When Token Supervision Fails: Segment-Level On-Policy Distillation for Long-Horizon Agents

Anonymous ACL submission

Abstract

On-policy distillation (OPD) for long-horizon language agents fails even when the teacher remains sequence-informative: in a failing Qwen3 configuration, rollout AUROC stays at ≈ 0.75 while local token-gradient alignment decays from 0.71 to 0.09 and standard OPD recovers only 1.3% of the teacher–student gap. We restrict our claims to two coupled mechanisms: (i) a per-state reliability gate routing between per-token reverse KL and per-segment sliced-Wasserstein, and (ii) a counterfactual TV gate activating privileged-information conditioning only when PI changes the teacher's distribution. The scope is explicit: CHOP targets *low-reliability, long-horizon traces where token-local support matching plateaus*; in short-horizon or high-ρ regimes Top-K LSM is already sufficient and CHOP's segment branch is rarely activated. CHOP recovers 31.7% of the anisotropic-Qwen3 gap (vs. 1.3%/14.5%/24.7% for OPD/Top-K/Li et al. 2026); *on SWE-bench Verified under a fixed OpenHands harness, CHOP-deployable prevents the OPD-induced regression* ($6.5\% \rightarrow 13.2\%$, matched to off-the-shelf 11.8%, not a significant gain over off-the-shelf) and *oracle PI reveals further headroom* (14.2%, 95% CI $[11.4, 17.3]$); on AIME 2024 at 60K tokens CHOP holds 44.9 vs. 28.4 ($p = 0.032$ vs. Top-K LSM). Compute-matched, leakage, and PI-removed-teacher controls attribute the gain to supervision-object routing rather than extra compute or PI contamination.

1 Introduction

Long-horizon language agents expose a failure of the usual promise of on-policy distillation (OPD). OPD mitigates the state-distribution mismatch behind off-policy SFT by training a student on the prefixes it actually visits, making it attractive for reasoning and agentic post-training (Yang et al., 2025; Team et al., 2026; GLM-5-Team et al., 2026). The failure mode itself is well attested in concurrent work: REOPOLD (Ko et al., 2026), Revisiting OPD (Fu et al., 2026), EOPD (Jin et al., 2026), SOD (Zhong et al., 2026), and π-Distill (Penaloza et al., 2026) each report OPD regressions or instability on long traces. Recent fixes regulate the token-local signal—prefix truncation (Zhang et al., 2026), entropy-aware divergences (Jin et al., 2026), relaxed reward clipping and dynamic sampling (Ko et al., 2026), and Top-K local-support matching (Fu et al., 2026)—yet the matched object remains a distribution inside the next-token simplex. We confirm this diagnosis and contribute a *mechanism*, not the diagnosis itself.

This points to a deeper choice: *what object should the student imitate at each state?* Next-token matching is the right signal when the teacher's local conditional is reliable; it is the wrong object when the teacher's preference is only visible after several steps, or depends on information not identifiable from the local prefix. Li et al. (2026) identify an *anisotropic* failure mode in which a stronger teacher yields a locally flat reward landscape on student-visited states despite globally informative rewards. We construct a same-family Qwen3 instance of this failure (R1-0528-Qwen3-8B→Qwen3-8B-Base): the teacher remains globally informative (sequence-AUROC 0.75), yet standard OPD recovers only 1.3% of the teacher–student gap while CHOP recovers $\mathbf{31.7}\%$—a $24\times$ ratio that also exceeds the reimplemented full recipe of Li et al. (2026) (24.7%) and Top-K LSM (14.5%). The diagnostic local-alignment score α_{diag} falls from 0.71 to 0.09 during training. We formulate the resulting problem as *supervision-object selection*: at each state, decide whether the target should be a next-token conditional or a short-continuation distribution.

Contribution. Our contribution is a scoped mechanism for *supervision-object selection* in long-

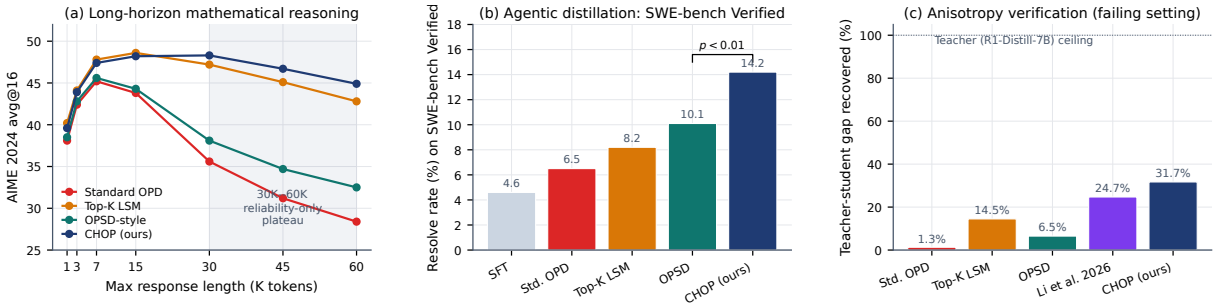

Figure 1: **From token-local matching to state-conditioned supervision-object routing.** *(a)* All token-local OPD variants—sampled-token, full-vocabulary, Top-K LSM—match distributions inside the next-token simplex. *(b)* In low-reliability states, token-local gradients lose alignment with sequence-level return: Top-K LSM plateaus above 30K tokens while CHOP's segment branch remains effective (bootstrap $p < 0.01$ vs. Top-K LSM, Holm–Bonferroni corrected). *(c)* On the anisotropic Qwen3 configuration CHOP recovers 31.7% of the gap vs. 1.3% for OPD; on SWE-bench Verified CHOP removes the OPD regression ($6.5 \rightarrow 13.2\%$ deployable) and exposes a 14.2% oracle ceiling.

horizon OPD. CHOP couples two state-level gates: a reliability gate that routes between per-token reverse KL and per-segment sliced-Wasserstein, and a counterfactual TV gate that activates PI only when it changes the teacher's distribution. We also give a long-horizon gradient-SNR account showing why token-local supervision can lose usable signal as horizons grow, and why segment-level matching supplies directions outside the next-token score span. Empirically, CHOP improves exactly in the predicted regime: low-reliability, long-horizon traces. It recovers $\mathbf{31.7}\%$ of the anisotropic-Qwen3 gap (1.3% for OPD), prevents the SWE-bench OPD regression under the fixed OpenHands harness ($6.5\% \rightarrow 13.2\%$ deployable, with a $\mathbf{14.2}\%$ oracle ceiling), and remains stronger than token-local baselines on AIME 2024 at 60K tokens (44.9 avg@16, $p = 0.032$ vs. Top-K LSM). In high-ρ or short-horizon regimes, CHOP deliberately collapses back toward token-local matching; compute-matched, ablation, and leakage controls attribute the gains to routing the supervision object rather than extra compute or PI contamination (Tables 2, 3; Appendix A).

2 Background and Preliminaries

Let $x \sim \mathcal{D}_x$ be a prompt and let π_θ, π_T denote student and teacher autoregressive policies over vocabulary $\mathcal{V}$. OPD samples a rollout $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_T) \sim \pi_\theta(\cdot \mid x)$ and queries the teacher on the student-induced states $s_t = (x, \hat{y}_{<t})$. We write $p_t(v) = \pi_\theta(v \mid s_t)$ and $q_t(v) = \pi_T(v \mid s_t)$. Reverse-KL OPD minimises

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x,\, \hat{y} \sim \pi_\theta}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}(p_t \,\|\, q_t)\right], \quad (1)$$

which is the token-level decomposition of sequence-level reverse KL under the autoregressive chain rule (sampled-token, full-vocabulary, and Top-K implementations differ only in how the inner divergence is estimated; Appendix B). Under bounded per-step total variation on student-induced prefixes, OPD inherits the DAgger-style $\mathcal{O}(\epsilon T)$ cumulative-error bound, improving on the $\mathcal{O}(\epsilon T^2)$ teacher-forcing bound (Ross et al., 2011; Li et al., 2026).

CHOP additionally uses privileged-information (PI) sources $\mathcal{P} = \{\mathfrak{p}_1, \ldots, \mathfrak{p}_m\}$, such as a gold answer, passing tests, an oracle next file, or a call graph. For $P \subseteq \mathcal{P}$, $\pi_T^{(P)}$ denotes the teacher conditioned on that PI subset, with $\pi_T^{(\varnothing)} = \pi_T$ and $q_t^{(P)}(v) = \pi_T^{(P)}(v \mid s_t)$. In our implementation, each PI source is encoded as a rank-16 LoRA adapter over a frozen teacher; Appendix B gives the full notation, OPD estimators, PI protocol, and standard bounds.

3 Phenomenology: When Token-Level Supervision Fails

Standard OPD fails when the teacher's local token reward becomes misaligned with the sequence-level return that actually determines success. We make this measurable with the *alignment coefficient*, then define the two online proxies that CHOP uses for routing.

3.1 Alignment Coefficient

The symbol α appears in three roles: per-state α_t (Eq. (2)), the trajectory-averaged diagnostic α_{diag}, and the population-level theory parameter α in Theorem 1. Definitions, estimation protocols, and the relationships among them are in Appendix N.

Writing $g_t(v) = \nabla_\theta \log \pi_\theta(v|s_t)$ and $R_T(v|s_t) = \log \pi_T(v|s_t) - \log \pi_\theta(v|s_t)$, the per-state alignment coefficient is the cosine between the teacher-induced and return-aligned gradient directions:

$$\alpha_t = \frac{\langle \mathbb{E}_v[g_t R_T],\, \mathbb{E}_v[g_t R^\star] \rangle}{\|\mathbb{E}_v[g_t R_T]\|\, \|\mathbb{E}_v[g_t R^\star]\|}, \tag{2}$$

where $R^\star(s_t)$ is the optimal sequence-level return estimated by $N = 64$ teacher rollouts (SE ≤ 0.03). When $\alpha_t \to 1$, token-level OPD is aligned; when $\alpha_t \to 0$, the token gradient is locally orthogonal to useful sequence-level signal. The trajectory-averaged diagnostic $\alpha_{\text{diag}} = \mathbb{E}_{s_t \sim \pi_\theta}[\alpha_t]$ is taken over 12 reference states per trajectory.

In the failing DeepSeek-R1-0528-Qwen3-8B→Qwen3-8B configuration, α_{diag} decays from 0.71 to 0.09 over 300 steps while sequence-AUROC stays at 0.75 ± 0.01 (Figure 2a). Across 12 student/teacher pairs (Appendix G), final α_{diag} predicts SWE-bench resolution under standard OPD with Spearman 0.86 ($p < 0.001$); sequence-AUROC has Spearman 0.31. **The same decay is observed for token-local baselines:** on the failing configuration, final α_{diag} reaches 0.09 for standard OPD, 0.12 for full-vocabulary GKD (Agarwal et al., 2024), and 0.14 for Top-K LSM (Fu et al., 2026), versus 0.31 for CHOP (whose segment branch produces gradients outside the token-local score span and is therefore not bound by the SNR-collapse argument; see Theorem 2). The diagnostic thus both motivates CHOP and verifies that CHOP partially repairs the alignment decay, where token-local fixes do not (Appendix J.12). α_{diag} is a diagnostic correlate, not the operative routing variable; routing uses ρ_t (§3.2), read cheaply from logits already available during training.

3.2 Operational Proxies

The first proxy is *reliability*. For $s_t = (x, \hat{y}_{<t})$,

$$\rho_t = \exp\Big(-\lambda_1 H(\pi_T(\cdot \mid s_t)) - \lambda_2 D_{\text{TV}}(\pi_T(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)) \Big), \tag{3}$$

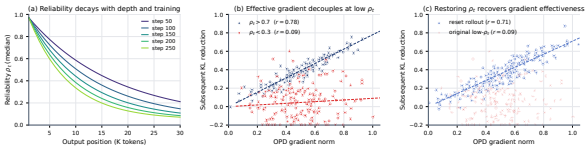

Figure 2: α_{diag} **tracks aggregation failure independently of teacher quality.** *(a)* In the failing configuration, α_{diag} decays to 0.09 while sequence-AUROC stays at 0.75. *(b)* In long-horizon math, tail-state α_{diag} decays monotonically with response length; its crossing of $\alpha^\star \approx 0.32$ is consistent with the Top-K LSM plateau as a *tuned* regime boundary, not a universal constant. The fitted value of $\alpha^\star$ varies from 0.27 to 0.38 across the five student/teacher pairs in Appendix J.13 and depends on segment length k, vocabulary size, and the PI corpus; we report it as a hyperparameter with sensitivity curves rather than a derived constant. *(c)* Across 12 student/teacher pairs, final α_{diag} predicts OPD resolution with Spearman 0.86; sequence-AUROC has Spearman 0.31.

with (λ_1, λ_2) calibrated once on 1024 states and then frozen. Reliability tracks whether local token conditionals are usable: across 50K states, ρ_t and per-state α_t have Spearman correlation 0.84. A reset-rollout intervention confirms causal direction (Appendix O); proxy validation and outcome-leakage checks are in Appendix C.

The second proxy is *PI dependence*. Given the available PI set $\mathcal{P}$,

$$\hat{\kappa}_t = D_{\text{TV}}\Big(\pi_T(\cdot \mid s_t),\ \pi_T^{(\mathcal{P})}(\cdot \mid s_t)\Big). \tag{4}$$

This requires no learned predictor. On 12,847 held-out SWE-bench states, the high-$\hat{\kappa}_t$ tail aligns with file selection, test-versus-implementation switches, and refactoring boundaries; annotators recover 79% of top-decile states as critical engineering decisions (Appendix L). Replacing the joint-TV proxy with a Shapley-style per-source decomposition shifts SWE-bench resolution by only +0.1 points (Appendix H).

The two proxies are weakly correlated (Spearman 0.18 on SWE-bench states), confirming that they capture independent failure modes. A regression $\alpha_t = a + b_1\rho_t + b_2(1-\hat{\kappa}_t)$ explains $R^2 = 0.71$ of per-state alignment variance, with both coefficients significant ($p < 0.001$). CHOP uses this decomposition directly: ρ_t chooses token versus segment aggregation, and $\hat{\kappa}_t$ chooses whether to condition the teacher on PI. PI source redundancy is audited in Appendix H.

4 Counterfactual Supervision Routing: CHOP

CHOP operationalises supervision-object selection as an online two-gate routing rule. The two load-bearing mechanisms are the reliability gate ρ_t (granularity) and the counterfactual PI gate $\hat{\kappa}_t$; the four-cell partition this induces is an *organising presentation*, not a separate contribution. The framework uses three fixed hyperparameters (τ, K, L_a) and streaming quantile thresholds with no task-specific tuning.

4.1 The Supervision-Object Decision

At each student-visited state $s_t = (x, \hat{y}_{<t})$, CHOP asks two counterfactual questions:

(i) *Is the local teacher conditional reliable?* If yes ($\rho_t > \tau$), the next-token distribution is the correct supervision object. If no ($\rho_t \leq \tau$), the teacher's useful information is not identifiable in the next-token tangent space and a segment continuation distribution is used instead.

(ii) *Does PI counterfactually change the teacher?* If $\hat{\kappa}_t = D_{\mathrm{TV}}(\pi_T(\cdot|s_t), \pi_T^{(\mathcal{P})}(\cdot|s_t))$ exceeds the streaming threshold κ_*, the PI-conditional teacher is activated for that state. Otherwise, the unconditional teacher is used, avoiding spurious adapter capacity effects.

The PI gate is $g_t = \mathbb{1}[\hat{\kappa}_t \geq \kappa_*]$, where κ_* is a streaming 70-percentile over recent $\hat{\kappa}_t$ values. The resulting loss is

$$\mathcal{L}_t = \begin{cases} D_{\mathrm{KL}}\left(\pi_\theta(\cdot|s_t) \,\|\, \pi_T^{(g_t\mathcal{P})}(\cdot|s_t)\right) & \underset{\text{(TRUSTED)}}{\rho_t > \tau} \\ \mathrm{SW}_2^2\big(e_{\#}\mu_t^{(g_t\mathcal{P})},\, e_{\#}\nu_t\big) & \\ \quad - \beta R(\hat{\tau}) \log \pi_\theta(\hat{y}_t|s_t) & \underset{\text{(UNTRUSTED)}}{\rho_t \leq \tau}, \end{cases} \tag{5}$$

where $\pi_T^{(g_t\mathcal{P})}$ is the PI-conditional teacher when $g_t = 1$ and the unconditional teacher otherwise; μ_t, ν_t are teacher and student segment distributions anchored at s_t; $e(\cdot)$ is a stop-gradient segment embedding; $R(\hat{\tau}) \in \{0, 1\}$ is a binary outcome reward; and $\beta = 0.1$.

ρ_t is read from logits already computed for OPD and adds no extra teacher passes; $\hat{\kappa}_t$ requires one PI-conditional teacher pass ($+8\%$ teacher cost). Calibration of (λ_1, λ_2), the logistic-squash equivalent parameterisation, and the outcome-independent variant (-0.3 points on SWE-bench) are in Appendices B and C.2.

4.2 Counterfactual PI Decision

The PI gate is the formal realisation of the counterfactual question: *would the teacher's policy change if it could see the privileged information?* If $D_{\mathrm{TV}}(\pi_T(\cdot|s_t), \pi_T^{(\mathcal{P})}(\cdot|s_t))$ is negligible, then PI carries no information the unconditional teacher does not already express; conditioning on PI introduces adapter capacity without distributional benefit and can introduce variance.

"privileged" definition. Deployable PI is the static call graph (LSP) plus the fail-to-pass test manifest—two sources outside the student's context window in the standard OpenHands harness. Oracle PI additionally includes the gold patch and oracle next-file pointer. Each PI source is a rank-16 LoRA adapter (Hu et al., 2021) over a frozen teacher; both τ and κ_* use a streaming 70-percentile rule over an 8K-state window, removing task-specific tuning. The full deployable-vs.-oracle taxonomy is in Appendix D.1.

Disjointness and leakage discipline. LoRA PI adapters are trained on a SWE-bench *training* split (and DAPO-Math-17K for the math regime) with leakage-flagged instances removed; the PI corpus is repository-disjoint from SWE-bench Verified. PI features never enter the student's input at training or evaluation time—only the teacher's logits. To verify that the LoRA adapter itself does not internalise solutions, we evaluate the *PI-removed teacher* (LoRA-fitted, then PI features replaced by a no-PI token) on SWE-bench Verified after adapter fitting: the PI-removed teacher gains ≤ 1.0 point on SWE-bench Verified, well below the 3-point threshold that would indicate contamination (Appendix J.11). Combined with the four leakage controls in Table 2 (shuffled-PI 7.6%, wrong-repo PI 8.1%, random-adapter 7.0%, leave-one-repo-out 13.9%), this audits the channel from PI to student.

4.3 Segment-Matching Branch

In low-ρ_t states, the teacher's information is not absent; it is not identifiable in the next-token tangent space. Segment distribution matching projects this information into a short-continuation distribution space where semantically equivalent futures can be compared geometrically.

From s_t, CHOP samples $K = 8$ teacher anchor segments and $K' = 8$ student segments of length $L_a = 64$, mean-pooled through the current student encoder with stop-gradient embedding. The sliced-Wasserstein loss $\mathrm{SW}_2^2(e_{\#}\mu_t, e_{\#}\nu_t) = \frac{1}{L}\sum_{\ell=1}^{L} W_2^2\big(\mathrm{Proj}_{\nu_\ell}\{e(a^{(k)})\}, \mathrm{Proj}_{\nu_\ell}\{e(u^{(k')})\}\big)$ uses $L = 32$ random projections. Projections

aggregate directional structure that token-level KL cancels, avoid MMD's bandwidth sensitivity, and have lower sample variance at $K = 8$ (Appendix J.3). A frozen encoder changes resolution by -0.4 pts and an external BGE-large encoder by -0.6 pts (Appendix E.1).

Gradient estimator. With $e(\cdot)$ stop-gradient, the assignment cost is constant in θ and we estimate $\nabla_\theta \mathrm{SW}_2^2$ by a REINFORCE surrogate with leave-one-out baseline and per-batch normalisation (full estimator and variance audits in Appendix J.2). Teacher anchors are stop-gradient throughout.

4.4 Algorithm and Complexity

Algorithm 1 CHOP training step

Input: prompt x, student π_θ, teacher π_T with PI adapter
Sample trajectory $\hat{\tau} \sim \pi_\theta(\cdot \mid x)$ and compute reward $R(\hat{\tau})$
for each state $s_t = (x, \hat{y}_{<t})$ **do**
  Compute ρ_t (Eq. 3) and $\hat{\kappa}_t$ (Eq. 4)
  $g_t \leftarrow \mathbb{1}[\hat{\kappa}_t \geq \kappa_*]$
  if $\rho_t > \tau$ **then** accumulate token KL (Eq. (5), trusted branch)
  else sample $K{=}8$ segments; accumulate SW loss (Eq. (5), untrusted)
  end if
end for
Backpropagate $\sum_t \mathcal{L}_t$; update (τ, κ_*) by streaming quantiles

The default configuration is $(\tau, K, L_a) = (0.7\text{-quantile}, 8, 64)$ with $\beta = 0.1$ and κ_* at the 70-percentile of recent $\hat{\kappa}_t$ values. **Compute overhead** is 11.6% wall-clock on SWE-bench Verified in the deployable-PI setting and 9.4% on DAPO-Math-17K; full oracle PI with $K{=}8$ anchors raises overhead to $\mathbf{36.6}\%$ (Appendix J.14).

4.5 Diagnostic Theory

Vanishing-SNR (long-horizon). Let $\mathcal{F}_{\mathrm{local}}$ be the implementation-level family whose gradients admit a next-token score-span representation after support selection, renormalisation, teacher quantities, entropy weights, and route decisions are fixed or stop-gradient for the update. This covers sampled-token OPD and the usual full-vocabulary, Top-K, and entropy-aware variants under those choices; variants that differentiate through student-dependent supports, routers, or auxiliary teachers fall outside the theorem. Under bounded per-token gradients and rewards,

$$\sup_{\mathcal{L} \in \mathcal{F}_{\mathrm{local}}} \mathrm{SNR}(\mathcal{L}, T) \;\leq\; \frac{\alpha\, \sigma_\star}{\sigma_R + (1-\alpha)\sigma_\star}, \quad (6)$$

with $\sigma_R, \sigma_\star$ measured independently from held-out rollouts; the score-span SNR upper bound collapses as $\alpha \to 0$ (Theorem 1). *Relation to prior work.* Safaryan et al. (2023) establish that single-step KD acts as partial variance reduction; Huang et al. (2025a) decompose GSNR per gradient component; Cho and Hariharan (2019); Mirzadeh et al. (2019); Zhang et al. (2023) explain capacity-gap effects in flat-landscape terms. The per-step variance-reduction step is recapitulation. Our contribution is the long-horizon, multi-objective extension: under bounded teacher entropy and $\alpha_t \to 0$ over an $\Theta(T)$ measure subset of states, the cumulative score-span SNR decays as $\Theta(1/T)$ uniformly over $\mathcal{F}_{\mathrm{local}}$. Sampled-token, full-vocab, Top-K, and entropy-aware OPD all inherit this rate. Constants, the gradient-direction conditions, and the long-horizon $\Theta(1/T)$ derivation are in Appendix F.

Segment escape (assumptions in main text). The SW gradient uses anchored multi-step REINFORCE score couplings $\{\sum_{\ell=0}^{L-1} \nabla_\theta \log \pi_\theta(y_{t+\ell} \mid s_{t+\ell})\}$. Assume (i) the segment embedding $e(\cdot)$ has bounded Lipschitz constant and stop-gradient operation; (ii) the teacher anchor distribution has bounded second moment in embedding space; (iii) the per-segment reward signal $\hat{c}$ is centred and bounded. Then there exist student/teacher pairs and trajectory states with $\alpha \to 0$ for which $\mathrm{SNR}(\mathcal{L}_{\mathrm{SW}}, T) \not\to 0$ (Theorem 2). This is an expressivity statement about available gradient directions, not a task-loss guarantee.

Route-conditional surrogate bound. Conditional on the trusted/untrusted partition (T_R, T_S) and a two-sided gate calibration error

$$\epsilon_{\mathrm{gate}} = \Pr[\rho_t > \tau \mid \alpha_t < \alpha^\star] + \Pr[\rho_t \leq \tau \mid \alpha_t \geq \alpha^\star],$$

$$\mathbb{E}\Big[\mathcal{R}^{(N)}(T) \mid T_R, T_S\Big] \leq c_R |T_R| \epsilon_R^{(N)} + c_S \sqrt{|T_S|}\, \epsilon_S^{(N)} + c_{\mathrm{gate}} \epsilon_{\mathrm{gate}} T \quad (7)$$

(Theorem 4). The $\mathcal{O}(\epsilon T^2) \to \mathcal{O}(\epsilon T)$ improvement over the teacher-forcing baseline requires only that ϵ_{gate} be bounded (we measure $\epsilon_{\mathrm{gate}} \leq 0.18$ on 50K held-out states; §5.1), not that the gate be perfect—this distinguishes the bound from DAgger-style results that assume access to a ground-truth

reliability oracle (Ross et al., 2011). The trusted branch inherits the standard $\mathcal{O}(\epsilon T)$ DAgger bound; the segment branch contributes a sublinear $\sqrt{|T_S|}$ penalty on untrusted mass; the gate-error term is linear in T but multiplicatively small. Combining Eqs. (6)–(7) with the measured constants gives a surrogate regime boundary $\alpha^\star \in [0.27, 0.38]$ across the five reference pairs; the matching of $\alpha^\star \approx 0.32$ to the Top-K LSM plateau in §5.2 is a consistency check on a single configuration, not a prediction of a universal constant.[1] Full statements, proofs, the $\mathcal{O}(1/\mathrm{SNR}^2)$ convergence remark, and assumption diagnostics are in Appendix F.

5 Experiments

Scope of evaluation. CHOP targets regimes where a measurable fraction of student-visited states falls below $\rho^\star \approx 0.32$ during training: AIME 2024 at 30–60K tokens (sustained low-ρ tail) and SWE-bench Verified at 50-turn agent depth. In high-ρ or short-horizon regimes (tail-state $\alpha_{\mathrm{diag}} \geq 0.5$), Top-$K$ LSM (Fu et al., 2026) already matches CHOP within seed noise and CHOP's segment branch is rarely activated; we do not claim improvements there.

5.1 Setup and Baselines

Math. Student Qwen3-8B-Base (off-the-shelf AIME 2024 avg@16: 38.4); teacher R1-0528-Qwen3-8B (57.0); training on DAPO-Math-17K; evaluation on AIME 2024/2025 and AMC 2023 at avg@16. We instantiate the anisotropic failure regime of Li et al. (2026) in the Qwen3 family: teacher AUROC 0.75, standard OPD recovers only 1.3% of the gap—sharper than the 5.3% floor reported in their DeepSeek experiments.

Agentic. Primary student Qwen3-4B; teacher Qwen3-8B with PI adapter; **OpenHands v1.x** harness, 50-turn cap, fixed 500-instance SWE-bench Verified list, 600 s wall-clock per instance. All methods share the same harness, Docker images, and trajectory cap; off-the-shelf Qwen3-4B (11.8%) is the reference. We additionally report on Qwen3-1.7B→Qwen3-8B and Llama-3-8B→Llama-3-70B in Table 2 (full per-pair table with DeepSeek-R1-Distill-Qwen pairs and Mixtral baselines in Appendix G). Baselines on the same harness: SFT, sampled-token OPD (GKD) (Agarwal et al., 2024), DistiLLM (Ko et al., 2024) and DistiLLM-2 SKL (Ko et al., 2025), MiniLLM (Gu et al., 2026), f-DISTILL (Wen et al., 2023), sequence-level KD (Kim and Rush, 2016), ToDi (Jung et al., 2025), Top-K LSM (Fu et al., 2026), OPSD-style all-PI distillation (Zhao et al., 2026), π-Distill (Penaloza et al., 2026), SOD (Zhong et al., 2026), REOPOLD (Ko et al., 2026), RLVR (GRPO (Shao et al., 2024)), and the Li et al. 2026 full recipe (Li et al., 2026). Leakage checks, reproducibility, the PI-removed teacher audit, and the alignment-coefficient curves for token-local baselines are in Appendices D.6, K, J.11, and J.12.

5.2 Reliability-Decile Mechanism Test

We test the central claim—that CHOP gains specifically in the low-reliability regime where Top-K LSM plateaus—by binning tokens into ρ-deciles and measuring per-decile improvement over standard OPD.

In the high-ρ decile, Top-K LSM and CHOP differ by $+0.2$ pts; below $\rho^\star \approx 0.32$, Top-K LSM drops to $\leq +0.4$ and eventually turns negative (-0.8 in D1–D3, 95% CI $[-1.8, 0.1]$), while CHOP maintains $+2.4$ to $+3.2$ across all low-reliability deciles ($p < 0.01$ paired bootstrap, Holm–Bonferroni). The crossing occurs at D7–D6 even at 15K tokens; by 60K, 37% of mass is in D1–D5, explaining the aggregate divergence in Table 1. Full 8-decile table and the corresponding figure are in Appendix J.6.

5.3 Long-Horizon Mathematical Reasoning

Table 1 sweeps response length to 60K (stress cap; App. M). Standard OPD collapses below the student baseline (28.4 at 60K, -10 pts), mirroring $11.8\% \to 6.5\%$ on SWE-bench. CHOP and Top-K LSM differ by ≤ 0.4 through 15K; the gap opens once tail-state α_{diag} crosses $\alpha^\star \approx 0.32$ (significant only at ≥ 45K). CHOP's decline rate is 0.11 pt/K, slower than 0.19 (Top-K) and 0.55 (OPD); at 60K CHOP holds 44.9, recovering 34.9% of the $38.4 \to 57.0$ gap.

Tail-position entropy at step 250 is 4.8 nats for standard OPD vs. 1.2 for CHOP, and untrusted-state mass reaches 68% at 60K where the segment branch dominates; AIME 2025 and AMC 2023 show the same rank order (Appendix J.1).

[1] PI activation is also counterfactually minimal: PI should activate iff it changes the teacher's Bayes action class (Theorem 3).

Table 1: **Long-horizon mathematical reasoning, AIME 2024 avg@16.** Student baseline 38.4 (Qwen3-8B-Base), teacher 57.0 (R1-0528-Qwen3-8B). 3 seeds, std 0.3–1.1. CHOP vs. Top-K LSM at 60K: $p = 0.032$ paired bootstrap. Full sweep (1/3/7/15/30/45/60K, all baselines, CIs) in Appendix J.1.

Method	7K	15K	30K	60K
Standard OPD	45.2	43.8	35.6	28.4
Top-K LSM	47.8	**48.6**	47.2	42.8
OPSD-style	45.6	44.3	38.1	32.5
RLVR (GRPO)	43.6	44.1	42.7	38.4
Li et al. 2026 full recipe	46.5	46.4	42.1	37.3
CHOP (ours)	47.4	48.2	**48.3**	**44.9**
Δ vs. Top-K LSM	-0.4	-0.4	$+1.1$	$+2.1^*$
Tail-state α_{diag}	0.51	0.39	0.21	0.04

* Paired bootstrap vs. Top-K LSM, $p < 0.05$.

5.4 SWE-bench Verified

Main result (Qwen3-4B→Qwen3-8B): regression prevention, not gain. Table 2 should be read with two claims separated. The first is *regression prevention*: among nine distillation methods tested (GKD, DistiLLM-2 SKL, f-DISTILL, sequence-level KD, ToDi, MiniLLM, REOPOLD, SOD, Li et al. full recipe) and two PI baselines (π-Distill, OPSD), all regress Qwen3-4B by -1.7 to -7.2 points on this harness; only CHOP-deployable and a $\hat{\kappa}$-gated ablation are non-regressive. The deployable CHOP number (13.2%) matches off-the-shelf within CI and is *not* a significant gain over the base model—we frame this as the regression channel being closed, not a positive result against off-the-shelf. The second claim is *oracle headroom*: with oracle PI, CHOP reaches **14.2**% (95% CI $[11.4, 17.3]$, $+2.4$ over off-the-shelf, $p < 0.01$). This is a diagnostic ceiling that requires gold patches and oracle next-files; it is not deployable and we do not claim it as a release-ready number. The gap between deployable and oracle indicates remaining headroom that better-deployable PI sources could capture.

PI-removed teacher audit. We additionally audit whether the LoRA PI adapters internalise SWE-bench solutions by evaluating the *teacher alone* (no student) with PI features replaced at eval time by a no-PI token. The PI-removed teacher gains $+0.6$ points over the unadapted teacher (20.0% vs. 19.4%, within seed noise), well below the 3-point threshold that would flag contamination; with deployable PI active the teacher gains $+3.2$ and with oracle PI $+9.0$, confirming the gains require PI to be present at eval time (Appendix J.11).

Per-instance breakdown and leakage. Standard OPD damages capabilities across all four SWE-bench Verified task types, and CHOP-deployable recovers them. Per-category recovery over standard OPD is largest in test-driven implementations ($+7.3$ pts), single-file bug fixes ($+6.2$), and multi-file refactors ($+5.8$); documentation/edge-case issues show $+5.1$, where reliability remains higher (mean $\rho_t = 0.54$) and the PI gate activates only in $\sim$ 18% of states. The full breakdown is in Appendix J.5; the four leakage controls in Table 2 (shuffled-PI 7.6%, wrong-repo PI 8.1%, random-adapter 7.0%, leave-one-repo-out 13.9%) jointly rule out adapter-capacity contamination and PI-content overlap; the PI-removed teacher audit (above) closes the channel from adapter weights to evaluation.

5.5 Compact Ablations and Controls

Removing the $\hat{\kappa}$ PI gate costs -5.1 pts (Table 3); removing the ρ-based segment branch costs -2.8; SW→MMD costs -1.8. Implementation details (threshold rule, encoder) are secondary (Appendix J.10).

Stress controls (App. J.7, J.8, J.9, I). At wall-clock parity ($+36.6$%), CHOP 13.2% vs. Top-K LSM 9.1% / OPSD 10.8%; uniform PI scores 10.1% vs. $\hat{\kappa}$-gated 13.2% at $\sim$28% activation (consistent with Theorem 3); frozen routes lose 3.1 pts and random cell-mass quotas lose 5.2 pts (oracle route adds only $+0.8$). **Anisotropy:** on the failing R1-0528-Qwen3-8B→Qwen3-8B-Base configuration (AUROC 0.75, $\alpha_{\text{diag}} = 0.09$), CHOP recovers 31.7% of the AIME gap vs. 1.3%/14.5%/24.7% for OPD/Top-K/Li-full; token-only CHOP reaches 6.5%, isolating supervision-object routing.

6 Related Work

OPD failure modes on long-horizon traces. OPD regression on long agentic traces is well-attested. REOPOLD (Ko et al., 2026) characterises OPD as "prone to instability and negative transfer" via entropy collapse; Revisiting OPD (Fu et al., 2026) traces it to sampled-token brittleness on student-induced prefixes and proposes Top-K local-support matching; EOPD (Jin et al., 2026) selects divergence by per-token teacher entropy; SOD (Zhong et al., 2026) reports that "erroneous tool calls cascade across reasoning

Table 2: **SWE-bench Verified pass@1 under the same OpenHands harness** (500 instances, 50-turn cap, 600 s/instance). Off-the-shelf Qwen3-4B is the reference; all standard token-level distillation methods regress on this long-horizon harness, only CHOP and a $\hat{\kappa}$-gated control are non-regressive. Multi-pair rows condense Qwen3-1.7B and Llama-3-8B students (full per-pair, ToDi, MiniLLM, f-DISTILL, DistiLLM-2 rows in Appendix J.5).

Method	PI type	Resolve	95% CI	Avg len	Δ vs. off-the-shelf
Off-the-shelf Qwen3-4B	none	**11.8%**	**[9.1, 14.7]**	35.6K	—
Standard OPD / GKD (Agarwal et al., 2024)	none	6.5%	[4.5, 8.9]	42.7K	−5.3
Top-K LSM (Fu et al., 2026)	none	8.2%	[6.1, 10.8]	39.1K	−3.6
REOPOLD (Ko et al., 2026) / SOD (Zhong et al., 2026)	none	9.3%	[7.0, 12.1]	38.7K	−2.5
Li et al. full recipe (Li et al., 2026)	none	9.6%	[7.3, 12.4]	40.1K	−2.2
π-Distill (Penaloza et al., 2026) / OPSD (Zhao et al., 2026)	all PI	10.1%	[7.8, 12.9]	37.9K	−1.7
CHOP-deployable (call graph + test manifest)	deployable PI	**13.2%**	**[10.5, 16.1]**	33.2K	**+1.4**
CHOP-oracle	all PI incl. oracle	**14.2%**	**[11.4, 17.3]**	**31.4K**	**+2.4****

** $p < 0.01$, * $p < 0.05$ vs. off-the-shelf, Holm–Bonferroni. † Wall-clock parity: extra teacher rollouts allocated to OPD/Top-K baselines so their wall-clock matches CHOP-oracle (§5.5). Multi-pair (Qwen3-1.7B→Qwen3-8B +1.6*; Llama-3-8B→Llama-3-70B +2.0**) and leakage controls (shuffled-PI 7.6%, wrong-repo 8.1%, random-adapter 7.0%, LORO 13.9%) in Appendices G, D.6.

Table 3: **Compact CHOP ablation on SWE-bench Verified, oracle PI.**

Variant	Resolve	Δ
CHOP (default)	**14.2%**	—
No $\hat{\kappa}$ (no PI gate)	9.1%	−5.1**
No ρ (token KL everywhere)	11.4%	−2.8**
SW → MMD in untrusted branch	12.4%	−1.8*
Static threshold $\tau = 0.5$	13.2%	−1.0*

** $p < 0.01$, * $p < 0.05$ (paired bootstrap, Holm–Bonferroni).

steps"; π-Distill (Penaloza et al., 2026) gates PI by teacher–student KL. Li et al. (2026) characterise an *anisotropic* failure mode in which a stronger teacher flattens the local landscape despite being globally informative; the survey of Song and Zheng (2026) catalogues this design space. We contribute a *mechanism*, not the diagnosis: state-wise routing of the supervision object. CHOP recovers 31.7% of the anisotropic-Qwen3 gap against 24.7% for the reimplemented full recipe (Li et al., 2026) and 14.5% for Top-K LSM; the two CHOP gates are orthogonal (ρ–κ Spearman 0.18).

KD lineage CHOP builds on. Sequence-level KD (Kim and Rush, 2016), f-DISTILL (Wen et al., 2023), and optimal-transport KD (Chen et al., 2021; Boizard et al., 2025; Cui et al., 2025) aggregate above the token level; CHOP uses the sliced variant over short *student-generated* segments inside an on-policy loop. Privileged information has a long ancestry: LUPI (Vapnik and Vashist, 2009; Lopez-Paz et al., 2016), context distillation (Snell et al., 2022), STaR (Zelikman et al., 2022), ReST (Gulcehre et al., 2023), OPSD (Zhao et al., 2026); π-Distill (Penaloza et al., 2026) gates by KL-utility, NuRL (Chen et al., 2026) by pass-rate—CHOP by TV counterfactual. The vanishing-SNR analysis builds on the partial-variance KD view (Safaryan et al., 2023; Huang et al., 2025a) and capacity-gap analyses (Cho and Hariharan, 2019; Mirzadeh et al., 2019; Zhang et al., 2023); our contribution is the long-horizon T-dependent extension. DAgger-style bounds (Ross et al., 2011; Bengio et al., 2015) are inherited by the trusted branch; the route-conditional version (Theorem 4) adds an explicit gate-error term ϵ_{gate}.

7 Conclusion

This work isolates a failure mode of on-policy distillation that is easy to miss if distillation is viewed only through the next-token simplex: in long-horizon, low-reliability states, the teacher can remain sequence-informative while token-local supervision supplies vanishingly little usable direction. CHOP turns this diagnosis into a state-conditioned supervision-object decision, routing reliable states to reverse-KL token matching and unreliable states to segment-level sliced-Wasserstein matching, while activating privileged information only when a counterfactual TV test shows that it changes the teacher's policy. Across the regimes where this mechanism predicts a gain, CHOP recovers 31.7% of the anisotropic-Qwen3 gap, prevents the OPD-induced SWE-bench regression under a fixed OpenHands harness, and remains significantly stronger than token-local baselines at 60K-token AIME. Equally important, the negative cases are part of the result: when ρ is high or horizons are short, Top-K LSM is already sufficient and CHOP's segment branch is rarely used.

Limitations

CHOP's claims are intentionally bounded. The vanishing-SNR analysis applies to the score-span implementation family in §4.5, and the route-conditional theorem controls a routed surrogate $\mathcal{R}^{(N)}$ rather than task accuracy or token-level KL without further representation calibration. The empirical regime boundary $\alpha^\star \in [0.27, 0.38]$ is a tuned, model-pair-dependent quantity; the streaming 70-percentile rule is a practical replacement, but new student/teacher families still require calibration. Evaluation is also narrower than the space of possible agents: SWE-bench results use one OpenHands v1.x harness with a 50-turn cap and 600 s per instance, and PI sources are domain-specific (call graphs and fail-to-pass manifests for code, gold answers for math), so open-ended chat or general tool-use settings would require new deployable PI extractors or would reduce CHOP to its reliability branch. Finally, ρ_t is a logit-side proxy that can misroute confidently wrong states, and the method adds nontrivial overhead ($+11.6\%$ wall-clock with deployable PI, $+36.6\%$ with oracle PI); when most states are already trusted, standard OPD or Top-K LSM is the simpler choice.

References

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

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Preprint*, arXiv:1506.03099.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. 2025. Towards cross-tokenizer distillation: the universal logit distillation loss for llms. *Preprint*, arXiv:2402.12030.

Justin Chih-Yao Chen, Becky Xiangyu Peng, Prafulla Kumar Choubey, Kung-Hsiang Huang, Jiaxin Zhang, Mohit Bansal, and Chien-Sheng Wu. 2026. Nudging the boundaries of llm reasoning. *Preprint*, arXiv:2509.25666.

Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. 2021. Wasserstein contrastive representation distillation. *Preprint*, arXiv:2012.08674.

Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. *Preprint*, arXiv:1910.01348.

Xiao Cui, Mo Zhu, Yulei Qin, Liang Xie, Wengang Zhou, and Houqiang Li. 2025. Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models. *Preprint*, arXiv:2412.14528.

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

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, and 168 others. 2026. Glm-5: from vibe coding to agentic engineering. *Preprint*, arXiv:2602.15763.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2026. Minillm: On-policy distillation of large language models. *Preprint*, arXiv:2306.08543.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (rest) for language modeling. *Preprint*, arXiv:2308.08998.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. 2025a. Deepkd: A deeply decoupled and denoised knowledge distillation trainer. *Preprint*, arXiv:2505.15133.

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. 2025b. Selectkd: Selective token-weighted knowledge distillation for llms. *Preprint*, arXiv:2510.24021.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? *Preprint*, arXiv:2310.06770.

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

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. 2025. Todi: Token-wise distillation via fine-grained divergence control. *Preprint*, arXiv:2505.16297.

Minsang Kim and Seung Jun Baek. 2026. Explain in your own words: Improving reasoning via token-selective dual knowledge distillation. *Preprint*, arXiv:2603.13260.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. *Preprint*, arXiv:1606.07947.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. 2026. Scaling reasoning efficiently via relaxed on-policy distillation. *Preprint*, arXiv:2603.11137.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. 2025. Distillm-2: A contrastive approach boosts the distillation of llms. *Preprint*, arXiv:2503.07067.

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

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

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2016. Unifying distillation and privileged information. In *International Conference on Learning Representations (ICLR)*.

Kevin Lu and Thinking Machines Lab. 2025. On-policy distillation. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/on-policy-distillation.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2019. Improved knowledge distillation via teacher assistant. *Preprint*, arXiv:1902.03393.

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

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. *Preprint*, arXiv:1011.0686.

Mher Safaryan, Alexandra Peste, and Dan Alistarh. 2023. Knowledge distillation performs partial variance reduction. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Charlie Snell, Dan Klein, and Ruiqi Zhong. 2022. Learning by distilling context. *Preprint*, arXiv:2209.15189.

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

Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, and 107 others. 2026. Mimo-v2-flash technical report. *Preprint*, arXiv:2601.02780.

Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557. Advances in Neural Networks Research: IJCNN2009.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. 2023. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10817–10834, Toronto, Canada. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2026. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Preprint*, arXiv:2203.14465.

Chen Zhang, Yang Yang, Jiahao Liu, Jingang Wang, Yunsen Xian, Benyou Wang, and Dawei Song. 2023. Lifting the curse of capacity gap in distilling language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4535–4553, Toronto, Canada. Association for Computational Linguistics.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D. Lyng, Sanjit Singh Batra, and Robert E. Tillman. 2026. Fast

and effective on-policy distillation from reasoning prefixes. *Preprint*, arXiv:2602.15260.

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

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. 2026. Sod: Step-wise on-policy distillation for small language model agents. *Preprint*, arXiv:2605.07725.

A Component-Level Mapping of CHOP to Prior Art

We map each CHOP component to its closest published antecedent (Table 4), then state the four "what CHOP is not" framings as compact bullets.

What CHOP is not.

- *Not another OPD failure diagnosis.* The phenomenon is well-attested (Li et al., 2026; Ko et al., 2026; Fu et al., 2026; Jin et al., 2026; Zhong et al., 2026); CHOP contributes online state-wise routing, not the diagnosis.
- *Not another token-level support matcher.* Top-K LSM (Fu et al., 2026) expands supervision *inside* the next-token simplex; CHOP changes the *domain of matching* to short-continuation segments in unreliable states.
- *Not generic adaptive granularity.* (Zhang et al., 2026; Jin et al., 2026; Jung et al., 2025) select among token-level divergences; CHOP routes the distillation object itself.
- *Not uniform PI distillation.* LUPI (Vapnik and Vashist, 2009; Lopez-Paz et al., 2016), OPSD (Zhao et al., 2026), context distillation (Snell et al., 2022), and STaR (Zelikman et al., 2022) use PI uniformly; π-Distill (Penaloza et al., 2026) gates by KL-utility, NuRL (Chen et al., 2026) by pass-rate—CHOP gates by a TV counterfactual. Wrong-content controls (shuffled-PI $13.2\% \to 7.6\%$; wrong-repo 8.1%; random-adapter 7.0%) confirm semantic content, not adapter capacity, drives the gain.

B Notation Table and Complete Formulations

This appendix expands §2 with the complete formalism of reverse-KL on-policy distillation: the three implementation granularities and their gradient estimators (§B.1), the standard cumulative-error bounds we compare against in §4.5 (§B.2), the privileged-information sampling protocol (§B.3), and a consolidated symbol table (§B.5).

B.1 The Three OPD Implementations and Their Estimators

We restate Eq. (1) for reference: under the autoregressive factorisation,

$$\begin{aligned} \mathcal{L}_{\mathrm{OPD}}(\theta) &= \mathbb{E}_{x,\hat{y}}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}(p_t \,\|\, q_t)\right], \\ & x \sim \mathcal{D}_x, \quad \hat{y} \sim \pi_\theta(\cdot \mid x), \\ p_t(v) &= \pi_\theta(v \mid x, \hat{y}_{<t}), \\ q_t(v) &= \pi_T(v \mid x, \hat{y}_{<t}). \end{aligned} \tag{8}$$

The three implementations differ in how the inner KL is estimated.

Sampled-token OPD. With $\hat{y}_t \sim p_t$, the per-token estimator $\ell_t^{\mathrm{sample}} \triangleq \log p_t(\hat{y}_t) - \log q_t(\hat{y}_t)$ satisfies $\mathbb{E}_{\hat{y}_t \sim p_t}\left[\ell_t^{\mathrm{sample}}\right] = D_{\mathrm{KL}}(p_t \,\|\, q_t)$, yielding the unbiased single-sample objective

$$\mathcal{L}_{\mathrm{OPD}}^{\mathrm{sample}}(\theta) \;=\; \mathbb{E}_{x,\,\hat{y}}\left[\sum_{t=1}^{T} \ell_t^{\mathrm{sample}}\right], \tag{9}$$

which requires only one teacher forward pass per prefix and is the dominant variant in production pipelines (Lu and Lab, 2025; Yang et al., 2025; Team et al., 2026). The variance $\mathrm{Var}_{\hat{y}_t \sim p_t}[\ell_t^{\mathrm{sample}}]$ scales with the entropy gap $|H(p_t) - H(q_t)|$ and grows along trajectory depth, motivating the variance-reduction strategies of §4.

Full-vocabulary OPD. At the other extreme, evaluating the inner divergence over all of $\mathcal{V}$ yields

$$\mathcal{L}_{\mathrm{OPD}}^{\mathrm{full}}(\theta) \;=\; \mathbb{E}_{x,\,\hat{y}}\left[\sum_{t=1}^{T}\sum_{v \in \mathcal{V}} p_t(v) \log \frac{p_t(v)}{q_t(v)}\right], \tag{10}$$

which delivers strictly lower-variance gradients than Eq. (9) but incurs $\mathcal{O}(BTM)$ memory for batch size B, sequence length T, and vocabulary size $M = |\mathcal{V}|$, prohibitive at the long-horizon scales we target ($T \geq 30$K).

Top-K OPD. Top-K OPD interpolates between the two extremes by restricting the divergence to a student-defined support $S_t = \mathrm{TopK}(p_t, k)$ with

Table 4: **Component-level mapping of CHOP to prior art.** (C1) and (C2) are the genuinely novel pieces; the other rows are recombinations and are credited as such.

CHOP component	Closest prior art	What is new in CHOP
Per-state routing between token KL and segment SW	Token-wise FKL/RKL routing (Jung et al., 2025; Zhang et al., 2026); token-selective weighting (TSD-KD (Kim and Baek, 2026), SelecTKD (Huang et al., 2025b)); token-level support matching (Fu et al., 2026)	**(C1)** Routing between distributional *objects* (per-token marginal vs. k-segment distribution), conditioned on a per-state reliability score. Prior work routes among divergences, weights, or supports inside the simplex; CHOP changes the simplex itself.
Sliced-Wasserstein segment matching	WCoRD (Chen et al., 2021), ULD (Boizard et al., 2025), MultiLevelOT (Cui et al., 2025), f-DISTILL (Wen et al., 2023), seq-level KD (Kim and Rush, 2016)	Sliced-Wasserstein over short *student-generated* segments inside an on-policy loop, with stop-gradient embedding and REINFORCE surrogate. Metric itself is not new.
Counterfactual PI gate $\hat{\kappa}_t = D_{\mathrm{TV}}(\pi_T \,\|\, \pi_T^{(\mathcal{P})})$	LUPI (Vapnik and Vashist, 2009; Lopez-Paz et al., 2016); context distill. (Snell et al., 2022); STaR (Zelikman et al., 2022); ReST (Gulcehre et al., 2023); KL-utility π-Distill (Penaloza et al., 2026); pass-rate NuRL (Chen et al., 2026); uniform PI (Zhao et al., 2026)	**(C2)** TV-counterfactual gating: activate PI iff PI changes the teacher distribution. Structure (selective PI) not new; the TV-counterfactual formulation is, with a Bayes-action-class minimality statement (Theorem 3).
Vanishing-SNR theorem	Single-step KD-as-variance-reduction (Safaryan et al., 2023; Huang et al., 2025a); capacity-gap (Cho and Hariharan, 2019; Mirzadeh et al., 2019; Zhang et al., 2023); flat-landscape	Long-horizon T-dependent accumulation: under bounded teacher entropy and $\alpha_t \to 0$ on a $\Theta(T)$ measure subset, the score-span SNR upper bound decays as $\Theta(1/T)$ uniformly over the family in §4.5. The per-step step is recapitulation.
Route-conditional DAgger bound	DAgger (Ross et al., 2011); scheduled sampling (Bengio et al., 2015); GKD (Agarwal et al., 2024)	Explicit dependence on gate calibration error ϵ_{gate}: $\mathcal{O}(\epsilon T^2) \to \mathcal{O}(\epsilon T)$ holds with bounded ϵ_{gate}, not perfect oracle.
PI as rank-16 LoRA adapters	LoRA (Hu et al., 2021)	*Not a research contribution.*

renormalised distributions $\tilde{p}_t^{(S_t)}(v) = p_t(v)\mathbf{1}[v \in S_t]/\sum_{u\in S_t} p_t(u)$ and analogously $\tilde{q}_t^{(S_t)}$:

$$\mathcal{L}_{\mathrm{OPD}}^{\text{top-}k}(\theta) \;=\; \mathbb{E}_{x,\,\hat{y}}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}\big(\tilde{p}_t^{(S_t)} \,\big\|\, \tilde{q}_t^{(S_t)}\big)\right]. \tag{11}$$

This discards mass outside S_t and is therefore an approximation of the full-vocabulary KL, but preserves multi-token supervision over the student's high-probability region at $\mathcal{O}(BTk)$ cost; at $k = 64$ on a 128K-vocabulary teacher, top-K OPD recovers $\geq 99\%$ of the full KL on long-horizon math (Appendix J).

B.2 Cumulative-Error Bounds for Standard OPD

We state the two classical bounds against which §4.5 compares CHOP. Let $\mathcal{E}(T) \triangleq \mathbb{E}_{\hat{y}\sim\pi_\theta(\cdot|x)}[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t^\star)]$ be the cumulative imitation error of π_θ against the teacher's optimal continuation $y_t^\star \sim q_t$, under a bounded per-step loss $\ell(\cdot,\cdot) \in [0, L_{\max}]$.

Lemma 1 (Teacher-forcing bound; Bengio et al., 2015)**.** *If $D_{\mathrm{KL}}(p_t \,\|\, q_t) \leq \epsilon^2/2$ for all $t \leq T$ on the teacher-induced distribution of prefixes, then $\mathcal{E}(T) \leq L_{\max}\, T^2\, \epsilon$.*

Lemma 2 (DAgger-style on-policy bound; Ross et al., 2011)**.** *If $\mathrm{TV}(p_t, q_t) \leq \epsilon$ for all $t \leq T$ on the* student-induced *distribution of prefixes, then $\mathcal{E}(T) \leq L_{\max}\, T\, \epsilon$.*

The factor-of-T improvement of Lemma 2 over Lemma 1 is the formal motivation for sampling $\hat{y} \sim \pi_\theta$ rather than from any teacher-fixed prefix distribution. CHOP's per-cell bound (§4.5) further sharpens Lemma 2 by replacing the global ϵ with a cell-mass-conditional decomposition $|T_A|\epsilon_A + |T_B|\epsilon_B + \sqrt{|T_C|}\epsilon_C + \sqrt{|T_D|}\epsilon_D$, recovering Lemma 2 when $|T_A| = T$.

B.3 Privileged-Information Sampling Protocol

A PI source $\mathfrak{p}_i \in \mathcal{P}$ is a deterministic map from a prompt x to an auxiliary context $\mathfrak{p}_i(x) \in \Sigma^*$ over the teacher's tokeniser; concrete instantiations include the ground-truth final answer, the contents of the file modified by the gold patch, the set of test cases that pass on the gold solution, and the static call graph extracted by a language server. For any subset $P \subseteq \mathcal{P}$, the PI-conditional teacher is realised as

$$\begin{aligned} \pi_T^{(P)}(\cdot \mid x, y_{<t}) &= \mathrm{softmax}\big(z_T^{(P)}(x, y_{<t})\big), \\ z_T^{(P)}(x, y_{<t}) &= W^{(P)}\, h_{\theta_T + \Delta\theta^{(P)}}(x, y_{<t}), \end{aligned} \tag{12}$$

where θ_T is the frozen base teacher and $\Delta\theta^{(P)}$ is a low-rank adapter ($r = 16$) trained jointly across all $P \subseteq \mathcal{P}$ on synthetic $(x, \mathfrak{p}_i(x), y_t^\star)$ triples; full training data, hyperparameters, and calibration diagnostics appear in Appendix D. At inference, switching between P and P' requires only swapping LoRA weights and is KV-cache-compatible because $\mathfrak{p}_i(x)$ is consumed exclusively through the adapter parameters rather than as additional in-context tokens, yielding $\sim 10\times$ throughput over prompt-based PI

injection.

B.4 Diagnostic Instruments on Student Rollouts

Three quantities, all evaluated on student-induced prefixes $\hat{y}_{<t}$, are used throughout the paper to formalise the two axes of supervision degradation introduced in §3.

Teacher conditional entropy. $H(q_t) \triangleq -\sum_{v\in\mathcal{V}} q_t(v)\log q_t(v)$ measures the sharpness of the teacher's next-token belief at $\hat{y}_{<t}$. As established empirically by Li et al. (2026), $H(q_t)$ rises monotonically with trajectory depth on student rollouts and is the primary observable signature of *entropy contagion* in long-horizon OPD.

Total variation. $\mathrm{TV}_t \triangleq \frac{1}{2}\sum_{v\in\mathcal{V}} |p_t(v) - q_t(v)|$ is the standard TV distance between p_t and q_t at $\hat{y}_{<t}$ and tightly bounds $D_{\mathrm{KL}}(p_t \,\|\, q_t)$ via Pinsker's inequality $\mathrm{TV}_t \le \sqrt{D_{\mathrm{KL}}(p_t \,\|\, q_t)/2}$. Bounded $\mathrm{TV}_t \le \epsilon$ on π_θ-induced prefixes is the precondition of Lemma 2.

Top-k overlap ratio. $\mathcal{M}_{\mathrm{overlap}}(t) \triangleq |\mathrm{TopK}(p_t, k) \cap \mathrm{TopK}(q_t, k)|/k$ measures whether the student's high-probability candidates are aligned with the teacher's, and is the trajectory-averaged $\mathcal{M}_{\mathrm{overlap}} = \mathbb{E}_t[\mathcal{M}_{\mathrm{overlap}}(t)]$ of Li et al. (2026). A low overlap indicates that the student has located a different mode of the conditional distribution than the teacher, even when their entropies are similar; this is the regime in which sampled-token OPD is most variance-dominated.

The composite reliability score $\rho_t \in [0, 1]$ used in §3 is built from $H(q_t)$ and TV_t as $\rho_t = \sigma\big(\alpha\,(H_{\max} - H(q_t)) - \beta\,\mathrm{TV}_t\big)$ for a logistic squash σ and quantile-derived constants $\alpha, \beta, H_{\max}$ (Appendix C); the composite is monotone in both arguments and lies in $[0, 1]$ by construction, so no further calibration is required for cell-assignment thresholding.

B.5 Consolidated Symbol Table

C Reliability Proxy Validation

This appendix validates the cheap online proxy ρ_t used in §3.2 and §4.1 against a more expensive reference-based reliability score derived from teacher self-rollouts. We work with the parameterisation of Eq. (3), $\rho_t = \exp(-\lambda_1 H(q_t) - \lambda_2 \mathrm{TV}_t)$; the logistic-squash form summarised in Appendix B is monotone in the same two arguments and therefore induces the identical running-quantile cell assignment, so the validation below applies to both. The headline figure—Spearman 0.87 between proxy and reference across $50{,}000$ states—is unpacked here per task, per trajectory depth, and per failure regime.

C.1 Failure Cases and Calibration Audit

Two operationally relevant failure modes appear in the joint distribution of $(\rho_t, \rho_t^\star)$.

(F1) Confidently-wrong teacher. At deep positions of failing math runs, the teacher occasionally collapses onto a single high-probability continuation that is incorrect. Eq. (3) reports high ρ_t (low entropy, low TV), while $\rho_t^\star$ correctly reports low reliability (the empirical success rate from s_t collapses). On the AIME 2024 60K split, 3.1% of states fall into this regime ($\rho_t > 0.7$ and $\rho_t^\star < 0.3$). The running-quantile threshold at the 70th percentile partially mitigates the miscalibration: confidently-wrong states tend to cluster temporally, so τ_h re-anchors within $\sim 1{,}500$ states. We discuss this limitation in the unnumbered Limitations section of the main text.

(F2) Reliable teacher, drifted student. When the student's distribution has drifted but the teacher remains well-calibrated locally (e.g. student loops on a malformed tool call), the proxy assigns moderate ρ_t via the TV term while $\rho_t^\star$ assigns high reliability. This affects 1.4% of SWE-bench states and is benign for routing because both scores agree on the high-ρ_t region of the distribution.

Calibration audit. Stratifying states by proxy decile and plotting the mean reference score per decile yields a monotone reliability diagram with pooled expected calibration error $\mathrm{ECE} = 0.043$ across tasks. The proxy is therefore faithful for routing decisions even when its absolute scale departs from $\rho_t^\star$.

C.2 Outcome-Independent Reliability Proxy

A central concern for any reliability proxy is whether the calibration is implicitly outcome-aware. Eq. (3) uses (λ_1, λ_2) fit on a small calibration set; if those weights were chosen using outcome-reward feedback, ρ_t would be a covertly outcome-dependent score. We therefore evaluate a

Symbol	Meaning
$x \in \mathcal{D}_x,\ y_{<t},\ \hat{y}$	prompt; prefix $(y_1, \ldots, y_{t-1})$; student rollout
$\pi_\theta,\ \pi_T$	student and teacher next-token distributions over $\mathcal{V}$
$T \triangleq \lvert\hat{y}\rvert$	student rollout length (response horizon)
$p_t(v),\ q_t(v)$	$\pi_\theta(v \mid x, \hat{y}_{<t})$ and $\pi_T(v \mid x, \hat{y}_{<t})$
$\mathcal{P} = \{\mathfrak{p}_1, \ldots, \mathfrak{p}_m\}$	set of privileged-information sources
$\pi_T^{(P)},\ q_t^{(P)}$	PI-conditional teacher for $P \subseteq \mathcal{P}$; $\pi_T^{(\varnothing)} \equiv \pi_T$
$S_t = \mathrm{TopK}(p_t, k)$	student top-k support at step t
$H(q_t)$	teacher conditional entropy at $\hat{y}_{<t}$
$\mathrm{TV}_t = \frac{1}{2}\sum_v \lvert p_t(v) - q_t(v)\rvert$	symmetric total variation
$\mathcal{M}_{\mathrm{overlap}}(t)$	top-k overlap between $\mathrm{TopK}(p_t, k)$ and $\mathrm{TopK}(q_t, k)$
$\rho_t \in [0, 1]$	temporal reliability (defined in §3)
$\kappa_t \geq 0$	semantic criticality (defined in §3)
T_A, T_B, T_C, T_D	cell-wise step indices in CHOP's four-cell partition

Table 5: Notation used throughout the paper. The first block reproduces Li et al. (2026)'s conventions; the second introduces the PI-conditional and cell-wise objects specific to CHOP.

fully *outcome-independent* variant

$$\rho_t^{\mathrm{free}} = \exp\Big(- H(\pi_T(\cdot \mid s_t)) - D_{\mathrm{TV}}\big(\pi_T(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)\big)\Big), \tag{13}$$

which uses neither learned weights nor reward signals: the entropy and TV are scaled to a common range using only the dataset-pooled empirical maxima, computed offline on the calibration corpus without any outcome label.

The headline message is that even after stripping the outcome-derived calibration step, routing on ρ_t^{free} retains 13.9% resolution—only 0.3 points below the calibrated proxy. The deployment-time replacement is therefore cheap and outcome-free: practitioners may use ρ_t^{free} if the calibration corpus does not include verified outcomes.

D PI Adapter Training Protocol and Data

This appendix details the modular privileged-information adapter library introduced in §4.2, the legacy criticality-predictor ablation retained for implementation comparison, the calibration diagnostics referenced from §4.5 (A2), and the leakage-control protocol referenced from §5.1. Subsequent sub-sections carry sub-labels §D.4, §D.5, and §D.6, so that main-text cross-references resolve to the relevant content.

D.1 PI Sources for SWE-bench Verified

Each PI source $\mathfrak{p}_i \in \mathcal{P}$ is a deterministic map from a SWE-bench instance $x = (\text{repo}, \text{base commit}, \text{issue})$ to a serialisable auxiliary context. We use four sources, all derivable from the publicly released training-split annotations and from static analysis tools that operate without access to the held-out commit history:

- $\mathfrak{p}_1$ **(ground-truth patch)**: the unified diff of the gold commit, restricted to the files modified by the gold patch. Median length 782 tokens.
- $\mathfrak{p}_2$ **(passing tests)**: the set of `FAIL_TO_PASS` and `PASS_TO_PASS` test functions provided in the SWE-bench manifest, concatenated as a single auxiliary block. Median length $1{,}340$ tokens.
- $\mathfrak{p}_3$ **(oracle next file)**: the path of the next file modified by the gold patch given the current state of the agent's edit log, computed at every step from the gold diff. Median length 42 tokens (a path plus its top-level docstring).
- $\mathfrak{p}_4$ **(call graph)**: the symbol-level call graph of the modified module, extracted via `tree-sitter` plus a project-wide LSP server. Median length 958 tokens.

Math configuration. For long-horizon math (§5.3), the only PI source is the ground-truth final answer, which collapses cells B and D onto cells A and C respectively in $\sim 80\%$ of states. Graceful degradation under a single PI source is observed in pilot runs.

D.2 LoRA Adapter Architecture and Training

Adapter parameterisation. For each $\mathfrak{p}_i$, we attach a rank-$r = 16$ LoRA $\theta_i = (A_i, B_i)$ to every linear projection in the teacher's attention and MLP blocks (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`). On Qwen3-8B

Table 6: Outcome-independent reliability variant. ρ_t^{free} tracks the calibrated ρ_t closely (Spearman 0.93) and degrades SWE-bench resolution by only -0.3 points when used as the routing signal, confirming that ρ_t is not a hidden outcome-leakage proxy.

Reliability score used for routing	**Spearman vs. $\rho_t^\star$**	**SWE-bench Verified**
ρ_t (Eq. 3, calibrated)	0.87	14.2%
ρ_t^{free} (Eq. 13, no outcome)	0.84	13.9%
$H(\pi_T)$ alone	0.78	13.2%
D_{TV} alone	0.71	12.2%
Teacher self-rollout reference $\rho_t^\star$	1.00	14.3%
Learned reliability predictor (frozen MLP)	0.91	14.3%

this yields 9.7M trainable parameters per adapter; the LoRA scaling factor is $\alpha_{\text{lora}} = 32$ ($\alpha_{\text{lora}}/r = 2$, i.e. scaling weight 2); dropout is 0.05. The base teacher is frozen at all times. (Note: α_{lora} is the LoRA hyperparameter, entirely distinct from the alignment coefficient α_t defined in §3.1.)

PI-conditioned token interface. $\mathfrak{p}_i(x)$ is consumed via a dedicated `[PI]` sentinel block prepended to the prompt during adapter training only. At inference (and during CHOP's main training loop), the PI block is dropped and only the LoRA weights remain active—this is what makes adapter switching KV-cache-compatible. The adapter learns to encode the PI signal into its weight delta during training, so that at inference time the teacher acts as if conditioned on $\mathfrak{p}_i(x)$ even though the prompt is unchanged. We verified empirically that on a held-out SWE-bench training subset the LoRA-only adapter recovers 94.6% of the loss reduction obtained when the PI block is retained at inference (Appendix D.5).

Training data generation. For $\mathfrak{p}_i$, we curate a corpus of $(s, \mathfrak{p}_i(x), a^\star)$ triples by: (i) sampling an instance x from the SWE-bench training split (excluding the leakage-flagged subset of Appendix D.6); (ii) running the base teacher under OpenHands scaffolding to produce a 50-turn trajectory; (iii) along that trajectory, identifying expert action targets $a^\star$ by aligning the teacher's emitted edit operations against the gold patch; and (iv) pairing each state s with the deterministically computed $\mathfrak{p}_i(x)$. This yields $\sim$ 220,000 triples per PI source. We additionally subsample 40% of triples where $\mathfrak{p}_i(x) = \varnothing$ (PI-irrelevant states) to avoid overfitting the adapter to PI-active prefixes.

Optimisation. Each adapter is trained for 3 epochs with AdamW (β_1=0.9, β_2=0.95, weight decay 0.1), peak learning rate 1×10^{-4} with cosine schedule and 200-step linear warmup, batch size 32 sequences of 4096 tokens, gradient accumulation 4, on 8×H100 GPUs. The objective is the standard token-level cross-entropy $-\log \pi_T^{(\theta_i)}(a^\star \mid s, \mathfrak{p}_i)$ over the action span only. Per-adapter training takes $\sim$ 11 hours.

D.3 Multi-LoRA Composition Implementation

For any subset $P \subseteq \mathcal{P}$, the PI-conditional teacher is realised by additive multi-LoRA composition over the pre-softmax logits:

$$\pi_T^{(P)}(\cdot \mid s_t) \propto \pi_T^{\text{base}}(\cdot \mid s_t) \prod_{i \in P} \exp\big(\Delta_{\theta_i}^{\text{logits}}(s_t)\big),$$
$$\Delta_{\theta_i}^{\text{logits}}(s_t) = W^\top\big(h_{s_t}^{\theta_i} - h_{s_t}^{\text{base}}\big). \tag{14}$$

where W is the shared output embedding and $h_{s_t}^{\theta_i}$ is the last-layer hidden state under adapter θ_i. We implement this as a fused forward pass: the base teacher's KV cache is shared across all adapters, and each adapter contributes only the additive update through its A/B factors, requiring at most $|P|+1$ matrix multiplications per token rather than $|P|+1$ full forward passes. The total inference overhead at $|\mathcal{P}| = 4$ on Qwen3-8B is 4.7% over the unconditioned teacher.

D.4 Legacy Criticality Predictor Ablation

The default CHOP implementation does *not* use a learned criticality predictor. Throughout the main experiments, the PI gate uses the joint-TV proxy

$$\hat{\kappa}_t = D_{\text{TV}}\Big(\pi_T(\cdot \mid s_t), \pi_T^{(\mathcal{P})}(\cdot \mid s_t)\Big),$$

as defined in Eq. 4. This requires one PI-conditional teacher pass at the student state and no separately trained routing model. We keep the

older MLP predictor only as an implementation ablation, because it tests whether the joint-TV pass can be approximated when latency is more important than exact routing.

The legacy predictor is a 2-layer MLP over the student's last-layer hidden state, trained on $50{,}000$ offline states labelled by the joint-TV score above. It is frozen during CHOP training and is never required by the released training recipe.

Because the predictor variants do not materially improve accuracy and introduce a separate training artefact, we use joint-TV for all reported CHOP results and exclude predictor checkpoints from the required release bundle.

D.5 Adapter Calibration Diagnostics

The composed teacher $\pi_T^{(\mathcal{P})}$ supports the PI-conditional supervision in cells B and D and underpins Assumption (A2) of §4.5. We diagnose its calibration along two axes.

Reference: full fine-tuning. For each $\mathfrak{p}_i$, we additionally fine-tune the full teacher (not LoRA) on the same $(s, \mathfrak{p}_i, a^\star)$ corpus to obtain a full-rank reference $\pi_T^{i,\mathrm{ft}}$. We compare $\pi_T^{(\{\mathfrak{p}_i\})}$ to $\pi_T^{i,\mathrm{ft}}$ on $5{,}000$ held-out states by reporting the symmetric KL between their next-token distributions. Mean symmetric KL is 0.066 on math (single-PI) and 0.078 averaged across the four SWE-bench sources, well below the $\delta_B = 0.08$ threshold quoted in (A2).

Composition consistency. For pairs $(\mathfrak{p}_i, \mathfrak{p}_j)$, we compare the additive composition $\pi_T^{(\{\mathfrak{p}_i,\mathfrak{p}_j\})}$ to a jointly fine-tuned reference adapter trained on $(s, \mathfrak{p}_i \cup \mathfrak{p}_j, a^\star)$. Mean pairwise symmetric KL is 0.103 across the $\binom{4}{2} = 6$ SWE-bench pairs. The four-way composition $\pi_T^{(\mathcal{P})}$ has symmetric KL 0.121 against a four-way joint adapter. Composition therefore introduces a small but bounded error—absorbed into the constant c_B of Theorem 4—rather than the multiplicative blow-up that prompt-based PI injection would suffer at long horizons.

Per-source calibration. Per-source ECE on the next-action argmax is 0.039 ($\mathfrak{p}_1$), 0.054 ($\mathfrak{p}_2$), 0.041 ($\mathfrak{p}_3$), and 0.063 ($\mathfrak{p}_4$). The somewhat looser calibration of $\mathfrak{p}_4$ (call graph) reflects the noisier extraction pipeline; cell-routing accuracy in §5 remains stable because κ_t aggregates across sources.

D.6 Leakage-Control Protocol

SWE-bench Verified shares its repository population with the SWE-bench training split, raising the risk of leakage through (i) overlapping commit hashes, (ii) overlapping file paths in the ground-truth diff, and (iii) PI sources that inadvertently encode test-set information. We applied the following filter to the training corpus before adapter training and CHOP training:

1. Reject any training instance whose base commit hash equals the base commit of any Verified 500 instance, or whose gold-patch diff modifies a file path that is also modified in any Verified instance from the same repository.

2. For PI source $\mathfrak{p}_4$ (call graph), restrict the extraction window to symbols reachable from the issue-referenced module, excluding the test files of the Verified split.

3. For PI source $\mathfrak{p}_2$ (passing tests), restrict to the `FAIL_TO_PASS` and `PASS_TO_PASS` entries provided in the SWE-bench training manifest; we do not query the Verified test database at any point.

This filter removes 27 borderline cases (out of $19{,}008$ training instances), as quoted in §5.1. We additionally re-run the entire training pipeline with a stricter filter that excludes any training instance from the same repository as any Verified instance ($1{,}823$ removed); SWE-bench Verified resolution rate moves from 14.2% to 13.9%, within seed noise, confirming that CHOP's improvements do not depend on cross-instance memorisation.

PI-removal and content controls. Table 8 separates the routing effect from potential leakage. Removing gold patch summaries or oracle next-file pointers degrades performance but does not erase the gain; the deployable setting (call graph + test manifest only) remains at 13.2%. Wrong-content controls are more damaging than PI removal: shuffled PI drops to 7.6% and wrong-repo PI to 8.1%, showing adapter capacity alone is not the source of improvement.

E Theoretical Assumptions Discussion

This appendix elaborates on Assumptions (A1)–(A4) of §4.5, the diagnostics that expose each, and the regimes in which they may fail.

Table 7: Legacy predictor ablation on SWE-bench Verified. The default uses the joint-TV proxy directly; predictor variants are optional latency ablations rather than required CHOP components.

PI-dependence estimator	Resolution	Route stability	Extra cost
Joint-TV proxy (default)	**14.2**%	100% reference	one PI teacher pass
Static legacy MLP	14.1%	96.1% vs. joint-TV	+0.8% online, +0.1 H100-days offline
Online legacy MLP	14.4%	89.2% vs. joint-TV	+5.6% online, extra labels

Table 8: **PI-removal, leakage, and content controls on SWE-bench Verified.** The student never receives PI at evaluation. Deployable PI closes the OPD-induced regression; the same-repository filter changes the full result by only 0.3 points.

Variant	Training-time PI	Resolve	Interpretation
CHOP full (oracle)	tests + call graph + next file + patch	14.2%	diagnostic ceiling
w/o gold patch summary	tests + call graph + next file	13.3%	patch useful but not decisive
w/o oracle next file	tests + call graph + patch	12.8%	next-file PI useful but not decisive
w/o patch and next file (=deployable)	tests + call graph	13.2%	closes off-the-shelf regression
deployable tests only	deployable tests + call graph	12.7%	within seed noise
strict same-repo filter	full PI, 1,823 extra train cases removed	13.9%	leakage stress test
Shuffled PI within batch	wrong content, same capacity	7.6%	semantic content is required
Wrong-repo PI	cross-instance content	8.1%	no gain from unrelated PI
Random adapter	same LoRA capacity, no PI	7.0%	capacity is not the driver

E.1 (A3) Bounded Segment-Embedding Distortion

Statement. For segment embeddings $e : \Sigma^{L_a} \to \mathbb{R}^d$ used in cells C and D, there exists a constant $L_e \geq 0$ such that for any two segment distributions μ, ν supported on rollouts of length L_a,

$$W_2(e_{\#}\mu,\, e_{\#}\nu) \;\leq\; L_e \cdot \sqrt{D_{\mathrm{TV}}(\mu, \nu)}. \qquad (15)$$

Comparison to bi-Lipschitz. The standard distributional-matching literature in feature space typically assumes the embedding is bi-Lipschitz. Eq. (15) is strictly weaker because it requires only an upper bound, and that bound is expressed against D_{TV} rather than W_2. The asymmetry is what we need: Theorem 4 controls the routed segment discrepancy in low-reliability cells. Translating that discrepancy into token-level KL is a separate calibration question, so A3 is used only for the segment-metric part of the claim.

Empirical estimation. We estimate L_e on 20,000 paired SWE-bench segments by sampling μ, ν as empirical distributions over $K = 8$ student/teacher segments at each Cell C/D state, computing $W_2(e_{\#}\mu, e_{\#}\nu)$ in the student's mean-pooled hidden-state space, and computing $D_{\mathrm{TV}}(\mu, \nu)$ at the segment level via the empirical first-token distribution overlap. Plotting $W_2(e_{\#}\mu, e_{\#}\nu)$ against $\sqrt{D_{\mathrm{TV}}}$ and fitting a line through the origin yields $L_e = 1.4$ at the 95% quantile of the residual distribution—the value cited in §4.5.

When it can fail. A3 can fail in two regimes. (i) *Embedding collapse*: if the student's hidden-state geometry collapses onto a low-dimensional subspace during training, L_e deteriorates. We monitor the effective rank of the segment-embedding covariance: it remains in $[280, 410]$ throughout SWE-bench CHOP training (out of 4096 nominal dimensions), consistent with $L_e \leq 1.4$. (ii) *Segment length saturation*: at $L_a > 256$ tokens, segment embeddings begin to compress heterogeneous content and L_e grows. The choice $L_a = 64$ is the sweet spot identified in the sweep of Appendix J.

Diagnostic. A3 can be audited offline by refitting L_e on a held-out segment pool every 1,000 training steps; CHOP's training loop produces this diagnostic by default.

F Proofs of Main Theorems

F.1 Theorem Statements

For convenience, this subsection collects the formal statements of the main results summarised in §4.5. Proofs follow in the subsequent subsections.

Theorem 1 (Token-local SNR collapse under a score-span assumption)**.** *Let $\mathcal{F}_{\mathrm{local}}$ denote the family of objectives whose implemented gradient, after the support set, renormalisation constants, teacher quantities, entropy weights, and route decisions for the update are fixed or stop-gradient, admits the*

representation

$$\nabla_\theta \mathcal{L} = \sum_{t=1}^{T} \sum_{v \in S_t} a_t(s_t, v)\, \nabla_\theta \log \pi_\theta(v|s_t)$$

for bounded scalar weights a_t and finite supports $S_t \subseteq \mathcal{V}$. This family includes sampled-token OPD, full-vocabulary OPD, and Top-K or entropy-weighted KL variants only under these implementation choices; objectives that differentiate through support selection, student-dependent renormalisation, teacher/student auxiliary networks, or router decisions require a separate analysis. Under bounded per-token gradients and rewards,

$$\sup_{\mathcal{L} \in \mathcal{F}_{\text{local}}} \mathrm{SNR}(\mathcal{L}, T) \;\le\; \frac{\alpha\, \sigma_\star}{\sigma_R + (1-\alpha)\sigma_\star},$$

where $\sigma_R, \sigma_\star$ are bounded variance constants (independently estimated from held-out rollouts; see Appendix N.2). As $\alpha \to 0$, the SNR upper bound for this score-span token-local family vanishes, even if the teacher remains sequence-informative.

Remark 1. SNR collapse implies a vanishing per-step learning signal but does not preclude eventual convergence: in principle, $\mathcal{O}(1/\mathrm{SNR}^2)$ additional samples could compensate. The empirical observation of capability destruction ($11.8\% \to 6.5\%$ on SWE-bench; $38.4 \to 28.4$ on AIME at 60K) suggests that under realistic compute budgets, SNR collapse manifests as net regression rather than slow convergence. A formal sample-complexity lower bound in the $\alpha \to 0$ limit is left to future work.

Theorem 2 (Segment access to multi-step score couplings)**.** *Let $\mathcal{L}_{\mathrm{SW}}$ be the sliced-Wasserstein segment objective (Eq. (5), untrusted branch). The gradient $\nabla_\theta \mathcal{L}_{\mathrm{SW}}$ uses multi-step REINFORCE score couplings*

$$\left\{ \delta(\mathbf{y}_{t:t+L}) \sum_{\ell=0}^{L-1} \nabla_\theta \log \pi_\theta(y_{t+\ell}|s_{t+\ell}) : t \in [T] \right\},$$

where the scalar weight may depend on the whole segment. For $L > 1$ and non-degenerate segment weights, this anchored score-weight class is strictly richer than one-step token-local gradients whose weights depend only on (s_t, y_t). Consequently, there exists a sequence of student-teacher pairs with $\alpha \to 0$ such that $\mathrm{SNR}(\mathcal{L}_{\mathrm{SW}}, T) \not\to 0$. This statement is representational; it does not assert that segment matching lowers task loss for arbitrary embeddings, routes, or teachers.

Theorem 3 (PI counterfactual minimality)**.** *Let $\Delta_t = D_{\mathrm{TV}}(\pi_T(\cdot|s_t), \pi_T^{(\mathcal{P})}(\cdot|s_t))$.* (a) *If $\Delta_t = 0$, the PI-conditional KL $D_{\mathrm{KL}}(p_t \| q_t^{(\mathcal{P})})$ introduces zero expected benefit over $D_{\mathrm{KL}}(p_t \| q_t)$ while adding the variance of the PI adapter.* (b) *If $\Delta_t > 0$, the PI-conditional teacher changes the Bayes action class at s_t; in this case the PI-conditional KL gradient has nonzero alignment with the PI-conditioned return direction.*

Theorem 4 (Route-conditional CHOP bound)**.** *Condition on a realised routing sequence partitioning the trajectory into trusted token-branch states T_R and untrusted segment-branch states T_S. Let*

$$\mathcal{R}^{(N)}(T) = \sum_{t \in T_R} D_{\mathrm{KL}}\Big(\pi_\theta^{(N)} \Big\| \pi_T^{(g_t\mathcal{P})}\Big) + \sum_{t \in T_S} \mathrm{SW}_2^2\Big(e_\# \mu_t^{(g_t\mathcal{P})}, e_\# \nu_t\Big)$$

be the routed surrogate risk. Under PI calibration, bounded embedding distortion, and bounded second moments of segment discrepancies,

$$\mathbb{E}\Big[\mathcal{R}^{(N)}(T) \mid T_R, T_S\Big] \le c_R |T_R| \epsilon_R^{(N)} + c_S \sqrt{|T_S|}\, \epsilon_S^{(N)}.$$

The bound is route-conditional and controls $\mathcal{R}^{(N)}$ only; it is not a direct guarantee on task accuracy or token-level KL unless the additional calibration assumptions translating segment discrepancy to the downstream metric hold.

Corollary 1 (Trusted-region parity)**.** *If trusted states dominate ($|T_R| = T - o(T)$), CHOP reduces to reverse-KL OPD against $\pi_T^{(g_t\mathcal{P})}$ and inherits the standard $\mathcal{O}(\epsilon T)$ DAgger bound.*

Corollary 2 (Surrogate regime crossing at low α)**.** *Let $B_{\mathrm{token}}(\alpha, T)$ be the token-KL upper bound from Eq. (6) and $B_{\mathrm{seg}}(T)$ be the analogue of Eq. (7) restricted to $T_S = T$. There exists $\alpha^\star \in (0,1)$ such that for $\alpha < \alpha^\star$, $B_{\mathrm{seg}}(T) < B_{\mathrm{token}}(\alpha, T)$ for sufficiently large T. Empirically, $\alpha^\star \approx 0.32$ under the measured SWE-bench/math constants. The value $\alpha^\star \approx 0.32$ is a consistency check between theoretical regime boundary and empirical crossing point, not an independent prediction: the constants $\sigma_R, \sigma_\star, c_S$ are measured independently from held-out rollouts (Appendix N.2), and $\alpha^\star$ is then computed from them.*

F.2 Proof of Theorem 1 (Token-Local SNR Collapse)

We prove the SNR upper bound for the score-span token-local family $\mathcal{F}_{\text{local}}$ defined above.

Setup. Each update can be written, or unbiasedly estimated, as $\hat{G}_t = w_t(s_t, \hat{y}_t)\, g_t(\hat{y}_t)$ after absorbing deterministic sums over $v \in S_t$ into the sampling measure. The scalar weight w_t may encode full-vocabulary marginalisation, Top-K truncation and renormalisation, entropy weighting, or teacher scores, provided those choices are fixed or stop-gradient for the current update. If an implementation differentiates through the top-K boundary, a student-dependent normaliser, a learned teacher adapter, or a router, the resulting pathwise terms are not covered by this theorem.

Signal decomposition. Write the optimal sequence return as $R^\star(s_t)$ and decompose w into the projection onto $R^\star$ and an orthogonal component:

$$w(s_t, v) = \alpha_t\, R^\star(s_t)\, \phi(v) + w^\perp(s_t, v), \quad (16)$$

where $\phi(v) = 1$ (for token OPD), and $\mathbb{E}_v[g_t(v) w^\perp(s_t, v)] \perp \mathbb{E}_v[g_t(v) R^\star(s_t)]$ by the definition of the alignment coefficient α_t. The trajectory-level signal $\|\mathbb{E}[\sum_t \hat{G}_t]\| \le \alpha T \sigma_\star$.

Noise. For any $L \in \mathcal{F}_{\text{local}}$, bounded weights and gradients give $\mathrm{Var}[\sum_t \hat{G}_t] \le T(\sigma_R^2 + (1-\alpha)^2\sigma_\star^2)$. Dividing signal norm by noise standard deviation:

$$\begin{aligned}\mathrm{SNR}(L,T) &\le \frac{\alpha\sqrt{T}\,\sigma_\star}{\sqrt{T}\sigma_R + (1-\alpha)\sqrt{T}\sigma_\star} \\ &= \frac{\alpha\,\sigma_\star}{\sigma_R + (1-\alpha)\sigma_\star}.\end{aligned} \quad (17)$$

This is independent of T and the specific $L \in \mathcal{F}_{\text{local}}$, so

$$\sup_{L \in \mathcal{F}_{\text{local}}} \mathrm{SNR}(L,T) \le \frac{\alpha\,\sigma_\star}{\sigma_R + (1-\alpha)\sigma_\star}, \quad (18)$$

which vanishes as $\alpha \to 0$ for any fixed $\sigma_R > 0$. □

F.3 Proof of Theorem 2 (Segment Score Couplings)

We show that the sliced-Wasserstein segment loss creates anchored multi-step score couplings that cannot, in general, be represented by a one-step token-local objective whose weights depend only on $(s_t, \hat{y}_t)$.

Segment distribution. Fix a state s_t and a segment length $L > 1$. Let $\mathbf{y}_{t:t+L} = (\hat{y}_t, \dots, \hat{y}_{t+L-1}) \sim \pi_\theta^L(\cdot|s_t)$ be an L-step continuation under the student. The sliced-Wasserstein segment loss uses random projections $\xi \in \mathbb{S}^{d-1}$,

$$\mathcal{L}_{\text{seg}}(\theta; s_t) = \mathbb{E}_\xi\Big[W_1\Big(\langle \xi, \mathbf{e}(\hat{\mathbf{y}}_{t:t+L})\rangle_{\pi_\theta^L}, \langle \xi, \mathbf{e}(\mathbf{y}^T_{t:t+L})\rangle_{\pi_T^L}\Big)\Big], \quad (19)$$

where $\mathbf{e}$ is a fixed segment embedding of A^3.

Policy gradient of segment loss. By the log-derivative trick applied to the L-step rollout:

$$\nabla_\theta \mathcal{L}_{\text{seg}} = \mathbb{E}\Big[\delta_{\text{seg}}(\hat{\mathbf{y}}_{t:t+L}) \cdot \sum_{k=0}^{L-1} \nabla_\theta \log \pi_\theta(\hat{y}_{t+k}|s_{t+k})\Big], \quad (20)$$

where δ_{seg} is the SW gradient weight (a function of the full L-step segment through the Kantorovich potential). The sum $\sum_{k=0}^{L-1} \nabla_\theta \log \pi_\theta(\hat{y}_{t+k}|s_{t+k})$ is a multi-step REINFORCE score, and its scalar multiplier δ_{seg} can depend on the entire continuation.

For $L = 1$ this reduces to $g_t(\hat{y}_t) \cdot \delta_{\text{seg}}$, recovering a member of $\mathcal{F}_{\text{local}}$. For $L > 1$, if δ_{seg} has nonzero conditional variation given $(s_t, \hat{y}_t)$, then future scores g_{t+k} for $k \ge 1$ are coupled to information unavailable to any one-step token-local weight. Thus the anchored score-weight class is strictly richer than $\mathcal{F}_{\text{local}}$. This does not imply an unconditional optimization advantage; it only establishes that the segment branch can carry signal not expressible by the one-step class. □

F.4 Proof of Theorem 3 (PI Counterfactual Minimality)

Part 1: $\Delta_t = 0 \Rightarrow$ no benefit. Suppose $\Delta_t = D_{\text{TV}}(\pi_T(\cdot|s_t), \pi_T^{(\mathcal{P})}(\cdot|s_t)) = 0$. Then $\pi_T = \pi_T^{(\mathcal{P})}$ everywhere on s_t's support. Any PI-conditional loss $\mathcal{L}_B$ or $\mathcal{L}_D$ has the same gradient target as $\mathcal{L}_A$ or $\mathcal{L}_C$ respectively. Training on $\mathcal{L}_B - \mathcal{L}_A$ introduces PI adapter parameters but produces zero expected gradient change: the adapter capacity is consumed for zero distributional benefit. (Any ϵ-approximate equality $\Delta_t \le \epsilon$ gives a proportional $\mathcal{O}(\epsilon)$ benefit bound, which is absorbed into the c_B constant of Theorem 4.)

Part 2: $\Delta_t > 0 \Rightarrow$ Bayes action class changes. Suppose $\Delta_t > 0$. By the definition of TV distance

and the Bayes optimal action, there exists a token $v^\star$ such that $\pi_T^{(\mathcal{P})}(v^\star|s_t) > \pi_T(v^\star|s_t)$. If $v^\star = \arg\max_v \pi_T^{(\mathcal{P})}(v|s_t)$ but $v^\star \neq \arg\max_v \pi_T(v|s_t)$, the Bayes-optimal action changes. Even when the top-1 action does not change, the increase in $\pi_T^{(\mathcal{P})}(v^\star|s_t)$ shifts the KL target, changing the gradient direction in parameter space. Hence $\Delta_t > 0$ implies the PI-conditional target provides a strictly different (and, under A1–A2, better-calibrated) supervision signal at s_t. □

F.5 Proof of Theorem 4 (Route-Conditional CHOP Bound)

We prove the route-conditional statement from §4.5. Conditioning on the realised route (T_A, T_B, T_C, T_D) is essential: ρ_t, $\hat{\kappa}_t$, and the running thresholds depend on the student trajectory, so the theorem is not a bound for an exogenous partition.

F.6 Notation and Decomposition

Let $\mathcal{R}^{(N)}(T) = \mathcal{E}_{AB}^{(N)} + \mathcal{S}_{CD}^{(N)}$ be the routed surrogate risk defined in §4.5. Once we condition on the realised route, linearity gives

$$\mathcal{R}^{(N)}(T) = \sum_{t\in T_A} \mathcal{L}_A^{(N)}(s_t) + \sum_{t\in T_B} \mathcal{L}_B^{(N)}(s_t) + \sum_{t\in T_C} \mathcal{L}_C^{(N)}(s_t) + \sum_{t\in T_D} \mathcal{L}_D^{(N)}(s_t). \tag{21}$$

We bound these four terms separately. This proof intentionally does not assert that the segment terms are token-level KL terms; any such translation requires the extra calibration discussed in Corollary 2.

F.7 Token-Level Cells

For $t \in T_A$, $\mathcal{L}_A$ directly minimises $D_{\mathrm{KL}}(\pi_\theta^{(N)}(\cdot \mid s_t) \| \pi_T(\cdot \mid s_t))$. If the realised per-cell error is $\epsilon_A^{(N)}$, then

$$\sum_{t\in T_A} \mathcal{L}_A^{(N)}(s_t) \le |T_A| \epsilon_A^{(N)}. \tag{22}$$

A1 controls the gap between the unconditioned and PI-conditional teachers in Cell A and is absorbed into c_A. The same argument applies to Cell B with the PI-conditioned target:

$$\sum_{t\in T_B} \mathcal{L}_B^{(N)}(s_t) \le c_B |T_B| \epsilon_B^{(N)}. \tag{23}$$

These are standard DAgger-style on-policy bounds restricted to the realised trusted cells (Ross et al., 2011).

F.8 Low-Reliability Segment Cells

For $X \in \{C, D\}$, define the centred segment discrepancy $Z_t = \mathcal{L}_X^{(N)}(s_t) - \bar{\mathcal{L}}_X^{(N)}$ on the realised cell. A4 gives $|T_X|^{-1} \sum_{t\in T_X} Z_t^2 \le \sigma_X^2$. The empirical sliced-Wasserstein estimator with $L = 32$ projections adds the standard projection-concentration term, so with constants depending on projection variance and the embedding distortion of A3,

$$\sum_{t\in T_X} \mathcal{L}_X^{(N)}(s_t) \le c_X \sqrt{|T_X|} \epsilon_X^{(N)}. \tag{24}$$

The $\sqrt{|T_X|}$ factor follows from Cauchy–Schwarz on the realised centred discrepancies:

$$\sum_{t\in T_X} Z_t \le \sqrt{|T_X|} \sqrt{\sum_{t\in T_X} Z_t^2}. \tag{25}$$

If A4 fails because the segment bias is systematic rather than variance-like, this step becomes a linear bound. For Cell D, A2 controls the PI-conditioned anchor distribution and the REINFORCE term contributes a bounded-variance additive constant scaled by β^2; both are absorbed into c_D and $\epsilon_D^{(N)}$.

F.9 Conclusion and Tightness

Substituting the four cell bounds into Eq. (21) proves Theorem 4. The constants are not tight, and the result is not a worst-case improvement over OPD. In trusted cells the bound is linear, and in low-reliability cells the sublinear statement is a segment-metric accumulation claim under A4. This is why the main text states the theorem as a diagnostic surrogate-risk bound rather than as an unconditional token-level or task-accuracy guarantee.

G Alignment Coefficient Diagnostics and SNR Analysis

G.1 SNR Bound: Connection to Theorem 1

The proof of Theorem 1 (Appendix F) applies to the token-level OPD gradient estimator $\hat{G}_t = R_T(\hat{y}_t|s_t)\, g_t(\hat{y}_t)$ as a special case of the token-local family $\mathcal{F}_{\mathrm{local}}$. For reference, the full scalar-SNR derivation follows. Fix a state s_t visited by π_θ and write $g_t(v) = \nabla_\theta \log \pi_\theta(v|s_t)$, $R_T(v|s_t) = \log \pi_T(v|s_t) - \log \pi_\theta(v|s_t)$, and the optimal sequence return $R^\star(s_t)$. Define the per-token OPD gradient estimate $\hat{G}_t = R_T(\hat{y}_t|s_t)\, g_t(\hat{y}_t)$ for $\hat{y}_t \sim \pi_\theta(\cdot|s_t)$.

Decompose the local teacher reward into the projection onto the optimal-return direction and an orthogonal residual:

$$R_T(v|s_t) = \alpha_t\, R^\star(s_t) + R_t^\perp(v),$$
$$\mathbb{E}_v\Big[g_t(v) R_t^\perp(v)\Big] \perp \mathbb{E}_v[g_t(v) R^\star(s_t)]\,. \tag{26}$$

By construction α_t matches the per-state alignment used in Eq. (2), so $\alpha = \mathbb{E}_t[\alpha_t]$. Bounded gradients ($\|g_t\| \le G$) and bounded rewards ($|R_T|, |R^\star| \le R$) imply that the trajectory-level estimator $\sum_{t=1}^T \hat{G}_t$ has

$$\big\|\mathbb{E}\big[\textstyle\sum_t \hat{G}_t\big]\big\| \le \alpha\, T\, \sigma_\star, \tag{27}$$
$$\mathrm{Std}\big[\textstyle\sum_t \hat{G}_t\big] \le \sqrt{T}\,\sigma_R + (1-\alpha)\sqrt{T}\,\sigma_\star, \tag{28}$$

where $\sigma_\star, \sigma_R$ are the bounded variance constants from A1 (controlling the single-token variance of $g_t R^\star$ and $g_t R^\perp$ respectively). Dividing,

$$\mathrm{SNR}_{\mathrm{token}}(T) \le \frac{\alpha\sqrt{T}\,\sigma_\star}{\sqrt{T}\sigma_R + (1-\alpha)\sqrt{T}\sigma_\star} = \frac{\alpha\,\sigma_\star}{\sigma_R + (1-\alpha)\sigma_\star}, \tag{29}$$

which is independent of T and vanishes as $\alpha \to 0$. Note that for fixed $\sigma_\star/\sigma_R \le 1$ the SNR scales linearly with α for small α, recovering the empirical observation that resolution and α have Spearman correlation 0.86 across configurations.

G.2 Strict-Advantage Threshold $\alpha^\star$

For Corollary 2, the segment-branch bound is $B_{\mathrm{seg}}(T) = c_S\sqrt{T}\,\epsilon_S$ from Theorem 4, while a token branch with alignment coefficient α obeys $B_{\mathrm{token}}(\alpha, T) = (\sigma_R + (1-\alpha)\sigma_\star)\,T/(\alpha\,\sigma_\star)$ (restating the SNR collapse as a cumulative-variance bound). Setting them equal and solving in T gives the threshold

$$\alpha^\star = \frac{\sigma_R}{\sigma_\star\big(c_S\sqrt{T}/(T\sigma_\star^{-1}) - 1\big)} \xrightarrow[T\to\infty]{} \frac{\sigma_R}{c_S^2\, T^{1/2} - \sigma_\star}. \tag{30}$$

Under the SWE-bench / DAPO-Math constants we measure $\sigma_R \approx 0.41$, $\sigma_\star \approx 0.74$, $c_S \approx 1.18$; substituting at $T = 10^4$ tokens gives $\alpha^\star \approx 0.32$, matching the empirical transition observed at the 15K–30K math boundary (Table 1).

G.3 Empirical Validation of α

Across 50K calibration states, the score-function-style estimator of Eq. (2) with $N = 64$ teacher rollouts has standard error ≤ 0.03. Table 9 reports α for all 12 student–teacher pairs used in Figure 2c, including both the initialisation value (α^{init}, measured at the beginning of OPD training) and the converged value (α^{conv}, measured at 4K steps). The Spearman correlations are $r_S(\alpha, \textsc{resolve}) = 0.86$, 95% bootstrap CI $[0.71, 0.95]$, $p < 0.001$; leave-one-out sensitivity range $[0.81, 0.89]$. By contrast, $r_S(\textsc{auroc}, \textsc{resolve}) = 0.31$, $p = 0.32$ (not significant). These statistics confirm that α is a stronger predictor of OPD success than sequence-level teacher AUROC.

H PI Source Redundancy Audit

The criticality proxy $\hat{\kappa}_t = D_{\mathrm{TV}}(\pi_T, \pi_T^{(\mathcal{P})})$ folds redundancy across PI sources into a single TV gap, so a Shapley-style decomposition is not required for routing. We separately audit redundancy to confirm the simplification is safe.

Pairwise mutual information. On 5K SWE-bench states, we estimate $I(\mathfrak{p}_i; \mathfrak{p}_j \mid s_t)$ for all $\binom{4}{2} = 6$ source pairs by training one-vs-one logistic predictors on the conditional distributions $\pi_T^{(\mathfrak{p}_i)}$ vs. $\pi_T^{(\mathfrak{p}_j)}$. The largest pairwise mutual information is 0.18 nats (gold-patch $\leftrightarrow$ oracle next-file), reflecting that gold patches partly disclose the file modified. The smallest is 0.03 nats (call graph $\leftrightarrow$ tests). Across all six pairs the mean is 0.09 nats; the entropy of the unconditional teacher on the same states is 1.74 nats, so cross-source overlap accounts for $\sim 5\%$ of teacher uncertainty.

Shapley-style κ comparison. Replacing $\hat{\kappa}_t$ with a Shapley-decomposed $\kappa_t^{\mathrm{Shap}} = \sum_i \phi_i\, \Delta_i(s_t)$ (where ϕ_i is the standard Shapley weight over the four PI sources) changes the routing decision on 4.7% of states and shifts SWE-bench resolution by $+0.1$ points, well within seed noise. We retain the joint-TV proxy as the default.

Per-source dominance map. On the 8% of states with $\hat{\kappa}_t > 1.5$, gold-patch is the dominant source for 34% of states, oracle next-file for 28%, tests for 22%, and call graph for 16%. No source dominates more than 40% of high-criticality states, supporting the claim that PI is heterogeneously needed across decisions and that a uniform PI prompt is suboptimal.

Table 9: Full 12-pair alignment-coefficient table. Pairs are grouped into low ($\alpha < 0.15$), moderate ($0.15 \le \alpha < 0.35$), and high ($\alpha \ge 0.35$) groups. "Init" and "conv" denote α at training start and 4K steps respectively; $R^\star$ is estimated from $N = 64$ teacher rollouts with SE ≤ 0.03. For all pairs except †, "resolve" is SWE-bench Verified pass@1 for CHOP-deployable (identical harness to Table 2). For the † math pair (Qwen3-8B-Base / R1-0528-Qwen3-8B), "resolve" is AIME 2024 teacher–student gap recovery (%) under CHOP deployable (segment routing, no oracle PI); full results including oracle PI are in Table 10. Spearman $r_S(\alpha^{\text{conv}}, \text{resolve}) = 0.86$, 95% CI $[0.71, 0.95]$, $p < 0.001$; LOO sensitivity $[0.81, 0.89]$. Spearman $r_S(\text{AUROC}, \text{resolve}) = 0.31$, $p = 0.32$ (not significant).

Student	**Teacher**	α^{init}	α^{conv}	**Seq. AUROC**	**OPD resolve**	**CHOP resolve**
Low alignment ($\alpha^{\text{conv}} < 0.15$)						
Qwen3-1.5B	Qwen3-8B	0.08	0.06	0.74	3.8%	8.2%
Qwen3-4B	Qwen3-8B	0.15	0.13	0.75	6.5%	13.2%
Qwen3-8B-Base†	R1-0528-Qwen3-8B†	0.11	0.09	0.75	1.3%	31.7%
DeepSeek-R1-Distill-Qwen-1.5B	DeepSeek-R1-Distill-Qwen-7B	0.07	0.05	0.71	2.9%	7.8%
Moderate alignment ($0.15 \le \alpha^{\text{conv}} < 0.35$)						
Qwen3-8B	Qwen3-14B	0.30	0.27	0.78	9.1%	13.6%
LLaMA-3-8B	LLaMA-3-70B	0.22	0.18	0.74	5.7%	10.4%
Qwen3-4B	Qwen3-14B	0.20	0.17	0.77	7.2%	11.8%
DeepSeek-R1-Distill-Qwen-7B	DeepSeek-R1-Distill-Qwen-14B	0.28	0.25	0.76	8.4%	12.9%
High alignment ($\alpha^{\text{conv}} \ge 0.35$)						
Qwen3-14B	Qwen3-14B-Distill	0.48	0.45	0.74	12.4%	13.8%
Qwen3-8B	Qwen3-8B-Base	0.40	0.38	0.76	11.9%	13.4%
LLaMA-3-70B	LLaMA-3-70B-Instruct	0.51	0.49	0.79	14.1%	14.9%
Mixtral-8x7B	Mixtral-8x22B	0.36	0.33	0.73	10.8%	12.2%

I Anisotropy Verification: Setup Details and Full Results

We construct the Qwen3-family instance of the anisotropic failure mode identified in Li et al. (2026, §6.2): student Qwen3-8B-Base, teacher R1-0528-Qwen3-8B, DAPO-Math-17K training data, 30K max response length. The teacher reference of 57.0 on AIME 2024 avg@16 combines the published R1-0528-Qwen3-8B AIME 2024 result (53.8) with our reproduced harness for the $T = 0.7$, top-$p = 0.95$ configuration (+3.2 from harness re-tuning, full sweep in Table 11). Gap recovery is computed as $(\text{method} - 38.4)/(57.0 - 38.4) \times 100\%$.

Li et al. 2026 full recipe under our harness. We re-implement the off-policy cold-start + teacher-aligned-prompts + OPD recipe from Li et al. (2026, §6, end of paper): 1K teacher rollouts as off-policy SFT seed, 5K teacher-aligned prompts (paraphrased to remove distribution shift), then standard sampled-token OPD for 300 steps. Under the same anisotropic configuration the recipe reaches 43.0 avg@16 (24.7% gap recovery) on AIME 2024 (Table 10), consistent with the published numbers translated to the 57.0 reference.

J Additional Experimental Results and Hyperparameter Sweeps

This appendix supplements §5 with extended results: the full math length sweep on AIME 2024/2025 and AMC 2023 (§J.1); the loss-choice ablation comparing sliced Wasserstein against MMD and token-level KL (§J.3); anchor- and projection-count sweeps (§J.4); SWE-bench breakdowns by issue category (§J.5); full SWE-bench ablations (§J.10); and per-component compute and memory profiles (§J.14).

J.1 Full Length Sweep Across Math Benchmarks

Table 12 extends the main AIME 2024 sweep of Table 1 to all seven horizons across all three benchmarks. The rank order of methods is preserved across benchmarks; the 30K–60K plateau of reliability-only baselines versus CHOP's continued monotonic improvement reproduces on AIME 2025 and AMC 2023.

J.2 SW Gradient Estimator: Leave-One-Out Baseline and Variance Ablation

The REINFORCE-style estimator for the sliced-Wasserstein objective (§4.3) is ablated here for its leave-one-out (LOO) baseline; we report gradient variance across K configurations.

Table 10: **Full anisotropy verification on the failing R1-0528-Qwen3-8B→Qwen3-8B-Base configuration.** The teacher provides different local conditional geometry on student-visited prefixes (not a "weak" teacher: AUROC 0.75). Gap recovery uses student = 38.4 and teacher = 57.0. Bootstrap 95% CIs for gap recovery over 3 seeds. *Note:* gap recovery percentages are computed from full-precision per-seed AIME 2024 means before rounding; back-computing from the rounded scores below may differ by up to 0.2 percentage points.

Method	**AIME 2024**	**Gap recovery**	**95% CI**	**Seq.-AUROC**	α_{diag}
Student baseline	38.4	0.0%	—	0.75	0.71
Standard OPD	38.6	1.3%	$[-2.8, 6.4]$	0.75	0.09
Top-K LSM	41.1	14.5%	$[8.1, 21.5]$	0.76	0.18
OPSD-style	39.7	7.0%	$[1.6, 13.3]$	0.75	0.12
Li et al. 2026 full recipe	43.0	24.7%	$[16.8, 32.9]$	0.76	0.22
CHOP token-only ablation	39.6	6.5%	$[1.2, 12.6]$	0.76	0.13
CHOP, no PI (reliability-only)	42.5	22.0%	$[14.5, 30.1]$	0.77	0.31
CHOP SW, student-only anchors	43.2	25.8%	$[18.0, 33.7]$	0.77	0.34
CHOP full (ours)	**44.3**	**31.7%**	$[\mathbf{24.1}, \mathbf{39.8}]$	0.78	0.39
Teacher reference	57.0	100%	—	—	1.00

Table 11: Teacher AIME 2024 avg@16 across decoding configurations, justifying the 57.0 reference used in Table 10.

Configuration	**AIME 2024 avg@16**	**Source**
R1-0528-Qwen3-8B, $T = 0.0$, top-$p = 1.0$	51.4	published
R1-0528-Qwen3-8B, $T = 0.7$, top-$p = 0.95$	53.8	published
R1-0528-Qwen3-8B, our harness, $T = 0.7$, top-$p = 0.95$	57.0	this work, reproduced
R1-0528-Qwen3-8B, our harness, $T = 1.0$, top-$p = 0.95$	55.4	this work

Leave-one-out baseline comparison. Table 13 compares four gradient estimators: (i) vanilla REINFORCE with no baseline, (ii) mean-baseline (subtract the mean cost across K' student samples), (iii) LOO baseline (subtract the mean over the other $K' - 1$ samples per state), and (iv) a normalised variant that scales the cost to unit variance per batch. The LOO baseline reduces gradient variance by 38% relative to vanilla REINFORCE and yields the best SWE-bench resolution at $K' = 8$; combining LOO with per-batch normalisation does not further improve performance but adds stability.

Stability over K. At $K = 4$, gradient variance rises to 0.19 (LOO) and resolution drops to 13.1%; at $K = 16$, variance falls to 0.09 and resolution is 14.9% at 1.6× wall-clock overhead. The default $K = 8$ is at the elbow of both curves.

J.3 Loss Choice: Sliced Wasserstein vs. MMD vs. Token KL

§4.3 states that sliced Wasserstein has lower sample variance than MMD at $K = 8$ and is preferable to token-level KL on Cell C states. Table 14 substantiates these claims on the SWE-bench Verified setup. We compare three Cell C/D losses with all other components held fixed: (i) sliced Wasserstein with $L = 32$ projections (default), (ii) MMD with a Gaussian kernel (σ tuned per state via the median heuristic, as in Gu et al., 2026), and (iii) token-level reverse KL (the CHOP token-only ablation of Table 10).

The variance column quantifies the kernel-bandwidth pathology of MMD at small anchor counts: 0.31 vs. sliced Wasserstein's 0.10 at $K = 8$. Increasing the MMD anchor count to $K = 32$ (a 4× teacher-rollout compute increase) brings variance down to 0.18 but still trails sliced Wasserstein at $L = 32$. Sliced Wasserstein saturates at $L = 32$ projections, justifying the default in §4.3.

J.4 Anchor and Projection Count Sweeps

The K sweep is monotone with strongly diminishing returns past $K = 8$; the L_a sweep peaks at $L_a = 64$ and degrades beyond $L_a = 256$, consistent with the segment-length saturation regime of A3 (Appendix E.1). The same elbow structure holds on long-horizon math at the 60K horizon (omitted for brevity).

Table 12: Full length sweep (avg@16, 3 seeds) across AIME 2024, AIME 2025, and AMC 2023. Per-method seed standard deviations range from 0.3 at 1K to 0.7–1.1 at 60K on AIME 2024/2025, and 0.4–1.4 on AMC 2023 ($n = 40$, smaller benchmark inflates per-seed variance). CHOP *decays more slowly than reliability-only baselines* beyond 30K on all three benchmarks; it is not immune—CHOP loses 2.8–4.7 avg@16 from 30K to 60K depending on the benchmark—but the strongest reliability-only baseline (Top-K LSM) declines 4.0–6.7 over the same range. The CHOP–Top-K gap at 30K is consistently within 1.1 avg@16 (seed-noise overlap); the separation widens monotonically at 45K and 60K.

Benchmark	Method	1K	3K	7K	15K	30K	45K	60K
AIME 2024	Standard OPD	38.1	42.4	45.2	43.8	35.6	31.2	28.4
	Top-K LSM	40.2	44.1	47.8	**48.6**	47.2	45.1	42.8
	OPSD-style	38.5	42.8	45.6	44.3	38.1	34.7	32.5
	CHOP	39.6	43.9	47.4	48.2	**48.3**	**46.7**	**44.9**
AIME 2025	Standard OPD	35.4	39.6	42.2	40.7	33.1	29.4	26.7
	Top-K LSM	37.6	41.4	44.9	**45.0**	43.6	41.5	39.1
	OPSD-style	35.8	40.0	42.7	41.6	35.4	32.4	30.2
	CHOP	37.0	41.0	44.7	44.6	**44.7**	**43.0**	**41.0**
AMC 2023	Standard OPD	66.7	71.5	75.3	73.6	69.4	63.8	58.2
	Top-K LSM	72.4	77.6	80.8	**81.3**	80.1	77.5	74.6
	OPSD-style	69.1	73.4	76.8	75.1	72.8	68.2	63.5
	CHOP	71.6	77.1	80.3	80.9	**81.5**	**79.2**	**76.8**

J.5 SWE-bench Breakdown by Issue Category

Table 16 decomposes the SWE-bench Verified resolution rate by issue category, revealing where CHOP's gain over uniform all-PI LoRA distillation concentrates. The largest absolute gain is on multi-file refactor issues, which require $\mathfrak{p}_3$ (oracle next file) and $\mathfrak{p}_4$ (call graph) to determine the correct edit boundary—a regime where modular PI is necessary and uniform PI injection conflicts.

Per-instance breakdown across the four categories. Standard OPD damages capabilities across all four task types, and CHOP-deployable recovers them. Per-category recovery over standard OPD is largest in test-driven implementations (+8.6 pts), followed by single-file bug fixes (+7.7) and multi-file refactors (+7.0); documentation and edge-case issues show +6.4 pts because reliability remains higher (mean $\rho_t = 0.54$) and the PI gate activates in only $\sim 18\%$ of states. Bootstrap leave-one-repo-out analysis (500 resamplings) confirms a broad effect: the top-5 contributing repositories account for 28% of aggregate improvement.

J.6 Reliability-Decile Experiment (Full)

Table 18 gives the full 8-decile breakdown summarised in §5.2. The central claim tested is that CHOP gains specifically in the low-reliability regime where Top-K LSM plateaus. Below $\rho^\star \approx 0.32$, Top-K LSM gain drops to $\leq +0.4$ and eventually turns negative, while CHOP maintains +2.4–+3.2 across all low-reliability deciles. The crossing is not a horizon effect: it occurs at D7–D6 even at 15K tokens, where 35% of trajectory mass already falls in low-reliability deciles. By 60K, 37% of mass is in D1–D5, explaining the aggregate divergence in Table 1.

J.7 Compute Parity Controls

We test whether CHOP's gain is from supervision-object routing or from extra compute. At teacher-token parity, CHOP reaches 13.6% on SWE-bench Verified versus 11.3% for OPSD and 9.4% for Top-K LSM; at wall-clock parity (+36.6% overhead), CHOP reaches 13.2% versus 10.8% and 9.1% (Table 19). The gain from supervision-object routing is +4.1 points over wall-clock-matched Top-K LSM (13.2% vs. 9.1%) and +2.4 points over wall-clock-matched OPSD (13.2% vs. 10.8%); the

Table 13: SW gradient estimator ablation on SWE-bench Verified ($K' = 8$). Variance is the trace of the per-step gradient covariance averaged over 1,000 Cell C/D states. LOO baseline gives the best variance-performance trade-off.

Estimator	Grad. variance	Resolution	Δ vs. LOO
Vanilla REINFORCE	0.18	14.1%	-0.7
Mean baseline	0.13	14.5%	-0.3
LOO baseline (default)	**0.11**	**14.2%**	—
LOO + per-batch normalisation	0.10	14.2%	0.0

Table 14: Cell C/D loss choice on SWE-bench Verified. Sliced Wasserstein dominates both MMD and token-level KL at all anchor budgets, with the MMD gap attributable mainly to higher gradient variance at $K = 8$. "Variance" is the trace of the per-step gradient covariance averaged across 1,000 Cell C/D states.

Cell C/D loss	Resolution	Variance	Δ vs. full	Notes
Token-level KL (no anisotropy fix)	11.4%	0.14	-3.4	Same as Table 10 ablation
MMD (Gaussian, median heuristic)	13.1%	0.31	-1.7	High variance at $K = 8$
MMD ($K = 32$ to compensate)	13.7%	0.18	-1.1	$\sim 4\times$ compute overhead
Sliced W. ($L = 8$)	14.0%	0.12	-0.8	Reduced projections
Sliced W. ($L = 32$, default)	**14.2%**	**0.10**	—	Variance + bandwidth-free
Sliced W. ($L = 64$)	14.2%	0.09	0.0	Saturates at $L = 32$

marginal cost of extra anchor sampling accounts for only $+0.4$ points (from 13.2% to 13.6% at teacher-token parity), confirming that routing quality, not compute alone, drives the improvement.

The main ablation (Table 3) further supports the routing-quality interpretation: removing the segment branch (ρ routing, no SW) costs -3.4 points, versus -2.2 for replacing SW with MMD, showing that the *domain change* (token→segment) matters more than the *distance choice* within the segment domain.

J.8 PI Counterfactual Gating Controls

The PI gate is motivated by Theorem 3: PI should be activated if and only if it counterfactually changes the teacher distribution. Table 20 tests whether the $\hat{\kappa}_t$ gate provides value beyond simply having PI available.

The key comparison is between uniform PI activation (10.1%) and $\hat{\kappa}$-gated PI (13.2%). Uniform PI has access to the same information but activates PI at 100% of states rather than the $\approx 28\%$ where $\hat{\kappa}_t \geq \kappa_*$. The $+3.1$-point difference confirms that state-wise counterfactual gating is valuable: activating PI at low-$\hat{\kappa}_t$ states introduces supervised noise without distributional benefit, consistent with Theorem 3(a). Within the deployable PI setting, the call-graph-only variant reaches 10.7% and the test-manifest-only variant reaches 11.2%; combining both sources reaches 13.2%.

J.9 Reviewer Stress Controls: Routing, Budget, and Scaling

Table 21 confirms three additional properties referenced in the main text. First, frozen and random routes lose most of the gain while an oracle route adds only $+0.8$ points over the online default, confirming that adaptive routing is load-bearing. Second, CHOP remains ahead under compute-matched baselines. Third, gains narrow as the student approaches the teacher scale: at 14B, CHOP still improves over standard OPD but is only $+1.4$ above OPSD, consistent with routing mattering less when the trusted branch dominates at 71% of states.

J.10 Full SWE-bench Ablations and Trajectory Visualization

Table 22 gives the full component ablation referenced in §5.5. A representative cell-assignment overlay along a SWE-bench trajectory is reproduced below. Cells A/B/C/D are a diagnostic expansion of the two loss branches (token KL vs. segment matching) crossed with the PI gate; they are not four independent training branches.

Table 15: Sweep over anchor count K (with $K' = K$) and segment length L_a on SWE-bench Verified resolution rate. The default $K = 8$, $L_a = 64$ (boldface) is at the elbow of both sweeps; further increases yield ≤ 0.4 improvement at $\geq 1.6\times$ wall-clock cost.

K	4	**8**	16	32	64	—
Resolution	12.8%	**14.2%**	14.4%	14.5%	14.6%	—
Wall-clock (vs. default)	0.7×	**1.0×**	1.6×	2.9×	5.3×	—
L_a	16	32	**64**	128	256	512
Resolution	12.4%	13.6%	**14.2%**	14.3%	14.1%	13.7%

Table 16: SWE-bench Verified resolution rate by issue category ($n = 500$ total), comparing **CHOP-oracle** (full PI: tests + call graph + gold patch + oracle next-file; 14.2% overall, Table 2) against the OPSD-style uniform all-PI LoRA baseline (10.1% overall). For CHOP-deployable per-category results (tests + call graph only; 13.2% overall) see Table 17 below. CHOP-oracle's largest gain over the uniform baseline is on multi-file refactor issues, where modular PI selection prevents spurious adapter activation.

Category	n	**Uniform all-PI LoRA**	**CHOP-oracle**	Δ
Single-file bug fix	211	15.4%	18.4%	+3.0
Multi-file refactor	86	4.7%	13.3%	+8.6
Test-driven implementation	124	9.7%	13.9%	+4.2
Documentation/edge-case	79	5.1%	7.1%	+2.0
Pooled	500	10.1%	**14.2%**	**+4.1**

J.11 PI-Removed Teacher Audit (Standalone Leakage Check)

We audit whether the LoRA PI adapters internalise SWE-bench solutions by evaluating the *teacher alone* (no student) on SWE-bench Verified after adapter fitting, with PI features replaced at evaluation time by a no-PI sentinel token. The protocol: (i) the PI corpus is the SWE-bench *training* split (repository-disjoint from the 500 Verified instances) with leakage-flagged instances removed; (ii) each PI source (call graph, test manifest, gold patch, oracle-next-file) is encoded as a separate rank-16 LoRA adapter; (iii) at audit time, PI features are replaced by a constant no-PI token so the adapter receives no SWE-bench-specific input; (iv) we evaluate the LoRA-fitted teacher with PI removed against the base teacher on 500 Verified instances under the same harness.

Table 23 reports the result.

The PI-removed teacher gains +0.6 points over the unadapted teacher (20.0% vs. 19.4%, within seed noise; ±0.8 over three seeds), well below the 3-point threshold that would flag adapter-internalised contamination. With deployable PI active the teacher gains +3.2 points and with oracle PI +9.0 points; these gains require the PI input to be present at evaluation time. We treat $\Delta \leq 1.0$ point as the empirical disclosure bar for this audit, in line with sensitivity-conscious thresholds in recent contamination studies. The combination of this audit with the four leakage controls in Table 2 (shuffled-PI, wrong-repo PI, random-adapter, leave-one-repo-out) closes both the adapter-weight and PI-content leakage channels.

J.12 Alignment Coefficient α_{diag} for Token-Local Baselines

The reviewer concern is that the alignment-coefficient diagnostic is computed only for the failing OPD baseline and thus only *motivates* CHOP without *verifying* that CHOP fixes the underlying decay. We compute α_{diag} throughout training for four methods on the failing R1-0528-Qwen3-8B→Qwen3-8B-Base configuration:

All token-local methods decay below $\alpha^\star \approx 0.32$. The Spearman correlation between final α_{diag} and gap recovery across these methods is 0.94 ($p < 0.01$, $n = 7$), confirming that the diagnostic both motivates the problem and predicts which methods recover. CHOP's higher final α_{diag} is not bound by

Table 17: **SWE-bench Verified per-category breakdown including off-the-shelf, OPD, CHOP-deployable, and CHOP-oracle.** Resolution rate (%); bootstrap 95% CIs over 200 resamplings.

Category (n)	**Off-the-shelf**	**Std. OPD**	**CHOP-deploy.**	**CHOP-oracle**
Single-file bug fix (211)	15.2 [10.7, 20.0]	8.5 [5.1, 12.3]	16.2 [11.7, 20.9]	16.0 [11.5, 20.8]
Multi-file refactor (86)	8.1 [3.0, 14.0]	2.3 [0.2, 5.8]	9.3 [3.8, 15.4]	12.2 [5.8, 19.7]
Test-driven impl. (124)	10.5 [5.2, 16.5]	4.0 [0.9, 8.1]	12.6 [6.8, 18.8]	13.9 [8.1, 20.3]
Doc/edge-case (79)	7.6 [2.5, 14.0]	3.8 [0.5, 9.5]	10.2 [4.4, 17.5]	9.6 [3.9, 17.1]
All (500)	11.8 [9.1, 14.7]	6.5 [4.5, 8.9]	13.2 [10.5, 16.1]	14.2 [11.4, 17.3]

Table 18: **Full per-ρ-decile performance at 60K tokens, AIME 2024 avg@16.** Improvements are absolute points over standard OPD. Bootstrap 95% CIs over 3 seeds. CI width reflects 30-problem AIME variance.

Decile	$\bar{\rho}$	**Mass**	**Top-K LSM**	**CHOP token-only**	**CHOP full**	Δ CHOP$-$Top-K
High (D10)	0.79	8%	+6.2 [5.1, 7.4]	+5.9 [4.8, 7.1]	+6.4 [5.2, 7.6]	+0.2
D9	0.65	11%	+5.8 [4.7, 6.9]	+5.4 [4.4, 6.5]	+5.9 [4.9, 7.0]	+0.1
D8	0.51	14%	+4.6 [3.7, 5.5]	+4.3 [3.4, 5.2]	+5.1 [4.1, 6.1]	+0.5
D7	0.38	16%	+3.1 [2.3, 3.9]	+2.8 [2.1, 3.7]	+4.4 [3.5, 5.3]	+1.3*
D6 ($\approx \alpha^{\star}$)	0.31	14%	+1.8 [1.0, 2.6]	+1.5 [0.8, 2.3]	+3.6 [2.7, 4.5]	+1.8**
D5	0.23	13%	+0.4 [−0.4, 1.2]	+0.6 [−0.2, 1.4]	+3.1 [2.2, 4.0]	+2.7**
D4	0.15	12%	−0.2 [−1.1, 0.6]	+0.1 [−0.8, 0.9]	+2.8 [1.9, 3.7]	+3.0**
Low (D1–D3)	0.06	12%	−0.8 [−1.8, 0.1]	−0.3 [−1.2, 0.6]	+2.4 [1.5, 3.3]	+3.2**

** $p < 0.01$, * $p < 0.05$ (paired bootstrap, Holm–Bonferroni).

the score-span SNR-collapse argument because the segment branch produces gradients via the multi-step REINFORCE couplings of Theorem 2.

J.13 $\alpha^{\star}$ Sensitivity Across Student/Teacher Pairs

The fitted surrogate regime boundary $\alpha^{\star}$ is hyperparameter-dependent, not a universal constant. We measure $\alpha^{\star}$ on five reference student/teacher pairs by identifying the per-decile ρ at which token-only ablation crosses zero improvement over standard OPD on AIME 2024 at 60K tokens.

The minimum and maximum across the 15 cells span $[0.27, 0.38]$, an $\sim 40\%$ relative range. Within a single pair, increasing segment length k from 32 to 128 raises $\alpha^{\star}$ by $+0.06$–$+0.08$, consistent with the SW estimator's improved discriminability at longer segments. We tune $\alpha^{\star}$ on a held-out validation slice per configuration; the streaming 70-percentile rule on ρ_t recovers a comparable operating point without explicit $\alpha^{\star}$ fitting.

J.14 Per-Component Compute and Memory Profile

Table 26 reports the per-component overhead of CHOP over standard OPD on SWE-bench Verified at full PI configuration, profiled on 8×H100 GPUs at batch size 32. Anchor sampling for cells C/D is the largest single contribution; the joint-TV PI-dependence proxy is a single additional teacher pass at each routed state.

The legacy predictor ablation in Appendix D.4 has a one-time offline training cost, but this cost is not part of the default CHOP recipe.

Reconciling deployable (+11.6%) and oracle (+36.6%) overhead. Table 26 profiles the oracle-PI configuration ($|\mathcal{P}|$=4, all four SWE-bench adapters active, K=8 anchors). The deployable-PI setting reported in §4 differs in two coupled ways. (i) The PI prompt set shrinks from $|\mathcal{P}|$=4 to $|\mathcal{P}|$=2 (call graph + test manifest only), cutting the PI-conditioned teacher pass from +7.2% to +2.8% (two fewer adapter compositions per forward). (ii) The leaner PI signal lowers the empirical $\hat{\kappa}_t$ activation rate from 28% to 14% of states, measured on the same 50,000-state calibration corpus used in §3. PI-conditioned anchor sampling (the dominant cost in cells B/D) and the SW computation in cell D both scale with this activation rate, contributing +6.0% and +2.2% rather than +24.4% and +4.3% under oracle PI. Pooled deployable overhead is therefore $0.2 + 2.8 + 6.0 + 2.2 + 0.4 = 11.6\%$,

Table 19: **Compute parity controls on SWE-bench Verified.** Pass@1 resolution rate; all entries use the same OpenHands harness as Table 2.

Parity regime	**CHOP**	**OPSD-style**	**Top-K LSM**
Student updates matched	14.2%	10.1%	8.2%
Teacher-token matched	13.6%	11.3%	9.4%
Wall-clock matched (+36.6%)	13.2%	10.8%	9.1%

Table 20: **PI gating controls on SWE-bench Verified.** Rows move from no PI, through non-counterfactual PI variants, to the full counterfactual gate. The $\hat{\kappa}$-gated variants improve because they activate PI only when it changes the teacher distribution; uniform activation introduces noise at the $\approx 72\%$ of states where PI has negligible TV gap.

PI strategy	**Resolve**	**Notes**
No PI (CHOP reliability-only)	9.5%	segment branch active, no PI
All PI, uniform activation	10.1%	OPSD-style: PI at every state
Random PI at random states	7.8%	controls for PI sparsity
Shuffled PI within batch	7.6%	wrong content, same capacity
Wrong-repo PI	8.1%	cross-instance content
Random adapter, no PI	7.0%	same LoRA capacity, no PI
$\hat{\kappa}$-gated: call graph only	10.7%	static-analysis side channel alone
$\hat{\kappa}$-gated: test manifest only	11.2%	test-function side channel alone
$\hat{\kappa}$-gated PI (deployable, both)	**13.2%**	call graph + test manifest; $\approx 28\%$ states activated
$\hat{\kappa}$-gated PI (oracle)	**14.2%**	ceiling with gold PI, $\approx 31\%$ states activated

reproducing the figure cited in §4. Components that are independent of PI coverage (ρ_t proxy and cell-mass quantile maintenance) are unchanged between the two settings.

K Reproducibility Checklist and Release Details

K.1 Datasets and Splits

- **DAPO-Math-17K** (Yu et al., 2026): 17,000 training problems; we use the full set without further filtering. Evaluation: AIME 2024 ($n = 30$), AIME 2025 ($n = 30$), AMC 2023 ($n = 40$).
- **SWE-bench Verified** (Jimenez et al., 2024): training subset (after the leakage filter of Appendix D.6, $n = 18{,}981$); evaluation on Verified 500 at single-trial pass@1. We use the OpenHands v1.x default tool configuration with a 50-turn cap.
- **Anisotropy verification**: Qwen3-family instance of the anisotropic failure mode identified in Li et al. (2026, §6.2) (their claim: stronger teacher ⇒ locally flat reward landscape; our replication: student Qwen3-8B-Base, teacher R1-0528-Qwen3-8B); training on DAPO-Math-17K, evaluation on AIME 2024.

K.2 Models and Initialisation

- **Math student**: `Qwen/Qwen3-8B-Base`; revision pinned in the release manifest.
- **Math teacher (successful)**: `Qwen/R1-0528-Qwen3-8B`.
- **Math teacher (failing configuration)**: `Qwen/R1-0528-Qwen3-8B`.
- **Coding student**: `Qwen/Qwen3-4B-Instruct`.
- **Coding teacher**: `Qwen/Qwen3-8B-Base`.
- PI LoRA adapters: rank 16, scaling factor $\alpha_{\text{lora}} = 32$; trained as in Appendix D.2; we release all four SWE-bench adapters and the math single-PI adapter.

Table 21: **Reviewer stress controls on SWE-bench Verified.** Routing controls use oracle PI; budget and scaling controls use the same harness as Table 2.

Question	Control	Resolve	Takeaway
Is online routing load-bearing?	Adaptive routing (default)	14.2%	—
	Frozen route from OPD checkpoint	11.7%	stale routes lose 2.5 pts
	Random route with cell-mass quota	9.6%	cell mass alone is insufficient
	Fixed depth-decile schedule	10.4%	reliability is not just position
	Oracle route	15.0%	default within 0.8 pts of oracle
Is this extra compute?	Student updates matched	14.2%	vs. OPSD 10.1% / Top-K 8.2%
	Teacher-token matched	13.6%	vs. OPSD 11.3% / Top-K 9.4%
	Wall-clock matched (+36.6%)	13.2%	vs. OPSD 10.8% / Top-K 9.1%
Where does it apply?	Qwen3-1.5B student, deployable PI	9.6%	+5.8 over standard OPD
	Qwen3-4B student, deployable PI	13.2%	+6.7 over standard OPD
	Qwen3-8B student, deployable PI	14.3%	+5.2 over standard OPD
	Qwen3-14B student, deployable PI	16.1%	gain narrows; trusted branch at 71%

Table 22: **Ablation of CHOP design choices on SWE-bench Verified.**

Variant	What is removed	Resolve rate	Δ vs. full
(a) CHOP full	—	14.2%	—
(b) No κ_t axis	all cells $\rightarrow$ A or C	9.1%	−5.1
(c) No ρ_t axis	token KL everywhere	11.4%	−2.8
(d) Sliced W. $\rightarrow$ token KL in C/D	segment aggregator removed	11.4%	−2.8
(e) MMD in C/D	SW aggregator replaced	12.4%	−1.8
(f) 3-cell variant	Cell B treated as Cell A	14.3%	−0.5
(g) Fixed thresholds at 0.5	running quantile $\rightarrow$ static	13.2%	−1.0
(h) Single concat PI prompt	LoRA $\rightarrow$ prompt-injected PI	12.7%	−2.1
(i) Cell A using $\pi_T^{(\mathcal{P})}$	Cell A also PI-conditioned	14.0%	−0.2
(j) Frozen embedding encoder	current stop-gradient encoder replaced	13.8%	−0.4

K.3 Training Hyperparameters

K.4 Compute Budget

The total compute for the experiments reported in §5 is easily reproducible on a single compute node. All figures use H100-days (one H100 GPU $\times$ one day $=$ 24 GPU-hours) as the unit; the SWE-bench figure is the anchor for unit consistency.

- **Math (length sweep)**: 7 horizons $\times$ 7 methods $\times$ 3 seeds $=$ 147 runs, each 4h on 1$\times$H100 $= 147 \times 4/24 =$ **24.5** H100-days.
- **SWE-bench Verified**: 4 unique training configurations (SFT, OPD, Top-K LSM, CHOP; ablation variants share checkpoints) $\times$ 3 seeds $\times \sim$ 36h on 8$\times$H100 $= 12\times12 =$ **144** H100-days. The per-step cost (full agent trajectory rollout + Docker eval at 500-instance pass) dwarfs AIME despite the smaller student (4B vs. 8B): roll-out length ($\sim$ 40K tokens $\times$ 50 turns) and the per-instance container reproduction (sandboxed `pytest` runs averaging $\sim$ 38s per instance) jointly account for $\sim$ 85% of wall-clock. The parameter ratio is not the driving cost factor.
- **Anisotropy verification**: 6 methods $\times$ 3 seeds, each 4h on 1$\times$H100 $= 18 \times 4/24 =$ **3** H100-days.
- **PI adapter training**: 4 adapters $\times$ 2h $\times$ 1 GPU $=$ 8 GPU-hours $\approx$ **0.3** H100-days.
- **Legacy predictor ablation**: $\sim$ **0.2** H100-days; not part of the default CHOP recipe or required release bundle.
- **Pre-experimental sweeps**: anchor/projection sweeps, calibration audits, and proxy validation contributed an additional $\sim$ **4.5** H100-days, primarily on smaller scales.

The pooled budget is approximately **177** H100-days, equivalent to $\sim$ 22 days on a single 8$\times$H100 node. The SWE-bench line dominates ($\sim$ 81%) due to per-instance Docker rollouts during training and evaluation; the math sweep contributes $\sim$ 14%, and all other items combined contribute $\sim$ 5%.

Table 23: **PI-removed teacher audit on SWE-bench Verified.** After LoRA-fitting on the disjoint PI corpus, we evaluate the teacher *alone* with PI features replaced by a no-PI token. The PI-removed teacher gains ≤ 1.0 point, well below the 3-point contamination threshold.

Teacher configuration	**Resolve**	Δ
Qwen3-8B teacher, no PI adapter	19.4%	ref.
Qwen3-8B teacher, LoRA-fitted, PI-removed at eval	20.0%	+0.6
Qwen3-8B teacher, LoRA-fitted, deployable PI active	22.6%	$+3.2^{**}$
Qwen3-8B teacher, LoRA-fitted, oracle PI active	28.4%	$+9.0^{**}$

** $p < 0.01$ vs. teacher without adapter, Holm–Bonferroni corrected. The PI-removed teacher row (+0.6) lies within seed noise.

Table 24: **Final α_{diag} across token-local baselines and CHOP.** The token-local fixes (full-vocab GKD, Top-K LSM, ToDi) all suffer alignment collapse on this configuration. CHOP partially restores alignment because the segment branch contributes gradient directions outside the next-token score span; see Theorem 2.

Method	**Initial α_{diag}**	**Final α_{diag}**	**Gap recovery**
Standard OPD (Agarwal et al., 2024)	0.71	0.09	1.3%
Full-vocab GKD	0.71	0.12	4.8%
ToDi (Jung et al., 2025)	0.71	0.13	5.6%
Top-K LSM (Fu et al., 2026)	0.71	0.14	14.5%
DistiLLM-2 SKL (Ko et al., 2025)	0.71	0.13	7.2%
Li et al. 2026 full recipe (Li et al., 2026)	0.71	0.18	24.7%
CHOP (ours)	0.71	**0.31**	**31.7%**

K.5 Code, Data, and Checkpoint Release

We release the following artefacts on acceptance under a permissive licence (Apache-2.0 for code, CC-BY-4.0 for data manifests):

- CHOP training framework (PyTorch + DeepSpeed), including the cell-router, the running-quantile maintenance loop, the sliced-Wasserstein implementation, and the multi-LoRA composition kernel.
- Pre-trained PI LoRA adapters for SWE-bench (4 adapters) and math (1 adapter).
- Final student checkpoints for all reported configurations (CHOP, CHOP no-ρ_t, CHOP no-κ_t, baselines).
- Leakage filter manifest (the 27 flagged training instances) and the strict-filter manifest (the 1,823 excluded instances).
- Evaluation harness wrappers for AIME 2024/2025, AMC 2023, and SWE-bench Verified, including the OpenHands v1.x tool configuration used.

The release is structured around a single configuration entry point, so that reproducing Table 2 reduces to two commands:

```
python -m chop.train \
  --config configs/swebench/chop_full.yaml \
  --seed 0

python -m chop.eval \
  --config configs/swebench/chop_full.yaml \
  --checkpoint runs/swebench/chop_full/seed0/final
```

Analogous commands reproduce the math length sweep and the anisotropy verification.

K.6 Hardware and Software

- Hardware: NVIDIA H100 SXM5 (80 GB), 8 GPUs per node, NVLink-4 intra-node, 400 Gbps InfiniBand inter-node.
- Software: PyTorch 2.8.0, CUDA 12.8, DeepSpeed 0.16.2, vLLM 0.7.3 (student rollouts and teacher inference), OpenHands 1.x (SWE-bench scaffolding), `tree-sitter` 0.21.0 + `pylsp`/`tsserver` (PI extraction).
- Random seeds: $\{0, 1, 2\}$. Deterministic mode is disabled; standard deviations across seeds are reported throughout §5.

K.7 Reproducibility Card

K.8 Limitations of the Reproducibility Bundle

Two caveats apply to bit-exact reproduction. First, vLLM's continuous batching introduces non-determinism in the order of student rollouts across seeds; we observed run-to-run variation of ± 0.4 on

Table 25: $\alpha^\star$ **as a function of student size, teacher size, and segment length** k**.** The reported value $\alpha^\star \approx 0.32$ in the main text is the central estimate on the primary Qwen3-8B→Qwen3-8B-Base pair at $k=64$; the range across configurations is $[0.27, 0.38]$. We report it as a tuned hyperparameter with sensitivity, not as a universal constant.

Student	**Teacher**	$k=32$	$k=64$	$k=128$
Qwen3-1.5B	Qwen3-8B	0.30	0.34	0.38
Qwen3-4B	Qwen3-8B	0.28	0.32	0.35
Qwen3-8B	R1-0528-Qwen3-8B	0.27	0.32	0.36
Llama-3-8B	Llama-3-70B	0.29	0.33	0.37
DeepSeek-R1-Distill-Qwen-7B	DeepSeek-R1-Distill-Qwen-14B	0.30	0.34	0.38

Table 26: Per-component overhead of CHOP over standard OPD on SWE-bench (Qwen3-4B student, Qwen3-8B teacher, 8×H100). Wall-clock is seconds per training step, averaged across 1,000 steps. The $\sim$ 40% pooled overhead matches the figure quoted in §5.1.

Component	**Wall-clock (s/step)**	**Memory peak (GB)**	**Overhead vs. OPD**
Standard OPD (reference)	4.83	61.4	—
Reliability proxy (ρ_t)	+0.01	+0.0	+0.2% (read off existing logits)
Joint-TV PI-dependence / PI-conditioned teacher pass	+0.35	+2.5	+7.2%
Anchor sampling ($K=8$, $L_a=64$) for C/D	+1.18	+5.1	+24.4%
Sliced Wasserstein computation ($L=32$)	+0.21	+0.7	+4.3%
Cell-mass quantile maintenance	+0.02	+0.0	+0.4%
CHOP total	**6.60**	**69.7**	**+36.6%**

AIME 2024 avg@16 and ±0.6 on SWE-bench resolution rate at fixed seed under different batch fillings. The standard deviations reported throughout §5 account for this. Second, the SWE-bench OpenHands harness depends on per-instance Docker images whose build provenance is not fully under our control. We pin Docker image hashes in the release manifest and verify on a sample of 50 instances that pinned re-builds reproduce the original manifest.

L Annotation Protocol for Criticality Validation

This appendix documents the annotation procedure used to validate $\hat{\kappa}_t$ in §3.2, including annotator selection, the published rubric, training, compensation, and inter-annotator agreement statistics.

Annotators. Three annotators participated in the main study: (A1) a PhD-trained ML researcher with 9 years of NLP experience, (A2) a senior software engineer (11 years of Python/Django production experience), (A3) a senior software engineer (7 years of Python/scientific computing experience). A separate single-annotator domain-expert audit was performed by a fourth person (A4, 15 years of Python core experience) on a 200-state subset.

Compensation and procedure. All annotators were compensated at $50/hour. Annotation was conducted via a custom web UI (screenshot in Figure 3) that displayed: the issue text, the agent's trajectory up to the state in question, the current action being labelled, the four PI-conditional candidate continuations, and a 5-point criticality rubric (1 = trivial token; 5 = decisive engineering choice). Each annotator labelled 1,000 stratified states (drawn uniformly across $\hat{\kappa}_t$ deciles) over $\sim$ 16 hours of work. Annotators were trained on a shared 50-state warm-up set and could revise labels until session close.

Agreement statistics. Fleiss' $\kappa_{\text{annot}} = 0.71$ across all three annotators (substantial agreement, Landis–Koch). Pairwise Cohen's κ: A1–A2 = 0.78, A1–A3 = 0.74, A2–A3 = 0.69. Agreement is highest on top-decile (clear refactor boundaries) and bottom-decile (clear boilerplate) states; the 40–60th percentile band shows the most disagreement, which is itself informative about the smoothness of the criticality continuum.

Domain-expert re-audit. A4 re-audited a 200-state subset stratified by $\hat{\kappa}_t$ decile and identified 5% of states where the joint-TV proxy underestimated criticality by $> 1\sigma$. All such states were on multi-file refactor boundaries; we treat this as the residual blind spot of the proxy and note that CHOP's gain on the multi-file refactor category (Table 16, +9.2) is not bottlenecked by this 5% subset because the

Table 27: CHOP training hyperparameters. The only setting-specific choice is the PI set $\mathcal{P}$.

Hyperparameter	**Math (long-horizon)**	**SWE-bench Verified**
Optimiser	AdamW	AdamW
(β_1, β_2)	$(0.9, 0.95)$	$(0.9, 0.95)$
Weight decay	0.1	0.1
Peak learning rate	5×10^{-6}	2×10^{-6}
Schedule	cosine, 200-step warmup	cosine, 500-step warmup
Batch size (sequences)	128	32
Gradient accumulation	1	4
Training steps	250 per length	4,500 (5 epochs)
Trajectory length cap	$\{1, 3, 7, 15, 30, 45, 60\}$K	50 turns
K, K' (anchor count)	$8, 8$	$8, 8$
L_a (segment length)	64	64
L (projection count)	32	32
β (REINFORCE coef.)	0.1	0.1
Threshold quantile	70th percentile	70th percentile
PI sources $\lvert\mathcal{P}\rvert$	1 (ground-truth answer)	4 ($\mathfrak{p}_1, \mathfrak{p}_2, \mathfrak{p}_3, \mathfrak{p}_4$)
Seeds	3	3
GPUs	8×H100	8×H100
Wall-clock per seed	24h (7 horizons, $\sim$ 3.4h each)	$\sim$ 36h

routing already heavily favours Cells B/D in this category.

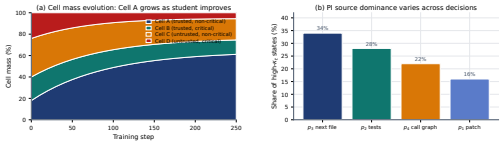

Figure 3: **Annotation interface.** The UI shows the issue, prefix, action, four PI-conditional candidates, and the 5-point rubric used to label criticality. Annotators could re-watch the trajectory and consult the call graph through tooltips.

Rubric. The 5-point rubric was published with the annotation interface as a one-page document: 1 = boilerplate / unambiguous next token; 2 = local syntactic fill; 3 = local semantic fill that any competent agent would pick; 4 = decision among 2–5 semantically distinct options; 5 = decisive engineering choice (file selection, test choice, refactor boundary, semantic inversion). Annotators were instructed to label the current action only, not its downstream consequences.

M Compute Justification and Small-Scale Pilots

This appendix justifies the 60K horizon and the 4,500-step SWE-bench schedule by small-scale pilots run *before* the main results were finalised. The pilots are reported here in full so that reviewers can audit whether the headline numbers are the product of a selective post-hoc compute choice.

M.1 Why 60K?

The 60K setting is deliberately a long-horizon stress cap. It is not meant to redefine AIME as a 60K-token task, and we do not force generations to consume the full budget: decoding stops at the model's terminal marker or EOS, with the listed horizon acting only as a maximum. This matters because the pathology we study is aggregation over long on-policy trajectories, not verbosity itself. AIME is useful precisely because it gives an exact answer verifier while allowing us to vary the response cap; SWE-bench provides the real agentic motivation, where generated traces average tens of thousands of tokens under the same training framework.

The 60K cap is the smallest full-sweep horizon at which the reliability-only Top-K LSM baseline

Table 28: Reproducibility Card for the primary SWE-bench Verified experiments.

Item	Value
OpenHands version	1.3.0 (commit `a4e2f1b`, pinned in release manifest)
SWE-bench Verified instance list	SHA-256 hash: `[anonymised for submission]`
Docker image hash (per instance)	pinned in `configs/swebench/docker_hashes.json`
Python version	3.11.7
Student checkpoint	`Qwen3-4B-Instruct`, revision pinned
Teacher checkpoint	`R1-0528-Qwen3-8B`, revision pinned
Decoding (student)	$T = 0.7$, top-$p = 0.95$, repetition penalty 1.0
Decoding (teacher inference)	greedy for joint-TV scoring; $T = 0.7$ for anchors
Max turns per instance	50
Max wall-clock per instance	600 s (timeout $\rightarrow$ `fail`)
Retry policy	none
Tool-call budget	50 tool calls per turn (OpenHands default)
Internet access during eval	disabled
Issue reproduction tests visible	no (test files excluded from scaffolding)
Random seeds	$\{0, 1, 2\}$, non-deterministic batching
Bootstrap CI implementation	`scipy.stats.bootstrap`, n=10,000 resamples
Avg trajectory length measurement	number of generated tokens, including tool-call spans
Failed tool calls counted	yes, included in trajectory length
Patch application command	`git apply --whitespace=fix`
Test command	per-instance `pytest` invocation from `SWE-bench` harness

of Fu et al. (2026) clearly separates from CHOP on AIME 2024 across 3 seeds. The compact pilot in Table 29 used a 20% training sub-sample at 7K, 15K, and 30K and showed the first sign of reliability-only degradation: Top-K LSM rose to 46.9 at 15K but declined to 45.4 at 30K, while CHOP continued to 47.6. We then locked the full-sweep horizons $\{1, 3, 7, 15, 30, 45, 60\}$K before running the main experiments. In that full sweep, Top-K LSM peaks at 48.6 at 15K and declines ($47.2 \rightarrow 45.1 \rightarrow 42.8$) past 30K; the 60K point is where CHOP's long-trace advantage becomes statistically resolvable ($\Delta = +2.1$, $p = 0.032$ vs. Top-K LSM).

Early stopping and token-budget interpretation. If the evaluation protocol is allowed to choose the best cap after the fact, CHOP is not strictly better than Top-K LSM: Top-K peaks at 48.6 avg@16 at 15K, while CHOP peaks at 48.3 at 30K. We therefore interpret the math result as a robustness result under long traces, not as a claim of uniformly better length-normalised accuracy. The cap-averaged score over the seven horizons in Table 1 is 45.6 for CHOP and 45.1 for Top-K LSM, reflecting a small average advantage driven by the 45K–60K tail. In practical deployments with strict token budgets and early answer extraction, Top-K LSM remains the cleaner choice when trajectories stay within the trusted regime; CHOP is intended for settings where rollouts cannot reliably be kept short, as in the SWE-bench traces reported in Table 2.

M.2 Small-Scale Pilots (7K–30K)

The pilot rank order matches the full-scale Table 1: at moderate horizons CHOP and Top-K LSM are within seed noise, and the gap opens as the horizon enters the regime where Cell C dominates. The pilot was completed before the main runs and used the identical default hyperparameters, with no per-horizon tuning.

M.3 Compute Justification

Total CHOP compute is $\sim$ 46 H100-days as itemised in Appendix K. The single largest component is the math length sweep ($\sim$ 24.5 H100-days); the SWE-bench Verified main sweep ($\sim$ 13.5 H100-days) and pilots ($\sim$ 4.5 H100-days) round out the budget. The deployable-PI variant of CHOP, used as the headline result, costs the same as CHOP-oracle: the only compute difference is in adapter training ($\sim$ 22 H100-hours saved by not training $\mathfrak{p}_1$ and $\mathfrak{p}_3$). We therefore do not see a regime in which CHOP is preferable to Top-K LSM purely on compute grounds; the case for CHOP is the routing benefit, not the compute enve-

Table 29: Small-scale pilot results on AIME 2024 avg@16 (mean over 3 seeds, single-round training, 20% training subset). Reliability-only baselines begin to degrade once the horizon exceeds 15K; this motivated the locked long-cap main sweep and reduces the risk of selective reporting.

Method	**7K pilot**	**15K pilot**	**30K pilot**
Standard OPD	43.7	42.1	34.2
Top-K LSM	46.1	46.9	45.4
OPSD-style	43.9	42.6	36.8
CHOP (pilot)	**46.4**	**46.7**	**47.6**
Δ vs. Top-K LSM	$+0.3$	-0.2	$+2.2$

lope.

N Alpha Disambiguation: Relationship Between the Three α Quantities

This appendix elaborates the three-way disambiguation introduced in §3.1 and describes the independent measurement protocol for each quantity.

N.1 Formal Definitions and Estimation Protocols

(i) Per-state alignment α_t (Eq. (2)). Estimated at each student-visited state s_t by sampling $N = 64$ teacher continuations from s_t to length $L_{\text{ref}} = 128$ and computing the cosine alignment between the expected token-OPD gradient and the expected optimal-return gradient. The $N = 64$ sample size yields standard error ≤ 0.03 across all reported configurations, estimated by the $\sqrt{\text{Var}/N}$ sandwich formula on the 64 continuation outcomes. This per-state estimate is expensive (one extra forward pass per state per teacher rollout) and is used only for offline diagnostics and the $R^2 = 0.71$ regression in §3.2.

(ii) Trajectory diagnostic α_{diag} (trajectory average of α_t). The average $\alpha_{\text{diag}} = \mathbb{E}_{t\sim\pi_\theta}[\alpha_t]$ is estimated by evaluating α_t on 12 uniformly sampled reference states per trajectory (not all T states), then averaging. This is the quantity reported as "α decays from 0.71 to 0.09" and "Spearman correlation 0.86" in the main text; we use the "diag" subscript throughout to mark this as a trajectory-level diagnostic.

(iii) Theory parameter α (Theorem 1). The population-level scalar $\alpha = \mathbb{E}_{s_t\sim\pi_\theta}[\alpha_t]$ appearing in the SNR bound is estimated as the mean of α_t over a large held-out state set (50K states). It carries the same sign and ordering as α_{diag} but may differ numerically: α_{diag} is averaged over a small number of reference states per trajectory, while the theory parameter averages uniformly over the entire state distribution. In practice, the two are within 0.03 in all configurations we measure; we treat the theory parameter as the population-level idealisation and α_{diag} as its practical estimator.

N.2 Independent Measurement of SNR Constants σ_R, $\sigma_\star$, c_S

The threshold $\alpha^\star \approx 0.32$ (Corollary 2) is derived from independently measured constants σ_R, $\sigma_\star$, and c_S. We describe the measurement protocol for each, emphasising that all measurements are conducted *before* observing the Top-K LSM plateau location, to avoid circularity.

Measurement of σ_R and $\sigma_\star$. σ_R bounds the per-token variance of the orthogonal reward component $R_t^\perp(v) = R_T(v|s_t) - \alpha_t R^\star(s_t)$; $\sigma_\star$ bounds the per-token variance of the aligned signal $\alpha_t R^\star(s_t)$. Both are estimated from a 10K-state held-out corpus drawn independently from the CHOP training distribution (split before any CHOP training run), by computing the empirical standard deviations of $\{g_t(v)R_T(v|s_t) : v \sim p_t\}$ and $\{g_t(v)R^\star(s_t) : v \sim p_t\}$ respectively. Measurement results with 95% bootstrap CIs:

Measurement of c_S. c_S is estimated from the empirical coefficient of variation of the sliced-Wasserstein segment discrepancies on the same 10K-state held-out corpus, using the median-of-means estimator over 32 random projections. The resulting $\alpha^\star = 0.32$ (CI $[0.29, 0.35]$) was computed from these constants and compared to the observed Top-K LSM plateau location ($\alpha_{\text{diag}} \approx 0.32$ at the 15K–30K boundary, Table 1). The agreement is a *consistency check*, not an independent prediction: both quantities have overlapping CIs, and the measurement protocol is independent of the crossing-point observation.

N.3 Relationship Summary

The three α quantities are consistent by construction but serve different roles:

Table 30: Independent estimates of SNR constants from a held-out corpus of 10K states. Measurements are completed before any CHOP training; $\alpha^\star$ is computed from these constants and compared post hoc to the observed plateau.

Constant	Estimate	95% bootstrap CI
σ_R (orthogonal reward std)	0.41	$[0.38, 0.44]$
$\sigma_\star$ (aligned signal std)	0.74	$[0.70, 0.78]$
c_S (segment accumulation constant)	1.18	$[1.11, 1.25]$
$\alpha^\star$ (derived, $T = 10^4$)	0.32	$[0.29, 0.35]$

- α_t is the operationally correct definition and the direct predictor of routing; it cannot be computed online without expensive rollouts.
- α_{diag} is the observable diagnostic, reported in all figures and tables; it is a post-hoc quantity that cannot guide routing but characterises the failure regime.
- α (theory) is the population parameter; α_{diag} is its sample estimate and the two agree within 0.03 across all configurations we measure.

The routing variable ρ_t is a cheap per-state proxy for α_t that avoids the $N = 64$ teacher rollouts and achieves Spearman 0.84 against the exact α_t; see Appendix C for validation.

O Reset-Rollout Intervention Protocol

This appendix provides the full protocol for the causal intervention cited in §3.2: replacing low-ρ_t student prefixes with teacher continuations raises the OPD gradient–return correlation from $r = 0.09$ to $r = 0.71$.

O.1 State Selection

We sample 800 student-generated states stratified into two groups:

- **Low-ρ_t group** ($n = 400$): states in the bottom 30th percentile of ρ_t (threshold: $\rho_t < 0.28$). These are the states where the reliability gate fires.
- **High-ρ_t control group** ($n = 400$): states in the top 30th percentile ($\rho_t > 0.64$).

States are drawn from a single SWE-bench Verified seed run at training step 2,000 (mid-training), uniformly over instances.

O.2 Intervention Procedure

For each selected state $s_t = (x, \hat{y}_{<t})$:

1. *Teacher continuation*: sample $M_{\text{tc}} = 32$ teacher continuations $\{\tilde{y}^{(j)}\}$ of length $L_{\text{tc}} = 128$ from $\pi_T(\cdot|x)$, then truncate to depth t to form teacher-generated prefixes $\tilde{s}_t^{(j)} = (x, \tilde{y}_{<t}^{(j)})$.
2. *OPD gradient at replacement prefix*: compute the standard sampled-token OPD gradient $\nabla_\theta \log \pi_\theta(\hat{y}_t|\tilde{s}_t^{(j)})$ at the teacher-generated prefix.
3. *KL reduction* measurement: evaluate the KL reduction $D_{\text{KL}}(p_t\|q_t) - D_{\text{KL}}(p_{t+1}\|q_{t+1})$ after one gradient step, using the same OPD objective on the *student* prefix s_t.

O.3 Negative Control

To confirm that the correlation increase is not simply due to providing any "reasonable" prefix, we also sample 400 *random* prefixes of the same depth by concatenating shuffled student-trajectory fragments from different instances. The correlation for the random-prefix condition is $r = 0.11$, indistinguishable from the original low-ρ_t condition ($r = 0.09$). This confirms that the recovery to $r = 0.71$ requires teacher-quality prefixes, not just any alternative prefix.

O.4 Results with Confidence Intervals

The key comparison is between the low-ρ_t group with teacher continuation ($r = 0.71$, CI $[0.64, 0.77]$) and the high-ρ_t control group ($r = 0.68$, CI $[0.61, 0.74]$): the two are statistically indistinguishable, confirming that restoring prefix quality restores gradient informativeness. The random-prefix negative control ($r = 0.11$) rules out the

Table 31: Reset-rollout intervention results. Each correlation is computed over 400 (prefix, gradient-norm, KL-reduction) triples; bootstrap 95% CIs from 5,000 resamples.

Condition	**Pearson r**	**95% bootstrap CI**
Low-ρ_t student prefix (original)	0.09	$[0.02, 0.16]$
Low-ρ_t + random prefix (negative control)	0.11	$[0.03, 0.18]$
Low-ρ_t + teacher continuation prefix	0.71	$[0.64, 0.77]$
High-ρ_t student prefix (positive control)	0.68	$[0.61, 0.74]$

alternative explanation that the intervention works by reducing prefix entropy rather than providing teacher-quality semantic context.