Beyond Absolute Imitation: Anchored Residual Guidance for Privileged On-Policy Distillation

Wenhao Zhang
South China University of Technology
vanhowe@outlook.com

Abstract

On-policy distillation (OPD) has demonstrated strong empirical gains in enhancing complex reasoning in LLMs by aligning a student model with a teacher's predictive distribution over the student's own trajectories. An emerging variant, Privileged OPD, further strengthens this paradigm by employing a self-teacher model augmented with privileged information, such as oracle traces, to mitigate teacher-student capacity gaps while providing dense, answer-directed supervision. However, current methods treat privileged information as a monolithic imitation target, failing to disentangle locally reachable reasoning steps from future-conditioned oracle signals. Consequently, the student is encouraged to match a hindsight-biased distribution that often falls outside its local predictive support. This reachability mismatch incentivizes the student model to skip valid intermediate reasoning in favor of locally unsupported shortcuts. To resolve this, we introduce **Anchored Residual On-Policy Distillation (AR-OPD)**, a dual-view framework that disentangles privileged supervision. Rather than enforcing strict full-view imitation, AR-OPD establishes a locally compatible anchor using a partially privileged teacher, isolating and injecting oracle foresight as a controlled residual to provide destination-directed guidance. Across diverse reasoning tasks, **AR-OPD outperforms full privileged OPD by 2.3 points and SFT by 7.9 points**. Crucially, this anchored residual mechanism reduces hindsight leakage by **21.7%** and mitigates late-stage drift, yielding up to a **7.2-point advantage** on challenging long-horizon trajectories exceeding 768 tokens.

1 Introduction

On-policy distillation (OPD) has rapidly emerged as an effective paradigm for large language model post-training, driving substantial gains in recent industry pipelines [Qwen Team, 2025, Xiaomi MiMo Team, 2025, GLM-5 Team, 2026]. By querying a teacher on states actively induced by the student, OPD addresses the off-policy exposure bias of supervised fine-tuning [Brown et al., 2020, Wei et al., 2022, Ouyang et al., 2022, Taori et al., 2023] and the reward sparsity of reinforcement learning [Schulman et al., 2017, Guo et al., 2025, Dong et al., 2024, Sheng et al., 2024, Thinking Machines Lab, 2025, Agarwal et al., 2024]. However, standard cross-model OPD often suffers from pattern mismatch: distilling a structurally incompatible teacher can degrade student performance [Li et al., 2026, Fu et al., 2026, Jang et al., 2026].

To circumvent this, *Privileged OPD* (e.g., OPSD, SDFT [Zhao et al., 2026]) restricts the teacher and student to the same base model but augments the teacher with auxiliary context, such as oracle traces. By allowing the self-teacher to "see more," privileged OPD improves teacher-student compatibility and provides dense supervision aligned with the student's native generation manifold. While Privileged OPD densifies the learning signal, it can also shift the teacher's distribution away from the student's causal predictive support.

Preprint.

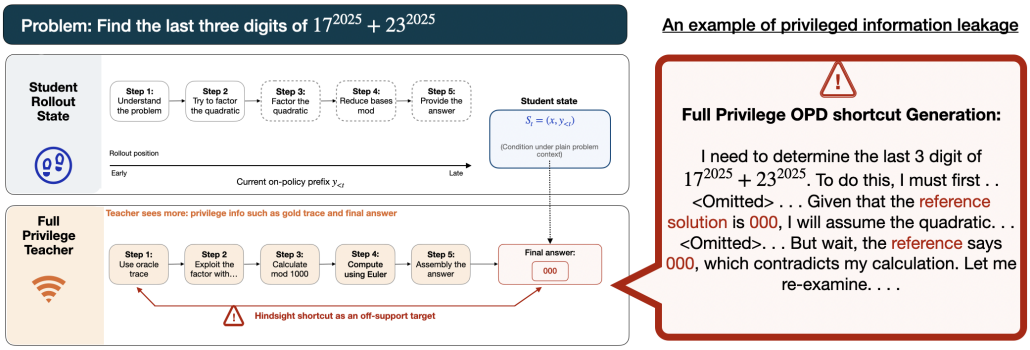

Figure 1: Privileged-information leakage as an off-support target. A representative example illustrating privileged-information leakage in a privileged OPD-trained model, where the model appeals to an invisible reference solution during inference. Because the full privileged teacher conditions on oracle information unavailable to the student, it can assign probability mass to answer-conditioned states. Direct imitation therefore turns factually correct privileged information into a shortcut target.

This creates a different failure mode from ordinary teacher-student mismatch. The teacher is not merely stronger or weaker; it is conditioned on information that the student cannot access at the same prefix. Thus, the central question is not whether privileged information is useful, but how much of it remains locally learnable as a token-level target.

Consequently, treating this fully privileged teacher as an absolute imitation target introduces a subtle but persistent optimization pathology. Because the full-view oracle conditions on the complete future context, its distribution implicitly incorporates hindsight-derived information and future-conditioned structure. It shifts probability mass toward tokens that, while factually correct, remain causally unsupported from the student's current prefix—such as anticipating a final answer branch before establishing the necessary intermediate rationale. We characterize these premature emissions as **privileged-information leakage** or *shortcut* events (Figure 1 illustrates this target-side failure, where a privileged OPD-trained model appeals to an invisible "reference solution" during inference, a phenomenon later quantified in Figure 6). In Section 4, we show that this mismatch becomes most visible near the rollout tail, where full-view targets assign increasing mass outside the student's local support. We further show that full-view reliability itself degrades with privileged-context length, making monolithic full-view imitation brittle precisely in long-horizon reasoning.

To overcome this absolute imitation trap, we propose **Anchored Residual On-Policy Distillation (AR-OPD)**, an objective designed to extract privileged guidance while preserving target reachability. Rather than forcing the assimilation of a monolithic full teacher, AR-OPD disentangles privileged supervision into two controllable components: a locally compatible ANCHOR and a destination-directed RESIDUAL. The ANCHOR uses a partially privileged teacher to provide a causally reachable target that isolates the student from hindsight-biased answers, while the RESIDUAL injects the full-view teacher's marginal foresight as a λ-scaled log-probability update over the anchor. This shifts imitation from an unsupported hindsight coordinate to a controlled, future-directed update.

Our contributions are summarized as follows:

- ❶ **Controlling the impact of privileged information.** We show that monolithic use of oracle-augmented teachers can induce optimization pathologies, and introduce a framework that extracts useful destination-directed foresight while controlling hindsight-biased manifold drift.
- ❷ **Anchored Residual On-Policy Distillation (AR-OPD).** We introduce a dual-view distillation method that operationalizes this control through a causally reachable partial-oracle ANCHOR and a λ-scaled log-space RESIDUAL, transferring privileged guidance without sacrificing local sequence compatibility.
- ❸ **Mechanistic evidence and cross-domain gains.** AR-OPD suppresses privileged-information leakage, reducing shortcut events by **21.7%**. It also achieves the highest average score across diverse reasoning benchmarks, improving over the base model by **12.1 points** and full privileged OPD by **2.3 points**.

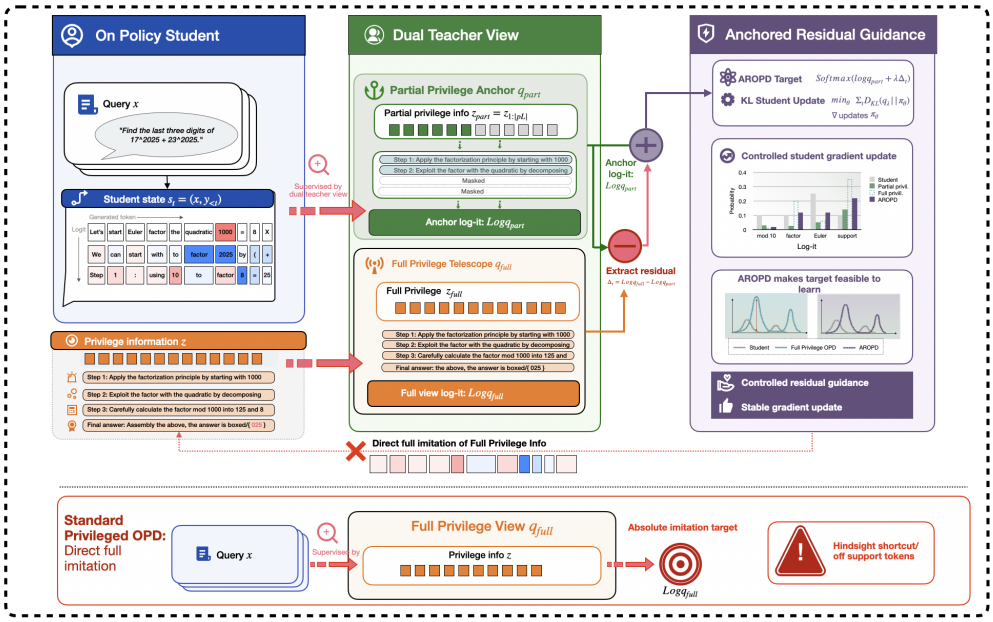

Figure 2: AR-OPD constructs a dual-view anchored residual target. The ANCHOR supplies the reachable reference, while the RESIDUAL transfers controlled full-view guidance before distillation to the student.

2 Related Work

Knowledge distillation and on-policy distillation. Knowledge distillation trains a student to match a teacher distribution or teacher-generated outputs [Hinton et al., 2015, Kim and Rush, 2016, Sanh et al., 2019, Gu et al., 2023]. For autoregressive generation, off-policy distillation and SFT can suffer from exposure mismatch because training targets are not conditioned on student-generated prefixes [Ross et al., 2011, Lamb et al., 2016]. OPD addresses this by querying teachers on student-visited states and has become a strong post-training recipe for reasoning models [Thinking Machines Lab, 2025, Agarwal et al., 2024, Fu et al., 2026, Li et al., 2026]. Recent variants study objective stability, entropy control, and teacher-student compatibility, showing that dense teacher targets are useful but can also become harmful when teacher distributions are locally incompatible with the student [Jang et al., 2026, Jin et al., 2026, Yang et al., 2026b].

Privileged OPD. Privileged OPD conditions the teacher on auxiliary information such as demonstrations, reference traces, verifier feedback, or oracle answers [Lightman et al., 2023, Shenfeld et al., 2026, Zhao et al., 2026, Penaloza et al., 2026, Yang et al., 2025]. Existing variants differ mainly in how they stabilize this privileged teacher. SDFT uses a delayed or EMA teacher to reduce training instability [Shenfeld et al., 2026], whereas OPSD keeps teacher and student synchronized and adds full-vocabulary JSD with per-token clipping [Zhao et al., 2026]. Despite these differences, they still treat the privileged distribution as an absolute imitation target, leaving target reachability and hindsight-induced support mismatch largely unaddressed.

3 Preliminaries

3.1 On-Policy Distillation

In autoregressive OPD, prompts $x \sim \mathcal{D}$ are paired with student rollouts $y = (y_1, \ldots, y_T) \sim \pi_\theta(\cdot \mid x)$. The state at step t is $s_t = (x, y_{<t})$, and the student policy factorizes as

$$\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid s_t). \tag{1}$$

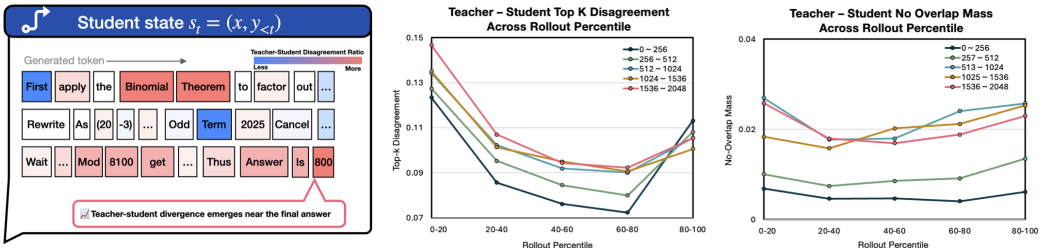

Figure 3: Target-reliability diagnostics. Left: A token-level illustration of late-rollout teacher–student divergence, where disagreement concentrates near the final-answer region. **Middle:** Top-k disagreement rises again near the rollout tail and increases with privileged-context length. **Right:** No-overlap mass shows the same tail-end elevation, indicating that the teacher assigns probability mass to tokens outside the student's local predictive support.

Given a teacher distribution $q(\cdot \mid s_t)$, the target of OPD is to minimize token-level KL divergence on student-visited states:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}\big(q(\cdot \mid s_t) \,\|\, \pi_\theta(\cdot \mid s_t)\big)\right]. \tag{2}$$

Different OPD implementations can provide full logits, top-k logits, or hard next-token labels. We use the logit-level view because it makes target construction observable: changing the privileged context changes the entire teacher distribution, not only the sampled next token.

3.2 Privileged OPD

Privileged OPD conditions the teacher on auxiliary information z, such as a solution trace, an oracle answer, or an expert demonstration. With a frozen or delayed teacher parameter copy $\bar{\theta}$, the privileged teacher is

$$q_z(\cdot \mid s_t) = \pi_{\bar{\theta}}(\cdot \mid x, z, y_{<t}). \tag{3}$$

The student is trained without z at inference time. Direct privileged imitation therefore uses the training objective

$$\mathcal{L}_{\text{P-OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}\big(q_z(\cdot \mid s_t) \,\|\, \pi_\theta(\cdot \mid s_t)\big)\right]. \tag{4}$$

4 Why Full-View Imitation Fails Local Alignment

Standard Privileged OPD assumes that oracle-augmented teachers provide uniformly reliable targets. We challenge this assumption by diagnosing the inherent reachability mismatch and signal fragility in absolute full-view imitation.

4.1 Target Unreachability: The Hindsight Support Gap

To quantify distributional misalignment, we measure Top-k Disagreement (D_k) and the Support Gap (M_τ^{out}). We compute these diagnostics on 1000 NuminaMath-style examples by sampling student rollouts from the public prompt and rescoring the same generated prefixes under the student, partial-view teacher, and full-view teacher prompts. We report top-k disagreement with $k = 16$, and compute support gap as the teacher mass assigned outside the student's locally plausible support; detailed estimator settings are provided in Section C.2. As shown in the middle and right panels of Figure 3, both metrics experience a tail-end elevation and scale with privileged-context length; the left panel illustrates the same late-rollout divergence pattern at the token level. Notably, for sequences exceeding 1024 tokens, the support gap increases substantially in the final 40% of the rollout. This tail-end elevation is consistent with **hindsight-induced manifold drift**. One explanation is that in early reasoning steps, the student and teacher distributions often share common ground, but as generation deepens, the student's prefix can accumulate slight deviations. The full-view teacher, conditioned on the gold future, remains anchored to the oracle trajectory. Rather than offering a localized correction, it assigns high probability to tokens that are justified by hindsight but logically disconnected from the

student's actual accumulated prefix. Consequently, absolute full-view targets can pull the student toward **locally unsupported states**. Under the forward KL $D_{\mathrm{KL}}(q \,\|\, \pi_\theta) = \sum_v q(v \mid s_t) \log \frac{q(v|s_t)}{\pi_\theta(v|s_t)}$, low-probability student terms $\pi_\theta(v \mid s_t)$ in the denominator can disproportionately dominate the expected loss. This can bias optimization toward oracle shortcuts instead of facilitating constructive, step-by-step reasoning transfer, consistent with recent analyses showing that privileged teacher-only signals can induce information leakage and that successful OPD depends on locally shared high-probability tokens at student-visited states [Yang et al., 2026a, Li et al., 2026].

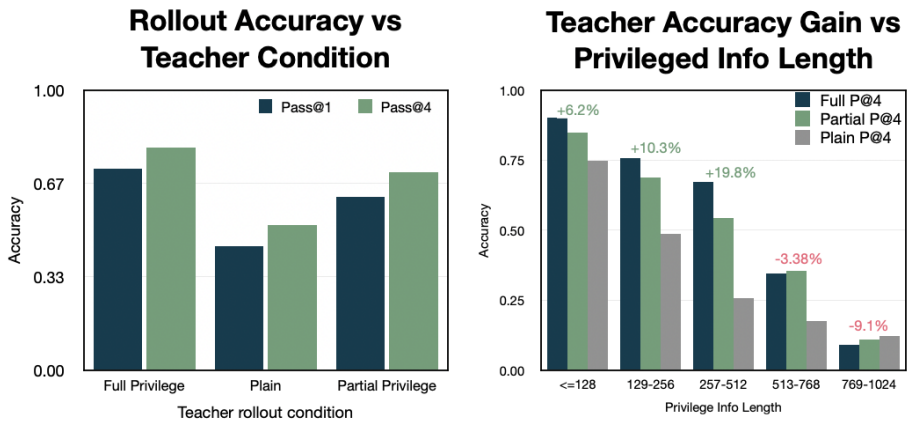

Figure 4: Reliability of privileged teachers in long-horizon contexts. Left: Even with gold traces, the privileged teacher is imperfect and its reliability decays sharply as privileged context length increases. **Right:** The marginal gain of full-view over partial-view conditioning peaks at 256–512 tokens but then turns into negative marginal returns (up to -9.1%) in longer horizons. This highlights that full-view targets in long-horizon reasoning are often noisier than partial-view anchors.

4.2 Signal Fragility: The Illusion of Full-View Reliability

Gold privileged context improves average teacher behavior, but it does not make the full-view teacher a uniformly reliable token target. On 5,000 NuminaMath problems with `Qwen2.5-7B-Instruct` [Qwen Team, 2024], we run four rollouts for each teacher/student condition and extract final answers for scoring. Figure 4 shows that full-view teacher pass@4 drops from 90.18% in the < 128-token bucket to 9.09% in the 768–1024-token bucket, and its marginal advantage over the partial view becomes negative by as much as -9.1%. This length sensitivity is consistent with recent OPD analyses showing that dense teacher targets can become locally incompatible when teacher-side information exceeds the student's reachable support [Li et al., 2026, Fu et al., 2026, Shenfeld et al., 2026, Penaloza et al., 2026].

The signal decomposition sharpens this point. Using the advantage scores defined in Equation (17), Figure A.1 shows that the partial privileged view is nearly as predictive of final correctness as the full-vs-student total advantage signal: partial advantage reaches ROC-AUC 0.766, close to the total signal at 0.779. By contrast, the marginal full-minus-partial signal falls to 0.585, indicating that the residual alone is a much weaker correctness classifier. Thus full-view information should not be treated as a standalone target; when used, it is better transferred as a controlled residual over a reliable partial anchor.

5 AR-OPD: Anchored Residual Guidance

AR-OPD replaces full-view imitation with a dual-view target. The method evaluates two teachers on the same student-generated state s_t: a partial teacher $q_{\mathrm{part}}(\cdot \mid s_t)$ and a full teacher $q_{\mathrm{full}}(\cdot \mid s_t)$. Figure 2 summarizes the target construction. The design follows a broader post-training pattern in which unstable long-horizon objects are decomposed into an anchor plus a controlled update, rather than optimized as a single monolithic target [Zhu et al., 2025, Li et al., 2025, Yang et al., 2026b].

5.1 Partial Privilege Defines the Local Anchor

Let the privileged information be $z = (z_1, \ldots, z_L)$. We construct a partial privileged context by deterministic truncation:

$$z_{\text{part}} = z_{1:\lfloor \rho L \rfloor}, \qquad \rho \in (0, 1). \tag{5}$$

The default setting is $\rho = 0.5$. The partial teacher is

$$q_{\text{part}}(\cdot \mid s_t) = \pi_{\bar{\theta}}(\cdot \mid x, z_{\text{part}}, y_{<t}). \tag{6}$$

The partial context exposes enough information to provide a structured setup while withholding later solution fragments, making the partial teacher a local anchor rather than an answer-leaking target. The front-partial choice is deliberately conservative: it resembles guided-prefix approaches that use early solution information to improve reachability, while avoiding direct supervision on the complete oracle trace [Qu et al., 2026, Shenfeld et al., 2026, Penaloza et al., 2026].

5.2 Full Privilege Enters as a Scaled Residual

The full teacher is

$$q_{\text{full}}(\cdot \mid s_t) = \pi_{\bar{\theta}}(\cdot \mid x, z, y_{<t}). \tag{7}$$

In log-probability space, the full-view learning signal decomposes as

$$\log q_{\text{full}} - \log \pi_\theta = (\log q_{\text{part}} - \log \pi_\theta) + (\log q_{\text{full}} - \log q_{\text{part}}). \tag{8}$$

The first term is the anchoring component, and the second term is the residual effect of giving the teacher the remaining privileged information.

Anchored residual target.

$$\boxed{\log q_\lambda(v \mid s_t) = \underbrace{\log q_{\text{part}}(v \mid s_t)}_{\text{partial anchor}} + \lambda \underbrace{\big(\log q_{\text{full}}(v \mid s_t) - \log q_{\text{part}}(v \mid s_t)\big)}_{\text{full-minus-partial residual}} - C_t.} \tag{9}$$

Target construction at each student state s_t.

1. Query the ANCHOR, $q_{\text{part}}(\cdot \mid s_t)$, using the front partial privileged trace.
2. Query the FULL VIEW, $q_{\text{full}}(\cdot \mid s_t)$, using the complete privileged trace.
3. Transfer only $\lambda(\log q_{\text{full}} - \log q_{\text{part}})$, then normalize to obtain q_λ.
4. Distill the student against q_λ on the same on-policy prefix s_t.

The normalizer C_t makes $q_\lambda(\cdot \mid s_t)$ a valid distribution. The form makes the object-level change explicit: imitation no longer targets q_{full} directly; it targets an anchored belief update around q_{part}. The student is trained by KL distillation to the anchored residual target:

$$\mathcal{L}_{\text{AR-OPD}}(\theta) = \mathbb{E}_{x \sim \mathcal{D},\, y \sim \pi_\theta}\left[\sum_{t=1}^{T} D_{\text{KL}}\big(q_\lambda(\cdot \mid s_t) \,\|\, \pi_\theta(\cdot \mid s_t)\big)\right]. \tag{10}$$

In implementation, targets are built on student-generated rollouts rather than fixed teacher traces. For each sampled prefix, the partial and full teacher prompts rescore the same next-token state with temperature 1.0; the anchored residual distribution is then formed in log space and optimized with forward KL. We use the same prompt and completion limits as the evaluation setup, and apply a small reference mixup and residual clipping only for numerical stability; the full configuration is listed in Table A.2.

The coefficient λ controls the amount of full-view residual transfer. When $\lambda = 0$, AR-OPD becomes partial privileged OPD; when $\lambda = 1$, it recovers the full privileged target by construction. Thus $\lambda < 1$ is a contractive residual regime that keeps the target anchored between the partial and full views, while $\lambda > 1$ extrapolates beyond the full-view log shift. Figure 5 visualizes this change: the full view is used as a directional residual around the partial anchor rather than as an absolute imitation target. The formulation makes λ closer to a residual-guidance strength than to a generic smoothing parameter, complementing work that studies reward extrapolation, target reformulation, and entropy-aware OPD control [Yang et al., 2026b, Jang et al., 2026, Jin et al., 2026].

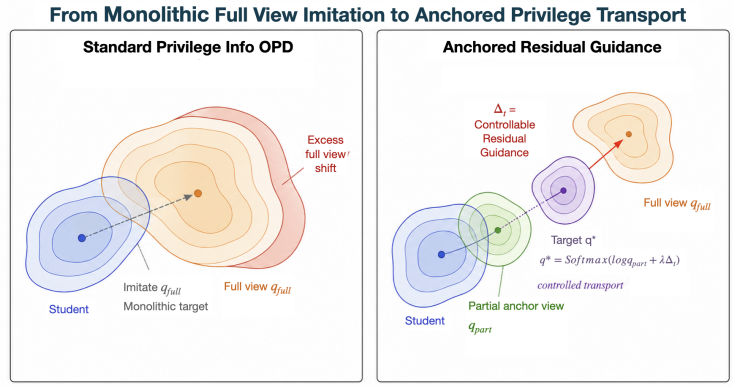

Figure 5: Constructing reachable targets via anchored residual guidance. AR-OPD anchors on the student-compatible partial view and uses the full view only as a controlled directional residual.

6 Experiments: Performance Gains Follow Target Reliability

The evaluation separates outcome evidence from mechanism evidence through four checks: cross-domain outcome, anchor compatibility, residual value, and training dynamics.

6.1 Experimental Setup

Tasks and datasets. We evaluate four reasoning task groups. For mathematics, we train on NuminaMath-style CoT data [AI-MO/Numina, 2024] and evaluate on MATH500 [Hendrycks et al., 2021, Lightman et al., 2023] and AMC23 [Mathematical Association of America, 2023]. For code, we train on Magicoder-Evol-Instruct data [Wei et al., 2024, Luo et al., 2023] and evaluate HumanEval and MBPP with pass@1 [Chen et al., 2021, Austin et al., 2021]. For science, we use SciKnowEval Chemistry L-3 [Feng et al., 2024] with exact-match multiple-choice accuracy. For medical QA, we train on MedMCQA [Pal et al., 2022] and evaluate on MedQA [Jin et al., 2020] and MedCQA [Pal et al., 2022]. We use the language-model evaluation harness for standardized evaluation plumbing where applicable [Gao et al., 2024].

Model and training. Unless stated otherwise, we use `Qwen2.5-7B-Instruct` as the base model [Qwen Team, 2024]. The student and teacher views share the same model family and differ only in the privileged context exposed to the teacher during target construction. Across tasks, we set the maximum completion length to 1024 tokens and use a 2048-token context window including the input prompt. The default partial rate is $\rho = 0.5$. We sweep $\lambda \in \{0.4, 0.6, 0.8, 1.0, 1.2\}$ to study residual strength, keeping other hyperparameters fixed across methods.

Compared methods. We compare the base model, supervised fine-tuning (SFT), full privileged OPD, partial privileged OPD, and AR-OPD. The comparison separates demonstration-only adaptation, full privileged self-distillation, and guided partial-privilege training [Ouyang et al., 2022, Shenfeld et al., 2026, Qu et al., 2026]. Full OPD trains directly against q_{full} with an EMA self-teacher. AR-OPD trains against q_λ, the anchor-residual target.

Evaluation protocol. All reported methods are evaluated with the same public prompts and without privileged context at inference time. Privileged information is used only to construct training targets, so the comparison isolates how different target constructions affect the same base model. For math and multiple-choice tasks, we extract the final answer with the task-specific evaluator; for code, we use the standard unit-test pass@1 protocol. All distillation variants use the same on-policy rollout budget, optimizer settings, prompt limits, and final-checkpoint reporting rule. Thus, differences in the main table reflect how the target distribution is constructed rather than changes in data scale, inference-time context, or evaluator choice.

6.2 Main Results: Performance

Table 1: Anchored residual guidance achieves the strongest average performance across domains. Metrics are accuracy or pass@1; code, science, and medical results remain part of the main cross-domain evidence, while checkpointed math-only diagnostics are separated in Section D.

Domain	Benchmark	Base	SFT	Partial OPD	Full OPD	AR-OPD
Math	MATH500	61.8	63.2	70.6	71.0	**74.6**
	AMC23	45.0	47.5	45.0	55.0	**57.5**
Code	HumanEval	79.2	76.8	79.4	78.8	**82.6**
	MBPP	77.5	79.4	77.8	78.5	**80.4**
Science	SciKnowEval	32.12	47.4	64.2	68.0	**68.5**
Medical	MedQA	58.7	66.0	64.3	65.2	**66.7**
	MedCQA	53.2	56.7	55.4	59.0	**61.5**
Average	All benchmarks	58.2	62.4	65.2	67.9	**70.3**

Performance. AR-OPD consistently improves over standard SFT and competitive on-policy distillation baselines across reasoning-intensive and knowledge-heavy domains. As shown in Table 1, AR-OPD obtains the best average score of 70.3, improving over the base model by **12.1 points**, SFT by **7.9 points**, and the strong Full OPD baseline by **2.3 points**. The gains are largest in mathematical reasoning, where hindsight-conditioned targets are most likely to induce support mismatch: AR-OPD reaches 74.6 on MATH500 and 57.5 on AMC23, exceeding Full OPD by **3.6 and 2.5 points**, respectively. Notably, even Partial OPD, which remains blinded to the final answer, outperforms standard full-trace SFT by an average of 2.8 points. While Partial OPD lacks destination awareness, AR-OPD adds residual guidance over a locally valid partial anchor, yielding a **5.1-point average gain over Partial OPD**. The improvement extends beyond math: AR-OPD also ranks first on HumanEval, MBPP, SciKnowEval, MedQA, and MedCQA.

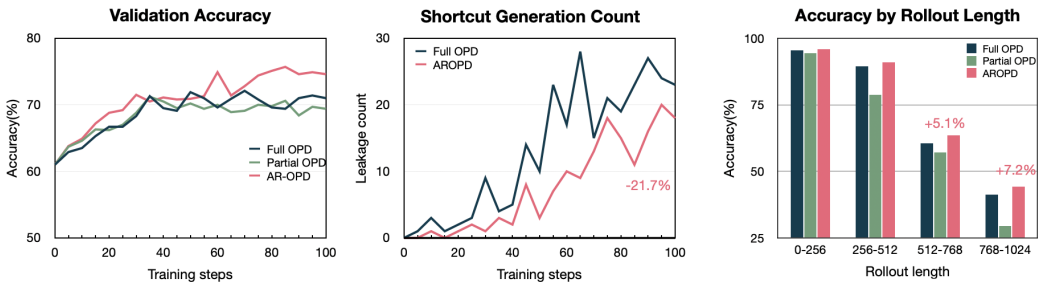

Figure 6: AR-OPD improves validation accuracy, reduces shortcuts, and is strongest on long rollouts. Left: Validation accuracy during Math10K training: AR-OPD shows consistent learning gains, while Full OPD peaks early and then largely plateaus. **Middle:** Shortcut-generation count over training steps; at the final training step (100%), AR-OPD has 18 shortcut events versus 23 for Full OPD, a 21.7% relative reduction. **Right:** Accuracy grouped by rollout length, where AR-OPD improves over Full OPD by 5.1 points on 512–768-token rollouts and 7.2 points on 768–1024-token rollouts.

6.3 Shortcut Dynamics and Long-Rollout Accuracy

Shortcut dynamics. The shortcut diagnostic quantifies the failure mode from Figure 1. By anchoring to the partial view, AR-OPD consistently suppresses shortcut events, reducing the average count across all 21 training checkpoints from 12.95 for Full OPD to 7.52. At the final checkpoint, AR-OPD produces only 18 shortcuts compared to Full OPD's 23, a **21.7% relative reduction**.

Long-horizon accuracy. AR-OPD's advantage scales with trajectory length (Figure 6, right). While competitive on short rollouts, it surpasses Full OPD by **5.1 points** on 512–768-token rollouts and **7.2 points** on 768–1024-token rollouts. These results suggest that, as reasoning horizons grow, anchored residual targets reduce the compounding hindsight drift that can degrade monolithic full-view imitation.

Empirical highlight. The results show that controlling the full-view residual improves over full privileged OPD, while the gain over Partial OPD confirms that the residual itself adds useful destination-directed signal.

6.4 Lambda Sweep and Residual Control

Table 2: Lambda sweep for the AR-OPD residual scale on the Math-10K checkpointed run. Metrics are relaxed accuracy. Contractive residual transfer ($\lambda < 1$) gives the strongest settings, while extrapolating past the full-view residual with $\lambda = 1.2$ degrades performance.

Residual scale	Regime	Math500	AMC23
$\lambda = 0.4$	Weak residual	71.2	50.0
$\lambda = 0.6$	Contractive residual	71.6	**57.5**
$\lambda = 0.8$	Contractive residual	**74.6**	55.0
$\lambda = 1.0$	Full-view limit	71.8	55.0
$\lambda = 1.2$	Extrapolated residual	69.0	50.0

Residual scale determines how aggressively AR-OPD transfers the full-view log shift. By Equation (9), $\lambda = 1$ recovers q_{full}, values below one keep the target contractively tied to the partial anchor, and values above one extrapolate beyond the full-view direction. The sweep in Table 2 shows that $\lambda = 0.8$ is strongest on Math500 and $\lambda = 0.6$ is strongest on AMC23, whereas the weak residual setting $\lambda = 0.4$ underuses answer-directed guidance and the extrapolated setting $\lambda = 1.2$ underperforms both contractive settings. We therefore treat λ as the main control knob balancing partial-anchor stability against full-view over-transfer, in line with recent OPD and RL-for-reasoning work showing that target scale and reformulation materially affect optimization behavior [Yang et al., 2026b, Jin et al., 2026, Wang et al., 2025].

7 Discussion: Decomposing Privileged Distribution Shift

Our results frame privilege as a target-design problem rather than a binary choice between using or discarding oracle information. Full-view oracle distributions can provide useful destination-directed signal, but absolute imitation can create a hindsight trap: tokens valid under the privileged future may remain unsupported from the student's current prefix. The diagnostics suggest that this mismatch grows near rollout tails and that longer privileged context does not necessarily improve teacher reliability. AR-OPD mitigates this by separating supervision into a locally valid anchor and a controlled residual, turning full-view information into a bounded update rather than a standalone coordinate.

Future work. Promising directions include adaptive anchors based on student uncertainty or gradient variance, broader privileged modalities beyond gold traces such as search heuristics or human preferences, and long-horizon settings such as multi-turn agents or repository-level coding where hindsight interference may be more pronounced.

8 Limitations

AR-OPD requires two teacher forward passes per student state, increasing target-generation cost relative to single-view OPD and motivating comparison with search- or verifier-based alternatives [Liu et al., 2025, Yang et al., 2025, Walder and Karkhanis, 2025]. Our experiments use one open model family, `Qwen2.5-7B-Instruct`, and modest data sizes; broader sweeps over data, scale, and families such as Llama [Touvron et al., 2023] remain important. Algorithmically, we use a fixed partial anchor ($\rho = 0.5$) and static residual scale; adaptive schedules, diagnostics beyond NuminaMath, and structured-output robustness remain open [Liu et al., 2023, Hassid et al., 2025, Madaan et al., 2023, Yuan et al., 2024].

References

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

AI-MO/Numina. Numinamath. https://huggingface.co/datasets/AI-MO/NuminaMath-CoT, 2024.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector, 2024. URL https://arxiv.org/abs/2408.09000.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024.

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, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

GLM-5 Team. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Michael Hassid, Gabriel Synnaeve, Yossi Adi, and Roy Schwartz. Don't overthink it. preferring shorter thinking chains for improved llm reasoning. *arXiv preprint arXiv:2505.17813*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. *arXiv preprint arXiv:2601.07155*, 2026.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1317–1327, 2016.

Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29, 2016.

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

Zhuofeng Li, Haoxiang Zhang, Seungju Han, Sheng Liu, Jianwen Xie, Yu Zhang, Yejin Choi, James Zou, and Pan Lu. In-the-flow agentic system optimization for effective planning and tool use. *arXiv preprint arXiv:2510.05592*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1qvx610Cu7.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*, 2025.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct, 2023. URL https://arxiv.org/abs/2306.08568.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

Mathematical Association of America. 2023 american mathematics competitions problems. https://maa.org/math-competitions/american-mathematics-competitions-amc/, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

Yuxiao Qu, Amrith Setlur, Virginia Smith, Ruslan Salakhutdinov, and Aviral Kumar. Pope: Learning to reason on hard problems via privileged on-policy exploration. *arXiv preprint arXiv:2601.18779*, 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

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

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://crfm.stanford.edu/2023/03/13/alpaca.html, 2023.

Thinking Machines Lab. On-policy distillation. https://thinkingmachines.ai/blog/on-policy-distillation/, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Christian Walder and Deep Karkhanis. Pass@ k policy optimization: Solving harder reinforcement learning problems. *arXiv preprint arXiv:2505.15201*, 2025.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Empowering code generation with OSS-instruct. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52632–52657. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wei24h.html.

Xiaomi MiMo Team. Mimo: Unlocking the reasoning potential of language model – from pretraining to posttraining, 2025. URL https://arxiv.org/abs/2505.07608.

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.

Wenkai Yang, Jingwen Chen, Yankai Lin, and Ji-Rong Wen. Deepcritic: Deliberate critique with large language models. *arXiv preprint arXiv:2505.00662*, 2025.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026b.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

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

He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. Anchored supervised fine-tuning. *arXiv preprint arXiv:2509.23753*, 2025.

Use of LLMs

Large language models were utilized as auxiliary tools to assist with code writing, manuscript editing, and figure preparation. The authors maintained full review and control over all research ideas, experimental designs, analyses, and final content. LLM assistance was used for LaTeX cleanup, wording refinement, caption compression, and drafting small analysis scripts during revision. LLMs were not used to generate benchmark measurements; all reported numbers come from the described experiments and were checked by the authors.

Code and Data Availability

The project page is available at `https://vanhowe.github.io/AR-OPD/`. It provides access to the paper assets, training and evaluation code, configuration files, launch scripts, lightweight evaluation data, and data-layout manifests. Large derived training datasets, checkpoints, and generated report bundles are not bundled in the release; the released scripts document the expected local layout and validation path.

A Prompt for Privileged Information

For each training example, we construct a standard student prompt and two privileged teacher prompts. The student prompt contains only the original programming problem. The privileged prompts additionally include an example response before asking the model to produce its own response.

Template for full and partial privileged information.

```
[Original problem]

This is an example for a response to the question:
[Privileged response]

Now answer with a response of your own, including the thinking
process.
```

For the full privileged condition, `[Privileged response]` is the complete reference trace, including the final code answer when available. For the partial privileged condition, `[Privileged response]` is a partial trace that removes the final answer. In the code data, the partial trace is constructed by removing fenced code blocks and inline code snippets from the reference response, then truncating the remaining explanation to the first 50% of characters without splitting a word.

Resulting training record.

- `prompt`: the original programming problem;
- `partial_teacher_prompt`: the original problem plus partial privileged information;
- `full_teacher_prompt`: the original problem plus full privileged information;
- `gold_trace`: the full reference response;
- `partial_trace`: the partial privileged response;
- `gold_answer`: the extracted final code answer.

Table A.1: Training data and privileged-view construction. Each row summarizes the source and dual-view construction used to build the corresponding training records.

Domain	Train size	Source	View construction
Math	10K	NuminaMath-style CoT subset	Partial teacher keeps the first 50% of the reasoning trace at a complete-word boundary and removes the final answer; full teacher keeps the complete trace with the gold final answer.
Code	4K	Magicoder-Evol-Instruct-110K [Wei et al., 2024]; upstream Evol-CodeAlpaca-style data	Partial trace removes fenced code blocks and inline code snippets, then keeps the first 50% of the remaining explanation without splitting a word; full trace keeps the complete reference response.
Science	2K	SciKnowEval Chemistry L-3 [Feng et al., 2024]	Multiple-choice demonstrations use the same dual-view construction; accuracy is computed by exact match on the final answer choice.
Medical	2K	MedMCQA subset [Pal et al., 2022]	Multiple-choice demonstrations use the same dual-view construction; accuracy is computed by exact match on the final answer choice.

Table A.2: Shared dual-GPU serving and training configuration. Code uses 4K training examples; science and medical use 2K examples.

Parameter	Value
Base model	`Qwen2.5-7B-Instruct`
Hardware topology	GPU 0 training, GPU 1 vLLM teacher serving
Learning rate / scheduler	2×10^{-5}, cosine, warmup ratio 0.1
Epochs / max steps	1 epoch, epoch-based training
Per-device batch / global batch	1 / 64
Gradient accumulation	64
Prompt / completion length	max prompt 1024, max completion 1024
Context window	vLLM max model length 2048, including input prompt
Training temperature	1.0
Precision	bf16 when GPU is available
Distillation objective	forward KL, distill alpha 0.0
Ref mixup / residual clip	mixup alpha 0.01, residual clip 5.0
Default residual scale	$\lambda = 0.6$, with sweep in Table 2
Teacher CPU offload	enabled via dual-teacher CPU offload

B Training and Evaluation Provenance

To make the main comparison reproducible, Tables A.1 to A.3 summarize the training data, shared optimization configuration, and evaluation protocol used for Table 1. Unless otherwise specified, all methods report the final checkpoint under the same evaluation protocol.

C Diagnostic Definitions

C.1 Connection to Classifier-Free Guidance

Classifier-Free Guidance (CFG) combines unconditional and conditional score estimates by scaling the residual effect of a condition [Ho and Salimans, 2022]. In diffusion models, this can be written abstractly as

$$\tilde{s}_\gamma(x_t, c) = s_\theta(x_t, \varnothing) + \gamma\big(s_\theta(x_t, c) - s_\theta(x_t, \varnothing)\big), \tag{11}$$

where the unconditional score provides a reference trajectory and the conditional residual steers sampling. Recent analyses further connect CFG to predictor-corrector behavior [Bradley and Nakkiran, 2024]. We use this only as a loose analogy: AR-OPD does not import a diffusion objective, but it adopts the same structural separation between an anchor distribution and a scaled directional residual.

Table A.3: Evaluation protocol for the main results. All methods report final-checkpoint performance under the same evaluator and prompt protocol.

Benchmark	Decoding	Evaluator	Metric
MATH500	Greedy, temperature 0.0	Math answer parser with relaxed normalization	Relaxed accuracy
AMC23	Greedy, temperature 0.0	Math answer parser with relaxed normalization	Relaxed accuracy
HumanEval	Greedy, temperature 0.0	Original HumanEval evaluator [Chen et al., 2021]	pass@1
MBPP	Greedy, temperature 0.0	Original MBPP evaluator [Austin et al., 2021]	pass@1
SciKnowEval	Greedy, temperature 0.0	Exact match on final multiple-choice answer	Accuracy
MedQA	Greedy, temperature 0.0	Exact match on final multiple-choice answer	Accuracy
MedCQA	Greedy, temperature 0.0	Exact match on final multiple-choice answer	Accuracy

C.2 Diagnostic Estimator

We estimate the target-reliability diagnostics on 1000 NuminaMath-style math examples. For each example, we sample a student rollout from the public prompt and treat every generated prefix as a student-visited state s_t. We then rescore the same generated token sequence under the public student prompt, partial privileged teacher prompt, and full privileged teacher prompt using the same checkpoint. Logits are converted to probabilities with temperature $T = 1.0$; generation uses top-$p = 1.0$, maximum generation length 1024, and maximum prompt length 2048. We report top-k disagreement with $k = 16$, averaged over token positions and grouped by five rollout-percentile bins and privileged-context-length buckets.

For a teacher distribution q, we first define the top-k overlap ratio as

$$O_k(q, \pi_\theta; s_t) = \frac{|\mathrm{Top}_k(q(\cdot \mid s_t)) \cap \mathrm{Top}_k(\pi_\theta(\cdot \mid s_t))|}{k}. \tag{12}$$

The corresponding top-k disagreement ratio is

$$D_k(q, \pi_\theta; s_t) = 1 - O_k(q, \pi_\theta; s_t). \tag{13}$$

A larger D_k indicates weaker local alignment between the teacher's high-probability tokens and the student's reachable predictive support on the same prefix.

The implementation-level diagnostic also records shared teacher mass over the overlapping top-k set:

$$\widehat{B}_k(q, \pi_\theta; s_t) = \sum_{v \in \mathrm{Top}_k(q(\cdot|s_t)) \cap \mathrm{Top}_k(\pi_\theta(\cdot|s_t))} q(v \mid s_t). \tag{14}$$

The corresponding top-k no-overlap mass proxy is the teacher mass assigned to top-k teacher tokens that do not appear in the student's top-k set:

$$\widehat{N}_k(q, \pi_\theta; s_t) = \sum_{v \in \mathrm{Top}_k(q(\cdot|s_t)) \setminus \mathrm{Top}_k(\pi_\theta(\cdot|s_t))} q(v \mid s_t). \tag{15}$$

This top-k proxy should be distinguished from the full-vocabulary thresholded support gap below.

We also define the teacher mass outside the student's local predictive support at threshold τ:

$$M_\tau^{\mathrm{out}}(q, \pi_\theta; s_t) = \sum_{v : \pi_\theta(v|s_t) < \tau} q(v \mid s_t). \tag{16}$$

This no-overlap mass, also referred to as the support gap, measures how much probability the teacher assigns to tokens that the student currently treats as locally implausible. Equivalently, since q is normalized, $M_\tau^{\mathrm{out}} = 1 - \sum_{v : \pi_\theta(v|s_t) \geq \tau} q(v \mid s_t)$. Thus the plotted no-overlap quantity is already the complement of in-support teacher mass. High values indicate that the teacher distribution may be correct under its privileged view but locally difficult for the student to imitate from s_t.

C.3 Teacher-Signal Advantage Scores

To measure which part of the privileged signal predicts final-answer correctness, we compute log-probability advantage scores over a token span $\mathcal{T}$. The total, partial, and marginal advantages are

$$
\begin{aligned}
A_{\text{full}}(\mathcal{T}) &= \sum_{t \in \mathcal{T}} \left(\log q_{\text{full}}(y_t \mid s_t) - \log \pi_\theta(y_t \mid s_t)\right), \\
A_{\text{part}}(\mathcal{T}) &= \sum_{t \in \mathcal{T}} \left(\log q_{\text{part}}(y_t \mid s_t) - \log \pi_\theta(y_t \mid s_t)\right), \\
A_{\text{marg}}(\mathcal{T}) = A_{\text{full}}(\mathcal{T}) - A_{\text{part}}(\mathcal{T}) &= \sum_{t \in \mathcal{T}} \left(\log q_{\text{full}}(y_t \mid s_t) - \log q_{\text{part}}(y_t \mid s_t)\right).
\end{aligned} \tag{17}
$$

We use each scalar score as a predictor of final-answer correctness and report ROC-AUC in Figure A.1.

C.4 Shortcut-Event Counting Protocol

To quantify privileged-information leakage in Figure 6, we evaluate trained privileged-OPD checkpoints on a held-out NuminaMath validation split and generate 5K responses per checkpoint. We use a deterministic response-level detector for shortcut-like evidence, such as references to unavailable solutions, answer-conditioned certainty without local derivation, or prompt-external answer cues. A response is counted once if any rule fires, so the reported count is a conservative response-level diagnostic of target-side leakage rather than a judge of mathematical correctness.

D Math Deep Dive and Training Diagnostics

This appendix reports supplemental diagnostics for the math setting; these results do not replace the main cross-domain comparison in Table 1. The Numina500 split exposes checkpoint and residual-scale sensitivity rather than redefining the headline benchmark table.

D.1 Teacher-Signal Diagnostics

Figure A.1 and Figure A.2 provide the appendix-level diagnostics behind the main mechanism claim: the partial view carries stable correctness-predictive signal, while the full-view target creates larger disagreement and support gaps against the student.

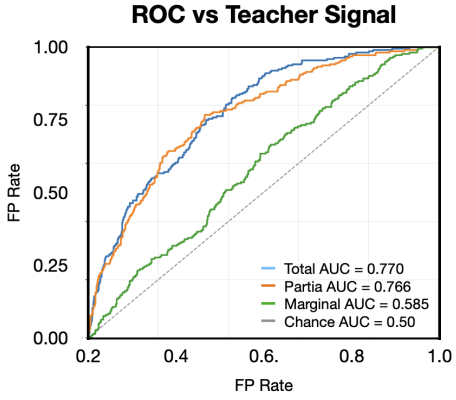

Figure A.1: Correctness-predictive teacher signal. The partial privileged signal nearly matches the total signal in ROC-AUC, while the marginal full-minus-partial residual is substantially weaker.

D.2 Checkpoint-Level Math Results

The checkpoint sweep is used only as supplemental evidence for residual-scale sensitivity: contractive AR-OPD settings are strongest on the headline math diagnostics, while the main cross-domain comparison remains the decisive result.

D.3 Failure Modes

Some raw failures reflect length and answer-format instability; relaxed-grading gains, especially for the partial-only model, should be read as formatting sensitivity rather than reasoning improvement.

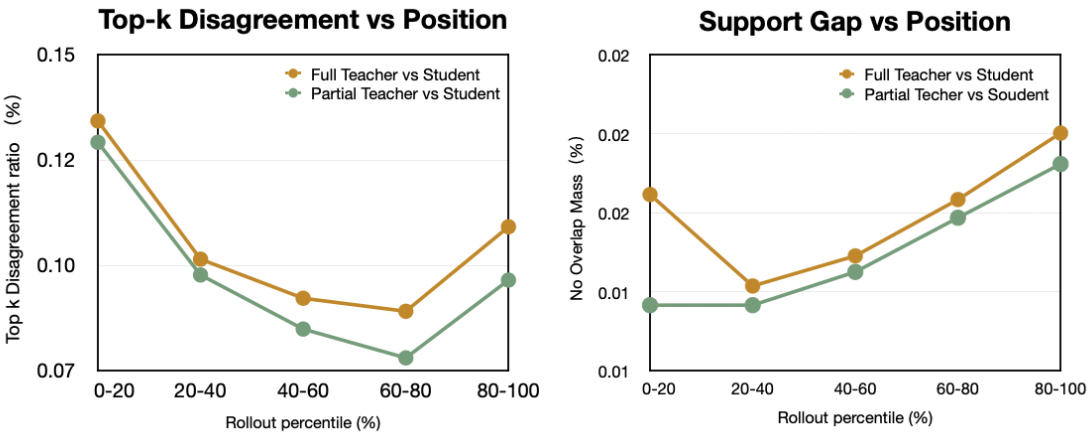

Figure A.2: Teacher–student support-gap diagnostics. Full-view targets show higher top-k disagreement and no-overlap mass than partial-view anchors across rollout positions.