OFF-POLICY POLICY OPTIMIZATION UNDER EXTREME POLICY MISMATCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Policy gradient methods such as proximal policy optimization (PPO) are typically trained from fresh on-policy rollouts. At the same time, strong teacher policies can provide high-quality demonstrations that expose the student to successful actions and trajectories it may rarely discover on its own. Directly incorporating such demonstrations into policy-gradient training, however, creates substantial policy and state-distribution mismatch, raising concerns about biased or unstable updates and whether learning signal from teacher trajectories remains useful for the student policy. Surprisingly, we find that PPO can learn from teacher demonstrations under extreme policy mismatch when teacher and student samples enter the same standard clipped PPO surrogate with token-level importance ratios. Under explicit assumptions on an unclipped population surrogate, we characterize how teacher demonstrations induce *selective imitation*, providing action guidance at student-visited prefixes and exploration guidance at prefixes the student has not yet reached, while still permitting improvement beyond the teacher. Empirically, off-policy demonstrations speed up PPO convergence by $22.4\times$ even when teacher-action importance ratios start as low as 10^{-6}, and achieve the highest greedy test accuracy on three language tasks among supervised fine-tuning, on-policy distillation, and on-policy PPO.

1 INTRODUCTION

LLM post-training typically incorporates teacher knowledge through supervised fine-tuning (SFT) or on-policy distillation (OPD) (Ouyang et al., 2022; Lu & Thinking Machines Lab, 2025; DeepSeek-AI, 2025), while policy-gradient RL is trained primarily from fresh rollouts of the current policy. Directly feeding teacher trajectories into a policy gradient method is seemingly problematic: when the teacher and student differ substantially, importance ratios can be orders of magnitude from one and induced prefix distributions can differ dramatically.

Recently, some works have proposed algorithms to address policy staleness and rollout-training mismatch that arises from varying sources such as asynchronous training and training-inference engine mismatch (Zhang et al., 2026; Lei et al., 2026; Zheng et al., 2025). Another line of work studies learning from external demonstrations, using new objectives or auxiliary losses. (Yan et al., 2025; Fu et al., 2025b). These works motivate the question: how far off-policy can standard PPO be pushed without modifying its objective?

Surprisingly, we find that it can be pushed very far. We mix teacher demonstrations with fresh student rollouts in the standard clipped PPO objective, using token-level importance ratios relative to the policy that generated each sample. We make no modification to the policy-gradient surrogate and introduce no auxiliary imitation loss. Despite teacher-action importance ratios as small as 10^{-6}, PPO learns reliably from the demonstrations in our controlled setting and remains effective across several language-model reasoning tasks.

We investigate this phenomenon from three perspectives:

1. **Sources of off-policy learning signal.** We analyze the mixed on-/off-policy surrogate and show how teacher demonstrations provide both *action guidance* at student-visited prefixes and *exploration guidance* at prefixes the student has not yet reached. We further show that useful teacher signal can persist even under extreme policy mismatch, with its relative

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.

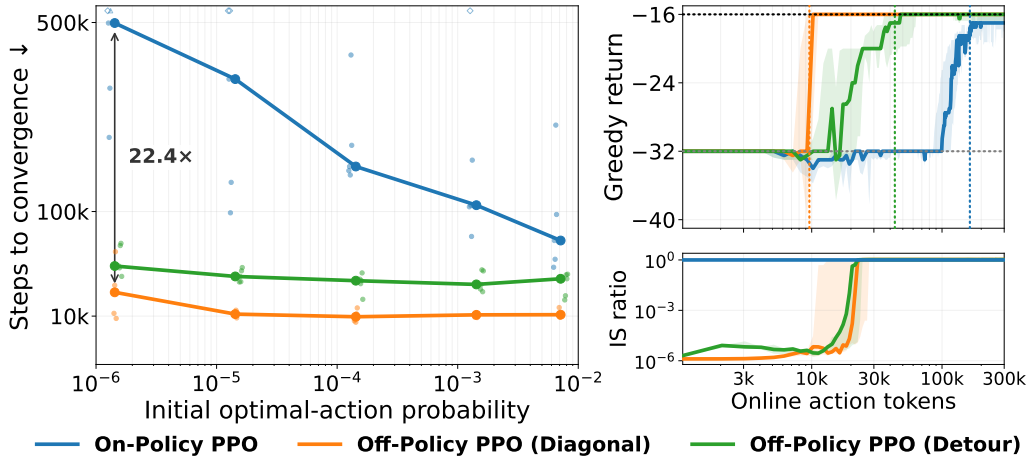


Figure 1: Off-policy PPO on GridMaze-AR (Section 4). **Left:** steps to convergence (↓, **square-root** scale) across four orders of magnitude of initial mismatch. Off-policy PPO converges consistently while on-policy PPO degrades steeply. **Top right:** a representative training curve (**log** scale). Off-policy PPO reaches optimality substantially faster. **Bottom right:** the importance ratio on teacher tokens starts near 10^{-6} and climbs toward one as the student converges.

strength governed by advantage margins and visitation rather than the importance-ratio magnitude alone (Section 3).

2. **Controlled experiments under extreme policy mismatch.** Using a simple autoregressive navigation environment (GridMaze-AR) in which teacher–student mismatch can be varied precisely, we find that off-policy PPO learns reliably across four orders of magnitude of initial teacher-action probability and reaches optimality up to $22.4\times$ faster than on-policy PPO. These experiments isolate when and how teacher trajectories accelerate learning under severe mismatch. (Section 4).
3. **Validation on language-model reasoning tasks.** Across word reversal, countdown, and GSM8K, mixing teacher demonstrations directly into PPO matches or outperforms SFT, on-policy distillation, and on-policy PPO. Despite large teacher–student mismatch, the same unmodified PPO objective remains effective in language-model training and achieves the highest greedy held-out accuracy across all three tasks (Section 5).

2 Background

We consider a finite-horizon MDP with horizon H, states s, actions a, and reward $R(s, a)$. A policy π_θ maps states to action distributions; the objective is $J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=0}^{H-1} R(s_t, a_t)]$. We set $\gamma = 1$ throughout; all experiments and the analysis in Section 3 use undiscounted returns. In the autoregressive setting, each state s_t is the token prefix generated so far. Each policy ρ induces a normalized state frequency $d_\rho(s) = \Pr_\rho(s_t = s)/H$. The advantage is $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$, estimated in practice via GAE (Schulman et al., 2015b). When data come from a behavior policy $\mu \neq \pi_\theta$, the token-level importance ratio $r_t(\theta) = \pi_\theta(a_t \mid s_t)/\mu(a_t \mid s_t)$ reweights each sample.

PPO (Schulman et al., 2017) clips r_t to stabilize updates:

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t\Big[\min\Big(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\Big)\Big], \tag{1}$$

where $\epsilon > 0$ is the clipping width. In the standard on-policy setting, μ is a recent snapshot of π_θ and ratios stay near one. When μ differs substantially—as with teacher demonstrations—ratios can start far from one and clipping can suppress gradient flow entirely.

3 NAIVE OFF-POLICY LEARNING

We study updates that mix fresh student rollouts and fixed teacher demonstrations in equal proportion. Write π_{old} for the frozen rollout policy, μ for the teacher, and π_θ for the policy being optimized ($\pi_{\theta_0} = \pi_{\text{old}}$). Each sample uses the token-level importance ratio $\pi_\theta/\pi_{\text{old}}$ (online) or π_θ/μ (demonstrations). To isolate how the two sources influence learning, we analyze the unclipped surrogate with frozen advantage targets and full action support. The results characterize this surrogate; they do not guarantee improvement from a full PPO update with a neural policy. Appendix D gives the precise assumptions.

3.1 WHERE OFF-POLICY LEARNING PUSHES

For each source $\rho \in \{\pi_{\text{old}}, \mu\}$, write $Q^\rho(s,a)$ for the expected remaining reward after taking a and following ρ, and let $b_\rho(s)$ be a state-only baseline so that $\overline{A}_\rho(s,a) = Q^\rho(s,a) - b_\rho(s)$. (Practical GAE targets need not satisfy this decomposition; see Appendix D.) Token-level importance weighting cancels the source policy's action probabilities, so each source's contribution depends on π_θ only through how it distributes probability over actions. Averaging over prefixes with frequencies $d_\rho(s)$ and giving the two sources equal weight gives the mixed objective

$$L(\theta) = \frac{1}{2} \sum_s \sum_a \pi_\theta(a \mid s) \Big[d_{\pi_{\text{old}}}(s)\, \overline{A}_{\pi_{\text{old}}}(s,a) + d_\mu(s)\, \overline{A}_\mu(s,a) \Big].$$

The quantity in brackets is the score assigned to action a: moving probability from a lower-scoring action to a higher-scoring one increases L, with more weight given to the source that visits the prefix more often (Proposition D.1).

Learning from the teacher. If the student's action values are nearly equal, its own rollouts provide little signal about which action to prefer. Teacher data can break this tie: actions with higher teacher values are favored, provided the teacher visits the prefix often enough for its contribution to dominate any opposing student preference. This is *selective imitation*: learning is guided by the teacher's action values, without requiring the student to reproduce the teacher's action probabilities (Corollary D.2).

More precisely, define the *advantage margin* $\Delta_\rho(s)$ as the difference between a chosen action's advantage and the student-weighted average over alternatives, under source ρ. If the teacher margin satisfies $\Delta_\mu(s) \geq m_s > 0$ while the student's margin has magnitude at most e_s, teacher data drive improvement whenever

$$d_\mu(s)\, m_s > d_{\pi_{\text{old}}}(s)\, e_s. \tag{2}$$

Condition (2) is a product of two quantities—margin and frequency—and either factor can dominate:

- *Action guidance.* The teacher's advantage margin is large enough to compensate for lower visitation: the teacher knows *what* to do at a prefix the student already visits.
- *Exploration guidance.* The teacher visits a prefix frequently enough to compensate for a weaker margin: the teacher shows *where* to act, supplying prefixes that are rare or unobserved in the student's online batch.

Exploration guidance is the more interesting mechanism: demonstrations provide learning signal at prefixes the student has not yet reached, before the student discovers them on its own. This does not require large teacher margins—only sufficient teacher visitation.

Improving beyond the teacher. If the student matches the teacher's policy exactly, both sources supply on-policy data. With unbiased advantage targets, the expected mixed gradient is then proportional to the ordinary return gradient. Unless that gradient is zero, a sufficiently small step improves return beyond the teacher's. PPO clipping is inactive near this point (Appendix D.1). Matching the teacher therefore need not stop learning, although this local result does not guarantee that training will reach or surpass the teacher.

Despite these useful learning signals, two practical concerns remain. First, importance sampling across full trajectories can introduce variance that grows exponentially with the horizon—can training remain stable? Second, when π_θ assigns near-zero probability to a teacher-favored action, the importance ratio r_t is tiny—is the learning signal too small, or too noisy, to be useful?

3.2 TOKEN-LEVEL CORRECTION PRESERVES TEACHER COVERAGE

Token-level importance sampling avoids the product-induced variance growth of trajectory-level correction. Each token contributes through its own ratio r_t rather than a product of ratios over the preceding trajectory. Under bounded per-token ratio second moments and advantage targets, the resulting gradient estimator has controlled second moments without an explicit exponential dependence on the horizon (Appendix D.2, Proposition D.5).

The price is bias. Token-level ratios correct the action distribution conditional on a prefix, but they do not correct which prefixes are observed or which continuation policy generated the target. Thus teacher samples retain the teacher visitation distribution d_μ and continuation values Q^μ, whereas the true on-policy gradient would involve d_π and Q^π. Consequently, the token-corrected teacher term is not an unbiased estimator of the student's return gradient.

This mismatch is not necessarily harmful. In particular, using teacher continuation values does not by itself prevent the student from outperforming the teacher. Define the teacher-relative margin

$$m_{\pi|\mu}(s) = \sum_a \pi(a \mid s) A^\mu(s, a). \tag{3}$$

The performance-difference identity (Cheng et al., 2020, Lemma 1) gives

$$J(\pi) - J(\mu) = H\, \mathbb{E}_{s\sim d_\pi}[m_{\pi|\mu}(s)]. \tag{4}$$

Hence, if the student's choices have nonnegative teacher-relative margin on the prefixes it visits, then its return is at least that of the teacher. Notably, this statement does not require π and μ to be close, nor does it require their state distributions or continuation values to match.

This result should not be read as a guarantee for the token-level update. Our surrogate weights prefixes according to d_μ, while the performance-difference identity averages margins under d_π. Increasing the surrogate therefore does not imply that $m_{\pi|\mu}(s) \geq 0$ at every student-visited prefix. Rather, the result shows that retaining teacher continuation values is compatible with improvement beyond the teacher.

The remaining state-distribution bias has a second interpretation that is central to our setting. Because teacher prefixes remain weighted by d_μ, the update receives gradient signal at prefixes that may be rare or absent under the current student policy. Thus the same bias that prevents exact recovery of the on-policy gradient also preserves the exploration guidance described in Section 3.1. Correcting all the way to d_π would remove precisely this source of teacher coverage.

3.3 SMALL RATIOS AND RELATIVE TEACHER SIGNAL

At the start of an update, when π_{old} assigns probability $\varepsilon \ll 1$ to a teacher-favored action, the ratio $\pi_\theta(a \mid s)/\mu(a \mid s)$ is tiny. This might suggest the teacher's gradient contribution is negligible relative to the online contribution. The ratio alone does not establish that: teacher sampling contributes $\mu(a \mid s)[\pi_{\text{old}}(a \mid s)/\mu(a \mid s)] = \varepsilon$ in expectation, while online sampling contributes $\varepsilon \times 1 = \varepsilon$.

Consider increasing one action's probability while decreasing all others proportionally—following the action's logit direction. Both the teacher and online source gradients of the unclipped surrogate along this direction take the form

$$g_\rho(s) = \varepsilon(1 - \varepsilon)\, \Delta_\rho(s), \tag{5}$$

where Δ_ρ is the advantage margin between the chosen action and the student's weighted alternatives under source ρ (Appendix D.3, Proposition D.6). The $\varepsilon(1-\varepsilon)$ factor is shared; it does not favor one source over the other. Writing $h_\rho(s) = \frac{1}{2}\, d_\rho(s)\, g_\rho(s)$ for each source's prefix-weighted contribution, the teacher-to-online ratio of gradient magnitudes is

$$\frac{|h_\mu(s)|}{|h_{\pi_{\text{old}}}(s)|} = \frac{d_\mu(s)\, |\Delta_\mu(s)|}{d_{\pi_{\text{old}}}(s)\, |\Delta_{\pi_{\text{old}}}(s)|}. \tag{6}$$

This comparison requires a nonzero online denominator. Both derivatives still vanish as $\varepsilon \to 0$: the statement is about relative strength, not a large absolute gradient. It is local to one prefix and one logit direction. Exact-prefix overlap can become scarce after teacher and student trajectories diverge, and the calculation does not establish alignment of their full gradients in a shared Transformer.



Figure 2: GridMaze-AR: three routes on a 17×17 grid with 8 movement directions. **Left:** the perimeter base policy (32 steps). **Center:** the diagonal teacher, following the optimal diagonal (16 steps). **Right:** the detour teacher, taking a suboptimal detour (20 steps).

When the sign can remain reliable. The same cancellation applies to sampling noise. The relative variance of the teacher gradient estimate along this direction—the ratio of variance to squared mean—equals $(S_\mu - \Delta_\mu^2)/\Delta_\mu^2$, where S_μ is a second-moment quantity capturing the teacher's action coverage and target noise (defined in Appendix D.3, Proposition D.7). This ratio has no explicit ε factor when teacher sampling, alternative proportions, and target distributions are held fixed (Proposition D.7). A small importance ratio alone therefore need not make the teacher's estimate noisier in relative terms. Poor action coverage or a vanishing margin can still make the required sample count large. With enough teacher samples at a given prefix and a nonvanishing advantage margin, the sign of the teacher gradient can be estimated reliably.

4 GridMaze-AR: A Controlled Testbed

The preceding theory applies to the mixed policy-gradient surrogate in general. We now instantiate it with PPO, the most widely used policy-gradient algorithm, and evaluate on GridMaze-AR, an autoregressive 17×17 grid navigation task designed to give precise control over policy mismatch. The agent starts at $(0, 0)$ and must reach $(16, 16)$ using 8 movement directions. The objective is to minimize episode length; we report return as $-L$ where L is the path length, so the optimal policy follows the diagonal in 16 steps (return -16). The setting is autoregressive: the actor and critic observe only the preceding action tokens, and the maze position is never an input. We use a 1-layer Transformer for both actor and critic. Full architecture and training details are in Appendix A.

Base policy and teachers. The base policy is pretrained via supervised learning to follow the *perimeter* route—north along the left edge, then east along the top—with preferred-action probability p, distributing $1 - p$ uniformly over the 7 remaining actions (Figure 2). By varying p, we control the initial importance ratio on the teacher's preferred action: at $p = 0.95$, the ratio on a diagonal action is $(0.05/7)/0.95 \approx 0.0075$.

We construct two demonstration sources. The *diagonal teacher* follows the optimal diagonal with preferred probability 0.95. The *detour teacher* follows a suboptimal detour ($\mathrm{N}^4\,\mathrm{NE}^{12}\,\mathrm{E}^4$, 20 steps, return -20)—still better than the perimeter but not optimal. The detour teacher uses preferred probability 0.95 in the initial-mismatch sweep and 0.999 in the remaining experiments. Each training step mixes equal amounts of fresh online rollouts and fixed demonstration tokens.

Off-policy PPO tolerates extreme mismatch. Figure 1 (left) sweeps the initial optimal-action probability across four orders of magnitude by varying the base-policy perimeter probability p, reporting the median steps to convergence across 5 seeds. The diagonal teacher converges fastest, reaching optimality in roughly 10,000 steps and remaining nearly flat across the entire range—consistent with logarithmic dependence on initial probability. The detour teacher is $\sim 3\times$ slower, reflecting its weaker advantage margins, but equally insensitive to the initial probability. Online

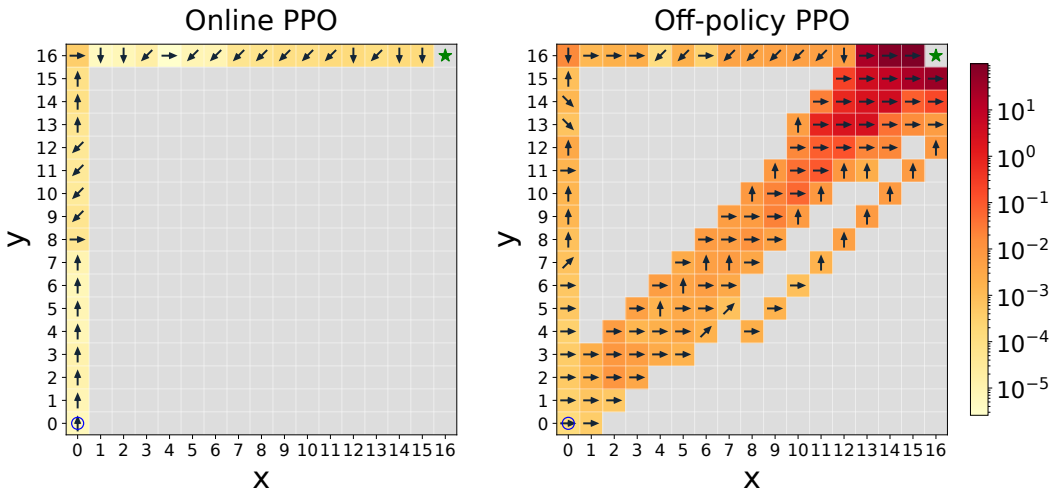


Figure 4: *Exploration guidance* at the first PPO update. Arrows indicate the action with the largest gradient-induced probability increase; color shows the RMS magnitude of the probability response across all eight actions (shared log scale). **Left:** online PPO receives samples only along the perimeter. **Right:** off-policy PPO additionally covers the diagonal corridor via teacher demonstrations, providing *guidance* at prefixes the student has never *explored*. Both panels share the same initial policy and online batch.

PPO is the slowest by a wide margin and degrades steeply as the probability decreases, with one of five seeds failing to converge at 10^{-6}. Figure 1 (bottom right) illustrates this recovery: the median importance ratio on teacher tokens starts as low as 10^{-6} and climbs toward one as the student converges.

Learning dynamics. Figure 1 (top right) shows training curves for the $p = 0.999$ base policy. Off-policy PPO with the diagonal teacher reaches optimal return at a median of 9,600 tokens—$4.5\times$ faster than the detour teacher (43,500) and $17\times$ faster than on-policy PPO's median among its four successful seeds (163,600). The diagonal teacher's curve rises sharply and stabilizes, while the detour teacher climbs in stages, consistent with the weaker advantage margins of the suboptimal detour route. Both off-policy methods reach optimal return at least once in all ten seeds, while on-policy PPO does so in four of ten and shows slower initial progress.

Compute-matched comparison. Off-policy PPO processes twice the tokens per update (online plus teacher), so the speedup could partly reflect extra compute. To control for this, we halve the number of PPO epochs from 4 to 2, roughly matching the per-update training budget. Even with matched compute, off-policy PPO reaches near-optimal return $\sim 10\times$ sooner in online environment tokens (Figure 3). The advantage comes from the teacher's learning signal, not from additional gradient steps.

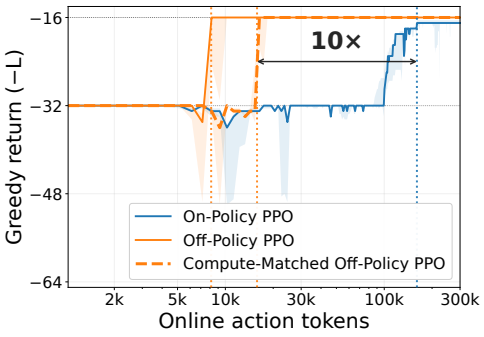


Figure 3: Compute-matched off-policy PPO (dashed) reaches near-optimal return $\sim 10\times$ faster than on-policy PPO.

Exploration guidance. Figure 4 illustrates the exploration guidance mechanism from Section 3.1 at the first PPO update: the teacher's demonstrations supply gradient signal along the diagonal corridor—prefixes the student has never visited—while online PPO receives signal only along the perimeter.

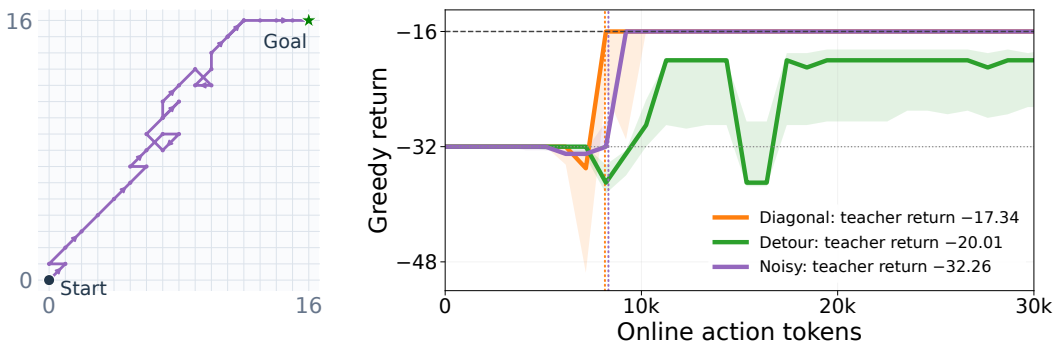


Figure 5: Coverage of the optimal route matters more than teacher return. **Left:** a sampled noisy-teacher trajectory that visits the diagonal corridor despite frequent mistakes (mean return -32). **Right:** off-policy PPO with this noisy teacher first reaches optimal return nearly as fast as with the diagonal teacher (return -17), and faster than the higher-return detour teacher (return -20), though it is less stable. Curves show median greedy return and interquartile range across three paired seeds; vertical dotted lines indicate median first optimality.

What makes a good teacher? Teacher return alone does not predict learning speed. We introduce a *noisy teacher* whose trajectories wander through the diagonal corridor with frequent mistakes, achieving a mean return of only -32—worse than the base policy's perimeter route. Yet off-policy PPO with this noisy teacher first reaches optimal return nearly as fast as with the diagonal teacher, and substantially faster than with the higher-return detour teacher, though it does not sustain optimality as reliably (Figure 5). The difference is coverage: the noisy teacher visits prefixes along the optimal route, providing exploration guidance at states that matter, while the detour teacher follows a fixed suboptimal corridor that never crosses the diagonal. This is consistent with the teacher-guidance condition in Corollary D.2: visitation of relevant prefixes can compensate for weaker advantage margins.

Token-level vs. sequence-level importance sampling. Section 3.2 argues that token-level importance sampling avoids the exponential variance of trajectory-level correction at the cost of a benign bias. Figure 6 confirms this empirically. With the diagonal teacher and the $p = 0.999$ base policy, token-level off-policy PPO reaches optimal return within $\sim$20,000 tokens. Sequence-level off-policy PPO eventually approaches optimality but with substantially higher variance and slower convergence. The on-policy comparison is starker: token-level on-policy PPO converges, while the median of sequence-level on-policy PPO remains at the perimeter return (-32), with only one of three seeds showing improvement and none reaching optimality.

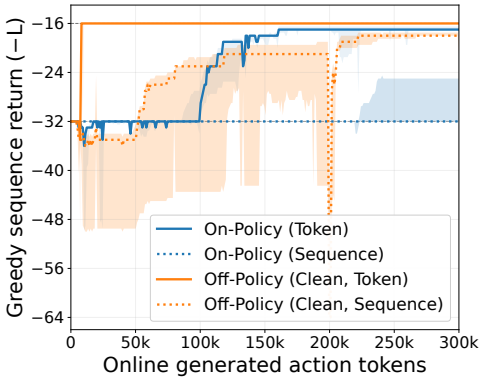


Figure 6: Sequence-level IS either stalls (on-policy) or converges slowly with high variance (off-policy), while token-level IS converges stably.

5 Language Experiments

We test off-policy PPO on three language tasks of increasing difficulty: word reversal (reorder 3–4 words), countdown (find arithmetic expressions hitting a target), and GSM8K (grade-school math word problems). In each task, a small Qwen2.5 (Team, 2025) student learns from demonstrations generated by a larger Qwen2.5 teacher (Table 3). Each training update mixes equal amounts of fresh student rollouts and fixed teacher demonstrations. We compare against supervised fine-tuning (SFT) on teacher demonstrations, on-policy distillation, and on-policy PPO without demonstrations. We

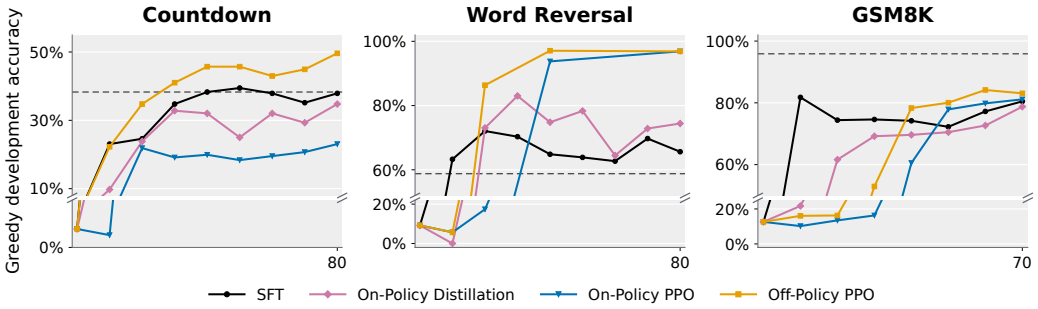


Figure 8: Greedy development accuracy on three language tasks. Off-policy PPO (orange) matches or exceeds all baselines. Dashed line: teacher accuracy. Student and teacher models vary by task (Table 3).

	Countdown		Word Reversal		GSM8K	
Method	Greedy	Sampled	Greedy	Sampled	Greedy	Sampled
SFT	33.0	22.3	64.3	61.1	68.5	**65.1**
OPD	37.1	27.0	73.8	71.3	68.2	63.9
On-Policy PPO	25.0	22.7	95.9	93.7	68.7	51.7
Off-Policy PPO	**48.6**	**44.7**	**97.5**	**95.3**	**72.0**	62.4

Table 1: Held-out test accuracy (%) with greedy and sampled decoding. Off-policy PPO achieves the highest greedy accuracy on all three tasks and the highest sampled accuracy on two of three.

additionally evaluate on a multi-agent voting task with a scripted teacher; off-policy PPO again outperforms baselines (Appendix C). Full setup details—prompt formats, hyperparameters, and evaluation protocols—are in Appendix B.

Main results. Figure 8 shows greedy development accuracy across all three tasks; Table 1 confirms these results on held-out test sets. Off-policy PPO matches or exceeds every baseline. On word reversal, it reaches 97% accuracy, matching on-policy PPO and surpassing both the teacher and SFT. On countdown, it is the only method to substantially exceed the teacher's accuracy. On GSM8K, it reaches comparable accuracy to SFT while using the same number of fresh rollouts. SFT learns quickly but saturates early, exceeding teacher accuracy on word reversal but plateauing on the other two tasks. On-policy PPO eventually reaches high accuracy on word reversal but trails off-policy PPO on countdown, consistent with the exploration-guidance mechanism from Section 4: teacher demonstrations provide gradient signal at prefixes the student would not visit on its own.

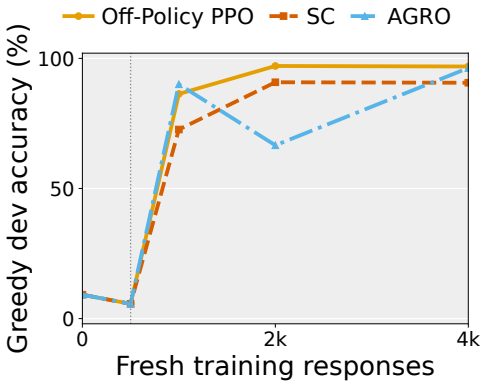


Figure 7: Off-policy PPO vs. AGRO and Score Centering on word reversal. Off-policy PPO converges fastest; AGRO exhibits a transient accuracy drop before recovering.

Comparison to AGRO and Score Centering.
We compare off-policy PPO against AGRO (Tang et al., 2025), designed for mixed on- and off-policy RL, and Score Centering (Marek & Ryabinin, 2026), originally proposed for training–inference mismatch, which we adapt to our mixed-data setting. Figure 7 shows word reversal with the same teacher demonstrations. All three methods eventually reach high accuracy, but off-policy PPO converges fastest. AGRO exhibits a transient accuracy drop mid-training before recovering,

mirroring the instability observed on GridMaze-AR (Section 4). Score Centering converges more slowly but remains stable. The policy-gradient surrogate provides competitive or better performance without additional algorithmic machinery.

6 Related Work

Off-policy reinforcement learning. Off-policy learning decouples data collection from policy optimization. Q-learning (Watkins & Dayan, 1992), off-policy actor–critic methods (Degris et al., 2012; Silver et al., 2014), and experience replay (Mnih et al., 2015; Schaul et al., 2015) established that reusing past experience can dramatically improve sample efficiency. The central difficulty is distribution mismatch: the exact per-decision correction—a product of per-step importance ratios—has variance that can grow exponentially with the horizon (Precup et al., 2000; Liu et al., 2018; Metelli et al., 2020). Practical methods keep the behavior state distribution and apply per-step, often truncated, ratios (Espeholt et al., 2018; Wang et al., 2017), or account for state-distribution effects through emphatic weightings (Imani et al., 2018) or density-ratio estimation (Liu et al., 2019). The surrogates of TRPO and PPO (Schulman et al., 2015a; 2017) make the same per-step approximation; TRPO provides a divergence-dependent improvement bound that weakens as the behavior policy diverges from the current one. External demonstrations have also been incorporated through demonstration-augmented policy gradients (Rajeswaran et al., 2018), guided policy search (Levine & Koltun, 2013), and advantage-gated teacher trajectories (Wu et al., 2026). These works establish methods for learning from off-policy data and managing distribution mismatch.

Post-training of language models. LLM post-training typically separates demonstration learning from reinforcement learning: supervised fine-tuning and distillation (Ouyang et al., 2022; DeepSeek-AI, 2025; Agarwal et al., 2024; Gu et al., 2024; Lu & Thinking Machines Lab, 2025) transfer teacher knowledge, while on-policy methods such as PPO-based RLHF (Ouyang et al., 2022), GRPO (Shao et al., 2024), and DAPO (Yu et al., 2025) optimize from fresh rollouts. As asynchronous and off-policy pipelines scale up training (Noukhovitch et al., 2025; Fu et al., 2025a; Piché et al., 2025), recent work modifies the objective to overcome staleness through prefix-aware updates (Zhang et al., 2026; Lei et al., 2026; Mao et al., 2026; Tian et al., 2026), token masking (Zheng et al., 2025), asymmetric weighting (Le Roux et al., 2025), and trust-region distillation (Xie et al., 2026), which primarily targets distillation stability but also supports off-policy reuse. A complementary direction incorporates external demonstrations into RL with modified importance weighting or auxiliary losses (Yan et al., 2025; Fu et al., 2025b; Zhang et al., 2025; Dong et al., 2025), or conditions rollouts on demonstration prefixes (Huang et al., 2025; Setlur et al., 2026). We show that an unmodified policy gradient surrogate can directly absorb teacher demonstrations under extreme policy mismatch, where importance ratios can be orders of magnitude from one.

7 Conclusion

We showed that vanilla PPO with token-level importance sampling can learn from teacher demonstrations under extreme policy mismatch, without requiring new objectives or auxiliary losses. The theoretical analysis identifies exploration guidance—teacher demonstrations supplying gradient signal at prefixes the student has not yet reached—as the key mechanism, and shows that the teacher's relative gradient contribution depends on advantage margins and visitation, not on the magnitude of the importance ratio. GridMaze-AR confirms this across four orders of magnitude of initial mismatch, and three language tasks show that the same approach matches or exceeds SFT, on-policy distillation, and on-policy PPO. Our experiments use small student models (0.5B–1.5B parameters) and fixed pre-generated demonstrations; whether these findings extend to larger-scale training with online teacher generation remains an open question.

AI Use Statement

We used generative AI tools to help develop and critique parts of the theoretical analysis, including mathematical claims and proofs; review experimental methods and interpret results; and write and inspect research code and configurations. We also used these tools to search and summarize related work, check bibliographic information, draft and edit manuscript text, prepare LaTeX, and generate

figures and figure captions. The authors take responsibility for the paper's final content, including AI-assisted text, claims, and artifacts.

Reproducibility Statement

Sections 4 and 5 describe the experimental comparisons. Appendices A, B, and C provide the task definitions, data partitions, model and teacher choices, training settings, evaluation procedures, and numbers of seeds used for all reported results. Appendix D states the assumptions and gives proofs for the theoretical results.

References

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

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 560–569, 2018.

Lukas Biewald, Thomas Stojnić, and Neil Houlsby. Reasoning gym: Verifiable reasoning environments for LLM RL. *arXiv preprint arXiv:2505.15277*, 2025.

Ching-An Cheng, Andrey Kolobov, and Alekh Agarwal. Policy improvement via imitation of multiple oracles. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/3c56fe2f24038c4d22b9eb0aca78f590-Abstract.html.

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. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. In *International Conference on Machine Learning (ICML)*, 2012.

Yihong Dong, Xue Jiang, Yongding Tao, Huanyu Liu, Kechi Zhang, Lili Mou, Rongyu Cao, Yingwei Ma, Jue Chen, Binhua Li, Zhi Jin, Fei Huang, Yongbin Li, and Ge Li. RL-PLUS: Countering capability boundary collapse of LLMs in reinforcement learning with hybrid-policy optimization. *arXiv preprint arXiv:2508.00222*, 2025.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018. URL https://arxiv.org/abs/1802.01561.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, Tongkai Yang, Binhang Yuan, and Yi Wu. AReaL: A large-scale asynchronous reinforcement learning system for language reasoning. In *Advances in Neural Information Processing Systems*, 2025a.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. SRFT: A single-stage method with supervised and reinforcement fine-tuning for reasoning. *arXiv preprint arXiv:2506.19767*, 2025b.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo M. Ponti, and Ivan Titov. Blending supervised and reinforcement fine-tuning with prefix sampling. *arXiv preprint arXiv:2507.01679*, 2025.

Ehsan Imani, Eric Graves, and Martha White. An off-policy policy gradient theorem using emphatic weightings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 267–274, 2002.

Nicolas Le Roux, Marc G. Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alex Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Toth, and Sam Work. Tapered off-policy REINFORCE: Stable and efficient reinforcement learning for large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Shiye Lei, Zhihao Cheng, and Dacheng Tao. A step back: Prefix importance ratio stabilizes policy optimization. *arXiv preprint arXiv:2601.22718*, 2026. URL https://arxiv.org/abs/2601.22718.

Sergey Levine and Vladlen Koltun. Guided policy search. In *International Conference on Machine Learning (ICML)*, 2013.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

Kevin Lu and Thinking Machines Lab. On-policy distillation. https://thinkingmachines.ai/blog/on-policy-distillation/, 2025.

Renjie Mao, Chao Qu, Yecheng Jason Wang, Bo Chen, Guang Wang, and Jianye Hao. Beyond uniform token-level trust region in LLM reinforcement learning. *arXiv preprint arXiv:2606.10968*, 2026.

Martin Marek and Max Ryabinin. Score centering stabilizes off-policy reinforcement learning. *arXiv preprint arXiv:2609.20807*, 2026.

Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. Importance sampling techniques for policy optimization. *Journal of Machine Learning Research*, 21(141):1–75, 2020. URL https://www.jmlr.org/papers/v21/20-124.html.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, Arian Hosseini, Rishabh Agarwal, and Aaron Courville. Asynchronous RLHF: Faster and more efficient off-policy RL for language models. In *ICLR*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 (NeurIPS)*, 35, 2022.

Jiayi Pan, Jipeng Zhang, Renjie Chen, Zipeng Li, Songyang Wu, and Haozheng Peng. Tinyzero: Reproducing DeepSeek-R1-Zero in countdown and multiplication tasks. https://github.com/Jiayi-Pan/TinyZero, 2025. Countdown task dataset: Jiayi-Pan/Countdown-Tasks-3to4.

Alexandre Piché, Ehsan Kamalloo, Rafael Pardinas, Xiaoyin Chen, and Dzmitry Bahdanau. PipelineRL: Faster on-policy reinforcement learning for long sequence generation. *arXiv preprint arXiv:2509.19128*, 2025.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pp. 759–766. Morgan Kaufmann, 2000.

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *Robotics: Science and Systems (RSS)*, 2018.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015. URL https://arxiv.org/abs/1511.05952.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *arXiv preprint arXiv:1502.05477*, 2015a. URL https://arxiv.org/abs/1502.05477.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Amrith Setlur, Zijian Wang, Andrew Cohen, Paria Rashidinejad, and Sang Michael Xie. Reuse your FLOPs: Scaling RL on hard problems by conditioning on very off-policy prefixes. *arXiv preprint arXiv:2601.18795*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, 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.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning (ICML)*, 2014.

Wen Sun, Arun Venkatraman, Geoffrey J. Gordon, Byron Boots, and J. Andrew Bagnell. Deeply AggreVaTeD: Differentiable imitation learning for sequential prediction. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3309–3318, 2017.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

Yunhao Tang, Taco Cohen, David W. Zhang, Michal Valko, and Rémi Munos. RL-finetuning LLMs from on- and off-policy data with a single algorithm. *arXiv preprint arXiv:2503.19612*, 2025. URL https://arxiv.org/abs/2503.19612.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Minghao Tian, Yunfei Xie, and Chen Wei. How off-policy can GRPO be? μ-GRPO for efficient LLM reinforcement learning. *arXiv preprint arXiv:2605.17570*, 2026.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. In *International Conference on Learning Representations (ICLR)*, 2017.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992.

Zhouchonghao Wu, Raymond Song, Vedant Mundheda, Luis E. Navarro-Serment, Christof Schoenborn, and Jeff Schneider. TADPO: Reinforcement learning goes off-road. *arXiv preprint arXiv:2603.05995*, 2026.

Zhengpeng Xie, Li Lyna Zhang, Zeke Xie, and Mao Yang. Trust region policy distillation. *arXiv preprint arXiv:2607.04751*, 2026.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance, 2025.

Qiying Yu et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. On-policy RL meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. *arXiv preprint arXiv:2508.11408*, 2025.

Yuheng Zhang, Chenlu Ye, Shuowei Jin, Changlong Yu, Wei Xiong, Saurabh Sahu, and Nan Jiang. Rethinking importance sampling in LLM policy optimization: A cumulative token perspective. *arXiv preprint arXiv:2605.07331*, 2026.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy RL reach with stale data on LLMs? *arXiv preprint arXiv:2510.01161*, 2025.

A GRIDMAZE-AR: EXPERIMENTAL DETAILS AND ABLATIONS

Task. GridMaze-AR is an autoregressive navigation task on a 17×17 grid. The agent starts at $(0, 0)$ and must reach $(16, 16)$ using one of eight movement directions (N, NE, E, SE, S, SW, W, NW) at each step. An action whose destination lies outside the grid leaves the position unchanged but still consumes a token. The episode terminates on the first goal visit or after $H = 64$ actions. The setting is fully autoregressive: both actor and critic observe only the preceding action tokens—the agent receives no coordinates, no environment observations, and no intermediate reward feedback. Grid position is reconstructed internally for reward computation but is never supplied as a network input. Different action prefixes that reach the same grid cell remain distinct model inputs. The optimal route follows the diagonal (NE^{16}, 16 steps); the perimeter base route takes 32 steps.

Architecture. The actor and critic are separate causal Transformers with no shared parameters. Each has one layer, embedding dimension 64, four attention heads, feedforward dimension 128, GELU activations, and learned positional embeddings. The maximum context is 64 tokens. A beginning-of-sequence token precedes the action history. The actor outputs eight logits whose softmax defines $\pi_\theta(a \mid h_t)$; the critic outputs a scalar value $V_\phi(h_t)$.

Base policy and mismatch sweep. The actor is pretrained with cross-entropy on soft labels to follow the perimeter route: north below the top row, east on the top row. The preferred perimeter action receives probability p and the remaining seven actions each receive $(1 - p)/7$. We train a single base policy at $p = 0.999$ and construct the initial-mismatch sweep by rescaling the output-layer logits with a temperature that maps the trained odds to each target probability. The resulting optimal-action (diagonal) probability $(1 - p)/7$ spans roughly four orders of magnitude across the sweep.

Teachers. The *diagonal teacher* is an epsilon-soft policy with preferred-action probability $p_T = 0.95$: it prefers NE in the interior and recovers toward the diagonal after deviations (16-step optimal route). The *detour teacher* follows $\mathrm{N}^4\mathrm{NE}^{12}\mathrm{E}^4$ (20 steps, return -20) with analogous recovery rules; it uses $p_T = 0.95$ in the initial-mismatch sweep and $p_T = 0.999$ in the remaining experiments. The *noisy teacher* wanders through the diagonal corridor with a preferred-action probability of approximately 0.60 and a deadline-feasibility mask that guarantees reaching the goal; its corpus averages return -32. Each teacher corpus consists of 256 episodes generated once with a fixed seed and reused across all training runs.

Training. We use PPO with clipping width $\epsilon = 0.2$, GAE with $\lambda = 0.95$, and discount $\gamma = 1$. The training reward is terminal: a completed episode of length L receives reward $(64 - L)/48$, mapping the optimal 16-step solution to reward 1 and timeouts to 0. Each update collects $\sim$1,024 fresh online tokens; with $\beta = 1$, an equal number of teacher tokens are sampled per update. Advantages are whitened jointly across online and offline tokens. We run 4 PPO epochs per update with sequence-preserving minibatches of 256 tokens. Actor and critic use separate Adam optimizers ($\epsilon_{\mathrm{Adam}} = 10^{-5}$). The critic learning rate is 3×10^{-4}; the actor starts at 10^{-5} for the first two updates, ramps linearly to 3×10^{-4} over the next eight updates, and holds. The entropy coefficient is 0.01 and the value-loss coefficient is 0.5. Gradients are clipped to norm 1.0. Before PPO begins, the critic is calibrated for 50 epochs on a held-out rollout set (256 training and 128 validation sequences) against Monte Carlo return targets. The standard training budget is $300{,}000$ online tokens.

Evaluation. We evaluate after every PPO update with a single greedy decode from the fixed start position. Convergence is defined as the first update achieving greedy return ≥ -17 (within one step of optimal). Learning curves report greedy return against newly generated online tokens; teacher-token reuse and evaluation episodes do not advance this counter. The main learning curves use ten seeds; the initial-mismatch sweep uses five seeds per condition; the compute-matched and teacher-source ablation experiments use three paired seeds.

A.1 COMPARISON TO AGRO ON GRIDMAZE-AR

AGRO (Tang et al., 2025) is a recent algorithm designed for mixed on- and off-policy RL fine-tuning. Figure 9 compares off-policy PPO and AGRO on GridMaze-AR with the $p = 0.999$ base

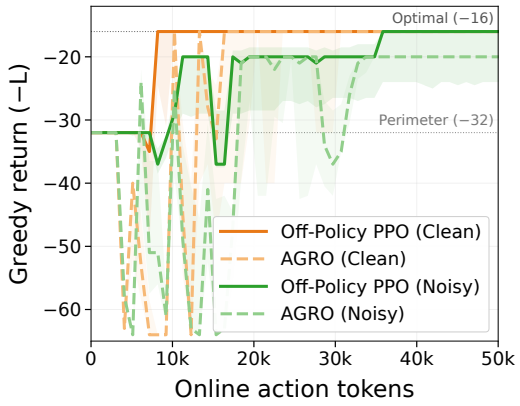


Figure 9: Off-policy PPO vs. AGRO on GridMaze-AR ($p = 0.999$). Solid: PPO. Dashed: AGRO. Both learn from demonstrations, but AGRO shows large training instability, especially with the detour teacher.

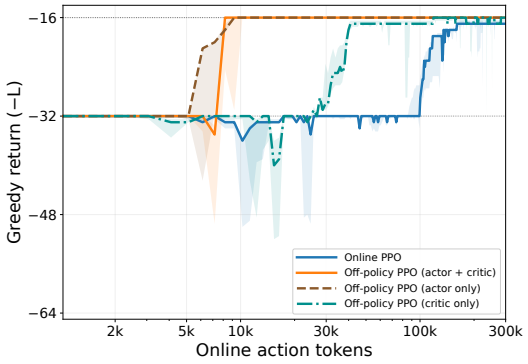


Figure 10: Actor vs. critic training with teacher data ($p = 0.999$, diagonal teacher). Training the actor on teacher demonstrations (dashed brown) converges nearly as fast as training both (solid orange). Training only the critic (dash-dot teal) is much slower and less stable.

policy. With the diagonal teacher, AGRO reaches near-optimal return but exhibits large transient dips during training. With the detour teacher, the instability is more pronounced: AGRO's greedy return drops sharply before partially recovering, while off-policy PPO converges monotonically. Both methods ultimately learn from demonstrations, but the policy-gradient surrogate provides more stable optimization under extreme policy mismatch.

A.2 Actor vs. critic training

Teacher demonstrations enter both the actor and critic losses. To isolate which channel drives the speedup, we train the actor only or the critic only on teacher data, keeping the other component on-policy. Training the actor on teacher demonstrations accounts for nearly all of the benefit: actor-only off-policy PPO converges almost as fast as the full actor+critic variant (Figure 10). Training only the critic is much slower and exhibits large transient dips. It still converges faster than online PPO ($\sim 4\times$), but accounts for only a fraction of the full off-policy speedup. This is consistent with the exploration-guidance mechanism: the actor directly receives gradient signal at teacher-visited prefixes, while a better critic baseline provides only an indirect benefit.

B LANGUAGE EXPERIMENTS: EXPERIMENTAL DETAILS AND ABLATIONS

Tasks. We use three tasks of increasing difficulty. *Countdown* (Pan et al., 2025): given three numbers and a target, produce an arithmetic expression that uses every number exactly once and evaluates to the target. Correctness is checked with exact rational arithmetic. *Word reversal* (Biewald et al., 2025): given a list of 3–4 words, return them in reverse order. The scorer requires an exact case-sensitive match of the reversed word list. *GSM8K* (Cobbe et al., 2021): solve a grade-school math word problem and report the numerical answer after a #### marker. The scorer extracts and compares the final number. In all three tasks, the reward is binary whole-answer correctness assigned at the final response token; prompt tokens are excluded from the training loss.

Data. Table 2 gives the prompt partitions. Training, critic-warmup, development, and test splits are disjoint. All plotted curves report development-set accuracy.

Task	Training prompts	Critic warmup	Development	Reserved test
Countdown	640	128	256	512
Word Reversal	4,000	256	512	1,024
GSM8K	3,500	1,000	461	1,319

Table 2: Prompt counts for each task partition.

Models and demonstrations. All models belong to the Qwen2.5 family (Table 3). Demonstrations are generated once per training prompt and retained without correctness filtering or rejection sampling; off-policy PPO and SFT use the same demonstration bank within each task. For Countdown, the teacher generates with a reasoning prompt and 1,024-token budget; we strip the reasoning and keep only the final answer expression in student-visible demonstrations. Word Reversal demonstrations are bare answer lists; GSM8K demonstrations retain the full solution and final answer.

Task	Student	Teacher	Student token cap
Countdown	0.5B-Instruct	7B-Instruct	64
Word Reversal	0.5B-Instruct	1.5B-Instruct	96
GSM8K	1.5B-Base	14B-Instruct	512

Table 3: Student and teacher checkpoints.

Training. PPO uses token-level importance ratios, clipping width $\epsilon = 0.2$, $\gamma = \lambda = 1$, and a frozen-initial-reference KL penalty with coefficient 0.01. On-policy PPO uses only fresh student responses. Off-policy PPO adds cached teacher demonstrations: 32 student plus 32 teacher responses per Countdown update, and 50 plus 50 for Word Reversal and GSM8K. Advantages are whitened jointly across online and offline response tokens. SFT minimizes negative log-likelihood on cached teacher responses. On-policy distillation generates fresh student responses and obtains frozen teacher feedback on their prefixes.

Per-task optimization details: Countdown uses learning rate 10^{-6} and 4 optimizer steps per update. Word Reversal PPO uses 3×10^{-7} and 10 steps per update; its SFT and distillation baselines use 10^{-6}. GSM8K PPO and distillation use 3×10^{-7} and 1 step per update; SFT uses 3×10^{-6}. All methods sample at temperature 0.8 with top-$p = 1$.

Unit-denominator convention. For Countdown, teacher demonstration tokens use a unit denominator in the PPO ratio:

$$r_t^{\text{demo}}(\theta) = \pi_\theta(a_t \mid s_t).$$

This convention was inherited from preliminary experiments and retained for comparability. It is an approximation: the actual teacher likelihood is not recorded, so the ratio does not perform exact importance-sampling correction. Word Reversal and GSM8K use recorded teacher log-probabilities as the denominator. Fresh student responses use their recorded collection probabilities in all three tasks. We explore the effect of this denominator choice in Section B.2.

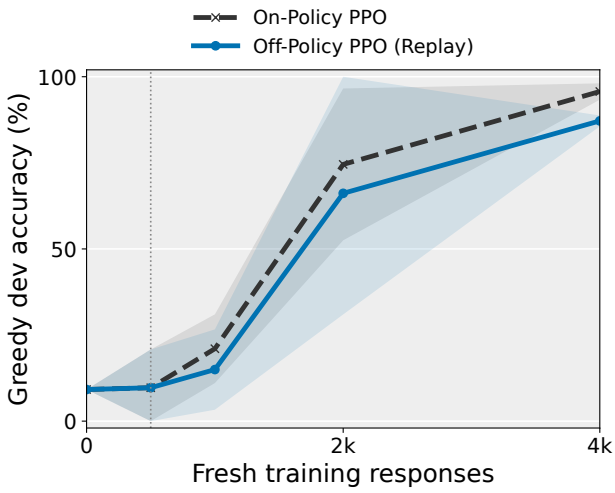


Figure 11: Experience replay on word reversal. Off-policy PPO with replay (the agent's own past generations) nearly matches on-policy PPO, suggesting the same surrogate extends beyond teacher demonstrations to self-generated data.

Evaluation. Learning curves report greedy whole-answer development accuracy from a single training seed, evaluated at saved checkpoints (generally every ten updates for Countdown and GSM8K; at updates 0, 10, 20, 40, and 80 for Word Reversal). Countdown and Word Reversal run for 80 updates; GSM8K runs to update 70. The teacher reference lines are 38.3% for Countdown (sampled generation on calibration puzzles), 58.8% for Word Reversal (greedy on the development set), and 95.9% for GSM8K (greedy on the development set).

Held-out test accuracy (Table 1) uses the final checkpoints: update 80 for Countdown and Word Reversal, update 70 for GSM8K (epoch 1 for SFT). We evaluate on 512 Countdown puzzles, 1,024 Word Reversal questions, and the 1,319 official GSM8K test problems. We report accuracy under both greedy and sampled decoding (one response per question at temperature 0.8, top-$p = 1$). All results are from a single training seed per method.

Additional Word Reversal comparisons. The AGRO and Score Centering runs use the same 50/50 online–offline batches and demonstration bank as off-policy PPO. The replay experiment replaces teacher demonstrations with 50 historical student responses sampled uniformly from the agent's own replay buffer, alongside 50 fresh responses per update; it uses 20 actor optimizer steps per update and reports mean $\pm$ standard deviation across three seeds.

B.1 Replay as off-policy data

An appealing application of off-policy learning is experience replay: reusing the agent's own past generations alongside fresh rollouts. We test this on word reversal, where each update uses 50 fresh rollouts and 50 historical responses sampled uniformly from a replay buffer (no teacher data). Figure 11 shows that off-policy PPO with replay nearly matches on-policy PPO on word reversal, suggesting that the mechanisms studied in this paper extend beyond fixed teacher demonstrations to self-generated replay. We defer a detailed investigation of replay strategies to future work.

B.2 Teacher choice and demonstration weighting for Countdown

Figure 12 crosses two teacher sizes (Qwen2.5-7B-Instruct and Qwen3.8-27B-FP8) with two importance-sampling denominator choices on Countdown. Two surprising patterns emerge. First, the larger 27B teacher does not produce a stronger student: under both denominator choices, the 7B teacher yields higher final accuracy (49.6% vs. 39.5% with unit denominators). Second, replacing the unit denominator with recorded teacher likelihoods does not help—and may hurt: with the 7B

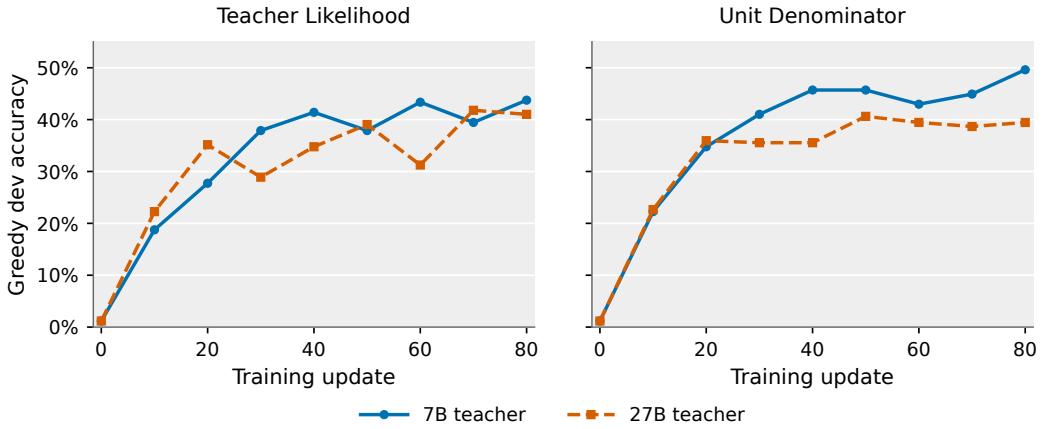


Figure 12: Teacher choice and IS denominator on Countdown. **Left:** teacher-likelihood denominators. **Right:** unit denominators ($r_t = \pi_\theta(a_t \mid s_t)$). The Qwen2.5-7B teacher outperforms the Qwen3.8-27B teacher under both denominator choices. Single seed.

teacher, teacher-likelihood PPO reaches 43.8% versus 49.6% for unit-denominator PPO. Since the 7B teacher and student share a tokenizer, this gap cannot be attributed to cross-tokenizer misalignment. These observations parallel the GridMaze finding that teacher quality alone does not predict student learning speed (Section 4), and hint at an interaction between teacher behavior and demonstration weighting. However, the teacher's likelihoods condition on internal reasoning hidden from the student, so this comparison does not cleanly isolate the effect of the IS denominator. These are single-seed runs; understanding the mechanisms behind these patterns is an open question.

C Multi-agent voting task

Task. We consider a multi-agent voting task with three agents, five candidate alternatives, and three rounds of voting. Each agent has an individual utility for each candidate. A candidate is selected if it receives at least two votes in the final round; otherwise, no candidate is selected. Each agent's game payoff is its utility for the selected candidate, or zero if no candidate is selected. We evaluate the *social optimum rate*: the fraction of games in which the selected candidate maximizes the sum of the three agents' utilities. Thus, selecting a candidate is not sufficient for success; the collective choice must also maximize total welfare.

Model and demonstrations. The student checkpoint is Qwen2.5-0.5B-Base. The teacher is a scripted policy based on the previous round's plurality, with a 5% probability of taking a uniformly random action and no privileged information. For each training seed, a fixed bank of 1,000 teacher games is retained without filtering. We compare on-policy PPO, off-policy PPO using this bank, and on-policy distillation (OPD) from the same teacher.

Training. All runs use 60 training updates, each collecting 32 fresh student games. Off-policy PPO additionally samples 32 teacher games per update and uses the unit-denominator convention, $r_t^{\text{demo}}(\theta) = \pi_\theta(a_t \mid s_t)$. PPO uses clipping width $\epsilon = 0.2$, discount $\gamma = 1$, GAE parameter $\lambda = 0.95$, and entropy coefficient 0.01. Actor and critic learning rates are 3×10^{-6} and 10^{-5}, respectively, with two optimization epochs per update, minibatches of 96 tokens, and gradient norm clipping at 1. Advantages are normalized jointly across fresh and teacher data. The OPD baseline uses the same hyperparameters. The critic is calibrated on held-out games before training begins; all methods start from the same initialization. The comparison matches the number of fresh games, not total optimization work.

Evaluation. We use three training seeds and evaluate on 256 development games at updates 0, 1, 5, 10, 20, 30, 40, 50, and 60. A separate set of 1,024 test games is evaluated at the final checkpoint;

the figure reports development results only. Curves show means across seeds and shaded bands show their minimum and maximum. The horizontal reference is the teacher's mean social optimum rate on its demonstration banks (51.3%), rather than a development-set teacher evaluation.

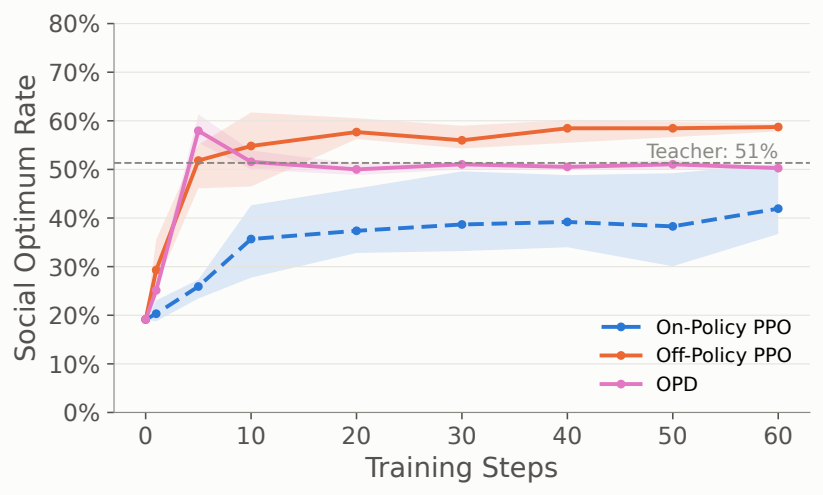


Figure 13: Learning with a scripted teacher in the voting task. Off-policy PPO uses a unit denominator for teacher actions. Training steps denote updates, each collecting 32 fresh games; off-policy PPO also uses 32 teacher games. Lines show mean development social optimum rates across three seeds; bands show the seed range. The dashed gray line is the teacher-bank reference.

Results. Figure 13 shows that teacher demonstrations improve learning at a fixed fresh-game budget. By update 10, off-policy PPO reaches a mean social optimum rate of 54.8%, compared with 35.7% for on-policy PPO. At update 60, off-policy PPO reaches 58.7%, versus 41.9% for on-policy PPO and 50.3% for OPD. OPD improves rapidly at first but subsequently settles near 50%, whereas off-policy PPO maintains a higher rate.

D THEORY OF OFF-POLICY LEARNING WITH TOKEN-LEVEL CORRECTION

This appendix formalizes the theoretical claims in Section 3. We first establish shared notation, then develop three results: (1) the mixed surrogate's learning direction, (2) bias and variance of token-level correction, and (3) gradient strength under small importance ratios. Each subsection corresponds to a subsection of the main text. All three results concern a frozen, unclipped population surrogate.

Notation and setup. Consider a finite-horizon task with horizon H. A *state* s is a time-indexed prefix—it encodes both the position in the sequence and all tokens generated so far. We write π_{old} for the frozen online rollout policy and μ for the teacher's (demonstration) policy. The candidate policy being optimized is π_θ, with $\pi_{\theta_0} = \pi_{\text{old}}$. Assume a common task, initial-state distribution, and reward function; these results do not establish transfer across different reward functions. We use total expected episode return J and normalized state frequencies, so the surrogate gradient at equality is $H^{-1}\nabla J$. All action probabilities are positive on the states considered. This support assumption makes the importance ratios and cancellations valid.

Each source $\rho \in \{\pi_{\text{old}}, \mu\}$ induces a distribution over states. We write $d_\rho(s) = \Pr_\rho(s)/H$ for the *normalized state frequency*: the fraction of time steps at which source ρ visits state s, averaged over the horizon. The action value $Q^\rho(s,a)$ is the expected remaining reward from taking action a at state s and then following source ρ's own policy. A *baseline* $b_\rho(s)$ is any function of the state alone (not the action); the *advantage* is $\overline{A}_\rho(s,a) = Q^\rho(s,a) - b_\rho(s)$.

The chosen-action direction. Several results below analyze a specific one-dimensional change to the policy: fix a state s and a single "chosen" action $a_\star$. Write $\varepsilon = \pi_{\text{old}}(a_\star \mid s)$ for the student's

current probability of choosing $a_\star$, and $w_b = \pi_{\text{old}}(b \mid s)/(1-\varepsilon)$ for the conditional probability of each alternative $b \neq a_\star$ given that $a_\star$ is not chosen, so $\sum_{b \neq a_\star} w_b = 1$. We vary $a_\star$'s probability through its *log-odds* x, defined by the logistic function $\sigma(x) = 1/(1+e^{-x})$:

$$\pi_x(a_\star \mid s) = \sigma(x), \qquad \pi_x(b \mid s) = [1-\sigma(x)]\, w_b, \qquad x_0 = \log \frac{\varepsilon}{1-\varepsilon}. \tag{7}$$

This holds the relative proportions among alternatives fixed while changing only $a_\star$'s share. In softmax logit coordinates, it corresponds to shifting only $a_\star$'s logit. The initial point x_0 recovers the student's current distribution.

Advantage margin. The *advantage margin* compares $a_\star$ against the student-weighted average of all alternatives:

$$\Delta_\rho(s) = \overline{A}_\rho(s, a_\star) - \sum_{b \neq a_\star} w_b\, \overline{A}_\rho(s, b). \tag{8}$$

A positive margin $\Delta_\rho(s) > 0$ means that source ρ's targets favor $a_\star$ over the student's typical alternative. A state-only shift in the targets (adding or changing the baseline) cancels from Δ_ρ because $\sum_b w_b = 1$.

Mixed surrogate. We analyze the unclipped population surrogate for a 50/50 online–offline batch, abstracting away PPO clipping. With frozen advantage targets—each assumed to have conditional mean $\overline{A}_\rho(s,a)$, an assumption not automatic for GAE—the equal-mixture surrogate for a candidate policy π_θ is

$$L(\pi_\theta; \pi_{\text{old}}, \mu) = \tfrac{1}{2}\mathbb{E}_{d_{\pi_{\text{old}}}, \pi_{\text{old}}}\left[\frac{\pi_\theta(a \mid s)}{\pi_{\text{old}}(a \mid s)}\, \overline{A}_{\pi_{\text{old}}}(s,a)\right] + \tfrac{1}{2}\mathbb{E}_{d_\mu, \mu}\left[\frac{\pi_\theta(a \mid s)}{\mu(a \mid s)}\, \overline{A}_\mu(s,a)\right]. \tag{9}$$

The first term is the standard unclipped policy-gradient surrogate for online data; the second adds the teacher's contribution. The importance ratios $\pi_\theta(a \mid s)/\rho(a \mid s)$ reweight each source's actions to evaluate the candidate π_θ. All source policies, frequencies, and targets are held fixed while varying π_θ.

What changes with practical advantage targets? For arbitrary frozen targets, write $B_\rho(s,a) = \mathbb{E}[\widehat{A}_\rho \mid s, a]$. Action correction still gives

$$L_B(\pi_\theta) = \tfrac{1}{2}\sum_{s,a} \pi_\theta(a \mid s)[d_{\pi_{\text{old}}}(s) B_{\pi_{\text{old}}}(s,a) + d_\mu(s) B_\mu(s,a)].$$

Consequently, the probability-transfer and directional-gradient identities hold with margins formed from B_ρ. Interpreting those margins as source action-value differences additionally requires $B_\rho = Q^\rho - b_\rho$. GAE with an approximate critic need not satisfy this condition. The equality-to-return-gradient result also requires its stated unbiased-target assumption. A fixed common positive rescaling and state-only shift preserve margin signs; random batch-dependent whitening need not preserve the population identities. PPO clipping changes which samples contribute derivatives. We therefore do not identify the frozen surrogate analysis with the complete practical update.

D.1 Off-policy learning direction

This subsection supports Section 3.1: teacher demonstrations can guide the student toward better actions, and reaching the teacher's level does not stop improvement.

A key simplification. The importance ratios cancel the source policies in the denominators. This means the surrogate depends on π_θ only through how it distributes probability over actions, not through how far π_θ is from either source:

Proposition D.1 (Mixed-surrogate action preference)**.** *Define the combined action score $F(s,a) = d_{\pi_{\text{old}}}(s)\, Q^{\pi_{\text{old}}}(s,a) + d_\mu(s)\, Q^\mu(s,a)$, which weights each source's evaluation of action a by how often that source visits state s. Then*

$$L(\pi_\theta; \pi_{\text{old}}, \mu) = \tfrac{1}{2}\sum_s \sum_a \pi_\theta(a \mid s)\, F(s,a) \;-\; C, \tag{10}$$

where C is a constant independent of π_θ. Transferring probability $\delta > 0$ from action a to action b at state s changes the surrogate by exactly

$$L(\pi'; \pi_{\text{old}}, \mu) - L(\pi_\theta; \pi_{\text{old}}, \mu) = \frac{\delta}{2}\big[d_{\pi_{\text{old}}}(s)\,\Delta_{\pi_{\text{old}}} + d_\mu(s)\,\Delta_\mu\big], \tag{11}$$

where $\Delta_\rho = Q^\rho(s,b) - Q^\rho(s,a)$ is how much source ρ's evaluation prefers b over a.

Proof. For either source ρ, expand the expectation: $\sum_a \rho(a \mid s)\frac{\pi_\theta(a|s)}{\rho(a|s)}\overline{A}_\rho(s,a) = \sum_a \pi_\theta(a \mid s)\,Q^\rho(s,a) - b_\rho(s)$. The importance ratio cancels the source policy in the denominator, leaving only the candidate's probabilities multiplied by the action values. Weighting by prefix frequencies and summing the two source halves gives (10). The transfer changes only the probabilities of a and b, contributing $\delta[F(s,b) - F(s,a)]/2$. □

In words: the surrogate prefers transferring probability toward whichever action has a higher frequency-weighted evaluation across both sources. The teacher's influence comes from its action-value differences at the states it visits frequently, not from the size of the importance ratio. The importance-sampling cancellation itself is standard. Our use of it is specific to the mixed online–offline setting: the surrogate combines two source-specific value functions $Q^{\pi_{\text{old}}}$ and Q^μ, each weighted by its own visitation distribution. This is related to the teacher-value-based learning in AggreVaTeD (Sun et al., 2017), which also uses teacher action values to guide a student; the difference is that our surrogate arises from a mixture of two data sources rather than from an interactive oracle.

Teacher guidance under weak student preferences. The main text considers increasing the probability of $a_\star$ while proportionally reducing all alternatives. This is the natural direction for selective imitation: adopt one teacher-favored action without committing to the teacher's full distribution.

Corollary D.2 (Teacher guidance under weak student preferences)**.** *Suppose the teacher strongly favors $a_\star$ ($\Delta_\mu(s) \geq m_s > 0$), the student has at most a weak preference ($|\Delta_{\pi_{\text{old}}}(s)| \leq e_s$), and the teacher visits this state often enough that $d_\mu(s)\,m_s > d_{\pi_{\text{old}}}(s)\,e_s$. Then increasing the probability of $a_\star$ by any $0 < \delta \leq 1 - \varepsilon$ while reducing alternatives proportionally strictly improves the surrogate:*

$$\begin{aligned} L(\pi_\delta; \pi_{\text{old}}, \mu) - L(\pi_{\text{old}}; \pi_{\text{old}}, \mu) &= \frac{\delta}{2}[d_{\pi_{\text{old}}}(s)\,\Delta_{\pi_{\text{old}}}(s) + d_\mu(s)\,\Delta_\mu(s)] \\ &\geq \frac{\delta}{2}[d_\mu(s)\,m_s - d_{\pi_{\text{old}}}(s)\,e_s] > 0. \end{aligned} \tag{12}$$

Proof. The proportional transfer removes δw_b from each alternative b and adds δ to $a_\star$. This is equivalent to summing pairwise transfers from each b to $a_\star$, weighted by w_b. Applying (11) to each pair and using the definition of Δ_ρ gives the identity. The assumed bounds on $\Delta_\mu(s)$ and $|\Delta_{\pi_{\text{old}}}(s)|$ yield the lower bound. □

Matching the teacher does not stop improvement. When $\pi_{\text{old}} = \mu$, both halves of the batch come from the same distribution. By the policy-gradient theorem (Sutton et al., 1999), the mixed gradient is proportional to the standard return gradient:

Corollary D.3 (Gradient at equality)**.** *Let $\pi_{\text{old}} = \mu = \pi_{\theta_0}$ with differentiable, positive action probabilities. If both sources have conditionally unbiased targets $\mathbb{E}[\widehat{A}_\rho \mid s,a] = A^{\pi_{\text{old}}}(s,a)$, then the mixed gradient at equality is proportional to the return gradient: $g_{\text{mix}}(\theta_0) = H^{-1}\nabla_\theta J(\theta_0)$. PPO clipping is locally inactive at equality and does not change this conclusion. If $\nabla J(\theta_0) \neq 0$, a sufficiently small step in the gradient direction strictly improves return beyond the teacher's.*

Proof. At $\pi_{\text{old}} = \mu$, all importance ratios are 1. The mixed gradient reduces to the standard policy-gradient expression $\mathbb{E}_{d_{\pi_{\text{old}}},\pi_{\text{old}}}[A^{\pi_{\text{old}}}(s,a)\,\nabla_\theta \log \pi_{\theta_0}(a \mid s)] = H^{-1}\nabla_\theta J(\theta_0)$ by the policy-gradient theorem. PPO's clip at $[1-\epsilon, 1+\epsilon]$ is inactive in a neighborhood of θ_0. □

D.2 TOKEN-LEVEL BIAS AND VARIANCE

Section 3.2 claims that token-level importance sampling introduces bias but avoids the exponential variance of trajectory-level correction, and that the resulting bias is compatible with improvement over the teacher.

Benign bias (restating an established result). The benign-bias argument in the main text applies the performance-difference lemma of Cheng et al. (2020). We restate the key result here for completeness.

Lemma D.4 (Performance-difference identity; Cheng et al. (2020, Lemma 1), cf. Kakade & Langford (2002); Schulman et al. (2015a))**.** *Let π and μ be policies in a finite-horizon task with horizon H and bounded rewards. Define the teacher-relative margin $m_{\pi|\mu}(s) = \sum_a \pi(a \mid s)\, A^\mu(s,a)$. Then*

$$J(\pi) - J(\mu) = H\, \mathbb{E}_{s\sim d_\pi}[m_{\pi|\mu}(s)].$$

In particular, if $m_{\pi|\mu}(s) \geq 0$ at every state the student visits, then $J(\pi) \geq J(\mu)$ (Cheng et al., 2020, Corollary 1). No closeness between the policies, their prefix distributions, or their continuation values is required.

For the off-policy surrogate, this means that retaining teacher continuation values in the gradient does not prevent improvement beyond teacher return, provided the student's action choices are at least as good as the teacher's under teacher evaluation. The condition concerns student-visited prefixes, while the teacher surrogate weights prefixes by teacher visitation. Improving the surrogate does not ensure nonnegative margins at every student-visited prefix, so this establishes compatibility with improvement rather than a guarantee for the biased token-level update.

Token-gradient variance. Full trajectory importance correction multiplies per-token ratios across the entire sequence, producing weights that can be exponentially large. Token-level correction avoids this by treating each token independently.

The following bound specializes standard importance-sampling moment analysis (cf. Metelli et al., 2020, Lemma 1) to the token-level gradient estimator in our setting. The conditional second-moment assumption on the importance weights (C_2 below) is equivalent to bounding the per-token χ^2-divergence, a standard condition in off-policy evaluation (Metelli et al., 2020).

Proposition D.5 (Token-gradient second moment)**.** *Define $\psi(s,a) = \nabla_\theta \log \pi_\theta(a \mid s)|_{\theta=\theta_0}$. Write $r = \pi_{\mathrm{old}}(a \mid s)/\mu(a \mid s)$ for the token-level importance ratio. Assume:*

1. *The conditional importance-weight second moment is bounded: $C_2(s) = \sum_a \pi_{\mathrm{old}}(a \mid s)^2/\mu(a \mid s) \leq C_2$ for all states. (This is equivalent to bounding the χ^2-divergence between $\pi_{\mathrm{old}}(\cdot \mid s)$ and $\mu(\cdot \mid s)$.)*
2. *Advantage targets have bounded second moment: $\mathbb{E}[\widehat{A}^2 \mid s,a] \leq M_A^2$.*
3. *Policy gradients are bounded: $\|\psi(s,a)\| \leq M_\psi$.*

Then for a single token-level gradient sample $X = r\,\widehat{A}\,\psi(s,a)$,

$$\mathbb{E}\,\|X\|^2 \leq C_2\, M_A^2\, M_\psi^2. \tag{13}$$

The bound has no explicit horizon factor, although its constants may depend on H. Averaging n independent trajectory-level estimates gives variance at most $C_2\, M_A^2\, M_\psi^2/n$.

Proof. Condition on the state and action:

$$\begin{aligned}\mathbb{E}\,\|X\|^2 &= \mathbb{E}_{s\sim d} \sum_a \mu(a \mid s)\, r^2\, \mathbb{E}[\widehat{A}^2 \mid s,a]\ \|\psi(s,a)\|^2 \\ &\leq M_A^2\, M_\psi^2\, \mathbb{E}_s[C_2(s)] \leq C_2\, M_A^2\, M_\psi^2.\end{aligned}$$

The key step is that $\sum_a \mu(a \mid s)\, r^2 = C_2(s)$: the teacher's sampling probabilities in the denominators combine with the squared ratio to give a bounded quantity, without any product over time steps.

For tokens within a single trajectory that may be correlated, Jensen's inequality gives $\left\|H^{-1}\sum_t X_t\right\|^2 \le H^{-1}\sum_t \|X_t\|^2$, so the same per-token bound holds for each trajectory's average. Averaging n independent trajectories then provides the $1/n$ variance reduction. □

Trajectory correction can be exponentially harder. Under full trajectory importance weighting, the weight is the product $W = \prod_{t=0}^{H-1} r_t$, and its second moment can grow exponentially with the horizon (see Metelli et al., 2020, for a detailed analysis). As a concrete example: when per-token ratios are conditionally independent with $\mathbb{E}_\mu[r_t^2] = 1 + 4\delta^2$ for a small δ, we get $\mathbb{E}_\mu[W^2] = (1+4\delta^2)^H$. Token-level correction avoids this product: each token contributes through its own ratio, not through a product of all preceding ratios. This does not make tokens statistically independent—they share a trajectory—but it removes the multiplicative accumulation of ratio variance across time steps.

D.3 Gradient strength and sign reliability

Section 3.3 claims that a small importance ratio does not make the teacher's gradient contribution negligible relative to the online contribution. We prove this using the chosen-action direction and advantage margin defined in the shared setup above.

Proposition D.6 (Relative source strength along the chosen-action direction)**.** *The surrogate derivative along* (7) *for each source is*

$$g_{\pi_{\text{old}}}(s) = \varepsilon(1-\varepsilon)\,\Delta_{\pi_{\text{old}}}(s), \qquad g_\mu(s) = \varepsilon(1-\varepsilon)\,\Delta_\mu(s). \tag{14}$$

Both sources share the same factor $\varepsilon(1-\varepsilon)$. The ratio of their prefix-weighted contributions $h_\rho(s) = \frac{1}{2} d_\rho(s)\, g_\rho(s)$ is therefore

$$\frac{|h_\mu(s)|}{|h_{\pi_{\text{old}}}(s)|} = \frac{d_\mu(s)\,|\Delta_\mu(s)|}{d_{\pi_{\text{old}}}(s)\,|\Delta_{\pi_{\text{old}}}(s)|}. \tag{15}$$

The ratio is defined when $d_{\pi_{\text{old}}}(s) > 0$ and $\Delta_{\pi_{\text{old}}}(s) \neq 0$. Both derivatives remain valid when this denominator vanishes, but their ratio does not. The token importance ratio does not appear in the relative-strength expression.

Proof. Substitute (7) into the conditional surrogate for source ρ:

$$L_\rho(x; s) = \sigma(x)\,\overline{A}_\rho(s, a_\star) + [1-\sigma(x)] \sum_{b \neq a_\star} w_b\,\overline{A}_\rho(s, b).$$

Differentiate with respect to x and evaluate at x_0. The only x-dependent terms are $\sigma(x)$ and $1-\sigma(x)$, and $\sigma'(x_0) = \varepsilon(1-\varepsilon)$. This gives (14). The common $\varepsilon(1-\varepsilon)$ cancels in the ratio, yielding (15). □

The intuition is straightforward: both the online and teacher gradients are "small" for the same reason—the student rarely chooses $a_\star$, so the $\varepsilon(1-\varepsilon)$ factor suppresses both equally. What determines which source dominates is the advantage margin Δ_ρ, weighted by the prefix frequency d_ρ. A teacher that strongly favors $a_\star$ at states it visits frequently contributes at least as much gradient signal as the online source, regardless of how small the importance ratio is. The $\varepsilon(1-\varepsilon)$ factor is the standard derivative of the logistic function; the calculation here applies it to characterize which data source dominates the mixed gradient, which is specific to our setting.

Local scope and prefix coverage. These are derivatives at one fixed prefix in one chosen-action direction. For an autoregressive policy, two trajectories cease to share exact prefixes after their first differing token. Positive population support does not ensure useful finite-sample overlap. At a teacher-only observed prefix there may be no online estimate to compare, and a shared neural update changes many prefixes together. The identities do not establish global gradient alignment or adequate coverage of student-visited states.

Sign reliability. The previous result concerns the expected gradient. In practice, we estimate the gradient from a finite batch. The next proposition shows that the *relative* noise in the teacher's gradient estimate—the ratio of standard deviation to mean—also does not depend on how small ε is.

Sign reliability of gradient estimates has been studied in the context of compressed optimization by Bernstein et al. (2018). The following proposition applies a similar analysis to the teacher's gradient in our mixed-data setting, showing that the key factorization—which makes the relative variance independent of ε—is specific to the teacher-sampling structure rather than a generic property of sign estimators.

Proposition D.7 (Sign reliability of the teacher gradient)**.** *Under the same setup, with finite S_μ and $\Delta_\mu \neq 0$ and teacher sampling, alternative proportions, and target distributions fixed as ε varies, let $M_\mu(a) = \mathbb{E}[\widehat{A}_\mu^2 \mid s, a]$ denote the conditional second moment of the teacher's advantage target. Define*

$$S_\mu = \frac{M_\mu(a_\star)}{\mu(a_\star \mid s)} + \sum_{b \neq a_\star} \frac{w_b^2\, M_\mu(b)}{\mu(b \mid s)}, \tag{16}$$

which captures the teacher's sampling variability along the chosen-action direction. For a single teacher draw, the scalar gradient estimate X_μ along this direction has mean and variance

$$\mathbb{E}[X_\mu] = \varepsilon(1{-}\varepsilon)\,\Delta_\mu, \qquad \mathrm{Var}(X_\mu) = \varepsilon^2(1{-}\varepsilon)^2\,(S_\mu - \Delta_\mu^2). \tag{17}$$

The factor $\varepsilon^2(1{-}\varepsilon)^2$ appears in both the squared mean and the variance, so the relative variance

$$\frac{\mathrm{Var}(X_\mu)}{(\mathbb{E}[X_\mu])^2} = \frac{S_\mu - \Delta_\mu^2}{\Delta_\mu^2} \tag{18}$$

does not depend on ε. For the average of n independent teacher draws, Chebyshev's inequality gives a sign-error bound:

$$\Pr(\overline{X}_{\mu,n}\,\Delta_\mu \leq 0) \leq \frac{S_\mu - \Delta_\mu^2}{n\,\Delta_\mu^2}. \tag{19}$$

Proof. Along (7), the log-probability derivative $\partial \log \pi_x(a \mid s)/\partial x$ evaluated at x_0 equals $1{-}\varepsilon$ for $a_\star$ and $-\varepsilon$ for each alternative b. The single-draw gradient estimate is therefore $X_\mu = \varepsilon(1{-}\varepsilon)\,Y_\mu$, where

$$Y_\mu = \begin{cases} \widehat{A}_\mu \,/\, \mu(a_\star \mid s) & \text{if the draw is } a_\star, \\ -w_b\,\widehat{A}_\mu \,/\, \mu(b \mid s) & \text{if the draw is } b \neq a_\star. \end{cases}$$

Taking moments under teacher sampling: $\mathbb{E}[Y_\mu] = \Delta_\mu$ and $\mathbb{E}[Y_\mu^2] = S_\mu$, which gives (17). The common factor $\varepsilon^2(1{-}\varepsilon)^2$ cancels from the relative variance, proving (18). Applying Chebyshev's inequality to the average of n independent draws—noting that a wrong-sign estimate differs from the mean by at least the mean's magnitude—gives (19). □