SWPO: SCORE-WEIGHTED CRITIC FITTING FOR MULTI-TURN POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As language agents generate longer reasoning and tool-use trajectories, learning from sparse outcome rewards becomes increasingly difficult. Critic-free group methods can receive weak or noisy learning signals in this setting, motivating PPO's learned state-dependent baseline. Turn-level PPO aligns optimization with complete reasoning-and-action decisions and can produce substantially shorter responses, but aggregating token scores and likelihood ratios can also increase update variability. Optimal-baseline theory motivates score-energy-weighted critic fitting to address this variability, yet existing OPT logit proxies require vocabulary-wide computation. We introduce SWPO, which constructs a cheap scalar proxy from stored sampled-token probabilities and sums it within each turn to weight critic fitting. The method retains the turn-level PPO actor and admits an exact factor-of-two logit-space bound, with a conditional interpretation of population baseline error. Experiments on search and embodied tasks support summed scalar weighting among the tested formulations, preserve shorter responses in the search-training comparison, and achieve competitive performance with the summed OPT proxy. SWPO reduces measured actor-update time by about 40% relative to summed OPT, with little additional time over the PPO baselines.

1 INTRODUCTION

Language agents increasingly rely on extended reasoning and repeated tool use to solve tasks. As their generated trajectories grow longer, a sparse final reward must guide many intermediate decisions. Critic-free methods such as GRPO and GiGPO estimate advantages from groups of sampled outcomes (Shao et al., 2024; Feng et al., 2025); when sampled outcomes offer little contrast or useful state-level groups are sparse, their learning signals can be weak or noisy. These conditions can make training less stable and motivate PPO-style methods, whose learned critic provides a state-dependent baseline for the return (Schulman et al., 2017; Ouyang et al., 2022).

Within PPO, turn-level optimization follows the semantic structure of agent interaction: one action consists of a reasoning segment followed by a tool call or environment action. The critic predicts a baseline at the start of that decision. This granularity can encourage concise responses, as observed in Figure 1, without guaranteeing higher final reward. Our turn-level actor multiplies token likelihood ratios within a turn and applies the original PPO clipping surrogate to the complete ratio.

The same aggregation introduces an optimization challenge. The complete-turn policy score is a sum of token scores, so its squared norm depends on turn length, token probabilities, and alignment among score directions. Multiplying token ratios also makes clipping sensitive to accumulated changes across the turn. These effects can increase update variability and make some critic errors more consequential for the actor than others. Established optimal-baseline theory addresses this issue by weighting return-prediction error according to policy-score energy (Dayan, 1991; Greensmith et al., 2004; Weaver & Tao, 2013); this gives a principled way to allocate critic-fitting effort to the turns that matter most for the variance of an individual on-policy contribution.

Recent Optimal Token Baseline (OTB) work makes this principle more practical through a forward-only logit-score proxy, which we call the OPT proxy (Li et al., 2026). That proxy still requires a vocabulary-wide probability reduction. We introduce SWPO, which replaces it with a scalar proxy computed from stored sampled-token probabilities and summed within each turn. The resulting weight guides a learned turn-start critic, whose predictions supply the actor's advantages while

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.

retaining the turn-level PPO objective; proxy construction uses simple scalar operations without an additional backward pass or vocabulary-wide reduction.

We evaluate GRPO, GiGPO, Token PPO, Turn PPO, and SWPO on search and embodied tasks, and compare uniform, length-only, mean scalar, summed scalar, and summed OPT critic weights (Tables 1–2; full ablations in Appendix B.4). The ablations support summed scalar weighting, with competitive average performance and substantially lower actor-update cost than summed OPT. Our contributions are:

- A cheap sampled-token proxy for score-weighted turn-level critic fitting, built on established optimal-baseline theory.
- An exact factor-of-two logit-space bound and a conditional baseline-error bound, with the full derivation in Appendix A.
- A controlled comparison of proxy formulations and their performance–computation tradeoff under a common turn-level PPO actor.

2 Related Work and Motivation

2.1 Learning from Outcome Rewards: Critic-Free Methods and Turn-Level PPO

Multi-turn agent training must assign credit to intermediate decisions from task feedback. Process-supervision methods provide intermediate scores using annotations or external judges (Uesato et al., 2022; Lightman et al., 2023; Bai et al., 2022; Lee et al., 2023), adding supervision and scoring costs. Our setting uses final task outcomes. GRPO constructs group-relative advantages from sampled responses (Shao et al., 2024), while GiGPO adds finer credit assignment through groups of repeated states (Feng et al., 2025). These methods avoid a learned critic, but their signals depend on variation in sampled rewards and the availability of informative groups. When outcomes are identical, GRPO's outcome-based group signal vanishes; GiGPO's finer signal also depends on suitable state matches. Other approaches refine credit assignment through additional update or scoring procedures, including SeeUPO and IGPO (Hu et al., 2026; Wang et al., 2025).

PPO fits a critic to returns and uses its state-dependent predictions as a baseline (Schulman et al., 2017; Ouyang et al., 2022). The update unit determines how this baseline is used. Token-level PPO operates on individual token decisions, whereas turn-level PPO treats a complete reasoning segment followed by one action as the environment-changing decision and queries the critic at its start. This aligns optimization with the agent's interaction structure and produces shorter responses in our search comparison. It also changes the dispersion of scores and importance ratios, motivating care in critic fitting. Section 3 defines the product-ratio actor and turn-start critic used throughout our comparison.

2.2 Optimal Baselines and Efficient Gradient Proxies

Score-energy-weighted baselines are established variance-reduction methods (Dayan, 1991; Greensmith et al., 2004; Weaver & Tao, 2013). OTB uses a forward-only vocabulary-wide logit proxy in a causal token baseline constructed from group statistics (Li et al., 2026). SWPO adapts the weighting principle to a learned state-only turn critic and simplifies the proxy to sampled-token probabilities. We use *OPT proxy* as shorthand for OTB's squared logit-score expression. Our *Turn PPO + Summed OPT* baseline places that expression in the same turn-level actor and critic as SWPO, allowing a controlled comparison of proxy cost and effectiveness. The full OTB algorithm uses a different baseline construction.

Length weighting provides a related alternative: OPO approximates gradient-energy weighting using response length (Hao et al., 2025). Our length-only control applies this idea to individual turns. Comparing it with the summed scalar proxy tests the contribution of token probabilities, while comparison with the mean scalar proxy tests the effect of retaining turn length in the weight. These ablations complement the theoretical proxy bound and assess which formulation works best for the evaluated turn-level PPO setting.

3 PRELIMINARIES

3.1 TURN-LEVEL VARIABLES

We first define the token-level objects and then group them into turn-level variables. We model a language-agent episode as a finite-horizon Markov decision process over turns. The index n denotes the turn number. At turn n, the state S_n is the full interaction history before generation. It includes the original prompt, previous model messages, tool calls or environment actions, and observations returned by the environment.

The set $\mathcal{T}_n$ contains token positions inside turn n. If the turn length is L_n, then $\mathcal{T}_n = \{1, \dots, L_n\}$. The token prefix before token t is $Z_{n,t}$, and the sampled token is $a_{n,t}$. The full turn action is

$$A_n = (a_{n,1}, \dots, a_{n,L_n}). \tag{1}$$

The token policy probability is $\pi_\theta(a_{n,t} \mid Z_{n,t})$, where θ denotes policy parameters. The probability of the whole generated turn is

$$\Pi_\theta(A_n \mid S_n) = \prod_{t \in \mathcal{T}_n} \pi_\theta(a_{n,t} \mid Z_{n,t}). \tag{2}$$

The return G_n is the Monte Carlo return assigned to the whole turn. In our sparse-reward setting, G_n is derived from the final episode reward.

The token-level policy score is

$$g_{n,t} = \nabla_\theta \log \pi_\theta(a_{n,t} \mid Z_{n,t}), \tag{3}$$

and the turn-level policy score is

$$\psi_n = \sum_{t \in \mathcal{T}_n} g_{n,t} = \nabla_\theta \log \Pi_\theta(A_n \mid S_n). \tag{4}$$

A turn-level scalar baseline is written $B(S_n)$ or $V_\phi(S_n)$, where ϕ denotes critic parameters.

3.2 TURN-LEVEL PPO BACKBONE

Given rollout data from π_{old}, where π_{old} is the fixed behavior policy used to generate the sampled trajectories, define the token ratio

$$r_{n,t}(\theta) = \frac{\pi_\theta(a_{n,t} \mid Z_{n,t})}{\pi_{\text{old}}(a_{n,t} \mid Z_{n,t})}. \tag{5}$$

Let $r_n^{\text{turn}}(\theta)$ denote the likelihood ratio of the complete sampled turn. Using the autoregressive factorization in Eq. 2,

$$\begin{aligned} r_n^{\text{turn}}(\theta) = \frac{\Pi_\theta(A_n \mid S_n)}{\Pi_{\text{old}}(A_n \mid S_n)} &= \prod_{t=1}^{L_n} r_{n,t}(\theta) \\ &= \exp\left(\sum_{t=1}^{L_n} \log r_{n,t}(\theta) \right). \end{aligned} \tag{6}$$

We compute the sum of log-ratios and exponentiate with numerically safe handling. There is no division by L_n and no tokenwise clipping before aggregation. Define the symmetric PPO clipping offset $\epsilon \in (0, 1)$ and the clipped turn ratio

$$r_n^c(\theta) = \text{clip}\big(r_n^{\text{turn}}(\theta), 1 - \epsilon, 1 + \epsilon\big). \tag{7}$$

Let A_n^{adv} be the detached turn advantage and $\mathbb{E}_n$ an empirical average over sampled turns. The actor loss retains the original PPO minimum between unclipped and clipped terms (Schulman et al., 2017):

$$\mathcal{L}_\pi(\theta) = -\mathbb{E}_n\big[\min\big(r_n^{\text{turn}}(\theta) A_n^{\text{adv}}, r_n^c(\theta) A_n^{\text{adv}}\big)\big]. \tag{8}$$

Clipping is applied once to the turn ratio; it is not a hard constraint on every token or a replacement of the objective by the clipped term alone. At the rollout policy all ratios equal one, so the local actor derivative uses the complete-turn score. Appendix A.7 derives this connection and the length sensitivity of the clipping boundaries.

3.3 TURN-LEVEL VALUE MODEL

The critic is also turn-level. Let $V_\phi : \mathcal{S} \to \mathbb{R}$ be a value model with parameters ϕ that maps a turn-start state to a scalar baseline. It is queried at the boundary before turn n:

$$V_n = V_\phi(S_n), \tag{9}$$

where S_n is the completed interaction history before the next turn is generated. Unlike a token-prefix critic, it does not predict $V_\phi(Z_{n,t})$ for each partial prefix inside the sampled turn.

If an episode ends at turn N with final outcome reward R_N, then in the sparse-reward case the Monte Carlo return assigned to turn n can be written as

$$G_n = \gamma^{N-n} R_N, \tag{10}$$

where γ is the discount factor. More generally, G_n denotes the return assigned to the whole turn. The turn advantage used in Eq. 8 is

$$A_n^{\mathrm{adv}} = G_n - V_\phi(S_n). \tag{11}$$

Pure turn-level PPO fits the critic with the unweighted turn-boundary value loss

$$\mathcal{L}_V^{\mathrm{PPO}}(\phi) = \mathbb{E}_n \left[(G_n - V_\phi(S_n))^2 \right]. \tag{12}$$

This design matches the policy-update unit: the actor treats A_n as the action taken from S_n, and the critic estimates the baseline for that same turn boundary.

4 METHOD: SWPO

4.1 TURN SCORES AND THE ORACLE CRITIC OBJECTIVE

We analyze one turn under a fixed rollout policy $\pi_{\mathrm{old}} = \pi_{\theta_{\mathrm{old}}}$. Here n indexes turns, S_n is the history before turn n, $A_n = (a_{n,1}, \ldots, a_{n,L_n})$ is its sampled action of L_n tokens, $\mathcal{T}_n = \{1, \ldots, L_n\}$, and $Z_{n,t}$ is the prefix before token t. With actor parameters θ, define the token score $g_{n,t}$ and the complete-turn score ψ_n explicitly as

$$g_{n,t} = \nabla_\theta \log \pi_\theta(a_{n,t} \mid Z_{n,t})|_{\theta_{\mathrm{old}}}, \qquad \psi_n = \nabla_\theta \log \Pi_\theta(A_n \mid S_n)|_{\theta_{\mathrm{old}}} = \sum_{t \in \mathcal{T}_n} g_{n,t}. \tag{13}$$

The complete-turn probability Π_θ is the product in Eq. 2. Expectations in this analysis use the fixed rollout policy and environment randomness. Let G_n be the turn return, $B(S_n)$ a fixed state-only baseline, and $\rho_n = \|\psi_n\|_2^2$ the parameter-space turn-score energy. The on-policy turn-gradient contribution is

$$\hat{h}_n = (G_n - B(S_n))\psi_n, \tag{14}$$

where the baseline does not depend on the sampled current turn. Since $\mathbb{E}[B(S_n)\psi_n \mid S_n] = 0$, its subtraction preserves the mean of this estimator. For a fixed state s and candidate baseline value $b \in \mathbb{R}$, minimizing trace variance is equivalent to minimizing the second moment,

$$\mathcal{J}_s(b) = \mathbb{E}[\rho_n (G_n - b)^2 \mid S_n = s], \qquad B_\rho^*(s) = \frac{\mathbb{E}[\rho_n G_n \mid S_n = s]}{\mathbb{E}[\rho_n \mid S_n = s]}. \tag{15}$$

This classical identity motivates weighting squared return-prediction error by ρ_n (Greensmith et al., 2004). Its specialization here concerns an individual on-policy turn contribution and is not a claim of a new optimal-baseline principle. Appendix A.2 gives the proof, and Appendix A.7 relates this population analysis to the implemented PPO update.

4.2 FORWARD-ONLY PROXY AND THE LOGIT-SPACE BOUND

Exact parameter-score energies require per-turn gradient instrumentation. Instead, let $z_{n,t} \in \mathbb{R}^{|\mathcal{V}|}$ be the rollout-policy logits over vocabulary $\mathcal{V}$, $p_{n,t}^{\mathrm{vocab}} = \mathrm{softmax}(z_{n,t})$ the probability vector, $p_{n,t}$ its sampled-token entry, and $e_{a_{n,t}}$ the corresponding one-hot vector. The logit score and scalar proxy are

$$s_{n,t} = \nabla_{z_{n,t}} \log p_{n,t} = e_{a_{n,t}} - p_{n,t}^{\mathrm{vocab}}, \qquad q_{n,t} = (1 - p_{n,t})^2. \tag{16}$$

For every token, the following bound holds exactly:

$$q_{n,t} \le \|s_{n,t}\|_2^2 \le 2q_{n,t}. \tag{17}$$

SWPO uses only the stored sampled-token probabilities to form

$$W_n = \sum_{t \in \mathcal{T}_n} q_{n,t}. \tag{18}$$

The state-only critic V_ϕ is fitted with the detached turn weights:

$$\mathcal{L}_V^{\mathrm{SWPO}}(\phi) = \mathbb{E}_n\left[W_n (G_n - V_\phi(S_n))^2\right], \tag{19}$$

where $\mathbb{E}_n$ averages over sampled turns and W_n is held fixed during critic optimization. We apply the raw weights without additional batch normalization or critic-weight clipping. The actor retains Eq. 8, using the critic's predictions to form its advantages. Section 4.3 gives the baseline-error bound, the structural rationale for summed weighting, and the practical validation strategy.

4.3 Baseline-Error Bound under Normalized-Energy Agreement

The relevant error is the difference between the proxy-weighted population target $B_W^*(s) = \mathbb{E}[W_n G_n \mid S_n = s]/\mathbb{E}[W_n \mid S_n = s]$ and the oracle $B_\rho^*(s)$. For states with positive mean weights, define the conditionally normalized energies and their discrepancy:

$$\bar{\rho}_n = \frac{\rho_n}{\mathbb{E}[\rho_n \mid S_n]}, \qquad \bar{W}_n = \frac{W_n}{\mathbb{E}[W_n \mid S_n]}, \qquad \eta(s) = \mathbb{E}[|\bar{\rho}_n - \bar{W}_n| \mid S_n = s]. \tag{20}$$

Assume bounded returns $G_n \in [G_{\min}, G_{\max}]$ and write $\Delta_G = G_{\max} - G_{\min} > 0$. Appendix A.6 establishes the final error bound

$$\boxed{\left|B_W^*(s) - B_\rho^*(s)\right| \le \frac{\Delta_G}{2}\eta(s).} \tag{21}$$

Assumption 1 (normalized-energy agreement). For a desired baseline-error tolerance $\varepsilon_{\mathrm{b}} > 0$, assume $\eta(s) \le 2\varepsilon_{\mathrm{b}}/\Delta_G$ on the states where the guarantee is invoked. Equation 21 then yields error at most ε_{b}. This assumption quantifies the joint effect of logit-to-parameter distortion and within-turn cross terms on the relative weights used for critic fitting.

Structural rationale. Summing token proxies follows the additive structure of the complete-turn score. Under the matched-policy and regularity conditions in Appendix A.4, token-score cross terms have zero conditional expectation, so expected turn-score energy decomposes into a sum of expected token-score energies. Together with the exact token-level logit bound, this motivates accumulating $q_{n,t}$ across a turn. The decomposition concerns expected score energy; realized cross terms and their dependence on returns are accounted for by the discrepancy $\eta(s)$ in Assumption 1.

Normalized-energy agreement is plausible when gradient propagation approximately preserves relative score energies across turns. The exact output-head identity connects logit-score energy to parameter-gradient energy through the hidden-state norm; normalization layers motivate approximate stability of this scale, following Li et al. (2026). Common multiplicative scales cancel in the conditionally normalized energies used in the analysis, and small policy changes motivate approximate consistency between rollout-derived weights and subsequent updates. Appendix A.5 formalizes this rationale through a sufficient residual condition accounting for backbone Jacobian distortion and within-turn cross terms.

Practical validation. Under the normalized-energy agreement assumption, the preceding analysis bounds the discrepancy between the proxy-weighted population baseline and the oracle baseline by the desired tolerance. To assess whether this theoretically motivated approximation is effective in actual training, we conduct extensive downstream ablations in Section 5.3, comparing uniform, length-only, mean scalar, summed scalar, and summed OPT weights (Appendix B.4). These experiments examine whether token-probability information and summed aggregation yield the expected practical benefits. Section 5.4 further evaluates the performance–computation tradeoff against the fuller OPT proxy. Together, the results provide empirical support for the proposed approximation and its use in turn-level PPO.

5 EXPERIMENTS

The main comparison retains GRPO, GiGPO, Token PPO, Turn PPO, and SWPO. A separate controlled study fixes the product-ratio actor and varies only the critic weight. The primary proxy-replacement comparison is summed scalar SWPO versus Turn PPO + Summed OPT, reporting final performance and actor-update time together.

Experimental scope. The turn-level comparisons use the product-ratio actor, with the critic-weight variants defined in Section 5.3. Tables 1–2 give the main PPO-family comparison, while Appendix B.4 gives the complete critic-weight ablation. The runtime table reports the matched actor-update measurements used for the scalar-versus-OPT comparison.

5.1 SETUP

Benchmarks. The search agent is trained on 2WikiMultiHopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and HotpotQA (Yang et al., 2018) with the `wikipedia` tool (Goldsmith, 2014). We evaluate these tasks plus held-out Bamboogle for OOD transfer (Press et al., 2023); WebShop and ALFWorld test web and embodied interaction (Yao et al., 2022; Shridhar et al., 2021).

Methods and models. We compare token-level PPO, turn-level PPO, GRPO (Shao et al., 2024), GiGPO (Feng et al., 2025), and SWPO on Qwen3-4B/8B (Yang et al., 2025). Turn PPO shares SWPO's product-ratio actor and critic architecture but uses uniform critic weights; GiGPO is a critic-free turn-level baseline. The Summed OPT ablation replaces only the scalar token proxy with Eq. 23, keeping the actor, turn aggregation, and critic update otherwise matched.

Protocol and metrics. Tables 1 and 2 use 125 Bamboogle examples and 500 validation examples for every other benchmark, with four independent evaluation rollouts per example. For each example, we place the four outcomes in a fixed per-example permutation before computing metrics. For binary rewards, Pass@k is the fraction of examples solved by the first k elements of that permutation; Avg@k is successful rollouts among those elements divided by Nk, for dataset size N. Pass@1 equals Avg@1; Avg@k need not increase with k. Gain is SWPO minus the strongest baseline in percentage points. Pass@1 and Avg@4 are the primary endpoints, while the remaining metrics provide complementary evidence. Appendix B gives the implementation details.

Training procedure. PPO/SWPO use 512 training instances per batch with one rollout each; GRPO uses 64 instances with eight rollouts each. Search, WebShop, and ALFWorld train for 160, 150, and 40 steps, respectively, and all runs converge within these schedules. A turn is one thinking segment followed by one tool call (search) or one selected action (WebShop/ALFWorld); GiGPO uses the same definition as Turn PPO and SWPO. The proxy ablations use detached raw critic weights without additional batch normalization or critic-weight clipping. We use PyTorch and veRL/HybridFlow on NVIDIA H200 GPUs (Paszke et al., 2019; Sheng et al., 2025); Appendix B gives the data, hyperparameters, and infrastructure details.

5.2 TURN-LEVEL PPO: RESPONSE EFFICIENCY AND REWARD

Figure 1 starts between 14,564 and 15,438 tokens and ends at 14,570 tokens for Token PPO, 10,901 for Turn PPO, and 11,264 for SWPO. Relative to Token PPO's endpoint, Turn PPO and SWPO reduce response length by 25.2% and 22.7%, respectively. SWPO thus preserves much of the turn-level response-length advantage, although its responses are slightly longer than Turn PPO's. Shorter responses do not guarantee higher reward: Turn PPO exceeds Token PPO on the search-task results in Table 1, but has lower Pass@1 and Avg@4 on both ALFWorld model scales in Table 2.

Figure 2 tracks Pass@4 on the Qwen3-4B Bamboogle evaluation set of 125 examples. SWPO shows a steadier upward trend in the displayed trace and finishes at 41.6%, compared with 38.4% for Turn PPO and 36.8% for Token PPO: gains of 3.2 and 4.8 percentage points. These endpoints match Table 1. The advantage is specific to this metric and model: SWPO does not improve Bamboogle Pass@1 or Avg@4 over Turn PPO at either scale, and its 8B Pass@4 is 47.2% versus Turn PPO's 48.0%. The displayed curve supports a qualitative account of training behavior, rather than a measurement of gradient variance or across-seed stability.



Figure 1: Search-training response-length trace. Turn PPO and SWPO finish with 25.2% and 22.7% shorter responses than Token PPO, respectively.

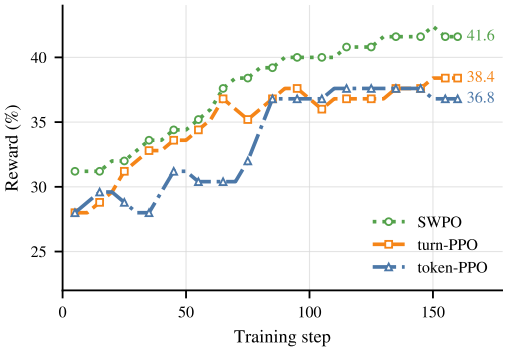

Figure 2: Measured Pass@4 reward curve on the Qwen3-4B Bamboogle evaluation set with 125 examples. The endpoints match Table 1.

Task	Model	Metric	Token PPO	Turn PPO	GRPO	GiGPO	SWPO	Gain
2Wiki	4B	Pass@1/2/4	24.0/31.0/38.0	25.0/32.0/39.0	22.0/29.0/36.0	21.6/27.6/37.8	**27.6/36.4**/37.6	+2.6/+4.4/-1.4
		Avg@1/2/4	24.0/20.8/19.5	25.0/21.3/20.0	22.0/18.6/18.8	21.6/18.4/19.7	**27.6/25.6**/19.0	+2.6/+4.3/-1.0
	8B	Pass@1/2/4	27.0/35.0/43.0	28.0/36.0/44.0	24.6/32.0/40.0	25.8/33.2/40.4	**30.6/39.0/48.2**	+2.6/+3.0/+4.2
		Avg@1/2/4	27.0/23.5/23.0	28.0/24.0/23.6	24.6/20.7/21.2	25.8/20.3/21.4	**30.6/26.7/26.8**	+2.6/+2.7/+3.2
MuSiQue	4B	Pass@1/2/4	12.0/18.0/27.0	12.8/18.8/27.8	10.6/16.0/24.6	12.2/16.0/24.4	**16.6**/17.6/**30.6**	+3.8/-1.2/+2.8
		Avg@1/2/4	12.0/9.2/11.0	12.8/9.8/11.5	10.6/8.1/9.7	12.2/8.5/10.2	**16.6/13.4/15.5**	+3.8/+3.6/+4.0
	8B	Pass@1/2/4	18.0/25.0/34.0	18.8/25.8/34.8	16.0/22.6/31.0	16.0/22.4/31.6	**22.2/29.8**/34.2	+3.4/+4.0/-0.6
		Avg@1/2/4	18.0/13.8/15.5	18.8/14.4/16.0	16.0/12.4/14.0	16.0/13.5/14.5	**22.2/17.1/19.4**	+3.4/+2.7/+3.4
HotpotQA	4B	Pass@1/2/4	19.0/25.0/32.0	19.8/25.8/32.8	17.0/23.0/30.0	15.6/21.6/30.8	**22.4/29.2**/31.6	+2.6/+3.4/-1.2
		Avg@1/2/4	19.0/17.8/18.3	19.8/18.2/18.8	17.0/15.6/16.8	15.6/16.1/18.6	**22.4**/17.9/**22.0**	+2.6/-0.3/+3.2
	8B	Pass@1/2/4	26.0/33.0/40.0	26.8/33.8/40.8	24.0/31.0/38.0	26.0/32.4/40.0	**30.4**/32.4/**44.2**	+3.6/-1.4/+3.4
		Avg@1/2/4	26.0/23.5/23.8	26.8/24.0/24.3	24.0/21.5/22.2	26.0/20.2/22.1	**30.4**/22.3/24.0	+3.6/-1.7/-0.3
Bamboogle (OOD)	4B	Pass@1/2/4	20.8/28.8/36.8	22.4/29.6/38.4	18.4/25.6/33.6	19.2/24.8/32.8	21.6/28.8/**41.6**	-0.8/-0.8/+3.2
		Avg@1/2/4	20.8/20.8/23.6	22.4/21.6/24.0	18.4/18.8/21.4	19.2/18.8/21.2	21.6/**24.4**/22.6	-0.8/+2.8/-1.4
	8B	Pass@1/2/4	30.4/37.6/47.2	31.2/39.2/48.0	27.2/35.2/44.0	28.0/35.2/45.6	30.4/37.6/47.2	-0.8/-1.6/-0.8
		Avg@1/2/4	30.4/28.4/31.0	31.2/29.2/31.8	27.2/24.8/28.6	28.0/24.4/29.8	30.4/**32.8**/30.8	-0.8/+3.6/-1.0

Table 1: Final measured search-task results: 125 Bamboogle examples and 500 examples for each other task. Entries are percentages for $k = 1/2/4$; Gain is SWPO minus the strongest baseline in points.

5.3 Main Comparison and Focused Ablations

Tables 1 and 2 compare PPO variants with the critic-free GRPO and GiGPO baselines under the same evaluation protocol. Token PPO and Turn PPO both exceed GRPO and GiGPO on Pass@1 and Avg@4 in all 12 task–model pairs. SWPO exceeds the strongest of these four baselines on Pass@1 in 10 of 12 pairs, with both Bamboogle scales as exceptions; on Avg@4, it does so in 6 of 12 pairs. The evidence therefore supports SWPO's practical effectiveness while showing that gains depend on the task and evaluation metric.

Proxy-weighting ablations. Appendix B.4 reports the complete five-way critic-weight ablation across both model scales. The compact summary below provides the practical validation described in Section 4.3; the Turn PPO and SWPO settings reproduce the corresponding entries in Tables 1 and 2. For a selected detached raw weight x_n, all variants use the same actor and critic architecture and fit the critic with

$$\mathcal{L}_V^x(\phi) = \mathbb{E}_n\left[x_n(G_n - V_\phi(S_n))^2\right], \tag{22}$$

without additional batch normalization or critic-weight clipping. SWPO is the choice $x_n = W_n$ in Eq. 19.

Task	Model	Metric	Token PPO	Turn PPO	GRPO	GiGPO	SWPO	Gain
WebShop	4B	Pass@1/2/4	43.6/52.2/58.6	44.2/52.8/59.2	37.8/48.8/55.0	39.2/47.4/55.4	**45.8**/**56.6**/58.4	+1.6/+3.8/-0.8
		Avg@1/2/4	43.6/42.8/42.8	44.2/43.5/43.8	37.8/36.8/37.2	39.2/39.4/38.1	**45.8**/42.9/**45.7**	+1.6/-0.6/+1.9
	8B	Pass@1/2/4	44.0/49.0/56.4	44.8/49.4/57.0	41.8/42.4/53.4	39.6/42.4/54.6	**47.6**/49.0/**59.6**	+2.8/-0.4/+2.6
		Avg@1/2/4	44.0/41.0/40.0	44.8/42.1/40.9	41.8/35.4/37.6	39.6/34.7/38.7	**47.6**/**44.7**/40.1	+2.8/+2.6/-0.8
ALFWorld	4B	Pass@1/2/4	74.4/80.6/84.4	73.4/79.2/83.6	69.4/74.4/78.2	69.8/77.0/79.6	**76.8**/80.2/**87.8**	+2.4/-0.4/+3.4
		Avg@1/2/4	74.4/76.4/77.0	73.4/75.5/75.6	69.4/65.2/67.7	69.8/63.4/69.5	**76.8**/**78.1**/76.5	+2.4/+1.7/-0.5
	8B	Pass@1/2/4	80.4/84.6/84.6	79.2/83.8/83.8	69.8/79.2/82.6	68.4/80.4/83.8	**83.0**/83.8/**86.8**	+2.6/-0.8/+2.2
		Avg@1/2/4	80.4/82.5/81.4	79.2/81.3/80.5	69.8/72.8/69.8	68.4/72.6/72.1	**83.0**/81.8/**84.3**	+2.6/-0.7/+2.9

Table 2: Final measured embodied-task results on 500 examples per task. Entries and gains use the same conventions as Table 1.

The fuller OPT proxy uses the squared logit-score norm from OTB (Li et al., 2026):

$$f_{n,t} = \|s_{n,t}\|_2^2 = 1 - 2p_{n,t} + \|p_{n,t}^{\text{vocab}}\|_2^2. \tag{23}$$

Let $F_n = \sum_{t=1}^{L_n} f_{n,t}$ be the summed OPT weight. The five choices are

$$x_n^{\text{uniform}} = 1, \qquad x_n^{\text{length}} = L_n, \quad x_n^{\text{mean}} = \frac{W_n}{L_n}, \\ x_n^{\text{sum}} = W_n, \quad x_n^{\text{OPT}} = F_n. \tag{24}$$

Table 3 summarizes the contrasts. The summed scalar and summed OPT variants differ only in their token proxy; the scalar proxy avoids the vocabulary-wide squared-probability reduction, and both avoid additional parameter-gradient backward passes. Appendix A.7 derives the weighted population targets and explains why averaging over a variable turn length can change the target.

Variant	Raw weight	Contrast isolated
Turn PPO	1	Unweighted baseline
Turn PPO + Length Only	L_n	Length dependence without token probabilities
Turn PPO + Mean Proxy	W_n/L_n	Removing the explicit length multiplier
SWPO (Summed Proxy)	W_n	Sampled-token probability signal plus length
Turn PPO + Summed OPT	F_n	Fuller logit proxy with identical turn aggregation

Table 3: Controlled product-ratio critic-weight ablation. Every row uses the same actor and critic design; the nonuniform variants differ only in their detached raw critic weights.

The sum-versus-mean contrast tests aggregation, the sum-versus-length contrast tests whether token probabilities add value beyond length, and the sum-versus-OPT contrast isolates the cheaper proxy. Averaging equally over the 12 task–model pairs in Appendix B.4, SWPO achieves 37.92% Pass@1 versus 36.37% for summed OPT, and 35.56% Avg@4 versus 35.05%. SWPO also has higher averages than the length-only and mean-proxy controls on these two metrics, supporting summed scalar weighting among the tested alternatives. Individual task–metric results favor different proxies, so this evidence supports the practical choice of SWPO without establishing universal superiority.

5.4 Scalar Proxy versus OPT: Performance and Training Time

The primary efficiency experiment compares SWPO with Turn PPO + Summed OPT, changing $q_{n,t}$ to $f_{n,t}$ while retaining the actor, critic, sum aggregation, precision, and update counts. The two proxies achieve competitive task performance, with modestly higher SWPO averages on Pass@1 and Avg@4 and individual results favoring each method. The central result is the actor-update cost: SWPO retains this level of performance while avoiding the vocabulary-wide reduction required by summed OPT. This comparison concerns proxy replacement within turn-level PPO; the complete OTB algorithm uses a different baseline construction (Li et al., 2026).

We report seconds per actor update because the fuller proxy's vocabulary-wide computation is incurred in this phase. This measurement characterizes update efficiency; total training time also depends on rollout generation, critic updates, and communication.

In Table 4, summed OPT takes approximately $1.66\times$ SWPO's actor-update time, corresponding to a 39.6% reduction for SWPO. SWPO takes 1.4% more time than Token PPO and 7.6% more than Turn PPO in the measured comparison. These results support a low-overhead scalar proxy that preserves the computational efficiency of the PPO backbone while offering competitive performance with the fuller proxy.

Method	Actor update time (s/update)	Relative cost	GPU computing utilization
SWPO (Scalar Proxy)	283.2345	1.00×	60.23%
Turn PPO + Summed OPT	469.2840	1.66×	42.12%
Turn PPO	263.2829	0.93×	65.00%
Token PPO	279.2034	0.99×	61.39%

Table 4: Measured actor-update time for the proxy comparison. Relative cost is each method's time divided by SWPO's time. SWPO and the PPO baselines have comparable cost, whereas summed OPT incurs the larger update cost shown by its relative-cost ratio.

6 Conclusion

SWPO adapts established score-weighted critic fitting to a learned turn-level critic whose action is a complete reasoning-and-action turn. The product-ratio actor applies the original PPO surrogate to the complete turn, while the scalar proxy uses stored sampled-token probabilities to weight critic fitting without the vocabulary-wide reduction of the fuller OPT proxy. The logit-space bound and conditional baseline-error analysis explain why this approximation can preserve useful score-energy information, while the ablations empirically support summed weighting among the tested scalar formulations. SWPO achieves competitive or better average performance than summed OPT, with individual results favoring each proxy, while requiring substantially less actor-update time and maintaining the shorter-response behavior observed in the search-training comparison.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073.

Peter Dayan. Reinforcement comparison. In *Connectionist Models*, pp. 45–51. Elsevier, 1991.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. *arXiv preprint arXiv:2505.10978*, 2025. URL https://arxiv.org/abs/2505.10978.

Jonathan Goldsmith. wikipedia. https://pypi.org/project/wikipedia/, 2014. Python package, version 1.4.0; install command: pip install wikipedia.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004. URL https://www.jmlr.org/papers/v5/greensmith04a.html.

Yaru Hao, Li Dong, Xun Wu, Shaohan Huang, Zewen Chi, and Furu Wei. On-policy RL with optimal reward baseline. *arXiv preprint arXiv:2505.23585*, 2025. URL https://arxiv.org/abs/2505.23585.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020. URL https://aclanthology.org/2020.coling-main.580/.

Tianyi Hu, Qingxu Fu, Yanxi Chen, Zhaoyang Liu, and Bolin Ding. SeeUPO: Sequence-level agentic-rl with convergence guarantees. *arXiv preprint arXiv:2602.06554*, 2026. URL https://arxiv.org/abs/2602.06554.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF: Scaling reinforcement learning from human feedback with AI feedback. *arXiv preprint arXiv:2309.00267*, 2023. URL https://arxiv.org/abs/2309.00267.

Yingru Li, Jiawei Xu, Ziniu Li, Jiacai Liu, Wei Liu, Yuxuan Tong, Longtao Zheng, Zhenghai Xue, Yaxiang Zhang, Tianle Cai, Ge Zhang, Qian Liu, and Baoxiang Wang. The optimal token baseline: Variance reduction for long-horizon LLM-RL. *arXiv preprint arXiv:2602.07078*, 2026. URL https://arxiv.org/abs/2602.07078.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2023. URL https://arxiv.org/abs/2210.03350.

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.

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. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, 2025. doi: 10.1145/3689031.3696075. URL https://dl.acm.org/doi/10.1145/3689031.3696075.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. URL https://aclanthology.org/2022.tacl-1.31/.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL https://arxiv.org/abs/2211.14275.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. *arXiv preprint arXiv:2510.14967*, 2025. URL https://arxiv.org/abs/2510.14967.

Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning. *arXiv preprint arXiv:1301.2315*, 2013. URL https://arxiv.org/abs/1301.2315.

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. URL https://arxiv.org/abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. URL https://aclanthology.org/D18-1259/.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

A Derivation and Scope of the SWPO Proxy Guarantee

This appendix develops a single proof chain: a legal turn baseline and its oracle variance objective, the exact logit-space proxy bound, the turn-score decomposition, and a conditional error guarantee for the proxy-weighted baseline. We follow the optimal-baseline and output-head reasoning of Li et al. (2026) and derive the baseline-error result for SWPO's state-conditioned turn critic. The score identities and algebraic bounds hold under the stated regularity conditions. Normalized-energy agreement supplies the additional quantitative condition linking the proxy to the oracle baseline, with structural motivation in Appendix A.5 and implementation scope in Appendix A.7.

A.1 Notation, Sampling, and Regularity Conditions

Fix a turn index n and actor parameters θ_{old}. The state S_n is the history before the turn, $A_n = (a_{n,1}, \ldots, a_{n,L_n})$ is the sampled turn, and $Z_{n,t}$ is the prefix before token t. The vocabulary $\mathcal{V}$ is finite. Expectations are over generation by $\pi_{\mathrm{old}} = \pi_{\theta_{\mathrm{old}}}$ and, for the return G_n, subsequent policy and environment randomness. All token probabilities and parameter scores refer to this same policy; the scores are evaluated at θ_{old}. The environment transition law has no direct dependence on actor parameters.

Assume differentiability with locally fixed support and interchangeability of differentiation and expectation. Generation ends at a stopping time $L_n \le L_{\max} < \infty$, including the sampled termination token when applicable, or at a fixed truncation limit. Let $\mathcal{F}_{n,t-1}$ denote the information available before token t, including S_n and whether the turn is still active. The indicator $I_{n,t} = \mathbf{1}\{t \le L_n\}$ is $\mathcal{F}_{n,t-1}$-measurable. Set $g_{n,t} = 0$ after termination; on active positions,

$$g_{n,t} = \nabla_\theta \log \pi_\theta(a_{n,t} \mid Z_{n,t})|_{\theta_{\mathrm{old}}}, \qquad \psi_n = \sum_{t=1}^{L_{\max}} g_{n,t}, \qquad \rho_n = \|\psi_n\|_2^2. \tag{A.1}$$

Assume finite conditional score second moments. These conditions justify all finite sums and conditional expectations below. If decoding changes the distribution through temperature or support truncation, the score identity uses the actual sampling distribution; mismatched probabilities require an appropriate correction.

Write $z_{n,t}$ for logits, $p_{n,t}^{\mathrm{vocab}}$ for their softmax vector, $p_{n,t}$ for its sampled-token entry, and $s_{n,t} = e_{a_{n,t}} - p_{n,t}^{\mathrm{vocab}}$ for the logit score, where $e_{a_{n,t}}$ is the sampled token's one-hot vector. Set $q_{n,t} = (1 - p_{n,t})^2$ on active positions and zero otherwise, so $W_n = \sum_t q_{n,t}$. Assume $G_{\min} \le G_n \le G_{\max}$ almost surely and define $\Delta_G = G_{\max} - G_{\min}$. For a state s, abbreviate

$$\mu_\rho(s) = \mathbb{E}[\rho_n \mid S_n = s], \qquad \mu_W(s) = \mathbb{E}[W_n \mid S_n = s]. \tag{A.2}$$

The conditionally normalized weights and their mean absolute discrepancy are

$$\bar{\rho}_n = \frac{\rho_n}{\mu_\rho(S_n)}, \qquad \bar{W}_n = \frac{W_n}{\mu_W(S_n)}, \qquad \eta(s) = \mathbb{E}[|\bar{\rho}_n - \bar{W}_n| \mid S_n = s]. \tag{A.3}$$

Normalized-weight statements concern states with $0 < \mu_\rho(s), \mu_W(s) < \infty$. If $\mu_\rho(s) = 0$, then $\psi_n = 0$ almost surely at that state and every baseline has zero turn-gradient variance. If $\mu_W(s) = 0$, then $W_n = 0$ almost surely at that state and the raw weighted objective does not select a unique baseline there; the normalized-target guarantee applies to states with positive mean weight. If $\Delta_G = 0$, both nondegenerate weighted targets equal the constant return.

A legal baseline $B(S_n)$ is a fixed measurable function of the turn-start state, independent of the sampled current action. For a learned critic, the population score identity treats its parameters as fixed; Appendix A.7 discusses finite-sample fitting and the implemented actor.

A.2 Legal Baselines and the Oracle Turn Objective

Proposition A.1 (baseline identity). For any legal baseline with the required integrability,

$$\mathbb{E}[B(S_n)\psi_n \mid S_n] = 0, \qquad \mathbb{E}[(G_n - B(S_n))\psi_n \mid S_n] = \mathbb{E}[G_n\psi_n \mid S_n]. \tag{A.4}$$

Proof. The stopped sequence probability $\Pi_\theta(A \mid s)$ is normalized over complete turns, so

$$\mathbb{E}[\psi_n \mid S_n = s] = \sum_A \Pi_{\text{old}}(A \mid s)\, \nabla_\theta \log \Pi_\theta(A \mid s)|_{\theta_{\text{old}}} = \nabla_\theta \sum_A \Pi_\theta(A \mid s)\Big|_{\theta_{\text{old}}} = 0. \quad \text{(A.5)}$$

Pulling $B(s)$ outside the conditional expectation proves the claim.

Proposition A.2 (oracle baseline and exact excess variance). For $\mathcal{J}_s(b) = \mathbb{E}[\rho_n(G_n - b)^2 \mid S_n = s]$, the unique minimizer when $\mu_\rho(s) > 0$ is

$$B^*_\rho(s) = \frac{\mathbb{E}[\rho_n G_n \mid S_n = s]}{\mu_\rho(s)}. \quad \text{(A.6)}$$

Moreover, for any scalar b,

$$\mathcal{J}_s(b) - \mathcal{J}_s(B^*_\rho(s)) = \mu_\rho(s)(b - B^*_\rho(s))^2. \quad \text{(A.7)}$$

Proof. By Proposition A.1 the conditional mean of $(G_n - b)\psi_n$ is independent of b. Its trace variance is $\mathcal{J}_s(b) - \|\mathbb{E}[G_n\psi_n \mid S_n = s]\|_2^2$, so it suffices to minimize $\mathcal{J}_s$. Differentiation gives

$$\mathcal{J}'_s(b) = -2\mathbb{E}[\rho_n(G_n - b) \mid S_n = s], \qquad \mathcal{J}''_s(b) = 2\mu_\rho(s) > 0.$$

Solving $\mathcal{J}'_s(b) = 0$ proves Eq. A.6. Expanding around $B^*_\rho(s)$ proves Eq. A.7, because $\mathbb{E}[\rho_n(G_n - B^*_\rho(s)) \mid S_n = s] = 0$.

For comparison, unweighted population MSE yields $B^*_0(s) = \mathbb{E}[G_n \mid S_n = s]$. Hence

$$B^*_\rho(s) - B^*_0(s) = \frac{\text{Cov}(\rho_n, G_n \mid S_n = s)}{\mu_\rho(s)}, \quad \mathcal{J}_s(B^*_0) - \mathcal{J}_s(B^*_\rho) = \frac{\text{Cov}(\rho_n, G_n \mid S_n = s)^2}{\mu_\rho(s)}. \quad \text{(A.8)}$$

This establishes the oracle's advantage for the individual-turn objective. The subsequent proxy-error and learned-critic decompositions connect this population target to practical fitting; cross-turn covariance belongs to a separate whole-trajectory objective.

A.3 Exact Logit-Score Expansion and Proxy Bound

Proposition A.3 (factor-of-two bound). For every active token, $q_{n,t} \le \|s_{n,t}\|_2^2 \le 2q_{n,t}$. *Proof.* Write $p_{n,t,v}$ for vocabulary entry v. The softmax score identity gives

$$\begin{aligned} \|s_{n,t}\|_2^2 &= (1 - p_{n,t})^2 + \sum_{v \neq a_{n,t}} p_{n,t,v}^2 \\ &= 1 - 2p_{n,t} + \|p^{\text{vocab}}_{n,t}\|_2^2. \end{aligned} \quad \text{(A.9)}$$

The omitted sum is nonnegative and at most $\big(\sum_{v \neq a_{n,t}} p_{n,t,v}\big)^2 = (1 - p_{n,t})^2$, proving both inequalities. With $F_n = \sum_{t \in \mathcal{T}_n} \|s_{n,t}\|_2^2$ the accumulated logit-score energy,

$$W_n \le F_n \le 2W_n \quad \text{(A.10)}$$

holds for each realized turn. The exact token expression in Eq. A.9 is the OPT logit proxy (Li et al., 2026). This factor-of-two interval gives a deterministic scale comparison; Assumption 1 characterizes the relative agreement relevant to the weighted baseline.

A.4 Turn-Score Decomposition and Return-Weighted Terms

Define the diagonal parameter-score energy D_n and the signed cross-term energy C_n:

$$\rho_n = D_n + C_n, \qquad D_n = \sum_{t=1}^{L_{\max}} \|g_{n,t}\|_2^2, \qquad C_n = 2\sum_{j<k} g_{n,j}^\top g_{n,k}. \quad \text{(A.11)}$$

Proposition A.4 (expected, not pathwise, decomposition). Under Appendix A.1's matched-policy conditions,

$$\mathbb{E}[C_n \mid S_n] = 0, \qquad \mathbb{E}[\rho_n \mid S_n] = \sum_{t=1}^{L_{\max}} \mathbb{E}[\|g_{n,t}\|_2^2 \mid S_n]. \tag{A.12}$$

Proof. Conditional on $\mathcal{F}_{n,k-1}$, $I_{n,k}$ is known and the active score has zero mean:

$$\mathbb{E}[g_{n,k} \mid \mathcal{F}_{n,k-1}] = I_{n,k} \sum_a \pi_{\text{old}}(a \mid Z_{n,k})\, \nabla_\theta \log \pi_\theta(a \mid Z_{n,k})|_{\theta_{\text{old}}} = 0.$$

For $j < k$, $g_{n,j}$ is $\mathcal{F}_{n,k-1}$-measurable. The tower property gives $\mathbb{E}[g_{n,j}^\top g_{n,k} \mid S_n] = \mathbb{E}[g_{n,j}^\top \mathbb{E}[g_{n,k} \mid \mathcal{F}_{n,k-1}] \mid S_n] = 0$. Summing the finite collection of terms proves the result; token-score independence is not required.

Equation A.12 supplies an exact additive interpretation of unweighted expected turn energy. For baseline fitting, G_n depends on later actions, so the return-weighted terms $\mathbb{E}[G_n C_n \mid S_n]$ and $\mathbb{E}[(G_n - b)^2 C_n \mid S_n]$ can remain nonzero, and individual turns can have appreciable C_n. Assumption 1 incorporates these contributions into the normalized-weight discrepancy. This separates the exact score identity from OPT's distinct approximation of advantage-weighted cross-temporal terms and motivates the residual decomposition below.

A.5 Parameter-Space Interpretation and a Sufficient Residual Condition

Exact chain-rule and output-head identities. Let P be the number of actor parameters and $J_{n,t} = \partial z_{n,t}/\partial\theta \in \mathbb{R}^{|\mathcal{V}|\times P}$ the logit Jacobian at θ_{old}. Then

$$g_{n,t} = J_{n,t}^\top s_{n,t}, \qquad \|g_{n,t}\|_2^2 = s_{n,t}^\top J_{n,t} J_{n,t}^\top s_{n,t}. \tag{A.13}$$

For a linear output head $z_{n,t} = U h_{n,t}$ with $U \in \mathbb{R}^{|\mathcal{V}|\times d}$ and hidden vector $h_{n,t} \in \mathbb{R}^d$, the direct head gradient satisfies

$$\nabla_U^{\text{head}} \log p_{n,t} = s_{n,t} h_{n,t}^\top, \qquad \|\nabla_U^{\text{head}} \log p_{n,t}\|_F^2 = \|s_{n,t}\|_2^2 \|h_{n,t}\|_2^2. \tag{A.14}$$

This is the full U-gradient when U is untied and $h_{n,t}$ does not depend on it. With tied embeddings it is only the direct output-head contribution; additional paths must be included in $J_{n,t}$.

Section 5.3 of Li et al. (2026) motivates approximate head-scale constancy using RMS normalization and assumes approximately isotropic backbone propagation to extend the proxy to the full model. This provides a structural rationale for SWPO: when hidden-state scale and directional gradient amplification vary moderately, logit-score energy can retain useful relative information about parameter-score energy. Normalization removes common scale, and local policy updates motivate approximate consistency during optimization. The residual decomposition below explicitly accounts for variation from normalization gains, inputs, the backbone Jacobian, and C_n.

To quantify this structural rationale, choose state-only positive scales $c_J(s)$ and $c_q(s)$, intended to describe the Jacobian scale and the logit-to-scalar scale, and set $c(s) = c_J(s)c_q(s)$. For example, if along each realized score direction

$$\left|s_{n,t}^\top J_{n,t} J_{n,t}^\top s_{n,t} - c_J(s)\|s_{n,t}\|_2^2\right| \le \alpha(s) c_J(s) \|s_{n,t}\|_2^2 \tag{A.15}$$

for a nonnegative distortion $\alpha(s)$, then

$$\begin{aligned}\mathbb{E}[|\rho_n - c(s)W_n| \mid S_n = s] \le{}& \alpha(s)c_J(s)\mathbb{E}[F_n \mid S_n = s] \\ &+ c_J(s)\mathbb{E}[|F_n - c_q(s)W_n| \mid S_n = s] + \mathbb{E}[|C_n| \mid S_n = s].\end{aligned} \tag{A.16}$$

This follows by substituting $\rho_n = D_n + C_n$ and applying the triangle inequality to $(D_n - c_J F_n) + c_J(F_n - c_q W_n) + C_n$. It identifies three contributions to normalized-energy agreement: directional Jacobian distortion, scalar-proxy distortion, and realized score cross terms. The logit bound supplies deterministic control of the proxy scale, while the residual condition below quantifies their combined effect on baseline accuracy.

Lemma A.5 (residual condition implies normalized agreement). For any state-only $c(s) > 0$, define the dimensionless residual

$$r(s) = \frac{\mathbb{E}[|\rho_n - c(s)W_n| \mid S_n = s]}{c(s)\mu_W(s)}. \tag{A.17}$$

If $r(s) < 1$, then

$$\eta(s) \le \frac{2r(s)}{1 - r(s)}. \tag{A.18}$$

Proof. Fix s and abbreviate $a = \mu_\rho(s)$ and $b = c(s)\mu_W(s)$. The definition implies $|a - b| \le r(s)b$, hence $a \ge (1 - r(s))b > 0$. Since $\bar{W}_n = c(s)W_n/b$,

$$\eta(s) \le \frac{\mathbb{E}[|\rho_n - c(s)W_n| \mid S_n = s]}{a} + \mathbb{E}[c(s)W_n \mid S_n = s]\left|\frac{1}{a} - \frac{1}{b}\right| \tag{A.19}$$

$$= \frac{\mathbb{E}[|\rho_n - c(s)W_n| \mid S_n = s] + |b - a|}{a} \le \frac{2r(s)}{1 - r(s)}. \tag{A.20}$$

This sufficient condition characterizes agreement of realized relative weights. Choosing $c(s) = \mu_\rho(s)/\mu_W(s)$ aligns the mean energies, and $r(s)$ then quantifies the remaining sample-dependent discrepancy.

A.6 Complete Baseline-Error Proof and Conditional Guarantee

For the unnormalized proxy-weighted population objective $\mathbb{E}[W_n(G_n - B(S_n))^2]$, the same quadratic argument as Proposition A.2 yields

$$B^*_W(s) = \frac{\mathbb{E}[W_n G_n \mid S_n = s]}{\mu_W(s)} = \mathbb{E}[\bar{W}_n G_n \mid S_n = s]. \tag{A.21}$$

Similarly $B^*_\rho(s) = \mathbb{E}[\bar{\rho}_n G_n \mid S_n = s]$. These are population minimizers over unrestricted state-only functions, not necessarily the outputs of a trained neural critic.

Theorem A.6 (baseline-error bound). Under the regularity and positivity conditions in Appendix A.1,

$$|B^*_W(s) - B^*_\rho(s)| \le \frac{\Delta_G}{2}\eta(s).$$

Proof. Both normalized weights have conditional mean one, so $\mathbb{E}[\bar{W}_n - \bar{\rho}_n \mid S_n = s] = 0$. Let $m_G = (G_{\min} + G_{\max})/2$. Subtracting this constant inside the expectation and using $|G_n - m_G| \le \Delta_G/2$ gives

$$\begin{aligned} |B^*_W(s) - B^*_\rho(s)| &= \left|\mathbb{E}[(\bar{W}_n - \bar{\rho}_n)(G_n - m_G) \mid S_n = s]\right| \\ &\le \mathbb{E}[|\bar{W}_n - \bar{\rho}_n|\,|G_n - m_G| \mid S_n = s] \\ &\le \frac{\Delta_G}{2}\mathbb{E}[|\bar{W}_n - \bar{\rho}_n| \mid S_n = s] = \frac{\Delta_G}{2}\eta(s). \end{aligned} \tag{A.22}$$

The proof accommodates dependence between returns and weights. Since both weights are nonnegative with mean one, $0 \le \eta(s) \le 2$; Assumption 1 sharpens the general return-range bound to the desired baseline-error tolerance.

Assumption-dependent corollary. For $\Delta_G > 0$, Assumption 1 is precisely the requirement $\eta(s) \le 2\varepsilon_{\mathrm{b}}/\Delta_G$ for a specified baseline-error tolerance $\varepsilon_{\mathrm{b}} > 0$. Under it,

$$|B^*_W(s) - B^*_\rho(s)| \le \varepsilon_{\mathrm{b}}, \qquad \mathcal{J}_s(B^*_W) - \mathcal{J}_s(B^*_\rho) \le \mu_\rho(s)\varepsilon_{\mathrm{b}}^2. \tag{A.23}$$

The second inequality follows from the exact excess-variance identity, Eq. A.7. Alternatively, Lemma A.5 supplies the sufficient condition and bound

$$|B^*_W(s) - B^*_\rho(s)| \le \frac{\Delta_G r(s)}{1 - r(s)}, \qquad r(s) \le \frac{\varepsilon_{\mathrm{b}}}{\Delta_G + \varepsilon_{\mathrm{b}}} \implies |B^*_W(s) - B^*_\rho(s)| \le \varepsilon_{\mathrm{b}}. \tag{A.24}$$

Thus normalized-energy agreement yields the desired baseline-error constraint and a bound on the individual-turn variance gap. A uniform guarantee requires the condition uniformly (or almost surely) over the states of interest; if $\mathbb{E}[\mu_\rho(S_n)] < \infty$, integrating Eq. A.23 also bounds the average gap. The theorem is conditional on score agreement and concerns population targets. Downstream improvements provide complementary evidence of practical effectiveness, rather than a numerical estimate of $\eta(s)$ or a proof of superiority over B^*_0.

A.7 Practical Weights, Learned Critics, and Megatron/PPO Scope

Weighted population targets and turn aggregation. For a nonnegative raw weight x_n from Eq. 24, assume $0 < \mathbb{E}[x_n \mid S_n = s] < \infty$ and bounded returns. Define the population objective for a candidate state-only baseline value b as

$$\mathcal{J}_s^x(b) = \mathbb{E}[x_n(G_n - b)^2 \mid S_n = s].$$

Expanding the quadratic gives a strictly positive coefficient $\mathbb{E}[x_n \mid S_n = s]$ on b^2. Its derivative is

$$\frac{\partial \mathcal{J}_s^x(b)}{\partial b} = 2b\,\mathbb{E}[x_n \mid S_n = s] - 2\mathbb{E}[x_n G_n \mid S_n = s].$$

Setting this derivative to zero yields the unique unrestricted population minimizer

$$B_x^*(s) = \frac{\mathbb{E}[x_n G_n \mid S_n = s]}{\mathbb{E}[x_n \mid S_n = s]}. \tag{A.25}$$

The uniform choice recovers the ordinary conditional return mean. Multiplying a weight by a positive state-only scale leaves this target unchanged because that scale cancels between numerator and denominator. By contrast, L_n depends on the sampled turn, so replacing W_n by W_n/L_n generally changes the target:

$$B_{W/L}^*(s) = \frac{\mathbb{E}[(W_n/L_n)G_n \mid S_n = s]}{\mathbb{E}[W_n/L_n \mid S_n = s]}.$$

This establishes why the mean and summed proxies are distinct ablations. The logit bound also implies $W_n \le F_n \le 2W_n$ for every realized turn, as proved in Eq. A.10. This scale comparison motivates the cheaper proxy, while the downstream ablations assess the relative effectiveness of the resulting critic targets.

Raw practical weights and learned critics. The implemented SWPO objective is Eq. 19; the general ablation objective is Eq. 22. Their raw weights are detached and held fixed during critic optimization. For SWPO, define the fitting error $e_\phi(s) = |V_\phi(s) - B_W^*(s)|$. The triangle inequality and Theorem A.6 give

$$|V_\phi(s) - B_\rho^*(s)| \le e_\phi(s) + \frac{\Delta_G}{2}\eta(s). \tag{A.26}$$

The fitting term includes finite-data, function-approximation, and optimization error. For another weight rule x_n, the same argument uses its own discrepancy $\eta_x(s) = \mathbb{E}[|x_n/\mathbb{E}[x_n \mid S_n] - \bar{\rho}_n| \mid S_n = s]$ and fitting error relative to $B_x^*(s)$. Thus the guarantee for the summed SWPO proxy does not automatically assign the same error bound to every ablation variant.

Connection to the product-ratio actor. Fix a sampled turn and its length during differentiation. Let r_n^{turn} be the exact ratio in Eq. 6, θ_{old} the rollout parameters, and $A_n^{\mathrm{adv}} = G_n - V_\phi(S_n)$ a detached advantage computed from a fixed critic. The denominator probabilities are constants, so

$$\begin{aligned}
\log r_n^{\mathrm{turn}}(\theta) &= \sum_{t=1}^{L_n} \left[\log \pi_\theta(a_{n,t} \mid Z_{n,t}) - \log \pi_{\mathrm{old}}(a_{n,t} \mid Z_{n,t})\right], \\
\nabla_\theta \log r_n^{\mathrm{turn}}(\theta) &= \sum_{t=1}^{L_n} \nabla_\theta \log \pi_\theta(a_{n,t} \mid Z_{n,t}), \\
\nabla_\theta r_n^{\mathrm{turn}}(\theta) &= r_n^{\mathrm{turn}}(\theta)\nabla_\theta \log r_n^{\mathrm{turn}}(\theta), \\
\nabla_\theta r_n^{\mathrm{turn}}(\theta)\big|_{\theta_{\mathrm{old}}} &= 1 \cdot \sum_{t=1}^{L_n} g_{n,t} = \psi_n.
\end{aligned} \tag{A.27}$$

For $\epsilon > 0$, all ratios equal one at the rollout policy and lie strictly inside the clipping interval. The unclipped and clipped branches coincide locally. Thus the negative actor-loss gradient contribution at that point is $(G_n - V_\phi(S_n))\psi_n$, exactly the contribution in Proposition A.2. There is no $1/L_n$ factor. The state-only baseline identity applies with the critic held fixed; weights used earlier to train that critic do not enter the actor loss directly.

This local identity does not make clipped, reused-data PPO globally unbiased or variance-optimal. Updated ratios, clipping, any advantage normalization, shared actor–critic parameters, same-batch critic fitting, and cross-turn covariance require separate analysis. The protocol must record these choices. In particular, fixing critic parameters for the population argument does not remove sample dependence if the critic is fitted to the same outcomes used to evaluate its gradients.

Length sensitivity of the clipping interval. Let d be a common token log-ratio for intuition, so $\log r_{n,t} = d$ within a turn. Then

$$r_n^{\text{turn}} = \prod_{t=1}^{L_n} e^d = e^{L_n d}.$$

For $0 < \epsilon < 1$, taking logarithms of $1 - \epsilon \le e^{L_n d} \le 1 + \epsilon$ and dividing by $L_n > 0$ yields

$$\frac{\log(1-\epsilon)}{L_n} \le d \le \frac{\log(1+\epsilon)}{L_n}. \tag{A.28}$$

Longer turns can therefore reach ratio clipping boundaries after smaller per-token log-ratio changes. This is an illustration of the objective, not a claim that all token ratios are equal or that PPO enforces a hard trust region. The reported experiments evaluate downstream outcomes and actor-update time; this algebraic observation is not a measured clipping-frequency result.

Proxy scope. The scalar/OPT comparison holds the critic and actor architectures and training protocol fixed while substituting the token proxy; their learned parameters evolve during training. Both proxies avoid additional gradient instrumentation. The current paper therefore interprets downstream performance and measured actor-update time as evidence for practical usefulness and efficiency, not as a direct estimate of full parameter-gradient variance.

B Experimental Setup and Implementation Details

B.1 Datasets and Training Schedules

Table B.1 lists the training data and schedules; search uses 15,000 training examples in total, and Bamboogle is held out for OOD evaluation. All runs fully converge within the stated final-checkpoint schedules.

Setting	Training data	Validation	Steps
Search	First 5,000 samples from each of HotpotQA, MuSiQue, and 2WikiMultiHopQA	500 per task	160
Bamboogle (OOD)	Held-out evaluation only	125	—
WebShop	5,000 items; 4,057 training examples	500	150
ALFWorld	2,500 training games	500	40

Table B.1: Training data and schedules. Validation sizes count examples or games, not rollouts; each receives four independent evaluation rollouts, ordered by a fixed per-example permutation for metric computation.

B.2 Optimization, Sampling, and Turn Definitions

The turn-level actor multiplies token ratios and clips the product once, retaining the minimum in Eq. 8; Token PPO retains its token-level surrogate. PPO uses symmetric clipping offsets $\epsilon_{\text{lower}} = \epsilon_{\text{upper}} = 0.2$, giving importance-ratio clipping bounds $[0.8, 1.2]$. Its KL penalty coefficient is 0.001 and discount factor is $\gamma = 1$. The proxy ablations use detached raw critic weights and no additional batch normalization or critic-weight clipping. Turn PPO + Summed OPT changes only the token proxy and retains summed aggregation. The mean-proxy and length-only controls follow Eq. 24. The KL coefficient is applied consistently across the PPO-family comparisons.

Setting	PPO/SWPO variants	GRPO
Training instances per batch	512	64
Rollouts per training instance	1	8
Trajectories per training batch	512	512
KL penalty coefficient	0.001	0.001
Search learning rate	1×10^{-6}	1×10^{-6}
PPO clipping offsets (lower/upper)	$0.2/0.2$	—
PPO discount factor γ	1	—
Critic-weight processing	Detached raw weights; no extra norm./clipping	—

Table B.2: Optimization and sampling for the PPO variants and GRPO. Nonuniform critic weights apply only to the weighted variants. Dashes denote parameters not assigned to GRPO here.

GRPO matches PPO's learning rate and KL coefficient. Both generate 512 trajectories per training batch, but differ in prompt coverage; equal trajectory counts do not imply equal token counts or cost. All methods use four rollouts per validation example, independently of training rollout multiplicity.

A turn consists of one thinking segment followed by one tool call (search) or selected action (WebShop/ALFWorld). GiGPO, Turn PPO, and SWPO share these boundaries; Token PPO retains token-level updates.

B.3 Evaluation Metrics and Aggregation

Evaluation-metric details are stated in the main experimental protocol above.

For validation example i, let $r_{i,j} \in \{0, 1\}$ denote the outcome at position j in its fixed per-example permutation of the four rollouts, with $i = 1, \ldots, N$ and $j = 1, \ldots, 4$. For $k \in \{1, 2, 4\}$, the reported percentage metrics are

$$\text{Pass@}k = \frac{100}{N} \sum_{i=1}^{N} \mathbf{1}\left\{ \sum_{j=1}^{k} r_{i,j} > 0 \right\}, \qquad \text{Avg@}k = \frac{100}{Nk} \sum_{i=1}^{N} \sum_{j=1}^{k} r_{i,j}. \tag{B.1}$$

Pass@1 and Avg@1 are identical. Because the prefixes are nested within the fixed permutation, Pass@k cannot decrease as k increases, whereas Avg@k can increase or decrease. Table B.3 makes the denominators explicit: 500 Bamboogle evaluation rollouts at $k = 4$ arise from 125 examples, not 500 distinct examples.

Validation set	N	Pass@k	Avg@1	Avg@2	Avg@4
Bamboogle	125	125	125	250	500
Each other benchmark	500	500	500	1,000	2,000

Table B.3: Example counts and metric denominators. Pass@k divides solved-example counts by N for every k; Avg@k divides successful-rollout counts by Nk.

Before rounding, Bamboogle Pass@k has 0.8-point increments and Avg@1/2/4 has $0.8/0.4/0.2$-point increments. For other benchmarks these are 0.2 and $0.2/0.1/0.05$ points, respectively. Tables display one decimal place. Figure 2's 4B Bamboogle Pass@4 endpoints match Table 1.

Gain is SWPO minus the strongest of Token PPO, Turn PPO, GRPO, and GiGPO. Boldface marks only SWPO cells that strictly exceed all comparators in that table. Section 5.3 averages each metric separately over the 12 task–model pairs with equal weight: it neither pools examples nor mixes metrics or double-counts Pass@1 and Avg@1. Dataset sizes determine the metric denominators, not an additional rescaling of the measured percentages.

B.4 Complete Proxy-Weighting Ablations

The following tables provide the full per-task and per-model results for the proxy-weighting ablation summarized in Section 5.3. They use the same fixed-permutation evaluation protocol and metric definitions as the main results.

Task	Model	Metric	Turn PPO	Turn PPO + Length Only	Turn PPO + Mean Proxy	Turn PPO + Summed OPT	SWPO
2Wiki	4B	Pass@1/2/4	25.0/32.0/39.0	23.0/31.8/35.6	25.8/32.0/38.6	25.6/32.6/39.8	**27.6**/**36.4**/37.6
		Avg@1/2/4	25.0/21.3/20.0	23.0/21.4/17.4	25.8/22.3/20.2	25.6/22.0/20.6	**27.6**/**25.6**/19.0
	8B	Pass@1/2/4	28.0/36.0/44.0	24.8/34.4/41.8	27.8/36.0/44.2	29.0/36.6/44.6	**30.6**/**39.0**/**48.2**
		Avg@1/2/4	28.0/24.0/23.6	24.8/23.1/22.1	27.8/24.5/24.0	29.0/24.6/24.4	**30.6**/**26.7**/**26.8**
MuSiQue	4B	Pass@1/2/4	12.8/18.8/27.8	12.6/17.0/24.8	13.8/18.6/28.2	13.6/19.4/28.8	**16.6**/17.6/**30.6**
		Avg@1/2/4	12.8/9.8/11.5	12.6/8.8/11.7	13.8/10.5/12.5	13.6/10.5/12.4	**16.6**/**13.4**/**15.5**
	8B	Pass@1/2/4	18.8/25.8/34.8	17.6/25.8/32.2	19.4/26.8/34.6	19.4/26.6/35.8	**22.2**/**29.8**/34.2
		Avg@1/2/4	18.8/14.4/16.0	17.6/14.3/15.7	19.4/15.0/16.9	19.4/15.0/16.8	**22.2**/**17.1**/**19.4**
HotpotQA	4B	Pass@1/2/4	19.8/25.8/32.8	19.4/24.0/33.0	20.8/25.4/32.8	20.6/26.4/33.6	**22.4**/**29.2**/31.6
		Avg@1/2/4	19.8/18.2/18.8	19.4/14.8/15.7	20.8/19.2/19.0	20.6/18.7/19.7	**22.4**/17.9/**22.0**
	8B	Pass@1/2/4	26.8/33.8/40.8	24.8/31.8/39.6	27.0/34.6/40.4	27.8/34.4/41.8	**30.4**/32.4/**44.2**
		Avg@1/2/4	26.8/24.0/24.3	24.8/23.6/23.1	27.0/24.2/25.3	27.8/24.7/24.9	**30.4**/22.3/24.0
Bamboogle (OOD)	4B	Pass@1/2/4	22.4/29.6/38.4	19.2/28.8/37.6	22.4/29.6/39.2	23.2/30.4/39.2	21.6/28.8/**41.6**
		Avg@1/2/4	22.4/21.6/24.0	19.2/21.2/22.6	22.4/21.2/23.6	23.2/22.4/25.0	21.6/**24.4**/22.6
	8B	Pass@1/2/4	31.2/39.2/48.0	29.6/39.2/48.0	31.2/39.2/48.0	32.0/40.0/48.8	30.4/37.6/47.2
		Avg@1/2/4	31.2/29.2/31.8	29.6/29.2/28.6	31.2/29.6/31.6	32.0/30.0/32.6	30.4/**32.8**/30.8

Table B.4: Final measured proxy-weighting ablations on search and OOD tasks for Qwen3-4B/8B. Entries are percentage triples for $k = 1/2/4$; Turn PPO and SWPO match Table 1. Boldface marks SWPO values that exceed all other settings in the corresponding cell; other cells may favor a different proxy.

Task	Model	Metric	Turn PPO	Turn PPO + Length Only	Turn PPO + Mean Proxy	Turn PPO + Summed OPT	SWPO
WebShop	4B	Pass@1/2/4	44.2/52.8/59.2	44.0/52.8/57.6	43.8/53.2/59.2	45.2/53.4/60.2	**45.8**/**56.6**/58.4
		Avg@1/2/4	44.2/43.5/43.8	44.0/40.3/41.7	43.8/43.2/44.7	45.2/44.2/44.7	**45.8**/42.9/**45.7**
	8B	Pass@1/2/4	44.8/49.4/57.0	44.6/48.6/56.6	45.2/50.4/57.2	45.6/50.0/57.8	**47.6**/49.0/**59.6**
		Avg@1/2/4	44.8/42.1/40.9	44.6/39.8/39.2	45.2/42.9/40.5	45.6/42.7/41.6	**47.6**/**44.7**/40.1
ALFWorld	4B	Pass@1/2/4	73.4/79.2/83.6	72.2/78.2/80.2	74.2/79.6/83.2	74.4/79.8/84.2	**76.8**/**80.2**/**87.8**
		Avg@1/2/4	73.4/75.5/75.6	72.2/72.6/75.7	74.2/75.4/76.5	74.4/76.3/76.6	**76.8**/**78.1**/76.5
	8B	Pass@1/2/4	79.2/83.8/83.8	77.0/83.4/83.6	79.2/83.4/83.8	80.0/84.4/84.6	**83.0**/83.8/**86.8**
		Avg@1/2/4	79.2/81.3/80.5	77.0/79.2/80.1	79.2/81.1/81.3	80.0/82.0/81.3	**83.0**/81.8/**84.3**

Table B.5: Final measured proxy-weighting ablations on WebShop and ALFWorld for Qwen3-4B/8B. Entries are percentage triples for $k = 1/2/4$; Turn PPO and SWPO match Table 2. Boldface marks SWPO values that exceed all other settings in the corresponding cell; other cells may favor a different proxy.

B.5 Infrastructure

Experiments use NVIDIA H200 GPUs with 144 GB of memory per GPU, PyTorch, and veRL/HybridFlow with the Megatron update backend (Paszke et al., 2019; Sheng et al., 2025). Table 4 reports seconds per actor update for Section 5.4.