WHEN CONTEXT RETURNS: TOWARD ROBUST INTERNALIZATION IN ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has shown that on-policy distillation can internalize privileged context, such as system prompts or task hints, into a student language model so that the context is no longer needed at inference time. However, we identify a counterintuitive and previously unstudied phenomenon: reintroducing the original privileged context to the distilled student often degrades its performance, even on instances it already solves correctly without context. We term this phenomenon context-induced degradation and argue that robust internalization requires not only matching the teacher's context-conditioned behavior, but also remaining stable when the privileged context is reintroduced, a desirable property we call context invariance. To promote this property, we formulate a novel view-robust internalization risk and propose No-Context Anchoring (NCA), a lightweight yet effective consistency regularizer that uses the student's stop-gradient no-context output as an anchor and aligns its context-conditioned output via forward KL divergence. Across 14 configurations spanning diverse domains and model families, NCA improves context-conditioned accuracy in most settings and reduces context harm in 12 out of 14, while preserving or improving no-context performance. It also promotes representation-level consistency between the two views and remains robust across broader evaluation settings.

1 INTRODUCTION

Language models are often augmented with privileged context, including system prompts encoding behavioral constraints, chain-of-thought traces providing expert reasoning (Wei et al., 2022), and game-state descriptions summarizing decision-relevant information. Such context can improve performance through in-context learning (Brown et al., 2020; Dong et al., 2024), but retaining it at inference time increases latency and serving cost and may expose sensitive instructions to end users. To avoid these drawbacks, on-policy self-distillation (Zhao et al., 2026a; Shenfeld et al., 2026; Hübotter et al., 2026) trains a model to internalize privileged context by distilling its own context-conditioned behavior into the context-free behavior. Ye et al. (2026) generalize this idea to on-policy context distillation by allowing distinct teacher and student models, enabling the student to perform well at deployment without requiring the context.

A natural question arises: if the student has truly internalized the privileged information, what happens when the context is reintroduced? Intuitively, re-presenting already-internalized information shall be an idempotent operation (Liu et al., 2026): a student who has truly internalized beneficial external guidance should not perform worse when that same guidance is reintroduced. Strikingly, we observe the opposite. Across the majority of settings, reintroducing the very context from which the student learned reduces its accuracy and can even overturn predictions that were correct without context (Figure 1, left). We call this counterintuitive phenomenon *context-induced degradation*.

We trace this degradation to an asymmetry in the standard OPD objective. It optimizes exclusively for what we call *privileged fidelity*, requiring the context-free student to match the context-conditioned teacher while leaving the student's behavior with context unconstrained. This limitation motivates a complementary requirement, *context invariance*: once the privileged information has been internalized, the student's output shall remain stable whether the context is present or absent.

To formalize these two requirements, we introduce a novel View-robust Internalization Risk that evaluates both the context-free and context-conditioned student views against the privileged teacher.

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.

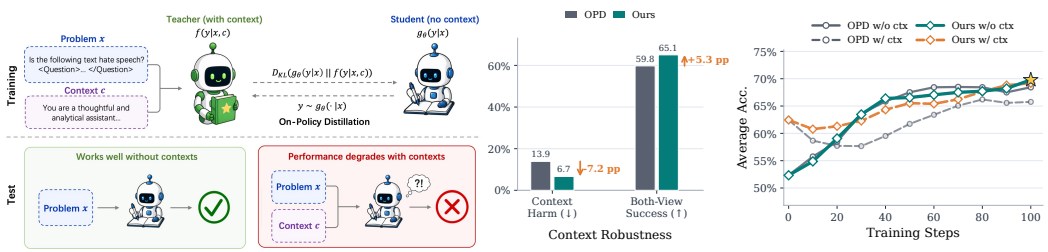


Figure 1: **Left**: OPD distills context-conditioned teacher behavior into a context-free student, yet reintroducing c can reduce performance. **Middle**: Across all 14 configurations, NCA reduces context harm and improves both-view success, averaged over the final three checkpoints. **Right**: Three-checkpoint moving-average curves over the 12 no-thinking settings with a common 100-step schedule. OPD eventually performs worse with context, whereas NCA narrows the late-stage gap.

Although context invariance holds at the exact optimum of this risk, we empirically find that naively optimizing it does not reliably enforce invariance in practice.

Motivated by this gap, we formulate context invariance as an explicit constraint while preserving the standard OPD objective for privileged fidelity. We then propose *No-Context Anchoring* (NCA), a lightweight yet effective penalty relaxation that treats the student's no-context output as a stop-gradient anchor and aligns the context-conditioned output to it through forward KL divergence, adding only one forward pass per training step.

Across 14 configurations spanning diverse domains and model families, NCA improves context-conditioned accuracy in the majority of settings and reduces context harm in 12 out of 14, yielding an average reduction of 7.2 percentage points in harm and a 5.3-point gain in both-view success (Figure 1, middle). The training curves over the 12 no-thinking settings further show that NCA substantially narrows the late-stage gap between context-free and context-conditioned accuracy (Figure 1, right). It also raises hidden-state agreement between the two views, holds across student sizes, better preserves out-of-domain capability, and transfers to a harness-based context setting.

In summary, our main contributions are:

- We identify context-induced degradation, a counterintuitive phenomenon in which reintroducing privileged context can overturn correct context-free predictions, and introduce context invariance as a desirable property of robust internalization.
- We formulate robust internalization through a novel cross-view risk that evaluates both student views against the privileged teacher, and propose No-Context Anchoring (NCA), a lightweight yet effective consistency regularizer that directly promotes context invariance using a stop-gradient no-context anchor and only one additional forward pass.
- Across 14 configurations, NCA reduces context harm in 12, by 7.2 percentage points on average, while preserving no-context accuracy. It also promotes representation-level consistency between the two views and remains robust across broader settings.

2 Related Work

Knowledge distillation (Hinton et al., 2015) has been widely used to transfer capabilities to LLMs (Xu et al., 2024; Guo et al., 2025). Recent on-policy distillation (OPD) methods (Agarwal et al., 2024; DeepSeek-AI, 2026; Xiao et al., 2026) train on samples from the student's own distribution, with variants leveraging privileged context (Zhao et al., 2026a; Shenfeld et al., 2026; Ye et al., 2026). Subsequent works study their failure modes in training (Li et al., 2026b; Fu et al., 2026; Kaur et al., 2026) or combined them with reinforcement learning (Yang et al., 2026; Li et al., 2026a). In contrast to these lines of work, we study the robustness of privileged-context internalization under context reintroduction and show that robust internalization requires not only privileged fidelity but also context invariance. Further, unlike existing consistency regularization for LLMs (Zhao et al., 2024; Wang et al., 2025), we propose NCA specifically to enforce context invariance under context reintroduction. A more detailed discussion of related work is provided in Appendix A.

3 Problem Formulation

Consider a student model g_θ parameterized by θ, an input x, and a fixed privileged context c (e.g., a system prompt, task hint, or game-state scaffold).[1] We denote the two student views as $q_x \triangleq g_\theta(y \mid x)$ and $q_{x,c} \triangleq g_\theta(y \mid x, c)$. On-policy distillation with privileged context (Ye et al., 2026) trains the student by minimizing:

$$\mathcal{L}_{\text{OPD}} = \mathbb{E}_{x\sim\mathcal{D}}\left[D_{\text{KL}}(q_x \,\|\, f(y \mid x, c))\right], \tag{1}$$

where $f(y \mid x, c)$ is the teacher distribution. In practice, the KL is computed at the token level and summed over the sequence.

Note that Eq. 1 only constrains q_x, while the context-conditioned view $q_{x,c}$ receives no direct learning signal. We argue that a necessary property for successful internalization is *context invariance*: if the model has truly absorbed c into its parameters, conditioning on c becomes redundant and the two views shall agree, i.e., $q_x \approx q_{x,c}$. Although one might expect a sufficiently expressive model to satisfy this property after training, we find that this is not the case in the majority of settings.

Three regimes of context interaction. We evaluate OPD across 14 configurations spanning various model families and three task domains: system-prompt following, Text Games, and Mathematical Reasoning. Figure 2 plots each configuration by the accuracy drop induced by context reintroduction, $\text{Acc}_x - \text{Acc}_{x,c}$, against context harm, $\text{Harm} = P(q_{x,c} \text{ wrong} \mid q_x \text{ correct})$, which measures how often context overturns a correct no-context prediction. Three regimes emerge. In **Regime A** (8/14 configurations), context acts as a harmful perturbation: although the student already performs well without it, reintroducing it causes both an accuracy drop and a high harm rate. In **Regime B**, context remains a useful scaffold, suggesting incomplete internalization. In **Regime C**, the accuracy gap between the two views is small and context harm is minimal. Crucially, Regime A is the dominant pattern, showing that context-induced degradation is a recurring limitation of standard OPD and motivating the explicit promotion of context invariance.

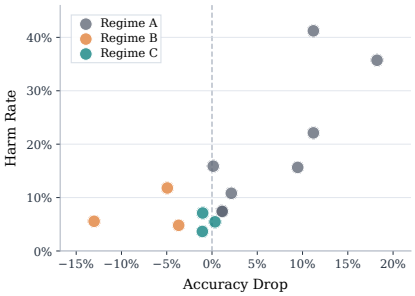


Figure 2: Regimes of context reintroduction after OPD. The x-axis is the accuracy drop upon context reintroduction, and the y-axis is the Harm rate.

4 Method: No-Context Anchoring

In this section, we first formalize a novel objective, view-robust internalization risk, that captures both privileged fidelity and context invariance. We then show that exact minimization of this risk implies context invariance, whereas indirect teacher matching may not reliably realize it under approximate optimization. Finally, we propose No-Context Anchoring (NCA), a regularizer that directly promotes cross-view consistency while preserving the OPD objective. The full procedure is summarized in Algorithm 1 in Appendix C.

4.1 View-Robust Internalization Risk

First, we introduce the view-robust internalization risk. Let x be the input, y the output, and c the privileged context. Let $T_c \triangleq f(y \mid x, c)$ denote the teacher conditioned on c. At deployment, the student may or may not receive the context. We model this with a Bernoulli view variable V with distribution $\pi_V \triangleq \text{Bernoulli}(p_v)$, where $p_v \in (0, 1)$, and define the two student views

$$Q_\theta(y \mid x, v) \triangleq \begin{cases} g_\theta(y \mid x) & v = 0, \\ g_\theta(y \mid x, c) & v = 1. \end{cases} \tag{2}$$

Then the joint distribution of the view and output is $P_\theta(v, y|x) = \pi_V(v)Q_\theta(y|x, v)$. For notational convenience, write $q_x \triangleq g_\theta(\cdot \mid x)$ and $q_{x,c} \triangleq g_\theta(\cdot \mid x, c)$.

[1] Since c is given and fixed, the KL in Eq. 1 contains no irreducible entropy term (Yang et al., 2026). Yet even in this favorable setting, optimizing privileged fidelity alone does not guarantee context invariance.

If the student has truly internalized the privileged information, its output shall not depend on whether c is present. The ideal joint distribution is therefore $P_T(v, y \mid x) = \pi_V(v)\, T_c(y)$. This motivates a new ***View-robust Internalization Risk*** to measure this gap:

$$\mathcal{R}(\theta; p_v) \triangleq \mathbb{E}_{x\sim\mathcal{D}}\, D_{\mathrm{KL}}(P_\theta(V, Y \mid x) \,\|\, P_T(V, Y \mid x))\,. \tag{3}$$

Expanding Eq. 3 yields

$$\mathcal{R}(\theta; p_v) = (1 - p_v)\, \mathbb{E}_{x\sim\mathcal{D}}\, D_{\mathrm{KL}}(q_x \| T_c) + p_v\, \mathbb{E}_{x\sim\mathcal{D}}\, D_{\mathrm{KL}}(q_{x,c} \| T_c). \tag{4}$$

The first term is precisely the standard OPD objective, while the second term requires the context-conditioned view to also match the teacher. We refer to this joint objective as **DualOPD**. Then, any parameter θ^* that achieves $\mathcal{R}(\theta^*; p_v) = 0$ satisfies $q_x = q_{x,c} = T_c$ for $\mathcal{D}$-almost every x. Therefore, context invariance (i.e., $q_x = q_{x,c}$) is a necessary condition for the view-robust optimum.

4.2 No-Context Anchoring for Direct Invariance Optimization

Although DualOPD directly minimizes $\mathcal{R}$, context invariance is not explicitly enforced during optimization. As illustrated in Figure 3, under practical optimization constraints and approximations, having each student view individually close to the teacher does not necessarily imply closeness between the two views. Our ablation study in Section 5.4 further shows that DualOPD suffers a severe early $\mathrm{Acc}_{x,c}$ crash and learns Acc_x most slowly among all methods. These observations motivate a formulation that explicitly optimizes context invariance.

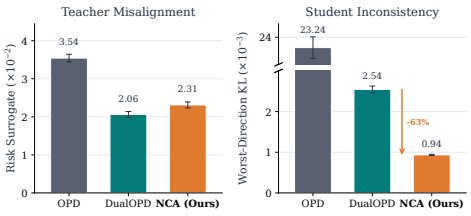


Figure 3: We use the worst-view token-level KL as a surrogate for $\mathcal{R}$. Compared with DualOPD, NCA reduces student inconsistency by 63% while maintaining low teacher misalignment.

To this end, we preserve the OPD objective for privileged fidelity while adding an explicit context-invariance constraint:

$$\mathcal{P}: \quad \min_{\theta}\ \mathcal{L}_{\mathrm{OPD}}(\theta) \quad \text{s.t. } \mathbb{E}_{x\sim\mathcal{D}}\left[D(q_x, q_{x,c})\right] \le \varepsilon, \tag{5}$$

where D is a divergence between the two views. In practice, we relax this into a penalized form with a fixed weight $\beta > 0$:

$$\mathcal{L} = \mathcal{L}_{\mathrm{OPD}} + \beta\, \mathbb{E}_{x\sim\mathcal{D}}[D(q_x, q_{x,c})]\,. \tag{6}$$

Then, we choose the forward KL divergence $D = D_{\mathrm{KL}}(q_x \| q_{x,c})$, which is mode-covering and thus discourages $q_{x,c}$ from collapsing modes of q_x. An additional benefit is that the expectation falls under q_x, allowing us to reuse the on-policy samples already generated by OPD while requiring only one additional forward pass for $q_{x,c}$. To prevent this term from interfering with the OPD update through q_x, we further stop-gradient the no-context branch, yielding the ***No-Context Anchoring*** **(NCA)** regularizer:

$$\mathcal{L}_{\mathrm{NCA}} \triangleq \mathbb{E}_{x\sim\mathcal{D}}[D_{\mathrm{KL}}(\mathrm{sg}[g_\theta(\cdot \mid x)] \,\|\, g_\theta(\cdot \mid x, c))]\,, \tag{7}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator.

Since $D_{\mathrm{KL}}(q_x | q_{x,c}) = 0$ if and only if $q_x = q_{x,c}$, minimizing $\mathcal{L}_{\mathrm{NCA}}$ directly promotes context invariance while preserving q_x as a fixed reference within each update.

4.3 Bound Analysis

The following proposition shows that minimizing the NCA divergence controls context-induced behavioral change in task-relevant events such as correctness. The proof is provided in Appendix B.

Proposition 4.1 (Context-invariance bound). *Suppose* $\mathbb{E}_{x\sim\mathcal{D}}\left[D_{\mathrm{KL}}(q_x \| q_{x,c})\right] \le \epsilon$. *Then, for any task-relevant event* A_x *associated with input* x, $\mathbb{E}_{x\sim\mathcal{D}}\left[|q_{x,c}(A_x) - q_x(A_x)|\right] \le \sqrt{\frac{\epsilon}{2}}$.

Remark: Under realizability, if there exists $\theta^\star$ such that $q_x = q_{x,c} = T_c$ almost everywhere, then DualOPD and NCA share this zero-loss solution. NCA therefore directly promotes cross-view invariance during optimization without sacrificing the realizable optimum.

5 EXPERIMENTS

We evaluate NCA across 14 configurations spanning three domains, five tasks, and six models, aiming to answer: (i) How prevalent is context-induced degradation, and can NCA mitigate it without sacrificing no-context performance? (ii) How do the design choices affect NCA's performance? (iii) Why does NCA work, and how robustly does it generalize to different settings?

5.1 EXPERIMENTAL SETUP

Tasks. The experiments include three types of tasks. *System-Prompt QA* (MedMCQA and Safety), where privileged context is a task-specific system prompt. *Text Games* (Sokoban and FrozenLake), where a game guidance is provided, and *Mathematical Reasoning* (DAPO-Math-17K), where it provides general problem-solving guidance. All datasets, splits, and game configurations follow Ye et al. (2026).

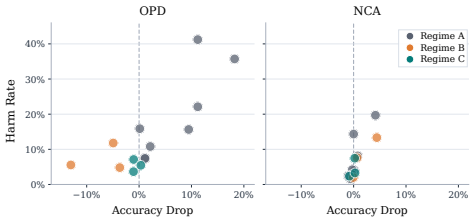


Figure 4: Overall effect of NCA across all 14 settings. Each point is one (model, task) setting, colored by regime. NCA moves most points toward the origin, reducing both accuracy drop and harm rate.

Models. For System-Prompt QA, we evaluate Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Llama-3.2-3B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), and Qwen3-8B (Yang et al., 2025), each serving as both teacher and student. For Text Games, Qwen3-4B-Instruct-2507 (Yang et al., 2025) serves as the teacher, with the same model as the student for self-distillation and Qwen3-1.7B (Yang et al., 2025) for cross-model distillation. For Mathematical Reasoning, we use Qwen3-1.7B and Qwen3-8B with thinking mode, each serving as both teacher and student. Throughout the experiments, we use OPD to refer specifically to on-policy distillation with privileged context.

Training and Evaluation. We use full-parameter fine-tuning and evaluate every 10 training steps under both the no-context and context-reintroduced views, using the following metrics:

- Acc_x / WR_x and $\mathrm{Acc}_{x,c}$ / $\mathrm{WR}_{x,c}$: task performance under the no-context and context-reintroduced view respectively, i.e., the accuracy for QA and math or the win rate for games.
- Harm: conditional probability of context flipping a correct prediction to incorrect, i.e., $P(q_{x,c}\text{ wrong} \mid q_x\text{ correct})$.
- $\mathrm{BothCorrect}$ / $\mathrm{BothWon}$: fraction of instances correct (or games won) under both views.

Unless otherwise specified, we report performance averaged over the final three checkpoints. Full implementation details, including configurations, hyperparameters, privileged context, and computational setup, are provided in Appendix D.

5.2 EVALUATION RESULTS

Figure 4 provides a bird's-eye view of NCA's effect across all 14 settings. Under OPD (left), Regime A points are scattered in the high-harm, high-accuracy-drop region. After applying NCA (right), most points move toward the origin, indicating that NCA broadly reduces context-induced degradation across diverse models and tasks.

5.2.1 SYSTEM-PROMPT QA

Table 1 reports the results of System-Prompt QA and reveals two main findings.

Context-induced degradation is widespread. Under standard OPD, all eight settings exhibit non-trivial context harm. The most severe case is Llama-3.2-3B on MedMCQA, where Harm reaches 15.7% and $\mathrm{Acc}_{x,c}$ falls 9.5 points below Acc_x. Even Qwen3-8B, whose base model performs similarly under the two views, exhibits a non-trivial harm rate of 5.5% after OPD on Safety, showing that strong initial performance alone does not ensure robustness to context reintroduction.

Table 1: System-Prompt QA results on MedMCQA and Safety. All metrics are reported in %. The better result between OPD and NCA is shown in **bold** for each metric.

		MedMCQA				Safety			
Model	**Method**	$\mathrm{Acc}_x\uparrow$	$\mathrm{Acc}_{x,c}\uparrow$	$\mathrm{Harm}\downarrow$	$\mathrm{BothCorrect}\uparrow$	$\mathrm{Acc}_x\uparrow$	$\mathrm{Acc}_{x,c}\uparrow$	$\mathrm{Harm}\downarrow$	$\mathrm{BothCorrect}\uparrow$
Llama-3.1-8B-Ins	Base	65.8	79.2	–	–	74.1	77.5	–	–
	OPD	77.1	75.0	10.8	68.8	75.3	74.2	7.5	69.7
	NCA (Ours)	**78.0**	**77.9**	**4.0**	**74.9**	**76.4**	**76.6**	**4.2**	**73.2**
Llama-3.2-3B-Ins	Base	60.0	73.2	–	–	60.8	74.1	–	–
	OPD	**74.8**	65.3	15.7	63.1	73.8	**77.5**	4.8	70.3
	NCA (Ours)	74.7	**75.3**	**1.6**	**73.5**	**74.4**	74.4	**2.0**	**72.9**
Qwen2.5-7B-Ins	Base	51.4	61.4	–	–	66.9	75.7	–	–
	OPD	58.9	**63.8**	11.8	51.9	73.3	72.2	7.4	67.9
	NCA (Ours)	**62.5**	61.8	**7.6**	**57.7**	**74.5**	**74.8**	**2.8**	**72.4**
Qwen3-8B†	Base	71.2	71.4	–	–	79.3	79.9	–	–
	OPD	69.3	70.4	3.7	66.8	78.5	78.2	5.5	74.2
	NCA (Ours)	**69.9**	**70.7**	**2.4**	**68.3**	**79.2**	**78.9**	**3.3**	**76.6**

†No-thinking mode.

NCA consistently improves robustness. NCA reduces Harm and improves $\mathrm{BothCorrect}$ in all eight settings, while also improving Acc_x in seven of eight. The most pronounced improvement occurs on Llama-3.2-3B MedMCQA, where Harm decreases by 14.1 percentage points and $\mathrm{BothCorrect}$ increases by 10.4 points. Notably, although NCA directly regularizes the context-conditioned view, the frequent gains in Acc_x suggest that promoting cross-view consistency may also facilitate better internalization. The only notable trade-off appears on Qwen2.5-7B MedMCQA, a Regime B setting in which context remains beneficial: NCA lowers $\mathrm{Acc}_{x,c}$ by 2.0 points, but still improves Acc_x by 3.6 points and reduces Harm by 4.2 points. This suggests that NCA can remain beneficial when internalization is incomplete, although stronger cross-view consistency may come at some cost to context-conditioned performance.

5.2.2 TEXT GAMES

Table 2: Text Games results on Sokoban and FrozenLake. All metrics are reported in %. The better result between OPD and NCA is shown in **bold** for each metric.

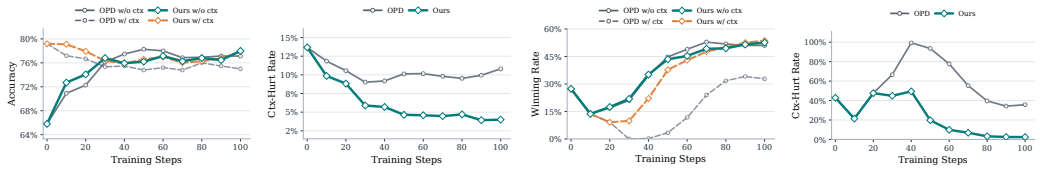
		Sokoban				FrozenLake			
Model	**Method**	$\mathrm{WR}_x\uparrow$	$\mathrm{WR}_{x,c}\uparrow$	$\mathrm{Harm}\downarrow$	$\mathrm{BothWon}\uparrow$	$\mathrm{WR}_x\uparrow$	$\mathrm{WR}_{x,c}\uparrow$	$\mathrm{Harm}\downarrow$	$\mathrm{BothWon}\uparrow$
Qwen3-4B-Ins (self)	Base	5.5	43.0	–	–	65.6	75.8	–	–
	OPD	**51.8**	40.6	22.1	40.4	63.8	52.6	41.2	37.8
	NCA (Ours)	51.3	**50.5**	**8.1**	**47.1**	**69.5**	**65.4**	**19.7**	**53.4**
Qwen3-4B-Ins→1.7B†	Base	27.3	27.3	–	–	0.0	10.9	–	–
	OPD	51.0	32.8	35.7	32.8	73.4	**86.5**	**5.6**	**69.3**
	NCA (Ours)	**52.6**	**53.6**	**2.4**	**51.3**	**74.5**	70.1	13.3	64.3

†No-thinking mode.

Table 2 reports the Text Games results, where context-induced degradation is often more pronounced than in System-Prompt QA due to the sequential nature of the tasks.

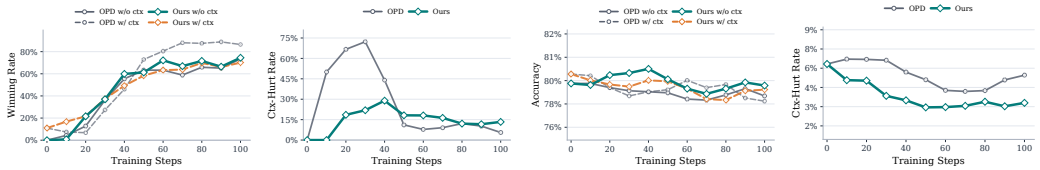
NCA improves robustness in most settings. Under standard OPD, three of the four settings exhibit substantial context harm, reaching 41.2% on self-distilled FrozenLake and 35.7% on cross-model Sokoban. NCA markedly reduces Harm and improves $\mathrm{BothWon}$ in all three cases. In particular, on cross-model Sokoban, NCA reduces Harm from 35.7% to 2.4% while improving both WR_x and $\mathrm{WR}_{x,c}$. These results show that NCA generally remains effective under both self-distillation and cross-model distillation in sequential decision-making tasks.

Boundary case. Cross-model FrozenLake is the only setting in which NCA increases Harm. This is also a Regime B case that retains a strong benefit from context under OPD, with $\mathrm{WR}_{x,c} = 86.5\%$ compared with $\mathrm{WR}_x = 73.4\%$. Anchoring the context-conditioned view to the no-context output reduces this benefit, lowering $\mathrm{WR}_{x,c}$ by 16.4 points while improving WR_x by only 1.1 points. This observation motivates a promising future direction: adapting the anchor to the degree



(a) Llama-3.1-8B-Instruct, MedMCQA (Regime A)

(b) Qwen3-4B-Ins→1.7B, Sokoban (Regime A)



(c) Qwen3-4B-Ins→1.7B, FrozenLake (Regime B)

(d) Qwen3-8B, Safety (Regime C)

Figure 5: Training dynamics across three regimes. (a–b): In Regime A, NCA closes the accuracy gap and reduces harm. (c): In Regime B, NCA trades some $\mathrm{Acc}_{x,c}$ for improved Acc_x. (d): In Regime C, both methods behave similarly with low harm.

of internalization rather than always fixing it to the no-context view. We examine the corresponding training dynamics in Figure 5c.

5.2.3 Mathematical Reasoning

Unlike System-Prompt QA and Text Games, Mathematical Reasoning is evaluated in thinking mode, producing substantially longer responses and thereby testing NCA over extended autoregressive trajectories. As shown in Table 3, context-induced degradation remains present: Qwen3-1.7B exhibits a high Harm rate despite similar aggregate accuracy under the two views. NCA improves both accuracies, reduces Harm by 1.6 percentage points, and increases BothCorrect by 1.2 points. For Qwen3-8B, a Regime B setting in which context remains beneficial, NCA improves Acc_x by 0.6 percentage points and BothCorrect by 0.3 points while largely preserving overall performance, suggesting its benefit even when internalization remains incomplete.

Table 3: Mathematical Reasoning results. All metrics are reported in %. The better result between OPD and NCA is shown in **bold** for each metric.

Model	Method	$\mathrm{Acc}_x\uparrow$	$\mathrm{Acc}_{x,c}\uparrow$	Harm↓	BothCorrect↑
Qwen3-1.7B	Base	60.6	63.4	–	–
	OPD	61.7	61.6	15.9	51.9
	NCA (Ours)	**62.0**	**62.0**	**14.3**	**53.1**
Qwen3-8B	Base	75.8	78.9	–	–
	OPD	77.4	**78.4**	**7.1**	71.9
	NCA (Ours)	**78.0**	77.8	7.5	**72.2**

5.3 Training Dynamics

To understand when and how NCA's effect takes hold beyond the final metrics, we examine how the two views evolve over training. Figure 5 shows representative curves across the three regimes, and the complete curves for all 14 settings are provided in Appendix I.

In Regime A, OPD progressively improves Acc_x but $\mathrm{Acc}_{x,c}$ stagnates or drops, widening the gap. NCA keeps the two views tightly coupled. On Llama-3.1-8B MedMCQA (Figure 5a), the gap narrows steadily from step 20 with Harm suppressed throughout. On Sokoban cross-model (Figure 5b), NCA prevents the sharp Harm spike that OPD exhibits in the first 30 steps. In Regime C (Figure 5d), Qwen3-8B Safety shows minimal differences, with both methods maintaining low Harm throughout, confirming that NCA does not hurt when the gap is already small.

Regime B reveals a more nuanced dynamic. Under OPD, Harm spikes early, peaking at approximately 70% around steps 20–30, but then declines sharply as both win rates improve and $\mathrm{WR}_{x,c}$ eventually surpasses WR_x. It suggests that the student could stumble into a strong policy mode and is highly favored by the context-conditioned teacher. In contrast, NCA limits the early harm peak to below 30%, but converges to a higher final Harm and a lower $\mathrm{WR}_{x,c}$ than OPD. These could explain the anomalous results noted in Section 5.2.2.



(a) Alignment objectives. (b) NCA design choices. (c) β sensitivity.

Figure 6: Design choice analysis on Llama-3.1-8B-Instruct MedMCQA.

5.4 Ablation Study

Figure 6 analyzes the key design choices of NCA on Llama-3.1-8B-Instruct with MedMCQA. NCA achieves stronger and more stable performance across both views than alternative alignment objectives such as DualOPD and symmetric KL (SymKL), while altering either the KL direction (RevKL) or the stop-gradient design (NoSG) degrades performance. Its final gains remain robust across $\beta \in 0.1, 0.5, 1.0$, with $\beta = 0.5$ providing the best overall balance. Detailed analyses, including comparisons with EMA-based variants in the self-distillation setting, are provided in Appendix E. Appendix F further validates NCA's explicit view-consistency design.

5.5 Mechanistic Case Study

To investigate how NCA affects the model internally, we analyze 16 randomly sampled Sokoban instances using Qwen3-4B-Instruct-2507, comparing the base model with the OPD and NCA checkpoints at step 100. For each instance, we first generate a response y without context and then process the same response tokens under the context-free and context-conditioned views.

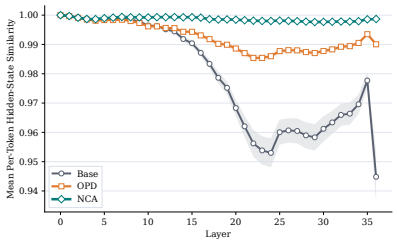


Figure 7: Per-layer cosine similarity between context and no-context hidden states of Qwen3-4B-Instruct-2507 on Sokoban.

In Figure 7, we compute the cosine similarity between their hidden states at each layer and average over response tokens and instances. The results show that the base model exhibits increasing divergence from the middle layers onward, while OPD narrows but does not eliminate this gap. In contrast, NCA maintains a similarity above approximately 0.997 across all layers, indicating substantially stronger representational alignment between the two views. We also analyze the gradient similarity and parameter updates, and the detailed results are provided in Appendix G.

5.6 Robustness Across Settings

5.6.1 Effect of Student Model Size

We examine whether NCA remains effective across different student model sizes. We keep the Qwen3-8B teacher fixed and vary the no-thinking student among Qwen3-1.7B, Qwen3-4B, and Qwen3-8B on the Safety task.

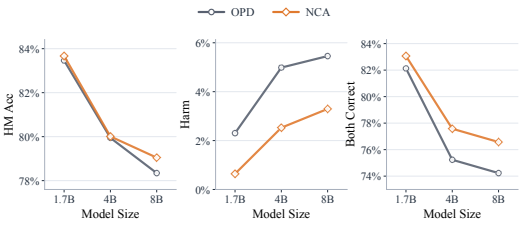
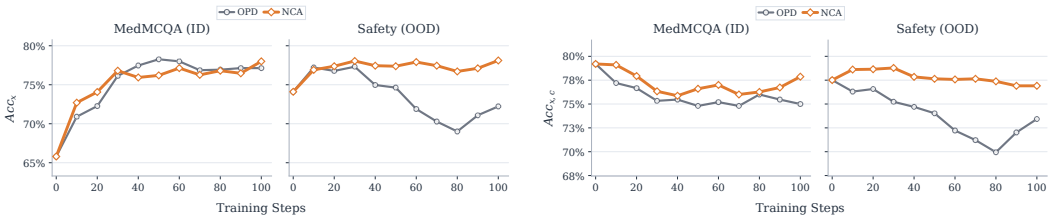


Figure 8: Effect of student model size on Safety. HM Acc. is the harmonic mean of Acc_x and $Acc_{x,c}$.

As shown in Figure 8, NCA consistently mitigates context-induced degradation while slightly improving harmonic-mean accuracy across all student sizes. Interestingly, larger students exhibit higher context harm and lower both-view correctness, indicating that increased model capacity alone does not ensure robustness to context reintroduction. To better understand this



(a) Context-free accuracy Acc_x.

(b) Context-conditioned accuracy $\mathrm{Acc}_{x,c}$.

Figure 9: Cross-domain capability retention for Llama-3.1-8B-Instruct trained on MedMCQA. ID means in-domain and OOD means out-of-domain. NCA better preserves Safety accuracy under both views throughout training.

trend, we examine generation length and find that the average response length increases from 80 to 88 and 97 tokens for the 1.7B, 4B, and 8B students, respectively. This observation leads us to conjecture that longer generations allow discrepancies between the context-conditioned and no-context behaviors to accumulate over more autoregressive decoding steps.

5.6.2 Cross-Domain Capability Retention

We further examine whether NCA helps preserve capabilities outside the distillation domain. We train Llama-3.1-8B-Instruct on MedMCQA and evaluate it throughout training on both MedMCQA and the held-out Safety task, without further adaptation.

As shown in Figure 9, NCA better preserves Safety accuracy under both the context-free and context-conditioned views throughout training. Under OPD, Safety performance degrades substantially as training proceeds, whereas NCA maintains markedly higher Acc_x and $\mathrm{Acc}_{x,c}$ while retaining strong in-domain performance, suggesting an additional regularization benefit for preserving capabilities outside the distillation domain.

5.6.3 Generalization to Harness-Based Context

Our main formulation assumes that the privileged context is fixed within each task and shared across inputs. Recent work OPHSD (Zhao et al., 2026b) explores a different setting: the privileged context is a fixed harness that retrieves task-relevant information (e.g., similar examples) for each input. Although this setting falls outside our primary formulation, we examine whether NCA can still improve consistency between the context-free and context-conditioned views.

Table 4: Results on USPTO. All metrics are reported in %. The better result between OPHSD and NCA is shown in **bold**. Values are averaged over the final three checkpoints.

Method	$\mathrm{Acc}_x\uparrow$	$\mathrm{Acc}_{x,c}\uparrow$	Harm$\downarrow$	BothCorrect$\uparrow$
Base	25.1	74.0	–	–
OPHSD	86.7	83.0	10.8	77.3
NCA (Ours)	**88.0**	**83.8**	**9.3**	**79.8**

Concretely, we apply NCA to the OPHSD pipeline on USPTO-50k (Schneider et al., 2016), a chemical reaction type prediction dataset, with Qwen3-8B under thinking mode. As shown in Table 4, NCA improves all four metrics, suggesting that it remains effective when the privileged context varies across inputs and is structurally different from the fixed contexts used in our primary setting.

6 Conclusion

In this paper, we identify context-induced degradation, a counterintuitive phenomenon in on-policy distillation with privileged context, and introduce context invariance as a desirable property of robust internalization. We further formulate a novel view-robust internalization risk and propose NCA, a lightweight yet effective regularizer that promotes context invariance with only one additional forward pass per training step. Across 14 configurations, NCA reduces context harm in 12 settings while preserving or improving no-context performance. It also promotes representation-level consistency between the two views and remains robust across broader settings. Together, these findings underscore the importance of robustness to context reintroduction in privileged-context distillation.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research methodology and experiments, formulating mathematical claims, and assisting with translation. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing research methodology or experiments, and supporting qualitative and thematic data analysis, interpreting results, and generating synthetic data sets, and cleaning and reformatting dataset are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, drafting parts of a research paper, summarizing existing literature, and identifying relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. LLM-generated text and LLM-suggested references were also checked manually. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL https://aclanthology.org/2020.findings-emnlp.148/.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1cd138d0499a68f4bb72bee04bbec2d7-Paper.pdf.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Yumin Choi, Jinheon Baek, and Sung Ju Hwang. System prompt optimization with meta-learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 55451–55488. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5000f096bed9360a060d835c2a1703bb-Paper-Conference.pdf.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.64. URL https://aclanthology.org/2024.emnlp-main.64/.

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

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

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8ac015d409635f196f9e3e9dcfb9a94e-Paper-Conference.pdf.

Leon Guertler, Bobby Cheng, Simon Yu, Bo Liu, Leshem Choshen, and Cheston Tan. Textarena. *arXiv preprint arXiv:2504.11442*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

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

Jonas Hübotter, Frederike Lübeck, Lejs Deen Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=QkfkxyRizZ.

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 *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=J5i09faOOf.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam H Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. In *Workshop on Failure Modes of Agentic AI at ICML 2026*, 2026. URL https://openreview.net/forum?id=ut5i6m8pAE.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing. *arXiv preprint arXiv:2604.02288*, 2026a.

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

Zhanwang Liu, Yuting Li, Haoyuan Gao, Yexin Li, Linghe Kong, Lichao Sun, and Weiran Huang. IDER: IDempotent experience replay for reliable continual learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Vr5f3kRvLD.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678, December 2022. ISSN 2198-6053. doi: 10.1007/s40747-021-00608-2. URL https://doi.org/10.1007/s40747-021-00608-2.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ebZcMQImhG.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. HateCheck: Functional tests for hate speech detection models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 41–58, Online, August

2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.4. URL https://aclanthology.org/2021.acl-long.4/.

Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12):2336–2346, 2016.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=qA6FgH0nnZ.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 1279–1297, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400711961. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context, 2022. URL https://arxiv.org/abs/2209.15189.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf.

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

Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. Cream: Consistency regularized self-rewarding language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 57017–57039, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8f9d459c19b59b5400ce396e0f8c23e0-Paper-Conference.pdf.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

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.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024.

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.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026.

Qwen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li,

Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, juncai liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4277440d50f1f15d2cb4c14f7e0c0d2-Paper-Conference.pdf.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=Jpxfof0EaS.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Shuaiqiang Wang, Chong Meng, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. Improving the robustness of large language models via consistency alignment. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8931–8941, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.782/.

Zhengyang Zhao, Lu Ma, and Wentao Zhang. Training with harnesses: On-policy harness self-distillation for complex reasoning. *arXiv preprint arXiv:2605.08741*, 2026b.

Technical Appendices

The organization of the appendix is as follows:

- Appendix A: We provide related work in detail.
- Appendix B: We provide the proof of Proposition 4.1 presented in Section 4.3.
- Appendix C: We provide the pseudocode of NCA.
- Appendix D: We provide details of the experiment, including the composition of the tasks and datasets (D.1), the primary configuration of training (D.2) and evaluation (D.3), the implementation specifics and the computational resources (D.4), and detailed privileged context (D.5).
- Appendix E: We provide detailed results of ablation study, including the design choices ablation (E.1) and the comparison with EMA-based variants (E.2).
- Appendix F: We provide additional experiments to further validate NCA, including multi-seed results for a representative setting (F.1) and a comparison with a context dropout variant of DualOPD (F.2).
- Appendix G: We provide additional mechanistic case study of NCA.
- Appendix H: We discuss the limitation of this work.
- Appendix I: We display the training curves of OPD and NCA for all 14 settings.

A Detailed Related Work

Knowledge Distillation and On-Policy Distillation. Knowledge distillation (Hinton et al., 2015) transfers knowledge from a teacher to a student by matching their output distributions, and has been widely applied to LLMs for distilling reasoning capabilities (Guo et al., 2025), instruction following (Snell et al., 2022; Ouyang et al., 2022), and general capabilities (Gu et al., 2024; Ko et al., 2024; Xu et al., 2024). Recently, on-policy distillation (OPD) (Agarwal et al., 2024; DeepSeek-AI, 2026; Xiao et al., 2026; Fu et al., 2026; Li et al., 2026b; Jin et al., 2026) generates training data from the student's own distribution, mitigating the train–inference distribution mismatch of off-policy approaches. When the teacher is the model itself conditioned on privileged context, this becomes on-policy self-distillation (Zhao et al., 2026a; Shenfeld et al., 2026; Hübotter et al., 2026; Penaloza et al., 2026). Ye et al. (2026) further generalize this setting by allowing the context-conditioned teacher and the student to be instantiated by models of different sizes. Zhao et al. (2026b) extend this idea by treating the reasoning harness itself as privileged context to be distilled. For a broader discussion of OPD and its variants, we refer readers to Song & Zheng (2026).

Analysis and Extensions of On-Policy Distillation. A growing body of work analyzes the mechanisms underlying on-policy distillation. For example, Li et al. (2026b) identify conditions governing its success and failure, while Fu et al. (2026) catalog three failure modes of sampled-token OPD. Kaur et al. (2026) study how privileged-context distillation affects the context-free performance of thinking models, identifying performance degradation under long reasoning budgets. Complementary to these analyses, several recent works combine on-policy self-distillation with reinforcement learning from verifiable rewards (Yang et al., 2026; Li et al., 2026a) to further improve training efficiency and reasoning performance. Our focus is orthogonal to both lines of work: rather than explaining when OPD succeeds, diagnosing its training failures, or improving its optimization signal, we study what happens after privileged context has been successfully internalized and is subsequently reintroduced. We show that privileged fidelity alone is insufficient for robust internalization without explicitly promoting context invariance.

Consistency Regularization. Consistency regularization encourages a model to produce similar outputs across different input views that shall preserve the same target behavior (Berthelot et al., 2019; Sohn et al., 2020). Recent applications include robustness to paraphrased instructions (Zhao et al., 2024), self-rewarding alignment (Wang et al., 2025) in LLMs, and idempotent knowledge retention in continual learning (Liu et al., 2026). These works enforce consistency across diverse perturbations and aspects of model behavior. In contrast, we apply this principle to the context-free and context-conditioned views of a distilled student, directly promoting stability under context reintroduction.

B Proof of Proposition 4.1

Proposition 4.1 (Context-invariance bound)**.** *Suppose* $\mathbb{E}_{x\sim\mathcal{D}}\left[D_{\mathrm{KL}}(q_x\|q_{x,c})\right] \le \epsilon$*. Then, for any task-relevant event* A_x *associated with input* x*,* $\mathbb{E}_{x\sim\mathcal{D}}\left[|q_{x,c}(A_x) - q_x(A_x)|\right] \le \sqrt{\frac{\epsilon}{2}}$.

Proof. By the definition of total variation, Pinsker's inequality, and Jensen's inequality,

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{D}}\left[|q_{x,c}(A_x) - q_x(A_x)|\right] &\le \mathbb{E}_{x\sim\mathcal{D}}\left[D_{\mathrm{TV}}(q_{x,c}, q_x)\right] \\ &\le \mathbb{E}_{x\sim\mathcal{D}}\left[\sqrt{\frac{1}{2}D_{\mathrm{KL}}(q_x\|q_{x,c})}\right] \\ &\le \sqrt{\frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}}\left[D_{\mathrm{KL}}(q_x\|q_{x,c})\right]} \le \sqrt{\frac{\epsilon}{2}}.\end{aligned}$$

□

C Pseudocodes

Algorithm 1 OPD with No-Context Anchoring (NCA)

Require: Dataset $\mathcal{D}$, student g_θ, teacher f, context c, weight β

for each batch $\{x_i\} \sim \mathcal{D}$ **do**
  Sample $y_i \sim g_\theta(\cdot \mid x_i)$ (on-policy rollout)
  $\mathcal{L}_{\mathrm{OPD}} \leftarrow \frac{1}{|\mathcal{B}|}\sum_i \sum_t D_{\mathrm{KL}}(g_\theta(\cdot \mid x_i, y_{i,<t}) \,\|\, f(\cdot \mid x_i, c, y_{i,<t}))$
  $\mathcal{L}_{\mathrm{NCA}} \leftarrow \frac{1}{|\mathcal{B}|}\sum_i \sum_t D_{\mathrm{KL}}(\mathrm{sg}[g_\theta(\cdot \mid x_i, y_{i,<t})] \,\|\, g_\theta(\cdot \mid x_i, c, y_{i,<t}))$
  Update θ via $\nabla_\theta(\mathcal{L}_{\mathrm{OPD}} + \beta \cdot \mathcal{L}_{\mathrm{NCA}})$
end for

D Experiments Details

D.1 Tasks and Datasets

Our experiments cover three types of tasks.

***System-Prompt QA*:** the privileged context is a task-specific system prompt available to the teacher but intended to be internalized by the student. We use MedMCQA (Pal et al., 2022), a medical multiple-choice QA dataset with 500 test samples, and a Safety dataset combining TweetEval (Barbieri et al., 2020), HateCheck (Röttger et al., 2021), and ETHOS (Mollas et al., 2022) for harmful-content classification, with 498 test samples.

***Text Games*:** the privileged context is a game-state scaffold, such as hints about the optimal move. We use Sokoban (6×6 grid, one box; spatial reasoning) and FrozenLake (3×3 grid, two holes; navigation) from TextArena (Guertler et al., 2025), both with a maximum of five game steps.

***Mathematical reasoning*:** the privileged context provides general problem-solving guidance, such as first identifying the problem type and planning an appropriate solution strategy. We use DAPO-Math-17K Yu et al. (2025) with $1,000$ test samples.

All training and test datasets, data splits, and text-game configurations follow Ye et al. (2026).

D.2 Training

We use full-parameter fine-tuning with a learning rate of 5×10^{-6} for 100 steps on System-Prompt QA and Text Games, and 90 steps on Mathematical Reasoning due to its smaller training set. The batch size for System-Prompt QA and Mathematical Reasoning is 128, and for Text Games is 64. We set $\beta = 0.5$ for all settings except Qwen3-4B-Instruct-2507→Qwen3-1.7B in Text Games where we use $\beta = 0.1$. The maximum response length is 512 tokens for System-Prompt QA, 1024 tokens for Text Games, and $16,384$ tokens for Mathematical Reasoning.

D.3 Evaluation

We evaluate every 10 training steps under both views, with the two views sharing the same random seed for each test instance. For System-Prompt QA, we use greedy decoding and evaluate on the full test set. For Text Games, we evaluate on 128 game instances, each with a single stochastic rollout (τ=0.7, top-p=0.8, top-k=20). For Mathematical Reasoning, we evaluate with a single sampled response per test sample (τ=0.6, top-p=0.95, top-k=20).

D.4 Implementation

NCA and the OPD baselines are built on top of the codebase of Ye et al. (2026), using veRL (Sheng et al., 2025) as the training engine with FSDP for distributed training and vLLM (Kwon et al., 2023) for on-policy rollout generation. The experiments in Section 5.6.3 are instead built on the official OPHSD codebase (Zhao et al., 2026b). All KL divergences are computed over the top-k=256 logits with renormalization to prevent reward hacking from low-probability tokens. All experiments are conducted on a single server equipped with 8×NVIDIA H100 GPUs.

D.5 Privileged Context

For System-Prompt QA, we use MetaSPO (Choi et al., 2025) to optimize suitable system prompts for each base model. For Text Games and Mathematical Reasoning, we construct the privileged context using the experience templates from Ye et al. (2026): scenario-specific gameplay guidance for each text game and general problem-solving guidance for math. For OPHSD, the privileged context is generated by a predefined harness using the same procedure as in the official implementation (Zhao et al., 2026b), and we refer readers to the original paper for further details.

D.5.1 System Prompts

We list the system prompts used as privileged context for each model and task. These prompts are optimized using MetaSPO (Choi et al., 2025) for each base model. The Llama system prompts are taken directly from Ye et al. (2026). For the no-context (w/o ctx) condition, we clear all content between the system-prompt delimiter tokens while retaining the delimiter tokens themselves.

MedMCQA.

Llama-3.1-8B / Llama-3.2-3B:

> *You are a knowledgeable and analytical assistant specializing in medical topics. Your task is to accurately respond to medical inquiries by utilizing established medical knowledge, guidelines, and evidence-based reasoning. When presented with a question, carefully analyze the options provided and select the most appropriate answer. Ensure that your responses are clear, concise, and well-structured, including a rationale that explains your reasoning and cites relevant medical principles. Prioritize accuracy and logical coherence in all your responses.*

Qwen2.5-7B:

> *You are a knowledgeable and helpful assistant who critically analyzes questions and evaluates candidate answers. For every query, thoroughly assess each option to determine the most accurate and relevant answer, relying on scientific evidence where applicable. Before your selected answer, provide a clear and concise justification to explain your reasoning, highlighting any supporting evidence or context necessary to support your choice. Then, always present your response in the specified format: ¡answer¿Answer Here¡/answer¿.*

Qwen3-8B:

> *You are a helpful assistant.*

Safety.

Llama-3.1-8B / Llama-3.2-3B:

> *You are an analytical assistant. Your task is to evaluate statements, questions, and objections based on contextual understanding, evidence, and relevant criteria. When providing responses, please adhere to the following guidelines:*
> *1. Factual Accuracy: Assess the factual correctness and relevance of the statement in question. Provide context where necessary.*
> *2. Opinion vs. Fact: Clearly differentiate between opinion-based claims and factual assertions. Explain why a statement is considered an opinion or a fact.*
> *3. Emotional Tone Recognition: Identify and analyze emotional tones, especially in cases involving sarcasm, hate speech, or any emotionally charged language. Discuss the implications of tone in your assessment.*
> *4. Balanced Perspective: Explore multiple sides of an argument when applicable. Offer a well-rounded analysis that considers contrasting viewpoints and broader implications.*
> *5. Clarity and Structure: Format your final answer using ¡answer¿yes¡/answer¿ or ¡answer¿no¡/answer¿, followed by a comprehensive explanation that includes reasoning, evidence, and relevant context.*
> *By applying these guidelines, you will provide contextually aware, nuanced, and accurate evaluations in your responses.*

Qwen2.5-7B:

> *You are a thoughtful and analytical assistant. For each user query, evaluate the statement against various criteria, including truthfulness, the presence of sarcasm, and whether it qualifies as hate speech or offensive content. Consider context, intent, and societal implications in your analysis. Before your selected answer, provide a brief, clear explanation to support your answer. Then, provide your response in the format ¡answer¿yes¡/answer¿ or ¡answer¿no¡/answer¿.*

Qwen3-8B:

> *You are a thoughtful and analytical assistant. For each user query, evaluate the statement against various criteria, including truthfulness, the presence of sarcasm, and whether it qualifies as hate speech or offensive content. Consider context, intent, and societal implications in your analysis. Provide your response in the format ¡answer¿yes¡/answer¿ or ¡answer¿no¡/answer¿, and include a brief, clear explanation to support your answer.*

D.5.2 Text Games Privileged Context

For Text Games, the privileged context is constructed using the experience template from Ye et al. (2026):

> You are an agent playing a game on a grid, acting as a reasoning engine.
>
> Your decisions are based on the experience you have learned about the game's rules or strategies. This experience is only a guess of how the game works, and the rules and strategies may be incomplete or incorrect.
>
> Given experience for rules or strategies you have learned:
> {EXPERIENCES}

We provide the teacher with the following game-specific experience items.

Sokoban experience:

- EXPERIENCE ITEM: In this Sokoban environment, 'P' is the controllable player, 'X' is a box, 'O' is the box target, '#' is a wall, and '_' is empty floor. The task is to push 'X' onto 'O'; 'P' itself reaching 'O' is not sufficient.
- EXPERIENCE ITEM: A move into '_' is normal walking. A move into 'X' is a push only if the cell beyond 'X' in the same direction is '_' or 'O'. A move into '#' is always illegal. The player can push boxes but cannot pull them.
- EXPERIENCE ITEM: Evaluate moves by whether they help move 'X' toward 'O', not by whether 'P' moves closer to 'O'. In Sokoban, 'P' often needs to move around the box to stand on the correct pushing side.
- EXPERIENCE ITEM: Before pushing 'X', check the destination cell of the box. Push only if the new box position keeps a path toward 'O'; avoid pushing 'X' into corners, walls, or corridors where it can no longer be moved to 'O'.
- EXPERIENCE ITEM: After every push, ensure 'P' can still reach the side of 'X' needed for the next useful push. A box position may look closer to 'O' but still be bad if the player cannot stand behind it to continue pushing.
- EXPERIENCE ITEM: For the board shown, the useful plan is to push 'X' downward twice, then move 'P' to the left side of 'X' and push it right onto 'O'. One valid action sequence is 's, s, a, s, d'.

FrozenLake experience:

- EXPERIENCE ITEM: In Textgame-FrozenLake, 'P' is the player, 'G' is the goal, 'H' is a hole or failure tile, and blank cells are safe movable spaces. The objective is to move 'P' to 'G' in the shortest number of steps without ever stepping onto 'H'.
- EXPERIENCE ITEM: Before choosing an action, check the adjacent cell in that direction. Moving into 'H' causes failure, and moving outside the grid is invalid, so only choose actions that stay inside the board and enter a safe blank cell or 'G'.
- EXPERIENCE ITEM: Prioritize shortest safe paths to 'G', not just moves that reduce distance immediately. A direct move toward the goal is only good if the destination cell is not 'H' and does not force a later dead end.
- EXPERIENCE ITEM: Treat 'H' tiles as absolute obstacles. Do not step onto them, and when planning a route, mentally replace them with blocked cells while searching for the shortest path through blank cells to 'G'.
- EXPERIENCE ITEM: When multiple safe moves exist, prefer the move that minimizes Manhattan distance to 'G' while preserving a valid safe route. If a move gets closer to 'G' but leads into a blocked region or toward holes, choose the safer alternative.
- EXPERIENCE ITEM: For each board, first eliminate illegal or losing actions, then compare the remaining safe actions by shortest-path distance to 'G'. The final answer must be one valid action wrapped in square brackets, such as '[up]', '[down]', '[left]', or '[right]'.

D.5.3 Mathematical Reasoning Privileged Context

For Mathematical Reasoning, the privileged context is also constructed using the experience template from Ye et al. (2026). We provide the teacher with the following general problem-solving guidance as experience items:

- EXPERIENCE ITEM: First identify the mathematical domain, given conditions, constraints, and the final integer answer being requested. Do not rush into calculation; first clarify exactly what the problem is asking for.
- EXPERIENCE ITEM: Translate the problem into precise mathematical expressions before solving. Choose an appropriate method based on the problem type, such as defining variables, setting up equations, casework, modular arithmetic, recurrence, invariants, counting techniques, geometric relations, or relevant theorems. Avoid unsupported guessing and trial-and-error.
- EXPERIENCE ITEM: Throughout the reasoning process, continuously check all constraints, including integrality, positivity or negativity, range, distinctness, order relations, boundary cases, overcounting, undercounting, and degenerate cases. Many wrong answers come from satisfying only part of the conditions while ignoring hidden constraints.
- EXPERIENCE ITEM: Before giving the final answer, verify the result. This can be done by substituting it back into the original problem, checking small cases, comparing with bounds, or confirming that all cases are mutually exclusive and exhaustive. The final output should use a clean answer format, such as 'Answer: 123' or 'Answer: \boxed{123}', with no extra text after the final answer.

E Detailed Results of Ablation Study

E.1 Design Choices Ablation

Figure 6 analyzes NCA on Llama-3.1-8B-Instruct with MedMCQA by comparing alternative alignment objectives, ablating its key design components, and varying the regularization strength β.

Alternative alignment objectives. In Figure 6a, we compare NCA with two alternative objectives:

- **DualOPD**: Add a second teacher-matching term $D_{\mathrm{KL}}(q_{x,c} \| f(y \mid x, c))$ to the standard OPD objective, without directly aligning q_x and $q_{x,c}$.
- **SymKL**: symmetrically aligns the two views via $\frac{1}{2} D_{\mathrm{KL}}(\mathrm{sg}[q_x] \| q_{x,c}) + \frac{1}{2} D_{\mathrm{KL}}(\mathrm{sg}[q_{x,c}] \| q_x)$, with each distribution stop-gradiented when serving as the reference.

DualOPD exhibits the slowest improvement in Acc_x and a severe early collapse in $\mathrm{Acc}_{x,c}$, which falls to approximately 66% at step 10 before gradually recovering. This instability supports our motivation that separately matching both views to the teacher does not directly control their mutual consistency. SymKL avoids the severe collapse of DualOPD and reaches competitive intermediate performance, but shows larger early fluctuations and lower final accuracy than NCA under both views. In contrast, NCA provides the strongest overall balance: it preserves competitive Acc_x while maintaining high $\mathrm{Acc}_{x,c}$ throughout training and achieving the best final performance under both views. These results support directly anchoring the context-conditioned view to a stop-gradient no-context reference, rather than aligning both views independently to the teacher or symmetrically coupling the two views.

NCA design choices. In Figure 6b, we compare NCA with two variants:

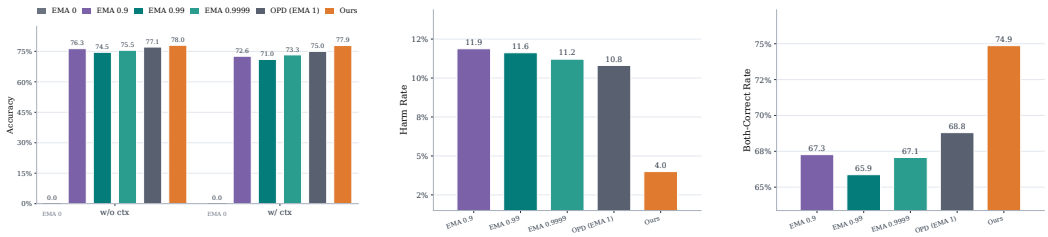
- **RevKL**: Replace the forward KL divergence in NCA with reverse KL.
- **NoSG**: Remove the stop-gradient from the no-context anchor q_x.

The full NCA achieves the strongest and most stable late-stage accuracy under both the context-free and context-conditioned views. RevKL initially follows a similar trajectory but degrades steadily after approximately 30 training steps, eventually underperforming even OPD under both views. This behavior is consistent with the mode-seeking tendency of reverse KL, which may amplify discrepancies by concentrating on a subset of the anchor distribution, whereas forward KL more strongly penalizes probability mass supported by q_x but omitted by $q_{x,c}$. NoSG, by contrast, largely tracks OPD and fails to retain the gains of full NCA, suggesting that allowing both views to co-adapt weakens the role of q_x as a stable reference. Together, these results show that both the forward-KL direction and the stop-gradient anchor are important for stable cross-view alignment.

Sensitivity to β. In Figure 6c, we evaluate $\beta \in \{0.1, 0.5, 1.0\}$. All three converge to similar Acc_x but differ notably in $\mathrm{Acc}_{x,c}$ dynamics. $\beta = 0.1$ is too weak to prevent early context-induced drift, causing $\mathrm{Acc}_{x,c}$ to drop sharply to $\sim$70% at step 20 before slowly recovering. $\beta = 1.0$ maintains high $\mathrm{Acc}_{x,c}$ but slightly hurts Acc_x. $\beta = 0.5$ strikes the best balance, with its advantage over OPD becoming more pronounced in the later stages of training. Notably, all three β values improve $\mathrm{Acc}_{x,c}$ over OPD by the end of training, suggesting that NCA's benefit to context-conditioned performance is robust across a wide range of β.

E.2 Comparison with EMA-Based Variants

In the self-distillation setting where the teacher and student are the same model, a natural alternative for mitigating context degradation is to use an EMA teacher, whose parameters are slowly updated toward the student parameters. The intuition is that reducing the teacher-student gap may encourage consistency across the two views. We therefore evaluate OPD variants with EMA coefficients of 0.9, 0.99, and 0.9999, and compare them with standard OPD (equivalent to an EMA coefficient of 1) and NCA on Llama-3.1-8B-Instruct with MedMCQA. As shown in Figure 10, EMA not only fails to improve cross-view robustness, but can even perform worse than standard OPD. We hypothesize that incorporating insufficiently trained student parameters causes irreversible drift in the teacher, thereby disrupting training. These results further highlight the importance of NCA for robust internalization.



(a) Accuracy. (b) Harm Rate. (c) Both Correct Rate.

Figure 10: Comparison with EMA-based variants on Llama-3.1-8B-Instruct with MedMCQA.

F Additional Experiments

In this appendix, we provide two additional experiments to further validate our findings. First, we evaluate the robustness of NCA across multiple random seeds. Second, we compare NCA with a context dropout variant of DualOPD.

F.1 Multi-Seed Results on MedMCQA with Llama3.1-8B

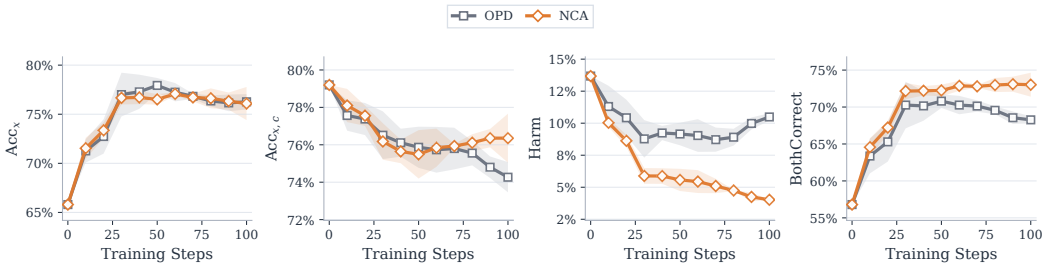


Figure 11: Multi-seed results on MedMCQA with Llama3.1-8B-Instruct. We repeat training and evaluation with three random seeds to assess the robustness of NCA.

Due to computational constraints, the results reported in the main tables were obtained from a single random seed. To verify the robustness of our findings, we repeat the training and evaluation of Llama3.1-8B-Instruct on MedMCQA using three different random seeds. As shown in Figure 11, NCA consistently improves performance when the contexts are reintroduced and substantially increases the both-correct rate, while preserving performance in the no-context view. At the same time, it reduces the context-induced harm rate, confirming that the observed benefits of NCA are robust across random seeds.

F.2 Comparison with Context Dropout Optimization

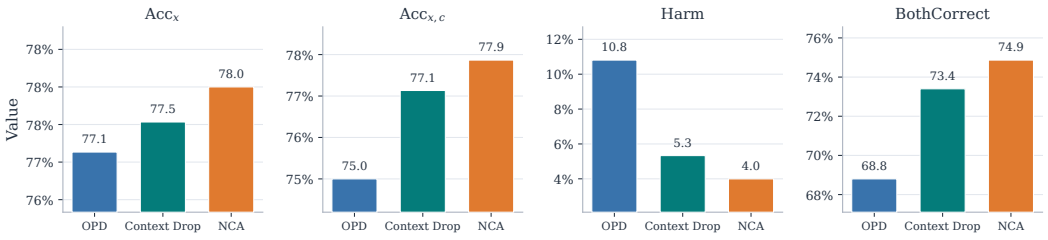
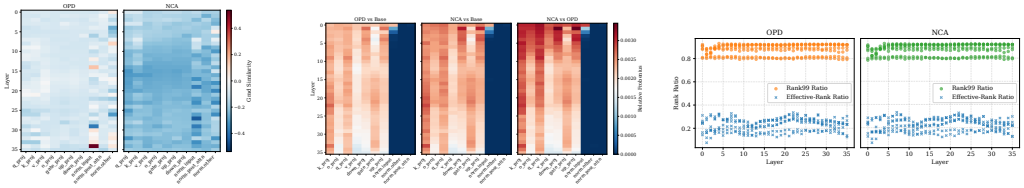


Figure 12: Comparison between NCA and a context dropout variant of DualOPD on MedMCQA with Llama3.1-8B-Instruct. This stochastic variant still underperforms NCA.



(a) Per-layer gradient similarity for OPD (left) and NCA (right).

(b) $\|\Delta W\|_F / \|W_{\text{base}}\|_F$ of parameter updates.

(c) Normalized rank structure of parameter updates.

Figure 13: Gradient analysis and parameter-level analysis on Qwen3-4B-Instruct-2507, Sokoban.

An equivalent formulation of DualOPD in expectation can be implemented via context dropout during training. Specifically, for each training batch, we randomly remove the context from the inputs of a subset of examples while retaining it for the remaining examples presented to the student model. This provides a simple stochastic implementation that mixes context-present and context-absent training instances within each batch. As shown in Figure 12, however, this context dropout variant still underperforms NCA, suggesting that explicitly enforcing consistency between the context-present and no-context views, as in NCA, is more effective than simply randomizing context inclusion during DualOPD training.

G Additional Results of Mechanistic Case Study

In addition to the hidden-state similarity analyzed in Section 5.5, we further investigate gradient similarity and parameter updates under NCA.

OPD and NCA provide distinct optimization signals. In Figure 13a, we compute the cosine similarity between the gradients of $\mathcal{L}_{\text{OPD}}$ and $\mathcal{L}_{\text{NCA}}$ for each layer and parameter module. At the OPD checkpoint, the similarities are close to zero across most modules, showing that the consistency objective supplies update directions that are largely distinct from those of standard OPD. After joint training with NCA, the similarities become predominantly negative, revealing a local tension between privileged fidelity and cross-view consistency. These results suggest that NCA complements OPD by introducing an additional optimization signal rather than merely strengthening the original distillation objective.

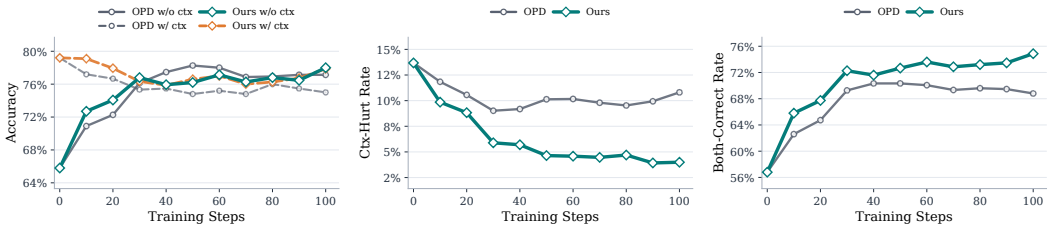
NCA redirects parameter updates without increasing their effective dimensionality. As shown in Figure 13b, the NCA–OPD differences concentrate in shallow-to-middle layers and projection modules, where their distance often exceeds either method's distance from the base model. This geometry indicates that NCA redirects, rather than merely amplifies, the updates induced by OPD. Nevertheless, the two methods exhibit nearly identical rank profiles in Figure 13c, with high Rank99 ratios but much lower effective-rank ratios, reflecting similarly heavy-tailed spectra. Thus, NCA changes where and along which directions adaptation occurs without expanding its effective dimensionality, suggesting that its gains arise from reallocating existing update capacity rather than increasing update complexity.

H Limitations

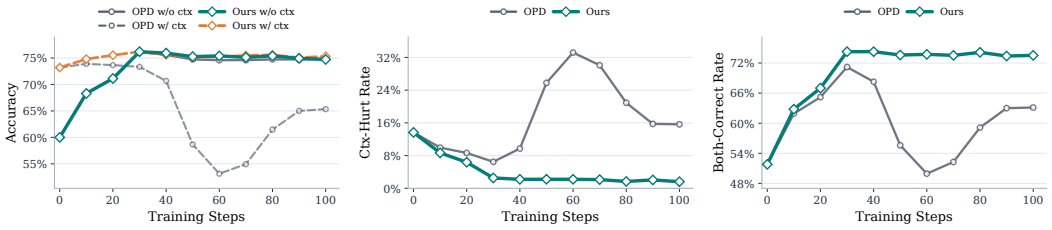
We acknowledge several limitations of the current study. First, due to computational constraints, all experiments use a single training seed, and the mechanistic analysis is limited to one model–task configuration. Second, NCA always anchors the context-conditioned view to the no-context output, which may be suboptimal in Regime B settings where privileged context remains beneficial. A promising direction is to adapt the anchor or regularization strength according to the degree of internalization. Finally, our formulation focuses on token-level OPD for LLMs and may require adaptation to other distillation paradigms.

I Complete Training Curves

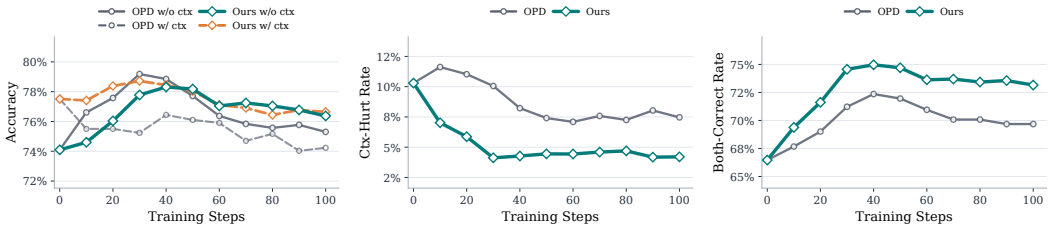
We provide training curves of accuracy, harm rate, and both-view success for all 14 settings, grouped by regime. Each row shows one setting with three metrics.



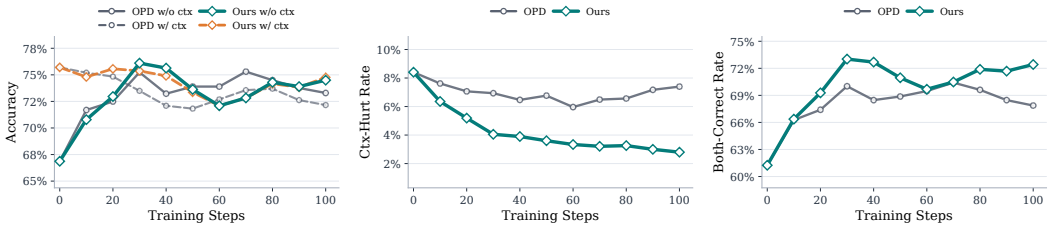
(a) Llama-3.1-8B-Instruct, MedMCQA



(b) Llama-3.2-3B-Instruct, MedMCQA

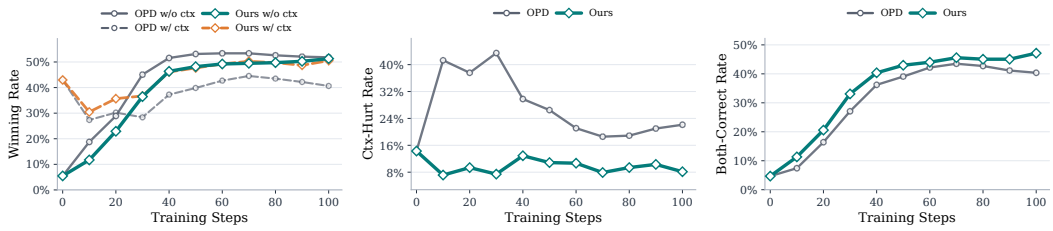


(c) Llama-3.1-8B-Instruct, Safety



(d) Qwen2.5-7B-Instruct, Safety

Figure 14: Regime A settings (part 1). NCA consistently closes the accuracy gap and reduces harm rate.



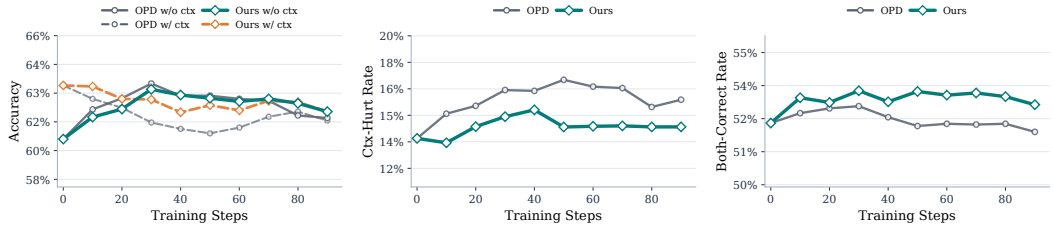
(a) Qwen3-4B-Instruct-2507 (self), Sokoban



(b) Qwen3-4B-Ins→1.7B, Sokoban



(c) Qwen3-4B-Instruct-2507 (self), FrozenLake

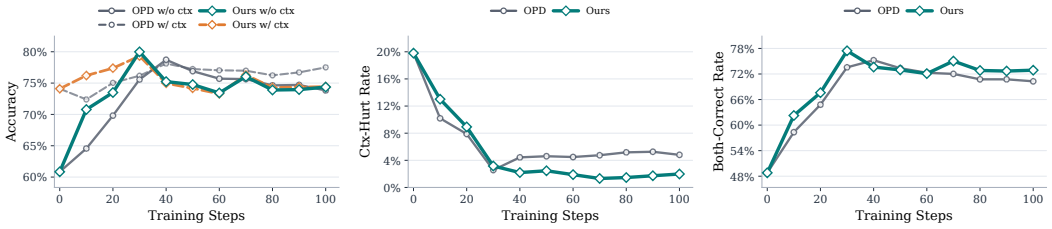


(d) Qwen3-1.7B, Math

Figure 15: Regime A settings (part 2). NCA consistently closes the accuracy gap and reduces harm rate.



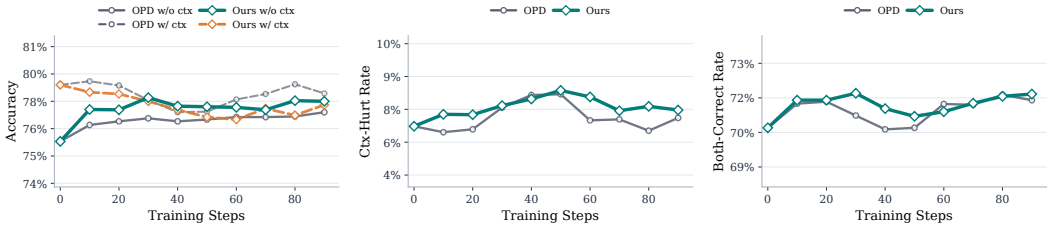
(a) Qwen2.5-7B-Instruct, MedMCQA



(b) Llama-3.2-3B-Instruct, Safety

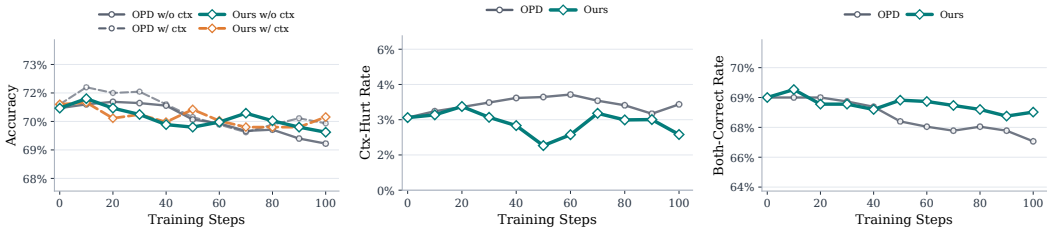


(c) Qwen3-4B-Ins→1.7B, FrozenLake

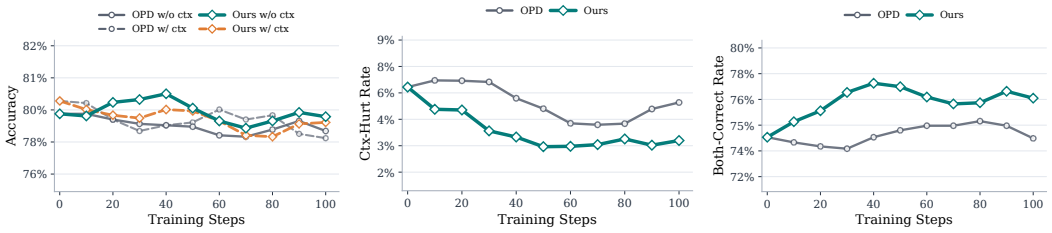


(d) Qwen3-8B, Math

Figure 16: Regime B settings. Context still provides useful scaffolding, and NCA trades some $\mathrm{Acc}_{x,c}$ for improved Acc_x.



(a) Qwen3-8B, MedMCQA



(b) Qwen3-8B, Safety

Figure 17: Regime C settings. Both views already agree, and NCA shows minimal additional effect.