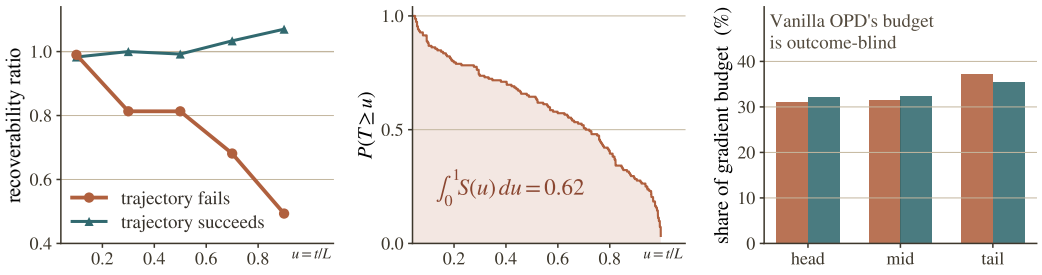
MATCHING WHILE ALIVE: A SURVIVAL VIEW OF ON-POLICY TRAJECTORY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) brings teacher supervision to student-generated prefixes, but its continued availability does not ensure that it remains equally relevant along each trajectory. Our analysis reveals differences in teacher recoverability across trajectories, while vanilla OPD treats teacher matching as an uninterrupted process along every rollout. We therefore formulate on-policy trajectory distillation as a **survival problem** with a **latent lifetime** for teacher matching. Marginalizing over the latent alive state converts the truncated matching objective into a survival-weighted formulation. This connects trajectory-level credit assignment to hazard modeling, with vanilla OPD recovered as the zero-hazard case. This framework admits a simple realization (**Cox-OPD**) based on a Cox proportional hazards model that combines a baseline hazard over prefix depth with observable student-teacher deviation. The accumulated hazard defines a trajectory-specific survival profile, allowing prefix history to shape how matching credit evolves along the rollout. Empirically, these weights reflect variation in independently measured teacher recoverability, and experiments on mathematical reasoning and code generation show gains in both single- and multi-teacher settings. Across all six single-teacher pairs, this survival-based extension improves average performance over vanilla OPD by 1.43 points on average, with 4B students gaining 2.13 points on average across four math benchmarks and both teacher settings. Ablations further highlight how the survival formulation structures the trade-off between credit horizon and trajectory separation.



(a) Recoverability by final outcome (b) Kaplan-Meier survival curve (c) Gradient budget by position

Figure 1: **The effective lifetime of teacher supervision varies across trajectories, while vanilla on-policy distillation (OPD) does not adapt its credit allocation accordingly. (a)** At relative depth $u = t/L$, let $R^S(u)$ and $R^T(u)$ denote the teacher's continuation accuracy from student and teacher prefixes of the same prompt, matched at the same relative depth within their respective rollouts. The ratio $R^S(u)/R^T(u)$ separates sharply by final outcome, reaching 0.49 on failing trajectories and 1.07 on successful ones near the end. **(b)** For each failing rollout that reaches $R^S(u) \leq 0.25$, let T denote the first relative depth at which $R^S(u)$ falls to 0.25 or below. The Kaplan-Meier curve $S(u) = \Pr(T \geq u)$ summarizes this subset, with mean normalized matching lifetime $\int_0^1 S(u)\,du = 0.62$, indicating that low teacher recoverability emerges with a substantial suffix still remaining. **(c)** Using $\sum_t |\ell_t|$ as a gradient-budget proxy, with ℓ_t the sampled-token student–teacher log-ratio, vanilla OPD's positional budget shares differ by at most 1.8 points across outcomes, showing that its credit allocation remains nearly outcome-blind.

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.

1 INTRODUCTION

Each token generated by an autoregressive model shapes the context for the next prediction and can influence the trajectory well beyond the current step (Ross et al., 2011). On-policy distillation (OPD) brings this dependence into training by querying the teacher along the student's own rollouts. Teacher supervision is therefore evaluated on student-generated prefixes, reducing the mismatch between training and generation (Lin et al., 2020, Agarwal et al., 2024, Gu et al., 2024). Although the teacher distribution remains available at every visited prefix, its value for matching may decline as deviations accumulate along the rollout (Li et al., 2026, Liu et al., 2026, Xie et al., 2026).

Vanilla OPD overlooks this distinction and treats every visited prefix as part of one uninterrupted matching process. We validate this intuition through recoverability, the teacher's continuation accuracy $R^S(u)$ from a student prefix at relative depth u. After normalization by accuracy $R^T(u)$ from matched teacher prefixes, recoverability stays stable on successful trajectories but declines on failing ones as generation proceeds (Figure 1a). Among failing rollouts that reach low recoverability, the Kaplan-Meier curve[1] (Kaplan & Meier, 1958) of first-crossing times shows that this can occur with a substantial suffix remaining (Figure 1b). Under vanilla OPD, a proxy for credit allocation shows nearly identical positional shares across outcome groups (Figure 1c). These observations suggest a trajectory-dependent lifetime of teacher matching that vanilla OPD does not explicitly model.

Recoverability provides an empirical diagnostic of the matching horizon, but measuring it throughout training requires additional teacher continuations. We therefore formulate OPD as a *survival problem* with a *latent lifetime* for teacher matching along each student-generated trajectory. We specify the lifetime distribution through per-step hazards that quantify the risk of matching ending at each step (Andersen et al., 1993). Marginalizing this lifetime weights each prefix's matching loss by its modeled survival probability, with accumulated hazard carrying earlier evidence into later credit assignment. We instantiate this framework with a Cox proportional hazards model (Cox, 1972) that combines a depth-dependent baseline with local student–teacher deviation from cached rollout scores. Across math and code, our survival-based extension of OPD consistently improves overall average performance over baselines. Our contributions are:

1. We formulate on-policy trajectory distillation as a latent survival problem where hazards define a matching lifetime distribution for each rollout, while zero hazard recovers vanilla OPD (§3).
2. Our Cox weights separate depth-dependent baseline and deviation-dependent risks, with a length-independent bound on each deviation's cumulative influence on suffix log-weights (§4).
3. We gain an average of 1.43 points over OPD across six single-teacher pairs and improve both multi-teacher domains at all three student scales. Teacher-continuation evaluations show that survival weights track recoverability beyond position alone, while analyses and ablations reveal self-annealing and the trade-off between trajectory separation and credit horizon (§§4–5).

2 PRELIMINARIES

Let x be a prompt and $y = (y_1, \ldots, y_L)$ a response generated autoregressively, with $s_t = (x, y_{<t})$ the context at position t. We write π_θ for the trainable student, π_T for the fixed teacher, and π_0 for the frozen snapshot of the student that generates the rollouts of the current training iteration ($\pi_0 \leftarrow \pi_\theta$). On-policy distillation (OPD) samples $y \sim \pi_0(\cdot \mid x)$ and minimizes the token-level reverse KL to the teacher on these rollouts,

$$\min_\theta \;\; \mathbb{E}_{y\sim\pi_0(\cdot|x)}\Big[\sum_{t=1}^{L} D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid s_t) \,\big\|\, \pi_T(\cdot \mid s_t)\big)\Big]. \tag{1}$$

Given a sampled rollout y, the score at position t is the log-ratio

$$\ell_t \;=\; \log \pi_0(y_t \mid s_t) - \log \pi_T(y_t \mid s_t), \tag{2}$$

whose expectation under $y_t \sim \pi_0(\cdot \mid s_t)$ is the token-level reverse KL, and θ is updated with the standard clipped policy-gradient surrogate on the ratio $\rho_t(\theta) = \pi_\theta(y_t \mid s_t)/\pi_0(y_t \mid s_t)$,

$$\mathcal{L}(\theta) \;=\; -\,\mathbb{E}\Big[\sum_{t=1}^{L} \min\big(\rho_t(\theta)\,\hat{A}_t,\; \mathrm{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\,\hat{A}_t\big)\Big], \tag{3}$$

[1]The Kaplan–Meier curve is a nonparametric estimate of the survival function, which here gives the probability that a rollout has not yet reached low recoverability.

with the unweighted advantage $\hat{A}_t = -\ell_t$ for vanilla OPD. Both log-probabilities in equation 2 are computed once per rollout batch and cached. However, equation 1 implicitly treats teacher matching as surviving the full rollout, assigning unit weight to every visited context regardless of accumulated deviation. This motivates viewing trajectory distillation through a survival lens, with a trajectory-dependent lifetime for teacher matching.

3 TRAJECTORY DISTILLATION AS A SURVIVAL PROBLEM

On-policy training observes the trajectories at risk We model matching failure as a latent event, using teacher-continuation recoverability as an empirical diagnostic. Figure 1a compares teacher continuation accuracy after student and teacher prefixes of the same prompt, matched at the same relative depth. On-policy training directly observes the current student's trajectories at risk of this failure. Each sampled trajectory is observed through its realized length, while the student and teacher score every sampled token from the same prompt and evolving context. OPD therefore records the complete sequence of local evidence along each realized rollout, from which a latent matching lifetime can be modeled.

A latent matching lifetime Given the prompt x and for $t \geq 1$, let $\pi_0^{<t}$ and $\pi_T^{<t}$ denote the distributions over $y_{<t}$ induced by π_0 and π_T, respectively. For any divergence D satisfying the data-processing inequality, $D(\pi_0^{<t} \,\|\, \pi_T^{<t})$ is non-decreasing with t. Prefix length thus provides a natural discrete clock, but this population-level separation does not identify a failure event along an individual trajectory. We therefore treat the matching lifetime as latent and model its per-step hazards along each realized trajectory.

Definition 1 (Lifetime model)**.** A *lifetime model* defines a discrete-time distribution for the latent matching lifetime τ along each realized trajectory y of length L. Matching remains alive at position t when $\tau \geq t$. Its per-step hazards $h_k = h_k(L, x, y_{\leq k}) \in [0, 1]$ are evaluated in generation order, with survival function

$$w_t \;=\; \Pr(\tau \geq t \mid y) \;=\; \prod_{k<t}(1 - h_k), \qquad w_1 = 1. \tag{4}$$

For a fixed realized length L, w_t depends only on the hazards accumulated before position t.

Distilling while alive The latent lifetime τ marks the point beyond which the evolving prefix is modeled as insufficiently teacher-relevant for full-strength distillation. Although the sampled trajectory continues, the survival view lets it contribute according to how long it remains alive under the chosen lifetime model, replacing equation 1 by the truncated objective

$$\min_{\theta} \;\; \mathbb{E}_{y\sim\pi_0(\cdot|x)}\, \mathbb{E}_{\tau}\Big[\sum_{t\leq\tau} D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid s_t) \,\big\|\, \pi_T(\cdot \mid s_t)\big)\Big]. \tag{5}$$

Survival weighting retains student error states while scaling their credit by modeled survival, whose relation to recoverability is evaluated in Figure 2.

Proposition 1 (Truncation identity)**.** *For a realized trajectory y of length L, let $(a_t)_{t=1}^{L}$ be any per-token scores determined by y. Then every lifetime model satisfies $\mathbb{E}_\tau\big[\sum_{t\leq\tau} a_t \,\big|\, y\big] = \sum_{t=1}^{L} w_t\, a_t$.*

Proof. Apply $\mathbb{E}_\tau[\cdot \mid y]$ to $\sum_{t\leq\tau} a_t = \sum_{t=1}^{L} \mathbf{1}\{\tau \geq t\}\, a_t$ and use equation 4. □

Proposition 1 exactly converts the truncated KL objective in equation 5 into a survival-weighted KL objective. For training, we optimize the sampled clipped surrogate in equation 3, with advantages

$$\hat{A}_t \;=\; -\,\mathrm{sg}[w_t]\;\ell_t, \tag{6}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient and w_t is the survival probability induced by the chosen lifetime model. Because $w_t = \mathbb{E}[\mathbf{1}\{\tau \geq t\} \mid y]$, the survival-weighted sum is the Rao-Blackwellization of its stochastically truncated counterpart. It therefore preserves the truncated objective in expectation while removing the conditional variance due to sampling τ. Appendix A.2 gives the corresponding variance decomposition, establishing survival weighting as a variance-reduced implementation of the same truncated objective.

Importantly, despite its multiplicative form, w_t plays a different role from an importance ratio. Sampling remains under $\pi_0(\cdot \mid x)$, while $w_t = \mathbb{E}[\mathbf{1}\{\tau \geq t\} \mid y]$ marginalizes the latent alive state. It therefore assigns credit along each trajectory according to the modeled survival of teacher matching rather than implementing a change of measure.

Corollary 1 (Worst-case credit discrepancy)**.** *Fix a lifetime model. Let $\mathcal{Y} = \sigma(x, y)$, $Z_t = \mathbf{1}\{\tau \geq t\}$, and $w_t = \mathbb{E}[Z_t \mid \mathcal{Y}]$. For deterministic bounds $M_t \geq 0$ satisfying $\mathbb{E}_y[\sum_{t=1}^{L} M_t] < \infty$, every $\mathcal{Y}$-measurable credit sequence $q_t \in [0, 1]$ satisfies*

$$\sup_{|a_t| \leq M_t} \left| \mathbb{E}_{y,\tau} \left[\sum_{t=1}^{L} (q_t - Z_t) a_t \right] \right| = \mathbb{E}_y \left[\sum_{t=1}^{L} M_t \, |q_t - w_t| \right], \tag{7}$$

where the supremum ranges over all real-valued, $\mathcal{Y}$-measurable token-score sequences satisfying the bounds almost surely.

Survival weighting gives zero worst-case discrepancy under the chosen lifetime model, whose ability to capture teacher relevance depends on the hazard design. For vanilla OPD, this discrepancy equals $\mathbb{E}_y[\sum_{t=1}^{L} M_t(1 - w_t)]$, the expected sum of token-score bounds beyond the modeled lifetime. The proof is given in Appendix A.1.

Remark 1 (Hazard timing in credit assignment). For a fixed realized length, the hazard h_t first appears in $w_{t+1} = w_t(1 - h_t)$ rather than w_t, so it leaves the credit at position t unchanged and discounts only later positions.

Credit assignment as hazard modeling Vanilla OPD is recovered as the degenerate hazard model $h \equiv 0$, for which $w_t \equiv 1$ and every context receives uniform credit. More generally, any positive, non-increasing credit sequence $w_t = w_t(L, x, y_{<t})$ with $w_1 = 1$ has a hazard representation $w_t = \prod_{k<t}(1 - h_k)$ with $h_k = 1 - w_{k+1}/w_k$ for $k < L$. Such credit sequences can therefore be specified through per-step hazards. The hazards in turn define equation 5, under which each trajectory is distilled for as long as it stays alive. Section 4 instantiates this view with the widely used Cox proportional-hazards model, which separates a baseline risk profile from the relative risk induced by student-teacher deviation.

4 COX-OPD: SURVIVAL WEIGHTING WITH PROPORTIONAL HAZARDS

The observable deviation statistic Building on this survival view, we first identify an observable signal of local student-teacher deviation. The sampled-token log-ratio ℓ_t in equation 2, already available during trajectory generation, yields the bounded statistic

$$\delta_t \;=\; (1 - e^{-\ell_t})^+ \;=\; \left(1 - \frac{\pi_T(y_t \mid s_t)}{\pi_0(y_t \mid s_t)}\right)^+ \in [0, 1], \qquad (z)^+ = \max(z, 0), \tag{8}$$

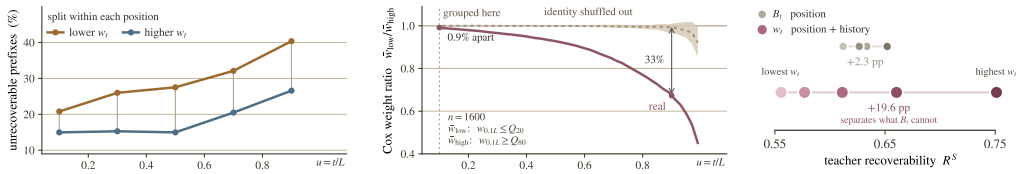
which measures the fraction of the student's probability mass at the sampled token that the teacher does not cover. It is zero when the teacher covers the student's probability at that token ($\ell_t \leq 0$), positive when the student assigns it more mass than the teacher, and saturates at one when the teacher assigns it zero probability, yielding a bounded covariate with a direct probabilistic interpretation.[2]

The speculative-decoding identity motivates the product-limit weight $w_t^{\mathrm{raw}} = \prod_{k<t}(1 - \delta_k)$, which, conditional on the realized trajectory, gives the probability that all preceding tokens pass teacher verification. Its log-weights are sensitive to near-certain rejection. For $w_t^{\mathrm{raw}} > 0$ and $k < t$,

$$\frac{\partial[-\log w_t^{\mathrm{raw}}]}{\partial \delta_k} = \frac{1}{1 - \delta_k}. \tag{9}$$

The derivative diverges as $\delta_k \to 1$, and a single rejection score close to one can suppress credit throughout the suffix. Trajectory distillation should preserve the monotonic risk ordering encoded by this rejection signal while controlling its influence on subsequent log-weights. We therefore use the soft rejection score δ_t as a bounded covariate in a Cox proportional-hazards model.

[2] δ_t is exactly the probability that the teacher would reject y_t under speculative-decoding verification, and its conditional mean is the token-level total variation, $\mathbb{E}_{y_t \sim \pi_0}[\delta_t \mid s_t] = \mathrm{TV}\left(\pi_0(\cdot \mid s_t), \pi_T(\cdot \mid s_t)\right)$. We treat these identities as a natural instantiation of the framework and defer their formal statement to the appendix.



(a) Teacher unrecoverability by w_t (b) Trajectory-specific separation (c) Hazard accumulation

Figure 2: **Modeled survival tracks teacher recoverability beyond position alone. (a)** Within each position bin, prefixes in the lower half of w_t are more often unrecoverable than those in the upper half. **(b)** For trajectories in the bottom and top quintiles of $w_{0.1L}$, the mean-weight ratio $\bar{w}_{\text{low}}/\bar{w}_{\text{high}}$ falls from ≈ 0.99 at $u = 0.1$ to ≈ 0.67 at $u = 0.9$. Shuffling δ across rollouts within position bins keeps the ratio near one over most depths. **(c)** After grouping tokens by t/L decile and $|\ell_t|$ quintile, recoverability stays flat when ranked by B_t, whereas ranking by w_t isolates a recoverability gradient from 56% to 75%.

Its per-token hazard rate is a baseline $\lambda_0(t)$ scaled by a relative risk $e^{\beta\delta_t}$,

$$\lambda_t \;=\; \lambda_0(t)\, e^{\beta\delta_t}. \tag{10}$$

The corresponding discrete-time hazard is $h_t^{\text{Cox}} = 1 - e^{-\lambda_t}$, which instantiates the lifetime model in Definition 1. For $\beta \geq 0$, the relative risk $e^{\beta\delta_t}$ scales the hazard rate according to local deviation, while $\lambda_0(t)$ specifies the rate at zero deviation. A zero value of δ_t only indicates that the teacher assigns at least as much probability to the sampled token as the student. The teacher may still struggle to produce a successful continuation from the resulting prefix. The baseline therefore encodes depth-dependent prior risk not fully captured by this one-sided local statistic.

We choose the baseline survival $S_0(t) = 1 - (t/L)^p$, with $p \geq 1$, so that $S_0(0) = 1$ and $S_0(L) = 0$. Its cumulative hazard is $\Lambda_0(t) = -\log S_0(t)$, with per-step increments $\lambda_0(t) = \Lambda_0(t) - \Lambda_0(t-1)$ for $1 \leq t < L$. Accumulating λ_k over positions $k < t$ gives the cumulative hazard H_t and the survival weight w_t used to scale the advantage in equation 6.

$$H_t \;=\; \sum_{k<t} \lambda_0(k)\, e^{\beta\delta_k}, \qquad w_t \;=\; e^{-H_t} \in (0, 1]. \tag{11}$$

Proposition 2 (Controlled adaptive survival)**.** *Let $\beta \geq 0$ and define the prefix-aligned baseline $B_t = \exp(-\sum_{k<t} \lambda_0(k)) = S_0(t-1)$. For every $k < t$,*

$$0 \;\leq\; \frac{\partial[-\log w_t]}{\partial \delta_k} = \beta\lambda_0(k)e^{\beta\delta_k} \;\leq\; \beta e^{\beta}\lambda_0(k), \tag{12}$$

so that, uniformly in L and $p \geq 1$,

$$\sum_{t>k} \frac{\partial[-\log w_t]}{\partial \delta_k} \;\leq\; \beta e^{\beta}, \tag{13}$$

and the survival weight satisfies

$$B_t^{e^{\beta}} \;\leq\; w_t \;\leq\; B_t. \tag{14}$$

At fixed L, β, p, pointwise larger deviations yield no larger survival weights or effective credit horizon $L_{\text{eff}}(y) = \sum_{t=1}^{L} w_t$. As deviations vanish pointwise, $w_t \to B_t$ and $L_{\text{eff}}(y) \to \sum_{t=1}^{L} B_t$.

Proposition 2, proved in Appendix A.4, characterizes how Cox-OPD allocates optimization credit. The product-limit construction has an exact verification interpretation, while the Cox link bounds the cumulative sensitivity of subsequent log-weights to any single deviation score by βe^{β}, uniformly in trajectory length. The corresponding product-limit sensitivity, $(L-k)/(1-\delta_k)$, grows with the suffix length and diverges as $\delta_k \to 1$.

The slope β controls how strongly the hazard responds to local deviation. The shape p controls how this sensitivity varies along the trajectory. At $p = 1$ the baseline failure mass is uniform over normalized length, while $p > 1$ pushes it toward the tail. The smaller early baseline hazard reduces the sensitivity in equation 12, helping preserve credit near the start of the trajectory. At fixed L, β, p, pointwise reductions in deviation cannot shorten the effective credit horizon. This self-annealing reduces the additional downweighting due to deviation and recovers the baseline profile B_t as deviations vanish, preserving the survival prior. Figure 3c illustrates this behavior during training. Algorithm 1 summarizes one training iteration.

Algorithm 1 Cox-OPD (one training iteration)

Require: student π_θ, teacher π_T, prompt batch, slope $\beta \geq 0$, shape $p \geq 1$

$\pi_0 \leftarrow \pi_\theta$ ▷ freeze rollout snapshot

for each prompt x **do**

  sample $y \sim \pi_0(\cdot \mid x)$ of length L and cache $\log \pi_0(y_t \mid s_t)$, $\log \pi_T(y_t \mid s_t)$

  $\ell_t \leftarrow \log \pi_0(y_t \mid s_t) - \log \pi_T(y_t \mid s_t)$ ▷ log-ratio equation 2

  $\delta_t \leftarrow (1 - e^{-\ell_t})^+$ ▷ deviation equation 8

  $\lambda_0(t) \leftarrow \Lambda_0(t) - \Lambda_0(t-1)$ for $t < L$, with $\Lambda_0(t) = -\log\big(1 - (t/L)^p\big)$ ▷ baseline hazard

  $w_t \leftarrow e^{-H_t}$ with $H_t = \sum_{k<t} \lambda_0(k)\, e^{\beta \delta_k}$ ▷ survival weight equation 11

  $\hat{A}_t \leftarrow -\operatorname{sg}[w_t]\, \ell_t$ ▷ weighted advantage equation 6, vanilla OPD uses $-\ell_t$

end for

update θ on the clipped objective equation 3 with advantages $\hat{A}_t$

An empirical check on the modeled lifetime Definition 1 specifies a latent lifetime that is never observed, and Figure 1a motivates this view by comparing recoverability across outcome groups. We compute survival weights from cached rollout scores and realized lengths without outcome labels. Figure 2 tests whether these weights track independently measured teacher recoverability. Within bins of normalized position, prefixes with lower weights are more often classified as unrecoverable (Figure 2a). Shuffling deviations across trajectories within position bins removes the weight separation between trajectory groups (Figure 2b). Figure 2c shows that higher survival weights remain associated with higher teacher recoverability within groups defined by normalized position and $|\ell_t|$.

5 Experiments

5.1 Experimental setup

Models and data We distill Qwen3 students at three scales (0.6B, 1.7B, and 4B) from either Qwen3-4B-Instruct-2507 or Qwen3-30B-A3B-Instruct-2507, with all models in non-thinking mode (Yang et al., 2025). Training uses 57K DeepMath problems of difficulty 6 or above and 25K Eurus-RL-Code problems (He et al., 2025, Cui et al., 2025).

Training details We follow the G-OPD training setup and build on its `verl` codebase.[3] For single-teacher distillation, we compare Cox-OPD with reproduced OPD and IW-OPD baselines under identical settings. Each update samples one on-policy response for each of 1,024 prompts, with the PPO mini-batch also set to 1,024. All methods are trained for two epochs using a learning rate of 10^{-5}, PPO clipping of 0.2, and maximum prompt and response lengths of 2,048 and 16,384 tokens. Token-level rollout correction uses a threshold of 5.0, while generation uses temperature 1.0 and top-p 1.0. For Cox-OPD, we fix the deviation sensitivity at $\beta = 3$ and the baseline shape at $p = 2$. Further training details are provided in Appendix B.

Evaluation Math results use Math-Verify mean@32 accuracy on AIME24/25/26 and HMMT25, averaging its February and November sets. Code results use EvalPlus pass@1 ($n = 4$) on HumanEval+ and MBPP+. Multi-teacher evaluation also includes LiveCodeBench (LCB). We use temperature 1.0 and top-p 1.0. Further details are provided in Appendix B.2.

5.2 Main results

Single-teacher distillation Across all six teacher-student pairs, Cox-OPD improves Avg over OPD by up to 1.98 points (Tables 1 and 2). The gains are strongest for the 4B student under both teachers and extend across math and code. Cox-OPD also shows consistent gains over IW-OPD, highlighting the value of modeling cumulative prefix evidence through a latent matching lifetime.

Multi-teacher distillation The multi-teacher experiment replaces a single fixed teacher with two domain experts, routing math and code prompts to Qwen3-4B RL-Math and RL-Code, respectively. Cox-OPD improves over OPD across all three student scales in both domains (Table 3). Distillation enables the 4B student to match the post-RL reference teachers on average in math ($\Delta = +0.14$) and code ($\Delta = +0.07$). Our survival-based extension further raises the math and code averages to 58.65 and 63.51, providing further evidence for the effectiveness of survival-based credit assignment.

[3] `https://github.com/RUCBM/G-OPD`

Table 1: **Single-teacher distillation from Qwen3-30B-A3B-Instruct-2507.** Math results are mean@32 accuracy (%), code results are EvalPlus pass@1 (%) estimated with $n = 4$. Best results are shown in **bold**.

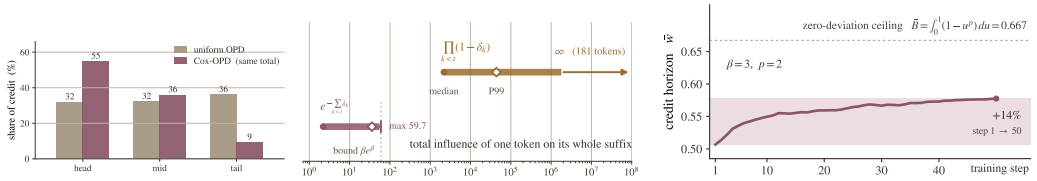
Student	Method	Math				Code		Avg
		AIME24	AIME25	AIME26	HMMT25	HE+	MBPP+	
Teacher Model		75.00	61.56	75.10	49.06	81.86	77.18	69.96
Qwen3-4B	OPD	56.35	46.15	54.06	33.02	79.88	74.07	57.26
	IW-OPD	56.46	49.69	54.48	33.18	78.83	74.37	57.84
	Cox-OPD (ours)	**57.40**	**50.73**	**54.79**	**35.63**	**81.10**	**74.83**	**59.08**
Qwen3-1.7B	OPD	36.88	28.54	26.35	14.58	69.51	58.99	39.14
	IW-OPD	34.27	29.38	26.25	15.26	68.24	58.85	38.71
	Cox-OPD (ours)	**37.19**	**29.69**	**29.38**	**17.08**	**70.10**	**59.99**	**40.57**
Qwen3-0.6B	OPD	9.69	17.50	13.02	8.07	40.85	43.12	22.04
	IW-OPD	9.48	16.15	13.30	8.18	40.67	**43.83**	21.94
	Cox-OPD (ours)	**11.25**	**19.27**	**13.33**	**8.40**	**40.97**	43.54	**22.79**

Table 2: **Single-teacher distillation from Qwen3-4B-Instruct-2507.** Math results are mean@32 accuracy (%), code results are EvalPlus pass@1 (%) estimated with $n = 4$. Best results are shown in **bold**.

Student	Method	Math				Code		Avg
		AIME24	AIME25	AIME26	HMMT25	HE+	MBPP+	
Teacher Model		61.98	47.50	53.23	36.56	78.66	73.15	58.51
Qwen3-4B	OPD	57.71	43.96	50.62	31.97	78.56	70.85	55.61
	IW-OPD	54.37	45.73	51.15	32.84	79.83	71.14	55.84
	Cox-OPD (ours)	**57.81**	**46.88**	**53.44**	**34.22**	**81.23**	**71.93**	**57.59**
Qwen3-1.7B	OPD	35.10	26.87	25.00	13.49	68.24	57.13	37.64
	IW-OPD	35.93	26.67	27.15	14.37	69.51	58.42	38.68
	Cox-OPD (ours)	**36.35**	**27.08**	**28.02**	**15.21**	**69.72**	**59.34**	**39.29**
Qwen3-0.6B	OPD	**12.71**	**18.75**	11.04	7.34	43.56	44.27	22.95
	IW-OPD	11.25	18.54	11.25	8.12	**46.34**	**45.59**	23.52
	Cox-OPD (ours)	12.19	17.81	**13.23**	**8.65**	46.10	45.34	**23.89**

5.3 A CLOSER LOOK AT THE SURVIVAL PRIOR

Credit assignment and training dynamics Beyond downstream accuracy, we analyze how survival weighting redistributes credit and adapts during training. With total credit matched to uniform OPD on each rollout, Cox-OPD allocates 55% of the budget to the head and 9% to the tail (Figure 3a), isolating credit redistribution from changes in total credit. In Figure 3b, Cox-OPD keeps the cumulative sensitivity of subsequent log-weights to each observed deviation below the βe^{β} bound in Proposition 2, whereas the median sensitivity under the raw product-limit rule exceeds this threshold. Figure 3c illustrates self-annealing under fixed β and p, with survival weighting retaining more credit on average as training progresses.



(a) Credit share by region (b) Bounded per-token influence (c) Training-time self-annealing

Figure 3: **(a)** The credit proxy $w_t|\ell_t|$ is rescaled per rollout to match uniform OPD's total budget. Grey bars pool Figure 1c over outcomes. **(b)** Cox suffix influence stays below βe^{β}, whereas the product-limit median exceeds it. Dots and diamonds mark medians and 99th percentiles, respectively. **(c)** With $\beta = 3$, $p = 2$ fixed (continuous-depth zero-deviation baseline $\bar{B} = 2/3$), the mean normalized credit horizon $\bar{w} = \mathbb{E}_y[L_{\text{eff}}(y)/L(y)]$ grows by 14% over 50 steps, consistent with self-annealing.

Table 3: **Multi-teacher distillation from Qwen3-4B RL-Math and RL-Code.** Prompts are routed to their domain teachers. Math scores are mean@32 accuracy; code scores are pass@1 ($n = 4$), using EvalPlus for HE+/MBPP+ and the LiveCodeBench protocol for LCB. Best results are in **bold**.

Student	Method	Math reasoning					Code generation			
		AIME24	AIME25	AIME26	HMMT25	Avg.	HE+	MBPP+	LCB	Avg.
Reference teachers (Qwen3-4B RL variants)		*RL-Math*					*RL-Code*			
4B		68.65	58.54	61.67	40.73	57.40	87.7	72.4	28.29	62.80
4B	Base	23.12	20.21	15.63	9.79	17.19	76.20	64.50	16.86	52.52
	OPD	67.40	**60.63**	60.73	41.41	57.54	87.96	72.42	28.23	62.87
	Cox-OPD (ours)	**69.58**	**60.63**	**62.08**	**42.30**	**58.65**	**89.02**	**73.08**	**28.43**	**63.51**
1.7B	Base	12.29	12.92	8.02	4.90	9.53	60.40	52.80	12.43	41.88
	OPD	44.90	**36.15**	36.67	20.94	34.67	75.61	62.96	22.14	53.57
	Cox-OPD (ours)	**46.67**	35.94	**38.85**	**21.77**	**35.81**	**75.91**	**63.03**	**23.43**	**54.12**
0.6B	Base	1.56	3.33	0.42	1.56	1.72	29.00	28.20	1.14	19.45
	OPD	18.54	25.94	19.27	13.28	19.26	56.10	47.88	17.43	40.47
	Cox-OPD (ours)	**19.27**	**27.19**	**19.27**	**14.64**	**20.09**	**57.93**	**49.67**	**17.71**	**41.77**

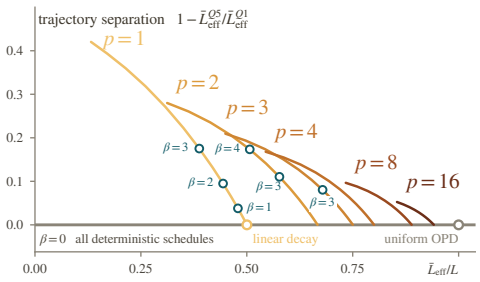
Hazard designs and parameters Following the 4B-student setup in Table 1, we report mean@32 gains over uniform OPD on AIME25/26. Figure 4a shows that Cox-OPD performs best across alternative hazard designs, with the raw product-limit rule reducing accuracy and the simplified variants yielding smaller gains, suggesting that both the shaped baseline and graded deviation contribute to the improvement. Figure 4b shows that the default $(\beta, p) = (3, 2)$ performs best, while several alternatives achieve comparable gains. More broadly, the results slightly favor moderately large values of β, consistent with sufficient sensitivity to deviation without excessive loss of credit horizon. Figure 4c illustrates the trade-off between normalized credit horizon and trajectory separation. At $\beta = 0$, p controls the deterministic horizon, whereas $\beta > 0$ introduces trajectory separation at the cost of some retained credit. These results suggest that effective credit assignment benefits from balancing trajectory separation with retained credit.

Hazard design	Δ_{25}	Δ_{26}
direct product-limit	−2.40	−0.65
clipped product-limit (clip each δ_t at 0.99)	+1.55	+0.24
constant baseline ($\lambda_0 = 1$)	+3.05	+0.48
binary deviation ($\delta_t \to \mathbf{1}\{\delta_t > 0\}$)	+3.18	+0.45
Cox-OPD (ours)	**+4.58**	**+0.73**

(a) Hazard variants

p	β	Δ_{25}	Δ_{26}
1	0	+1.20	+0.10
	1	+2.40	+0.32
	2	+4.12	+0.62
	3	+3.65	+0.51
2	**3**	**+4.58**	**+0.73**
	4	+4.28	+0.67
3	3	+2.98	+0.43
	4	+4.18	+0.64

(b) Parameter ablation



(c) Credit-separation trade-off

Figure 4: **Ablations and credit-assignment trade-off.** **(a)** Mean@32 gains over uniform OPD on AIME25 (Δ_{25}) and AIME26 (Δ_{26}) for alternative hazard designs. **(b)** Parameter sweep including position-only linear decay at $\beta = 0$, $p = 1$. **(c)** Normalized credit horizon versus trajectory separation. Q_1 and Q_5 denote the lowest and highest quintiles of trajectories ranked by mean token deviation $\bar{\delta} = L^{-1} \sum_{t=1}^{L} \delta_t$, respectively.

Toward mitigating cross-domain interference We apply the Cox parameterization to math only, code only, both domains, or neither (vanilla OPD). All variants jointly train a Qwen3-4B student on math and code with the same teacher routing. Figure 5 shows that both single-domain variants generally improve over vanilla OPD, with the Cox parameterization in both domains performing best overall on AIME24 and MBPP+. Full results are reported in Appendix C.1. These results suggest that survival-based credit assignment can improve joint learning from multiple domain experts, with the potential to reduce interference between their supervision signals.

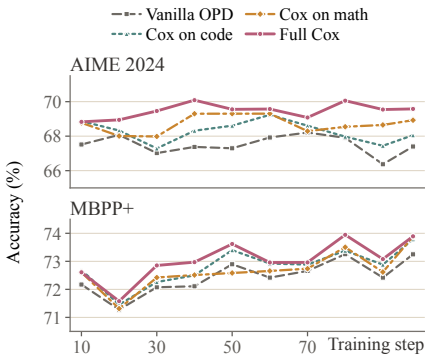


Figure 5: **Multi-teacher training trajectories.**

A bonus where the lifetime is observed. Speculative decoding makes the latent lifetime in Definition 1 observable at the first rejected draft token. Within each block, DFlash (Chen et al., 2026) uses fixed exponential position weights $w_k = \exp(-(k-1)/\gamma)$, corresponding to a constant hazard. Cox-DFlash instead conditions the hazard on target-draft deviation along the realized prefix. Figure 6 shows that Cox-DFlash achieves a higher mean acceptance length than DFlash. This extension motivates future work on applying the survival view to a broader class of trajectory-matching problems.

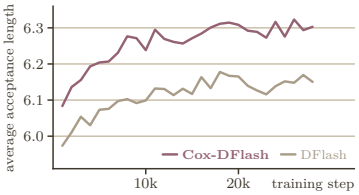


Figure 6: Cox-DFlash improves mean acceptance length over DFlash across five benchmarks. Results for each benchmark are in Appendix C.2.

6 Related Work

Survival analysis and hazard modeling Survival analysis represents event times through survival probabilities and hazards (Andersen et al., 1993), with nonparametric survival and cumulative hazard estimators (Kaplan & Meier, 1958, Nelson, 1972). Hazard models capture covariate effects (Cox, 1972, Aalen, 1989, Andersen & Gill, 1982) and accommodate flexible baseline shapes (Royston & Parmar, 2002). In this view, a latent failure time represents how long teacher matching remains relevant along a rollout, motivating trajectory-level survival modeling through hazard accumulation.

On-policy distillation OPD learns from teacher feedback on student-generated trajectories to reduce training-inference mismatch (Agarwal et al., 2024), following imitation learning on learner-visited states (Ross et al., 2011, Lin et al., 2020). Existing methods introduce teacher intervention in rollouts (Xu et al., 2025) or adapt distillation objectives through divergence design (Wen et al., 2023, Gu et al., 2024, Ko et al., 2024, Jin et al., 2026) and reward formulation (Yang et al., 2026). Other approaches refine supervision through token selection (Xu et al., 2026, Wang et al., 2026) or mismatch correction (Yu et al., 2026). Analysis highlights teacher-student compatibility and reports diminishing gains from teacher continuation at later prefixes in long rollouts (Li et al., 2026). Ding & Zhang (2026) attribute OPD's gains to low-probability-token suppression under noisy supervision.

Credit assignment in trajectory matching Trajectory-matching methods differ in how they aggregate and weight teacher feedback. Sampled token-level reverse-KL assigns credit locally (Lu & Thinking Machines Lab, 2025), whereas MiniLLM uses reward-to-go to optimize sequence-level reverse-KL (Gu et al., 2024). Other methods truncate supervision based on teacher margins (Liu et al., 2026) or apply prefix weighting motivated by a KL-constrained projection (Xie et al., 2026). In speculative decoding, D-PACE (Wu et al., 2026) extends DFlash (Chen et al., 2026) with weights based on a surrogate of expected accepted draft length. We obtain survival weights by marginalizing a latent matching lifetime, connecting trajectory-level credit assignment to hazard modeling.

7 Conclusion

This work aims to understand on-policy trajectory matching through a survival lens by modeling teacher matching along each student rollout as a process with a latent lifetime. The survival formulation provides a principled bridge from latent matching lifetimes to trajectory-level credit assignment by averaging over the corresponding latent trajectory truncations, with vanilla OPD recovered at zero hazard. Within this framework, a proportional-hazards model combines prefix depth with observed student-teacher deviation, giving rise to Cox-OPD. The resulting objective retains this survival interpretation in a deterministic form suitable for training. Across mathematical reasoning and code generation, this survival-based extension consistently improves vanilla OPD across student scales in both single- and multi-teacher settings. Teacher-continuation evaluations show that survival weights reflect variation in prefix recoverability, while ablations demonstrate the advantage of the proposed weighting rule over alternative credit designs.

Limitations and future work The proportional-hazards formulation provides a tractable realization of the survival view, while richer trajectory-dependent hazards could incorporate evidence beyond prefix depth and sampled-token deviation. Future work can test such alternatives across model families and domains. Speculative decoding is a natural extension because rejection exposes the stopping event. More broadly, the survival perspective may extend to other autoregressive trajectory-matching problems with latent or observed stopping events.

AI USE STATEMENT

Generative AI tools assisted with literature search, code implementation, manuscript drafting, and visualization. All data points used in the visualizations were obtained from our training or evaluation runs. The authors verified all cited sources and take full responsibility for the paper.

REFERENCES

Odd O. Aalen. A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8): 907-925, 1989. doi: 10.1002/sim.4780080803.

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

Per Kragh Andersen and Richard D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100-1120, 1982. doi: 10.1214/aos/1176345976.

Per Kragh Andersen, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. doi: 10.1007/978-1-4612-4348-9.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

Jian Chen, Yesheng Liang, and Zhijian Liu. DFlash: Block diffusion for flash speculative decoding, 2026. URL https://arxiv.org/abs/2602.06036.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

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

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187-202, 1972. doi: 10.1111/j.2517-6161.1972.tb00899.x.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards, 2025. URL https://arxiv.org/abs/2502.01456.

Yi Ding and Ruqi Zhang. Does on-policy distillation really distill? from noisy teacher to self-improvement, 2026. URL https://arxiv.org/abs/2608.31046.

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

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. DeepMath-103K: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning, 2025. URL https://arxiv.org/abs/2504.11456.

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, 2021. URL https://arxiv.org/abs/2103.03874.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code, 2024. URL https://arxiv.org/abs/2403.07974.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.07079.

E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457-481, 1958. doi: 10.1080/01621459.1958.10501452.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872-24895. PMLR, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

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

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6121-6133. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.494. URL https://aclanthology.org/2020.emnlp-main.494/.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation, 2023. URL https://arxiv.org/abs/2305.01210.

Kaiyuan Liu, Ziyuan Zhuang, Yang Bai, Bing Wang, Rongxiang Weng, and Jieping Ye. Prefix teach, suffix fade: Local teachability collapse in strong-to-weak on-policy distillation, 2026. URL https://arxiv.org/abs/2605.13643.

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

Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945-966, 1972. doi: 10.1080/00401706.1972.10488991.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627-635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Patrick Royston and Mahesh K. B. Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175-2197, 2002. doi: 10.1002/sim.1203.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation, 2026. URL https://arxiv.org/abs/2605.26844.

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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.605. URL https://aclanthology.org/2023.acl-long.605/.

Tianyu Wu, Yu Yao, Zhenting Qi, Han Zheng, Zhuohan Wang, Haoran Ma, Lawrence Liao, Himabindu Lakkaraju, Ju Li, and Yilun Du. D-PACE: Dynamic position-aware cross-entropy for parallel speculative drafting, 2026. URL https://arxiv.org/abs/2605.18810.

Yan Xie, Sijie Zhu, Tiansheng Wen, Bo Chen, and Yifei Wang. On the position bias of on-policy distillation, 2026. URL https://arxiv.org/abs/2606.22600.

Tengyu Xu, Eryk Helenowski, Karthik Abinav Sankararaman, Di Jin, Kaiyan Peng, Eric Han, Shaoliang Nie, Chen Zhu, Hejia Zhang, Wenxuan Zhou, Zhouhao Zeng, Yun He, Karishma Mandyam, Arya Talabzadeh, Madian Khabsa, Gabriel Cohen, Yuandong Tian, Hao Ma, Sinong Wang, and Han Fang. The perfect blend: Redefining RLHF with mixture of judges, 2024. URL https://arxiv.org/abs/2409.20370.

Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.11325.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation, 2026. URL https://arxiv.org/abs/2604.14084.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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

Zichao Yu, Chengzhi Yu, Shengze Xu, Yujin Han, Bingqing Jiang, Xu Wang, and Difan Zou. Mismatch matters: On-policy distillation beyond token agreement, 2026. URL https://arxiv.org/abs/2608.09836.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

A PROOFS AND ADDITIONAL PROPERTIES

A.1 WORST-CASE CREDIT DISCREPANCY

Proof of Corollary 1. Let $\mathcal{Y} = \sigma(x, y)$ and fix an admissible credit sequence q_t. Define

$$\mathfrak{R}(q) = \sup_{|a_t| \le M_t} \left| \mathbb{E}_{y,\tau} \left[\sum_{t=1}^{L} (q_t - Z_t) a_t \right] \right|. \tag{15}$$

The supremum ranges over the $\mathcal{Y}$-measurable score sequences in the corollary. Since $w_t = \mathbb{E}[Z_t \mid \mathcal{Y}]$, the tower property and the triangle inequality give

$$\begin{aligned} \left| \mathbb{E}_{y,\tau} \left[\sum_{t=1}^{L} (q_t - Z_t) a_t \right] \right| &= \left| \mathbb{E}_{y} \left[\sum_{t=1}^{L} (q_t - w_t) a_t \right] \right| \\ &\le \mathbb{E}_{y} \left[\sum_{t=1}^{L} M_t |q_t - w_t| \right]. \end{aligned} \tag{16}$$

The admissible choice $a_t = M_t \operatorname{sign}(q_t - w_t)$ attains this bound, proving equation 7. Setting $q_t = w_t$ gives zero discrepancy, while $q_t = 1$ gives $\mathbb{E}_y[\sum_{t=1}^{L} M_t(1 - w_t)]$. □

A.2 RAO-BLACKWELLIZATION OF STOCHASTIC TRUNCATION

Let a_t be $\mathcal{Y}$-measurable token scores satisfying $\mathbb{E}_y[(\sum_{t=1}^{L} |a_t|)^2] < \infty$, and define

$$X = \sum_{t=1}^{L} Z_t a_t, \qquad \bar{X} = \sum_{t=1}^{L} w_t a_t. \tag{17}$$

Both sums are square-integrable, and Proposition 1 gives $\bar{X} = \mathbb{E}[X \mid \mathcal{Y}]$. The tower property and the law of total variance yield

$$\mathbb{E}[\bar{X}] = \mathbb{E}[X], \qquad \operatorname{Var}(X) = \operatorname{Var}(\bar{X}) + \mathbb{E}[\operatorname{Var}(X \mid \mathcal{Y})]. \tag{18}$$

Marginalizing τ therefore removes the conditional variance introduced by sampling the lifetime while preserving the expected truncated sum.

A.3 REJECTION AND TOTAL-VARIATION IDENTITIES

Fix a context s_t and write $p(v) = \pi_0(v \mid s_t)$ and $q(v) = \pi_T(v \mid s_t)$. A sampled token $v \sim p$ satisfies $p(v) > 0$ almost surely. Conditional on s_t and v, speculative verification accepts the token with probability $\min(1, q(v)/p(v))$, so its rejection probability is

$$1 - \min\left(1, \frac{q(v)}{p(v)}\right) = \left(1 - \frac{q(v)}{p(v)}\right)^{+} = \delta_t. \tag{19}$$

Averaging over $v \sim p$ at this fixed context gives

$$\mathbb{E}_{v \sim p}[\delta_t] = \sum_{v : p(v) > q(v)} \big(p(v) - q(v)\big) = \frac{1}{2} \sum_{v} |p(v) - q(v)| = \mathrm{TV}(p, q). \tag{20}$$

A.4 CONTROLLED ADAPTIVE SURVIVAL

Proof of Proposition 2. Fix L, $\beta \geq 0$, and $p \geq 1$. For $k < L$, the Cox rate $\lambda_k = \lambda_0(k)e^{\beta\delta_k}$ induces the discrete hazard $h_k^{\mathrm{Cox}} = 1 - e^{-\lambda_k} \in [0, 1)$. Hence

$$\prod_{k<t}(1 - h_k^{\mathrm{Cox}}) = \prod_{k<t} e^{-\lambda_k} = e^{-\sum_{k<t}\lambda_0(k)e^{\beta\delta_k}} = w_t. \tag{21}$$

The restricted lifetime $T = \min(\tau, L)$ is specified by

$$\Pr(T = k \mid y) = w_k - w_{k+1} \quad (1 \leq k < L), \qquad \Pr(T = L \mid y) = w_L.$$

These masses are nonnegative and sum to one, with $T = 1$ when $L = 1$. Thus w_t is a valid survival probability under Definition 1, and Proposition 1 applies. Appendix B.4 gives the corresponding identity for the clipped training surrogate.

For $k < t$, differentiating with respect to δ_k while holding the other deviation scores fixed gives

$$0 \leq \frac{\partial[-\log w_t]}{\partial\delta_k} = \beta\lambda_0(k)e^{\beta\delta_k} \leq \beta e^{\beta}\lambda_0(k), \tag{22}$$

where the last inequality uses $\delta_k \in [0, 1]$. Summing over the suffix gives

$$\sum_{t>k} \frac{\partial[-\log w_t]}{\partial\delta_k} = (L - k)\beta\lambda_0(k)e^{\beta\delta_k}. \tag{23}$$

Convexity of Λ_0 gives $\lambda_0(k) \leq \Lambda_0'(k)$. With $u = k/L$, convexity of $u \mapsto u^p$ then yields

$$(L - k)\lambda_0(k) \leq \frac{p\,u^{p-1}(1 - u)}{1 - u^p} \leq 1.$$

This proves equation 13 uniformly in L and $p \geq 1$. The product-limit sensitivity $(L - k)/(1 - \delta_k)$ has no analogous uniform bound.

For the weight envelope, let $\Lambda_t^B = \sum_{k<t}\lambda_0(k) = -\log B_t$. Since $1 \leq e^{\beta\delta_k} \leq e^{\beta}$, we have $\Lambda_t^B \leq H_t \leq e^{\beta}\Lambda_t^B$. Exponentiating the negatives gives

$$B_t^{e^{\beta}} = e^{-e^{\beta}\Lambda_t^B} \leq w_t \leq e^{-\Lambda_t^B} = B_t, \tag{24}$$

which proves equation 14. For finite ℓ_t, multiplying this envelope by $|\ell_t|$ gives $B_t^{e^{\beta}}|\ell_t| \leq |\hat{A}_t| \leq B_t|\ell_t|$. The weight lower bound $B_t^{e^{\beta}}$ remains positive for finite L and β, even when a deviation score equals one.

For two trajectories of the same length with the same β, p, $\delta_k^A \geq \delta_k^B$ at every position implies $H_t^A \geq H_t^B$ and $w_t^A \leq w_t^B$. Consequently,

$$L_{\mathrm{eff}}(y^A) = \sum_{t=1}^{L} w_t^A \leq \sum_{t=1}^{L} w_t^B = L_{\mathrm{eff}}(y^B). \tag{25}$$

The tail-sum identity also gives

$$L_{\mathrm{eff}}(y) = \sum_{t=1}^{L} w_t = \mathbb{E}[\min(\tau, L) \mid y], \tag{26}$$

so L_{eff} is the conditional restricted mean lifetime. As deviations vanish pointwise at fixed L, β, p, we have $H_t \to \Lambda_t^B$, $w_t \to B_t$, and $L_{\mathrm{eff}}(y) \to \sum_{t=1}^{L} B_t$. This recovers the baseline credit profile in the zero-deviation limit. □

Early-hazard scaling For fixed k, p, β as $L \to \infty$, the expansion $-\log(1 - z) = z + O(z^2)$ gives

$$\lambda_0(k) = \log\frac{1 - ((k-1)/L)^p}{1 - (k/L)^p} = \frac{k^p - (k-1)^p}{L^p} + O(L^{-2p}). \tag{27}$$

By equation 12, the sensitivity of each subsequent log-weight to this early deviation is $O(L^{-p})$, uniformly over $\delta_k \in [0, 1]$. Summing over the suffix gives $O(L^{1-p})$, which remains bounded for $p = 1$ and vanishes for $p > 1$.

B IMPLEMENTATION DETAILS

B.1 TRAINING DATA, MODELS, AND OPTIMIZATION

We use the G-OPD training setup (Yang et al., 2026) and its `verl` implementation on 57K DeepMath problems of difficulty 6 or above and 25K Eurus-RL-Code problems (He et al., 2025, Cui et al., 2025). The students are Qwen3-0.6B, Qwen3-1.7B, and Qwen3-4B, with all models in non-thinking mode (Yang et al., 2025). Single-teacher runs pair each student with Qwen3-4B-Instruct-2507 or Qwen3-30B-A3B-Instruct-2507, comparing Cox-OPD, vanilla OPD (Agarwal et al., 2024), and IW-OPD (Xie et al., 2026) under shared settings (Table 4). Multi-teacher runs jointly train each student on both domains, routing math and code prompts to Qwen3-4B RL-Math and RL-Code, respectively.

Table 4: **OPD training configuration.** The Cox parameters below apply to the main experiments. Figure 4b reports the parameter ablation.

Setting	**Value**
Prompts per rollout batch	1,024
Responses per prompt	1
PPO mini-batch size	1,024
Training epochs	2
Learning rate	10^{-5}
PPO clipping parameter ϵ	0.2
Prompt token limit	2,048
Response token limit	16,384
Token-level rollout-correction threshold	5.0
Sampling temperature	1.0
Sampling top-p	1.0
Cox baseline shape p	2
Cox deviation sensitivity β	3

B.2 BENCHMARKS AND EVALUATION METRICS

Table 5 lists the benchmarks and metrics. OPD evaluation uses temperature 1.0 and top-p 1.0. Math mean@32 averages correctness over 32 responses per problem using Math-Verify. HMMT25 averages the February and November sets. Code results report pass@1 estimated from four responses per problem, with EvalPlus (Liu et al., 2023) used for HumanEval+ and MBPP+. Multi-teacher code evaluation additionally includes LiveCodeBench v6 (Jain et al., 2024). For speculative drafting, we evaluate on GSM8K and MATH-500 (Cobbe et al., 2021, Hendrycks et al., 2021), HumanEval and MBPP (Chen et al., 2021, Austin et al., 2021), and MT-Bench (Zheng et al., 2023). Reported averages are computed over the corresponding benchmarks in each table.

Table 5: **Evaluation tasks and metrics.** Single and multi denote the OPD teacher setting. Accuracy and pass@1 are reported as percentages.

Task / setting	**Benchmarks**	**Metric**
On-policy distillation		
Math (single / multi)	AIME24, AIME25, AIME26, HMMT25	Mean@32 accuracy
Code (single)	HumanEval+, MBPP+	Pass@1, $n = 4$
Code (multi)	HumanEval+, MBPP+, LiveCodeBench (LCB)	Pass@1, $n = 4$
Speculative drafting		
Math	GSM8K, MATH-500	
Code	HumanEval, MBPP	Mean acceptance length
Dialogue	MT-Bench	

B.3 COMPUTING SURVIVAL WEIGHTS

After generation, we compute each rollout's weights once from its cached student and teacher log-probabilities and realized length L, then use them as stop-gradient multipliers in equation 6. No outcome labels or additional teacher queries are required. For finite cached scores, the deviation and baseline cumulative hazard are evaluated in numerically stable form,

$$\delta_t = -\operatorname{expm1}(-\max(\ell_t, 0)), \qquad \Lambda_0(k) = -\operatorname{log1p}(-(k/L)^p), \quad k < L, \tag{28}$$

where $\operatorname{expm1}(z) = e^z - 1$ and $\operatorname{log1p}(z) = \log(1+z)$.

An exclusive cumulative sum of $\lambda_0(k)e^{\beta\delta_k}$ gives H_t in equation 11, so δ_t first affects w_{t+1}. This requires $O(L)$ scalar operations and only increments with $k < L$, avoiding the divergent endpoint $\Lambda_0(L)$. For $L = 1$, the sum is empty and $w_1 = 1$.

B.4 SURVIVAL WEIGHTING OF THE CLIPPED SURROGATE

Fix a completed rollout and its cached scores. Write $A_t^0 = -\ell_t$ and define

$$\phi_\epsilon(\rho, A) = \min\{\rho A, \operatorname{clip}(\rho, 1-\epsilon, 1+\epsilon)A\}. \tag{29}$$

For $c \geq 0$, $\phi_\epsilon(\rho, cA) = c\,\phi_\epsilon(\rho, A)$. With $Z_t = \mathbf{1}\{\tau \geq t\}$, this gives

$$\begin{aligned} \mathbb{E}_{\tau|y}\left[\sum_{t=1}^{L} Z_t\,\phi_\epsilon(\rho_t(\theta), A_t^0)\right] &= \sum_{t=1}^{L} w_t\,\phi_\epsilon(\rho_t(\theta), A_t^0) \\ &= \sum_{t=1}^{L} \phi_\epsilon(\rho_t(\theta), w_t A_t^0). \end{aligned} \tag{30}$$

Negating and averaging over rollouts recovers equation 3 using the advantages in equation 6. The identity also holds under a shared normalization independent of τ, yielding an exact marginalization of the sampled surrogate.

B.5 SPECULATIVE DECODING SETUP

We train DFlash (Chen et al., 2026) and its extension with our Cox parameterization (Cox-DFlash) in SpecForge,[4] using Qwen3-4B (Yang et al., 2025) as the frozen target and initializing both methods from the released five-layer checkpoint `z-lab/Qwen3-4B-DFlash-b16`. Training uses a 10% subset of Open-PerfectBlend (Xu et al., 2024) sampled proportionally across its data sources, with responses regenerated by the target in non-thinking mode. Cox-DFlash adds stop-gradient survival weights to the 15 prediction positions in each 16-token block, while all other training and evaluation choices remain fixed. We compare mean acceptance length under greedy decoding on the five benchmarks in Table 5, reporting benchmark averages in Figure 6 and individual curves in Appendix C.2. Table 6 summarizes the training and evaluation configuration.

Table 6: **Speculative-drafting configuration.** Training and evaluation settings are shared by both methods. Cox parameters apply only to Cox-DFlash.

Setting	Value	Setting	Value
Block size	16 tokens	Random seed	42
Training epochs	6	Maximum sequence length	3,072 tokens
Precision	bfloat16	Maximum anchors	512 per sequence
Optimizer	AdamW	Cox sensitivity β	3
Learning rate	10^{-4}	Cox baseline shape p	2
Learning-rate schedule	Cosine decay	Decoding	Greedy
Warmup	4% linear	Prompts per benchmark	256
Gradient norm clipping	1.0	Generation limit	1,024 tokens
Global batch size	14	Checkpoint interval	1,000 training steps

[4] `https://github.com/sgl-project/SpecForge`

C ADDITIONAL RESULTS

C.1 MULTI-TEACHER TRAINING TRAJECTORIES

Figure 7 extends Figure 5 to all seven benchmarks for the Qwen3-4B student. All four variants jointly train on math and code with the same teacher routing. They differ only in where Cox weighting is applied: math, code, both domains (Full Cox), or neither (vanilla OPD). Math and code scores are mean@32 accuracy and pass@1 ($n = 4$), respectively, both reported as percentages. Full Cox generally outperforms vanilla OPD at matched training checkpoints, with the largest gains on AIME24 and AIME26. On HumanEval+ and LiveCodeBench v6, this advantage generally persists even when performance fluctuates or declines.

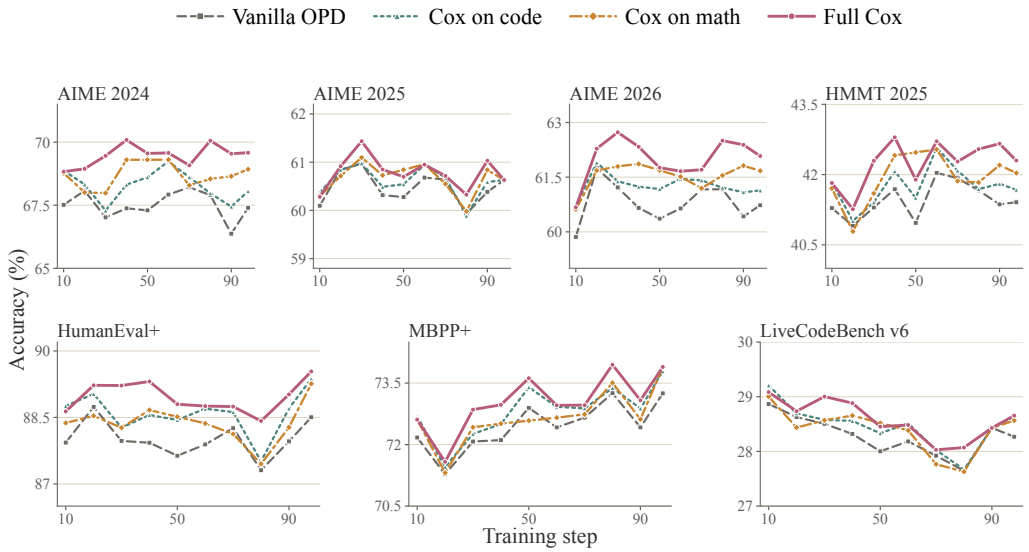


Figure 7: **Multi-teacher training trajectories.**

C.2 SPECULATIVE-DRAFTING RESULTS

Figure 8 reports mean acceptance length on each of the five benchmarks. Cox-DFlash generally outperforms DFlash at matched training checkpoints, including when acceptance length declines on HumanEval and MBPP or fluctuates on MT-Bench. At the final evaluated checkpoint, the gains range from 0.10 to 0.25 tokens across benchmarks.

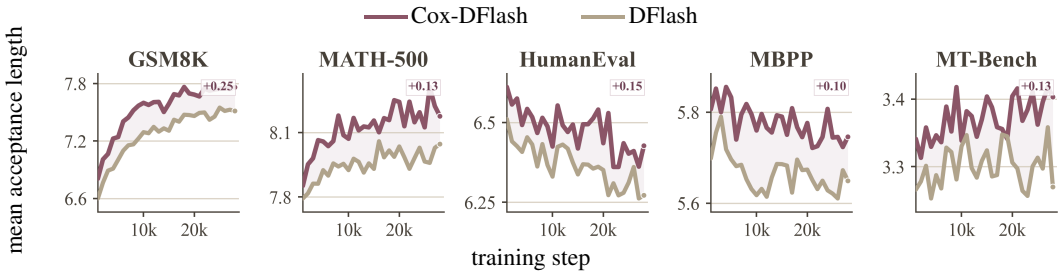


Figure 8: **Acceptance length across benchmarks.** Shading shows the gap between methods. Upper-right values show the final gain of Cox-DFlash over DFlash.