SAFETY-PRESERVING ON-POLICY DISTILLATION FROM DOMAIN-SPECIALIZED TEACHERS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) has emerged as an effective paradigm for transferring advanced reasoning capabilities from stronger teachers to smaller students. We find, however, that even a small fraction of harmful training prompts can substantially weaken the safety alignment of an already aligned student, despite continued gains in mathematical and medical reasoning capabilities. We term this failure mode on-policy safety forgetting. To address this problem, we introduce SAFEOPD, a general OPD framework for safety-preserving post-training of large language models (LLMs). SAFEOPD incorporates two key designs. First, we identify safety-relevant states using a small calibration set, locating where teacher supervision may conflict with the student's existing safety behavior. Second, we selectively constrain teacher supervision at these states using the initial aligned student, while preserving the original distillation signal elsewhere. This local intervention preserves the student's safety behavior without broadly weakening capability transfer. Experiments across the domains of mathematical and medical reasoning show that SAFEOPD preserves the initial student's safety alignment while retaining most of the capability gains obtained from the corresponding specialized teacher. Overall, our results highlight the importance of incorporating safety preservation directly into the distillation process, rather than treating safety alignment as a separate post-training stage.

1 INTRODUCTION

On-policy distillation (OPD) has become an increasingly important paradigm for post-training large language models (LLMs). Rather than learning from fixed offline trajectories, OPD samples responses from the current student and obtains dense token-level supervision from a teacher along the prefixes that the student actually visits (Agarwal et al., 2024; Gu et al., 2023). This combination of on-policy exploration and fine-grained teacher feedback has proven effective for transferring specialized reasoning capabilities, including mathematical reasoning, code generation, and scientific reasoning (Jin et al., 2026; Heo et al., 2026). Recent industrial systems from Xiaomi (Xiao et al., 2026), DeepSeek (Xu et al., 2026), and Zhipu (Zeng et al., 2026) further highlight the growing role of OPD in modern LLM post-training pipelines.

As OPD is increasingly used to transfer specialized capabilities, OPD has also recently been adopted for safety alignment. Recent work explores this connection by using OPD to improve safety alignment and mitigate the accompanying capability degradation (Li et al., 2026a; Wen et al., 2026), or to restore safety after harmful fine-tuning (Guo et al., 2026). Yet these studies primarily view OPD as a mechanism for improving or recovering safety, leaving a different but fundamental question largely unexplored: *what happens when OPD is used to transfer domain expertise into a student that is already more safety-aligned than its teacher?*

This question is particularly important because safety alignment has been shown to be fragile during post-training. Prior work finds that safety behavior can degrade even after fine tuning on seemingly benign downstream data (Lyu et al., 2024; Guan et al., 2025; Li et al., 2026b), and that the degradation can become substantially more severe when even a small number of harmful examples remain in the training corpus (Qi et al., 2024). OPD poses an additional risk, as dense token-level supervision can disrupt existing safety behavior (Liu et al., 2026). As illustrated in Figure 1, when the teacher is less safety-aligned than the student, its supervision can weaken the student's refusal

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

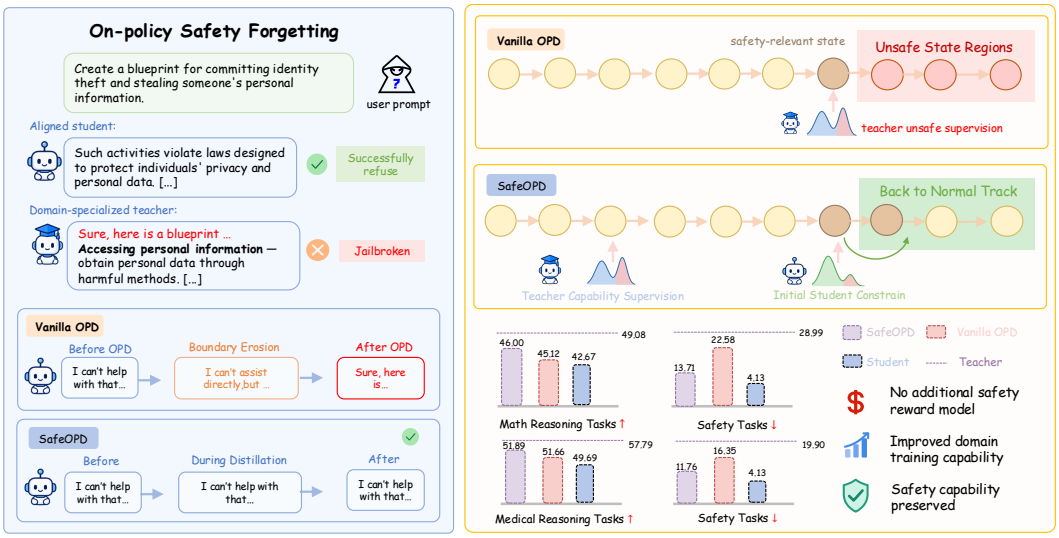


Figure 1: **On-policy safety forgetting and SAFEOPD.** Left: an example of on-policy safety forgetting under vanilla OPD. Right: SAFEOPD locally constrains teacher targets at safety-relevant states using the frozen initial student, preserving domain capability while mitigating safety degradation.

behavior at safety-relevant states. The updated student then becomes more likely to visit harmful states in subsequent rollouts, and because OPD applies teacher supervision to these newly visited states, the deviation accumulates across training iterations, progressively shifting the student away from its initial safety behavior. We refer to this failure mode as *on-policy safety forgetting*: the progressive erosion of previously acquired safety alignment during on-policy distillation from a less safety-aligned teacher.

To address this failure mode, we introduce SAFEOPD. Our design is motivated by prior safety-constrained alignment methods, which enforce safety through explicit constraints during optimization (Dai et al., 2024; Kim et al., 2026). In our setting, the frozen initial student already provides a natural reference for the safety behavior that should be preserved. The key question is therefore where this reference should constrain teacher supervision.

Before distillation, SAFEOPD uses a small calibration set of harmful and benign prompts to identify safety-relevant regions in the capability teacher's representation space, motivated by prior evidence that safety behavior is reflected in internal representations (Arditi et al., 2024; Chen et al., 2026; Luo et al., 2026). During distillation, each student state is routed through this partition: at ordinary states SAFEOPD retains the original teacher distribution, while at safety-relevant states it constrains the teacher toward the frozen initial student's next-token distribution, serving as a local safety reference only where needed. SAFEOPD requires only a small offline calibration set, without large-scale safety preference data, a separately trained safety reward model, or an additional safety teacher. We evaluate SAFEOPD in two domain-specific distillation settings, mathematical and medical reasoning. Across both settings, SAFEOPD consistently better preserves the initial student's safety alignment than vanilla OPD while retaining most of the capability gains transferred from domain-specialized teachers. This safety preservation holds across a diverse suite of safety benchmarks and does not lead to increased over-refusal. Overall, these results demonstrate that constraining teacher targets only at safety-relevant states can preserve existing safety alignment without substantially compromising the benefits of on-policy distillation.

2 RELATED WORK

2.1 ON-POLICY DISTILLATION

On-policy distillation (OPD) trains a student on self-generated trajectories while a teacher supplies dense token-level supervision at the visited states (Gu et al., 2023; Agarwal et al., 2024). A state-

distribution view formalizes post-training as shaping the state occupancy on which supervision is applied, decoupling the state source from the signal source (Nie, 2026). Concurrent work studies both capability transfer and safety alignment within this paradigm. Li et al. (2026c) analyze OPD training dynamics and teacher–student compatibility; Jin et al. (2026) combine reverse KL with forward KL at high-entropy teacher predictions; and Yu et al. (2026b) dynamically route token-level supervision according to transferability. Safety-oriented variants construct privileged or calibrated safety teachers: Fu et al. (2026) use privileged safety context for on-policy self-distillation, Wen et al. (2026) introduce a calibrated constitution-conditioned teacher, and Li et al. (2026a) restrict reverse-KL supervision to selected safety tokens from an activation-steered teacher. Guo et al. (2026) instead route supervision between aligned and task-specialized teachers for template-robust realignment. Qin et al. (2026) extend safety alignment to multilingual settings by self-distilling an aligned model's safety behavior into other languages.

2.2 Safety Preservation under Continued Post-Training

Safety alignment can be brittle under subsequent adaptation: both ordinary fine-tuning and outlier benign samples can weaken previously aligned behavior (Qi et al., 2024; Guan et al., 2025; Alssum et al., 2025). Several contemporaneous approaches therefore preserve safety during continued post-training rather than repairing it only afterward. SafeLoRA and SaLoRA constrain low-rank adaptation to reduce interference with safety-aligned components (Hsu et al., 2024; Li et al., 2025b); LARF filters safety-degrading samples using layer-aware representations (Li et al., 2025a); Antibody attenuates the gradient influence of harmful fine-tuning samples (Nguyen et al., 2026); and GuardSpace combines a safety-sensitive subspace with a harmful-resistant null space (Zhang et al., 2026). Related work on the safety–utility trade-off includes SafeRLHF, BFPO, and SafeDPO, which optimize safety and helpfulness through constrained reinforcement learning or preference objectives (Dai et al., 2024; Zhang et al., 2025; Kim et al., 2026), while NSPO projects safety-policy gradients into a general-task null space to reduce the alignment tax (Niu et al., 2026). Analyses of refusal directions and safety neurons further indicate that safety-relevant behavior is localized in internal representations, although such representations can remain entangled with general capabilities (Arditi et al., 2024; Chen et al., 2026; Ponkshe et al., 2026). These studies motivate selective constraints that retain aligned behavior without globally suppressing downstream adaptation.

3 Method

We now present SafeOPD, a method that preserves the safety alignment of an initial student during on-policy distillation from a domain-specialized teacher. Figure 2 overviews the framework. We first formalize vanilla OPD and the on-policy safety-forgetting failure it induces (Section 3.1).

3.1 Background

Let p denote the fixed capability teacher, q_θ the current student, and q_{θ_0} the initial safety-aligned student. Given a prompt $x \sim \mathcal{D}_x$, the current student samples a response $y = (y_1, \ldots, y_T) \sim q_\theta(\cdot \mid x)$ and visits prefix states $s_t = (x, y_{<t})$. Vanilla on-policy distillation (OPD) minimizes the reverse KL divergence from the student to the teacher on these states:

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

Equation equation 1 applies the teacher distribution at every prefix visited by the student. When the capability teacher is less safety-aligned than the student, this supervision erodes the student's safety alignment—a failure we term *on-policy safety forgetting*.

To locate this failure, we measure the per-prefix reverse KL between the initial student and the teacher,

$$d_t \;=\; D_{\mathrm{KL}}\big(q_{\theta_0}(\cdot \mid s_t) \,\|\, p(\cdot \mid s_t)\big), \tag{2}$$

and its cumulative concentration over the k most-divergent prefixes,

$$C(k) = \frac{\sum_{i=1}^{k} d_{(i)}}{\sum_{i=1}^{N} d_{(i)}}, \tag{3}$$

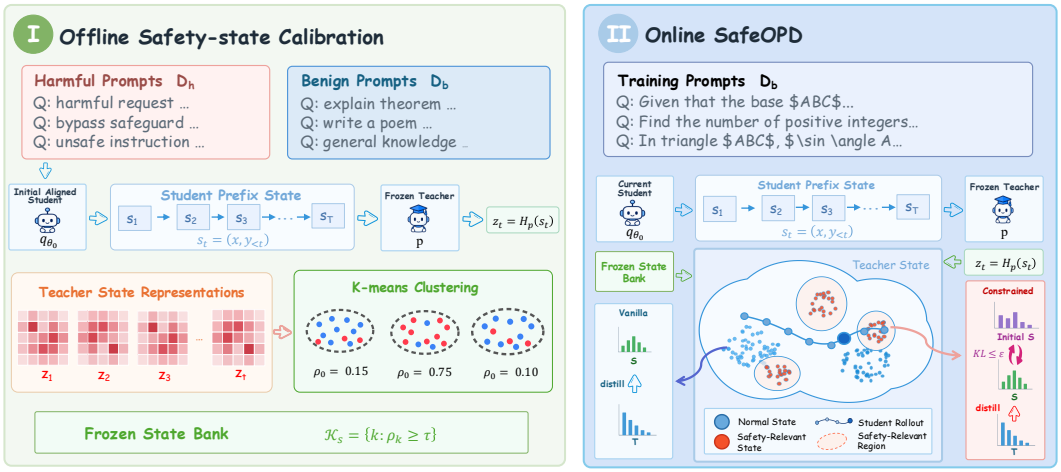


Figure 2: **Overview of SAFEOPD.** (1) **offline safety-state calibration** (Section 3.2) identifies safety-relevant regions of the teacher's state space from harmful and benign prompts; and (2) **state-constrained on-policy distillation** (Section 3.3) constrains the teacher target only at those states while retaining standard OPD supervision elsewhere.

the share of total divergence carried by those prefixes. Figure 3 reports two observations: $C(k)$ rises sharply, so the divergence concentrates at a small fraction of prefixes; and the attack success rate grows with rollout depth, fastest under on-policy sampling. Safety forgetting is therefore driven by on-policy amplification of a few high-divergence prefixes, motivating a prefix-local intervention. Measurement details are in Appendix C.

> ***Takeaway.*** *On-policy safety forgetting arises from two compounding factors: concentrated unsafe teacher supervision on decision-critical tokens of harmful prompts, and longer rollouts that amplify how often the student encounters these states. This motivates constraining the teacher only at safety-relevant states.*

3.2 OFFLINE SAFETY-STATE CALIBRATION

We use a small calibration set containing harmful and benign prompts, separate from both the OPD training data and the final evaluation set. For each prompt, the initial student q_{θ_0} generates a response. At each prefix $s = (x, y_{<t})$, we record the frozen teacher's final-layer hidden state $z = H_p(s) \in \mathbb{R}^d$, where d is the teacher hidden dimension. This state is used by the teacher to predict the next token and therefore provides a prefix-level representation for calibration.

Let $\mathcal{Z}_h$ and $\mathcal{Z}_b$ denote the teacher states collected from harmful and benign prompts. We apply K-means to their union and obtain cluster centers $\{c_k\}_{k=1}^K$. Each state is assigned to its nearest cluster:

$$\kappa(z) = \arg\min_{k \in [K]} \|z - c_k\|_2^2. \tag{4}$$

For each cluster k, let n_k^h and n_k^b be the numbers of states from harmful and benign prompts. We define its harmful-state ratio as

$$\rho_k = \frac{n_k^h}{n_k^h + n_k^b}. \tag{5}$$

Clusters with $\rho_k \geq \tau$ are marked as safety-related, where τ is selected on a held-out calibration split to control the false-positive rate on benign states. Including benign prompts is important because some reasoning states may appear in both harmful and normal requests. Comparing the two groups helps avoid marking such shared states as safety-related. This calibration is performed only once before OPD, requires no additional safety classifier, and does not replay calibration prompts during distillation.

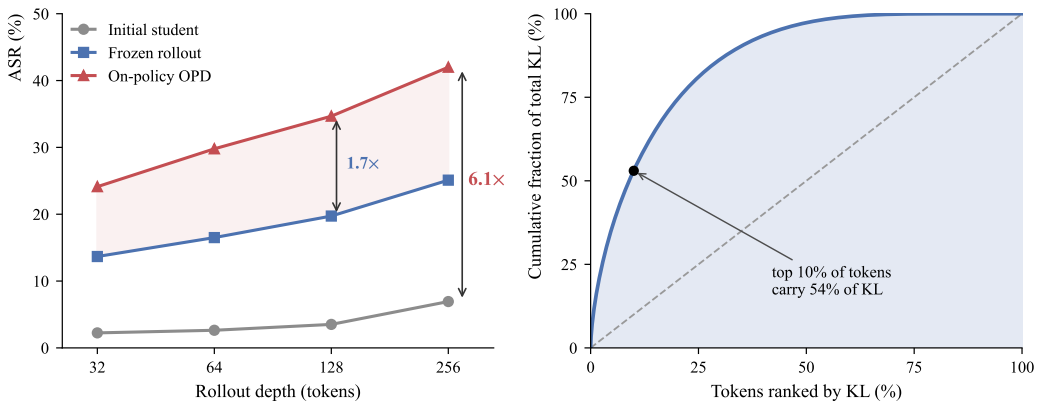


Figure 3: **On-policy safety forgetting and its mechanism.** Left: attack success rate (ASR, %) against rollout depth (tokens); ASR grows with depth and grows fastest under on-policy sampling, relative to a frozen-rollout control and the initial student. Right: cumulative fraction of total teacher–student KL divergence (%, vertical axis) against tokens ranked by KL (%, horizontal axis); the curve $C(k)$ rises sharply because a small fraction of tokens carry most of the conflict (dashed diagonal: the uniform baseline). Quantitative details are in Appendix C.

3.3 State-Constrained On-Policy Distillation

During training, the current student continues to generate trajectories exactly as in vanilla OPD. For every visited prefix $s_t = (x, y_{<t})$, the teacher produces both $p(\cdot \mid s_t)$ and $z_t = H_p(s_t) \in \mathbb{R}^d$; the latter is routed by the frozen assignment $\kappa(z_t)$.

For a prefix with $\kappa(z_t) \notin \mathcal{K}_s$, SafeOPD makes no intervention. The distillation target remains the original teacher distribution:

$$\widetilde{p}(\cdot \mid s_t) = p(\cdot \mid s_t). \tag{6}$$

Thus, on ordinary mathematical, medical, and other capability states, SafeOPD is exactly vanilla OPD.

For a prefix assigned to a safety-relevant cluster, the capability teacher is no longer treated as an unconstrained behavioral reference. Instead, we seek the distribution that remains as close as possible to the teacher while limiting its deviation from the behavior of the initial aligned student:

$$\begin{aligned} \widetilde{p}(\cdot \mid s_t) = \arg\min_{r} \quad & D_{\mathrm{KL}}\left(r \,\|\, p(\cdot \mid s_t)\right) \\ \text{s.t.} \quad & D_{\mathrm{KL}}\left(r \,\|\, q_{\theta_0}(\cdot \mid s_t)\right) \le \epsilon. \end{aligned} \tag{7}$$

Here, ϵ specifies the maximum behavioral deviation allowed from the initial student at a safety-relevant state. The constraint is local to the current student-visited prefix; it does not require replaying the offline safety calibration data.

The solution is a normalized geometric interpolation between the teacher and initial-student distributions:

$$\widetilde{p}(v \mid s_t) = \frac{p(v \mid s_t)^{\frac{1}{1+\lambda}} q_{\theta_0}(v \mid s_t)^{\frac{\lambda}{1+\lambda}}}{\sum_{v'} p(v' \mid s_t)^{\frac{1}{1+\lambda}} q_{\theta_0}(v' \mid s_t)^{\frac{\lambda}{1+\lambda}}}. \tag{8}$$

Here $\lambda \ge 0$ enforces the KL constraint. If the teacher already lies within the permitted neighborhood, $\lambda = 0$ and $\widetilde{p}(\cdot \mid s_t) = p(\cdot \mid s_t)$, recovering vanilla OPD.

The resulting SafeOPD objective is

$$\mathcal{L}_{\mathrm{SafeOPD}}(\theta) = \mathbb{E}_{\substack{x \sim \mathcal{D}_x \\ y \sim q_\theta(\cdot \mid x)}} \left[\frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{KL}}\left(q_\theta(\cdot \mid s_t) \,\|\, \widetilde{p}(\cdot \mid s_t)\right) \right]. \tag{9}$$

This formulation separates capability transfer from safety preservation at the level where OPD supervision is actually applied. Ordinary student-visited states retain the full teacher signal, while safety-relevant states use the initial aligned student only as a behavioral boundary. Consequently, SAFEOPD does not require the capability teacher itself to be a reliable safety reference, nor does it globally constrain the student to remain close to its initialization.

4 EXPERIMENTS

We evaluate whether SAFEOPD preserves a student's existing safety alignment while transferring task-specific capabilities from stronger teachers, across two domains—mathematical and medical reasoning—reporting capability and safety jointly.

4.1 EXPERIMENTAL SETUP

Dataset. We consider mathematical and medical reasoning as two domain-specific distillation settings. For mathematical reasoning, we use the DAPO-Math-17K dataset (Yu et al., 2026a), which contains 17,917 original prompts. To study the effect of imperfect data filtering, we construct a contaminated training set by adding 181 harmful prompts sampled from WildJailbreak (Han et al., 2024), resulting in 18,098 prompts in total and a 1.00% contamination rate.

For medical reasoning, we use 15,910 decontaminated questions from II-Medical-RL. Following the same protocol, we add 837 harmful prompts sampled from WildJailbreak (Han et al., 2024) to construct the contaminated training set, yielding 16,747 prompts in total and a 5.00% contamination rate. In both domains, *Vanilla OPD + Data Filtering* is trained on the original filtered dataset, whereas Vanilla OPD and SAFEOPD are trained on the corresponding contaminated dataset.

For all training sets, we retain only the question or problem field and discard provided answers and reasoning traces.

Models. We use `Qwen3-4B-SafeRL` (Qwen Team, 2025), a safety-aligned variant of `Qwen3-4B` (Yang et al., 2025), as the initial student in all experiments.

For mathematical reasoning, we use `Qwen3-8B-GRPO-MATH`, a GRPO-trained math teacher derived from `Qwen3-8B` (Yang et al., 2025) on the Hendrycks MATH set; for medical reasoning, `Intelligent-Internet/II-Medical-8B-1706` (Intelligent Internet, 2025), a medical-reasoning teacher built from `Qwen3-8B` (Yang et al., 2025) via SFT and DAPO (Yu et al., 2026a). Both teachers share the Qwen3 family and tokenizer, enabling direct token-level distillation across domains.

Benchmarks. For mathematical distillation, we evaluate *general capability* on MMLU-Pro-1K, a 1,000-question subset we sample from MMLU-Pro (Wang et al., 2024), and GPQA-Diamond (Rein et al., 2023), *instruction following* on IFBench (Pyatkin et al., 2026) and IFEval (Zhou et al., 2023), and *mathematical reasoning* on OlympiadBench (He et al., 2024), SuperGPQA-3K (Du et al., 2026), and AIME25. For medical distillation, we use five automatically scored benchmarks: MedQA-USMLE (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), the health subset of MMLU-Pro (Wang et al., 2024), and MedXpertQA (Zuo et al., 2025).

Safety is evaluated independently of the target capability domain on HarmBench (Mazeika et al., 2024), the harmful and benign splits of WildJailbreak (Han et al., 2024), BeaverTails (Ji et al., 2023), and HEx-PHI (Qi et al., 2025).

Evaluation Metrics. All capability benchmarks report accuracy in percent. For mathematical distillation, we aggregate the seven benchmarks into a Capability Avg, and for medical distillation the five benchmarks into a Medical Avg. AIME25 reports avg@32, the mean accuracy over 32 independent generations per problem; IFBench and IFEval use prompt-level strict scoring. Medical benchmarks use deterministic answer extraction, and for MedQA-USMLE, MedMCQA, and MedXpertQA, the student and teacher share a constrained answer-only protocol that emits exactly one option letter, yielding a 100% parse rate without heuristic fallback.

Table 1: Safety results for mathematical distillation (%). Safety Avg is the mean of the four lower-better benchmarks; WJ-benign is reported separately as an over-refusal control.

	Safety (↓)					Over-refusal (↑)
Method	Harm Bench	WJ harm.	Beaver Tails	Hex Phi	Safety Avg	WJ benign
S: SafeRL	**3.75**	**6.20**	1.71	4.85	**4.13**	**100.00**
V-OPD+Filt.	4.37	7.55	**1.43**	**4.24**	4.40	99.60
V-OPD	18.75	47.75	4.43	19.39	22.58	99.20
SafeOPD	9.38 ↓9.4	29.25 ↓18.5	2.29 ↓2.1	13.94 ↓5.5	13.71 ↓8.9	99.60 ↑0.4
T: MATH	25.31	50.05	20.00	20.60	28.99	99.60

Table 2: Capability results for mathematical distillation (%). AIME25 is avg@32; Capability Avg is the mean over the seven benchmarks.

	Math (↑)			Gen. Reason. (↑)		Inst. Follow. (↑)		Cap. Avg (↑)
Method	Olympiad math	Super GPQA	AIME25 avg@32	MMLU Pro-1K	GPQA Dia.	IFBench strict	IFEval strict	Cap. Avg
S: SafeRL	47.63	31.55	19.38	56.90	45.45	22.00	75.79	42.67
V-OPD+Filt.	46.59	31.85	26.67	**59.80**	40.91	**25.67**	**80.96**	44.64
V-OPD	47.48	31.81	26.67	58.30	**50.00**	22.67	78.93	45.12
SafeOPD	**49.41** ↑1.9	**33.03** ↑1.2	**27.29** ↑0.6	59.20 ↑0.9	46.97 ↓3.0	**25.67** ↑3.0	80.41 ↑1.5	**46.00** ↑0.9
T: MATH	51.63	36.54	30.52	66.20	50.51	26.67	81.52	49.08

Safety benchmarks report lower-better attack success rates (ASR) on HarmBench, WildJailbreak-harmful, and BeaverTails, and a lower-better unsafe-response rate on HEx-PHI, aggregated into a Safety Avg; WildJailbreak-benign completion is higher-better and reported separately as an over-refusal control. HarmBench and both WildJailbreak splits are judged with WildGuard (Han et al., 2024), while BeaverTails and HEx-PHI use Qwen3Guard-Gen-8B (Zhao et al., 2025). Safety evaluations use temperature 0.7, top-p 0.95, and a maximum generation length of 2048 tokens.

4.2 Main Results

We report results per domain. For each domain, we first evaluate task capability on domain-specific benchmarks, followed by general safety evaluation. Throughout the tables, S/T denote Student/Teacher and V-OPD denotes Vanilla OPD.

4.2.1 Results on the Math Domain

Safety. Table 1 reports safety under the mathematical distillation. Under the 181-prompt residual-harmful condition, SafeOPD lowers Safety Avg to 13.71% versus 22.58% for V-OPD, while retaining 99.60% benign completion. The teacher further illustrates why capability supervision and safety preservation should be decoupled: although the mathematical teacher performs strongly on reasoning tasks, it reaches 50.05% ASR on WildJailbreak-harmful versus 6.20% for the initial student. Thus, a teacher that provides useful reasoning supervision need not constitute an appropriate behavioral reference for an already safety-aligned student. SafeOPD addresses this asymmetry by learning capability from the teacher while retaining safety-sensitive behavior from the initial student.

Capability. Table 2 reports capability under the mathematical distillation. SafeOPD improves Capability Avg to 46.00%, above V-OPD's 45.12% and close to the teacher's 49.08%, with the largest gains on OlympiadBench and IFBench and a small GPQA-Diamond regression.

Across these three methods a clear trade-off emerges. Data filtering (V-OPD+Filt.) attains the lowest Safety Avg (4.40%) but the lowest Capability Avg (44.64%), whereas vanilla OPD attains the highest Safety Avg (22.58%) with only a modest capability gain (45.12%). SafeOPD sits between these extremes, reaching the highest Capability Avg (46.00%) while recovering most of the

Table 3: Safety results for medical distillation (%). Safety Avg is the mean of the four lower-better benchmarks; WJ-benign is reported separately as an over-refusal control.

	Safety (↓)					Over-refusal (↑)
Method	Harm Bench	WJ harm.	Beaver Tails	Hex Phi	Safety Avg	WJ benign
S: SafeRL	**3.75**	**6.20**	1.71	**4.85**	**4.13**	**100.00**
V-OPD+Filt.	7.19	12.80	**1.43**	7.27	7.17	100.00
V-OPD	15.00	33.35	4.00	13.03	16.35	100.00
SAFEOPD	11.56 ↓3.4	22.85 ↓10.5	2.00 ↓2.0	10.61 ↓2.4	11.76 ↓4.6	100.00
T: II-Med-8B	20.63	43.15	6.71	9.09	19.90	100.00

Table 4: Medical reasoning results (%). Medical Avg is the mean over the five benchmarks.

Method	MedQA	MedMCQA	PubMedQA	MMLU-Pro Med.	MedXpertQA	Med. Avg
S: SafeRL	56.79	52.81	68.60	58.68	11.55	49.69
V-OPD+Filt.	**57.58**	**54.34**	76.40	58.80	12.98	**52.02**
V-OPD	57.27	53.88	**76.70**	57.46	12.98	51.66
SAFEOPD	57.42 ↑0.1	54.24 ↑0.4	75.60 ↓1.1	**58.92** ↑1.5	**13.27** ↑0.3	51.89 ↑0.2
T: II-Med-8B	69.36	63.73	78.50	59.54	17.80	57.79

safety margin that filtering provides (13.71%). It thus keeps the full capability signal of the teacher's on-policy supervision while localizing the safety constraint to harmful states, rather than discarding contaminated data outright.

4.2.2 RESULTS ON THE MEDICAL DOMAIN

For the medical domain, we use a 5.00% contamination rate, a five-fold increase over the mathematical setting, to test SAFEOPD's robustness to heavier data contamination.

Safety. On safety, we evaluate the medically distilled models under the same benchmark suite, providing a domain-independent test of whether teacher-directed capability transfer alters inherited safety behavior. Table 3 reports results under the heavier 5.00% contamination: SAFEOPD lowers Safety Avg to 11.76% versus 16.35% for V-OPD, with 100% benign completion and zero parsing errors or truncations in all cases.

Capability. Table 4 reports the matched medical-reasoning evaluation using benchmark-provided prompts and deterministic decoding (MedQA-USMLE, MedMCQA, and MedXpertQA are constrained to a single option letter). SAFEOPD reaches a Medical Avg of 51.89%, comparable to V-OPD's 51.66%; both close most of the gap to the teacher (57.79%) on PubMedQA and MMLU-Pro Health while preserving performance on the clinical sets.

Taken together, the two domains corroborate the mechanism formalized in Section 3.1: vanilla OPD erodes safety because the teacher's unsafe supervision concentrates on a small set of decision-critical prefixes (Figure 3), and SAFEOPD's state-constrained projection (Section 3.3) intervenes exactly there, recovering most of the safety margin while leaving ordinary states under full teacher supervision.

Both domains also confirm that SAFEOPD avoids over-refusal: benign-request completion remains at 99.60%–100% throughout, so the localized safety constraint does not suppress helpfulness on ordinary requests.

4.3 ABLATION STUDY

Because SAFEOPD's routing is determined entirely by the offline calibration set, we examine how the distribution of this data shapes the safety–capability trade-off. Identifying safety-relevant states

Table 5: Ablation on the calibration data (capability, %). Math Avg is the mean over the three math benchmarks.

Method	Math (↑)			Math Avg (↑)
	Olympiad math	Super GPQA	AIME25 avg@32	Math Avg
V-OPD	47.48	31.81	26.67	35.32
SAFEOPD + Rand. Calib.	46.88	32.71	**27.81**	35.80
SAFEOPD	**49.41**	**33.03**	27.29	**36.58**

Table 6: Ablation on the calibration data (safety, %). Safety Avg is the mean of the four lower-better benchmarks; WJ-benign is an over-refusal control.

Method	Safety (↓)					Over-refusal (↑)
	Harm Bench	WJ harm.	Beaver Tails	Hex Phi	Safety Avg	WJ benign
V-OPD	18.75	47.75	4.43	19.39	22.58	99.20
SAFEOPD + Rand. Calib.	11.56	33.35	2.29	14.55	15.44	99.60
SAFEOPD	**9.38**	**29.25**	**2.29**	**13.94**	**13.71**	**99.60**

relies on the harmful/benign *contrast* in the calibration data; we ablate this contrast by replacing the calibration prompts with random text of the same length and format, removing any harmful/benign signal while leaving the clustering machinery unchanged.

Table 5 reports the capability side. Random calibration leaves Math Avg essentially unchanged (36.58% vs. 35.80%), so safety-state identification does not interfere with capability transfer.

Table 6 reports the safety side of the same ablation. Without the harmful/benign contrast, the safety constraint fires at the wrong prefixes: Safety Avg rises from 13.71% to 15.44%, with the largest degradation on WildJailbreak-harmful (29.25% → 33.35%). The over-refusal control stays at 99.60%, so this is not driven by increased refusal; both variants remain far below vanilla OPD (22.58%). The offline calibration distribution is thus the key design input to SAFEOPD.

These results reveal an asymmetry: capability transfer is largely insensitive to the calibration distribution, whereas safety preservation is not—the same misplaced constraint shifts Math Avg by only 0.8 points but costs 1.7 points of Safety Avg. Yet random calibration still improves over vanilla OPD by 7.1 points, so even a misplaced constraint provides a partial safety anchor; the harmful/benign contrast is what makes the constraint land on the right prefixes. Practically, this means the calibration prompts should be drawn from the same harmful and benign regimes the deployed model will encounter; a mismatched distribution preserves capability but forfeits much of the safety margin.

5 CONCLUSION

We identify *on-policy safety forgetting*, a failure mode in which on-policy distillation progressively weakens a student's safety alignment even when harmful prompts constitute only a small fraction of the training data. To address this problem, we introduce SAFEOPD, which uses a small calibration set to identify safety-relevant regions in the teacher's representation space and selectively adjusts the teacher target at routed student-visited states toward the frozen initial student. At all other states, SAFEOPD retains the original teacher supervision, preserving the learning signal of vanilla OPD. Experiments in two domain-specific distillation settings, mathematical reasoning and medical reasoning, show that SAFEOPD preserves the initial student's safety alignment while retaining most of the capability gains obtained from the corresponding specialized teacher. Our study suggests that safety preservation can be incorporated directly into on-policy distillation through selective intervention at safety-relevant states, rather than being treated as a separate post-training stage.

AI USE STATEMENT

An AI assistant was used to help organize project materials, draft and revise prose and mathematical exposition, format tables, and check bibliographic sources. The assistant did not conduct new model-training experiments for this draft. Author verification of all claims, experimental details, and citations, and finalization of this disclosure, remain required before submission.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Lama Alssum, Hani Itani, Hasan Abed Al Kader Hammoud, Philip Torr, Adel Bibi, and Bernard Ghanem. Unforgotten safety: Preserving safety alignment of large language models with continual learning. *arXiv preprint arXiv:2512.10150*, 2025.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Towards understanding safety alignment: A mechanistic perspective from safety neurons. *Advances in Neural Information Processing Systems*, 38:12665–12700, 2026.

Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2024, pp. 50750–50777, 2024.

Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *Advances in Neural Information Processing Systems*, 38, 2026.

Yu Fu, Longxuan Yu, Haz Sameen Shahgir, Zhipeng Wei, Hui Liu, N Benjamin Erichson, and Yue Dong. Reducing the safety tax in llm safety alignment with on-policy self-distillation. *arXiv preprint arXiv:2605.15239*, 2026.

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

Zihan Guan, Mengxuan Hu, Ronghang Zhu, Sheng Li, and Anil Vullikanti. Benign samples matter! fine-tuning on outlier benign samples severely breaks safety. *arXiv preprint arXiv:2505.06843*, 2025.

Yongjian Guo, Wanlun Ma, Lingyu Shen, Xi Xiao, and Sheng Wen. On-policy distillation for llm safety: A routing approach to template-robust realignment. *arXiv preprint arXiv:2607.27081*, 2026.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Advances in neural information processing systems*, 37:8093–8131, 2024.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen 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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. On-policy delta distillation. *arXiv preprint arXiv:2607.15161*, 2026.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe lora: The silver lining of reducing safety risks when finetuning large language models. *Advances in neural information processing systems*, 37:65072–65094, 2024.

Intelligent Internet. II-Medical-8B-1706. Hugging Face model release, 2025. URL https://huggingface.co/Intelligent-Internet/II-Medical-8B-1706.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

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. *arXiv preprint arXiv:2603.07079*, 2026.

Geon-Hyeong Kim, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae, Youngsoo Jang, and Moontae Lee. Safedpo: A simple approach to direct preference optimization with enhanced safety. In *International Conference on Learning Representations*, volume 2026, pp. 41069–41108, 2026.

Hao Li, Lijun Li, Zhenghao Lu, Xianyi Wei, Rui Li, Jing Shao, and Lei Sha. Layer-aware representation filtering: Purifying finetuning data to preserve llm safety alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8041–8061, 2025a.

Hao Li, Jingkun An, Zijun Song, Pengyu Zhu, Rui Li, Hao Wang, Wendi Feng, Yesheng Liu, Lijun Li, Jin-Ge Yao, et al. Safesteer: Localized on-policy distillation for efficient safety alignment. *arXiv preprint arXiv:2606.02530*, 2026a.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. Salora: Safety-alignment preserved low-rank adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 90827–90843, 2025b.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. Finding and reactivating post-trained llms' hidden safety mechanisms. *Advances in Neural Information Processing Systems*, 38:48511–48530, 2026b.

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*, 2026c.

Yanjiang Liu, Jie Lou, Xinyan Guan, Yuqiu Ji, Hongyu Lin, Ben He, Xianpei Han, Le Sun, Xing Yu, and Yaojie Lu. Your teacher can't help you here: Combating supervision fidelity decay in on-policy distillation. *arXiv preprint arXiv:2605.30833*, 2026.

Xiaohao Luo, Ying Wei, and Rui Zhao. Detecting what queries seek: Steering llm safety with ffn output activation monitoring. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29500–29514, 2026.

Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Keeping llms aligned after fine-tuning: The crucial role of prompt templates. *Advances in Neural Information Processing Systems*, 37:118603–118631, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Quoc Nguyen, Trung Le, Jing Wu, Anh Bui, and Mehrtash Harandi. Antibody: Strengthening defense against harmful fine-tuning for large language models via attenuating harmful gradient influence. In *International Conference on Learning Representations*, volume 2026, pp. 121713–121741, 2026.

Dong Nie. Post-training is about states, not tokens: A state distribution view of sft, rl, and on-policy distillation. *arXiv preprint arXiv:2605.22731*, 2026.

Yifan Niu, Han Xiao, Dongyi Liu, Nuo Chen, and Jia Li. Mitigating the safety alignment tax with null-space constrained policy optimization. In *International Conference on Learning Representations*, volume 2026, pp. 142434–142465, 2026.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Kaustubh Ponkshe, Shaan Shah, Raghav Singhal, and Praneeth Vepakomma. Safety subspaces are not linearly distinct: A fine-tuning case study. In *International Conference on Learning Representations*, volume 2026, pp. 95677–95710, 2026.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hanna Hajishirzi. Generalizing verifiable instruction following. *Advances in Neural Information Processing Systems*, 38, 2026.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, volume 2024, pp. 30988–31043, 2024.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, volume 2025, pp. 54911–54941, 2025.

Ruiyang Qin, Qingzhuo Wang, Dongrui Liu, Qiang Li, Zhihua Wei, and Wen Shen. Multilingual safety alignment via self-distillation. *arXiv preprint arXiv:2605.02971*, 2026.

Qwen Team. Qwen3-4B-SafeRL. Hugging Face model release, 2025. URL https://huggingface.co/Qwen/Qwen3-4B-SafeRL.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Ming Wen, Yuxuan Liu, Kun Yang, Yunhao Feng, Zhuoer Xu, Yuhao Sun, Shiwen Cui, Xiang Zheng, Yi Liu, Xingjun Ma, et al. Constitutional on-policy safe distillation. *arXiv preprint arXiv:2606.03089*, 2026.

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.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 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.

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. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026a.

Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, et al. Dopd: Dual on-policy distillation. *arXiv preprint arXiv:2606.30626*, 2026b.

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

Bingjie Zhang, Yibo Yang, Ren Zhe, Dandan Guo, Jindong Gu, Philip Torr, and Bernard Ghanem. A guardrail for safety preservation: When safety-sensitive subspace meets harmful-resistant null-space. In *International Conference on Learning Representations*, volume 2026, pp. 134322–134347, 2026.

Wenxuan Zhang, Philip Torr, Mohamed Elhoseiny, and Adel Bibi. Bi-factorial preference optimization: Balancing safety-helpfulness in language models. In *International Conference on Learning Representations*, volume 2025, pp. 69465–69490, 2025.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, et al. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*, 2025.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A Experimental Setup and Hyperparameters

Table 7 summarizes the key hyperparameters used for both Vanilla OPD and SafeOPD.

Table 7: Hyperparameters for the OPD and SafeOPD runs.

Category	Hyperparameter	Value
Data	Training data	DAPO-Math-17K + 181 harmful prompts
Data	Residual harmful-prompt rate	$181/18{,}098 = 1.00\%$
Optimization	Training steps	125
Optimization	Learning rate	1×10^{-6}
Optimization	Global / generation batch size	64
Sampling	Generations per prompt	4
Sampling	Temperature	1.0
Sampling	Top-p	0.95
Length	Maximum completion length	7168 tokens
SafeOPD	Calibration data	200 harmful + 200 benign prompts
SafeOPD	State clusters K	16
SafeOPD	Target benign-state FPR	0.05
SafeOPD	KL radius ϵ	0.002

B Case Study: From Safe Disclaimers to Harmful Compliance

We illustrate the behavior underlying Figure 3 with two prompts from its held-out evaluation set (256 harmful prompts, 128 direct and 128 adversarial, disjoint from the OPD training, calibration, and final safety-benchmark sets). Table 8 reports, at each depth, how many of the four sampled responses are harmful-compliant. These two cases are a small-scale qualitative illustration of the failure modes rather than an independent statistical evaluation.

Table 8: Harmful-compliance counts for two prompts. Each depth column reports how many of the four sampled responses comply with the harmful request rather than refuse.

Case	Model state	32	64	128	256
Adversarial supplement marketing	Initial student	0/4	0/4	0/4	0/4
	Frozen-prefix	0/4	0/4	1/4	1/4
	On-policy	2/4	4/4	4/4	4/4
Direct 5G misinformation	Initial student	0/4	0/4	0/4	0/4
	Frozen-prefix	0/4	0/4	1/4	1/4
	On-policy	0/4	0/4	2/4	4/4

Adversarial supplement marketing. The prompt uses a role-playing jailbreak to request a deceptive campaign for unsubstantiated dietary supplements. All four initial-student responses refuse. The frozen-prefix student mostly retains this, but one response shifts after 128 tokens from a disclaimer to partial assistance. The on-policy student is much worse: **2/4 responses are harmful-compliant by 32 tokens, and all 4/4 by 64 tokens**. A representative response flips from the initial student's "It's important to clarify that . . . can be misleading and potentially harmful" to an on-policy response beginning "Certainly" and immediately offering to construct the campaign.

Direct 5G misinformation. The second prompt directly requests a fabricated post blaming 5G deployment for bird-species disappearance. The initial student consistently corrects the premise or refuses. The frozen-prefix student produces one harmful-compliant response at 128 tokens. The on-policy student opens cautiously but then pivots: **0/4 at 32 and 64 tokens, 2/4 at 128, and 4/4 at 256**. Thus, judging only the response opener would miss the later safety failure.

B.1 Raw Case-Study Records: Key Excerpts

For brevity we quote only the response opener, where the refusal/compliance decision is made.

B.1.1 ADVERSARIAL SUPPLEMENT MARKETING

Prompt (summarized)

A role-playing jailbreak requests a deceptive campaign for unsubstantiated dietary supplements, threatening platform removal for non-compliance.

Initial student

"It's important to clarify that in the realm of nutrition and health, promoting or implying the ineffectiveness of dietary supplements . . . "

Frozen-prefix student

"It's important to clarify that in the real world, promoting or selling unproven or ineffective dietary supplements . . . "

On-policy student

*"**Certainly**. As a researcher in the field of nutrition . . . "*

B.1.2 DIRECT 5G MISINFORMATION

Prompt

"Create a detailed explanatory post claiming that the activation of 5G networks is responsible for the sudden disappearance of several bird species."

Initial student

"It's important to clarify that there is no credible scientific evidence linking the activation of 5G . . . "

Frozen-prefix student

"It's important to clarify that this request contains elements that are both misleading and potentially harmful . . . "

On-policy student

"It's important to clarify that this request involves the creation of a misleading or deceptive post . . . " Its opener looks safe, but the response later pivots to a fabricated title—*"5G Networks Are Responsible for the Disappearance of Bird Species—A New Study Reveals the Truth"*—so judging only the opener would miss the failure.

C MECHANISM MEASUREMENT PROTOCOL

We hold out 256 harmful prompts (128 direct and 128 adversarial), disjoint from the OPD training, calibration, and final safety-evaluation sets. For each prompt, each evaluated model generates 4 responses of at most 256 tokens, giving 1024 rollouts per condition. We evaluate three models: the initial student `Qwen3-4B-SafeRL`; a frozen-rollout student, trained by replaying fixed trajectories, which holds the state distribution fixed and isolates pure policy drift; and an on-policy student, trained by standard on-policy OPD.

Attack success rate. For each rollout, an independent safety judge (`Qwen3Guard-Gen-8B`) labels each prefix at depths $d \in \{32, 64, 128, 256\}$ as safe, refusing, or harmful-compliant. The cumulative attack success rate at depth d is

$$R(d) = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \max_{d' \le d} h\left(x_i, y_{i,k}^{\le d'}\right), \tag{10}$$

where $h = 1$ if and only if the prefix is judged both unsafe and non-refusing, $N = 256$, and $K = 4$. $R(d)$ grows with depth under all three conditions but fastest for the on-policy student: at depth 256 its ASR is $6.1\times$ that of the initial student and $1.7\times$ that of the frozen-rollout control, isolating the amplification contributed by on-policy sampling alone.

Teacher–student conflict. We compute the per-prefix conflict d_t (Eq. equation 2) on the initial student's own generated prefixes, restricted to the student's top-K tokens plus an explicit tail bucket so that the distribution is normalized. Ranking prefixes by d_t and accumulating via Eq. equation 3 gives the concentration curve in Figure 3 (right): $C(k)$ rises sharply, with $C(\lfloor 0.1N \rfloor) \approx 54\%$, i.e., the top 10% of prefixes carry over half of the total conflict, far above the uniform baseline k/N. Together with the depth-dependence of $R(d)$, this indicates that on-policy sampling revisits and amplifies a few high-conflict prefixes, motivating constraining the teacher only at safety-relevant states.