OMNIOPD: LOGIT-FREE ON-POLICY DISTILLATION VIA SPECULATIVE VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-Policy Distillation (OPD) trains a student model on its own generative trajectories under dense token-level feedback from a stronger teacher, mitigating both the off-policy distribution shift of Supervised Fine-Tuning (SFT) and the sparse credit assignment of outcome-based Reinforcement Learning (RL). However, standard OPD faces two coupled limitations. First, it requires direct access to the teacher's token-level logits, excluding a broad class of capable proprietary models. Second, the token-level logit signal itself is brittle, depending on a narrow overlap of plausible next tokens between teacher and student, and prone to amplifying degenerate patterns such as repetition loops. We introduce **OMNIOPD**, a novel framework that addresses both limitations through a logit-free, chunk-level supervision signal. OMNIOPD replaces deterministic logit matching with Monte Carlo rollouts that approximate the teacher's preferences through a continuous semantic similarity metric over multi-token chunks, and concentrates this supervision via a peak-entropy scheduler that audits the student only at its high-uncertainty reasoning forks. Across mathematical reasoning and competitive programming benchmarks, OMNIOPD outperforms SFT by up to +45.31% relative on math and exceeds standard white-box OPD by up to +28.64% on mathematical reasoning, demonstrating that effective on-policy teacher supervision can be recovered from text alone without access to teacher logits.

1 INTRODUCTION

On-Policy Distillation (OPD) has emerged as a central paradigm for post-training open-weight reasoning models (Agarwal et al., 2024; Lu & Lab, 2025; Song & Zheng, 2026). By having the student generate its own reasoning trajectories and querying a stronger teacher to score them, OPD combines the dense token-level supervision of Supervised Fine-Tuning (SFT) with the on-policy exploration of Reinforcement Learning (RL), mitigating both the distribution shift of imitating offline demonstrations (Huan et al., 2025) and the sparse credit assignment of outcome-based RL methods such as GRPO (Shao et al., 2024; Chen et al., 2025).

Despite these strengths, standard OPD inherits two coupled limitations that both stem from using the teacher's token-level log-probabilities as the supervision signal. First, standard OPD approaches (Agarwal et al., 2024; Lu & Lab, 2025) requires direct read-out of the teacher's exact next-token logits. This limitation is increasingly restrictive, as a broad class of capable proprietary models, such as Claude (Anthropic, 2025b), and the wider commercial LLM ecosystem (Aubakirova et al., 2026), return only text, not logits, and are therefore excluded from the OPD pipeline entirely (Song & Zheng, 2026).

Second, even when full logit access is available, token-level matching is a surprisingly brittle target. Recent studies of OPD failure modes (Li et al., 2026; Luo et al., 2026; Fu et al., 2026) show that the effective supervision signal lives in the narrow overlap between the teacher's and student's plausible next-token sets — a small set of plausible next tokens that concentrates the bulk of the gradient mass. Token-level supervision can therefore become brittle when teacher and student assign probability mass to different local continuations, particularly under differences in reasoning style or tokenization (Fu et al., 2026). Degenerate prefixes such as repetition loops can further induce misleading local supervision signals. This motivates our central question: Can dense on-policy

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

distillation be performed without access to the teacher's logits, and can text-only supervision remain effective without requiring exact token-level distribution matching?

We answer both questions with **OMNIOPD**, a logit-free framework that supervises the student through a multi-token, semantic-level signal estimated from text alone. Our approach builds on *Monte Carlo estimation* (Kroese et al., 2014). Rather than requiring deterministic logit distributions, OMNIOPD approximates the teacher's local preferences by sampling multiple discrete rollouts and estimating a teacher-compatibility signal from the semantic similarity between sampled teacher continuations and the student's generation. To make this computationally viable under restricted query budgets, we propose a peak-entropy *chunk scheduler*. Instead of querying the teacher at every token, OMNIOPD selectively audits chunks of student tokens at the high-uncertainty decision boundaries in the student's reasoning trajectory, where additional teacher supervision is likely to be most informative. Finally, to counteract the high variance and zero-probability traps inherent in discrete sampling, we introduce a *Bayesian smoothing step* (Robbins, 1992). By anchoring the empirical Monte Carlo estimates with the student's policy prior, OMNIOPD extracts a variance-reduced gradient signal, while a separate base-model KL anchor on un-audited tokens prevents policy collapse across the trajectory's unsupervised gaps.

Across mathematical reasoning and competitive programming benchmarks, OMNIOPD substantially outperforms SFT and, in several settings, also exceeds white-box OPD (Lu & Lab, 2025) despite operating without access to teacher logits. These results show that chunk-level text-only supervision can provide a competitive on-policy learning signal without requiring exact token-level distribution matching. When paired with stronger black-box teachers such as Claude-4.5-haiku (Anthropic, 2025b) and Gemini-2.5-flash (Comanici et al., 2025), OMNIOPD further improves the student beyond the corresponding GRPO (Shao et al., 2024) baseline, illustrating the complementary value of transferring information from a stronger external teacher.

In summary, our main contributions are as follows:

- We propose OMNIOPD, a logit-free, chunk-level OPD framework that removes the white-box access requirement of standard OPD and replaces exact token-level distribution matching with a text-only chunk-level supervision signal.
- We introduce three components that make OMNIOPD tractable and stable: a peak-entropy scheduler that concentrates teacher verification at the trajectory's most uncertain reasoning forks, a Bayesian-style shrinkage estimator that stabilizes sparse Monte Carlo evidence, and a trust-region KL penalty that prevents policy collapse across un-audited tokens. We provide formal guarantees on estimation error, gradient variance, and trajectory stability.
- We demonstrate empirically that OMNIOPD outperforms SFT baseline by up to +45.31% relative on mathematical reasoning and +18.52% on competitive programming, exceeds white-box OPD by up to +28.64% on mathematical reasoning, and outperforms the corresponding GRPO baseline when paired with stronger black-box teachers, such as Claude-4.5-haiku and Gemini-2.5-flash.

2 METHODOLOGY

2.1 PRELIMINARIES

Offline Supervised Fine-Tuning (SFT). The standard paradigm for transferring reasoning capabilities from a strong teacher model to a smaller student model is SFT. Given a dataset of instructions x and corresponding reasoning trajectories $y = (y_1, y_2, \ldots, y_T)$ generated by a teacher, SFT optimizes the student policy π_θ via maximum likelihood estimation. The objective minimizes the standard cross-entropy loss:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{t=1}^{T} \log \pi_\theta(y_t \mid x, y_{<t})\right] \tag{1}$$

While efficient, offline SFT inherently suffers from exposure bias and distribution shift, as the student is trained on the teacher's static state distribution, not its own (Lu & Lab, 2025).

Online On-Policy Distillation (OPD). OPD addresses this exposure bias by training on the student's own generations rather than the teacher's static trajectories. Let $\hat{y} \sim \pi_\theta(\cdot \mid x)$ be a trajectory sampled from the student policy. Standard OPD minimizes a per-token divergence (most commonly

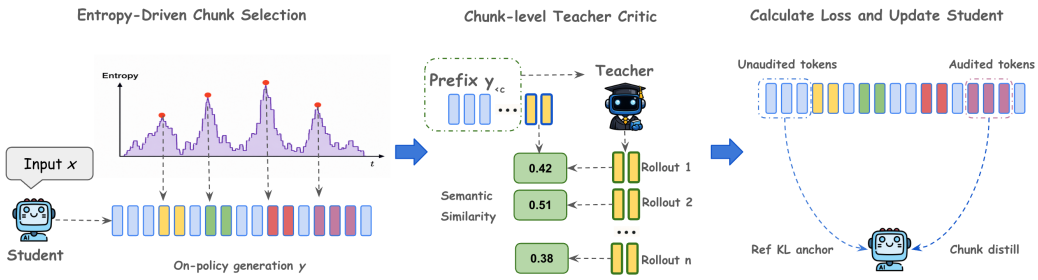


Figure 1: Overview of OMNIOPD. **Entropy-Driven Chunk Selection:** Given an input prompt x, the student model generates an on-policy trajectory. OMNIOPD first identifies critical reasoning forks by locating local peaks in predictive entropy. **Chunk-level Teacher Critic:** For each selected chunk, the trajectory prefix $y_{<c}$ is passed to the teacher. The teacher generates multiple Monte Carlo rollouts, which are then evaluated against the student's generated chunk via a continuous semantic similarity metric. **Calculate Loss and Update Student:** The student policy is updated by applying the distillation loss exclusively to the audited tokens. Simultaneously, a Reference KL anchor is applied across the unaudited tokens to constrain the policy shift.

the Reverse KL divergence) between the student and the true teacher policy $\pi^*_{teacher}$ (Agarwal et al., 2024; Lu & Lab, 2025):

$$\mathcal{L}_{\text{OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\hat{y}\sim\pi_\theta}\left[\sum_{t=1}^{|\hat{y}|} D_{\text{KL}}\Big(\pi_\theta(\cdot \mid x, \hat{y}_{<t}) \,\|\, \pi^*_{\text{teacher}}(\cdot \mid x, \hat{y}_{<t})\Big)\right] \tag{2}$$

By exploring its own action space and receiving corrective feedback from the teacher distribution $\pi^*_{teacher}$, the student learns a more robust policy that bridges the gap between training and inference distributions.

2.2 OMNIOPD

OMNIOPD addresses both limitations of standard OPD identified in section 1 through replacing the token-level logit signal with a chunk-level semantic signal estimated from text alone. As illustrated in figure 1, OMNIOPD has three components: a peak-entropy scheduler that selects high-uncertainty chunks from the student's trajectory (section 2.2.3), a chunk-level teacher critic that estimates the teacher's preference through Monte Carlo rollouts evaluated by semantic similarity (section 2.2.1 and 2.2.2), and a trust-region anchor that prevents policy drift on un-audited tokens (section 2.2.4). The full algorithm is presented in Algorithm 1 in Appendix D.

2.2.1 CHUNK-LEVEL DISTILLATION VIA SPECULATIVE VERIFICATION

Our logit-free distillation is built on Monte Carlo estimation of the teacher's preference. A naive token-level instantiation — querying the teacher at every step t and estimating the empirical frequency of the student's exact next token — is infeasible because the per-token query volume scales as $O(T)$ over reasoning trajectories of length T. Tokenization differences between distinct model families may also cause the student's exact next token to not appear in any teacher rollout, producing sparse, zero-gradient updates that destabilize optimization. To resolve these issues, OMNIOPD instead groups consecutive tokens into chunks of C tokens and applies a single teacher query per chunk, reducing the query volume by a factor of $O(T/C)$ and shifting the matching signal from exact tokens to semantic similarity. This chunk-level paradigm is structurally inspired by the verification phase of Speculative Decoding (Leviathan et al., 2023; Chen et al., 2023), where a large target model parallel-verifies a sequence drafted by a smaller model. In our framework, we treat each C-token block as a single, cohesive verification unit.

Let the student model generate an on-policy reasoning trajectory $y \sim \pi_\theta(\cdot \mid x)$. From this trajectory, we select M non-overlapping chunks of length C (the selection mechanism is later detailed in section 2.2.3). For a student chunk c, the black-box teacher receives the historical prefix $y_{<c}$ and generates N independent Monte Carlo rollouts $\{y^{(i)}_{\text{teacher}}\}_{i=1}^{N}$. Because these rollouts are conditioned

on the student's own prefix, severely erroneous histories may also bias the teacher continuation. We empirically quantify this prefix-conditioning effect in Appendix E.5, finding that direct inheritance of the student's error is concentrated primarily in late-stage prefixes of already failed trajectories. Notably, as frontier teacher models and local student models rarely share the same tokenizer, even at chunk level, computing exact probability matches between teacher rollouts and the student's chunk remains infeasible. To resolve this tokenizer mismatch, OMNIOPD evaluates the student's chunk against the teacher rollouts using a continuous semantic similarity function $\phi : (y_c, y_{\text{teacher}}) \mapsto [0, 1]$ (e.g., ROUGE-1 unigram overlap (Lin, 2004)). Here, ϕ is used as a practical proxy for teacher compatibility rather than an independent verifier of mathematical or functional correctness. The aggregate semantic similarity $k_{\text{sem}}^{(c)}$ for chunk c is then computed as:

$$k_{\text{sem}}^{(c)} = \sum_{i=1}^{N} \phi(y_c, y_{\text{teacher}}^{(i)}) \tag{3}$$

Importantly, we do not interpret ϕ as an independent verifier of step-level correctness. Rather, it serves as a dense surrogate for the teacher's local preference: student chunks that are more compatible with continuations sampled from the teacher receive stronger supervision. This role parallels the token-level probabilities used in standard OPD, which provide a teacher-derived learning signal without themselves constituting a ground-truth correctness verifier. We empirically evaluate how well this semantic surrogate preserves the teacher's underlying preference signal in Appendix E.1. By shifting the measurement from tokens to semantics over a continuous C-token window, we provide a teacher-compatibility signal over conceptual milestones across distinct model architectures without requiring exact token matches or penalizing benign stylistic deviations.

2.2.2 BAYESIAN SMOOTHING FOR SPARSE SIGNALS

While the chunk-level formulation reduces query overhead, estimating the teacher-compatibility signal from a limited sampling budget N can introduce substantial sampling variance. A naive frequentist estimator $\hat{\pi}_{\text{freq}}^{(c)} = k_{\text{sem}}^{(c)}/N$ can become exactly zero when the sampled teacher continuations exhibit no compatibility with the student chunk. Because this estimate later acts as a multiplicative weight in our optimization objective, a zero estimate removes the teacher-derived learning signal for that audited chunk.

To stabilize this sparse estimate, we introduce a Bayesian-style shrinkage estimator that combines the Monte Carlo evidence with a student-policy prior. Taking the student's normalized per-chunk probability as the prior,

$$\bar{\pi}_\theta^{(c)} = \left(\prod_{t \in c} \pi_\theta(y_t \mid x, y_{<t}) \right)^{\frac{1}{C}}, \tag{4}$$

the smoothed teacher-compatibility proxy, parameterized by prior strength α, is

$$\hat{\pi}_{\text{teacher}}^{(c)} = \frac{k_{\text{sem}}^{(c)} + \alpha \cdot \bar{\pi}_\theta^{(c)}}{N + \alpha}. \tag{5}$$

Equivalently, this estimator is a convex combination of the empirical Monte Carlo estimate and the student prior, with α controlling the corresponding bias–variance trade-off. Theorem B.1 shows that the resulting compatibility multiplier remains bounded and strictly positive even when $k_{\text{sem}}^{(c)} = 0$, while its variance is reduced relative to the unsmoothed frequentist estimator. Consequently, zero sampled compatibility cannot by itself eliminate the chunk-level update, while the smoothing reduces estimator-induced gradient variance under limited rollout budgets.

2.2.3 ENTROPY-DRIVEN CHUNK SELECTION

Under a budget of M chunks per trajectory, the method by which we select these windows determines distillation efficiency. Reasoning trajectories oftentimes contain long segments of low-complexity transitions (e.g., formatting and deterministic arithmetic), where teacher verification provides negligible gradient signal. To maximize the marginal utility of each critic, we aim to target teacher verification at the high-uncertainty reasoning forks where the student's policy diverges.

As the student model π_θ is fully accessible, we use its internal predictive entropy as a proxy for this local uncertainty. At each generation step t, we compute the entropy of the student's output vocabulary distribution:

$$\mathcal{H}_t = -\sum_{v \in \mathcal{V}} \pi_\theta(v \mid x, y_{<t}) \log \pi_\theta(v \mid x, y_{<t}) \tag{6}$$

where $\mathcal{V}$ is the vocabulary space. A low entropy H_t indicates a relatively deterministic transition, whereas high entropy indicates greater local predictive uncertainty.

To construct the audited chunks, we rank all token positions $t \in \{1, \ldots, T\}$ by their predictive entropy $\mathcal{H}_t$ in descending order. Starting from the position with the largest $\mathcal{H}_t$, we form a contiguous C-token candidate chunk around each anchor. If the candidate chunk overlaps with any previously selected chunk, we skip that anchor and proceed to the next-highest-entropy position. Otherwise, the candidate is added to the audited chunk set. We repeat this procedure until M non-overlapping chunks are selected. This greedy selection prioritizes regions with high predictive uncertainty while preserving a fixed audit budget and avoiding redundant supervision over overlapping trajectory segments.

2.2.4 Trust Region Anchoring

While our peak-entropy selection places the M audited chunks at the most informative positions, it leaves the remaining $|\mathcal{U}| = T - M \cdot C$ tokens un-audited. Optimizing a policy on these sparse, localized signals risks policy degeneration, as the student model π_θ might exploit structural shortcuts, generate incoherent text between audited chunks, or collapse its entropy to artificially shorten the trajectory length.

For all un-audited tokens $t \in \mathcal{U}$, we apply a KL divergence penalty against the reference policy π_{ref}, which is the frozen copy of the student's initial weights. By Pinsker's inequality (Pinsker, 1964), minimizing $D_{KL}(\pi_{\text{ref}} \,\|\, \pi_\theta)$ bounds the Total Variation shift of the student's policy, acting as a trust region that constrains drift in the unsupervised regions (formalized in Theorem B.3). Consequently, policy updates concentrate on the audited chunks while the trust region holds the un-audited tokens close to π_{ref}.

The OmniOPD optimization objective. Unlike standard reverse-KL OPD, the black-box teacher in our setting provides only sampled continuations rather than a normalized output distribution. We therefore use the Bayesian semantic score as a bounded compatibility weight on the likelihood of the realized student chunk, yielding a logit-free on-policy surrogate objective:

$$\mathcal{L}_{\text{OmniOPD}}(\theta) = \underbrace{-\mathbb{E}_{\hat{y}\sim\pi_\theta}\left[\sum_{c=1}^{M} \text{sg}\left[\hat{\pi}_{\text{teacher}}^{(c)}\right] \sum_{t \in c} \log \pi_\theta(y_t \mid x, y_{<t})\right]}_{\mathcal{L}_{\text{chunk}}(\theta)} \underbrace{+\beta \sum_{t \in \mathcal{U}} D_{\text{KL}}\big(\pi_{\text{ref}}(\cdot \mid x, y_{<t}) \,\|\, \pi_\theta(\cdot \mid x, y_{<t})\big)}_{\text{trust-region anchor}}. \tag{7}$$

where $\text{sg}[\cdot]$ denotes stop-gradient, such that the Bayesian semantic score is treated as a fixed compatibility weight during each student update, and β controls the strength of the trust-region anchor. By minimizing $\mathcal{L}_{\text{OmniOPD}}$, the student model systematically aligns its critical, high-entropy reasoning junctions with the frontier black-box teacher, while the trust-region penalty ensures the global stability and coherence of the resulting long-horizon trajectory. This objective does not assign explicit negative labels to low-compatibility chunks; instead, teacher disagreement suppresses their update by assigning a smaller compatibility weight. A comprehensive theoretical analysis is provided in Appendix B, where we show that this logit-free surrogate yields bounded gradient signals, characterize rollout sufficiency under limited sampling budgets, and bound the distributional drift of un-audited tokens.

3 Experiments

3.1 Experiment Setup

Models. We employ Qwen3-4B and Qwen3-1.7B (Yang et al., 2025) as the student models for policy optimization. To demonstrate the generalizability of our framework across both open-weight and closed-source models, we utilize a diverse suite of frontier models as oracles. Specifically, our

teacher ensemble comprises the open-weights Qwen3-32B and Qwen3-30B-A3B-Instruct (Yang et al., 2025), alongside the proprietary, API-gated Gemini-2.5-Flash (Comanici et al., 2025) and Claude-4.5-Haiku models (Anthropic, 2025a).

Datasets and Evaluation. We utilize the DAPO-Math-17K dataset (Yu et al., 2025) to train mathematical reasoning capabilities, benchmarking performance on AIME-2024 (Li et al., 2024), AIME-2025, AMC23 (Li et al., 2024), MATH-500 (Hendrycks et al., 2021), and OlympiadBench (He et al., 2024). Following the established setup of prior work (Liang et al., 2025), our code generation models are trained on the PRIME-RL compilation (Cui et al., 2025), including APPS, CodeContests, TACO, and Codeforces, and evaluated against 100 sampled validation instances per dataset category. To reduce evaluation variance, model performance across all domains is quantified using the Pass@1 metric, with reported results averaged over 16 independent sampling trials.

Baselines. We compare OMNIOPD against baselines spanning offline teacher distillation, white-box on-policy distillation, black-box distillation, self-distillation, and reinforcement learning. For offline teacher distillation, we use sequence-level knowledge distillation (Seq-KD), implemented as SFT on complete long-CoT trajectories generated by the teacher. When teacher logits are accessible, we additionally compare against standard white-box OPD. For black-box teacher distillation, we include GAD (Ye et al., 2025), which learns from teacher-generated feedback through an auxiliary discriminator. We also consider two forms of self-distillation that do not require an external teacher: Best-of-N SFT, which samples multiple trajectories from the student and fine-tunes on correct trajectories, and STaR-style self-training (Zelikman et al., 2022), which iteratively trains on the student's own successful reasoning trajectories. Finally, we compare against verifier- or reward-based reinforcement learning methods, including GRPO and DAPO (Shao et al., 2024; Yu et al., 2025), as complementary baselines for policy improvement through outcome-based feedback.

Implementation Details. For our OMNIOPD framework, the default hyperparameters are configured to $M = 10$ intermediate process chunks, a chunk length of $C = 50$ tokens, and $N = 10$ teacher MC rollouts per fork. The Reference KL anchoring penalty is weighted at $\beta = 0.1$, and we employ Edit Distance (Navarro, 2001) as the default semantic similarity metric (ϕ). We systematically explore the impact of the chunk count, chunk length, rollout budget, KL weight, and choice of semantic measurer in section 4.2. Other details are shown in Appendix D.1

Empirical Validation of the Semantic Teacher Proxy. Before evaluating downstream reasoning performance, we directly examine whether the semantic rollout signal used by OMNIOPD preserves the local preference of the teacher. Using the open-weight Qwen3-32B teacher, for which the underlying model probabilities are accessible, we compare our black-box semantic estimators against the corresponding teacher probability signal over student-generated chunks. This analysis evaluates the intended role of the semantic rollout signal induced by ϕ as a surrogate for teacher preference, rather than as an independent measure of step-level correctness. As detailed in Appendix E.1, Bayesian smoothing consistently reduces estimation error and improves Pearson correlation with the teacher signal across semantic metrics and rollout budgets. We further observe a trade-off between absolute estimation accuracy and preference ranking: ROUGE-1 provides the lowest estimation error, whereas Edit Distance better preserves the relative ranking of student reasoning paths. We systematically compare the downstream impact of this trade-off in our sensitivity analysis (section 4.2). We further analyze this proxy under varying levels of teacher–student token overlap in Appendix E.6, finding that the semantic signal remains informative even in low-overlap regions, although accessible teacher likelihood remains a stronger reference signal.

3.2 MAIN RESULTS

Significant Improvements over Black-Box Distillation Baselines. We present the main empirical results for mathematical reasoning and competitive programming in Table 1 and Table 6 in Appendix E.2. Across both student scales, OMNIOPD substantially improves over the corresponding Seq-KD baseline. With the Qwen3-32B teacher, OMNIOPD improves the Qwen3-1.7B student from 53.76% to 60.67% average accuracy and the Qwen3-4B student from 63.80% to 69.08%. Compared with the black-box GAD baseline, OMNIOPD improves average accuracy from 52.69% to 60.67% for Qwen3-1.7B/Qwen3-32B, from 67.48% to 69.08% for Qwen3-4B/Qwen3-32B, and from 59.12%

Table 1: Main results on mathematical reasoning benchmarks. Best results for each setup are in **bold**.

Student Model	Teacher Model	Method	MATH-500	AIME-24	AIME-25	AMC-23	OlympiadBench	Average
Reference Models (Zero-Shot / Base Inference)								
-	Qwen3-32B	Base	73.60	75.42	68.33	87.81	52.93	71.62
-	Qwen3-30B-A3B-Inst.	Base	69.81	61.04	44.79	78.59	56.38	62.12
-	Claude-4.5-haiku	Base	89.95	55.20	34.79	87.50	62.69	66.03
-	Gemini-2.5-flash	Base	86.83	74.79	65.20	93.13	61.86	76.36
Student: Qwen3-1.7B								
Qwen3-1.7B	-	Base	60.26	48.54	37.50	75.16	37.90	51.87
Qwen3-1.7B	-	STaR	85.15	44.58	33.13	80.00	57.31	60.03
Qwen3-1.7B	-	DAPO	75.44	44.79	38.33	81.72	51.85	58.43
Qwen3-1.7B	-	GRPO	80.35	47.50	44.45	84.06	55.67	62.41
Qwen3-1.7B	-	Best-of-16 SFT	80.80	45.63	35.00	85.16	57.53	60.82
Qwen3-1.7B	Qwen3-32B	Seq-KD	79.91	35.42	30.20	71.09	52.19	53.76
		OPD	**84.82**	**48.95**	**37.58**	79.06	58.08	**61.70**
		GAD	83.44	33.13	26.46	70.00	50.41	52.69
		OMNIOPD	84.61	47.08	33.33	**79.68**	**58.64**	60.67
Student: Qwen3-4B								
Qwen3-4B	-	Base	61.55	50.83	41.88	75.47	40.32	54.01
Qwen3-4B	-	STaR	62.70	60.80	51.90	88.10	59.00	64.50
Qwen3-4B	-	DAPO	72.20	66.50	60.60	94.40	51.00	68.94
Qwen3-4B	-	GRPO	73.75	68.62	60.70	95.93	52.19	70.24
Qwen3-4B	-	Best-of-16 SFT	65.50	72.50	60.20	95.80	54.90	69.78
Qwen3-4B	Qwen3-32B	Seq-KD	68.46	60.83	58.95	83.59	47.19	63.80
		OPD	67.50	61.45	60.25	82.65	48.96	64.16
		GAD	70.00	59.80	53.80	90.00	**63.80**	67.48
		OMNIOPD	**72.60**	**68.67**	**60.42**	**93.44**	50.27	**69.08**
Qwen3-4B	Qwen3-30B-A3B-Inst.	Seq-KD	66.77	34.58	31.04	70.16	46.29	49.77
		OPD	69.78	48.54	36.87	75.78	50.14	56.22
		GAD	**78.49**	45.63	39.38	75.94	**56.16**	59.12
		OMNIOPD	74.98	**75.00**	**62.92**	**92.65**	56.05	**72.32**
Qwen3-4B	Claude-4.5-haiku	Seq-KD	80.56	60.83	53.33	85.16	57.71	67.52
		OMNIOPD	**81.28**	**72.08**	**65.63**	**91.25**	**64.35**	**74.92**
Qwen3-4B	Gemini-2.5-flash	Seq-KD	80.10	71.25	63.13	90.44	62.61	73.51
		OMNIOPD	**83.30**	**73.13**	**64.37**	**92.96**	**64.58**	**75.67**

to 72.32% for Qwen3-4B/Qwen3-30B-A3B-Instruct. The gains become larger with frontier black-box teachers: distillation from Claude-4.5-haiku reaches 74.92%, compared with 67.52% for its corresponding Seq-KD baseline. These results suggest that supervising the student's own generated reasoning trajectories with localized teacher feedback can provide a more effective learning signal than relying solely on offline teacher demonstrations.

Comparison with White-Box OPD. On mathematical reasoning, OMNIOPD remains competitive with standard white-box OPD despite operating without teacher logits, and exceeds it in the Qwen3-4B settings. For the Qwen3-1.7B/Qwen3-32B pair, OMNIOPD achieves 60.67% average accuracy compared with 61.70% for OPD. As student capacity increases to Qwen3-4B, OMNIOPD reaches 69.08% with Qwen3-32B and 72.32% with Qwen3-30B-A3B-Instruct, compared with 64.16% and 56.22% for white-box OPD, respectively. We hypothesize that these gains arise primarily from replacing token-level reverse-KL matching with a bounded chunk-level compatibility signal over the student's realized continuation. This formulation avoids extreme gradients under low teacher support (Theorem B.1) and can reduce sensitivity to tokenization and surface-form mismatches (Theorem B.4). Our ablations suggest that Bayesian smoothing and entropy scheduling provide smaller, complementary gains on top of this chunk-level objective.

Comparisons to Teacher-Free Self-Training and Reward-Based RL. For the Qwen3-1.7B student, OMNIOPD achieves 60.67% average accuracy, outperforming STaR (60.03%) and DAPO (58.43%) and remaining comparable to Best-of-16 SFT (60.82%), while trailing GRPO (62.41%). A similar pattern holds at the 4B scale with the Qwen3-32B teacher, where OMNIOPD reaches 69.08%, exceeding STaR (64.50%) and DAPO (68.94%), but remaining slightly below Best-of-16 SFT (69.78%) and GRPO (70.24%). Importantly, unlike these teacher-free methods, OMNIOPD can leverage additional capability from stronger external teachers. With Claude-4.5-haiku and Gemini-2.5-flash, the Qwen3-4B student reaches 74.92% and 75.67%, respectively, substantially exceeding the 70.24% GRPO baseline. We note that teacher standalone accuracy is not perfectly predictive of distillation effectiveness, since on-policy supervision also depends on teacher behavior under student-induced prefixes; we discuss this further in Appendix E.4.

Table 2: Ablation study of OMNIOPD components on mathematical reasoning.

Variant	MATH	AIME-24	AIME-25	AMC-23	Olympiad	Avg
OMNIOPD (Full)	72.60	68.67	**60.42**	**93.44**	**50.27**	**69.08**
w/o Entropy Selection	70.35	70.42	58.54	92.69	50.26	68.45
w/o Bayesian Smoothing	72.95	70.00	58.29	91.87	50.04	68.63
Uniform + Frequentist	**72.90**	**70.70**	56.10	92.00	49.70	68.28
w/o KL Anchoring	9.27	8.13	6.25	12.03	5.71	8.28

Training Dynamics. We further examine the optimization behavior of OMNIOPD under both open-weight and frontier black-box teachers. As detailed in Appendix E.3, the chunk-level training loss decreases smoothly and the reference KL remains bounded throughout training, while downstream AIME-2025 accuracy improves over the same period. These dynamics are consistent with the bounded-gradient and trust-region behavior characterized in Theorem B.1 and Theorem B.3.

4 ABLATION STUDIES AND ANALYSIS

4.1 ABLATION STUDIES

To disentangle the contribution of the main components in OMNIOPD, we conduct ablations using a Qwen3-4B student and Qwen3-32B teacher. The results are summarized in Table 2.

The most substantial degradation occurs when removing the base-model KL anchor ($\beta = 0$), which reduces average accuracy from 69.08% to 8.28%. Because teacher supervision is applied only to a sparse set of audited chunks, unconstrained updates outside these regions can rapidly destabilize the policy. This result is consistent with the role of the trust-region term formalized in Theorem B.3: the KL anchor is essential for preserving the behavior of the base policy across un-audited portions of the trajectory. A detailed analysis in Appendix E.7 shows that this failure manifests as a length- and entropy-collapse feedback loop, rather than merely generic policy drift.

In contrast, entropy-based scheduling and Bayesian smoothing provide smaller but consistent improvements. Replacing peak-entropy chunk selection with uniform sampling reduces the average accuracy from 69.08% to 68.45%, while replacing the Bayesian estimator with its frequentist counterpart yields 68.63%. Removing both mechanisms simultaneously, using uniform chunk sampling together with frequentist teacher estimation under the same teacher-query budget, further reduces performance to 68.28%. These results indicate that neither entropy selection nor Bayesian smoothing is the sole source of OMNIOPD's improvement. Rather, they act as complementary refinements to the underlying logit-free teacher critic: entropy selection concentrates a fixed verification budget on more uncertain regions of the student trajectory, while Bayesian smoothing stabilizes the teacher-compatibility estimate under limited Monte Carlo sampling.

Overall, the ablations suggest a clear division of roles. The chunk-level Monte Carlo teacher signal and teacher-weighted on-policy objective form the core supervision mechanism, the KL anchor is critical for optimization stability, and entropy scheduling and Bayesian smoothing provide additional efficiency and robustness on top of this backbone. We provide qualitative examples of the high-entropy regions selected by the scheduler in Appendix E.8.

4.2 SENSITIVITY ANALYSIS

To understand the robustness of OMNIOPD, we analyze its sensitivity to the verification hyperparameters (M, N, C), the Bayesian prior strength α, and the semantic measurement function ϕ using the Qwen3-4B student and Qwen3-32B teacher. Detailed results for (M, N, C) are provided in Appendix E.9, while sensitivity to α and ϕ is reported in the main text in figure 2 and Table 3, respectively. Below, we summarize the key trends across these design choices.

Sensitivity to Verification Hyperparameters. The three verification hyperparameters exhibit distinct effects. The method is most sensitive to chunk size: reducing C from 50 to 25 decreases average accuracy from 69.08% to 24.48%, whereas increasing it to 100 improves accuracy to 71.58%,

Table 3: Sensitivity analysis of the semantic measurement function (ϕ). We compare normalized Edit Distance against ROUGE-1 unigram overlap across different student-teacher pairings.

Student Model	Teacher Model	Metric (ϕ)	MATH	AIME-24	AIME-25	AMC-23	Olympiad	Avg
Qwen3-1.7B	Qwen3-32B	Edit Distance	84.61	47.08	33.33	79.68	**58.64**	60.67
		ROUGE-1	**86.30**	**48.13**	**36.46**	**85.31**	58.16	**62.87**
Qwen3-4B	Qwen3-32B	Edit Distance	**72.60**	68.67	60.42	**93.44**	50.27	69.08
		ROUGE-1	70.36	**72.71**	**61.04**	92.18	**50.52**	**69.36**
Qwen3-4B	Qwen3-30B	Edit Distance	**74.98**	**75.00**	**62.92**	92.65	**56.05**	**72.32**
		ROUGE-1	74.78	71.67	61.88	**93.75**	55.72	71.56

indicating that sufficiently long chunks are important for meaningful local supervision. In contrast, increasing N from 10 to 20 yields only 0.49 points at $C = 50$ and essentially no gain at $C = 100$. Performance is also relatively stable to M: reducing M from 10 to 5 retains 68.38%, while increasing it to 20 reaches 70.96%. Full sensitivity results are reported in Appendix E.9, and the corresponding teacher-inference cost analysis induced by (M, N, C) are analyzed in Appendix E.10.

Impact of the Bayesian Prior (α). The parameter α controls the strength of the student-policy prior during our Bayesian smoothing step. In figure 2, OMNIOPD exhibits a distinct performance peak at $\alpha = 1.0$. Deviating from this optimal balance in either direction results in performance degradation, illustrating the bias-variance tradeoff inherent in discrete sampling. When the prior is too weak ($\alpha = 0.1$), the average accuracy drops to 67.52%, the framework overfits to the severe statistical variance and zero-probability traps of the $N = 10$ discrete Monte Carlo rollouts. Conversely, an overly large α causes the compatibility estimate to be dominated by the student's own policy prior, reducing the influence of the teacher-rollout evidence and weakening the effective distillation signal.

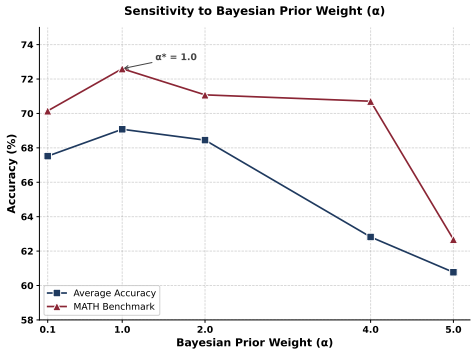


Figure 2: Sensitivity of OMNIOPD to the Bayesian prior (α). $\alpha = 1.0$ achieves the optimal balance between smoothing Monte Carlo variance and policy updates.

Robustness to Semantic Measurement Functions. While OMNIOPD is highly robust to the choice of semantic metric ϕ (Table 3), the optimal selection depends heavily on the student-teacher capacity gap. For heavily constrained students (1.7B), ROUGE-1 significantly outperforms Edit Distance (+2.20% average). Because smaller models struggle to perfectly mimic verbose teacher structures, Edit Distance is overly punitive; ROUGE-1 instead provides crucial linguistic flexibility. As student capacity scales to 4B, this gap narrows (+0.28% average). Conversely, when distilling a highly aligned instruction-tuned teacher into a 4B student, Edit Distance yields a slight edge (+0.76% average). This suggests ROUGE-1 is ideal for bridging extreme capacity mismatches, while Edit Distance excels at enforcing the strict structural mimicking required for aligned reasoning.

5 CONCLUSION

In this work, we introduced OMNIOPD, a logit-free, chunk-level OPD framework that addresses both the white-box access requirement of standard OPD and the brittleness of its token-level supervision signal. By combining peak-entropy chunk scheduling, Bayesian-smoothed speculative verification, and base-model KL anchoring, OMNIOPD provides a stable text-only surrogate for on-policy teacher supervision while controlling policy drift. Across mathematical reasoning and competitive programming benchmarks, OMNIOPD consistently improves over Seq-KD and is competitive with or exceeds white-box OPD in some settings, while also extending naturally to proprietary teachers without logit access. Ultimately, OMNIOPD helps bridge the closed-source capability divide by enabling smaller open-weight models to transfer reasoning capabilities from frontier black-box teachers and, in favorable settings, outperform strong baselines spanning offline distillation, self-training, and reward-based policy optimization.

AI USE STATEMENT

Generative AI tools were used for language polishing, proofreading, and assistance with coding during the preparation of this work. All AI-assisted text and code were reviewed, verified, and, where necessary, revised by the authors. We take responsibility for the final content of this work, including all text, claims, analyses, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies training methods for knowledge distillation using public benchmark environments and does not introduce new datasets containing personal or sensitive information. OMNIOPD relies on teacher judgments, so biases or unsafe preferences in the teacher model may be propagated to the student. Practical deployment should therefore include appropriate action validation, safety constraints, logging, and human oversight.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

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

Anthropic. Claude-4, 2025a. URL `https://www.anthropic.com/news/claude-4`. Accessed: 2025-07-10.

AI Anthropic. System card: Claude opus 4 & claude sonnet 4. *Claude-4 Model Card*, 2025b.

Malika Aubakirova, Alex Atallah, Chris Clark, Justin Summerville, and Anjney Midha. State of ai: An empirical 100 trillion token study with openrouter. *arXiv preprint arXiv:2601.10088*, 2026.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

Peter Chen, Xiaopeng Li, Ziniu Li, Wotao Yin, Xi Chen, and Tianyi Lin. Exploration vs exploitation: Rethinking rlvr through clipping, entropy, and spurious reward. *arXiv preprint arXiv:2512.16912*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The twelfth international conference on learning representations*, 2024.

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

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026.

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

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*, 2022.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. *arXiv preprint arXiv:2507.00432*, 2025.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 4, 2025.

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

Dirk P Kroese, Tim Brereton, Thomas Taimre, and Zdravko I Botev. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):386–392, 2014.

Hojae Lee, Junho Kim, and SangKeun Lee. Mentor-kd: Making small language models better multi-step reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17643–17658, 2024.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9, 2024.

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

Xiao Liang, Zhongzhi Li, Yeyun Gong, Yelong Shen, Ying Nian Wu, Zhijiang Guo, and Weizhu Chen. Beyond pass@ 1: Self-play with variational problem synthesis sustains rlvr. *arXiv preprint arXiv:2508.14029*, 2025.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

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

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying opd: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*, 2026.

Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pp. 380–384, 2013.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Carlos Miguel Patiño, Kashif Rasul, Quentin Gallouédec, Ben Burtenshaw, Sergio Paniego, Vaibhav Srivastav, Thibaud Frere, Ed Beeching, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Unlocking on-policy distillation for any model family, 2025.

Mark S Pinsker. Information and information stability of random variables and processes. *Holden-Day*, 1964.

Tiberiu Popoviciu. Sur certaines inégalités qui caractérisent les fonctions convexes. *Analele Stiintifice Univ."Al. I. Cuza", Iasi, Sectia Mat*, 11:155–164, 1965.

Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pp. 388–394. Springer, 1992.

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

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7059–7073, 2023.

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

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

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

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

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

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

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

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

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

Yuhang Zhou and Wei Ai. Teaching-assistant-in-the-loop: Improving knowledge distillation from imperfect teacher models in low-budget scenarios. *arXiv preprint arXiv:2406.05322*, 2024.

APPENDIX

A RELATED WORK

Offline policy distillation of LLMs. The traditional paradigm for transferring complex reasoning capabilities from frontier models to smaller, more deployable models is offline knowledge distillation (Hsieh et al., 2023; Shridhar et al., 2023; Kim & Rush, 2016). Early approaches primarily relied on sequence-level distillation from black-box APIs, querying more capable models to generate high-quality Chain-of-Thought (CoT) (Wei et al., 2022) trajectories and training the student model to mimic these sequences via SFT (Ho et al., 2022; Zhou & Ai, 2024; Lee et al., 2024). While effective for basic instruction following, this sequence-level imitation often fails to capture the nuanced, token-by-token reasoning dynamics of the teacher. With the emergence of highly capable open-source models (Grattafiori et al., 2024; Abdin et al., 2024; Yang et al., 2025; Guo et al., 2025; Kamath et al., 2025; Team et al., 2025; Xiao et al., 2026), the field increasingly shifted toward more fine-grained, white-box distillation techniques that directly optimize the student against the teacher's continuous output distributions via Forward Kullback-Leibler (KL) divergence or token-level logit matching (Gu et al., 2024; Wen et al., 2023). However, whether utilizing black-box textual trajectories or white-box logits, these approaches are fundamentally off-policy: the student is trained on the teacher's static state distribution, leading to inevitable distribution shift and exposure bias during deployment.

On-policy distillation and its limitations. To mitigate the exposure bias inherent to offline methods, OPD shifts the learning objective to the student's own generated distribution (Agarwal et al., 2024; Lu & Lab, 2025). In standard OPD frameworks, the student model actively generates reasoning trajectories that are then evaluated and scored by the teacher model, bridging the gap between training and inference distributions. Recent variants have expanded this paradigm to self-OPD, where a model iteratively improves through its own on-policy feedback (Zhao et al., 2026; He et al., 2026; Hübotter et al., 2026), as well as generalized OPD frameworks designed to handle architectural differences such as mismatched tokenizers between the student and teacher model (Patiño et al., 2025). Despite these advancements, recent analyses have revealed that the token-level logit signal underlying standard OPD is surprisingly brittle. Li et al. (2026) show that the effective supervision signal concentrates in the narrow overlap between teacher and student plausible next-token sets, and that a teacher with strong benchmark accuracy can fail to improve a student whose reasoning trajectories fall outside this overlap. Luo et al. (2026) document a systematic collapse mode in which degenerate prefixes such as repetition loops accidentally receive large positive gradient signal, since locally predictable tokens

are assigned high probability by the teacher. Fu et al. (2026) further show that single-sampled-token feedback is statistically noisy, and that tokenizer mismatches between teacher and student introduce a class of spurious disagreements that no amount of logit access can resolve. These findings establish token-level logit matching as a structurally fragile supervision signal, motivating the chunk-level semantic approach we develop in this paper. Beyond brittleness, standard OPD also faces a structural access barrier that computing the Reverse KL divergence between student and teacher requires explicit access to the teacher's token probabilities (Yang et al., 2026). Consequently, OPD with proprietary teachers has remained largely unexplored. Ye et al. (2025) recently introduced Generative Adversarial Distillation (GAD), which trains an auxiliary discriminator as an evolving reward model in a minimax game with the student. While GAD bypasses the logit requirement, its discriminator yields only a sequence-level scalar score, providing coarser feedback than the per-step supervision of standard OPD, and requiring an additional model to train and maintain. OMNIOPD instead recovers dense, chunk-level supervision directly from teacher rollouts through Monte Carlo semantic estimation, without any auxiliary network, and can match or exceed white-box logit matching in several evaluated settings

B THEORETICAL ANALYSIS OF OMNIOPD

In this section, we provide a theoretical analysis of OMNIOPD's three core design components — chunk-level loss formulation (Eq. (7)), Bayesian-smoothed estimation (Eq. (5)), and trust-region anchoring on un-audited tokens — by tying each one to a specific failure mode it prevents.

Setup. We assume the semantic similarity function ϕ (as in Eq. (3)) takes values in $[0, 1]$, teacher rollouts $\{y^{(i)}_{\text{teacher}}\}_{i=1}^{N}$ are i.i.d. samples from $\pi^{*}_{\text{teacher}}(\cdot \mid x, y_{<c})$, the student prior $\bar{\pi}^{(c)}_{\theta} > 0$ (as in Eq. (4)) as a geometric mean of strictly positive per-token probabilities, and the score function $\nabla_\theta \log \pi_\theta(y_t \mid \cdot)$ is bounded, which is a standard assumption for softmax-output transformer policies. For each chunk c, we define the true expected similarity as

$$\mu_c := \mathbb{E}_{y' \sim \pi^{*}_{\text{teacher}}(\cdot|x, y_{<c})}[\phi(y_c, y')] \in [0, 1]. \tag{8}$$

The empirical frequentist estimator $\hat{\pi}^{(c)}_{\text{freq}} := k^{(c)}_{\text{sem}}/N$ is the unbiased sample-mean estimator of μ_c with no prior smoothing. The Bayesian proxy in Eq. (5) admits the convex form

$$\hat{\pi}^{(c)}_{\text{teacher}} = \frac{N}{N+\alpha}\,\hat{\pi}^{(c)}_{\text{freq}} + \frac{\alpha}{N+\alpha}\,\bar{\pi}^{(c)}_{\theta}. \tag{9}$$

All expectations are conditional on the student trajectory $\hat{y}$; only teacher rollouts are random. For the optimization analysis below, the sampled trajectory, audited chunk set, and Bayesian compatibility weights are treated as fixed within each student update, consistent with the stop-gradient operation in Eq. (7).

B.1 GRADIENT STABILITY OF OMNIOPD

Standard OPD and an unsmoothed frequentist variant of OMNIOPD exhibit distinct local failure modes: reverse-KL supervision can produce extreme gradients under low teacher support, while a zero empirical compatibility estimate can remove the contribution of an audited chunk. Standard OPD's reverse-KL objective explodes when teacher mass vanishes on student-sampled tokens (Fu et al., 2026); OMNIOPD with empirical estimation collapses when teacher rollouts fail to match the student's chunk. Both failures concentrate on the same place where the teacher disagrees with the student, which are precisely the places that carry the most useful learning signal. A correct design must close both failure modes simultaneously. The theorem in this subsection establishes that OMNIOPD does so through two distinct mechanisms operating on two distinct components of the design.

Standard OPD's reverse-KL objective in Eq. (2) contains per-token terms of the form $\log\big(\pi_\theta(v)/\pi^{*}_{\text{teacher}}(v)\big)$, which diverge whenever the teacher assigns near-zero probability mass to a token the student samples. Recent analyses (Fu et al., 2026; Li et al., 2026; Luo et al., 2026) document this as a real failure mode of token-level OPD. The chunk-level loss in Eq. (7) was constructed to rule out this behavior by design. The teacher estimator enters as a bounded multiplier in $[0, 1]$ on the

student's log-likelihood, not as a denominator or inside a log, so the per-chunk gradient is controlled by the student's score function alone—independent of what the estimator returns or how badly the teacher disagrees with y_c.

This loss change does not, however, prevent every gradient failure. Under the most direct logit-free estimator—the empirical sample mean $\hat{\pi}^{(c)}_{\text{freq}} = k^{(c)}_{\text{sem}}/N$—the estimator collapses to *exactly* zero whenever none of the N teacher rollouts achieves semantic match with the student chunk. A zero multiplier in Eq. (7) removes the contribution of that audited chunk from the teacher-weighted objective. The Bayesian prior in Eq. (5) closes this gap. Because $\bar{\pi}^{(c)}_{\theta} > 0$ contributes a strictly positive term to the estimator at every chunk, the multiplier remains strictly positive for the fixed update regardless of the rollout outcome. Essentially, the student's own policy confidence acts as a fallback signal when the sampled teacher continuations exhibit no compatibility with the student chunk.

The Bayesian prior does more than prevent collapse. The convex combination $\hat{\pi}^{(c)}_{\text{teacher}} = \frac{N}{N+\alpha}\hat{\pi}^{(c)}_{\text{freq}} + \frac{\alpha}{N+\alpha}\bar{\pi}^{(c)}_{\theta}$ introduces a bias-variance tradeoff parameterized by the prior strength α: larger α reduces estimator noise but pulls the estimate further from the true expected compatibility μ_c, while smaller α inherits the empirical frequentist estimator's variance under small N. Neither extreme is optimal—the empirical frequentist limit ($\alpha \to 0$) is dominated by sampling noise, while the pure student-prior limit ($\alpha \to \infty$) ignores teacher rollouts entirely. Since $\hat{\pi}^{(c)}_{\text{teacher}}$ enters Eq. (7) as a multiplier on the per-chunk gradient, the estimator variance and its induced gradient variance contract as $\mathcal{O}(1/N)$, while the corresponding concentration width scales as $\mathcal{O}(1/\sqrt{N})$.

Theorem B.1 (Gradient stability of OMNIOPD)**.** *Under the chunk-level loss in Eq. (7) and the Bayesian path estimator from Eq. (5), the following three properties hold.*

(a) Loss design eliminates the gradient explosion of standard OPD. *For every realization of teacher rollouts and every estimator value $\hat{\pi}^{(c)}_{\textit{teacher}} \in [0, 1]$, the per-chunk gradient is bounded by the student's score function:*

$$\begin{aligned}\Big\|\hat{\pi}^{(c)}_{\textit{teacher}} \sum_{t\in c} \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})\Big\| \\ \leq \sum_{t\in c} \big\|\nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})\big\|.\end{aligned} \tag{10}$$

In contrast, the per-token gradient under the reverse-KL objective in Eq. (2) is unbounded as $\pi^{*}_{\textit{teacher}}(v \mid x, y_{<t}) \to 0$ for any sampled token v.

(b) Bayesian smoothing prevents zero-weight supervision collapse. *For every realization, including the worst case $k^{(c)}_{\textit{sem}} = 0$,*

$$\hat{\pi}^{(c)}_{\textit{teacher}} \;\geq\; \frac{\alpha\,\bar{\pi}^{(c)}_{\theta}}{N+\alpha} \;>\; 0, \tag{11}$$

so the compatibility multiplier remains strictly positive. Thus, zero sampled compatibility cannot by itself eliminate the chunk-level update, unlike the empirical frequentist estimator, whose multiplier vanishes when $k^{(c)}_{\textit{sem}} = 0$. The empirical frequentist estimator $\hat{\pi}^{(c)}_{\textit{freq}}$ does not satisfy this property: under $k^{(c)}_{\textit{sem}} = 0$, the chunk's gradient contribution vanishes entirely.

(c) Bias-variance characterization. *The Bayesian estimator's mean squared error admits the closed-form bias-variance decomposition*

$$\begin{aligned}\mathrm{MSE}\big(\hat{\pi}^{(c)}_{\textit{teacher}}\big) = \underbrace{\frac{N\sigma^2_{\phi}}{(N+\alpha)^2}}_{\textit{variance}} \\ + \underbrace{\frac{\alpha^2\big(\bar{\pi}^{(c)}_{\theta} - \mu_c\big)^2}{(N+\alpha)^2}}_{\textit{squared bias}},\end{aligned} \tag{12}$$

where $\sigma_\phi^2 := \mathrm{Var}_{y' \sim \pi^*_{teacher}}[\phi(y_c, y')] \leq 1/4$ *by Popoviciu's inequality (Popoviciu, 1965). The variance strictly contracts relative to the empirical frequentist baseline:* $\mathrm{Var}\big(\hat{\pi}^{(c)}_{teacher}\big) = \big(\frac{N}{N+\alpha}\big)^2 \mathrm{Var}\big(\hat{\pi}^{(c)}_{freq}\big) < \mathrm{Var}\big(\hat{\pi}^{(c)}_{freq}\big)$.

Proof. (a) Since $\hat{\pi}^{(c)}_{\text{teacher}} \in [0,1]$ as a convex combination of $\hat{\pi}^{(c)}_{\text{freq}} \in [0,1]$ and $\bar{\pi}^{(c)}_\theta \in [0,1]$, the triangle inequality together with $\hat{\pi}^{(c)}_{\text{teacher}} \leq 1$ gives Eq. (10). For the reverse-KL contrast, the gradient

$$\begin{aligned}&\nabla_\theta D_{\mathrm{KL}}\big(\pi_\theta \,\|\, \pi^*_{\text{teacher}}\big)\\ &\quad= \sum_v \nabla_\theta \pi_\theta(v)\, \log \frac{\pi_\theta(v)}{\pi^*_{\text{teacher}}(v)}\end{aligned}$$

contains terms whose magnitude diverges as $\pi^*_{\text{teacher}}(v) \to 0$ for any token v with $\pi_\theta(v) > 0$.

(b) From the convex form $\hat{\pi}^{(c)}_{\text{teacher}} = (k^{(c)}_{\text{sem}} + \alpha\bar{\pi}^{(c)}_\theta)/(N+\alpha)$, the setup assumption $\bar{\pi}^{(c)}_\theta > 0$, and the non-negativity of the semantic match count $k^{(c)}_{\text{sem}} \geq 0$,

$$\hat{\pi}^{(c)}_{\text{teacher}} \;=\; \frac{k^{(c)}_{\text{sem}} + \alpha\bar{\pi}^{(c)}_\theta}{N+\alpha} \;\geq\; \frac{\alpha\bar{\pi}^{(c)}_\theta}{N+\alpha} \;>\; 0.$$

The empirical frequentist estimator $\hat{\pi}^{(c)}_{\text{freq}} = k^{(c)}_{\text{sem}}/N$ admits no such lower bound: when $k^{(c)}_{\text{sem}} = 0$, $\hat{\pi}^{(c)}_{\text{freq}} = 0$.

(c) Conditional on the student trajectory $\hat{y}$, the prior $\bar{\pi}^{(c)}_\theta$ is fixed and only the empirical frequentist estimator is random. Since the rollouts are i.i.d. with $\phi(y_c, y^{(i)}_{\text{teacher}}) \in [0,1]$,

$$\begin{aligned}\mathbb{E}\big[\hat{\pi}^{(c)}_{\text{freq}}\big] &= \mu_c,\\ \mathrm{Var}\big(\hat{\pi}^{(c)}_{\text{freq}}\big) &= \frac{\sigma^2_\phi}{N},\end{aligned}$$

where $\sigma^2_\phi \leq \frac{1}{4}$ by Popoviciu's inequality (Popoviciu, 1965). Applying the convex form from Eq. (9), the bias and variance of the Bayesian estimator decompose as

$$\begin{aligned}\mathrm{Bias}\big(\hat{\pi}^{(c)}_{\text{teacher}}\big) &= \frac{N}{N+\alpha}\mu_c + \frac{\alpha}{N+\alpha}\bar{\pi}^{(c)}_\theta - \mu_c\\ &= \frac{\alpha\big(\bar{\pi}^{(c)}_\theta - \mu_c\big)}{N+\alpha},\\ \mathrm{Var}\big(\hat{\pi}^{(c)}_{\text{teacher}}\big) &= \Big(\frac{N}{N+\alpha}\Big)^2 \mathrm{Var}\big(\hat{\pi}^{(c)}_{\text{freq}}\big)\\ &= \frac{N\sigma^2_\phi}{(N+\alpha)^2}.\end{aligned}$$

Summing squared bias and variance yields Eq. (12). The variance contraction is immediate: $(N/(N+\alpha))^2 < 1$ for all $\alpha > 0$. □

B.2 Rollout Sufficiency of OmniOPD

OmniOPD's efficiency advantage over standard OPD rests entirely on the rollout budget per chunk being small. The per-sample cost analysis in Appendix E.10 shows that, under the prefix-reuse regime reported in Table 12, the default configuration ($M = 10$, $N = 10$, $C = 50$) requires approximately $1.13\times$ the normalized inference effort of long-CoT SFT, while the sparser variant ($M = 5$, $N = 10$, $C = 50$) requires approximately $0.56\times$. The theorem in this subsection characterizes how estimator uncertainty decreases with the rollout budget N, providing a theoretical basis for the diminishing returns observed empirically when increasing N beyond the default value.

The remaining question is whether a sparse configuration such as $N = 10$ gives a usable signal. Two effects work in OmniOPD's favor. First, the empirical estimator $\hat{\pi}^{(c)}_{\text{freq}}$ is a sample mean over N

i.i.d. rollouts bounded in $[0, 1]$, so by Hoeffding's inequality (Hoeffding, 1963) its deviation from μ_c shrinks at sub-Gaussian rate in N. Second, the Bayesian shrinkage in Eq. (5) tightens this further by a factor of $N/(N+\alpha) < 1$. Since $\hat{\pi}^{(c)}_{\text{teacher}}$ enters Eq. (7) as a multiplier on the per-chunk gradient, the estimator variance and its induced gradient variance contract as $\mathcal{O}(1/N)$, while the corresponding concentration width scales as $\mathcal{O}(1/\sqrt{N})$. This predicts diminishing statistical returns from increasing the rollout budget, consistent with the empirical sensitivity results.

Theorem B.2 (Concentration of the chunk-level estimator)**.** *For any $\varepsilon > 0$, the Bayesian chunk-level estimator concentrates around its shrunken target $\tilde{\mu}_c := \frac{N}{N+\alpha}\mu_c + \frac{\alpha}{N+\alpha}\bar{\pi}^{(c)}_\theta$ with sub-Gaussian rate*

$$\begin{aligned} &\Pr\Big(\big|\hat{\pi}^{(c)}_{\textit{teacher}} - \tilde{\mu}_c\big| \geq \varepsilon\Big) \\ &\leq 2\exp\left(-\frac{2(N+\alpha)^2\varepsilon^2}{N}\right). \end{aligned} \tag{13}$$

Consequently, the estimator-induced contribution to the per-chunk gradient variance is bounded by

$$\begin{aligned} &\mathrm{Var}\left(\hat{\pi}^{(c)}_{\textit{teacher}} \sum_{t\in c} \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t}) \,\Big|\, \hat{y}\right) \\ &\leq \frac{N}{4(N+\alpha)^2} \Big\| \sum_{t\in c} \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t}) \Big\|^2, \end{aligned} \tag{14}$$

shrinking at rate $\mathcal{O}(1/N)$ in the rollout budget.

Proof. Conditional on the student trajectory $\hat{y}$, decompose $\hat{\pi}^{(c)}_{\text{teacher}} - \tilde{\mu}_c = \frac{N}{N+\alpha}\big(\hat{\pi}^{(c)}_{\text{freq}} - \mu_c\big)$. Applying Hoeffding's inequality (Hoeffding, 1963) to $\hat{\pi}^{(c)}_{\text{freq}}$ as an average of N i.i.d. samples in $[0, 1]$,

$$\Pr\Big(\big|\hat{\pi}^{(c)}_{\text{freq}} - \mu_c\big| \geq \delta\Big) \leq 2\exp(-2N\delta^2),$$

and substituting $\delta = \frac{N+\alpha}{N}\varepsilon$ yields Eq. (13). For the variance bound, the score sum $\sum_{t\in c} \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})$ is deterministic conditional on $\hat{y}$, so the gradient variance factorizes as $\mathrm{Var}\big(\hat{\pi}^{(c)}_{\text{teacher}}\big) \cdot \big\|\sum_{t\in c} \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})\big\|^2$. Substituting $\mathrm{Var}\big(\hat{\pi}^{(c)}_{\text{teacher}}\big) \leq N/4(N+\alpha)^2$ from Theorem B.1(c) gives Eq. (14). □

B.3 Un-audited Token Stability of OmniOPD

OmniOPD audits only $M \cdot C$ tokens per trajectory and leaves the remaining $T - M \cdot C$ tokens un-audited. This sparsity is the design feature that makes OmniOPD efficient—one teacher query per chunk rather than per token. However, without any constraint on the un-audited region, the student can exploit the sparse audited supervision in ways that degrade the trajectory globally, including collapsing entropy on un-audited tokens to shorten trajectories and reduce penalty mass, generating incoherent connective text between audited chunks, or shifting probability mass on un-audited tokens to satisfy gradients from the audited chunks at the cost of long-horizon fluency.

The KL anchor against the frozen base policy π_{ref} closes these failure modes by constraining the un-audited region directly. At each un-audited token, Pinsker's inequality (Pinsker, 1964) bounds the total-variation distance between π_{ref} and π_θ by the per-token KL, so reducing per-token KL through the anchor term directly bounds the per-token drift. At the aggregate level, θ_{ref} is always a feasible point of the objective at which the anchor's per-token KL terms vanish, so the optimality of θ^* forces an aggregate constraint. The total KL summed over un-audited tokens at θ^* cannot exceed the chunk-level loss improvement $\mathcal{L}_{\text{chunk}}(\theta_{\text{ref}}) - \mathcal{L}_{\text{chunk}}(\theta^*)$ divided by β. Beyond a regularization weight, β therefore has a concrete operational meaning, i.e., β^{-1} is the exchange rate at which OmniOPD converts chunk-level loss improvement into permitted un-audited drift. Theorem B.3 below makes both bounds precise—the pointwise total-variation control from Pinsker's inequality, and the aggregate KL budget that scales inversely with β.

Theorem B.3 (Trust-region bound on un-audited tokens)**.** *For a fixed sampled trajectory, audited chunk set, and corresponding stop-gradient compatibility weights, consider the conditional* OMNIOPD *objective in Eq. (7). At any iterate θ and for every un-audited position $t \in \mathcal{U}$,*

$$\begin{aligned} &D_{\mathrm{TV}}\big(\pi_{\mathit{ref}}(\cdot \mid x, y_{<t}) \,\|\, \pi_\theta(\cdot \mid x, y_{<t})\big) \\ &\quad \le \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}\big(\pi_{\mathit{ref}}(\cdot \mid x, y_{<t}) \,\|\, \pi_\theta(\cdot \mid x, y_{<t})\big)}. \end{aligned} \tag{15}$$

Furthermore, any minimizer θ^* of this conditional objective satisfies the aggregate budget

$$\begin{aligned} &\sum_{t \in \mathcal{U}} D_{\mathrm{KL}}\big(\pi_{\mathit{ref}}(\cdot \mid x, y_{<t}) \,\|\, \pi_{\theta^*}(\cdot \mid x, y_{<t})\big) \\ &\quad \le \frac{1}{\beta}\Big[\mathcal{L}_{\mathit{chunk}}(\theta_{\mathit{ref}}) - \mathcal{L}_{\mathit{chunk}}(\theta^*)\Big]. \end{aligned} \tag{16}$$

As $\beta \to \infty$, the aggregate un-audited drift vanishes and $\pi_{\theta^*} \to \pi_{\mathit{ref}}$ on $\mathcal{U}$.

Proof. The pointwise total-variation bound in Eq. (15) is Pinsker's inequality applied at each $t \in \mathcal{U}$. For the aggregate budget in Eq. (16), evaluate $\mathcal{L}_{\text{OMNIOPD}}$ at the feasible point θ_{ref}, where the KL anchor term vanishes by construction:

$$\begin{aligned} \mathcal{L}_{\text{chunk}}(\theta^*) + \beta \sum_{t \in \mathcal{U}} D_{\mathrm{KL}}\big(\pi_{\text{ref}} \,\|\, \pi_{\theta^*}\big) &\le \mathcal{L}_{\text{OMNIOPD}}(\theta^*) \\ &\le \mathcal{L}_{\text{OMNIOPD}}(\theta_{\text{ref}}) \;=\; \mathcal{L}_{\text{chunk}}(\theta_{\text{ref}}), \end{aligned}$$

using optimality of θ^* and the vanishing of the KL term at θ_{ref}. Rearranging yields the stated budget. The limiting statement $\pi_{\theta^*} \to \pi_{\text{ref}}$ as $\beta \to \infty$ follows because the right-hand side is bounded above by a constant independent of β, so the aggregate KL must tend to zero. □

B.4 TOKENIZER- AND STYLE-INVARIANT SUPERVISION

In the logit-free setting that OMNIOPD addresses, the natural token-level alternative is sequence-level distillation (Kim & Rush, 2016)—sample N teacher rollouts and apply token-level cross-entropy on each rollout as if it were a ground-truth sequence. The resulting supervision forces the student to match the teacher's surface-form choices token by token, conflating semantic content with stylistic and formatting decisions. Standard OPD, although not directly comparable in the logit-free setting, exhibits a structurally analogous property at the distributional level—its reverse-KL objective in Eq. (2) contains terms $\log \pi^*_{\text{teacher}}(v \mid x, y_{<t})$ summed against every vocabulary token v, so the student inherits the teacher's full per-token preferences, including distinctions that are semantically irrelevant.

OMNIOPD's chunk-level loss decouples semantic content from surface form by routing all teacher information through the semantic similarity function ϕ. The Bayesian estimator $\hat{\pi}^{(c)}_{\text{teacher}}$ depends on rollouts only through $k^{(c)}_{\text{sem}}$, and the loss in Eq. (7) depends on rollouts only through this estimator. Any invariance property ϕ has is therefore inherited by the chunk-level loss. Theorem B.4 below formalizes this inheritance and contrasts it with the two token-level alternatives above.

Theorem B.4 (Tokenizer and stylistic invariance of the chunk-level loss)**.** *Let $\sim_\phi$ denote the equivalence relation on token sequences induced by ϕ: $y \sim_\phi y'$ iff $\phi(z, y) = \phi(z, y')$ for every z. For any two sets of teacher rollouts $\{y^{(i)}_{\mathit{teacher}}\}_{i=1}^N$ and $\{y'^{(i)}_{\mathit{teacher}}\}_{i=1}^N$ that are $\sim_\phi$-equivalent (i.e., $y^{(i)}_{\mathit{teacher}} \sim_\phi y'^{(i)}_{\mathit{teacher}}$ for every i), the chunk-level Bayesian estimator and the chunk-level loss in Eq. (7) are invariant:*

$$\begin{aligned} \hat{\pi}^{(c)}_{\mathit{teacher}}\big(\{y^{(i)}_{\mathit{teacher}}\}\big) &= \hat{\pi}^{(c)}_{\mathit{teacher}}\big(\{y'^{(i)}_{\mathit{teacher}}\}\big), \\ \mathcal{L}_{\mathit{chunk}}\big(\theta; \{y^{(i)}_{\mathit{teacher}}\}\big) &= \mathcal{L}_{\mathit{chunk}}\big(\theta; \{y'^{(i)}_{\mathit{teacher}}\}\big). \end{aligned} \tag{17}$$

In contrast, the sequence-level distillation loss

$$\begin{aligned} &\mathcal{L}_{\mathit{SFT}}\big(\theta; \{y^{(i)}_{\mathit{teacher}}\}\big) := \\ &\quad -\frac{1}{N} \sum_{i=1}^{N} \sum_{t} \log \pi_\theta\big(y^{(i)}_{\mathit{teacher},t} \mid x, y^{(i)}_{\mathit{teacher},<t}\big) \end{aligned} \tag{18}$$

depends on specific teacher tokens and is not invariant under $\sim_\phi$-substitution. The standard OPD loss in Eq. (2) depends on the teacher's per-token distribution $\pi^*_{teacher}$ and distinguishes $\sim_\phi$-equivalent tokens whenever the teacher assigns them different probabilities.

Proof. The chunk-level loss in Eq. (7) depends on the teacher rollouts $\{y^{(i)}_{\text{teacher}}\}$ only through the Bayesian estimator $\hat{\pi}^{(c)}_{\text{teacher}}$ from Eq. (5), which in turn depends on the rollouts only through $k^{(c)}_{\text{sem}} = \sum_i \phi(y_c, y^{(i)}_{\text{teacher},c})$. By the definition of $\sim_\phi$, $\phi(y_c, y^{(i)}_{\text{teacher},c}) = \phi(y_c, y'^{(i)}_{\text{teacher},c})$ for every i, so $k^{(c)}_{\text{sem}}$ is invariant, $\hat{\pi}^{(c)}_{\text{teacher}}$ is invariant, and $\mathcal{L}_{\text{chunk}}$ is invariant. The non-invariance of the sequence-level distillation loss in Eq. (18) is immediate from its definition—it contains $\log \pi_\theta(y^{(i)}_{\text{teacher},t} \mid \cdot)$ evaluated at the specific teacher tokens, which generally differ between $\sim_\phi$-equivalent rollouts. For standard OPD, Eq. (2) sums $\pi_\theta(v \mid \cdot) \log \pi^*_{\text{teacher}}(v \mid \cdot)$ over the entire vocabulary, so the loss is sensitive to any per-token distinction in π^*_{teacher}, including those between $\sim_\phi$-equivalent tokens. □

Remark B.5 (Three invariance classes). The choice of ϕ determines which surface-level variations the chunk-level loss treats as equivalent, yielding three practically important invariance properties.

Tokenizer invariance. If ϕ operates on detokenized text—for instance, an embedding-based semantic similarity over decoded strings—then the chunk loss is invariant to the teacher's tokenizer. Teachers with different vocabularies and tokenization schemes yield the same supervision signal up to ϕ.

Paraphrase invariance. If ϕ captures chunk-level semantic equivalence, teacher rollouts that paraphrase the same content yield identical chunk loss. The student is not penalized for choosing a different surface realization of the same semantic content.

Stylistic invariance. If ϕ is insensitive to register, formatting, or hedging conventions, teacher rollouts that differ only in stylistic dimensions yield identical chunk loss.

The empirical generalization of OMNIOPD across the Qwen3-32B (Yang et al., 2025) and Gemini-2.5-flash (Comanici et al., 2025) teachers in later section 3.2—two model families with different tokenizers, training data, and stylistic preferences—is consistent with these invariance properties.

C PROMPT TEMPLATES

To ensure a controlled comparison across student and teacher models, we apply the same plain-text prompt template across both training (DAPO-Math-17K) and all evaluation benchmarks. We do not apply model-specific chat templates or system prompts during evaluation; the prompt below is passed directly to every evaluated model. Consequently, the teacher scores reported in our experiments reflect this unified evaluation protocol and should not be interpreted as reproductions of model-specific numbers reported under their original evaluation recipes. The template additionally specifies the final-answer format to enable reliable automated parsing. As shown in Table 4, this template includes a specific directive to format the final answer, ensuring reliable automated parsing via regular expressions.

Table 4: Prompt template utilized for all mathematical reasoning tasks (both training and evaluation). The {Question} placeholder is replaced with the specific problem text during inference.

Application	Prompt Template
Training & Evaluation	`Solve the following math problem step by step. The last line of your response should be of the form Answer: \$Answer (without quotes) where \$Answer is the answer to the problem.\n\n{Question}`

D SUPPLEMENTARY DETAILS OF OMNIOPD

D.1 ADDITIONAL IMPLEMENTATION DETAILS

We optimize all student models for 1 epoch using a learning rate of 1×10^{-6} and a per-device batch size of 2. To accommodate extended reasoning trajectories, the maximum sequence length is capped at 20K tokens during training and extended to 32K tokens during evaluation. During trajectory generation, the student model's sampling temperature is set to 1.0. For experiments with hosted teachers (Claude-4.5-haiku and Gemini-2.5-flash), we use a matched 50-step training protocol for both Seq-KD and OMNIOPD to control teacher-inference exposure across methods. This setting is specific to our hosted-teacher experiments rather than an algorithmic requirement of OMNIOPD. As shown in Appendix E.3, the hosted-teacher training loss approaches its plateau within this optimization window.

Training and Hyperparameter Protocol. For the main student and teacher comparison in Table 1, we use a matched training protocol to reduce confounding from optimization-budget differences. Seq-KD, Best-of-N SFT, STaR, standard OPD, and OMNIOPD use the same student initialization and DAPO-Math-17K training split and are trained for one epoch with the same learning rate of 1×10^{-6}, a global batch size of 64, and a maximum generation length of 20K tokens. All methods are evaluated using the same downstream evaluation pipeline. Reward-based baselines (GRPO and DAPO) use the same peak learning rate, generation length, group size 8, and one-epoch training budget, together with their standard method-specific rollout and optimization settings.

E SUPPLEMENTARY RESULTS

E.1 EMPIRICAL VALIDATION OF THE SEMANTIC TEACHER PROXY

A central ingredient of OMNIOPD is the use of semantic agreement with sampled teacher continuations as a surrogate for the teacher's local preference. We emphasize that this surrogate is not intended to establish whether an individual reasoning step is independently correct. Instead, the relevant question for on-policy distillation is whether the resulting black-box signal preserves useful information about how the teacher evaluates student-generated reasoning paths.

To examine this relationship directly, we use the open-weight Qwen3-32B model as the verification oracle. Because its internal output distributions are accessible, we can compute the corresponding teacher probability signal for student-generated chunks and compare it against the black-box estimates derived solely from sampled textual continuations. For each student-generated chunk y_c, we define the white-box teacher reference signal as the geometric-mean probability assigned by the teacher to the realized student tokens:

$$s^{(c)}_{\text{teacher}} = \exp\left(\frac{1}{C}\sum_{t\in c}\log \pi_{\text{teacher}}(y_t \mid x, y_{<t})\right) = \left(\prod_{t\in c}\pi_{\text{teacher}}(y_t \mid x, y_{<t})\right)^{1/C}. \tag{19}$$

We use this quantity only as a diagnostic reference for evaluating how well the logit-free semantic proxy tracks teacher preference. We evaluate both the frequentist estimator $\hat{\pi}_{freq}$ and our Bayesian-smoothed estimator $\hat{\pi}_{teacher}$ using Mean Squared Error (MSE) and Pearson correlation across several semantic measurement functions ϕ and Monte Carlo rollout budgets N. Table 5 summarizes the results.

The results establish two critical dynamics within the OMNIOPD framework. First, the Bayesian estimator strictly dominates the frequentist baseline across every evaluated metric and sample size. By stabilizing the sparse teacher signals with the student's normalized prior, the Bayesian update systematically reduces the MSE and yields a stronger positive correlation with the white-box teacher reference signal. This empirical trend is consistent with the variance-reduction effect characterized in Theorem B.1.

Second, the data reveals a clear trade-off between magnitude accuracy and rank correlation. Rigid criteria like exact match suffer from severe sparsity, producing the highest estimation errors (MSE > 0.24). Strict structural metrics like Edit Distance also underestimate absolute magnitudes (MSE ≈ 0.148) but achieve the highest Pearson correlation (0.72), exceptionally preserving the relative

Algorithm 1 The OMNIOPD Optimization Framework

Input: Base reference policy π_{ref}, student policy π_θ, black-box teacher π^*_{teacher}
Hyperparameters: chunk number M, chunk size C, Monte Carlo rollouts N, prior strength α, KL penalty weight β, learning rate η
Output: Optimized student policy π_θ

while not converged **do**
  Sample instruction $x \sim \mathcal{D}$
  // 1. On-Policy Trajectory Generation
  Generate student trajectory $y = (y_1, \ldots, y_T) \sim \pi_\theta(\cdot \mid x)$
  // 2. Peak-Entropy Anchored Scheduling
  for $t = 1$ to T **do**
    Compute predictive entropy: $\mathcal{H}_t = -\sum_{v \in V} \pi_\theta(v \mid x, y_{<t}) \log \pi_\theta(v \mid x, y_{<t})$
  end for
  Rank candidate anchor positions in descending order of $\mathcal{H}_t$
  Initialize audited chunk set $\mathcal{C} \leftarrow \emptyset$
  for each candidate anchor t in ranked order **do**
    if $|\mathcal{C}| = M$ **then**
      break
    end if
    Form a contiguous C-token candidate chunk c_t containing anchor t
    if $c_t \cap c = \emptyset$ for all $c \in \mathcal{C}$ **then**
      $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_t\}$
    end if
  end for
  Identify un-audited tokens: $\mathcal{U} = \{1, \ldots, T\} \setminus \bigcup_{c \in \mathcal{C}} c$
  // 3. Speculative Verification & Bayesian Smoothing
  for each chunk $c \in \mathcal{C}$ **do**
    Query π^*_{teacher} with prefix $y_{<c}$ and sample N independent C-token continuations $\{y^{(i)}_{\text{teacher}}\}_{i=1}^N$
    Compute semantic compatibility evidence: $k^{(c)}_{\text{sem}} = \sum_{i=1}^N \phi\left(y_c, y^{(i)}_{\text{teacher}}\right)$
    Compute student-policy prior: $\bar{\pi}^{(c)}_\theta = \left(\prod_{t \in c} \pi_\theta(y_t \mid x, y_{<t})\right)^{1/C}$
    Compute smoothed teacher-compatibility proxy: $\hat{\pi}^{(c)}_{\text{teacher}} = \dfrac{k^{(c)}_{\text{sem}} + \alpha \bar{\pi}^{(c)}_\theta}{N + \alpha}$
  end for
  // 4. Trust-Region Anchoring & Optimization
  Compute teacher-weighted chunk loss: $\mathcal{L}_{\text{chunk}} = -\sum_{c \in \mathcal{C}} \text{sg}\left[\hat{\pi}^{(c)}_{\text{teacher}}\right] \sum_{t \in c} \log \pi_\theta(y_t \mid x, y_{<t})$
  Compute un-audited trust-region penalty: $\mathcal{L}_{\text{KL}} = \beta \sum_{t \in \mathcal{U}} D_{\text{KL}}\left(\pi_{\text{ref}}(\cdot \mid x, y_{<t}) \,\|\, \pi_\theta(\cdot \mid x, y_{<t})\right)$
  Compute total objective: $\mathcal{L}_{\text{OMNIOPD}}(\theta) = \mathcal{L}_{\text{chunk}} + \mathcal{L}_{\text{KL}}$
  Update student: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{OMNIOPD}}(\theta)$
end while
return π_θ

ranking of reasoning paths. Conversely, unigram metrics like ROUGE-1 bridge the tokenization gap to yield the most accurate absolute estimations, achieving the lowest MSE (≈ 0.022). We systematically compare the empirical impact of these two metrics in our sensitivity analysis (section 4.2).

E.2 Results on Programming Benchmarks

We present the main empirical results for competitive programming in Table 6.

Table 5: Comparison of estimation accuracy between Frequentist and Bayesian estimators across different semantic similarity metrics. Evaluated over 1000 samples with chunk size $C = 50$, number of chunks $M = 20$, and prior strength $\alpha = 1.0$. The Bayesian estimator consistently reduces MSE and improves correlation with the white-box teacher reference signal.

Semantic Metric	Rollouts (N)	MSE (Freq) ↓	MSE (Bayes) ↓	Corr (Freq) ↑	Corr (Bayes) ↑
`ROUGE-1` (Lin, 2004)	10	0.0279	**0.0217**	0.5530	0.6042
	20	0.0274	0.0241	0.5569	0.5838
`ROUGE-L` (Lin, 2004)	10	0.0667	0.0521	0.6396	0.6875
	20	0.0663	0.0584	0.6466	0.6726
`Jaccard` (Niwattanakul et al., 2013)	10	0.0827	0.0650	0.5865	0.6464
	20	0.0826	0.0730	0.5885	0.6210
`BLEU-1` (Papineni et al., 2002)	10	0.0279	0.0217	0.5539	0.6048
	20	0.0273	0.0240	0.5571	0.5842
`BLEU-2` (Papineni et al., 2002)	10	0.1206	0.0952	0.6402	0.6835
	20	0.1202	0.1066	0.6476	0.6707
`Edit distance` (Navarro, 2001)	10	0.1857	0.1478	0.6625	**0.7215**
	20	0.1851	0.1647	0.6713	0.7033
`Exact match`	10	0.3009	0.2417	0.4625	0.6994
	20	0.3006	0.2688	0.5051	0.6468

Table 6: Main results on competitive programming benchmarks.

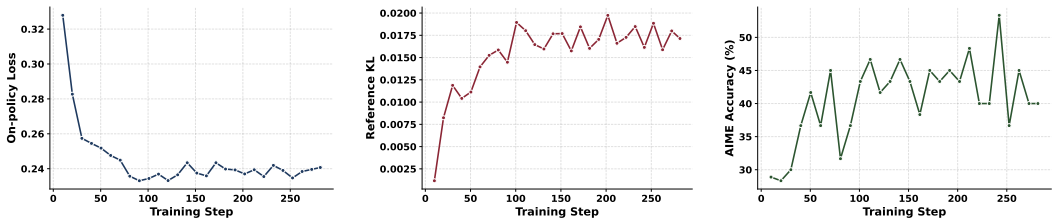
Student Model	Teacher Model	Method	APPS	CodeContests	Codeforces	TACO	Average
Qwen3-1.7B	Qwen3-32B	Seq-KD	46.78	45.16	43.39	26.43	40.44
		OPD	52.67	52.09	**52.89**	30.61	47.06
		OMNIOPD	**54.28**	**52.94**	51.55	**32.96**	**47.93**
Qwen3-4B	Qwen3-32B	Seq-KD	69.82	65.59	68.60	44.12	62.03
		OPD	**72.72**	**68.86**	**74.33**	**45.10**	**65.26**
		OMNIOPD	71.72	68.52	70.49	44.37	63.78

E.3 TRAINING DYNAMICS

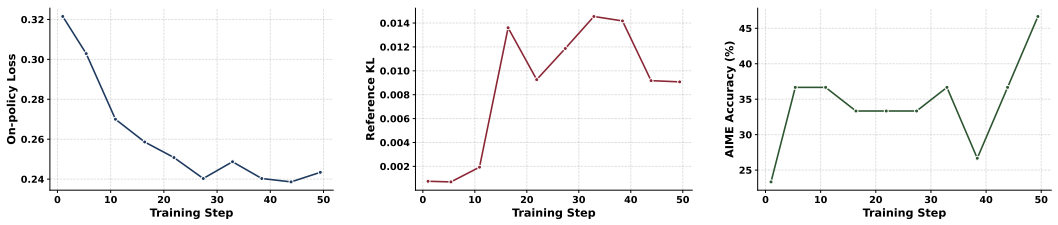
To understand the optimization stability and learning progression of OMNIOPD, we visualize the training dynamics of the Qwen3-4B student in figure 3. We track the process under both an open-weights teacher (Qwen3-32B) and a frontier black-box teacher (Gemini-2.5-flash). The chunk-level on-policy loss exhibits rapid and stable convergence across both setups. For the Qwen3-32B teacher, the loss drops sharply from approximately 0.33 to 0.24 within the first 100 steps before stabilizing. Remarkably, when utilizing the Gemini-2.5-flash teacher, the student achieves a similar loss plateau in merely 30 steps. This steady convergence is consistent with the bounded-gradient behavior characterized in Theorem B.1.

As the student rapidly deviates from its sub-optimal base behavior to align with the teacher's superior reasoning structures, we observe a corresponding initial rise in the KL divergence between the active policy and the frozen base model. Crucially, rather than growing unboundedly, the KL curves safely plateau. This plateau is consistent with the role of the trust-region penalty characterized in Theorem B.3, which constrains policy drift on un-audited tokens.

This stable optimization directly translates to continuous progression on downstream evaluations. Across the training steps, the pass-rate accuracy on the challenging AIME-2025 dataset climbs substantially alongside the reduction in on-policy loss (note that the performance gap between this validation curve and the final results in Table 1 is due to maximum token truncation during intermediate evaluation). Under the Qwen3-32B teacher, AIME accuracy rises from a sub-30% baseline to peak above 50% near step 240. Furthermore, the Gemini-2.5-flash teacher drives the student from 23% to roughly 47% accuracy within a strictly constrained 50-step budget. These curves show that downstream accuracy improves alongside optimization under both teacher settings. The specific contribution of peak-entropy scheduling is evaluated separately through the controlled ablation in Table 2.



(a) On-Policy Loss (Qwen3-32B) (b) Reference KL (Qwen3-32B) (c) AIME Accuracy (Qwen3-32B)



(d) On-Policy Loss (Gemini-2.5-Flash) (e) Reference KL (Gemini-2.5-Flash) (f) AIME Accuracy (Gemini-2.5-Flash)

Figure 3: Training dynamics of the Qwen3-4B student distilled via OMNIOPD. The top row utilizes the open-weight Qwen3-32B teacher, while the bottom row utilizes the frontier Gemini-2.5-flash model. We track the on-policy loss (left), the base-model KL divergence penalty (center), and the continuous improvement in AIME 2025 evaluation accuracy (right) across training steps.

E.4 TEACHER ACCURACY AND DISTILLATION EFFECTIVENESS

Teacher standalone benchmark accuracy and downstream distillation effectiveness are related but not identical. In OMNIOPD, the teacher is queried on student-induced prefixes and provides localized supervision at selected reasoning chunks rather than complete trajectory imitation. Consequently, the usefulness of a teacher also depends on its behavior under the student's state distribution and the compatibility of its local reasoning signal with the student policy. This can lead to non-monotonic teacher–student trends, where a teacher with lower standalone average accuracy yields a stronger distilled student. Similar effects have been observed in generalized on-policy distillation (Yang et al., 2026).

E.5 ROBUSTNESS TO STUDENT-PREFIX CONDITIONING

A defining property of on-policy distillation is that teacher supervision is evaluated on histories induced by the student policy. In OMNIOPD, the teacher therefore generates its Monte Carlo continuations from the student's prefix $y_{<c}$. While this preserves the on-policy training distribution, it introduces a potential source of bias: if the student has already entered an erroneous reasoning path, the teacher may inherit that error rather than recover from it.

We empirically examine how frequently this occurs using the same high-entropy fork-selection procedure employed during training. We sample 200 training problems using a Qwen3-4B student with 63.0% solution accuracy and collect 3,200 teacher continuations from high-entropy student prefixes using the Qwen3-32B teacher. We categorize each teacher continuation according to whether it (i) recovers to the correct final answer, (ii) reaches the same incorrect final answer as the student trajectory, or (iii) reaches a different incorrect answer.

As shown in Table 7, conditioning on student-generated histories does not generally force the teacher to reproduce the student's trajectory. Among continuations conditioned on prefixes from failed student solutions, the teacher recovers to the correct final answer in 26.4% of cases, while reproducing the student's same incorrect answer in 24.1%. The remaining 49.5% reach a different incorrect answer. Thus, erroneous student histories introduce meaningful noise into the teacher signal, but direct inheritance of the student's particular error is not the dominant outcome even among failed

Table 7: Teacher continuation behavior under student-conditioned prefixes. "Same student error" denotes a teacher continuation that reaches the same incorrect final answer as the corresponding failed student trajectory.

Prefix source	Continuations	Teacher correct	Same student error	Other wrong
All trajectories	3,200	65.9	8.9	25.2
Failed student trajectories	1,184	26.4	24.1	49.5
Passed student trajectories	2,016	89.0	–	11.0

Table 8: Teacher continuation behavior across fork depth for prefixes from failed student trajectories.

Fork depth	Continuations	Teacher correct	Same student error	Other wrong
Early	352	22.2	6.2	71.6
Middle	378	28.8	10.8	60.3
Late	454	27.8	48.9	23.3

trajectories. For prefixes drawn from successful student trajectories, the teacher remains correct in 89.0% of continuations.

We further examine whether this effect depends on how deeply the student has progressed along an unsuccessful reasoning trajectory. We partition the high-entropy forks from failed trajectories into early, middle, and late positions and report the corresponding teacher continuation behavior in Table 8.

The effect is strongly dependent on trajectory depth. As shown in Table 8, direct inheritance of the student's incorrect final answer occurs in only 6.2% and 10.8% of teacher continuations from early and middle forks, respectively, but rises sharply to 48.9% for late forks. This suggests that prefix-induced bias is concentrated after the student trajectory has already committed substantially to an erroneous solution path. At earlier high-entropy decision points, the teacher rarely reproduces the student's exact error, although a substantial fraction of continuations can still remain incorrect. Student-conditioned teacher rollouts should therefore be interpreted as noisy teacher-preference signals rather than guaranteed corrective trajectories.

Overall, these results characterize rather than eliminate the limitation of student-prefix conditioning. The teacher is generally able to deviate from the student's erroneous trajectory at early and intermediate forks, while late-stage prefixes are substantially more susceptible to inherited errors. This behavior supports the use of teacher continuations as localized preference signals at uncertain reasoning forks, while identifying late, already-committed trajectories as a regime in which the semantic teacher critic becomes less reliable.

E.6 Robustness under Low Teacher–Student Token Overlap

To examine how the semantic teacher proxy behaves when teacher and student generations have limited token-level agreement, we conduct a diagnostic in the Qwen3-4B/Qwen3-32B setting, where the teacher's token likelihood is directly accessible. We sample 1,600 audited chunks and partition them into terciles according to teacher–student token overlap. For each bucket, we measure the correlation between teacher log-likelihood and downstream correctness, between the semantic compatibility score k_{sem} and correctness, and between k_{sem} and teacher log-likelihood.

The semantic proxy remains informative even when token-level agreement is limited. In the lowest-overlap tercile, where average token overlap is only 0.347, k_{sem} remains significantly correlated with both teacher log-likelihood ($r = 0.353$) and downstream correctness ($r = 0.105$). Its correlation with teacher likelihood becomes stronger as overlap increases, indicating that the text-only approximation becomes more faithful when the two models' local generations are more closely aligned.

Importantly, teacher log-likelihood itself remains more strongly associated with correctness than k_{sem} in every overlap bucket. We therefore do not interpret the semantic proxy as improving upon white-box teacher likelihood when logits are available. Rather, this diagnostic shows that the chunk-level score preserves non-trivial information about teacher preference even in low-overlap regions,

Table 9: Behavior of the semantic teacher signal across teacher–student token-overlap regimes. Correlations are Pearson coefficients with corresponding p-values.

Overlap	n	**Avg. overlap**	**Teacher LL**	$r(\mathrm{LL}, \mathrm{correct})$	$r(k_{\mathrm{sem}}, \mathrm{correct})$	$r(k_{\mathrm{sem}}, \mathrm{LL})$
Low	533	0.347	-0.73	0.204 ($p = 2.0\times10^{-6}$)	0.105 ($p = 1.5\times10^{-2}$)	0.353 ($p = 4.4\times10^{-17}$)
Mid	533	0.447	-0.62	0.161 ($p = 1.9\times10^{-4}$)	0.102 ($p = 1.9\times10^{-2}$)	0.438 ($p = 2.3\times10^{-26}$)
High	534	0.581	-0.41	0.156 ($p = 2.8\times10^{-4}$)	0.071 ($p = 1.0\times10^{-1}$)	0.639 ($p = 1.3\times10^{-62}$)

Table 10: Training dynamics of the no-KL variant. Removing the reference anchor produces rapid length and entropy collapse even as the optimization objective continues to decrease.

Training Step	**Mean Length**	**Entropy**	**Semantic Score (k/N)**	**Training Loss**
1	9,232	0.399	4.3/20	0.195
51	3,326	–	–	–
151	1,001	–	–	–
201	78	0.033	7.7/20	0.017

where exact token-level agreement is limited. This is the regime relevant to OMNIOPD: when teacher probabilities are inaccessible, semantic rollout agreement provides a usable text-only surrogate without requiring a shared token-level distribution.

E.7 FAILURE ANALYSIS WITHOUT TRUST-REGION ANCHORING

The ablation in Table 2 shows that removing the reference KL anchor causes a severe performance collapse, reducing average mathematical reasoning accuracy from 69.08% to 8.28%. We further analyze the training dynamics of this variant to characterize the underlying failure mode. Rather than exhibiting unstable loss or immediate divergence, the no-KL policy initially appears to optimize normally but gradually collapses toward extremely short and low-entropy generations.

Length and entropy collapse. The most visible symptom is a rapid reduction in generation length. The mean student trajectory length decreases from 9,232 tokens at training step 1 to 3,326 at step 51, 1,001 at step 151, and only 78 tokens by step 201. Predictive entropy decreases in parallel, from 0.399 at the beginning of training to approximately 0.033 at the collapsed checkpoint. The resulting policy is therefore not only substantially shorter, but also nearly deterministic. In contrast, with the KL anchor enabled, mean generation lengths remain in the 7,600–9,500 token range and predictive entropy remains between approximately 0.18 and 0.40 throughout training.

Misleading improvement in the local teacher signal. Importantly, the semantic compatibility signal does not reveal this collapse. As the trajectories become progressively shorter, the mean semantic agreement score increases from 4.3/20 to 7.7/20, while the training loss simultaneously decreases from 0.195 to 0.017. Thus, according to the local optimization objective, training appears to improve even as the model loses its ability to produce valid long-form reasoning. These dynamics are consistent with a self-reinforcing length-collapse mechanism: shorter and more generic generations can become easier to match under the local semantic critic, which in turn increases their compatibility weight and further reinforces the shortened policy. Without a reference constraint on the surrounding trajectory, sparse chunk-level supervision alone provides no direct pressure to preserve the original model's long-form reasoning structure.

Downstream manifestation of the collapse. The final checkpoint confirms that this behavior is not merely a reduction in reasoning verbosity. The no-KL model obtains only 9.27% on MATH-500, 8.13% on AIME-2024, 6.25% on AIME-2025, 12.03% on AMC-23, and 5.71% on OlympiadBench. Across these evaluations, only approximately 12–14% of generations can be successfully parsed, and average output lengths fall to roughly 80–330 tokens. Qualitative inspection similarly shows frequent premature termination and incomplete reasoning rather than concise but valid solutions. The catastrophic accuracy degradation is therefore primarily associated with loss of valid reasoning structure and answer formatting.

Role of the trust-region anchor. These diagnostics clarify that the KL term serves a more specific role than generic regularization. On un-audited tokens, OMNIOPD minimizes $D_{\mathrm{KL}}(\pi_{\mathrm{ref}} \,\|\, \pi_\theta)$, which penalizes the student when probability mass assigned by the reference model is removed by the updated policy. This discourages premature EOS concentration, excessive entropy reduction, and loss of long-form reasoning behavior between audited chunks. Empirically, a small coefficient $\beta = 0.1$ is sufficient: with KL anchoring, the reference divergence remains approximately 0.005–0.01 while the model can still update substantially at teacher-audited regions.

Overall, the no-KL failure is better characterized as a length- and entropy-collapse feedback loop than as generic parameter drift. The local semantic objective can continue to improve during this collapse, making the reference-policy constraint essential for preserving global trajectory structure while OMNIOPD performs sparse, localized on-policy updates.

E.8 CASE STUDIES: QUALITATIVE ANALYSIS OF ENTROPY-SELECTED CHUNKS

This section presents how OMNIOPD isolates critical reasoning junctures. Rather than auditing the entire trajectory, our peak-entropy scheduling concentrates the teacher's verification budget on "forks in the road" where the student's policy exhibits high uncertainty.

Pattern Summarization: Why Entropy Works By analyzing the selected chunks across various domains, we identify three recurring patterns that justify the effectiveness of entropy-based selection:

- **Strategic Transitions:** Spikes in entropy frequently occur when the model finishes a setup phase and must choose a specific theorem or tool (e.g., deciding between the *Law of Cosines* vs. *Area Formulas* in Geometry).
- **Boundary and Constraint Validation:** In algebraic tasks, uncertainty peaks at logical branching points, such as defining cases for absolute values or checking if a solution $x = 1$ satisfies the initial domain constraints.
- **Combinatorial State Tracking:** In multi-step counting, entropy rises when the model must keep track of previously assigned states (e.g., vertex colors), representing a "memory bottleneck" where teacher correction is most vital.

In the following examples, we show a subset of the 10 selected chunks for three problems. The symbol —— marks the start of the verification chunk.

Case 1: Geometry — Angle Bisector *Problem: Find AD in* $\triangle ABC$ *where* $\sin A = 3/5$*,* $AB = 6$*,* $AC = 8$*.*

- **Fork #2 (Pos: 212, Entropy: 1.16):** `... But I don't know BC yet. Maybe I can find BC using the Law of Cosines? But —— wait, I know sin A = 3/5. Since angle A is acute, cos A should be positive...`
- **Fork #7 (Pos: 915, Entropy: 1.21):** `...called the formula for the length of an angle bisector: AD^2 = AB * AC - BD * DC. —— Wait, let me check that. Let me recall Stewart's theorem. Stewart's theorem relates...`
- **Analysis:** The model exhibits high uncertainty exactly when it needs to verify its own memory of geometric theorems (Stewart's vs. Bisector theorem), ensuring the teacher corrects any "hallucinated" formulas.

Case 2: Combinatorics — Hexagon Coloring *Problem: Ways to color hexagon vertices with 3 colors (no adjacent same).*

- **Fork #2 (Pos: 514, Entropy: 1.69):** `...Wait, maybe not. Alternatively, using recurrence relations. But maybe I can —— just trust the formula here. For n=4, the formula gives 18 for k=3. Let me check...`

Table 11: Sensitivity analysis of OMNIOPD verification hyperparameters: number of audited chunks (M), Monte Carlo rollouts per chunk (N), and chunk size (C).

M	N	C	MATH	AIME-24	AIME-25	AMC-23	Olympiad	Avg
10	10	50	72.60	68.67	60.42	93.44	50.27	69.08
5	10	50	72.76	69.58	55.63	93.44	50.48	68.38
10	10	25	52.73	10.42	4.79	32.19	22.26	24.48
10	20	50	73.35	68.54	61.04	93.13	51.81	69.57
20	10	50	75.01	70.63	**62.71**	90.94	55.50	70.96
10	10	100	**75.94**	**72.29**	62.50	**93.59**	53.58	**71.58**
10	20	100	75.00	72.08	61.88	93.13	**55.72**	71.56

- **Fork #10 (Pos: 1383, Entropy: 1.50):** `...vertex 4 is not adjacent to vertex 2. So —— if vertex 4 is B, is that okay? Since vertex 4 is adjacent to vertex 3 (A) and vertex 1...`
- **Analysis:** Entropy selection captures the model's struggle with cyclic constraints. It identifies the "wrap-around" check (vertex 6 vs vertex 1) as the highest-entropy region, which is the most common point of failure for smaller models.

Case 3: Algebra — Absolute Value Equations *Problem: Solve* $|x^2 - 3x + 2| = x - 1$.

- **Fork #3 (Pos: 395, Entropy: 1.56):** `... x^2 - 4x + 3 = 0. —— Factorizing this quadratic equation: Looking for two numbers that multiply to 3 and add to -4...`
- **Fork #6 (Pos: 837, Entropy: 1.64):** `...absolute value is A when A is non-negative. Therefore, —— the solutions from Case 1 must satisfy A >= 0. So, let's check for x = 1 and x = 3...`
- **Fork #10 (Pos: 1381, Entropy: 1.58):** `...x = 1 is a solution to both equations. However, —— we need to check if x = 1 is a solution to the original equation because...`
- **Analysis:** The selection focuses heavily on the *verification* of solutions. While the calculation of $x^2 - 4x + 3$ is relatively low-entropy, the *logical validity* check (testing if the results fit Case 1 or Case 2) triggers a spike, allowing OMNIOPD to enforce rigorous case-split logic.

E.9 SENSITIVITY TO VERIFICATION HYPERPARAMETERS

We provide the full sensitivity results for the verification hyperparameters (M, N, C) in Table 11. The experiments use the Qwen3-4B student and Qwen3-32B teacher, with all remaining hyperparameters fixed to their default values. The results show strong sensitivity to chunk length C, but substantially weaker sensitivity to the number of teacher rollouts N and audited chunks M. In particular, increasing N beyond 10 provides limited additional benefit, whereas sufficiently long chunks are necessary for reliable chunk-level supervision.

E.10 TEACHER INFERENCE COST ANALYSIS

The primary additional cost of OMNIOPD arises from teacher inference, i.e., the number of tokens that the teacher must process and generate for each student trajectory. This cost applies to both self-hosted teachers, where it determines inference compute and wall-clock time, and hosted teachers, where it translates into API usage. Because the monetary cost of hosted models depends on provider-specific pricing, caching, and batching policies, we report provider-agnostic input/output token requirements and normalized inference effort rather than absolute monetary expenditure.

Token complexity. Let P denote the task-prompt length and L the length of the student-generated trajectory. For each trajectory, OMNIOPD audits M chunks, samples N teacher continuations from each audited chunk, and generates at most C teacher tokens per continuation. The teacher output-token volume is therefore

$$T_{\text{out}} = MNC. \tag{20}$$

Thus, for fixed (M, N, C), the teacher decoding budget is independent of the total student trajectory length L and is directly controlled by the verification hyperparameters.

The teacher input-token volume additionally depends on the length of the student prefix supplied at each audited fork. Assuming that the M audited chunks are approximately distributed across the trajectory, the expected prefix length is approximately $L/2$. When the N continuations sampled from the same fork share the corresponding prefix computation through batching or prefix reuse, the expected teacher input volume is

$$T_{\text{in}}^{\text{reuse}} \approx M\left(P + \frac{L}{2}\right). \tag{21}$$

In contrast, if the same prefix must be processed independently for every Monte Carlo continuation, the corresponding input volume becomes

$$T_{\text{in}}^{\text{no-reuse}} \approx MN\left(P + \frac{L}{2}\right). \tag{22}$$

Practical deployments can lie between these two regimes depending on the inference backend and its support for batching or prefix caching. This decomposition makes the long-CoT scaling explicit: teacher decoding scales as $\mathcal{O}(MNC)$, whereas prefix processing scales linearly with the student trajectory length, between $\mathcal{O}(ML)$ and $\mathcal{O}(MNL)$ depending on prefix reuse.

Comparison with offline SFT. The relevant baseline for long-CoT distillation is teacher-generated long-form SFT on the same task distribution. In our mathematical reasoning setup, constructing an SFT example requires the teacher to generate a complete reasoning trajectory, with approximately $8{,}000$ output tokens per sample. Its teacher-inference cost therefore grows primarily through full-trajectory decoding as reasoning traces become longer. In contrast, OMNIOPD reuses the student trajectory as context and restricts teacher generation to short continuations around selected reasoning forks. Consequently, OMNIOPD trades additional prefix processing for a decoding budget that can be controlled explicitly through M, N, and C.

To compare these two forms of inference on a common scale, we separately account for teacher input and output tokens. Let r denote the effective per-token cost ratio between teacher decoding and prefix processing. For a configuration with teacher input volume T_{in} and output volume T_{out}, we define the normalized inference effort relative to SFT as

$$R_{\text{cost}} = \frac{T_{\text{in}} + rT_{\text{out}}}{P_{\text{SFT}} + rT_{\text{out}}^{\text{SFT}}}, \tag{23}$$

where P_{SFT} and $T_{\text{out}}^{\text{SFT}}$ denote the teacher input and output volumes required to construct an SFT trajectory. Table 12 instantiates this comparison under the prefix-reuse regime in Eq. (21), using $P = 200$, $L = 6{,}340$, and a representative decode-to-prefill weighting of $r = 8.33$. The resulting values should therefore be interpreted as provider-agnostic relative inference-effort estimates under this deployment regime rather than provider-specific monetary costs. Without prefix reuse, teacher input instead follows Eq. (22) and consequently scales with N.

Cost–performance trade-off. The resulting inference cost is directly tunable through the verification hyperparameters. Reducing the number of audited chunks M lowers both prefix processing and teacher decoding, while reducing N or C directly decreases the teacher generation budget. Under the prefix-reuse regime reported in Table 12, the default configuration ($M = 10$, $N = 10$, $C = 50$) requires approximately $1.13\times$ the normalized inference effort of long-CoT SFT. Reducing the audited-chunk budget to $M = 5$ lowers this estimate to approximately $0.56\times$, while retaining 68.38% average accuracy compared with 69.08% for the default configuration. Conversely, increasing the verification budget can improve downstream performance at greater teacher-inference cost. Thus, OMNIOPD exposes an explicit cost–performance trade-off through (M, N, C) rather than requiring a fixed teacher-inference budget.

Table 12: Teacher-inference requirements across OMNIOPD configurations. We report teacher input and output token volumes separately because their effective inference cost depends on the deployment backend, batching strategy, and support for prefix reuse. Teacher input in this table assumes prefix reuse across the N continuations sampled from the same audited fork. Relative cost is normalized against long-CoT SFT using $R_{\text{cost}} = (T_{\text{in}} + rT_{\text{out}})/(P_{\text{SFT}} + rT_{\text{out}}^{\text{SFT}})$ with a representative decode-to-prefill weighting of $r = 8.33$. These values should therefore be interpreted as provider-agnostic relative inference estimates rather than absolute monetary costs.

Method	**Hyperparameters**	**Teacher Input (Reuse)**	**Teacher Output**	**Relative Cost**
Standard SFT	—	200	8,000	1.00×
OMNIOPD	$M = 5, N = 10, C = 50$	16,850	2,500	0.56×
	$M = 10, N = 10, C = 25$	33,700	2,500	0.82×
	$M = 10, N = 10, C = 50$	33,700	5,000	1.13×
	$M = 10, N = 10, C = 100$	33,700	10,000	1.75×
	$M = 10, N = 20, C = 50$	33,700	10,000	1.75×
	$M = 20, N = 10, C = 50$	67,400	10,000	2.25×
	$M = 10, N = 20, C = 100$	33,700	20,000	3.00×

For experiments using hosted teachers, we use the same 50-step optimization budget for both the corresponding Seq-KD baseline and OMNIOPD to control teacher exposure across methods. This protocol is specific to our hosted-teacher evaluation and is not an algorithmic constraint of OMNIOPD. As shown in Appendix E.3, the hosted-teacher optimization curves approach their loss plateau within this training window.