SCALING PROPERTIES OF CAPABILITY TRANSFER IN SAME-FAMILY ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) can induce substantial reasoning capabilities in large language models (LLMs), but how much of this capability transfers across model scales, and how quickly, remains unclear. We study the scaling properties of *on-policy distillation (OPD)* across *weak-to-strong*, *same-base*, and *strong-to-weak* teacher–student setups. We find that early OPD training dynamics uniformly exhibit a regular *useful-transfer* regime, in which held-out accuracy (the *gold score*, G) rises approximately linearly in $d = \sqrt{\mathrm{KL}(\pi_\theta \| \pi_{\mathrm{ref}})}$, the square root of token-level reverse KL divergence from the student initialization. In every observed weak-to-strong pair, the student's peak gold score exceeds its teacher's own, so a compact RL expert can transfer capability to a much larger student via OPD. To estimate OPD outcomes, we fit *power laws* for how G_{peak} and the slope of the useful-transfer regime scale with student and teacher parameter counts and with teacher gold score. These laws show that peak gold score improves with teacher scale only up to roughly the student's scale, and that at a matched gold score smaller teachers transfer better, so a teacher's score alone does not define its supervision value. We also study the scaling effects of two OPD variants, bootstrapping weak-to-strong OPD, and the degree of on-policy supervision.

1 INTRODUCTION

Modern reasoning models often acquire task expertise through *reinforcement learning (RL)* (Shao et al., 2024; Olmo Team, 2025; Yang et al., 2025). *On-policy distillation (OPD)* is increasingly used in LLM post-training pipelines to transfer such expertise between models: a teacher policy provides token-level supervision on rollouts sampled from the student, without forcing the student to imitate the teacher's own trajectories (Agarwal et al., 2024; Gu et al., 2024; Lu & Thinking Machines Lab, 2025). Recent work has produced a growing family of OPD variants that reshape token rewards and distillation objectives (Song & Zheng, 2026; Yang et al., 2026c; Ko et al., 2026), and empirical analyses examine OPD training dynamics along axes such as rollout length and token-overlap ratio (Fu et al., 2026; Li et al., 2026b). However, the scaling properties of OPD, especially how outcomes depend on teacher and student scale, remain poorly understood. If these dependencies were captured in law-like form, the outcome of an OPD run could be predicted rather than discovered after training. Therefore, **the central question** we ask is: *can we estimate the performance of the outcome student policy from teacher and student scale, before performing OPD?*

Studies of reward model overoptimization offer an empirical roadmap: in RL with *proximal policy optimization (PPO)* against a learned reward model, gold reward dynamics can be described as functions of the KL divergence from the initial policy, with coefficients predictable from reward-model scale (Gao et al., 2023). A similar phenomenon arises in direct alignment algorithms such as DPO (Rafailov et al., 2024). OPD also optimizes the student against a proxy, the token-level implicit reward induced by a fixed teacher. We adopt the same approach: we characterize held-out accuracy (the *gold score* G, by analogy with the gold reward above) along the KL-indexed training progress $d := \sqrt{\mathrm{KL}(\pi_\theta \| \pi_{\mathrm{ref}})}$ from the initial student. We then try to predict the fitted coefficients from student and teacher scale. We find that OPD dynamics share an initial regular *useful-transfer regime* where gold score rises approximately linearly with d. Subsequent dynamics are noisy and separate into attenuated improvement, saturation, and regression, unlike the regular overoptimization regime of PPO and direct alignment.

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.

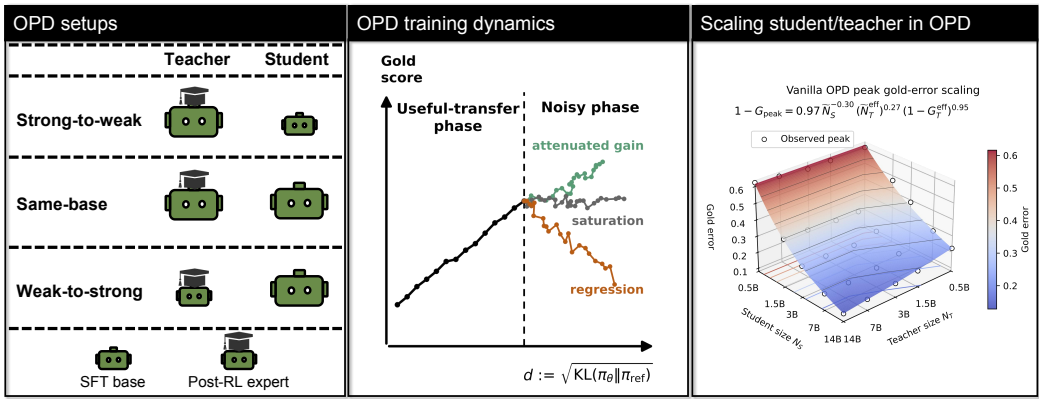

Figure 1: OPD transfers expertise from an RL-trained teacher to a student policy. **Left:** the three teacher–student setups, with post-RL experts as teachers and SFT bases as students. **Center:** gold score rises linearly with KL-indexed training progress d during the useful-transfer phase, after which dynamics turn noisy. **Right:** the fitted power law predicts peak gold error from teacher and student scale (top), shown with cross-sections at a fixed teacher or student size (bottom).

OPD applies across three teacher–student configurations: *strong-to-weak*, *same-base*, and *weak-to-strong*. The weak-to-strong direction is especially attractive because it amortizes task-specific post-training across a model family: instead of repeating RL at every scale, a weak expert serves as a cheap proxy for task expertise, while the strong student contributes knowledge and reasoning strategies its teacher lacks. OPD strengthens prior weak-to-strong supervision (Burns et al., 2024; Yuan et al., 2026) by querying the teacher on fresh rollouts from the evolving student rather than relying on teacher demonstrations. Recent weak-to-strong OPD methods show that policy contrasts can improve transfer from weak RL experts (Yu et al., 2026; Feng et al., 2026; Park et al., 2026), but do not systematically characterize how much and how fast capability transfers across scales.

In this work, we study the scaling properties of OPD by asking three questions. **First**, do OPD training dynamics contain a regular, predictable regime, and how long does it last? **Second**, how do student and teacher scale shape peak gold score, useful-transfer rate, and transfer extent, and can the fitted laws predict outcomes at held-out scales? **Third**, how do design choices such as the OPD objective (*Vanilla-OPD* vs. *Delta-OPD*) and the degree of on-policy supervision change transfer, and does bootstrapping weak-to-strong OPD along a model family improve on direct transfer from the smallest expert?

We investigate these questions via controlled experiments on math reasoning with Qwen2.5 models (0.5B–14B), fitting power laws in student and teacher parameter count and measured teacher score for peak gold score and useful-transfer slope (Figure 1). Our **main findings**:

- **Capability transfer has a regular initial regime and heterogeneous later dynamics.** Initial OPD training dynamics exhibit a regular useful-transfer regime, with gold score increasing linearly with training progress (d) (§4).
- **Scale predicts peak capability and local useful-transfer rate.** Peak remaining error and useful-transfer rate follow joint power laws in student size, effective teacher size, and measured teacher score, with the peak law extrapolating to the largest held-out scales within one accuracy point. At a matched score, smaller teachers transfer better (§5).
- **The distillation objective changes capability transfer.** Delta-OPD produces a larger local matched-KL gain slope in 15 shared pairs and a larger observed peak gain in 12, with the advantage concentrated in weak-to-strong pairs (§6.1).
- **Off-policy cold start harms weak-to-strong OPD.** An off-policy SFT phase harms OPD increasingly with the weak-to-strong capability gap (§6.2).
- **Bootstrapping weak-to-strong OPD does not improve on direct transfer.** Every bootstrapped chain peaks below direct OPD from the smallest post-RL expert at the same student scale, even though its intermediate teachers attain higher gold scores (§6.3).

2 PRELIMINARIES

This section fixes notation and formalizes the three distillation objectives used throughout the paper.

Vanilla-OPD. Let $x \sim \mathcal{D}$ be a prompt, y a student rollout, π_{T} a fixed teacher, and π_θ the policy parameterized by θ. The original OPD recipe minimizes the sequence-level reverse KL divergence between the policy and the teacher (Agarwal et al., 2024; Gu et al., 2024) (subscript V for "Vanilla"):

$$\min_\theta \mathcal{L}_{\mathrm{V}}(\theta) = \min_\theta \mathop{\mathbb{E}}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)} \left[\mathrm{KL}(\pi_\theta(y|x)\|\pi_{\mathrm{T}}(y|x))\right]. \tag{1}$$

By the chain rule of KL divergence, minimizing $\mathcal{L}_{\mathrm{V}}$ is equivalent to maximizing an RL-like objective in which token t receives the teacher-induced token reward $r_t = \log \pi_{\mathrm{T}}(y_t|x, y_{<t}) - \log \pi_\theta(y_t|x, y_{<t})$, and the exact policy gradient assigns token t the return-to-go advantage $\sum_{t'=t}^{|y|} r_{t'}$. Modern implementations instead apply a *zero-discount update*, discounting future rewards to zero so that each token receives only its immediate reward, $A_t^{\mathrm{V}} = r_t$ (Lu & Thinking Machines Lab, 2025). This practice is widely adopted in frontier post-training pipelines (Yang et al., 2025; Xiao et al., 2026; Zeng et al., 2026):

$$\max_\theta \mathcal{J}_{\mathrm{V}}(\theta) = \max_\theta \mathop{\mathbb{E}}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)} \left[\sum_{t=1}^{|y|} \underbrace{\log \pi_{\mathrm{T}}(y_t|x, y_{<t}) - \log \pi_\theta(y_t|x, y_{<t})}_{A_t^{\mathrm{V}}(x,y)}\right]. \tag{2}$$

In expectation, $-A_t^{\mathrm{V}}$ equals the full-vocabulary conditional reverse KL at each student-visited prefix. Estimating this divergence with only the sampled token's logprobs, rather than a sum over the full vocabulary, is known as *sampled-token estimation* (Li et al., 2026b), which we adopt for Vanilla-OPD under the policy-gradient framework by default.

Delta-OPD. The second variant, which we call Delta-OPD, derives its token reward from the policy shift the teacher acquired during RL. This reward construction is the common core of variants proposed for both transfer directions: OPD^2 (Heo et al., 2026) applies it to strong-to-weak distillation, and Direct-OPD (Feng et al., 2026) and W2S-OPD (Yu et al., 2026) apply it to weak-to-strong distillation. We therefore adopt it as the representative alternative objective for a design that spans weak-to-strong, same-base, and strong-to-weak pairs. Let $\pi_{\mathrm{T}}^{\mathrm{base}}$ be the SFT checkpoint from which π_{T} was produced by RL, and π_{ref} be the student base policy. The objective is as follows (superscript Δ for "Delta"):

$$\begin{aligned} A_t^\Delta(x,y) = r_t^\Delta &:= \log \pi_{\mathrm{T}}(y_t|x, y_{<t}) - \log \pi_{\mathrm{T}}^{\mathrm{base}}(y_t|x, y_{<t}), \\ \max_\theta \mathcal{J}_\Delta(\theta) &= \max_\theta \mathop{\mathbb{E}}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)} \left[\sum_{t=1}^{|y|} A_t^\Delta(x,y) - \mathrm{KL}(\pi_\theta(\cdot|x, y_{<t})\|\pi_{\mathrm{ref}}(\cdot|x, y_{<t}))\right]. \end{aligned} \tag{3}$$

Off-policy distillation (OffPD). When rollouts are instead sampled from the teacher and the student maximizes their likelihood, as in SFT on teacher demonstrations, distillation becomes sequence-level *off-policy distillation (OffPD)* (Hinton et al., 2015; Kim & Rush, 2016):

$$\min_\theta \mathcal{L}_{\mathrm{Off}}(\theta) = \min_\theta \mathop{\mathbb{E}}_{x\sim\mathcal{D},\, y\sim\pi_{\mathrm{T}}(\cdot|x)} \left[-\log \pi_\theta(y|x)\right]. \tag{4}$$

3 EXPERIMENTAL DESIGN

Models and training pipeline. The study uses Qwen2.5 Base (not instruction-tuned) models (Qwen Team, 2024) at 0.5B, 1.5B, 3B, 7B, and 14B parameters. All models undergo an initial SFT phase on a subset of Dolci-SFT (Olmo Team, 2025) for basic instruction-following under a chat template. Teacher models are obtained by GRPO RL on the mixed GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) training split of 14.8K examples and evaluated on the corresponding mixed test split of 6.3K examples (results in Figure 11). OPD and RL use the same training prompts, and each OPD run lasts at most ten epochs (580 updates) with periodic held-out evaluation. All training phases are implemented in the verl framework (Sheng et al., 2025) (hyperparameters in §E). The design contains 25 teacher–student combinations, including weak-to-strong, same-base, and strong-to-weak setups.

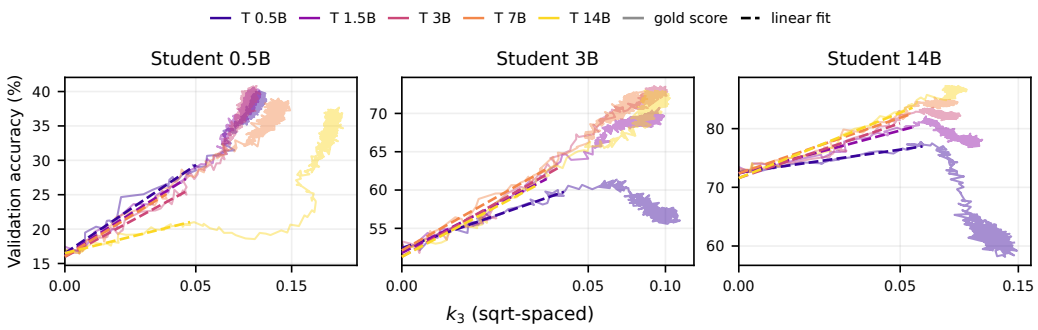

Figure 2: Gold score for the 0.5B, 3B, and 14B students as teacher scale varies. Dashed lines show linear fits to the first 30 observations. Full results are in Figures 9 and 10.

Metrics. Gold score is accuracy on the held-out test set, on which both the RL teachers and the OPD students are evaluated. It plays the role of the gold reward in Gao et al. (2023), with the teacher-induced token reward as the proxy. For prompts $x \sim \mathcal{D}$ and student rollouts y, define

$$k_3(\pi_\theta, \pi_{\text{ref}}) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} \exp(\delta_t) - \delta_t - 1 \right], \quad \text{where } \delta_t = \log \pi_{\text{ref}}(y_t|x, y_{<t}) - \log \pi_\theta(y_t|x, y_{<t}). \tag{5}$$

This is the nonnegative, unbiased k_3 estimator of reverse KL (Schulman, 2020; Tang & Munos, 2025), averaged over response tokens. Unlike Gao et al. (2023), who measure sequence-level KL, we measure token-mean KL as this matches the immediate-token objective in Equation (2). Since KL is a quadratic measure, we use its square root, $d := \sqrt{k_3}$, as a proxy for training progress.

4 Characterizing Capability-Transfer Dynamics in OPD

This section addresses our first question: *do OPD training dynamics contain a regular, predictable regime, and how long does it last?* We examine the checkpoint trajectories of the 25 Vanilla-OPD runs, which span weak-to-strong, same-base, and strong-to-weak setups, reading each run as a curve of gold score G against training progress d (§3) following Gao et al. (2023). The quantities introduced here, namely the transfer rate, the transfer extent, and the peak, are what §5 later characterizes across teacher and student scale.

Two phases emerge consistently. Transfer begins with a regular useful-transfer regime, in which gold score rises approximately linearly in d at a rate m. We define the *transfer endpoint* d_{transfer} as the last d before the trajectory falls below the 95% predictive band of its initial linear fit for three consecutive checkpoints. Beyond d_{transfer}, dynamics are noisy and heterogeneous, mixing attenuated improvement, saturation, and regression around the gold-score maximum at d_{peak}. Figure 2 overlays all teachers for three representative students, and Figures 9 and 10 show every pair. Every trajectory improves initially, and peak gold score increases with teacher scale for the larger students, from 72.7% to 81.9% for the 7B student and from 77.7% to 87.3% for the 14B student. In every panel the smallest teacher shows the clearest post-peak regression, while other tails attenuate or saturate. For smaller students, larger teachers do not always improve the peak: the 0.5B student peaks at 40.8% with the 3B teacher but reaches only 39.0% under the 7B teacher and 37.7% under the 14B teacher.

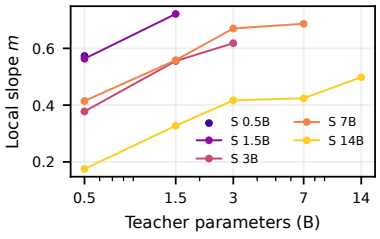

Figure 3: Fitted initial slope m across teacher scale, including only teachers no larger than each fixed student.

Initial transfer is approximately linear in the KL coordinate. Across the 25 runs, linear fits $G(d) = c + md$ to the first 30 checkpoints obtain $R^2 \in [0.932, 0.988]$ and RMSE in $[0.0026, 0.0129]$

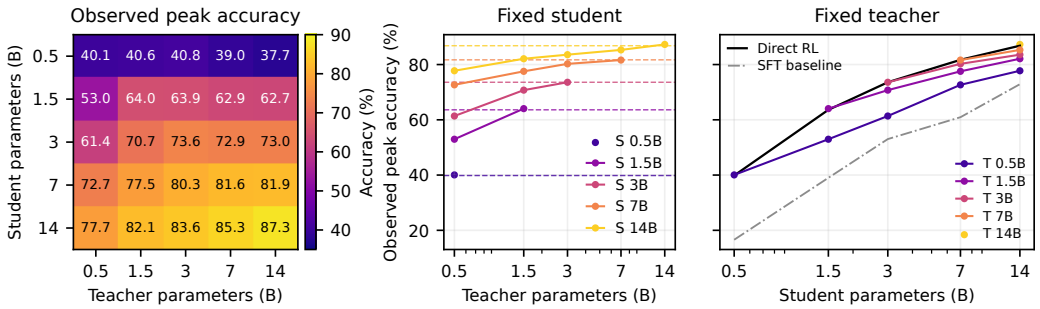

Figure 4: Peak gold score across model scale. Left: the complete 25-cell Vanilla grid. Middle: fixed-student Vanilla trends using teachers no larger than the student, with color-matched dashed levels marking each student's direct-RL endpoint. Right: Vanilla peaks grouped by teacher, together with the direct-RL endpoint and the SFT baseline, on the same y axis as the middle panel. Parameter axes are logarithmic and all curves are one-run descriptions.

accuracy units. The fitted slopes span $[0.175, 0.721]$: weak teachers generally induce smaller initial gains for larger students, while teachers closer to the student's scale tend to raise gold score faster per unit d (Figure 3). The observed tails include attenuated improvement, near-saturation, and regression, with no shared functional form.

Why square-root KL linearizes local transfer. Local KL geometry explains the observed linearity: along a smooth training path, gold score changes to first order in the parameter perturbation while KL changes to second order, so $G(d) = G(0) + md + O(d^2)$ with a slope set by the Fisher-normalized alignment of the update direction with the gold-score gradient (statement and proof in §D).

4.1 How scale shapes transfer and later dynamics

Figure 4 relates the complete Vanilla-OPD grid to direct student RL. At fixed teacher size, observed peak gold score increases with student scale. At fixed 7B and 14B students, it also increases monotonically with teacher size. The fixed-3B-student series instead peaks with the 3B teacher rather than the 7B teacher, and the 0.5B and 1.5B students lose accuracy under their largest teachers. Vanilla-OPD same-base peaks lie within 0.5 percentage points of the final direct-RL reference at all five scales, exceeding it at three of them.

The locations d_{peak} of the gold-score maximum are less monotone: for the 7B student they are 0.274, 0.282, 0.308, 0.323, and 0.323 as teacher size increases, and other student series likewise contain reversals. We therefore track the peak value G_{peak}, initial rate m, and transfer extent d_{transfer} rather than imposing a parametric law on the location of the observed maximum.

Does the teacher proxy outlive the gold gain? The teacher-induced implicit reward yields a continuous proxy score P, logged as the token-mean reward on training rollouts. Whether P keeps improving after gold score peaks distinguishes implicit-reward overoptimization from simple over-imitation. Figure 9 overlays the logged proxy curves with gold score: wherever gold score regresses, P keeps rising, so late-stage regression carries the signature of implicit-reward overoptimization.

When does the initial law cease to predict? The transfer endpoint is a sequential predictive quantity rather than a fitted turning point. Under the primary 95% band and three-consecutive-deviation rule, nine of 25 Vanilla-OPD trajectories depart within their observed support. Changing the confidence level and persistence requirement varies this count only from eight to twelve (Table 15).

5 Fitting power laws for peak gold score and useful-transfer rate

This section addresses our second question: *how do student and teacher scale shape peak gold score, useful-transfer rate, and transfer extent, and do the fitted laws extrapolate to held-out scales?* Since dynamics are regular only up to the transfer endpoint (§4), we fit the rate m and the peak G_{peak} as

Table 1: Fitted joint power-law coefficients (Equation (6)).

(a) Peak capability (Equation (6), left).

OPD method	A	α	β	ζ
Vanilla-OPD	0.97	0.30	-0.27	0.95
Delta-OPD	1.02	0.34	-0.33	1.01

(b) Useful-transfer rate (Equation (6), right).

OPD method	B	γ	δ	ξ
Vanilla-OPD	0.12	0.19	-0.61	1.90
Delta-OPD	0.13	0.20	-0.73	2.01

Table 2: Validation of the peak laws, in accuracy points. Leave-one-scale-out RMSE refits each law with all cells sharing one student or teacher scale withheld and predicts the withheld cells (mean over the two split types); extrapolation RMSE withholds the largest student (S) or teacher (T) scale. Protocols and further diagnostics are in §H. Best results are highlighted in bold.

OPD method	Peak law	Leave-one-scale-out (↓)	Extrapolation S / T (↓)
Vanilla-OPD	Scales only (Equation (7))	3.43	.64 / .75
	Teacher score only	2.55	2.54 / .70
	Joint (Equation (6))	**1.66**	**.55** / **.68**
Delta-OPD	Scales only (Equation (7))	2.47	.21 / **.29**
	Teacher score only	2.07	.93 / .46
	Joint (Equation (6))	**0.82**	**.20** / .32

separate targets and evaluate every law by withholding the largest models, while the extent d_{transfer} is summarized as an empirical KL budget.

We normalize parameter counts by one billion and cap teacher scale at student scale, $\widetilde{N}_S = N_S/(1\text{B})$ and $\widetilde{N}_T^{\text{eff}} = \min(N_T, N_S)/(1\text{B})$. The cap approximates the saturation and reversal observed above student scale (Figure 4); the smallest students' reversals remain unmodeled residuals.

Joint laws in scale and teacher score. Parameter count summarizes a teacher only when every teacher is trained to its RL endpoint. An undertrained teacher scores below the size trend. Our primary model therefore conditions the peak and rate on the remaining error of the effective teacher, whose gold score is G_T^{eff}, fitting Vanilla-OPD and Delta-OPD independently with the shared functional families

$$1 - G_{\text{peak}} = A\widetilde{N}_S^{-\alpha}(\widetilde{N}_T^{\text{eff}})^{-\beta}\left(1 - G_T^{\text{eff}}\right)^{\zeta}, \qquad m = B\widetilde{N}_S^{-\gamma}(\widetilde{N}_T^{\text{eff}})^{\delta}\left(1 - G_T^{\text{eff}}\right)^{-\xi}. \tag{6}$$

On the grid whose teachers are all RL endpoints, the two covariates are collinear ($r = -0.9999$), so the exponents of Equation (6) cannot be separated there. The bootstrapped chains of §6.3 break this collinearity with teachers that are themselves OPD products of a previous chain stage, whose measured gold scores and initial slopes sit below the size trend. Adding their five Vanilla-OPD and three Delta-OPD cells lowers the collinearity to -0.95 and -0.97 and identifies every teacher exponent, with all bootstrap intervals excluding zero (§H). Table 1 reports the fitted joint laws.

Motivation for multiplicative power law. Scaling the student removes a constant fraction of whatever error the teacher-induced supervision leaves, and the multiplicative family encodes exactly this interaction. An additive alternative instead posits a teacher-induced error that persists as $N_S \to \infty$ and yields higher fitting errors (§H).

Scale-only baseline. Dropping the teacher-score factor yields the family that reads scaling off parameter counts alone,

$$1 - G_{\text{peak}} = A\widetilde{N}_S^{-\alpha}(\widetilde{N}_T^{\text{eff}})^{-\beta}, \qquad m = B\widetilde{N}_S^{-\gamma}(\widetilde{N}_T^{\text{eff}})^{\delta}. \tag{7}$$

This family is the baseline that Table 2 compares the joint laws against.

Remark (On parameter count as the scaling variable)**.** Parameter count is a confounded measure that implicitly incorporates model scale, data scale, and training compute. Since Qwen Team (2024) tuned architecture and data scale by scaling laws, our power laws should be read at compute-optimal settings.

Peak capability. In every observed weak-to-strong pair, the student's peak gold score exceeds its teacher's own, by margins that shrink as teacher scale approaches student scale (Figure 4). With

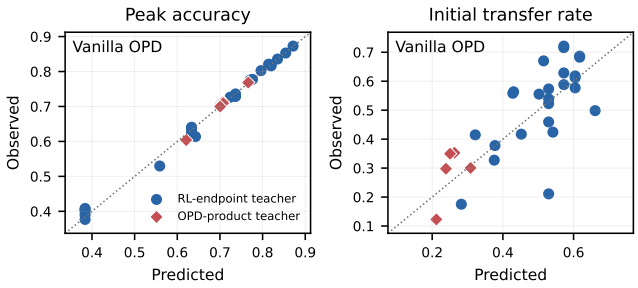

Figure 5: Observed Vanilla-OPD peak gold score and initial transfer rate against full-fit predictions of the joint laws of Equation (6). Red diamonds mark cells with bootstrapped-OPD teachers (§6.3).

$\zeta \approx 1$ for both methods, peak student error is nearly proportional to the capped teacher's remaining error, scaled down by a power of student size; the negative β means that at matched score the smaller teacher transfers better. Table 2 validates these laws: conditioning on teacher score halves the leave-one-scale-out RMSE of the scale-only baseline of Equation (7), and the joint law extrapolates to the held-out largest student and teacher within 0.7 (Vanilla-OPD) and 0.4 (Delta-OPD) accuracy points. A controlled comparison tests the negative β out of sample: we distill the 7B student from an intermediate 3B teacher checkpoint scoring 66.0, slightly above the 1.5B RL endpoint's 63.6. The joint law predicts its peak within 0.3 points (74.1 against the observed 73.8) and the correct ordering below the 1.5B teacher's 77.5. The scale-only and score-only laws instead both favor the larger, higher-scoring teacher (§H).

Useful-transfer rate. The rate law mirrors the peak law with stronger score dependence: $\xi = 1.9$ for Vanilla-OPD and 2.0 for Delta-OPD, so transfer slows sharply as the teacher's remaining error grows, and at matched score the larger teacher transfers more slowly. Rate fits are noisier than peak fits (log-space R^2 of 0.59 and 0.76); the rate law is better read as an interpretable scale summary than a precise predictor (§H).

Useful-transfer extent. The transfer extent resists a comparable law. Observed departures and censored lower bounds together span d of 0.20–0.36 for Vanilla-OPD and 0.27–0.34 for Delta-OPD, within a factor of two across the grid, with medians of 0.30 and 0.29. A censored accelerated-failure-time fit finds only weak scale dependence, and its Delta-OPD exponents are not identifiable (§H). We therefore read the extent as an approximately scale-free KL budget rather than a scaling target. Figure 5 plots predicted against observed Vanilla-OPD quantities; the Delta-OPD counterpart and the extent diagnostics are in §H.

6 Effects of Design Choices on OPD Transfer

This section addresses our third question: *how do design choices change OPD transfer?* We vary one choice at a time: the OPD objective, the degree of on-policy supervision, and bootstrapping.

6.1 Effect of OPD Variant: Delta-OPD

Our analyses above focus on Vanilla-OPD. Since many OPD variants have appeared recently, we take Delta-OPD (§2) as the alternative condition and ask how the objective changes transfer.

We compare 17 Delta-OPD runs with Vanilla-OPD cells having the same teacher and student scales. Ten pairs are weak-to-strong, five are same-base, and 0.5B←1.5B and 1.5B←3B are strong-to-weak controls. Because separately sampled

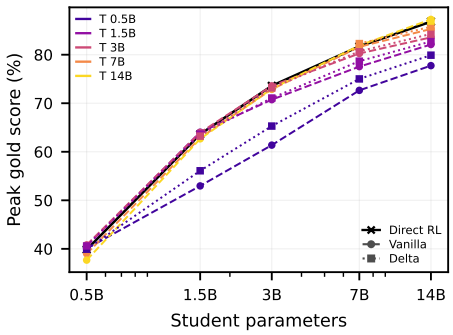

Figure 6: Peak gold score against student scale for Delta-OPD, Vanilla-OPD, and direct RL, grouped by teacher.

step-zero accuracies differ, the primary quantity is gain over each run's own initialization, $\Delta G_{\mathrm{V}}(d) = G_{\mathrm{V}}(d) - G_{\mathrm{V}}(0)$ and $\Delta G_{\Delta}(d) = G_{\Delta}(d) - G_{\Delta}(0)$. The local slope compares matched-KL gold gain. We estimate it from 30 Vanilla-OPD and 40 Delta-OPD observations (Delta-OPD's regular phase spans more checkpoints) and restrict fitted comparisons to their common d support.

First-40 Delta-OPD lines obtain $R^2 \in [0.916, 0.987]$, RMSE in $[0.0049, 0.0116]$ (Figure 10). Their slopes exceed Vanilla-OPD in 15 of the 17 shared cells; the exceptions are the two smallest students under the 0.5B teacher (Table 8). Delta-OPD attains the larger baseline-normalized peak gain and the larger absolute peak in 12 cells each, mostly in weak-to-strong pairs (nine of ten; Figure 6). The benefit shrinks with teacher scale: the 0.5B teacher yields two to four extra points of absolute peak, whereas larger teachers stay within about one point. Five of the 17 runs depart from the initial line.

6.2 Degree of on-policy supervision

Our previous results are based on pure OPD, where a teacher policy directly supervises a base model. Since SFT on teacher rollouts is often used as a cold start that aligns the student's output distribution with the teacher's (Li et al., 2026b), we compare three degrees of on-policy supervision: purely on-policy, on-policy with off-policy cold start, and purely off-policy.

The three conditions share teachers and the evaluation protocol of §3. Pure OPD trains as before. The cold-start condition first performs one epoch of SFT on teacher rollouts over 20K Dolci-Think-RL prompts (Olmo Team, 2025), then runs the full 580 OPD updates. Pure OffPD trains on teacher rollouts alone, five sampled responses per training prompt for five epochs (§2). We report each condition's peak gold score over its training budget, with the 0.5B expert as teacher in Figure 7 and the full weak-to-strong and same-base grid in Figure 20.

Pure OPD attains the best peak gold score in every cell (Figure 7). The off-policy cold start grows increasingly harmful with the weak-to-strong capability gap, costing 6.6, 15.7, and 19.4 points for the 3B, 7B, and 14B students taught by the 0.5B expert. The damage is done by the cold start itself: one epoch of SFT on the 0.5B expert's rollouts pins every student near the teacher's own score, erasing 32 points of the 14B student's initialization. The subsequent OPD phase does not recover the loss (dynamics in §J). Pure OffPD underperforms pure OPD in every cell, by 28.4 points in the extreme weak-to-strong cell and by less than one point in same-base cells (Figure 20), so on-policy supervision matters most where weak-to-strong transfer is most attractive.

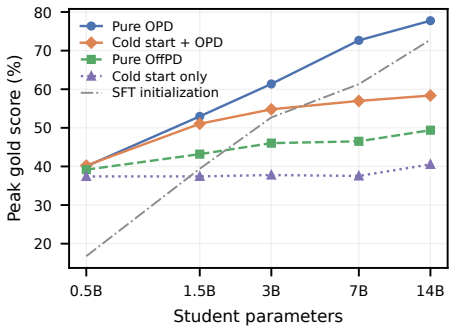

Figure 7: Peak gold score under three degrees of on-policy supervision (0.5B expert as teacher).

6.3 Bootstrapping weak-to-strong OPD

Burns et al. (2024) report that bootstrapping weak-to-strong fine-tuning through intermediate model sizes can outperform a single direct step in their chess puzzle setting, and we test whether the same holds for OPD. The preceding experiments obtain a teacher at scale X by RL on the X-scale SFT model. Here RL is applied only to the smallest (0.5B) model, and every larger teacher in the chain is the previous stage's best OPD checkpoint: $\mathcal{M}_1^{\mathrm{SFT}} \xrightarrow{\mathrm{RL}} \mathcal{M}_1 \xrightarrow{\mathrm{OPD}} \mathcal{M}_2 \xrightarrow{\mathrm{OPD}} \cdots \xrightarrow{\mathrm{OPD}} \mathcal{M}_n$. We run five Vanilla-OPD chains and three Delta-OPD chains, stepping through every intermediate size (up to 0.5B→1.5B→3B→7B→14B) or skipping sizes (e.g. 0.5B→1.5B→7B).

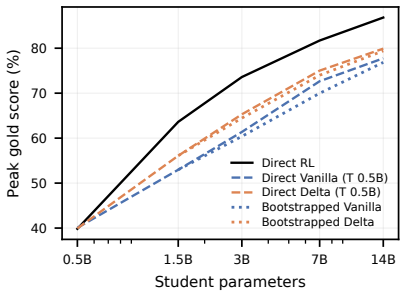

Figure 8: Full 0.5B→14B bootstrapped chains against direct baselines.

We compare each bootstrapped student against direct OPD from the 0.5B expert at the same scale, with direct RL endpoints as references (Figure 8). Bootstrapping

does not improve on direct transfer: every chain peaks below direct OPD from the 0.5B expert (60.4 versus 61.4 at 3B, 70.0–71.3 versus 72.7 at 7B, and 76.8 versus 77.7 at 14B, in %), and longer chains end slightly lower than shorter ones. The comparison also dissociates a teacher's score from its teaching value: the 1.5B OPD product scores 53.0% against the 0.5B RL expert's 39.8%, yet its 3B student peaks below the 0.5B expert's direct student.

This reverses the bootstrapping gains of Burns et al. (2024) and echoes Li et al. (2026b): a higher-scoring teacher helps only when it offers new capabilities. A candidate explanation is the feature-slots picture of Huang et al. (2026), in which each OPD stage bottlenecks the features inherited from the RL expert. The three Delta-OPD chains behave alike, the full one peaking below direct Delta-OPD from the 0.5B expert at every stage (64.5 versus 65.3 at 3B, 74.0 versus 75.0 at 7B, and 79.4 versus 79.9 at 14B). Direct RL on the student itself remains above every weak-teacher variant, so the appeal of weak-to-strong OPD rests on amortizing one expert across a model family rather than on beating direct RL (per-chain dynamics in §J).

7 Related work

On-policy distillation. In OPD, a teacher policy provides token-level supervision on rollouts sampled from the student, avoiding the exposure bias of offline teacher data (Agarwal et al., 2024; Gu et al., 2024; Lu & Thinking Machines Lab, 2025). Variants reshape the token reward and objective (Song & Zheng, 2026; Yang et al., 2026c; Ko et al., 2026), and empirical analyses dissect failure modes and recipes (Fu et al., 2026; Li et al., 2026b). Closest to us, OPD^2 (Heo et al., 2026), Direct-OPD (Feng et al., 2026), and W2S-OPD (Yu et al., 2026) supervise the student with contrasts between post-RL and pre-RL teacher checkpoints, a family our Delta-OPD instantiates, and OPRD instead rescales the student's verifier-driven RL gradient along this shift (Park et al., 2026). These works design and diagnose objectives; we characterize how fixed objectives scale.

Reward model overoptimization. Gao et al. (2023) model gold reward as a function of KL from the initial policy with coefficients predicted from reward-model scale, a roadmap extended to direct alignment (Rafailov et al., 2024) and surveyed as reward hacking (Wang et al., 2026b). Distillation inherits this proxy-optimization hazard: repeatedly fitting fixed offline teacher data induces teacher hacking (Tiapkin et al., 2025), while noisy-expert theory favors online interaction (Sriraman et al., 2026). The teacher-induced token reward is likewise a gold-score proxy, motivating our analysis.

Weak-to-strong generalization. Weak supervision can elicit capabilities beyond the supervisor's own (Burns et al., 2024; Ildiz et al., 2024; Yuan et al., 2026), with feature-learning theory attributing the gains to knowledge latent in the strong model (Awano & Suzuki, 2026). In pretraining distillation, small teachers also improve larger students, with gains saturating or reversing as teachers grow (Lu & Liu, 2026). Our peak scaling law quantifies this regime via the effective-teacher cap.

Predicting model performance with scaling laws. Scaling laws first predicted pretraining loss from model size, data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022) and have since been fit for fine-tuning (Zhang et al., 2024), pretraining distillation (Busbridge et al., 2025), and RL compute (Khatri et al., 2026), and the pretrained state predicts later RL gains (Shen et al., 2026). These laws predict endpoints from resource budgets, while ours target KL-indexed trajectories.

8 Conclusion

We systematically study the scaling properties of OPD in math reasoning. OPD dynamics form a two-phase process: a clean useful-transfer phase, in which gold score (G) is approximately linear in the square root of token-level reverse KL (d), followed by a noisy tail. Joint power laws in student scale, teacher scale, and teacher gold score predict peak gold score (G_{peak}) and, more approximately, useful-transfer slope for both OPD variants. Peak student error is nearly proportional to teacher remaining error, with smaller teachers transferring better at a matched score. Bootstrapping does not improve on direct transfer from the smallest expert, OffPD underperforms OPD, and an off-policy cold start harms weak-to-strong transfer. OPD outcomes can thus be estimated from student scale, teacher scale, and teacher score before training, so RL expertise can be trained once at small scale and transferred predictably across a model family.

AI USE STATEMENT

We used generative AI tools to refine hypotheses, provide feedback on the research, and support qualitative and thematic data analysis. We also used them for manuscript language polishing, literature search for §C, and result visualization (including analysis and plotting code). We did not use AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims or provide critical ingredients for proving them, write proofs, propose hypotheses, design research methodology or experiments, implement methods, translate text, or clean and reformat datasets, and AI tools did not interpret results. We reviewed all AI-assisted prose and code, and independently verified the reported results. We take responsibility for the final text, code, claims, and artifacts in this work.

ETHICS STATEMENT

Distillation can propagate a teacher's spurious task knowledge, hallucinations, biases, and unsafe patterns into a student. Also, (bootstrapped) weak-to-strong distillation may amplify the downstream effects of these defects even when aggregate task accuracy initially improves. Gold-score evaluation on a held-out test set provides partial safeguards, but it does not establish that the resulting model is reliable or safe beyond the evaluated setting. Before deployment, distilled models should undergo broader evaluations of bias, safety, robustness, and distribution shift.

REPRODUCIBILITY STATEMENT

The experimental objectives, evaluation procedure, KL measurement, checkpoint alignment, and fitting protocols are described in §§3–5. The appendix documents trajectories, extended experiments, and the full reproducibility specification. Training is built on verl[1] v0.8.0 (Sheng et al., 2025). Core analysis and plotting code, run manifests, derived summaries, and experiment configurations will be released upon acceptance. Raw checkpoints and datasets will be documented and shared where their licenses and access conditions permit.

We do not repeat training across random seeds due to limited computational resources (§E.5). Preliminary reruns show that OPD is more stable and predictable than RL, so the results are less susceptible to training stochasticity.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Ryoya Awano and Taiji Suzuki. The mechanism of weak-to-strong generalization: Feature elicitation from latent knowledge, 2026. URL https://arxiv.org/abs/2605.12908.

Stephen H Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the structure of generative models without labeled data. In *International Conference on Machine Learning*, pp. 273–282. PMLR, 2017.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4971–5012. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/burns24b.html.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russell Webb. Distillation scaling laws. In *Proceedings of the 42nd International Conference on Machine Learn-*

[1] https://github.com/verl-project/verl

ing, volume 267 of *Proceedings of Machine Learning Research*, pp. 5977–6045. PMLR, 2025. URL https://proceedings.mlr.press/v267/busbridge25a.html.

Ethan Caballero, Priyank Jaini, David Krueger, and Irina Rish. Unified neural scaling laws. *arXiv preprint arXiv:2605.26248*, 2026.

Qiye Cai, Yichuan Ma, Linyang Li, Peiji Li, Yongkang Chen, Qipeng Guo, Yicheng Zou, Tao Gui, Xiaocheng Feng, and Bing Qin. H^2 sd: Hybrid hindsight self-distillation. *arXiv preprint arXiv:2607.18955*, 2026.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien (eds.). *Semi-Supervised Learning*. MIT Press, 2006.

Leshem Choshen, Yang Zhang, and Jacob Andreas. A hitchhiker's guide to scaling law estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10683–10699. PMLR, 2025. URL https://proceedings.mlr.press/v267/choshen25a.html.

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

Shiyuan Feng, Huan-ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Weak-to-strong generalization via direct on-policy distillation, 2026. URL https://arxiv.org/abs/2607.05394.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

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

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

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

Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. On-policy delta distillation, 2026. URL https://arxiv.org/abs/2607.15161.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jing Huang, Daniel Wurgaft, Rachit Bansal, Laura Ruis, Naomi Saphra, David Alvarez-Melis, Andrew Kyle Lampinen, Christopher Potts, and Ekdeep Singh Lubana. Why larger models learn more: Effects of capacity, interference, and rare-task retention. *arXiv preprint arXiv:2605.29548*, 2026.

M. Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws, 2024. URL https://arxiv.org/abs/2410.18837.

Can Jin, Tristan J. Li, Rui Wu, Eddy Z. Zhang, and Dimitris N. Metaxas. Weak critics make strong learners: On-policy critique distillation for scalable oversight. In *3rd AI for Math Workshop: Toward Self-Evolving Scientific Agents*, 2026. URL https://openreview.net/forum?id=oEfedgUChS.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio Calmon. Inference-time reward hacking in large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 61720–61760. Curran Associates, Inc., 2025.

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=FMjeC9Msws.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation, 2026. URL https://arxiv.org/abs/2603.11137.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing. In *Third Conference on Language Modeling*, 2026a. URL https://openreview.net/forum?id=P2OuWwZspP.

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, 2026b. URL https://arxiv.org/abs/2604.13016.

Wenze Lin, Jiale Zhao, Xitai Jiang, Songde Rao, Yining Li, Shenzhi Wang, Bingxiang He, and Gao Huang. On-policy distillation with verifiable reward. *arXiv preprint arXiv:2608.24696*, 2026.

Chris Yuhao Liu et al. Awesome On-Policy Distillation, 2026. URL https://github.com/chrisliu298/awesome-on-policy-distillation.

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

Taiming Lu and Zhuang Liu. Strong teacher not needed? on distillation in llm pretraining. *arXiv preprint arXiv:2605.23857*, 2026.

Olmo Team. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Youngrok Park, Sangmin Bae, Hojung Jung, Jongwoo Ko, Yunseon Choi, Young Jin Kim, Pashmina Cameron, Aaron Courville, and Se-Young Yun. Eliciting weak-to-strong generalization with on-policy reverse distillation, 2026. URL https://arxiv.org/abs/2609.08798.

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

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *Advances in Neural Information Processing Systems*, 37:126207–126242, 2024.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB endowment. International conference on very large data bases*, volume 11, pp. 269, 2017.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason Lee, and Sanjeev Arora. What makes a reward model a good teacher? an optimization perspective. In *Advances in Neural Information Processing Systems*, volume 38, pp. 59162–59222. Curran Associates, Inc., 2025.

John Schulman. Approximating kl divergence. http://joschu.net/blog/kl-approx.html, March 2020. Accessed: 2026-08-17.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897. PMLR, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jingyan Shen, Ang Li, Salman Rahman, Yifan Sun, Micah Goldblum, Matus Telgarsky, and Pavel Izmailov. Understanding reasoning from pretraining to post-training, 2026. URL https://arxiv.org/abs/2607.16097.

Wei Shen, Guanlin Liu, Yu Yue, Ruofei Zhu, Qingping Yang, Chao Xin, and Lin Yan. Exploring data scaling trends and effects in reinforcement learning from human feedback. In *Advances in Neural Information Processing Systems*, volume 38, pp. 15283–15319. Curran Associates, Inc., 2025.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, 2025. verl framework, https://github.com/verl-project/verl.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Ved Sriraman, Peihan Liu, Daniel Hsu, and Adam Block. Behavior cloning is not all you need: The optimality of on-policy distillation for noisy expert feedback, 2026. URL https://arxiv.org/abs/2606.30923.

Yunhao Tang and Rémi Munos. On a few pitfalls in kl divergence gradient estimation for rl. *arXiv preprint arXiv:2506.09477*, 2025.

Daniil Tiapkin, Daniele Calandriello, Johan Ferret, Sarah Perrin, Nino Vieillard, Alexandre Rame, and Mathieu Blondel. On teacher hacking in language model distillation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59552–59569. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/tiapkin25a.html.

Chen Wang, Zhaochun Li, Jionghao Bai, Yining Zhang, Hexuan Deng, Ge Lan, and Yue Wang. Distilled reinforcement learning for llm post-training. *arXiv preprint arXiv:2607.17247*, 2026a.

Xiaohua Wang, Muzhao Tian, Yuqi Zeng, Zisu Huang, Jiakang Yuan, Bowen Chen, Jingwen Xu, Mingbo Zhou, Wenhao Liu, Muling Wu, et al. Reward hacking in the era of large models: Mechanisms, emergent misalignment, challenges. *arXiv preprint arXiv:2604.13602*, 2026b.

Lilian Weng. Scaling laws, carefully. *lilianweng.github.io*, June 2026. URL https://lilianweng.github.io/posts/2026-06-24-scaling-laws/.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

Junyao Yang, Yucheng Shi, Zhongzhi Li, Ruhan Wang, Zongxia Li, Haitao Mi, and Leowei Liang. T1: Terminal agent reinforcement learning for long-horizon tasks. *arXiv preprint arXiv:2609.11042*, 2026b.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation, 2026c. URL https://arxiv.org/abs/2602.12125.

Zhiqin Yang, Jingwen Fu, Yuhan Liu, Hengyu Liu, Yonggang Zhang, Kainan Cao, Zizhuo Zhang, Chenxin Li, Ruibin Yuan, Jiahao Pan, et al. Scaling large reasoning models beyond human supervision: A path toward superintelligence. *arXiv preprint arXiv:2608.31075*, 2026d.

Fangxu Yu, Weijia Xu, Michael Xu, Tianyi Zhou, and Zinan Lin. Weak-to-strong on-policy distillation, 2026. URL https://arxiv.org/abs/2607.26246.

Yige Yuan, Teng Xiao, Shuchang Tao, Xue Wang, Jinyang Gao, Bolin Ding, and Bingbing Xu. Incentivizing strong reasoning from weak supervision. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7138–7156, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.336. URL https://aclanthology.org/2026.eacl-long.336/.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets LLM finetuning: The effect of data, model and finetuning method. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5HCnKDeTws.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1): 44–53, 2018.

Appendix

Table of Contents

A Discussion **16**

B Limitations **17**

C Extended related work **17**

D Local geometry of capability transfer **19**
D.1 Statement 19
D.2 A local square-root-KL law 20
D.3 Method-specific directions and slopes 20
D.4 Assumptions and scope 21

E Hyperparameters **21**
E.1 SFT hyperparameters 21
E.2 RL hyperparameters 21
E.3 OPD hyperparameters 22
E.4 Evaluation hyperparameters 22
E.5 Random seeds 22

F Qwen2.5 math trajectory grid **22**

G Expanded Delta-OPD and RL results **25**

H Scaling law estimation and diagnostics **26**

I Endpoint and estimator sensitivity **33**

J Extended results for design choices **33**

K Lessons learned and what did not work **34**

Table 3: Summary of notations.

Symbol	Meaning
$x, \mathcal{D}$	Prompt and training-prompt distribution.
$y, y_{<t}, y_t, \lvert y\rvert$	Complete rollout, prefix before token t, token t, and rollout length.
$\pi_\theta, \pi_{\text{ref}}$	Student policy parameterized by θ and its fixed SFT reference.
$\pi_{\text{T}}, \pi_{\text{T}}^{\text{base}}$	RL-trained teacher and its pre-RL reference checkpoint.
r_t, r_t^{Δ}	Vanilla-OPD teacher-induced token reward and Delta-OPD teacher-shift reward.
$A_t^{\text{V}}, A_t^{\Delta}$	Advantages of Vanilla-OPD and Delta-OPD, equal to the token rewards under the zero-discount update.
$\mathcal{L}_{\text{V}}, \mathcal{J}_{\text{V}}, \mathcal{J}_{\Delta}, \mathcal{L}_{\text{Off}}$	Sequence-level Vanilla-OPD loss, token-level Vanilla-OPD objective, Delta-OPD objective, and OffPD loss.
$G(\theta), G(d)$	Gold score as a function of parameters or of the KL coordinate.
P	Teacher-induced proxy score, logged as the token-mean reward on training rollouts.
δ_t, k_3	Token log-probability difference and primary token-mean KL estimator.
d	KL-indexed amount of optimization, $d = \sqrt{k_3}$.
d_{transfer}	Last checkpoint compatible with the initial line before three sustained lower deviations; right-censored when no departure is observed.
d_{peak}	Earliest observed value of d attaining maximum gold score.
$N_S, N_T, \widetilde{N}_S, \widetilde{N}_T^{\text{eff}}$	Student scale, teacher scale, the student scale normalized by one billion parameters, and the teacher scale capped at student scale and likewise normalized.
G_T, G_T^{eff}	Held-out gold accuracy of the teacher RL endpoint, and of the teacher capped at student scale.
G_{peak}	Peak gold score of a trajectory.
A, α, β, ζ	Joint peak-law coefficients (Equation (6)); ζ acts on the capped-teacher error $1 - G_T^{\text{eff}}$.
B, γ, δ, ξ	Joint transfer-rate-law coefficients (Equation (6)), fitted independently for Vanilla-OPD and Delta-OPD.
c, m	Intercept and slope of the local linear fit $G(d) = c + md$.
$G_{\text{V}}(d), G_{\Delta}(d), \Delta G_{\text{V}}, \Delta G_{\Delta}$	Method-specific gold scores and gains over each run's own initialization.
$\mathcal{M}_1, \ldots, \mathcal{M}_n$	Models along the bootstrapping chain in increasing scale; $\mathcal{M}_1$ is the 0.5B RL endpoint and each later $\mathcal{M}_i$ is the OPD product taught by $\mathcal{M}_{i-1}$.

A Discussion

Capability dynamics are indexed by KL. Across the Qwen2.5 grid, every trajectory begins with useful transfer, while later checkpoints exhibit attenuated improvement, saturation, or regression. Both d_{transfer} and d_{peak} are properties of the teacher–student pair. Their reversals across scale show why a universal optimizer-step budget is insufficient.

Teacher quality is more than teacher size. The monotone peak trends for 7B and 14B students coexist with counterexamples at smaller scales. A useful scaling model should therefore condition on measured teacher competence as well as parameter count. Within the RL-endpoint grid the two are collinear at $r = -0.9999$ and cannot be distinguished. The bootstrapped chains break this collinearity with OPD-product teachers whose scores sit off the size trend, and the resulting joint law (Equation (6)) attributes peak error mainly to teacher remaining error, with a negative size exponent at matched score (Table 10). A held-out comparison confirms this out of sample: an intermediate 3B checkpoint scoring slightly above the 1.5B RL endpoint teaches the 7B student worse, as only the joint law predicts (§H). The bootstrapping comparison (§6.3) shows the same dissociation at the trajectory level: the 1.5B OPD teacher outscores the 0.5B RL expert yet teaches the 3B student worse.

Prediction connects the two regimes. The first 30 Vanilla-OPD and first 40 Delta-OPD observations estimate local transfer rate. Across scale, remaining peak error follows a stable joint power law in student scale and teacher score, while transfer extent stays within a narrow band. Peak capability therefore follows scale and teacher quality, whereas the extent of the regular regime behaves as an approximately scale-free KL budget.

Regression and overoptimization are not synonymous. Gold degradation can arise because the student over-imitates a teacher that lacks some of its capabilities. Teacher-induced implicit-reward overoptimization additionally requires fixed-prompt evidence that P continues to improve as G falls. Collapse denotes an abrupt optimization failure. The diagnostics separate three testable mechanisms.

Implications for OPD design. The peak law forecasts attainable capability from student scale and measured teacher quality, while the rate law quantifies how quickly a method spends KL to approach it. The empirical transfer budget supplies a separate KL allowance, after which gold measurements determine whether additional optimization yields attenuated gain, saturation, or regression. At the currently observed scales, Delta-OPD converts token-mean KL into local gold gain more rapidly than Vanilla-OPD in 15 of 17 shared cells and improves peak gain in 12. Additional seeds are needed to determine whether this descriptive advantage is stable.

Toward scaling beyond human supervision. Yang et al. (2026d) outline how large reasoning models may continue to improve as human supervision recedes from the learning loop. Our results bear on the supervision side of this program: a compact RL expert transfers capability to much larger students at rates and peaks predictable from scale, and the returns to teacher scaling taper off near the student's own scale. Within OPD, supervision competence therefore need not grow with the policy it trains. Feature-learning theory supports this picture: weak-to-strong training can succeed by eliciting knowledge already latent in the strong model (Awano & Suzuki, 2026). If this regularity persists at frontier scale, weak-to-strong OPD is a candidate mechanism for eliciting capability beyond the supervisor's own level, with human-level supervision as the fixed weak teacher.

B Limitations

Scope. The evidence is limited to one pretrained family (Qwen2.5, 0.5B–14B), one math reasoning mixture, one canonical run per teacher–student pair, responses of at most 2K tokens, and intentionally stopped trajectories. Evaluation records are aggregate rather than per prompt, so the present figures have neither prompt-bootstrap nor optimization-seed intervals, and cell-bootstrap intervals quantify variation across scale pairs but not across training seeds. Whether the fitted coefficients transfer to other families, tasks, or post-training recipes is untested.

Underspecified variables. Several recipe constants are held fixed and may matter. Rollout length is capped at 2K tokens, although prior work shows that rollout length affects OPD training stability (Li et al., 2026b), and results may also depend on the shared SFT recipe, the teacher RL checkpoint, the KL estimator, validation frequency, and the same-step logging convention. The token-mean coordinate weights tokens rather than responses; neither archive contains full-sequence k_3, so estimator robustness remains untested. Within the RL-endpoint grid, teacher scale and teacher competence vary together; the bootstrapped chains break this collinearity, but with only five auxiliary Vanilla-OPD cells and three Delta-OPD cells, whose slopes come from more sparsely logged evaluations. One matched-accuracy comparison distills the 7B student from an intermediate 3B teacher checkpoint and confirms the joint law's ordering (§H); repeating such controlled comparisons across the grid remains future work.

Mechanism and theory. The local expansion in §D explains why initial transfer is linear in d, but only its leading exponent. The extent of the linear window, the heterogeneous tails that follow it, and the power-law forms of the fitted coefficients remain empirical regularities without a mechanism. The Delta-OPD endpoint law has only five observed departures, the transfer-rate power law does not outperform the raw-affine baseline under leave-one-scale-out RMSE, and the fixed 30- and 40-observation windows are exploratory.

Early termination and finite evaluation frequency can right-censor d_{transfer} and place d_{peak} at the observation boundary. Multiple seeds, denser evaluation, and prompt-level resampling would quantify both sources of uncertainty.

C Extended related work

OPD variants. A large number of OPD variants have emerged recently (Song & Zheng, 2026; Liu et al., 2026). Generalized Knowledge Distillation trains on student-generated sequences and supports multiple divergences (Agarwal et al., 2024). ExOPD changes the relative reward and KL

weights (Yang et al., 2026c), while relaxed OPD interprets the log-likelihood ratio as a token reward and introduces clipping and sampling controls for instability (Ko et al., 2026). W2S-OPD, Direct-OPD, and OPD^2 independently use contrasts between teacher checkpoints to isolate a learned policy shift (Yu et al., 2026; Feng et al., 2026; Heo et al., 2026). OPRD amplifies the component of the student's verifier-driven policy gradient along the teacher-shift direction, rescaling only verifier-supported updates so that teacher guidance accelerates rather than redirects the student's own optimization. It reaches higher performance with fewer student updates than RL and distillation baselines in successor-generation transfer and multi-teacher consolidation (Park et al., 2026). OPD has also been combined with RL on verifiable rewards, coupling dense token-level distillation with outcome-level correctness signals, using either a separate teacher (Lin et al., 2026; Wang et al., 2026a) or the policy itself given hindsight access to verified solutions (Cai et al., 2026; Yang et al., 2026a; Li et al., 2026a). Empirical analyses further examine OPD training dynamics, failure modes, and recipes (Li et al., 2026b; Fu et al., 2026). We use the contrast family of objectives as an experimental condition, which we call Delta-OPD (Equation (3)), and compare methods at matched KL rather than matched optimizer step.

Proxy overoptimization and reward quality. The closest empirical precedent is the gold–proxy scaling study of Gao et al. (2023), which separates RL and best-of-n functional forms and relates their coefficients to reward-model scale, data, policy scale, and KL regularization. Direct-alignment work finds related overoptimization laws without an explicit RL loop (Rafailov et al., 2024), and reward hacking in large models has been surveyed systematically (Wang et al., 2026b). Complementary studies ask which reward-model properties yield useful optimization signals (Razin et al., 2025), how preference-data scale changes RLHF (Shen et al., 2025), and how inference-time search can exploit reward models (Khalaf et al., 2025). These results motivate the proxy diagnostics used to interpret regressing OPD trajectories.

Offline teacher hacking versus online distillation. Tiapkin et al. (2025) define teacher hacking in a controlled oracle–teacher–student hierarchy: repeated fixed teacher data can degrade ground-truth fit, while online generation and data diversity mitigate the effect. This differs from our on-policy rollouts, but supplies an important mechanism-level control. Sriraman et al. (2026) study a noisy-expert model and prove a separation between offline imitation and online interaction. Their findings motivate OPD under imperfect supervision; they do not imply the empirical scale or shape of capability-transfer dynamics.

Weak-to-strong generalization. Burns et al. (2024) establish the empirical weak-to-strong problem and show that strong students can outperform weak supervisors. In a tractable high-dimensional regression model, Ildiz et al. (2024) derive precise risk asymptotics and show that weak supervision can improve finite-data performance without improving the data-scaling exponent. Awano & Suzuki (2026) analyze two-layer networks in the feature-learning regime and show that weak-to-strong fine-tuning elicits features latent in the strong model's pretrained knowledge while avoiding the forgetting induced by standard supervised fine-tuning. Yuan et al. (2026) show that off-policy fine-tuning on chain-of-thought traces from much weaker reasoners recovers a large fraction of the gains of RL. Jin et al. (2026) use the weak model as a critic for scalable oversight, distilling revisions of the strong model's own outputs guided by the weak critiques. We study elicitation through on-policy token-level scoring and find that peak gold score improves with teacher scale only up to roughly the student's own scale, captured by the effective-teacher cap in the peak law.

Weakly supervised learning. Weak-to-strong learning can be viewed as an instance of weakly supervised learning, the long-studied setting where training signals are incomplete, inexact, or inaccurate (Zhou, 2018), including programmatic supervision that aggregates noisy labeling sources (Ratner et al., 2017; Bach et al., 2017). Learning from noisy labels is the closest classical problem (Song et al., 2023), but a weak teacher's mistakes concentrate on the instances it lacks competence for—the instance-dependent regime that is substantially harder than uniform label noise (Frénay & Verleysen, 2014)—so standard noise-correction techniques do not directly apply. The setting also differs from semi-supervised learning, where ground-truth labels are available for a subset of the training data (Chapelle et al., 2006): here no ground-truth reasoning supervision exists at the student's level, and gold labels serve only for held-out evaluation. Within this family, weak-to-strong OPD is distinctive in that the weak signal supervises tokens of the student's own rollouts rather than static labels or demonstration trajectories, so supervision quality evolves with the student policy.

Scaling laws for LLM training. Foundational neural scaling laws model pretraining loss as a power law of model size, data, and compute, enabling compute-allocation predictions (Kaplan et al., 2020). Hoffmann et al. (2022) systematically study compute-optimal scaling laws for LLM pretraining, finding that model size and data scale should be scaled equally. Unified functional forms extend such fits across architectures, tasks, and resource axes (Caballero et al., 2026), and Weng (2026) reviews how seemingly minor fitting choices can change extrapolated predictions. Moving from pretraining to adaptation, Zhang et al. (2024) find a multiplicative power-law relationship between fine-tuning performance and the amount of fine-tuning data together with model size, pretraining-data size, or the number of trainable parameters. Their comparison of full-model and parameter-efficient fine-tuning further shows that the preferred method depends on the task and data regime. Busbridge et al. (2025) fit compute-allocation laws for pretraining distillation, including cases with an existing teacher and cases where teacher training is charged to the budget. Lu & Liu (2026) further find that small or undertrained teachers improve larger pretraining students while stronger teachers saturate or reverse the gains, an offline analogue of the effective-teacher cap in our peak law. These results predict performance across resource and method configurations, rather than along a single KL-indexed optimization trajectory.

Post-training studies predict still different objects. Gao et al. (2023) model gold reward along a KL-indexed optimization trajectory and find optimization-specific functional forms whose coefficients vary smoothly with reward-model scale. Rafailov et al. (2024) extend this perspective to direct-alignment objectives, where KL-indexed degradation can arise without a separately trained proxy reward model. By contrast, Shen et al. (2026) connect the pretrained state to returns from subsequent RL: pretraining loss predicts performance at fixed RL compute, and early RL improvement rates vary systematically with pretraining data.

Khatri et al. (2026) instead fit sigmoidal performance–compute curves for LLM RL and distinguish changes in asymptotic performance from changes in compute efficiency across training recipes. Their extrapolation from smaller runs makes RL performance predictable in a compute coordinate, whereas our trajectories use token-mean KL as the amount of optimization. Collectively, these scaling properties concern different response variables and optimization coordinates; their functional forms should therefore not be transferred directly across settings.

Our targets are how teacher scale, student scale, and the OPD objective shape the training dynamics: how much capability transfers, and how fast. The local trajectory fits define these response variables; leave-one-scale-out prediction determines whether their cross-pair relationships constitute scaling laws. This standard follows Choshen et al. (2025), who show that checkpoint selection, model-size proximity, architecture, and seed variation materially affect scaling estimates. We therefore use leave-one-scale-out pair splits, checkpoint-level prediction, simple baselines, architecture controls, and seed-aware uncertainty. We report cross-objective results at matched KL rather than treating them as matched-compute or algorithm-independent efficiency comparisons, since objectives can accumulate KL at different rates for the same compute budget.

D Local geometry of capability transfer

This section derives Equation (8) for the population counterpart of the token-mean estimator in Equation (5). The result explains the leading exponent for both objectives; it does not claim that the empirical fitting windows are asymptotically small.

D.1 Statement

For any differentiable initial path with update direction h and nonzero token-Fisher norm, local KL geometry (Schulman et al., 2015) gives

$$G(d) = G_0 + md + O(d^2), \qquad m = \sqrt{2}\,\frac{g_0^\top h}{\sqrt{h^\top F_{\mathrm{tok}} h}}, \tag{8}$$

where $g_0 = \nabla_\theta G(\theta_0)$ and F_{tok} is the token-weighted Fisher matrix. Gold score changes to first order in the perturbation while KL changes to second order, so both objectives share the leading exponent and their slopes differ through the Fisher-normalized alignment of h with g_0. How far linearity persists, and the ensuing d_{transfer} and d_{peak}, are empirical properties.

D.2 A LOCAL SQUARE-ROOT-KL LAW

Let θ_0 parameterize π_{ref}, and define the reference score $s_0(x, y_{<t}, y_t) = \nabla_\theta \log \pi_\theta(y_t \mid y_{<t}, x)|_{\theta_0}$. The population divergence corresponding to token-mean aggregation is

$$D_{\text{tok}}(\theta) = \frac{\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\left[\sum_{t=1}^{|y|} \mathrm{KL}(\pi_\theta(\cdot \mid x, y_{<t}) \| \pi_{\text{ref}}(\cdot \mid x, y_{<t}))\right]}{\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}[|y|]}. \tag{9}$$

Indeed, for $\delta_t = \log \pi_{\text{ref}}(y_t \mid x, y_{<t}) - \log \pi_\theta(y_t \mid x, y_{<t})$,

$$\mathbb{E}_{y_t\sim\pi_\theta(\cdot|y_{<t},x)}[e^{\delta_t} - \delta_t - 1] = \mathrm{KL}(\pi_\theta \| \pi_{\text{ref}}), \tag{10}$$

because $\mathbb{E}_{\pi_\theta} e^{\delta_t} = 1$. Thus Equation (9) is the population quantity estimated by the logged token-mean k_3 statistic.

Consider a one-sided, differentiable training path,

$$\theta(\tau) = \theta_0 + \tau h + O(\tau^2), \qquad \tau \geq 0. \tag{11}$$

Define the token-weighted Fisher matrix

$$F_{\text{tok}} = \frac{\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\text{ref}}(\cdot|x)}\left[\sum_{t=1}^{|y|} s_0(x, y_{<t}, y_t) s_0(x, y_{<t}, y_t)^\top\right]}{\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\text{ref}}(\cdot|x)}[|y|]}. \tag{12}$$

The conditional KL is zero at θ_0, its first derivative is zero, and its Hessian is the conditional Fisher matrix (Schulman et al., 2015). Consequently,

$$D_{\text{tok}}(\theta(\tau)) = \frac{\tau^2}{2} h^\top F_{\text{tok}} h + O(\tau^3). \tag{13}$$

The on-policy prefix distribution also changes with θ, but this drift contributes only $O(\tau^3)$: its $O(\tau)$ change multiplies a conditional KL that is already $O(\tau^2)$.

Let $G(\theta)$ be expected sampled mean@1 accuracy, rather than its finite validation-set estimate, and write $g_0 = \nabla_\theta G(\theta_0)$. Smoothness gives

$$G(\theta(\tau)) = G_0 + \tau g_0^\top h + O(\tau^2). \tag{14}$$

If $h^\top F_{\text{tok}} h > 0$, then $d = \tau\sqrt{h^\top F_{\text{tok}} h/2} + O(\tau^2)$. Inverting this relation and substituting it into Equation (14) proves Equation (8). Positive initial transfer is equivalent to $g_0^\top h > 0$. If this inner product vanishes, the linear coefficient is zero and the leading gold-score change can instead be quadratic in d.

D.3 METHOD-SPECIFIC DIRECTIONS AND SLOPES

At initialization, define the immediate token signals

$$\begin{aligned} r_{\text{V}}(x, y_{<t}, y_t) &= \log \pi_{\text{T}}(y_t \mid y_{<t}, x) - \log \pi_{\text{ref}}(y_t \mid y_{<t}, x), \\ r_{\Delta}(x, y_{<t}, y_t) &= \log \pi_{\text{T}}(y_t \mid y_{<t}, x) - \log \pi_{\text{T}}^{\text{base}}(y_t \mid y_{<t}, x). \end{aligned} \tag{15}$$

Under the actor-update convention that sampled prefixes are held fixed within an update, the population policy-gradient signal is

$$b = \frac{\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\text{ref}}(\cdot|x)}\left[\sum_{t=1}^{|y|} s_0(x, y_{<t}, y_t) r(x, y_{<t}, y_t)\right]}{\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\text{ref}}(\cdot|x)}[|y|]}, \qquad h = B_0 b, \tag{16}$$

where B_0 represents the optimizer's local preconditioning. The reference-KL penalty has zero gradient at θ_0, so it does not alter this initial tangent. Setting $r = r_{\text{V}}$ or $r = r_{\Delta}$ selects the Vanilla-OPD or Delta-OPD direction h without changing the leading power of d. By Equation (8), the local slope measures the Fisher-normalized alignment of this direction with the gold gradient, rather than the magnitude of the token reward.

Table 4: Hyperparameters for Qwen2.5 SFT models.

Hyperparameter	Qwen2.5-0.5B	Qwen2.5-1.5B	Qwen2.5-3B	Qwen2.5-7B	Qwen2.5-14B
Learning rate	2.00e-5	1.52e-5	1.28e-5	1.03e-5	8.70e-6
Batch size	200				
Epochs	2				
Samples per epoch	300K				
Learning rate schedule	Linear decay to 1/10 maximum LR				
Warmup ratio	3%				
Max response length	2048				
Optimizer	AdamW				
Weight decay	0.1				

Table 5: Hyperparameters for Qwen2.5 RL teacher models.

Hyperparameter	Value
Learning rate	1e-6
Learning rate schedule	Constant
Global batch size	256
Global mini-batch size	64
GRPO group size	8
Epochs	10
Rollout temperature	1.0
Rollout top-p	1.0
Rollout top-k	-1
Max rollout length	2048
KL penalty strength	0.0
Optimizer	AdamW
Weight decay	0.01

D.4 Assumptions and scope

The derivation assumes common support and differentiable log probabilities; a locally differentiable generated-prefix distribution with finite moments; a differentiable initial optimizer path with nonzero Fisher norm; and a smooth expected gold utility. Relating the asymptotic result to the fitted windows additionally requires rollout-policy lag, clipping, response-length changes, the Fisher geometry, and the normalized gold alignment to vary slowly. These assumptions are plausible for small early updates but are empirically testable rather than guaranteed. The theorem characterizes the leading term as $d \to 0$. Global shape, peak formation, estimator choice, and cross-method slope ordering remain empirical properties.

E Hyperparameters

All SFT, RL, and OPD training runs are implemented on verl v0.8.0 (Sheng et al., 2025).

E.1 SFT hyperparameters

For SFT, the models share the majority of hyperparameters except for the learning rate (Table 4). The learning rate scales as the inverse fourth root of parameter count (N):

$$\mathrm{LR}(N) = 2 \times 10^{-5} \left(\frac{N}{0.5\mathrm{B}} \right)^{-1/4}. \tag{17}$$

E.2 RL hyperparameters

All RL teachers share the hyperparameters in Table 5.

Table 6: Hyperparameters for OPD on Qwen2.5 models.

Hyperparameter	Value
Learning rate	1e-6
Learning rate schedule	Constant
Batch size	256
Epochs	$\leq$10
Max rollout length	2048
Optimizer	AdamW
Weight decay	0.01

Table 7: Hyperparameters for evaluation.

Hyperparameter	Value
Temperature	1.0
Top-p	1.0
Top-k	-1
Max rollout length	2048

E.3 OPD HYPERPARAMETERS

All OPD runs share the hyperparameters in Table 6. The intermediate-teacher validation run (§H) reuses this recipe for the 7B student and changes only the teacher checkpoint.

E.4 EVALUATION HYPERPARAMETERS

Table 7 lists the sampling configuration shared by all gold-score evaluations.

E.5 RANDOM SEEDS

Every SFT, RL, and OPD run in this study uses a single random seed, and we do not repeat any configuration across seeds. The primary reason is compute. The study spans five model scales with 25 Vanilla-OPD cells, 17 Delta-OPD cells, five RL teachers, and the bootstrapping and on-policy-supervision conditions, so each additional seed would multiply the training and evaluation cost of the entire grid. Single runs per configuration are also standard in scaling-law studies, whose statistical support comes from regularity across many configurations rather than from per-configuration replication. Kaplan et al. (2020) train one model per configuration and estimate seed-level loss variation at roughly 0.02 nats, small against their fitted trends; Hoffmann et al. (2022) fit compute-optimal laws on over 400 single-run models; and Busbridge et al. (2025) report one distillation run per teacher–student configuration. Our preliminary trial reruns show OPD trajectories to be more stable than RL trajectories, and the cell-bootstrap intervals reported with the fitted laws quantify cross-pair variation rather than seed uncertainty (§B).

F QWEN2.5 MATH TRAJECTORY GRID

Figure 9 reports every canonical pair with the same alignment as the main text. Vanilla-OPD fit windows use the first 30 observations. Figure 10 reports the corresponding grid for the 15 weak-to-strong and same-base Delta-OPD runs, whose fit window is 40 observations (§6.1); the two strong-to-weak controls are omitted.

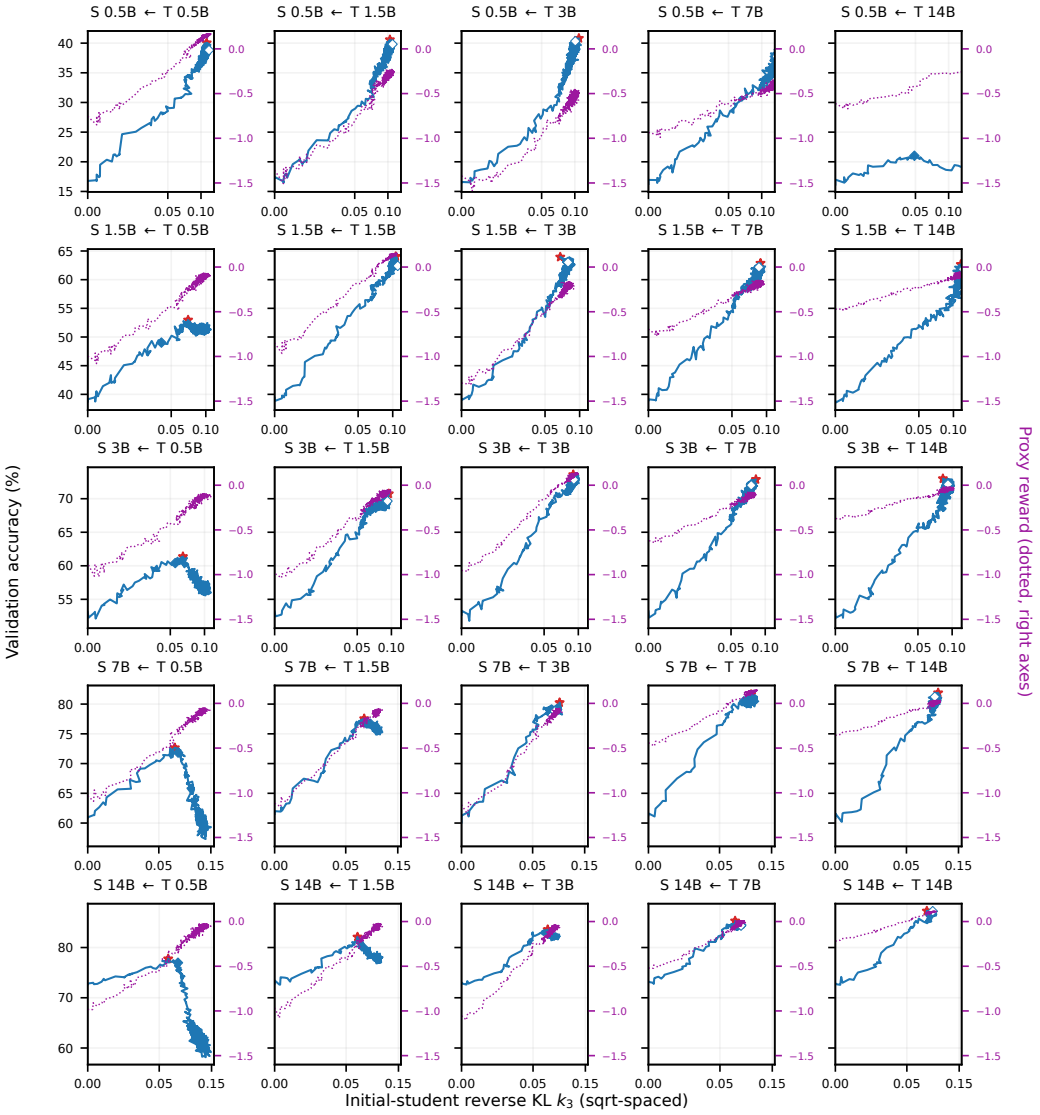

Figure 9: Qwen2.5 mathematics trajectory grid with token-mean-KL ticks on square-root-spaced axes. Diamonds mark transfer endpoints (hollow when right-censored), and stars mark global gold-score maxima. Dotted purple curves show the teacher-induced proxy reward on right-hand axes for the pairs where it was logged. Rows share the student scale and columns share the teacher scale; panels within a row share axis ranges.

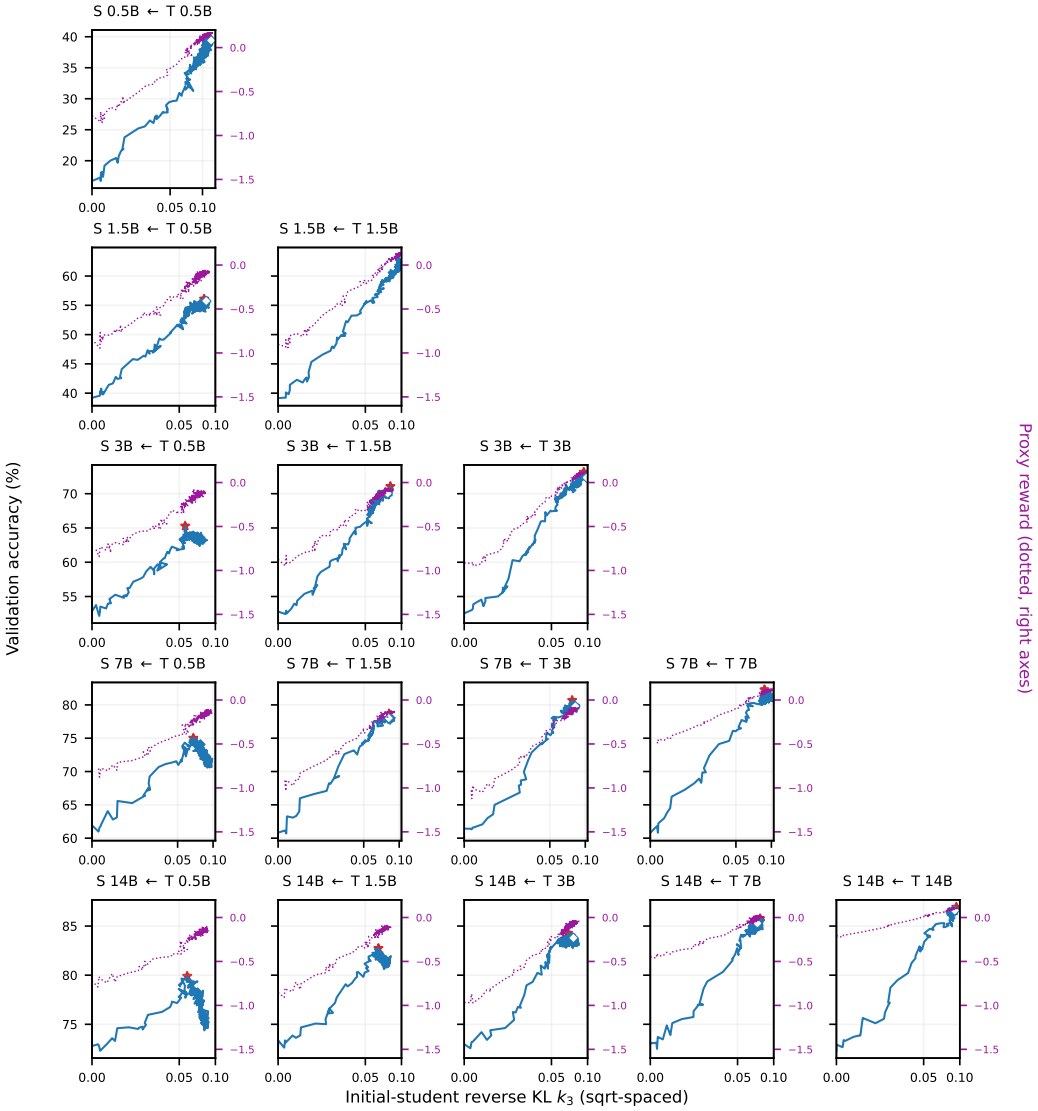

Figure 10: Delta-OPD trajectory grid over weak-to-strong and same-base cells with token-mean-KL ticks on square-root-spaced axes. Diamonds mark transfer endpoints (hollow when right-censored), and stars mark global gold-score maxima. Dotted purple curves show the token-mean Delta-OPD reward on right-hand axes, with the same ticks as Figure 9. Rows share the student scale and columns share the teacher scale; panels within a row share axis ranges.

G EXPANDED DELTA-OPD AND RL RESULTS

Complete Delta-OPD comparison. Table 8 gives the per-pair statistics summarized in the main text. The five observed transfer endpoints occur in four weak-to-strong cells and one same-base cell; the remaining twelve endpoints are right-censored.

Direct-RL trajectories. Figure 11 shows the teacher-construction runs used for the final-checkpoint reference in Figure 4. Four-point local fits obtain $R^2 \in [0.991, 0.998]$, while extrapolation RMSE over later checkpoints ranges from 0.016 to 0.041 accuracy units. The growing error reflects diminishing marginal improvement along the direct-RL trajectory.

Table 8: Per-pair Delta-OPD results. Peak gains are percentage points relative to each run's own initialization; endpoint status is observed (Obs.) or right-censored (Cens.).

Pair	Relation	m_{V}	m_{Δ}	R^2_{Δ}	Gain V	Gain Δ	Endpoint
0.5B←0.5B	Same-base	.573	.557	.985	23.3	22.6	Cens.
0.5B←1.5B	S2W	.539	.570	.984	23.3	23.6	Cens.
1.5B←0.5B	W2S	.564	.554	.984	14.2	16.8	Cens.
1.5B←1.5B	Same-base	.721	.744	.981	24.6	24.6	Cens.
1.5B←3B	S2W	.629	.723	.973	24.8	24.6	Cens.
3B←0.5B	W2S	.378	.444	.963	8.2	12.4	Obs.
3B←1.5B	W2S	.555	.642	.973	17.6	18.2	Cens.
3B←3B	Same-base	.618	.711	.944	20.9	20.5	Cens.
7B←0.5B	W2S	.415	.506	.967	11.7	13.1	Obs.
7B←1.5B	W2S	.558	.630	.982	16.0	17.8	Cens.
7B←3B	W2S	.670	.754	.964	20.2	19.3	Cens.
7B←7B	Same-base	.687	.712	.987	20.3	21.5	Obs.
14B←0.5B	W2S	.175	.268	.916	5.1	6.7	Obs.
14B←1.5B	W2S	.327	.387	.961	9.3	9.9	Obs.
14B←3B	W2S	.417	.470	.942	10.6	11.4	Cens.
14B←7B	W2S	.424	.470	.971	12.1	12.6	Cens.
14B←14B	Same-base	.498	.515	.964	14.5	14.6	Cens.

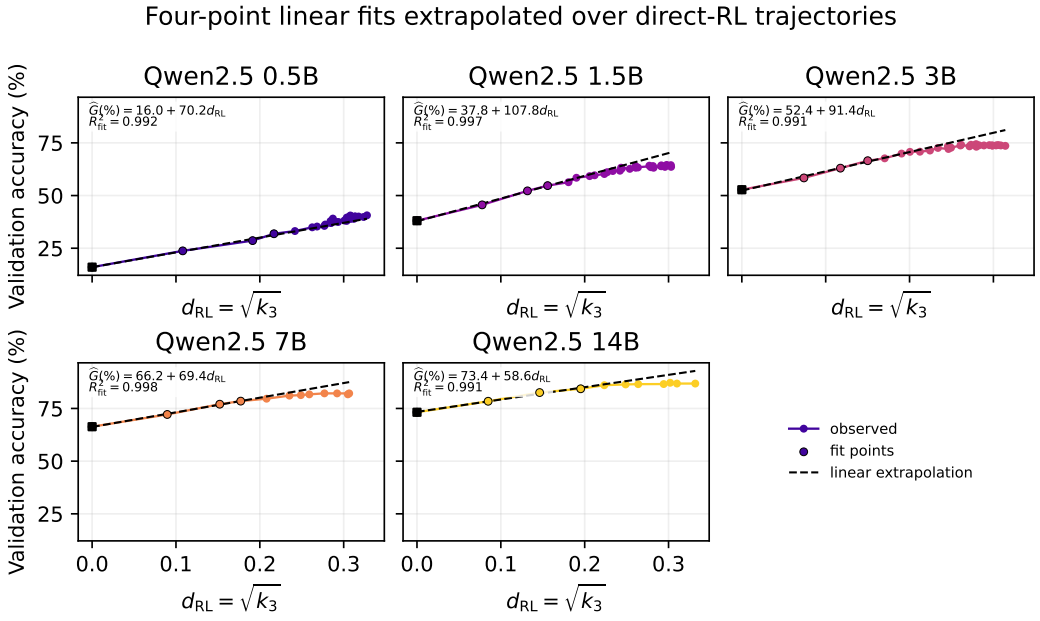

Figure 11: Direct-RL gold score against token-mean reverse KL. Each panel emphasizes the first four chronological observations, annotates the free-intercept local fit, and extrapolates it across the remaining trajectory. The black square marks the SFT initialization.

H Scaling law estimation and diagnostics

This appendix details the fitting protocols and diagnostics behind §5. Figure 12 shows the Delta-OPD counterpart of the Vanilla-OPD diagnostics in Figure 5.

SFT baseline and peak capability. We average the Vanilla-OPD step-zero accuracies within each student scale to obtain $G_{\mathrm{SFT}}(N_S)$ and fit the student-only remaining-error law

$$1 - G_{\mathrm{SFT}} = 0.682\, \widetilde{N}_S^{-0.326}. \tag{18}$$

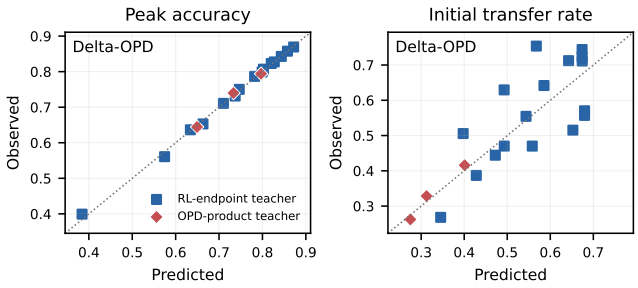

Figure 12: Observed Delta-OPD peak gold score and initial transfer rate against full-fit predictions of the joint laws of Equation (6), complementing the Vanilla-OPD diagnostics of Figure 5. Red diamonds mark cells whose teachers are OPD products of the bootstrapped chains (§6.3).

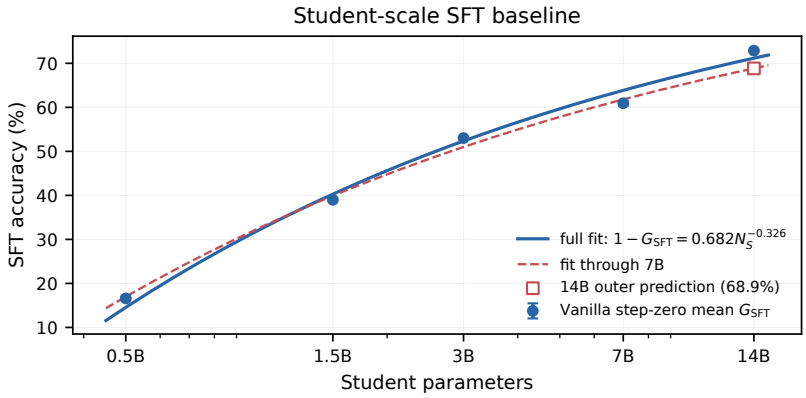

Figure 13: Student-scale SFT baseline. Points show equally weighted Vanilla step-zero means with within-scale standard deviations. The solid curve is the five-scale remaining-error fit in Equation (18); the dashed curve is refit through 7B and its open square is the 14B outer-scale prediction.

Each of the five scale means averages five Vanilla-OPD cells. Withholding the 14B mean gives a 4.0 percentage-point extrapolation error, the refit law underpredicting the observed 14B baseline. The vertical difference between this fitted baseline and a teacher-conditioned peak in Figure 4 visualizes capability added by OPD; the fitted means and outer-scale check are shown in Figure 13. We use this difference for interpretation and retain per-run peak gains for the matched-KL comparison, without fitting a separate gain law.

Model comparison. Peak-performance candidates use raw accuracy, logit accuracy, or log remaining error with student-only, uncapped two-scale, and capped two-scale covariates. Within the capped log-error family, we compare the multiplicative law in Equation (7), a separately weighted additive law $A_S \widetilde{N}_S^{-\alpha} + A_T (\widetilde{N}_T^{\mathrm{eff}})^{-\beta}$, and the equal-amplitude sensitivity $A[\widetilde{N}_S^{-\alpha} + (\widetilde{N}_T^{\mathrm{eff}})^{-\beta}]$. The multiplicative law has the lowest mean leave-one-scale-out error and AICc for both methods (Table 9). Figure 14 overlays the multiplicative and weighted-additive fits on the observed peak series.

Table 9: Peak-error functional-form comparison. S/T are leave-one-scale-out RMSE over student and teacher scales, in percentage points.

Method	Peak-error law	S/T RMSE	AICc
Vanilla	Multiplicative	1.81 / 1.60	−155.6
	Weighted additive	1.97 / 1.76	−140.8
	Equal additive	4.02 / 2.67	−100.0
Delta	Multiplicative	.91 / .85	−108.3
	Weighted additive	1.66 / 1.26	−88.7
	Equal additive	3.99 / 3.04	−53.9

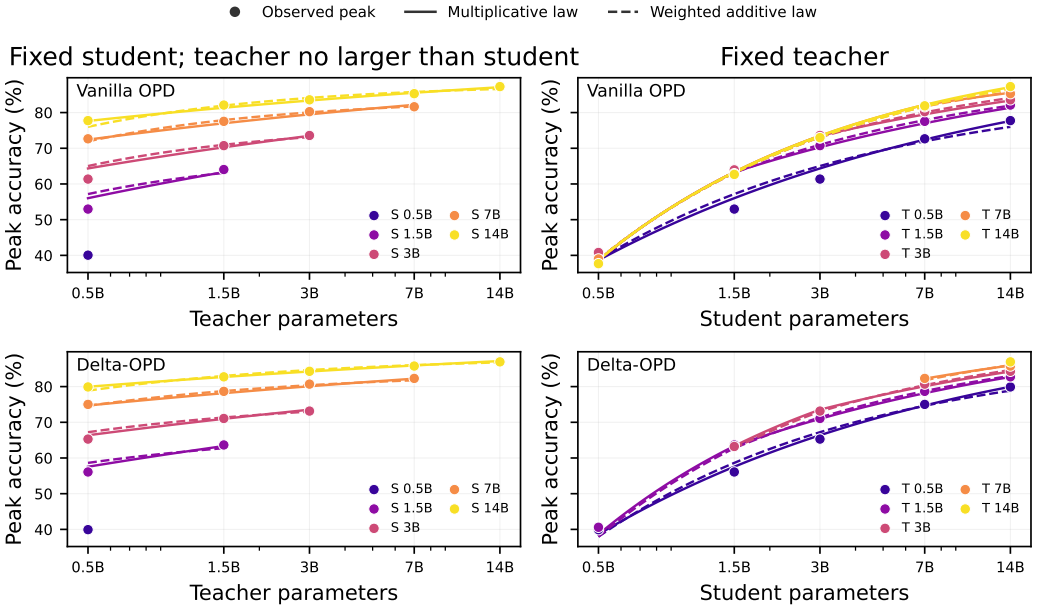

Figure 14: Peak-performance scaling under the multiplicative and weighted-additive laws. Rows separate Vanilla-OPD and Delta-OPD. The left column fixes student scale and retains teachers no larger than the student; the right column fixes teacher scale. Points are observed peaks, solid curves are multiplicative-law fits, and dashed curves are weighted-additive fits over each series' support.

Teacher size versus teacher performance. Parameter count is one of two natural teacher covariates; the teacher's own held-out gold accuracy G_T is the other. For a fair single-covariate comparison we replace $\widetilde{N}_T^{\text{eff}}$ in Equation (7) by the error of the same capped teacher, and fit

$$1 - G_{\text{peak}} = A\widetilde{N}_S^{-\alpha}(1 - G_T^{\text{eff}})^{\zeta}. \tag{19}$$

Within the RL-endpoint grid alone the two teacher covariates are collinear at $r = -0.9999$, so the size and performance laws cannot be separated there. The teacher-variable fits therefore add the bootstrap-chain cells of §6.3, whose teachers are OPD products with measured gold scores below the size trend: five auxiliary Vanilla-OPD cells and three Delta-OPD cells. Each chain stage also yields an initial slope, fitted on its method's standard initial window; chain evaluations are logged more sparsely than the grid's, so these windows span a longer d-range while remaining inside the useful-transfer regime. The added teachers reduce the collinearity to $r = -0.949$ for Vanilla-OPD and $r = -0.973$ for Delta-OPD (Figure 15, right). For the peak target the joint law in Equation (6) becomes identifiable for Vanilla-OPD, $\zeta = 0.952$ with 95% interval $[0.90, 1.04]$ and $\beta = -0.267$ with interval $[-0.31, -0.23]$, and its leave-one-scale-out RMSE improves to 1.66 points from 2.55 (score only) and 3.43 (size only) (Table 10). Peak student error is thus nearly proportional to the capped teacher's remaining error, and at matched score the smaller teacher transfers better. The three Delta-OPD chain cells identify the peak law as well: $\zeta = 1.009$ with interval $[0.90, 1.11]$ and $\beta = -0.325$ with interval $[-0.37, -0.28]$, with leave-one-scale-out RMSE of 0.82 points from 2.07

Table 10: Teacher-size, teacher-performance, and joint laws for the peak and rate targets, fitted on the grid cells plus the auxiliary bootstrap-teacher cells (five for Vanilla-OPD, three for Delta-OPD). Leave-one-scale-out RMSE is the mean of the student-scale and teacher-scale hold-out splits, in accuracy points for the peak target and in points per unit d for the rate target. The teacher exponent acts on $\widetilde{N}_T^{\text{eff}}$ for the size law and on $1 - G_T^{\text{eff}}$ for the performance law; the joint law fits both, with 95% bootstrap intervals in Table 14.

Method	Target	Teacher law	Student exp.	Teacher exponents	Hold-out RMSE	AICc	R^2
Vanilla-OPD	Peak	Size	0.278	$\beta = 0.147$	3.43	-101.2	0.915
		Perf.	0.269	$\zeta = 0.425$	2.55	-129.7	0.967
		Joint	0.304	$\beta = -0.267,\ \zeta = 0.952$	1.66	-190.2	0.996
	Rate	Size	0.240	$\delta = 0.217$	19.1	-13.0	0.243
		Perf.	0.269	$\xi = 0.697$	17.0	-21.3	0.427
		Joint	0.189	$\delta = -0.609,\ \xi = 1.896$	16.3	-28.2	0.587
Delta-OPD	Peak	Size	0.332	$\beta = 0.105$	2.47	-69.3	0.929
		Perf.	0.320	$\zeta = 0.303$	2.07	-80.3	0.961
		Joint	0.336	$\beta = -0.325,\ \zeta = 1.009$	0.82	-129.3	0.998
	Rate	Size	0.216	$\delta = 0.135$	16.0	-14.7	0.354
		Perf.	0.242	$\xi = 0.438$	15.2	-19.5	0.492
		Joint	0.203	$\delta = -0.725,\ \xi = 2.010$	14.1	-30.6	0.757

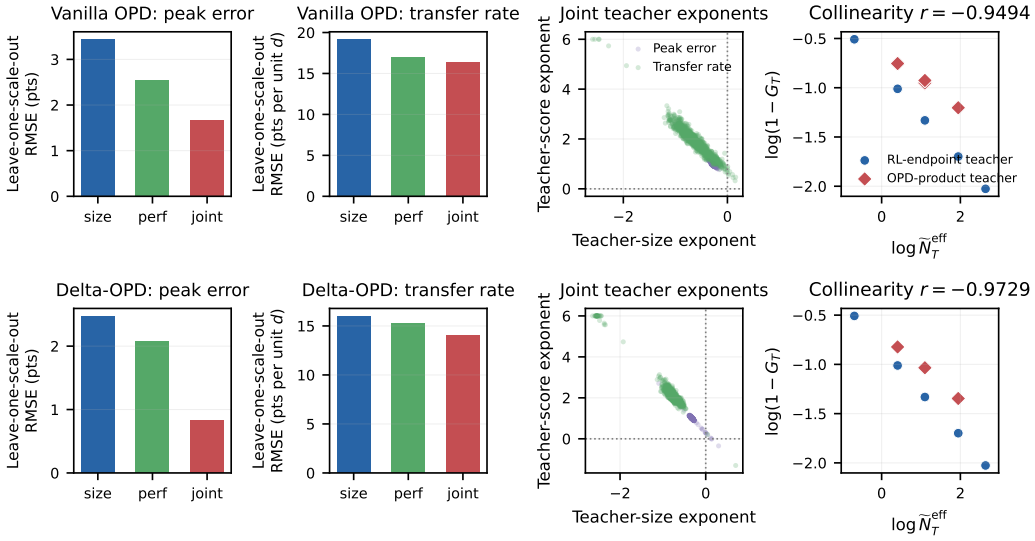

Figure 15: Teacher size versus teacher performance, fitted on the grid plus the auxiliary bootstrap-teacher cells; rows are Vanilla-OPD (top) and Delta-OPD (bottom). Left two columns: leave-one-scale-out RMSE of the size, performance, and joint laws for the peak and rate targets. Third column: bootstrap joint-law teacher exponents for both targets, concentrated away from zero on anticorrelated ridges. Right: capped teacher size against capped-teacher error; red diamonds mark OPD-product teachers, which sit off the RL-endpoint size trend and break the collinearity.

(score only) and 2.47 (size only). The rate target repeats the pattern with larger score exponents and noisier fits: $\xi = 1.90$ with interval $[0.93, 2.84]$ and $\delta = -0.61$ with interval $[-1.10, -0.10]$ for Vanilla-OPD, and $\xi = 2.01$ with interval $[1.62, 5.78]$ and $\delta = -0.73$ with interval $[-2.44, -0.54]$ for Delta-OPD, so a worse teacher slows transfer much more strongly than it limits the eventual peak.

Validation with an intermediate teacher checkpoint. The joint peak law predicts that when two teachers reach the same gold score, the larger one yields the worse student. We test this prediction out of sample by distilling the 7B student from an intermediate checkpoint of the 3B teacher's RL run, taken at step 58, the earliest saved epoch, whose directly evaluated gold score of 66.0 approxi-

Table 11: Out-of-sample validation of the Vanilla-OPD peak laws with an intermediate 3B teacher checkpoint, in gold-score points. Each row is a 7B-student cell; predictions use the full-fit coefficients of the scale-only (Equation (7)), score-only (Equation (19)), and joint (Equation (6)) laws. The intermediate-teacher cell is held out from every fit; only the joint law orders it correctly relative to the 1.5B endpoint teacher.

Teacher	Teacher score	Observed peak	Scales only	Score only	Joint
1.5B endpoint (step 580)	63.6	77.5	77.0	75.6	**77.1**
3B intermediate (step 58)	66.0	73.8	79.5	76.3	**74.1**
3B endpoint (step 580)	73.6	80.3	79.5	78.7	**79.6**

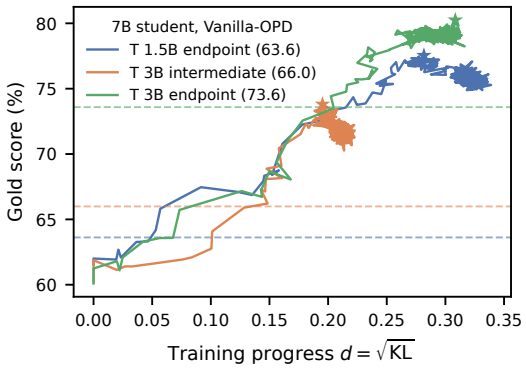

Figure 16: Trajectories of the intermediate-teacher validation. Gold score against d for the 7B student distilled from the 1.5B RL endpoint, the score-matched intermediate 3B checkpoint at step 58, and the 3B RL endpoint. Stars mark peaks and dashed levels each teacher's own gold score.

mately matches the 1.5B RL endpoint's 63.6. The run reuses the canonical Vanilla-OPD recipe (§E) and changes only the teacher checkpoint, and its cell enters no fit. Table 11 compares observed peaks with the full-fit predictions of the three Vanilla-OPD peak laws, and Figure 16 overlays the trajectories. The intermediate 3B teacher scores 2.4 points above the 1.5B endpoint yet its student peaks 3.7 points below (73.8 against 77.5). The joint law predicts the held-out cell within 0.3 points and reproduces the ordering, whereas the scale-only law predicts the opposite ordering and the score-only law also favors the higher-scoring intermediate teacher. The archived trajectory reaches update 430 of 580 and peaks at update 46, far before the truncation.

Outer-scale extrapolation. We refit the size-only peak laws and the joint teacher law after withholding the largest student scale and the largest observed teacher scale as separate tests. The largest-teacher scale is 14B for both methods. All models share each split's grid test cells; the joint teacher law additionally trains on the chain cells below the held-out scale. The size-only joint law improves every axis–method comparison over the additive law, and the joint teacher law extrapolates comparably, improving the Vanilla-OPD holdouts while staying within 0.03 points of the Delta-OPD ones (Table 12 and Figure 17). Bootstrap intervals resample training cells and measure fit stability rather than variation across training seeds. Nine replicates are non-identifiable and excluded; the other 11,991 model–split replicates converge and pass the rank check.

Table 12: Outer-scale peak extrapolation. Entries are accuracy-point RMSE with 95% training-cell-bootstrap intervals. Train counts refer to the size-only laws; the joint teacher law adds the chain cells below the held-out scale (four Vanilla-OPD and two Delta-OPD for the student holdout, five and three for the teacher holdout).

Holdout	Method	Train/Test	Joint	Weighted additive	Joint teacher
Largest student	Vanilla	20/5	.64 [.33, 2.27]	1.50 [.67, 3.60]	.55 [.28, 1.80]
	Delta	12/5	.21 [.12, .89]	.92 [.46, 2.35]	.20 [.11, .96]
Largest teacher	Vanilla	20/5	.75 [.32, 1.36]	1.06 [.67, 1.64]	.68 [.33, 1.25]
	Delta	16/1	.29 [.04, .72]	.37 [.02, 1.49]	.32 [.07, .68]

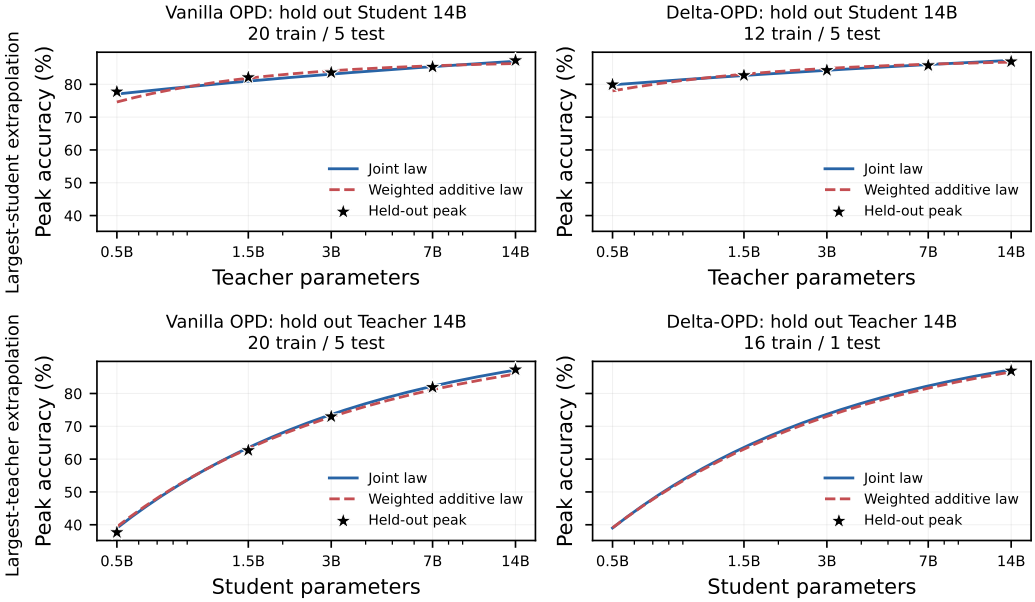

Figure 17: Peak-law extrapolation after fitting only smaller scales. Rows hold out the largest student or teacher scale; columns separate Vanilla-OPD and Delta-OPD. Stars are held-out peaks, solid curves are joint-law predictions, and dashed curves are weighted-additive predictions.

Direct accuracy-gain sensitivity. Direct, weighted-additive, and finite-horizon direct-RL-gap parameterizations of peak gain produced less consistent outer-scale extrapolation than the separate SFT and peak endpoints. We therefore retain observed per-run gains for descriptive matched-KL comparisons, but report no standalone gain law. The exploratory fits and derived predictions remain available in the analysis outputs for reproducibility.

Table 13 aggregates the validation summary of the joint teacher laws across the peak, rate, and extent targets. For transfer rate, the scale-only positive power law is also compared with a raw-affine model using identical covariates and held-out cells. The raw baseline retains a modest predictive advantage (leave-one-scale-out RMSE .186/.117 against .208/.126 for Vanilla-OPD and .181/.112 against .196/.116 for Delta-OPD), so the fitted exponents are read as an interpretable scale summary rather than the best predictor.

Endpoint candidates use log-normal accelerated-failure-time likelihoods with constant, student-only, and two-scale predictors, fitted to the observed and right-censored departures. Mean leave-one-scale-out log likelihood over student and teacher splits selects the capped two-scale model for Vanilla-OPD, whose exponents are bootstrap-identifiable with 997 of 1,000 stable samples, and the uncapped two-scale model for Delta-OPD, whose smaller exponents are not. The Delta-OPD fit remains exploratory: five events yield 775 stable cell-bootstrap samples out of 1,000. The selected

Table 13: Scale-law validation under the joint teacher laws of Equation (6): peak uses outer-scale holdout errors (S/T) in points; rate uses leave-one-scale-out RMSE (S/T) in accuracy per unit d.

Method	Peak S/T	Rate S/T	Endpoint model
Vanilla	.55 / .68	.203 / .124	two-scale $C = .37$
Delta	.20 / .32	.177 / .104	two-scale $C = .34$

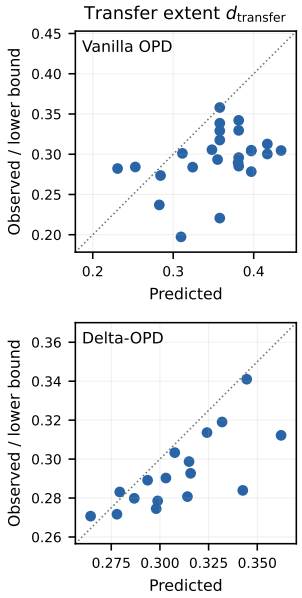

Figure 18: Transfer-extent diagnostics. Observed departures or censored lower bounds against the accelerated-failure-time fits of Equation (20) for Vanilla-OPD (top) and Delta-OPD (bottom).

laws are

$$\text{median}(d_{\text{transfer,V}}) = 0.37\,\widetilde{N}_S^{-0.13}(\widetilde{N}_T^{\text{eff}})^{0.19}, \qquad \text{median}(d_{\text{transfer},\Delta}) = 0.34\,\widetilde{N}_S^{-0.08}\widetilde{N}_T^{0.05}. \tag{20}$$

Both laws vary the median budget by less than a factor of two across the grid, shrinking slowly with student scale and growing slowly with teacher scale, and Figure 18 plots their predictions against observed or censored endpoints. Given these small and partly unidentifiable exponents, the main body reports the transfer extent as an approximately scale-free KL budget rather than a fitted target.

Sensitivity. Replacing $\widetilde{N}_T^{\text{eff}}$ by uncapped teacher size or omitting teacher scale worsens leave-one-scale-out peak prediction. Adding a learnable irreducible error to Equation (7) places that term at its zero lower bound for both methods, providing no evidence for an additional floor over the observed scale range. Endpoint-rule sensitivity remains governed by the confidence-band and persistence grid in Table 15.

Table 14: Power-law coefficients with 95% cell-bootstrap intervals. Peak and rate report the joint teacher laws of Equation (6); intervals quantify cross-pair variation rather than seed uncertainty. Extent rows report the median-budget laws of Equation (20), $\mathrm{median}(d_{\mathrm{transfer}}) = C\widetilde{N}_S^{-u}\widetilde{N}_T^{v}$, with the teacher scale capped or uncapped as marked.

Method	Target	Coefficients
Vanilla	Peak error	$A = .97\,[.93, 1.04],\ \alpha = .304\,[.286, .317],$ $\beta = -.267\,[-.311, -.229],\ \zeta = .952\,[.898, 1.035]$
Delta	Peak error	$A = 1.02\,[.93, 1.11],\ \alpha = .336\,[.327, .351],$ $\beta = -.325\,[-.372, -.276],\ \zeta = 1.009\,[.904, 1.114]$
SFT	Baseline	$A_{\mathrm{SFT}} = .682\,[.680, .683],\ \alpha_{\mathrm{SFT}} = .326\,[.324, .328]$
Vanilla	Rate	$B = .116\,[.050, .263],\ \gamma = .189\,[.011, .306],$ $\delta = -.61\,[-1.10, -.10],\ \xi = 1.90\,[.93, 2.84]$
Delta	Rate	$B = .129\,[.006, .181],\ \gamma = .203\,[.087, .304],$ $\delta = -.73\,[-2.44, -.54],\ \xi = 2.01\,[1.62, 5.78]$
Vanilla	Extent	$C = .373\,[.305, .480],\ u = .132\,[.030, .328],\ v = .190\,[.036, .601]$ (capped)
Delta	Extent	$C = .337\,[.309, .455],\ u = .080\,[.039, .106],\ v = .045\,[.018, .527]$ (uncapped)

I Endpoint and estimator sensitivity

Transfer-endpoint sensitivity. The primary endpoint uses a one-sided 95% predictive band and three consecutive lower deviations after the prespecified fit window. The machine-readable analysis records the endpoint checkpoint, KL coordinate, deviation start, and right-censoring status for every run. Confidence levels of 90%, 95%, and 99% crossed with persistence requirements of two, three, and four checkpoints produce the sensitivity grid in Table 15.

Alternate alignment, training seeds, and prompt bootstrap resamples provide additional validation axes. Held-out endpoint evaluation reports point error for observed departures and one-sided consistency for right-censored trajectories.

For the 17 Delta-OPD first-40 windows, no unconstrained quadratic has both statistically identifiable negative curvature and a meaningful in-support vertex under the prespecified condition-number and signal-to-noise checks. Their later checkpoints are therefore presented as observed trajectories.

Estimator sensitivity. The primary token-mean coordinate is defined in Equation (5). As a length-sensitive diagnostic, the archive also records the response mean of tokenwise k_3 sums,

$$k_{3,\mathrm{sum}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t}\left[\exp(\delta_{i,t}) - \delta_{i,t} - 1\right]. \tag{21}$$

Here, $\delta_{i,t}$ is the token log-probability difference already defined in Equation (5). This `actor/response_reverse_kl` scalar is not sequence-level KL: full-sequence k_3 would exponentiate $\delta_i = \sum_t \delta_{i,t}$ only after summation, and neither archive logs that quantity.

The sensitivity analysis compares token-mean and token-summed coordinates through correlations, curve overlays, endpoint changes, and local-fit coefficients. Full-sequence measurements remain a separate validation target because the archives lack that scalar.

Future estimator validation will compare full-sequence k_3 with sampled k_1 and partial full-vocabulary KL on common prompts and checkpoints. Abrupt failures will be diagnosed using update KL, gradient norm, entropy, and response length as a separate collapse category.

J Extended results for design choices

Bootstrap training dynamics. Figure 19 shows the full gold-score trajectory of every bootstrapped run in §6.3 against its own KL-indexed training progress. Most chains reach their peak early and then decline or saturate, mirroring the weak-to-strong dynamics of the direct grid.

Full on-policyness grid. Figure 20 expands Figure 7 to every weak-to-strong and same-base cell, one panel per student. The orderings reported in §6.2 hold across the grid.

Table 15: Endpoint-rule sensitivity: number of Vanilla-OPD trajectories with an observed departure from the initial line.

One-sided band	Two deviations	Three deviations	Four deviations
90%	12	9	9
95%	9	9	9
99%	8	8	8

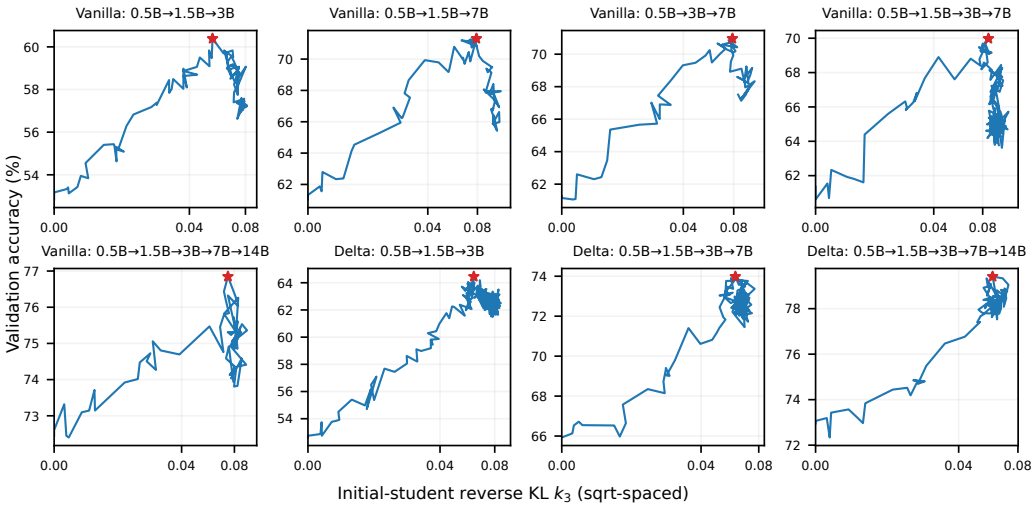

Figure 19: Gold-score dynamics of the five Vanilla-OPD and three Delta-OPD bootstrap chains. Stars mark the observed peak.

Cold-start dynamics. Figure 21 overlays OPD from the off-policy cold start with pure OPD on the shared cells of §6.2, each in the KL coordinate measured from its own initialization. In weak-to-strong cells the cold-started runs begin from a degraded initialization and track a lower trajectory throughout, so the peak deficit originates in the cold start rather than in the subsequent on-policy phase.

K Lessons learned and what did not work

Motivated by Yang et al. (2026b), this section records problems we encountered and approaches that did not work, together with the lessons we drew from them.

Differentiating the k_3 estimator gives forward-KL gradients. Delta-OPD regularizes the student toward its SFT reference with a differentiable reverse-KL term (Equation (3)), and our first implementation used the k_3 estimator (Schulman, 2020) for that term, as is common in RL frameworks. The k_3 estimate of the reverse-KL value is accurate, but its gradient with respect to the policy is the gradient of the forward KL, so training regularized the wrong divergence. We corrected this with the k_3^+ construction, which keeps the k_3 value estimate while backpropagating the k_2 gradient, matching the reverse-KL gradient in expectation. A KL estimator that enters the loss must therefore be validated at the gradient level, since value-level agreement is not sufficient.

Optimizer numerics can mask regression. Early weak-to-strong Vanilla-OPD runs on an FSDP training backend showed no post-peak regression, while the same recipes on a Megatron backend regressed reliably. The configurations differed in optimizer-state precision, BF16 under FSDP and FP32 under Megatron, and we suspect that the FP32 optimizer follows the small late-stage gradient signal more faithfully while BF16 rounding damps it. We therefore ran the study on a single backend with fixed numerics, and caution that late-stage dynamics can be sensitive to such details.

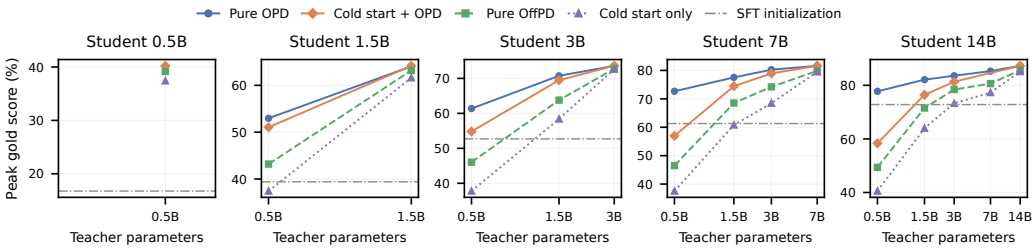

Figure 20: Peak gold score under three degrees of on-policy supervision, per student, over weak-to-strong and same-base cells. Dash-dotted levels mark SFT initializations.

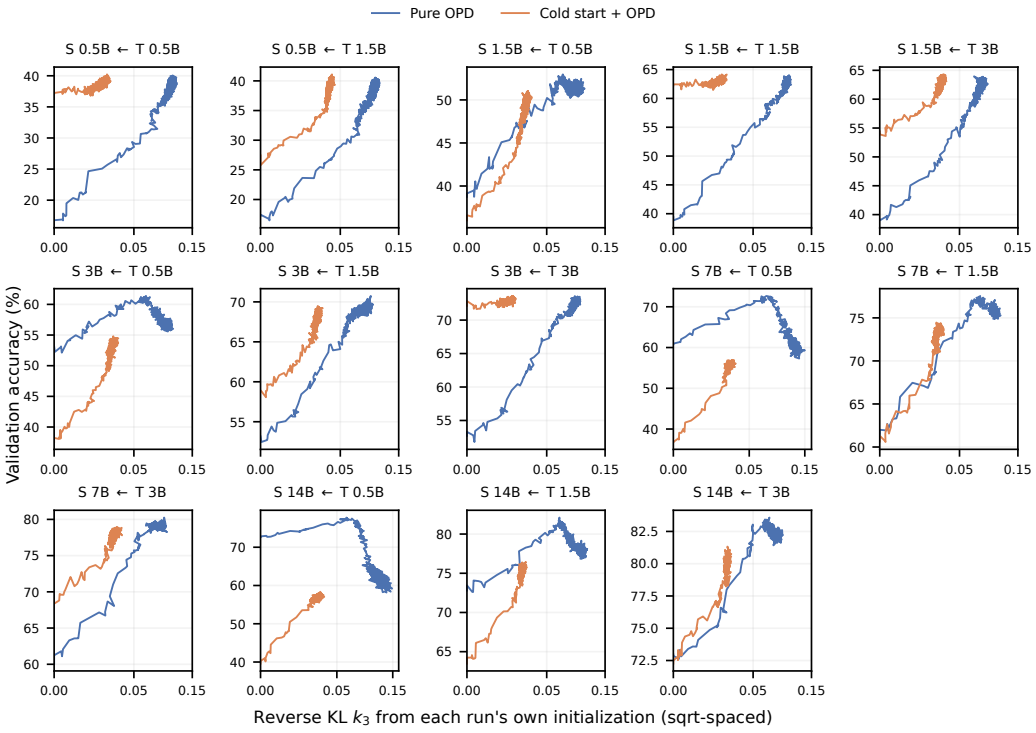

Figure 21: Cold-start OPD against pure OPD on the 14 shared cells. Each trajectory uses the reverse KL from its own initialization, so the two curves in a panel have different reference policies.

Sequence-level KL does not expose the regular regime. We first indexed trajectories by sequence-level reverse KL, following the coordinate of Gao et al. (2023). Under this coordinate the initial regime is not consistently regular across teacher–student pairs. Token-mean KL is also the principled choice for OPD: the immediate-token objective in Equation (2) carries no return-to-go, so the update controls the conditional next-token divergence rather than the joint sequence divergence (§3).

PPO and best-of-n functional forms do not transfer. We initially fitted the regular regime with the forms of Gao et al. (2023), the quadratic $ad - bd^2$ from best-of-n and the subquadratic $ad - bd \log d$ from RL. Across the grid the curvature terms are not statistically identifiable within the regular window, echoing the quadratic diagnostics in §I. The linear law in d describes the regular regime, and the heterogeneous tails outside it follow neither functional form.

Full-vocabulary loss accelerates weak-to-strong regression. For Vanilla-OPD we compared minimizing the full-vocabulary reverse KL as a differentiable loss against the sampled-token policy-gradient objective of §2. Under the full-vocabulary loss, post-peak regression in weak-to-strong pairs began earlier and was more severe. We therefore use the sampled-token form throughout.

No parametric law for the peak location. We attempted to model the location d_{peak} of the gold-score maximum as a function of student and teacher scale. The observed locations are non-monotone across the grid (§4.1), and no simple parametric family survived identifiability checks. We therefore model the peak value G_{peak}, the initial rate m, and the transfer extent d_{transfer}, and report the maximum's location as an observed quantity.