-checked. Untrained and trained rows of each pair are evaluated in the same environment.

What the eager failure looks like. Table 23 shows the step at which five of the 200 problems go wrong for the 1.5B model. Under eager attention in bfloat16 the model sets up the right operation with the right operands and then writes a wrong result. The same model with SDPA solves all five,

and so does eager attention in float32, which scores 62.0 on the 200 problems against 64.0 for SDPA and 30.5 for eager attention in bfloat16.

Table 23: Eager attention in bfloat16 corrupts arithmetic in Qwen2.5-1.5B-Instruct. Each row quotes the same step of the two greedy solutions to one GSM8K test problem.

Problem	Reference	Eager, bfloat16	SDPA, bfloat16
Bananas picked over three days	22	$4 + 6 + 12 = 12$	$4 + 6 + 12 = 22$
Toy soldiers sold at two prices	72	$30 + 42 = 38$	$30 + 42 = 72$
Stickers bought and given away	17	$45 - 28 = 7$	$45 - 28 = 17$
Average of two boot sizes	15	$5 + 25 = 29$	$5 + 25 = 30$
Medicine for eight doses	448	$14 + 42 = 16$	$14 + 42 = 56$

Comparison with model reports. Absolute accuracies are specific to this harness. With Qwen's instruction to put the answer in `\boxed{}` and a 2,048-token budget, Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct score 71.3 and 91.7 on GSM8K, close to the 73.2 and 91.6 of their report (Qwen et al., 2024), against 66.2 and 91.7 under our answer format and 512-token budget. The short budget also makes our MATH500 scores lower bounds. 12% of the 1.5B model's responses on the two benchmarks hit the 512-token limit, and with Qwen's instruction and 2,048 tokens it reaches 41.4 instead of 34.0. The MATH500 cell marked in Table 2 uses a fixed 200-problem subset.

L Models and Data

Tables 24 and 25 list every model and dataset with its public identifier. The two Llama checkpoints are the unsloth re-uploads of Meta's weights. The held-out SVAMP, ASDiv and GSM-Hard items are the last 20% of each set after a fixed shuffle.

Table 24: Models.

Model	Identifier	Role
Qwen2.5-0.5B-Instruct	`Qwen/Qwen2.5-0.5B-Instruct`	student
Qwen2.5-1.5B-Instruct	`Qwen/Qwen2.5-1.5B-Instruct`	student
Qwen2.5-3B-Instruct	`Qwen/Qwen2.5-3B-Instruct`	teacher
Qwen2.5-7B-Instruct	`Qwen/Qwen2.5-7B-Instruct`	teacher and student
Qwen2.5-14B-Instruct	`Qwen/Qwen2.5-14B-Instruct`	teacher and judge
Qwen2.5-32B-Instruct	`Qwen/Qwen2.5-32B-Instruct`	teacher
Llama-3.2-1B-Instruct	`unsloth/Llama-3.2-1B-Instruct`	student
Llama-3.1-8B-Instruct	`unsloth/Meta-Llama-3.1-8B-Instruct`	teacher
OLMo-2-7B-Instruct	`allenai/OLMo-2-1124-7B-Instruct`	teacher

Table 25: Datasets and the part of each that we use.

Dataset	Identifier	Part used
GSM8K	`openai/gsm8k`	main set, a 4,000-problem training pool per seed, 3,000 problems for the SeqKD demonstrations, and all 1,319 test problems
MATH500	`HuggingFaceH4/MATH-500`	all 500 problems
MBPP	`google-research-datasets/mbpp`	full set, 374 training and 500 test problems
GSM-Symbolic	`apple/GSM-Symbolic`	main set, 1,000 of 5,000 items
SVAMP	`ChilleD/SVAMP`	60 of 300 test problems
ASDiv	`yimingzhang/asdiv`	61 of 301 test problems
GSM-Hard	`reasoning-machines/gsm-hard`	264 of 1,319 problems
Alpaca	`tatsu-lab/alpaca`	a 4,000-prompt training pool per seed, 2,500 prompts with teacher responses for SeqKD, and 120 judged prompts

M PER-SEED RESULTS

Table 26 lists every seed behind Tables 2, 5 and 6. The four access rungs vary by at most 3.6 points across seeds on any pair, while the text-level baselines vary most, OmniOPD by 15.3 points and GAD by 6.6 on 7B→1.5B. LadderOPD beats SeqKD on every seed of the three Qwen pairs, and every seed of the shuffled control ends below the untrained student in all six settings.

Table 26: Per-seed accuracy in %, GSM8K except for the MBPP block. Values follow the seed order 42, 123, 456, 789, 1234, 2024, and runs with three seeds use the first three.

Method	Seeds	Per-seed accuracy	Mean±std
7B→1.5B			
SeqKD	6	71.2, 71.0, 69.2, 70.4, 71.3, 71.0	70.7±0.8
OmniOPD	3	71.3, 58.2, 56.0	61.8±8.3
GAD	3	64.4, 71.0, 70.0	68.5±3.5
Top-1 CE (L0)	6	73.5, 72.0, 72.0, 73.4, 72.2, 72.6	72.6±0.7
LadderOPD (L1)	6	74.6, 74.6, 75.4, 71.8, 74.4, 75.4	74.4±1.3
Top-k values (L2)	6	72.8, 74.1, 73.6, 71.9, 75.5, 73.0	73.5±1.2
Full logits (L3)	6	72.8, 74.4, 74.0, 75.3, 73.1, 73.6	73.9±0.9
Shuffled order	3	54.8, 59.4, 56.0	56.7±2.4
3B→0.5B			
SeqKD	6	42.3, 43.4, 44.4, 43.1, 42.0, 42.6	42.9±0.8
OmniOPD	3	31.9, 41.8, 38.8	37.5±5.1
GAD	3	40.5, 42.2, 40.0	40.9±1.1
Top-1 CE (L0)	6	47.4, 48.8, 48.9, 47.4, 48.4, 47.2	48.0±0.8
LadderOPD (L1)	6	51.3, 52.6, 53.2, 51.7, 52.2, 50.2	51.9±1.1
Top-k values (L2)	6	50.5, 50.9, 51.7, 51.0, 51.0, 49.5	50.8±0.7
Full logits (L3)	6	50.4, 52.0, 51.0, 50.2, 50.0, 51.3	50.8±0.8
Shuffled order	3	16.5, 19.2, 22.1	19.3±2.8
14B→1.5B			
SeqKD	6	68.3, 69.2, 69.6, 68.8, 69.8, 70.0	69.3±0.6
OmniOPD	3	61.5, 68.2, 62.9	64.2±3.5
GAD	3	63.9, 70.2, 70.2	68.1±3.6
Top-1 CE (L0)	6	72.7, 73.8, 73.2, 72.6, 72.8, 73.8	73.1±0.5
LadderOPD (L1)	6	73.1, 74.3, 74.3, 74.3, 74.4, 74.8	74.2±0.6
Top-k values (L2)	6	74.1, 73.2, 73.2, 73.0, 74.6, 73.5	73.6±0.6
Full logits (L3)	6	73.8, 74.1, 73.1, 74.1, 73.8, 73.0	73.6±0.5
Shuffled order	3	54.4, 60.0, 57.5	57.3±2.8
Llama-3.1-8B→Llama-3.2-1B			
Top-1 CE (L0)	3	53.6, 51.3, 50.6	51.8±1.6
LadderOPD (L1)	6	52.7, 53.8, 52.5, 51.7, 53.8, 51.8	52.7±0.9
Top-k values (L2)	3	53.5, 53.1, 54.2	53.6±0.6
Full logits (L3)	6	53.2, 50.0, 53.2, 52.9, 51.9, 52.5	52.3±1.2
Shuffled order	3	31.8, 28.8, 36.8	32.5±4.1
MBPP, 7B→1.5B			
Top-1 CE (L0)	3	46.6, 45.4, 43.2	45.1±1.7
LadderOPD (L1)	3	46.8, 47.2, 46.6	46.9±0.3
Top-k values (L2)	3	48.6, 47.0, 46.4	47.3±1.1
Full logits (L3)	3	47.6, 45.8, 47.8	47.1±1.1
Shuffled order	3	23.6, 28.0, 27.6	26.4±2.4
32B→7B			
LadderOPD (L1)	3	90.8, 90.5, 91.1	90.8±0.3
LadderOPD, no fork gate	3	90.5, 91.0, 89.8	90.4±0.6
Top-k values (L2)	3	90.2, 90.4, 90.7	90.4±0.2
Full logits (L3)	3	90.3, 90.8, 90.4	90.5±0.2
Shuffled order	3	83.9, 83.7, 86.1	84.6±1.3