WHEN DOES ONLINE IMITATION LEARNING HELP IN LLM POST-TRAINING? THE ROLE OF (NON-)REALIZABILITY BEYOND HORIZON

Huaqing Zhang[*1] **Jingchu Gai**[*2] **Juno Kim**[3] **Bingbin Liu**[4] **Andrej Risteski**[2]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University
[2]Machine Learning Department, Carnegie Mellon University
[3]Department of Electrical Engineering and Computer Sciences, UC Berkeley
[4]Kempner Institute, Harvard University
zhanghq22@mails.tsinghua.edu.cn, jgai@cs.cmu.edu, junokim@berkeley.edu, bliu@g.harvard.edu, aristesk@cs.cmu.edu

ABSTRACT

Online imitation learning (IL), particularly on-policy distillation with a reverse-KL objective, has emerged as a strong approach for LLM post-training, often outperforming offline baselines such as supervised fine-tuning (SFT). However, a principled understanding of when and why online interaction is beneficial remains unclear. In this work, we show that the benefits of online interaction depend critically on whether the setting is realizable, i.e., whether the student policy class can represent the expert policy. Under realizability, we empirically show that offline IL already matches the expert's performance, challenging the common explanation that online IL benefits from mitigating error accumulation. In contrast, in non-realizable (misspecified) settings, we prove that offline IL encounters an information-theoretic bottleneck even when horizon $H = 1$, and propose a structural characterization of misspecification relative to the reward, under which online IL provably achieves high performance despite large expert–student discrepancy.

1 INTRODUCTION

Imitation Learning (IL) broadly refers to learning a student policy from expert demonstrations, enabling the student to achieve strong performance without task-specific reward design. For post-training of large language models (LLMs), the canonical offline imitation learning method is supervised fine-tuning (SFT), where the learner imitates an expert policy (e.g., human-generated data or traces generated by a stronger model) using a fixed dataset of expert trajectories. However, such offline approaches are reported to suffer from limitations such as poor generalization (Chu et al., 2025; Gudibande et al., 2024; Kumar et al., 2022) and severe catastrophic forgetting (Chen et al., 2025).

Post-training methods based on *online* imitation learning have emerged as an alternative promising to mitigate these issues. In the online IL setting, the learner actively interacts with the environment and queries an expert to obtain feedback Ross et al. (2011). A prominent online IL approach for LLM post-training is *on-policy distillation*, where the student samples from its current policy and minimizes the reverse-KL divergence (or another divergence) to the expert distribution to the expert distribution Kim & Rush (2016); Agarwal et al. (2024); Gu et al. (2024). Although recent studies report the empirical superiority of online IL in post-training (Yang et al., 2025; Lu & Lab, 2025), a principled understanding of when and why online IL outperforms its offline counterpart for LLM post-training remains limited.

A classical explanation for the advantage of online IL is its ability to mitigate error accumulation: If the learner incurs a per-step prediction error of ϵ under the expert distribution, then offline IL

[*]Equal contribution.

can suffer a compounding error scaling as $O(H^2\epsilon)$ where H is the horizon length. In contrast, DAgger-style online IL algorithms reduce this dependence to $O(\mu H\epsilon)$, where μ is a recoverability coefficient (Ross & Bagnell, 2010; Ross et al., 2011). However, recent theory by Foster et al. (2024) shows that under the standard log-likelihood objective used in SFT, offline imitation learning can match the sample-complexity guarantees of DAgger in *realizable* settings, where the expert policy is contained in the student policy class. This result calls into question whether error accumulation alone can fully explain the practical advantage of online IL in LLM training.

In this work, we identify an orthogonal source of advantage of online over offline IL, which we argue arises from *non-realizability* (or *misspecification*): the student policy class is too restricted (for example, with a much smaller model size) to faithfully represent the expert policy. Unlike the classical error-accumulation view which concerns long-horizon compounding, such an advantage already arises in contextual bandits with $H = 1$.

We first study the realizable setting as a sanity check, and empirically show that in LLM post-training, offline IL suffices to produce a student that fully matches the expert's performance, with online IL offering no further gains in either accuracy or training speed (Section 2). Specifically, we consider both synthetic (Countdown (Pan et al., 2025)) and real-world math reasoning (GSM8K (Cobbe et al., 2021) and DeepScaleR (Luo et al., 2025)) tasks . For each task, we first train an expert by applying reinforcement learning (RL) to a base model. We then use this RL-trained model as the expert, and compare SFT and on-policy distillation when both are initialized from the same base model. By construction, this setup guarantees realizability, since the expert policy is itself obtained within the same model family. Our results show that the SFT-trained student matches the expert's performance, with learning speed that matches or exceeds on-policy distillation. This finding is consistent with the theory of Foster et al. (2024) and also rules out alternative explanations that are not covered by their theory, such as optimization obstacles or problem-dependent artifacts. Taken together, these results further motivate our focus on the *non-realizable* setting as the key regime for understanding when online interaction benefits imitation learning.

In the non-realizable setting, we develop an orthogonal understanding of the advantage of online over offline IL relative to the classical error-accumulation argument, through a simplified yet informative mathematical framework. We first show, both theoretically and empirically, that prior discrepancy-based analyses (Foster et al., 2024; Rohatgi et al., 2025; Rajaraman et al., 2020), which control the performance gap via distributional discrepancies between the expert and student policies, can be insufficient to characterize the effectiveness of imitation learning under misspecification, and that reward-dependent analyses are necessary. For offline imitation learning, we identify an *information-theoretic* limitation specific to the non-realizable regime: the expert policy may fail to assign sufficient probability mass to responses that are in principle learnable within the restricted student policy class. For online imitation learning, we further introduce a structural condition that explicitly captures the misspecification pattern relative to the reward, under which online IL can provably achieve small excess risk despite large distributional discrepancy. At a high level, the condition states that the expert's signal should either be aligned with the actual reward, or deviate from it only in ways that are the restricted student class cannot realize. Due to space constraints, we defer the preliminaries and related work to the appendix.

2 Offline IL Successfully Recovers Expert Policy Under Realizability

The seminal work of Ross & Bagnell (2010) showed that offline imitation learning can suffer from error accumulation: if the learner incurs a per-step error of ϵ under the expert distribution, then when rolled out under its own policy, the resulting compounding error can scale quadratically with the horizon H, i.e., as $O(H^2\epsilon)$. Online imitation learning algorithms such as DAgger (Ross et al., 2011) improve this dependence to $O(\mu H\epsilon)$, where a small recoverability coefficient μ indicates that the expert can effectively recover from any suboptimal intermediate actions. However, recent work by Foster et al. (2024) shows that, under the log-likelihood objective used in SFT, offline behavior cloning already achieves optimal statistical complexity in the realizable setting, and no online imitation learning algorithm can improve upon it in the worst case. That said, this theory does not cover every possible advantage of online IL in realizable settings. For example, it studies

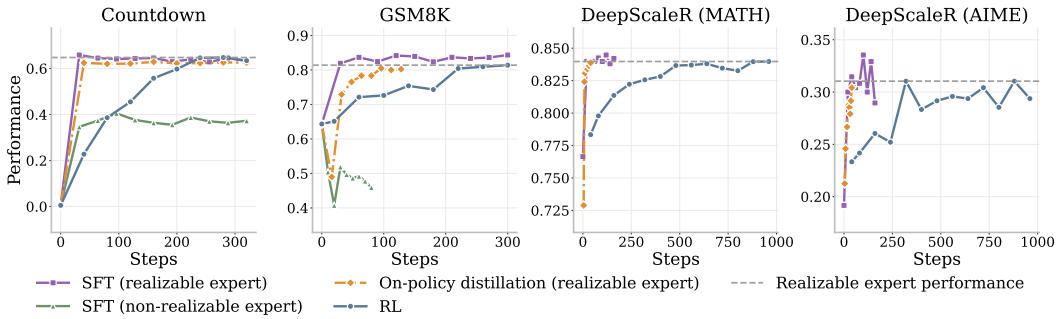

Figure 1: **In-distribution evaluation on different tasks.** SFT with a realizable expert fully matches the expert's performance, leaving litter room for further improvement from online interaction. On-policy distillation provides no accuracy gains or faster training speed. In contrast, when the expert is unrealizable, SFT exhibits a significant performance gap.

statistical complexity but does not consider optimization dynamics, and it does not rule out the possibility that online IL may still benefit from favorable problem-specific structure.

In this section, as a sanity check for the realizable setting, we empirically demonstrate that SFT (offline IL) is indeed sufficient to fully recover expert performance in this regime, and on-policy distillation does not yield further gains in either final performance or training efficiency. Conversely, under model misspecification, where the expert lies outside the student's class, SFT fails to match the expert and exhibits strictly inferior performance. This suggests that misspecification is likely required to see an online-offline gap in practice.

2.1 Experimental Setup

We conduct experiments on three tasks: a synthetic Countdown (Pan et al., 2025) task and two Math reasoning dataset: GSM8K (Cobbe et al., 2021) and DeepScaleR (Luo et al., 2025) (details in D). We use Qwen2.5-3B-Instruct (Team, 2024) as the base model for Countdown, Llama-3.2-3B (Grattafiori et al., 2024) fine-tuned on OpenR1 (Face, 2025) for GSM8K, and DeepSeek-R1-Distill-Qwen-1.5B as the backbone for DeepScaleR. We compare three training paradigms:

• **RL (GRPO):** The base model is optimized directly via GRPO (Shao et al., 2024).

• **SFT with Realizable Expert:** The student model is supervised on a teacher from its own model class. Specifically, we generate samples from the RL-ed model described above, which have been trained for 320 GRPO steps for Countdown, 80 steps for GSM8K, and 80 steps for DeepScaleR.

• **SFT with Unrealizable Expert:** The student is supervised by a highly capable unrealizable teacher, DeepSeek-V3.2-Exp (Liu et al., 2024), which achieves near-perfect accuracy on Countdown and GSM8K dataset.

2.2 Experimental Results

This section presents experimental results on both in-distribution and out-of-distribution evaluations. Additionally, we provide an empirical analysis of the DeepScaleR-1.5B model.

In-Distribution Evaluation. As shown in Figure 1, on both the synthetic Countdown task and real-world math reasoning benchmarks, offline imitation learning (SFT) with a realizable expert matches the expert's performance, leaving little room for additional gains from online IL. Moreover, SFT is substantially more sample-efficient than training the RL-based expert. In contrast, when the expert is non-realizable, SFT falls well short of the expert's performance.

Out-of-Distribution Generalization and Catastrophic Forgetting. We further evaluate the model's out-of-distribution (OOD) generalization on Countdown, where test instances use a larger number range than seen during training, and assess catastrophic forgetting during GSM8K training by measuring general language capability on MMLU benchmark. Figure 2 show that SFT with a realizable expert achieves OOD performance that matches the expert, whereas SFT with a non-

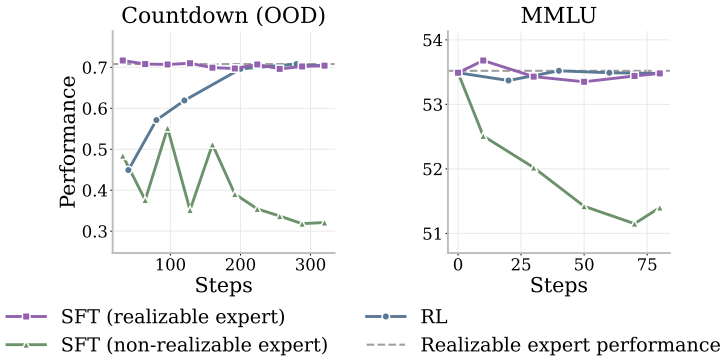

Figure 2: **OOD generalization and catastrophic forgetting.** Left: Models are evaluated on Countdown instances with larger number range. Right: Catastrophic forgetting during GSM8K training. Models are evaluated on MMLU benchmark. Overall, SFT from a realizable expert achieves strong OOD performance and exhibiting little to no MMLU degradation, whereas SFT from a non-realizable expert harms OOD generalization during training and leads to severe forgetting.

realizable expert deteriorates with fine-tuning. We observe a similar pattern for forgetting: under a realizable expert, the MMLU score remains stable throughout training, but under a non-realizable expert it drops substantially,.

These findings refine prior observations that SFT tends to generalize worse OOD (Chu et al., 2025) and forget more than RL (**?**Chen et al., 2025). Instead, our results suggest this gap is not inherent to the algorithm, but arises primarily under non-realizability. It is the training data distribution, rather than the SFT algorithm itself, that drives inferior OOD generalization and severe forgetting.

3 Comparison of Online and Offline IL in Non-Realizable Settings

Since the advantage of online imitation learning does not arise under realizability as shown in Section 2, we turn to the *non-realizable* setting, where the expert policy cannot be represented by the student policy class, i.e. $\pi^{\mathrm{e}} \notin \Pi$. Online imitation learning methods such as on-policy distillation have been empirically observed to significantly outperform offline approaches (SFT) in this setting (Yang et al., 2025; Lu & Lab, 2025)[1]. However, despite these successes, a theoretical understanding of *why and when* online interaction provides an advantage over offline imitation learning in non-realizable settings remains limited.

Previous work framed imitation learning as minimizing the distributional discrepancy (e.g., Hellinger distance) between the student and expert policy (Rohatgi et al., 2025). In this section, we first discuss and empirically verify the key limitations of such discrepancy-based analyses. In particular, they can be overly pessimistic in non-realizable settings, and, because minimizing discrepancy is misaligned with the true objective of maximizing expected reward, they do not adequately explain the widely observed empirical advantage of online imitation learning over offline methods.

Motivated by these limitations, we develop a new perspective on the advantage of online over offline imitation learning in non-realizable settings by directly taking the reward structure into account. For offline imitation learning, we identify a fundamental information-theoretic barrier in this regime, which differs from classical error-accumulation arguments (Ross & Bagnell, 2010; Ross et al., 2011). We show that the sample complexity of any offline imitation learning algorithm that learns solely from i.i.d. expert demonstrations must scale with a coverage coefficient $C^{\mathrm{e}}_{\star}$ (**??**), which can be viewed as a measure of the mismatch between the expert and the student policies. For online imitation learning, to move beyond reward-agnostic discrepancy-based analyses, we characterize the interplay among (i) the student policy class Π, (ii) the expert π^{e}, and (iii) the reward structure,

[1] Yang et al. (2025); Lu & Lab (2025) study distilling a smaller model from a much larger expert, and their setup therefore falls into the non-realizable setting.

and identify a general condition under which online imitation learning is provably efficient despite severe misspecification.

3.1 Limitations of Discrepancy-based Analyses in Non-Realizable Settings

A standard approach to analyzing imitation learning is to upper bound the performance suboptimality gap by a distributional discrepancy between the expert policy π^{e} and the learned policy $\hat{\pi} \in \Pi$ (Rajaraman et al., 2020; Foster et al., 2024; Rohatgi et al., 2025). For instance, one typically argues that

$$V(\pi^{\mathrm{e}}) - V(\hat{\pi}) \;\le\; \mathbb{E}_{x\sim\mathcal{D}_x}[D_{\mathrm{TV}}(\pi^{\mathrm{e}}(\cdot \mid x), \hat{\pi}(\cdot \mid x))]\,, \tag{1}$$

and D_{TV} can in turn be upper bounded by other statistical distances or divergences, such as KL divergence, Hellinger distance, and χ^2 divergence.[2] The discrepancy term is reward-agnostic, and can be minimized by any policy-matching procedure. For a given discrepancy measure D, standard learning-theoretic analyses typically yield finite-sample guarantees of the form

$$\mathbb{E}_{x\sim\mathcal{D}_x}[D(\pi^{\mathrm{e}}(\cdot \mid x), \hat{\pi}(\cdot \mid x))] \;\lesssim\; C_{\mathrm{apx}} \cdot \inf_{\pi\in\Pi} \mathbb{E}_{x\sim\mathcal{D}_x}[D(\pi^{\mathrm{e}}(\cdot \mid x), \pi(\cdot \mid x))] \;+\; \varepsilon_{\mathrm{stat}} \tag{2}$$

for some $C_{\mathrm{apx}} \ge 1, \varepsilon_{\mathrm{stat}} > 0$.

In realizable settings, the approximation term vanishes: $\inf_{\pi\in\Pi} \mathbb{E}_{x\sim\mathcal{D}_x}[D_{\mathrm{TV}}(\pi^{\mathrm{e}}(\cdot \mid x), \pi(\cdot \mid x))] = 0$. As a result, with sufficient expert data, policy matching can drive the discrepancy arbitrarily close to zero, and hence also drive the performance suboptimality $V(\pi^{\mathrm{e}}) - V(\hat{\pi})$ to zero.

In contrast, in non-realizable settings, even the best-in-class discrepancy is strictly positive, so discrepancy-based bounds can never imply a performance suboptimality smaller than this irreducible distance. While such bounds are tight in the worst case over all possible reward functions, they can be overly pessimistic for a particular reward function: a large discrepancy between the student policy and expert policy does not necessarily imply a large performance gap. For example, on a given math problem, a stronger expert may produce a correct solution directly, whereas a weaker student may rely on more trial and error; their response distributions can differ substantially even if both ultimately solve the problem with similar success rates. As shown empirically in Section 3.1.1, this phenomenon indeed appears in LLM post-training.

Furthermore, discrepancy-based analyses also fail to explain why online imitation learning often outperforms offline imitation learning in practice. Prior work has shown that offline behavior cloning is already statistically optimal for minimizing the discrepancy even in non-realizable settings (Rohatgi et al., 2025, Theorem 3.1), and the computational benefits of online interactions are also unclear. However, while this result indicates that online interaction does not generally help identify the student policy closest to the expert under a distributional discrepancy, it does not rule out the possibility that online interaction may still help identify the student policy with the highest expected reward under a particular reward function.

These limitations of discrepancy-based analyses motivate us to go beyond the reward-agnostic discrepancy-based analysis in non-realizable settings, and instead ask when offline imitation learning becomes inefficient (Section 3.2) and when online imitation learning can yield improvements under a specific reward function (Section 3.3).

3.1.1 Empirical Validation

In this section, we empirically validate that the distributional discrepancy between an expert and a student learned via imitation learning can be much larger than their performance gap. Specifically, we consider DeepSeek-R1-0528 as the expert and DeepSeek-R1-0528-Qwen3-8B as the student (DeepSeek-AI, 2025), which is distilled from DeepSeek-R1-0528 using Qwen3-8B-Base (Yang et al., 2025) as the base model. We evaluate both models on the AIME 2024 and AIME 2025 benchmarks. For each benchmark, each model generates 14 responses per question, yielding 840 responses per model in total.

[2] Foster et al. (2024) establishes a tighter upper bound based on Hellinger distance for deterministic experts or experts with bounded reward variance (their Theorems 2.1 and 3.1), compared with the direct total variation bound. Our discussion is not sensitive to this distinction.

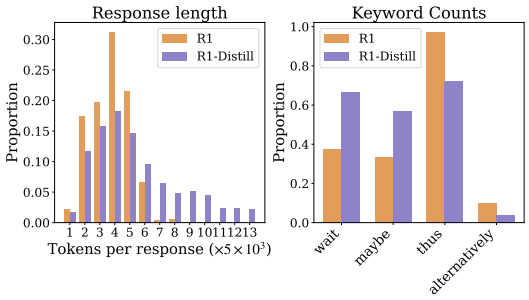

Figure 3: Response-length distribution and keyword frequencies for DeepSeek-R1-0528 (expert) and DeepSeek-R1-0528-Qwen3-8B (student) on AIME 2024 and AIME 2025 (14 responses per question). The results indicate a substantial distributional discrepancy between the expert and the student.

As shown in Figure 3, the two models exhibit substantial differences in their response distributions, as reflected by two simple and interpretable features: response length (measured using the Qwen3-8B tokenizer for both models) and the frequency of keywords associated with reasoning behavior in the responses Zeng et al. (2025); Gandhi et al. (2025). Table 1 further reports the approximate total variation distance obtained by discretizing response length into 5,000-token bins and the frequency gaps for the selected keywords. Note that these metrics are lower bounds of the total variation distance between the response distributions of two models, since distinct responses may fall into the same length bin or contain the same selected keywords. Therefore, the true distributional discrepancy can only be larger than what we report. Since even these lower bounds already exceed the observed performance gap by a wide margin (Table 1), our results provide empirical validation that distributional discrepancies can be much larger than performance differences.

We conclude that guarantees based solely on distributional discrepancies can be overly loose, motivating finer-grained characterizations that capture how policy differences interact with the reward structure.

3.2 Lower Bounds for Offline IL in Non-Realizable Settings

We now present an information-theoretic limitation of offline imitation learning under misspecification. Previous empirical results have identified settings in which supervised fine-tuning (i.e., offline imitation learning) struggles to improve a weak student in the non-realizable regime, where the expert is a much stronger reasoning model. In such cases, the student often fails to learn from expert responses that are either too compact or involve reasoning patterns too complex for the student to represent (Li et al., 2025; Jiang et al., 2025). Intuitively, under misspecification, the expert's responses cannot be faithfully represented by the student policy class, while those responses that are representable by the student are rarely revealed in i.i.d. expert samples.

Theorem 1 formalizes this intuition by showing that, even when the student class contains a near-optimal policy and the expert itself achieves perfect performance, learning from i.i.d. expert demonstrations can still be statistically hard. Specifically, for any offline imitation learning algorithm that observes only i.i.d. expert demonstrations, the sample complexity required to achieve vanishing excess risk must scale with the coverage coefficient $C^{\mathrm{e}}_{\star}$ defined as $C^{\mathrm{e}}_{\star} := \sup_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{\pi^{\star}(y|x)}{\pi^{\mathrm{e}}(y|x)}$, where $\pi^{\star} \in \Pi$ is a reference policy achieving $\max_{\pi\in\Pi} V(\pi)$. Such lower bound is **specific to the non-realizable setting**, as under realizability, one can take $\pi^{\star} = \pi^{e}$, in which case $C^{\mathrm{e}}_{\star} = 1$. Under misspecification, however, $C^{\mathrm{e}}_{\star}$ can be arbitrarily large, serving as a measure of the degree of mismatch between the expert and $\pi^{\star}$.

Limitation of offline IL under misspecification

Theorem 1. *For any integer $K \geq 2$, any coverage parameter $C^{\mathrm{e}}_{\star} \geq 1$, and any (possibly randomized) offline imitation learner $\mathcal{A}$ with access to N i.i.d. samples $(x_i, y_i)_{i=1}^{N}$ with $x_i \sim \mathcal{D}_x$ and $y_i \sim \pi^{\mathrm{e}}(\cdot \mid x_i)$, there exist a contextual bandit problem $(\mathcal{D}_x, r)$, an expert policy π^{e}, and a student policy class Π with $|\Pi| = K$ such that the following hold:*

1. *The expert is optimal, i.e., $V(\pi^{\mathrm{e}}) = 1$.*

2. *There exists an optimal student policy* $\pi^\star \in \Pi$ *with* $V(\pi^\star) = 1$.
3. *The expert satisfies a pointwise coverage condition with respect to* $\pi^\star$*, that is* $C^{\mathrm{e}}_\star < \infty$.

Moreover, the policy $\hat{\pi}$ *output by* $\mathcal{A}$ *satisfies*

$$1 - \mathbb{E}[V(\hat{\pi})] \gtrsim \min\Big\{1, \frac{C^{\mathrm{e}}_\star \log|\Pi|}{N}\Big\},$$

where the expectation is over the randomness of learner and sampling.

The proof follows a similar idea to Theorem 6.2 of Rajaraman et al. (2020) and is presented in Section C.1. The hard instance is constructed so that, at each context, the expert places most of its mass on an action that attains optimal reward but is unavailable to the student class due to misspecification, while the unique optimal student action is revealed only through a rare expert sample with probability $1/C^{\mathrm{e}}_\star$.

3.3 A Structural Condition for Effective Online IL.

The information bottleneck discussed in Section 3.2 can be alleviated by actively querying expert feedback, rather than relying solely on i.i.d. expert samples. However, in non-realizable settings, an expert with high performance may not necessarily be a good expert, as being a good expert also requires providing meaningful feedback on the learner's queries. In this section, we consider a setting in which the feedback on a response y is given by the expert's conditional probability $\pi^{\mathrm{e}}(y \mid x)$, and study a specific class of online imitation learning objectives (Definition 3). The widely used on-policy distillation algorithm for language model post-training is an instantiation of such objective with $f(\pi^{\mathrm{e}}(y \mid x)) = \log(\pi^{\mathrm{e}}(y \mid x))$ together with an additional entropy regularization term (Section A.2). Under this setup, we aim to seek structural assumptions under which imitation learning can achieve a performance-suboptimality guarantee (measured by excess risk $V(\hat{\pi}) - \min_{\pi\in\Pi} V(\pi)$) that could be smaller than the distributional discrepancy between the expert and the learned policy (as discussed in Section 3.1). As the following motivating example shows, this requires characterizing the interplay between the reward function and the misspecification structure.

A motivating synthetic example. We use a simple 2D Gaussian toy example (Figure 4) to motivate our condition. Actions are points $a = (x, y) \in \mathbb{R}^2$, and policies are Gaussians $\pi_\mu = \mathcal{N}(\mu, I_2)$. The reward depends only on the y-coordinate: $r(x, y) = \mathbf{1}[y \geq 0]$. The expert is $\pi^{\mathrm{e}} = \mathcal{N}(\mu^{\mathrm{e}}, I_2)$ with $\mu^{\mathrm{e}} = (-2, 2)$. We consider two student policy classes: $\Pi_A = \{\mathcal{N}(\mu, I_2) : \mu_x = 0\}, \Pi_B = \{\mathcal{N}(\mu, I_2) : \mu_y = \mu_x\}$. Both classes are misspecified, yet the expected rewards resulting from IL (with reverse-KL objective as in on-policy distillation) differ sharply. Let $\hat{\pi}_A, \hat{\pi}_B$ be the reverse-KL projections of π^{e} onto Π_A and Π_B, respectively. Then $V(\hat{\pi}_A) \approx 0.977, V(\hat{\pi}_B) = 0.5$. corresponding to excess risks 0.023 and 0.5, respectively.

The reason why distilling from the same expert yields a much higher expected reward in Π_A than in Π_B is because the two classes impose different *misspecification structures relative to the reward*. Here the reward $r(x, y) = \mathbf{1}[y \geq 0]$ depends only on the y-coordinate. Class Π_A is agnostic to the reward-irrelevant x-direction by fixing $\mu_x = 0$ while leaving μ_y free, hence the reverse-KL projection can match the expert's reward-relevant mean $\mu^{\mathrm{e}}_y = 2$. In contrast, Class Π_B couples the coordinates via $\mu_y = \mu_x$. Consequently, minimizing the reverse-KL objective tends to match the expert along the reward-irrelevant x-direction, resulting in non-optimal performance with respect to expected reward.

As illustrated in the above toy example, to characterize when imitation learning is effective under misspecification, one must account for misspecification structure between the student and expert under a particular reward function. Below we propose a general sufficient condition, which we term the *Misspecification–Reward Alignment Condition*.

Effectiveness of online IL under misspecification

Assumption 1 (Misspecification-Reward Alignment Condition)**.** *There exist* $\alpha > 0$*,* $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$*, and* $b : \mathcal{X} \to \mathbb{R}$ *such that*

$$f(\pi^{\mathrm{e}}(y \mid x)) = \alpha\, r(x, y) + h(x, y) + b(x).$$

Let $\pi^\star \in \Pi$ be a reference policy. For any $\epsilon > 0$, assume that for all $\pi \in \Pi$ satisfying $J_{\mathrm{on}}(\pi) \geq J_{\mathrm{on}}(\pi^\star) - \epsilon$,

$$\mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathbb{E}_{y\sim\pi(\cdot|x)}h(x,y) - \mathbb{E}_{y\sim\pi^\star(\cdot|x)}h(x,y)\right] \leq \delta(\epsilon).$$

Theorem 2. *Under Assumption 1, let $\pi \in \Pi$ be any policy that is ϵ-suboptimal for the online IL objective Equation* (4)*, i.e.,*

$$J_{\mathrm{on}}(\pi) \geq \max_{\pi'\in\Pi} J_{\mathrm{on}}(\pi') - \epsilon.$$

Then we conclude: $V(\pi) \geq V(\pi^\star) - \frac{\epsilon+\delta(\epsilon)}{\alpha}$.

The proof of Theorem 2 is provided in Section C.2. Assumption 1 views the reshaped expert signal $f(\pi^{\mathrm{e}}(y \mid x))$ as decomposing into three components. The first term $\alpha r(x, y)$ is the *reward-aligned* part, where $\alpha > 0$ quantifies the strength of alignment. The second term $h(x, y)$ is a *residual* that captures preferences encoded by the expert beyond reward. And the last term $b(x)$ depends only on the state. Crucially, Assumption 1 requires that within the ϵ-optimal set under J_{on}, the residual term cannot *inflate* $J_{\mathrm{on}}(\pi)$ for a reward-suboptimal policy. Two specific cases where Assumption 1 holds include:

1. When the reshaped expert signal $f(\pi^{\mathrm{e}}(y \mid x))$ is highly calibrated with the true reward $r(x, y)$. Concretely, if for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, it holds that residual term $\alpha \gg |h(x, y)|$, then the resulting suboptimality $\frac{\epsilon+\delta(\epsilon)}{\alpha} \ll 1$ as $\epsilon \to 0$.
2. The expert signal encodes preferences beyond reward (captured by the residual $h(x, y)$), but these preferences do not distinguish among near-optimal policies in the student class. Formally, for any π_1, π_2, with $J_{\mathrm{on}}(\pi_1), J_{\mathrm{on}}(\pi_2) < \epsilon$,
$$\left|\mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathbb{E}_{y\sim\pi_1(\cdot|x)}h(x,y) - \mathbb{E}_{y\sim\pi_2(\cdot|x)}h(x,y)\right]\right| \ll 1.$$
This characterization explains why the expert is effective for Π_A in the synthetic example.

Remark 3 (Comparison with the error-accumulation view)**.** A classic account of the benefits of online IL is that online interaction mitigates error accumulation, leading to DAgger-style algorithms (Ross & Bagnell, 2010; Ross et al., 2011; Lu & Lab, 2025): under offline imitation, small one-step prediction errors can compound over a horizon of length H, yielding an $O(H^2\epsilon)$ performance gap, whereas online interaction can correct for this distribution shift and improve the dependence to $O(\mu H\epsilon)$, where μ is a recoverability coefficient. In contrast, this work studies an orthogonal limitation of offline imitation learning, namely an information bottleneck that arises in non-realizable settings. Such a limitation appears even in contextual bandits (with $H = 1$). We further characterize the interplay between the reward function and the misspecification structure that enables online imitation learning in non-realizable settings, a perspective that is absent from previous DAgger-style analyses.

Finite-Sample Guarantees. Using standard pessimism techniques (Wang et al., 2024), one can extract a finite-sample result of the objective Equation (4) with the presense of a base model with good coverage (such condition is also consisitent with empircal since the success on-policy distillation requires the presence of a strong base model). Details of the algorithm and results are deferred to Section C.3.

4 Conclusion and Discussion

In this work, we show that misspecification is a key factor contributing to benefits of online imitation learning. We hope our results help clarify what leads to the empirical performance gap between offline and online IL methods, and suggest several directions that may be worth exploring in future work. Empirically, when realizability holds, offline IL can already recover the expert's performance, leaving no room for online interaction to help. On the theoretical front, we characterize an information-theoretic limitation of offline IL under misspecification, and give structural conditions under which online IL can still be effective even when the discrepancy between expert and student is large. This complements prior viewpoints based on error accumulation and reward-agnostic discrepancy analyses.

REFERENCES

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jonathan Chang, Masatoshi Uehara, Dhruv Sreenivas, Rahul Kidambi, and Wen Sun. Mitigating covariate shift in imitation learning via offline data with partial coverage. *Advances in Neural Information Processing Systems*, 34:965–979, 2021.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*, 2025.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

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

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Nicolas Espinosa-Dice, Sanjiban Choudhury, Wen Sun, and Gokul Swamy. Efficient imitation under misspecification. *arXiv preprint arXiv:2503.13162*, 2025.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, 2025.

Dylan J Foster and Alexander Rakhlin. Foundations of reinforcement learning and interactive decision making. *arXiv preprint arXiv:2312.16730*, 2023.

Dylan J Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. *Advances in Neural Information Processing Systems*, 37:120602–120666, 2024.

Dylan J Foster, Zakaria Mhammedi, and Dhruv Rohatgi. Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration. *arXiv preprint arXiv:2503.07453*, 2025.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

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

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Kz3yckpCN5.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786*, 2023.

Wangyi Jiang, Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. Teach small models to reason by curriculum distillation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7412–7422, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.376. URL https://aclanthology.org/2025.emnlp-main.376/.

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

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners. *arXiv preprint arXiv:2502.12143*, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

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

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, et al. Deepscaler: Surpassing o1-preview with a 1.5 b model by scaling rl. *Notion Blog*, 2025.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

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.

Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. Tinyzero. https://github.com/Jiayi-Pan/TinyZero, 2025. Accessed: 2025-01-24.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. *arXiv preprint arXiv:2402.14811*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Nived Rajaraman, Lin Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33:2914–2924, 2020.

Dhruv Rohatgi, Adam Block, Audrey Huang, Akshay Krishnamurthy, and Dylan J Foster. Computational-statistical tradeoffs at the next-token prediction barrier: Autoregressive and imitation learning under misspecification. *arXiv preprint arXiv:2502.12465*, 2025.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

Stephane Ross and J Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

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

Yuda Song, Dhruv Rohatgi, Aarti Singh, and J. Andrew Bagnell. To distill or decide? understanding the algorithmic trade-off in partially observable reinforcement learning. *arXiv preprint arXiv: 2510.03207*, 2025.

Umar Syed and Robert E Schapire. A game-theoretic approach to apprenticeship learning. *Advances in neural information processing systems*, 20, 2007.

Qwen Team. Qwen2. 5: A party of foundation models, september 2024. *URL https://qwenlm. github. io/blog/qwen2*, 5(4), 2024.

Lequn Wang, Akshay Krishnamurthy, and Alex Slivkins. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 766–774. PMLR, 2024.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Hassan Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *arXiv preprint arXiv:2312.11456*, 2023.

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.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

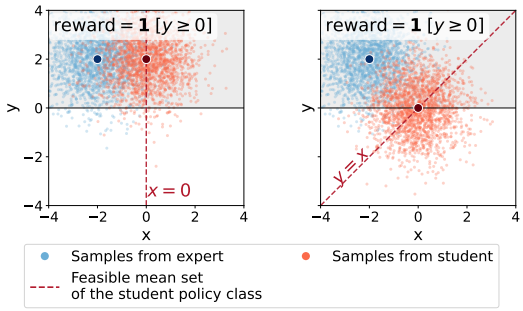

Figure 4: **Synthetic example.** Actions are $(x, y) \in \mathbb{R}^2$ and policies are Gaussians $\mathcal{N}(\mu, I_2)$; reward is $r(x, y) = \mathbf{1}[y \geq 0]$ (shaded region). Expert $\pi^{\mathrm{e}} = \mathcal{N}((-2, 2), I_2)$ (blue). **Left:** student class $\Pi_A : \mu_x = 0$ (red dashed line) yields a reverse-KL projection with $V(\hat{\pi}_A) \approx 0.977$. **Right:** student class $\Pi_B : \mu_y = \mu_x$ (red dashed line) yields $V(\hat{\pi}_B) = 0.5$.

Table 1: **Distributional discrepancy vs. performance gap.** The TV distance induced by discretized response length (5,000-token bins), keyword-frequency gaps, and the performance gap between DeepSeek-R1-0528 (expert) and DeepSeek-R1-0528-Qwen3-8B (student) on AIME 2024 and AIME 2025. Both the length-based TV distance and the keyword gaps are **lower bounds** on the TV distance between the full response distributions, yet they already far exceed the performance gap, indicating that discrepancy-only guarantees can be overly loose.

Metric	**Gap (%)**
Response length (TV distance)	29.99
Keyword (wait)	29.12
Keyword (maybe)	23.29
Keyword (thus)	24.88
Keyword (alternatively)	6.20
Performance gap	5.46

A Preliminaries

For a set $\mathcal{X}$, let $\Delta(\mathcal{X})$ denote the set of all probability distributions over $\mathcal{X}$. For two distributions P, Q over $\mathcal{X}$, their Kullback–Leibler divergence is defined by $D_{\mathrm{KL}}(P\|Q) := \mathbb{E}_{x\sim P}\left[\log \frac{P(x)}{Q(x)}\right]$, and their total variation distance is defined by $D_{\mathrm{TV}}(P, Q) := \sup_{A\subseteq\mathcal{X}} |P(A) - Q(A)| = \frac{1}{2}\sum_{x\in\mathcal{X}} |P(x) - Q(x)|$. We use standard asymptotic notation throughout, and we write $a \lesssim b$ as shorthand for $a = O(b)$, and $a \gtrsim b$ as shorthand for $a = \Omega(b)$.

A.1 Problem Setting

To provide a clean theoretical sandbox, we abstract language model post-training as a contextual bandit problem (Rafailov et al., 2023; Xiong et al., 2023; Foster et al., 2025). Let $\mathcal{X}$ represent the set of all possible prompts (contexts) and $\mathcal{Y}$ the set of generated responses (actions). A policy $\pi : \mathcal{X} \to \Delta(\mathcal{Y})$ specifies a conditional distribution $\pi(y \mid x)$ over responses y given a prompt x. Let $r(x, y) : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be the reward function. We aim to find a policy π within a policy class Π that achieves high expected reward $V(\pi) := \mathbb{E}_{x\sim\mathcal{D}_x}\mathbb{E}_{y\sim\pi(\cdot|x)} r(x, y)$, where $\mathcal{D}_x$ is the distribution over prompts. We follow standard convention in RL theory and assume the policy class $|\Pi| < \infty$ is finite (Foster & Rakhlin, 2023; Foster et al., 2024; Rohatgi et al., 2025).

Realizability. Realizability is a widely studied assumption in RL theory that asks whether the student policy class is expressive enough to represent the expert policy. In this work, we identify realizability as a crucial factor governing the advantage (or lack theoreof) of online IL over offline IL.

Definition 1 (Realizability and Misspecification)**.** We say that the expert policy π^{e} is *realizable* with respect to the student policy class Π if $\pi^{\mathrm{e}} \in \Pi$. Otherwise if $\pi^{\mathrm{e}} \notin \Pi$, the setting is *non-realizable* (or *misspecified*).

A.2 Imitation Learning

Imitation learning is a learning paradigm in which the learner improves its policy by leveraging information provided by an expert policy π^{e}, without direct access to the reward feedback during training. In this work, we study two settings: offline and online imitation learning.

Offline Imitation Learning. The learner is given a static dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ of i.i.d. expert demonstrations, where $x_i \sim \mathcal{D}_x$ and $y_i \sim \pi^{\mathrm{e}}(\cdot \mid x_i)$. The dominant approach in this setting is **behavior cloning**, which reduces imitation learning to a supervised learning problem with the following objective:

Definition 2 (Population Objective of Behavior Cloning)**.** Given an expert policy π^{e} and a loss function $\ell : \Delta(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$, the population behavior cloning objective is defined as

$$L_{\mathrm{BC}}(\pi) \coloneqq \mathbb{E}_{x\sim\mathcal{D}_x}\, \mathbb{E}_{y\sim\pi^{\mathrm{e}}(\cdot|x)}\big[\ell(\pi(\cdot \mid x), y)\big]. \tag{3}$$

The goal of behavior cloning is to find a policy $\pi \in \Pi$ that maximizes $L_{\mathrm{BC}}(\pi)$.

A standard instantiation of behavior cloning in LLM post-training uses the negative log-likelihood loss $\ell(\pi(\cdot \mid x), y) = -\log \pi(y \mid x)$, which corresponds to the **supervised fine-tuning (SFT)** objective.

Online Imitation Learning. In the online setting, the learner can actively query the expert and receive feedback on its own actions. Learning proceeds in episodes: in the i^{th} episode, the learner observes a context $x_i \sim \mathcal{D}_x$, executes a certain policy π_i to generate a response $y_i \sim \pi_i(\cdot \mid x_i)$ and subsequently receives expert-provided information evaluated on this response. In existing online imitation learning formulations based on MDPs (e.g., (Ross & Bagnell, 2010; Foster et al., 2024)), the expert provides step-wise guidance along a trajectory. In contrast, we adopt a simpler contextual bandit formulation: in each episode, the learner produces a complete response y_i (a sequence of tokens), and the expert reveals its conditional probability $\pi^{\mathrm{e}}(y_i \mid x_i)$ on that response. This form of density access is natural in language model distillation, where the expert model's logits are available. Furthermore, it captures the core interaction pattern of widely used on-policy distillation algorithms (Agarwal et al., 2024; Gu et al., 2024; Lu & Lab, 2025), and enables a meaningful characterization of the benefit of online access to expert information beyond a static offline dataset (Section 3).

While the objective and algorithm of online IL can be flexible, in this work we focus in particular on the following formulation:

Definition 3 (Population Objective of Online Imitation Learning)**.** Given an expert policy π^{e} and a *shaping function* $f : \mathbb{R} \to \mathbb{R}$ that maps the expert density to a scalar score, the population objective of online imitation learning is

$$J_{\mathrm{on}}(\pi) \coloneqq \mathbb{E}_{x\sim\mathcal{D}_x}\mathbb{E}_{y\sim\pi(\cdot|x)}\left[f(\pi^{\mathrm{e}}(y \mid x))\right]. \tag{4}$$

The goal of online imitation learning is to find a policy $\pi \in \Pi$ that maximizes $J_{\mathrm{on}}(\pi)$.

The key difference to the offline objective (Equation (3)) is that the inner expectation is taken over the student policy π, rather than the expert policy π^{e}.

A prominent instantiation of online imitation learning in language model post-training is **on-policy distillation** (Agarwal et al., 2024; Gu et al., 2024), which uses a reverse-KL objective and is optimized in an online, on-policy manner:

$$\max_{\pi\in\Pi} -\mathbb{E}_{x\sim\mathcal{D}_x}\left[D_{\mathrm{KL}}\left(\pi(\cdot \mid x) \| \pi^{\mathrm{e}}(\cdot \mid x)\right)\right]. \tag{5}$$

This reverse-KL objective is an instance of Definition 3 with $f\left(\pi^{\mathrm{e}}(y \mid x)\right) = \log\left(\pi^{\mathrm{e}}(y \mid x)\right)$, together with an additional entropy regularization term that promotes exploration and prevents policy

collapse:[3]

$$-\mathbb{E}_{x\sim\mathcal{D}_x}[D_{\mathrm{KL}}\left(\pi(\cdot\mid x)\|\pi^{\mathrm{e}}(\cdot\mid x)\right)]=\mathbb{E}_{x\sim\mathcal{D}_x}\mathbb{E}_{y\sim\pi(\cdot|x)}[\log\pi^{\mathrm{e}}(y\mid x)]+\mathbb{E}_{x\sim\mathcal{D}_x}[\mathcal{H}(\pi(\cdot\mid x))],$$

where $\mathcal{H}(\pi(\cdot\mid x)) \coloneqq -\mathbb{E}_{y\sim\pi(\cdot|x)}[\log\pi(y\mid x)]$ denotes the policy entropy.

Remark 4 (Why use a contextual bandit formulation?)**.** Previous analyses of imitation learning are typically framed in the Markov Decision Process (MDP) setting, which involves sequential decision-making, stochastic state transitions, and rewards assigned at each step. The auto-regressive language generation we consider is a special case: the state is accumulative with deterministic transition $s_t = (x, y_{\le t}) = (s_{t-1}, y_t)$, and the reward (e.g., whether a math problem is solved correctly) is defined over the entire trajectory. As a result, this setup is essentially a contextual bandit: given a context x, the policy π specifies the distribution over the entire response trajectories y. This formulation also matches the practical objectives used in LLM post-training. In particular, both SFT (Equation (3)) and on-policy distillation (Equation (5)) are defined at the full-trajectory level, and can therefore be naturally captured within the contextual bandit framework.

Moreover, in this work, we identify *non-realizability* as a source of advantage of online over offline imitation learning that is distinct from the classical error-accumulation view. Unlike the latter, which attributes the benefit of online interaction to the mitigation of compounding errors over long horizons (Ross & Bagnell, 2010), the advantage arising from non-realizability already appears in contextual bandits with $H = 1$. By removing the multi-step structure, the contextual bandit formulation provides a cleaner setting for isolating this source of advantage.

B Related Work

Imitation Learning for LLM Post-Training. Imitation learning (IL), in which a student model learns from expert demonstrations (e.g., human annotations or distillation from a stronger model), is a central paradigm for post-training language models. Supervised fine-tuning (SFT) can be viewed as **offline imitation learning**, where the model is trained with a log-likelihood objective on the demonstration data (Brown et al., 2020; Radford et al., 2021; Ouyang et al., 2022). Previous work shows that SFT amplify existing mechanisms and knowledge acquired during pretraining (Jain et al., 2023; Prakash et al., 2024). With a small amount of carefully curated data, SFT can already yield strong instruction-following performance and even unlock more advanced capabilities such as reasoning (Zhou et al., 2023; Ye et al., 2025; Muennighoff et al., 2025).

Recent studies comparing SFT with reinforcement-learning (RL) fine-tuning suggest that SFT can exhibit weaker out-of-domain generalization (Chu et al., 2025) and more severe catastrophic forgetting than RL (Chen et al., 2025; **?**). However, for relatively weak base models, SFT can be more sample-efficient than RL for improving reasoning performance (DeepSeek-AI, 2025). Moreover, SFT (or mid-training) often serves as a warm-start for subsequent RL fine-tuning by improving the initial policy (Yeo et al., 2025; Gandhi et al., 2025).

Rencently, **online imitation learning** has emerged as a powerful paradigm for LLM post-training. A representative instantiation in this context is on-policy distillation, where the student model actively generating responses and queries the expert for probabilities along these self-generated trajectories Kim & Rush (2016); Gu et al. (2024). Unlike supervised fine-tuning that optimize a forward-KL objective on fixed data, online distillation typically minimizes the reverse-KL divergence (or variations thereof) between the student and expert distributions Gu et al. (2024).These methods have proven effective in practice Yang et al. (2025); Lu & Lab (2025). However, a theoretical understanding of when and why on-policy methods outperform offline SFT remains limited, as discussed below.

Theoretical Analysis on Online and Offline IL. Seminal work Ross & Bagnell (2010) shows that offline imitation learning occur error accumulation, specifically a quadratic dependence with respect to horizong, and `Dagger` like algorithms can mitigate error amplification through online interactions under recoverability assumptions (Ross et al., 2011; Ross & Bagnell, 2014). However, this line of work only provide algorithm-dependent lower bound, and does not provide end-to-end separation between offline and online IL (Foster et al., 2024; Rajaraman et al., 2020).

[3]The shaping function f can be generalized to depend on both $\pi^{\mathrm{e}}(y\mid x)$ and $\pi(y\mid x)$ without affecting our analysis. This extension fully captures the reverse-KL objective. For simplicity, we focus on shaping functions that depend only on the expert density.

Foster et al. (2024) shows that under *realizability* offline imitation learning with log likelihood loss (which is adopted in SFT for LLM post-training) cannot be improved by a Dagger-like algorithm in the worst case in terms of statistical complexity. While online interaction could theoretically yield improvements under additional assumptions (see the discussion in Section 4 of Foster et al. (2024)), our empirical results do not provide evidence of such gains under realizability (Section 2). This motivates our focus on the misspecified setting, where the benefits of online interaction can be more clearly manifested.

Under *misspecification*, Rohatgi et al. (2025) frames imitation learning as minimizing the Hellinger distance between the learned policy and the expert policy. In this setting, they show that offline learning already achieves optimal statistical complexity for minimizing the Hellinger distance, and no computational advantage of online interaction is established. However, distributional discrepancy measures such as the Hellinger distance generally provide only an *upper bound* on reward suboptimality and do not constitute valid *lower bounds*: a large discrepancy does not necessarily imply poor expected reward of the learned policy.

Another line of work is inverse reinforcement learning (IRL), which aims to recover an underlying reward function from expert demonstrations (Abbeel & Ng, 2004; Ziebart et al., 2008; Syed & Schapire, 2007; Chang et al., 2021). Recent work shows that interaction with the environment can enable efficient imitation learning under misspecification through IRL, given certain structural assumptions (Espinosa-Dice et al., 2025). However, these results typically assume that the reward function lies in a finite and realizable hypothesis class, and the IRL algorithms are not commonly adopted in LLM post-training. Finally, a recent study (Song et al., 2025) also empirically identifies misspecification, rather than sampling error, as a more fundamental cause of behavior cloning's suboptimal performance in partially observed MDPs.

C Proofs and Additional Results

C.1 Proof of Theorem 1

Proof. Without loss of generality, assume $K := |\Pi| = 2^l$ for some $l \in \mathbb{N}^*$. Otherwise, let $K' = 2^{\lfloor \log_2 K \rfloor} \leq K$, construct a hard instance for K' policies, and then add $K - K'$ dummy policies; this can only make learning harder and does not decrease the minimax risk.

To prove the lower bound of the optimal suboptimality, we construct a joint distribution $\mathcal{P}$ over instances $\mathcal{C}$ and experts π^{e} such that for any offline learner $\hat{\pi}(D)$ based on N i.i.d. expert demonstrations D,

$$\mathbb{E}_{(\pi^{\mathrm{e}},\mathcal{C})\sim\mathcal{P}}\left[1 - \mathbb{E}\left[V(\hat{\pi})\right]\right] \gtrsim \min\left\{1,\ \frac{C^{\mathrm{e}}_\star \log|\Pi|}{N}\right\}.$$

Let $\mathcal{X} = \{x_1, \ldots, x_l\}$ and $\mathcal{Y} = \{y_1, y_2, y_3\}$. Define the student policy class Π to consist of deterministic policies that, for each context x_i, choose either y_2 or y_3:

$$\forall i \in [l],\ \pi(\cdot \mid x_i) \in \{\delta_{y_2}, \delta_{y_3}\}.$$

Thus $|\Pi| = 2^l = K$.

Let the context distribution $\mathcal{D}_x$ satisfy $\mathcal{D}_x(x_i) = \zeta$ for $i \leq l-1$ and $\mathcal{D}_x(x_l) = 1 - (l-1)\zeta$.

Let $\mathcal{S} := \{y_2, y_3\}^l$, and index environments by $s = (s_1, \ldots, s_l) \in \mathcal{S}$. For each $s \in \mathcal{S}$, define the expert policy π^{e}_s by, for all $i \in [l]$,

$$\pi^{\mathrm{e}}_s(\cdot \mid x_i) \ = \ \left(1 - \tfrac{1}{C^{\mathrm{e}}_\star}\right)\delta_{y_1} \ + \ \tfrac{1}{C^{\mathrm{e}}_\star}\,\delta_{s_i}.$$

Define the reward function r_s by

$$r_s(x_i, y) \ = \ \mathbf{1}[y = y_1 \text{ or } y = s_i]\,.$$

The optimal student policy is $\pi^\star_s(\cdot \mid x_i) = \delta_{s_i}$ for all $i \in [l]$.

Let $\mathcal{P}$ denote the uniform prior over instances induced by $\mathcal{S}$: draw $S \sim \mathrm{Unif}(\mathcal{S})$ and set the instance to be $(\mathcal{D}_x, r_S, \pi^{\mathrm{e}}_S)$. Let $\mathcal{X}(D)$ be the context which the secret action s_i is covered by i.i.d. samples in dataset D.

Then

$$\begin{aligned}\mathbb{E}_{(\pi^{\mathrm{e}},\mathcal{C})\sim\mathcal{P}}\mathbb{E}[1-V(\hat{\pi})] &= \mathbb{E}\mathbb{E}_{(\pi^{\mathrm{e}},\mathcal{C})\sim\mathcal{P}}[1-V(\hat{\pi})] \\ &\geq \mathbb{E}\mathbb{E}_{(\pi^{\mathrm{e}},\mathcal{C})\sim\mathcal{P}}\left[\left(1-\tfrac{1}{2}\right)(1-\mathcal{D}_x(\mathcal{X}(D)))\right] \\ &\geq \tfrac{1}{2}\mathbb{E}\left[1-\mathcal{D}_x(\mathcal{X}(D))\right]\end{aligned}$$

It remains to upper bound $\mathbb{E}|\mathcal{D}_x(\mathcal{X}(D))|$. For each $i \leq l-1$, the probability that i becomes covered in one sample is $\mathbf{Pr}(x = x_i, y \neq y_1) = \zeta/C^{\mathrm{e}}_\star$. Setting $\zeta = C^{\mathrm{e}}_\star/N$ gives

$$\mathbf{Pr}(i \notin \mathcal{X}(D)) = \left(1-\frac{\zeta}{C^{\mathrm{e}}_\star}\right)^N \gtrsim \frac{1}{e},$$

and therefore

$$\mathbb{E}|\mathcal{D}_x(\mathcal{X}(D))| = \sum_{i=1}^{l-1}\mathbf{Pr}(i \in \mathcal{X}(D)) \lesssim \zeta l.$$

Plugging back yields

$$\mathbb{E}_{(\pi^{\mathrm{e}},\mathcal{C})\sim\mathcal{P}}\mathbb{E}_D[1-V(\hat{\pi}(D))] \gtrsim \zeta l \asymp \frac{\log|\Pi|}{C^{\mathrm{e}}_\star N}.$$

□

C.2 Proof of Theorem 2

Proof. Fix any $\pi \in \Pi$ that is ϵ-suboptimal for the online IL objective Equation (4), i.e., $J_{\mathrm{on}}(\pi) \geq \max_{\pi'\in\Pi} J_{\mathrm{on}}(\pi') \geq J_{\mathrm{on}}(\pi^\star) - \epsilon$.

By Assumption 1, there exist $\alpha > 0$, h, and b such that for all (x, y),

$$f(\pi^{\mathrm{e}}(y \mid x)) = \alpha\, r(x,y) + h(x,y) + b(x).$$

Taking expectation over $x \sim \mathcal{D}_x$ and $y \sim \pi(\cdot \mid x)$ yields

$$\begin{aligned}J_{\mathrm{on}}(\pi) &= \mathbb{E}_{x\sim\mathcal{D}_x}\mathbb{E}_{y\sim\pi(\cdot|x)}f(\pi^{\mathrm{e}}(y \mid x)) \\ &= \alpha\,\mathbb{E}_{x\sim\mathcal{D}_x}\mathbb{E}_{y\sim\pi(\cdot|x)}r(x,y) + \mathbb{E}_{x\sim\mathcal{D}_x}\mathbb{E}_{y\sim\pi(\cdot|x)}h(x,y) + \mathbb{E}_{x\sim\mathcal{D}_x}b(x) \\ &= \alpha\, V(\pi) + \mathbb{E}_{x\sim\mathcal{D}_x}\mathbb{E}_{y\sim\pi(\cdot|x)}h(x,y) + \mathbb{E}_{x\sim\mathcal{D}_x}b(x).\end{aligned}$$

The same decomposition holds for $\pi^\star$:

$$J_{\mathrm{on}}(\pi^\star) = \alpha\, V(\pi^\star) + \mathbb{E}_{x\sim\mathcal{D}_x}\mathbb{E}_{y\sim\pi^\star(\cdot|x)}h(x,y) + \mathbb{E}_{x\sim\mathcal{D}_x}b(x).$$

Subtracting the two identities cancels the $b(x)$ term and gives

$$J_{\mathrm{on}}(\pi) - J_{\mathrm{on}}(\pi^\star) = \alpha\big(V(\pi) - V(\pi^\star)\big) + \mathbb{E}_{x\sim\mathcal{D}_x}\Big[\mathbb{E}_{y\sim\pi(\cdot|x)}h(x,y) - \mathbb{E}_{y\sim\pi^\star(\cdot|x)}h(x,y)\Big].$$

Rearranging,

$$\alpha\big(V(\pi^\star) - V(\pi)\big) = J_{\mathrm{on}}(\pi^\star) - J_{\mathrm{on}}(\pi) + \mathbb{E}_{x\sim\mathcal{D}_x}\Big[\mathbb{E}_{y\sim\pi(\cdot|x)}h(x,y) - \mathbb{E}_{y\sim\pi^\star(\cdot|x)}h(x,y)\Big].$$

Now use $J_{\mathrm{on}}(\pi) \geq J_{\mathrm{on}}(\pi^\star) - \epsilon$, which implies $J_{\mathrm{on}}(\pi^\star) - J_{\mathrm{on}}(\pi) \leq \epsilon$. Moreover, since $J_{\mathrm{on}}(\pi) \geq J_{\mathrm{on}}(\pi^\star) - \epsilon$, the second part of Assumption 1 applies and yields

$$\mathbb{E}_{x\sim\mathcal{D}_x}\Big[\mathbb{E}_{y\sim\pi(\cdot|x)}h(x,y) - \mathbb{E}_{y\sim\pi^\star(\cdot|x)}h(x,y)\Big] \leq \delta(\epsilon).$$

Combining these two bounds,

$$\alpha\big(V(\pi^\star) - V(\pi)\big) \leq \epsilon + \delta(\epsilon),$$

and dividing by $\alpha > 0$ finishes the proof:

$$V(\pi) \geq V(\pi^\star) - \frac{\epsilon + \delta(\epsilon)}{\alpha}.$$

□

C.3 Finite Sample Guarantee of Learning Objective Eq. (4)

In this section, we present a finite-sample guarantee for the online IL objective Equation (4). This objective is typically optimized in an online and on-policy manner, as in the on-policy distillation (Gu et al., 2024; Lu & Lab, 2025), where the learner adaptively generates responses according to the current student policy (similarly to reinforcement learning). Such learning algorithms crucially rely on a strong base model (DeepSeek-AI, 2025; Zeng et al., 2025), and are typically characterized theoretically by a *coverage coefficient* with respect to the base policy (e.g., Foster et al. (2025)):

$$C_\star^{\mathrm{base}} \coloneqq \sup_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{\pi^\star(y|x)}{\pi^{\mathrm{base}}(y|x)},$$

$$C^{\mathrm{base}}(\Pi) \coloneqq \sup_{\pi\in\Pi}\sup_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{\pi(y|x)}{\pi^{\mathrm{base}}(y|x)},$$

where $\pi^\star \in \Pi$ is a reference policy (e.g. an optimal student policy).

The sample complexity analysis of fully on-policy algorithms is challenging. Existing results (e.g., Xie et al. (2025)) typically require additional assumptions, such as an additional reverse-KL regularization and realizability of the policy class, which are not suitable for our setting. Therefore, we instead present a simpler algorithm based on *pessimistic policy optimization* (Wang et al., 2024), which only samples from the base policy and queries expert feedback, yet already yields meaningful improvements over offline imitation learning.

Concretely, the learner generates N i.i.d. samples $(x_i, y_i) \sim \pi^{\mathrm{base}}$, and queries the expert signal $f(\pi^{\mathrm{e}}(y_i \mid x_i))$. For any policy $\pi \in \Pi$, define the inverse-propensity weighted estimator

$$\widehat{J}_{\mathrm{IPW}}(\pi) \;\coloneqq\; \tfrac{1}{N}\textstyle\sum_{i=1}^N \frac{\pi(y_i|x_i)}{\pi^{\mathrm{base}}(y_i|x_i)}\, f(\pi^{\mathrm{e}}(y_i \mid x_i)),$$

and the empirical regularizer

$$\widehat{\mathrm{PL}}(\pi) \;\coloneqq\; \tfrac{1}{N}\textstyle\sum_{i=1}^N \sum_{y\in\mathcal{Y}} \frac{\pi(y|x_i)}{\pi^{\mathrm{base}}(y|x_i)}.$$

The learned policy $\hat{\pi}$ is defined as

$$\hat{\pi} \;\in\; \arg\max_{\pi\in\Pi} \big\{ \widehat{J}_{\mathrm{IPW}}(\pi) \;-\; \beta\, \widehat{\mathrm{PL}}(\pi) \big\}, \tag{6}$$

where $\beta > 0$ is a pessimism parameter. With an additional bounded signal value assumption (similar ones also appear in previous works, e.g., (Xie et al., 2025)), the learned policy in Equation (6) satisfies the following guarantee:

Theorem 5 (Theorem 3.5 of Wang et al. (2024))**.** *Assume for any $x \in \mathcal{X}, y \in \mathcal{Y}$, $|f(\pi^{\mathrm{e}}(y \mid x))| \le V_{\max}$ For a base model π^{base}, a reference student policy $\pi^\star$, and some $\beta > 0$, the learned policy in Equation* (6) *satisfies*

$$J_{\mathrm{on}}(\pi^\star) - J_{\mathrm{on}}(\hat{\pi}) \le O\left(\frac{C^{\mathrm{base}}(\Pi) V_{\max} \log|\Pi|}{N} + \sqrt{\frac{C_\star^{\mathrm{base}} V_{\max} \log|\Pi|)}{N}} \right).$$

The positive result in Theorem 5 depends on the coverage coefficients with respect to the *base policy* π^{base}, whereas the lower bound for offline IL in Theorem 1 depends on the coverage coefficient with respect to the *expert policy*. Since the base policy is the model before fine-tuning, it is reasonable to expect that it has much better coverage over student policies reachable during training. Whether on-policy algorithms can further relax the requirement of a strong base model remains an interesting open question.

D Experimental Details

D.1 Countdown Task

Experimental Setup We utilize the dataset introduced by Pan et al. (2025), which comprises 327,680 training instances and 1,024 test samples.[4] To illustrate the task format, we provide a representative training prompt.

[4] https://huggingface.co/datasets/Jiayi-Pan/Countdown-Tasks-3to4

Countdown Task Example

[INST] Using the numbers [5, 94, 9, 44], create an equation that equals 93. You can use basic arithmetic operations (+, -, *, /) and each number can only be used once. Show your work in `<think> </think>` tags. And return the final answer in `<answer> </answer>` tags, for example `<answer>`(1 + 2) / 3`</answer>`. **[/INST]**
Let me solve this step by step.

Our implementation is built upon the official codebase of Pan et al. (2025),[5] incorporating specific adaptations for NVIDIA A100 execution from a community fork.[6]

Training We fine-tune the `Qwen2.5-3B-Instruct` model (Team, 2024) using Reinforcement Learning (RL) for 320 steps on 2 NVIDIA A100 GPUs. The training configuration employs a global batch size of 128 with 5 rollouts per prompt, and a mini-batch size of 64 for optimization. We set the learning rate to 1×10^{-6} and the KL penalty coefficient to $\beta_{\text{KL}} = 1 \times 10^{-3}$. The maximum generation length is capped at 1024 tokens. We use a reward function that assigns 1.0 to correct responses, 0.1 to incorrect but format-compliant responses, and 0 to invalid outputs.

Table 2: Training details of the Countdown task.

Parameter	Value	Parameter	Value
Pretrained model	Qwen2.5-3B	Batch size	128
Generations/prompt	5	Mini-batch size	64
Max prompt length	2,048	Max response len	1,024
Learning rate	1×10^{-6}	Training steps	320
Entropy coeff	0.001	Clip ratio	0.2
Rollout engine	`vllm`	Rollout temp	1
Validation temp	1	Validation top-k	50
Validation top-p	0.7	Device	NVIDIA A100

OOD Evaluation Task In Section 2, we evaluate both in-distribution (ID) and out-of-distribution (OOD) performance in the Countdown task. For OOD evaluation, we use instances with four operands only, with larger input and answer ranges.

Table 3: Detailed settings for Countdown tasks used in ID and OOD evaluation.

Setting	Number of Operands	Range of Input	Range of Answer
Train / ID	3–4	100	100
OOD	4	125	300

D.2 Math Reasoning Tasks

GSM8K. We use Llama-3.2-2B (Grattafiori et al., 2024) as the base model. To enable RLVR, we first fine-tune the model on the OpenR1 dataset (Face, 2025) for 3,064 steps with batch size 64 and learning rate 10^{-5}. For RL training, we run GRPO on GSM8K with global batch size 1,024, mini-batch size 64, learning rate 10^{-6}, and maximum response length 4,096. For SFT from the RL expert, we use learning rate 10^{-5} and batch size 64.

[5] `https://github.com/Jiayi-Pan/TinyZero`
[6] `https://github.com/JerryWu-code/TinyZero`

DeepScaleR. We use DeepSeek-R1-Distill-Qwen-1.5B as the base model. For RL, we train with GRPO on the DeepScaleR dataset (Luo et al., 2025) with context length 8k for 960 steps, using batch size 64 and learning rate 10^{-6}. For SFT, we train for $15,360$ steps.

Step alignment across methods. One RL training step consists of multiple gradient updates. In **????**, we report progress in terms of RL steps. To align SFT with RL on the horizontal axis, we rescale SFT steps by a constant factor (/10 for Countdown and /16 for GSM8K and DeepScaleR).

D.3 Experimental Details for Section 3.1.1

For both models, we generate responses with temperature 0.6, top-p 0.95, and a maximum generation length of $100{,}000$ tokens.