# DUAL-AXIS POLICY OPTIMIZATION FOR LLM AGENTS: BAYESIAN FEEDBACK ATTRIBUTION AND TRAJECTORY MASS NORMALIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A reinforcement-learning objective for LLM agents makes two weighting choices that are usually left implicit: which decisions inside a trajectory receive credit, and how complete trajectories are weighted in the batch. We make both choices explicit. For the first, we propose Bayesian Feedback Attribution (BFA): we score the observed environment feedback under the executed action and under a counterfactual action sampled from the same policy, and use the resulting posterior lift to reallocate the host learner's advantages across decisions while conserving each trajectory's total coefficient mass. The policy model itself serves as the feedback model, so no learned dynamics model, return predictor, or auxiliary loss is required. For the second, we show that global token averaging weights trajectories by a size-biased transform of the episode distribution, and we derive the exact finite-batch discrepancy to uniform per-trajectory averaging as a length–gradient covariance term. Trajectory Mass Normalization (TMN) weights episodes uniformly, matching how benchmarks evaluate: one vote per episode. The resulting framework, BATON, introduces no new hyperparameters and no additional environment interaction, and drops into GRPO and GiGPO. On ALFWorld, WebShop, and search-augmented question answering across 1.5B–7B models, each component improves its host on its own, and the combination is best or tied for best in every comparison we ran, at 10–17% additional wall-clock cost measured on ALFWorld at the 1.5B scale.

## 1 INTRODUCTION

LLM agents solve tasks through repeated interaction with environments, including embodied instruction following, web navigation, and information retrieval (Yao et al., 2023; Shridhar et al., 2021; Yao et al., 2022; Jin et al., 2025). Reinforcement learning increasingly trains these interaction policies (Qi et al., 2024; Wang et al., 2025; Feng et al., 2025). Each rollout contains a sequence of decisions, observed responses, and a final task return. This structure raises two optimization questions: how should feedback allocate learning signals among decisions, and how should trajectories of different lengths enter the batch objective? We call these axes **Intra-Trajectory Feedback Attribution** and **Inter-Trajectory Objective Aggregation**.

The two questions become especially relevant when interactions vary widely in length. Figure 1 shows that failed trajectories are often longer than successful ones. Under global valid-token averaging, the longest 25% of trajectories receive 60.3% of the aggregation mass in ALFWorld and 42.2% in WebShop. Meanwhile, the environment responses along each trajectory provide local evidence for distinguishing actions taken under the same history. BATON uses this evidence within trajectories and controls length-dependent weighting across trajectories.

We introduce **BATON** (Bayesian Attribution and Trajectory Mass Normalization), which makes both weighting choices explicit. Bayesian Feedback Attribution (BFA) asks, for each decision, how much better the executed action explains the feedback that was actually observed than an alternative the same policy could have taken. This is a counterfactual question, and the policy itself answers it: scoring the observed feedback under both actions uses the language model's own next-token likelihoods, so no learned dynamics model, return predictor, or extra environment interaction is



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.

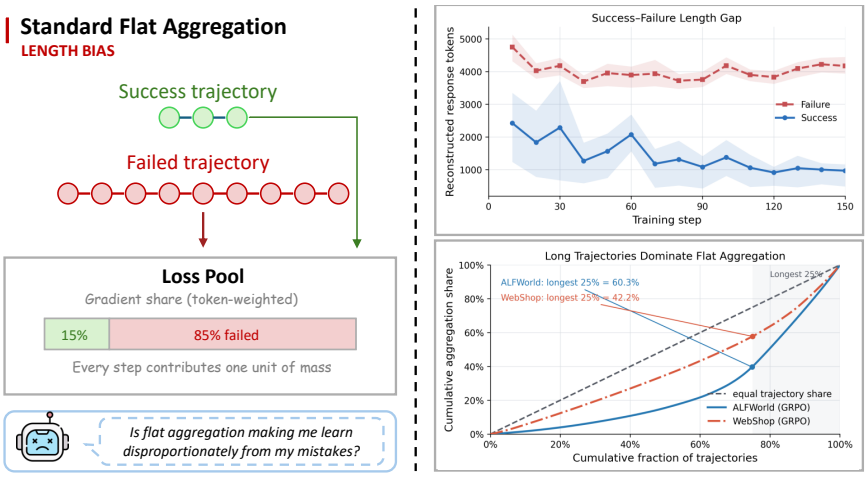


Figure 1: Trajectory-length imbalance in agent training. Global token averaging assigns each trajectory aggregation mass proportional to its valid policy-token count.

needed. The resulting posterior lift reallocates the host learner's advantage across decisions while conserving each trajectory's total coefficient mass. Trajectory Mass Normalization (TMN) then averages the complete-trajectory surrogates uniformly, so that each sampled task attempt contributes one share to the batch objective.

We analyze both choices. For BFA, the induced allocation is the unique maximizer of allocation-weighted log lift under a KL penalty against the original token allocation (Proposition 1), and it reduces exactly to the host allocation whenever the observed feedback is equally likely under the candidate actions. For TMN, token-proportional and uniform aggregation differ by exactly the empirical length–gradient covariance divided by mean trajectory length (Proposition 2); token averaging is the size-biased transform (Patil & Rao, 1978) of the episode distribution over surrogates. We evaluate BATON with GRPO and GiGPO on ALFWorld, WebShop, and search-augmented question answering across 1.5B–7B scales. Each component improves its host on its own; the combination is best or tied for best in all sixteen ablation cells (Section 4.4), at 10–17% additional wall clock on ALFWorld at the 1.5B scale, no extra environment interaction, and no new hyperparameters.

## 2 Preliminaries

**Problem setup.** A language-model agent interacts with an environment over multi-step tasks. At decision step $t$ it conditions on the history $h_t = (o_0, a_0, \ldots, a_{t-1}, o_t)$, generates $a_t \sim \pi_\theta(\cdot \mid h_t)$, and receives the next observation $o_{t+1}$. A complete trajectory $\tau_i$ has $T_i$ decisions, a task return $R_i$, $n_{i,t}$ valid policy tokens at step $t$, and $N_i = \sum_t n_{i,t}$ tokens in total. We optimize over batches of $B$ complete trajectories.

**Group-based policy optimization.** GRPO (Shao et al., 2024) and GiGPO (Feng et al., 2025) optimize groups of sampled trajectories without a value model; GRPO assigns each trajectory a group-relative advantage $\hat{A}_i$ shared by its tokens, while GiGPO refines the signal with turn-level advantages within the trajectory. With $r_{i,t,k}(\theta) = \pi_\theta(y_{i,t,k} \mid h_{i,t,k}) / \pi_{\theta_{\text{old}}}(y_{i,t,k} \mid h_{i,t,k})$ the likelihood ratio of the $k$-th valid policy token, the GRPO objective under global valid-token averaging is

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{S} \sum_{i=1}^{B} \sum_{t=1}^{T_i} \sum_{k=1}^{n_{i,t}} \min\Big[ r_{i,t,k}(\theta)\hat{A}_i,\ \text{clip}(r_{i,t,k}(\theta), 1-\epsilon, 1+\epsilon)\,\hat{A}_i \Big], \tag{1}$$

where $S = \sum_{i=1}^{B} N_i$. We omit separately reduced regularizers.

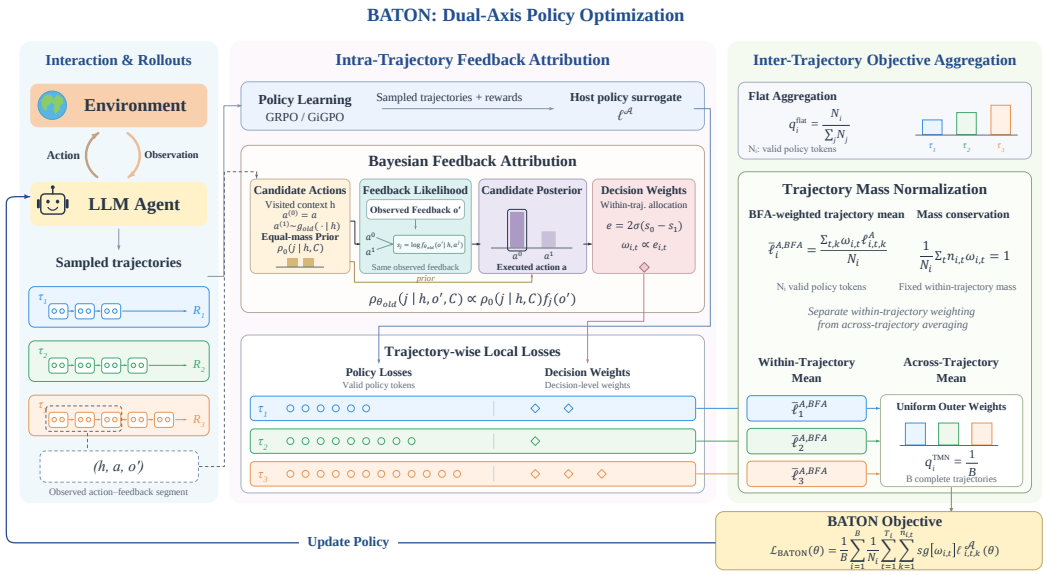


Figure 2: BATON. BFA scores the observed feedback under the executed action and a counterfactual alternative sampled from the policy itself, and converts the posterior lift into mass-conserving stop-gradient decision weights. TMN aggregates the weighted trajectory surrogates with uniform per-episode outer coefficients; token-proportional (flat) aggregation is its size-biased counterpart (Proposition 2).

# 3 METHOD

BATON separates **Intra-Trajectory Feedback Attribution** from **Inter-Trajectory Objective Aggregation**. On the first axis, Bayesian Feedback Attribution (BFA) reallocates coefficients among decisions by a counterfactual comparison scored by the policy itself: how much better does the executed action explain the observed feedback than an alternative the same policy could have taken? On the second axis, Trajectory Mass Normalization (TMN) replaces token-proportional aggregation with uniform per-trajectory averaging.

## 3.1 DUAL-AXIS POLICY OPTIMIZATION

For a batch of complete trajectories $\mathcal{B} = \{\tau_i\}_{i=1}^{B}$, let $n_{i,t}$ be the number of valid policy tokens at decision $t$ and $N_i = \sum_t n_{i,t} > 0$. Writing $\ell^{\mathcal{A}}_{i,t,k}$ for the host learner's local policy-surrogate loss, we define

$$\mathcal{L}(\theta) = \sum_{i=1}^{B} q_i \frac{1}{N_i} \sum_{t=1}^{T_i} \sum_{k=1}^{n_{i,t}} \mathrm{sg}[\omega_{i,t}] \ell^{\mathcal{A}}_{i,t,k}(\theta), \tag{2}$$

where $q_i \geq 0$, $\sum_i q_i = 1$, and sg denotes stop-gradient. The outer coefficient $q_i$ controls trajectory aggregation; the positive weight $\omega_{i,t}$ controls allocation among decisions. We impose

$$\frac{1}{N_i} \sum_t n_{i,t} \omega_{i,t} = 1, \tag{3}$$

so that the token coefficients within trajectory $i$ sum to $q_i$. For decisions with valid policy tokens, define $b_{i,t} = n_{i,t}/N_i$ and $\mu_{i,t} = b_{i,t}\omega_{i,t}$. Then $\sum_t \mu_{i,t} = 1$, and each local policy term has coefficient $q_i \mu_{i,t}/n_{i,t}$. The reference $b_i$ is the host learner's own token allocation rather than a free design choice: BFA is a penalized departure from it (Proposition 1), and the decision-uniform outer variant (turn-mean reduction) underperforms TMN in Appendix C. BFA determines the decision allocation $\mu_i$, while TMN selects the trajectory coefficients $q_i$. Both are held fixed when differentiating the policy surrogate.

### 3.2 Definition of Bayesian Feedback Attribution

Credit assignment within a trajectory asks which decisions made a difference to what happened next. We formalize "made a difference" as a counterfactual comparison: does the observed feedback support the executed action more than the alternatives the policy itself could have taken from the same history? The key observation is that a language-model policy already contains everything needed to answer this question: its next-token likelihoods score arbitrary text continuations, so the frozen policy doubles as a feedback model. No learned dynamics model (Mesnard et al., 2021), return predictor (Arjona-Medina et al., 2019), or auxiliary loss (Wang et al., 2026) is required.

We define *feedback attribution* as the change in support over candidate actions induced by observing the next environment response. For history $h$, let $\mathcal{C} = (a^{(j)})_j$ be an indexed candidate list, $\rho_0(j \mid h, \mathcal{C}) > 0$ its prior, and $f_j(o') = f_{\theta_{\text{old}}}(o' \mid h, a^{(j)})$ its feedback likelihood. Conditioning on $o'$ gives

$$\rho(j \mid h, o', \mathcal{C}) = \frac{\rho_0(j \mid h, \mathcal{C}) f_j(o')}{m_{\mathcal{C}}(o' \mid h)}, \qquad m_{\mathcal{C}}(o' \mid h) = \sum_r \rho_0(r \mid h, \mathcal{C}) f_r(o'). \tag{4}$$

Here $m_{\mathcal{C}}$ is the prior predictive feedback likelihood. The posterior-to-prior ratio is

$$e_j = \frac{\rho(j \mid h, o', \mathcal{C})}{\rho_0(j \mid h, \mathcal{C})} = \frac{f_j(o')}{m_{\mathcal{C}}(o' \mid h)}. \tag{5}$$

BFA uses the executed candidate's lift $e_j$ to determine its decision weight. Eq. 5 is a hindsight ratio in the sense of Harutyunyan et al. (2019), evaluated densely on next-step observations rather than on episodic outcomes, and computed exactly by teacher forcing the observed text—the outcome model that prior counterfactual-credit methods had to learn (Harutyunyan et al., 2019; Mesnard et al., 2021) is replaced by the policy itself.

### 3.3 Pairwise Bayesian Feedback Attribution

Let $g_{\text{old}}$ denote the frozen behavior distribution, including the host's rollout decoding settings and action constraints. For each visited decision $(h, a, o')$, we retain the executed action and independently sample one alternative from the same distribution:

$$a^{(0)} = a, \qquad a^{(1)} \sim g_{\text{old}}(\cdot \mid h), \qquad \mathcal{C} = (a^{(0)}, a^{(1)}), \qquad \rho_0(j \mid h, \mathcal{C}) = \frac{1}{2}. \tag{6}$$

The samples define an equal-mass empirical prior over candidate indices.

The frozen model scores the same feedback event under both candidates: For $j \in \{0, 1\}$, we compute

$$s_j = \log f_{\theta_{\text{old}}}(o' \mid h, a^{(j)}) = \sum_{m \in \mathcal{V}(o')} \log \pi_{\theta_{\text{old}}}(o'_m \mid h, a^{(j)}, o'_{<m}), \tag{7}$$

where $\mathcal{V}(o')$ contains all target positions of this observation event. Bayes' rule yields the executed candidate's lift

$$e(h, a, o') = \frac{2 \exp(s_0)}{\exp(s_0) + \exp(s_1)} = 2\sigma(s_0 - s_1), \tag{8}$$

where $\sigma$ is the logistic function. Common positive factors in the two likelihoods cancel in this ratio, so the lift depends on their relative support for the observed response.

**Posterior-guided allocation.** For each decision with valid policy tokens, let $e_{i,t} = e(h_{i,t}, a_{i,t}, o_{i,t+1})$. We choose weights proportional to posterior lift, $\omega_{i,t} = c_i e_{i,t}$, and Eq. 3 determines $c_i$:

$$\omega_{i,t} = \frac{e_{i,t}}{\frac{1}{N_i} \sum_s n_{i,s} e_{i,s}}, \qquad \mu^*_{i,t} = b_{i,t} \omega_{i,t}. \tag{9}$$

Let $\Delta_i$ denote the probability simplex over decisions with valid policy tokens.

**Proposition 1** (Feedback Evidence and Conserved Allocation)**.** (i) *Fix $h$ and $\mathcal{C} = (a, \tilde{a})$, and let $p = p_{\text{env}}(\cdot \mid h, a)$, $f_0 = f_{\theta_{\text{old}}}(\cdot \mid h, a)$, $f_1 = f_{\theta_{\text{old}}}(\cdot \mid h, \tilde{a})$, $m = (f_0 + f_1)/2$. If $D_{\text{KL}}(p \| f_0)$ is finite, then $D_{\text{KL}}(p \| m) \le D_{\text{KL}}(p \| f_0) + \log 2$ is finite as well, and*

$$\mathbb{E}_{o' \sim p}[\log e(h, a, o')] = D_{\text{KL}}(p \| m) - D_{\text{KL}}(p \| f_0), \tag{10}$$

*which equals $D_{\text{KL}}(p \| m) \ge 0$ when $f_0 = p$.*

(ii) *For a fixed trajectory with $b_{i,t} > 0$ and $e_{i,t} > 0$, the BFA allocation $\mu_i^*$ is the unique solution of*

$$\mu_i^* = \arg\max_{\mu \in \Delta_i} \left\{ \sum_t \mu_t \log e_{i,t} - D_{\mathrm{KL}}(\mu \| b_i) \right\}. \tag{11}$$

Part (i) identifies expected log lift with the executed-action model's predictive log-score advantage over the candidate mixture; the expectation is nonnegative exactly when the model predicts the observed feedback better under the executed action than under the mixture. Part (ii) shows the allocation maximizes log lift under a KL penalty against the original token allocation (proof in Appendix A.1). The resulting trajectory surrogate is

$$\bar{\ell}_i^{\mathcal{A},\mathrm{BFA}}(\theta) = \frac{1}{N_i} \sum_{t=1}^{T_i} \sum_{k=1}^{n_{i,t}} \mathrm{sg}[\omega_{i,t}] \ell_{i,t,k}^{\mathcal{A}}(\theta), \tag{12}$$

where the host learner supplies the advantage and clipping rule, and BFA only rescales each decision's surrogate gradient by its positive weight, preserving the host signal's sign.

**Scope: influence, not value.** The lift measures a decision's *influence on the subsequent observation*, not the value of the action: a wrong action can produce highly predictable feedback. The host advantage alone supplies value and sign; BFA only decides how the signed update is distributed across decisions. The case where the two disagree—a helpful action inside a failed trajectory, which influence-based weighting would penalize—is measurable directly: with rollout-estimated reference advantages (Appendix E), BFA reduces the coefficient mass on such decisions from 6.9% to 3.2% at the final checkpoint (Section 4.6).

**Neutral point and exact no-op case.** A decision keeps exactly its host share when its lift equals the trajectory's token-weighted mean lift, since $\omega_{i,t} = e_{i,t}/(\frac{1}{N_i}\sum_s n_{i,s} e_{i,s}) = 1$ at that point; mass moves only between decisions whose lifts differ. If the sampled alternative coincides with the executed action, the two scores agree and the lift is exactly 1. If every decision in a trajectory has the same lift—for example, when the feedback is templated and does not depend on the action—the conservation rule gives $\omega_{i,t} = 1$ and $\mu_i = b_i$: BFA reduces *exactly* to the host's token allocation.

**The contrast shares targets, not probabilities.** Both $s_0$ and $s_1$ in Eq. 8 sum over the same target positions $\mathcal{V}(o')$ of the same string, so $s_0 - s_1$ is a sum of per-token log-likelihood ratios. A position contributes to the contrast only where the two candidates assign different probability to the same observed token; positions whose continuation does not depend on the action, including templated boilerplate, contribute nothing, however many of them there are. Absolute likelihood scores shrink mechanically with feedback length and need explicit per-token normalization, as in the executed-likelihood control of Appendix D; the contrast does not. What does grow with length is genuine per-token disagreement between the candidates, and a long stretch of it saturates the sigmoid at 0 or 2.

### 3.4 Trajectory Mass Normalization

We use each complete interaction trajectory as an outer aggregation unit. Let $M_i(\theta) = \bar{\ell}_i^{\mathcal{A},\mathrm{BFA}}(\theta)$ denote the trajectory surrogate in Eq. 12. Under global valid-token averaging,

$$q_i^{\mathrm{flat}} = \frac{N_i}{\sum_j N_j}, \qquad \mathcal{L}_{\mathrm{flat}}(\theta) = \sum_i q_i^{\mathrm{flat}} M_i(\theta). \tag{13}$$

TMN instead averages the complete-trajectory surrogates uniformly:

$$q_i^{\mathrm{TMN}} = \frac{1}{B}, \qquad \mathcal{L}_{\mathrm{TMN}}(\theta) = \frac{1}{B} \sum_i M_i(\theta). \tag{14}$$

**Two aggregation measures.** Rollouts are drawn i.i.d. from the policy, so the outer weights are a property of the reducer, not of the sampler. Flat aggregation averages the surrogate over the pooled generated tokens; equivalently, it computes the expectation of the trajectory-mean surrogate under the size-biased transform (Patil & Rao, 1978) of the episode distribution, $\mathbb{E}^*[M] = \mathbb{E}[NM]/\mathbb{E}[N] =$

$\mathbb{E}[M] + \mathrm{Cov}(N, M)/\mathbb{E}[N]$. TMN computes the episode-uniform expectation $\mathbb{E}[M]$. Flat aggregation also matches the relative weighting of the vanilla episodic policy gradient at the rollout policy, up to a batch-dependent scalar; TMN departs from it for four reasons: (i) evaluation is episode-uniform—success rate and exact match give each episode one vote, regardless of how many tokens it generated; (ii) $N_i$ mixes decision mass with decoding verbosity, so token-proportional weighting couples a trajectory's influence to a proxy rather than to its content; (iii) with one scalar advantage per trajectory, the episode is the informative unit—spreading the same signed signal over more tokens adds no information, only weight; and (iv) the trust-region budget is finite: Figure 1 shows flat aggregation hands 60.3% of the update mass on ALFWorld to the longest quartile of episodes. The following proposition quantifies exactly what the choice amounts to.

**Proposition 2** (Length–Gradient Coupling)**.** *Fix the rollout batch, host advantages, and attribution weights, with $N_i > 0$. Let $g_i = \nabla_\theta M_i(\theta)$, $\bar{g} = B^{-1}\sum_i g_i$, and $\bar{N} = B^{-1}\sum_i N_i$. The gradients $g_{\mathrm{flat}} = \nabla_\theta \mathcal{L}_{\mathrm{flat}}$ and $g_{\mathrm{TMN}} = \nabla_\theta \mathcal{L}_{\mathrm{TMN}}$ satisfy*

$$g_{\mathrm{flat}} - g_{\mathrm{TMN}} = \frac{\mathrm{Cov}_{\mathcal{B}}(N, g)}{\bar{N}}, \tag{15}$$

*where* $\mathrm{Cov}_{\mathcal{B}}(N, g) = B^{-1}\sum_i (N_i - \bar{N})(g_i - \bar{g})$.

Proposition 2 is the exact finite-batch instance of the size-bias identity above: it characterizes the update term on which the two reducers differ, per batch, in terms of how episode length covaries with episode gradient. The trajectory gradients $g_i$ include the host advantage, clipping, and BFA weights. The proof is given in Appendix A.2.

### 3.5 Unified BATON Objective

Combining BFA and TMN gives

$$\mathcal{L}_{\mathrm{BATON}}(\theta) = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{N_i}\sum_{t=1}^{T_i}\sum_{k=1}^{n_{i,t}} \mathrm{sg}[\omega_{i,t}]\,\ell^{\mathcal{A}}_{i,t,k}(\theta). \tag{16}$$

Separately reduced host regularization terms retain their original reductions. For hosts with turn-level advantages such as GiGPO, the host computes all advantages before weighting; BFA rescales their coefficients without altering the advantage values or the host's centering. For a host using global valid-token averaging, the two coefficient choices yield a $2 \times 2$ design: the host uses $\omega_{i,t} = 1$ and $q_i = q_i^{\mathrm{flat}}$; BFA changes only $\omega_{i,t}$; TMN changes only $q_i$; BATON combines both changes.

## 4 Experiments

### 4.1 Experimental Setup

We evaluate Qwen2.5-Instruct models at 1.5B and 7B scales on ALFWorld and WebShop, and at 3B and 7B scales on search-augmented QA. Hyperparameters, feedback serialization, the counterfactual behavior distribution, and numerical safeguards are in Appendix H. We instantiate the framework with both GRPO and GiGPO as host learners and compare against prompting and reinforcement-learning baselines, including EnvRL. Matched variants share the same initial checkpoint, rollout budget, reward and advantage construction, optimization settings, and evaluation protocol, differing only in the component under study. All reported results are averaged over four independent training seeds, and we report mean $\pm$ standard deviation across seeds. Because matched variants share seeds, all comparisons between variants are paired: Appendix F reports per-seed paired bootstrap confidence intervals and Wilcoxon signed-rank tests for the margins that matter.

### 4.2 Main Results

Table 1 shows that the full dual-axis framework consistently improves both GRPO and GiGPO across ALFWorld and WebShop and at both model scales. With Qwen2.5-1.5B-Instruct, the ALFWorld success rate increases from 72.9% to 85.6% for GRPO and from 86.8% to 94.1% for GiGPO. On WebShop, the corresponding success rates improve from 57.1% to 68.2% and from 67.1% to 76.4%, respectively. The gains remain at the 7B scale: the framework improves ALFWorld from

Table 1: Main results on ALFWorld and WebShop. We report success rate (%) for each ALFWorld task domain and the overall result, and task score and success rate (%) on WebShop. Within each backbone, bold marks the highest mean in each metric column separately within the GRPO-based and GiGPO-based method groups.

| Method | ALFWorld | | | | | | | WebShop | |
|---|---|---|---|---|---|---|---|---|---|
| | Pick | Look | Clean | Heat | Cool | Pick2 | All | Score | Succ. |
| *Closed-Source Models* | | | | | | | | | |
| GPT-4o | 75.5 | 60.5 | 31.6 | 56.5 | 21.9 | 49.4 | 48.1 | 32.1 | 23.5 |
| Gemini-2.5-Pro | 92.5 | 63.7 | 61.9 | 69.3 | 26.2 | 58.9 | 60.4 | 42.2 | 36.1 |
| *Qwen2.5-1.5B-Instruct* | | | | | | | | | |
| Qwen2.5 | 6.1 | 5.2 | 3.4 | 9.5 | 4.5 | 0.0 | 4.2 | 22.8 | 5.4 |
| ReAct | 17.1 | 20.9 | 15.5 | 6.5 | 7.6 | 2.2 | 12.9 | 40.4 | 11.1 |
| Reflexion | 35.7 | 22.0 | 22.0 | 13.3 | 19.6 | 3.6 | 22.0 | 55.5 | 22.1 |
| PPO (with critic) | $65.1_{\pm 3.3}$ | $40.1_{\pm 7.1}$ | $57.3_{\pm 5.0}$ | $60.3_{\pm 6.4}$ | $46.8_{\pm 4.1}$ | $47.2_{\pm 2.1}$ | $54.5_{\pm 3.0}$ | $73.5_{\pm 3.2}$ | $51.7_{\pm 2.8}$ |
| RLOO | $88.0_{\pm 3.2}$ | $53.2_{\pm 8.4}$ | $70.8_{\pm 6.1}$ | $63.1_{\pm 8.6}$ | $66.0_{\pm 5.7}$ | $57.1_{\pm 4.5}$ | $69.6_{\pm 2.6}$ | $74.2_{\pm 5.4}$ | $51.9_{\pm 6.8}$ |
| GRPO | $85.7_{\pm 1.6}$ | $53.4_{\pm 8.2}$ | $84.7_{\pm 6.6}$ | $77.8_{\pm 8.1}$ | $60.0_{\pm 4.8}$ | $53.3_{\pm 5.7}$ | $72.9_{\pm 3.4}$ | $75.6_{\pm 3.6}$ | $57.1_{\pm 3.6}$ |
| GRPO + EnvRL | $85.9_{\pm 2.3}$ | $53.9_{\pm 6.9}$ | $85.0_{\pm 4.5}$ | $82.0_{\pm 4.7}$ | $76.8_{\pm 3.9}$ | $71.9_{\pm 4.4}$ | $77.4_{\pm 3.8}$ | $83.0_{\pm 2.1}$ | $67.0_{\pm 2.6}$ |
| GRPO + BATON (Ours) | $\mathbf{91.0}_{\pm 3.6}$ | $\mathbf{80.1}_{\pm 4.9}$ | $\mathbf{91.9}_{\pm 3.4}$ | $\mathbf{83.6}_{\pm 3.8}$ | $\mathbf{83.9}_{\pm 3.5}$ | $\mathbf{78.5}_{\pm 4.7}$ | $\mathbf{85.6}_{\pm 2.8}$ | $\mathbf{85.2}_{\pm 2.0}$ | $\mathbf{68.2}_{\pm 2.8}$ |
| GiGPO | $94.6_{\pm 5.7}$ | $67.2_{\pm 4.8}$ | $94.5_{\pm 3.9}$ | $94.7_{\pm 7.6}$ | $79.4_{\pm 4.9}$ | $76.6_{\pm 5.3}$ | $86.8_{\pm 1.8}$ | $83.7_{\pm 1.9}$ | $67.1_{\pm 4.7}$ |
| GiGPO + EnvRL | $95.0_{\pm 3.3}$ | $70.8_{\pm 6.4}$ | $94.9_{\pm 4.6}$ | $97.7_{\pm 3.5}$ | $\mathbf{94.1}_{\pm 4.8}$ | $\mathbf{94.7}_{\pm 4.7}$ | $91.8_{\pm 3.5}$ | $88.2_{\pm 2.5}$ | $74.2_{\pm 3.7}$ |
| GiGPO + BATON (Ours) | $\mathbf{97.2}_{\pm 1.8}$ | $\mathbf{83.5}_{\pm 4.6}$ | $\mathbf{98.0}_{\pm 1.5}$ | $\mathbf{99.8}_{\pm 0.4}$ | $94.0_{\pm 3.1}$ | $85.6_{\pm 4.2}$ | $\mathbf{94.1}_{\pm 1.7}$ | $\mathbf{90.3}_{\pm 1.6}$ | $\mathbf{76.4}_{\pm 2.7}$ |
| *Qwen2.5-7B-Instruct* | | | | | | | | | |
| Qwen2.5 | 33.7 | 21.4 | 19.7 | 6.6 | 3.0 | 3.1 | 14.9 | 26.1 | 8.0 |
| ReAct | 48.2 | 35.8 | 34.1 | 13.5 | 17.9 | 17.8 | 31.3 | 46.5 | 19.3 |
| Reflexion | 62.4 | 41.3 | 45.1 | 30.5 | 36.6 | 23.6 | 42.8 | 57.9 | 29.1 |
| PPO (with critic) | $92.5_{\pm 3.8}$ | $63.7_{\pm 8.6}$ | $92.8_{\pm 2.5}$ | $89.2_{\pm 6.8}$ | $80.7_{\pm 2.2}$ | $68.6_{\pm 8.1}$ | $80.5_{\pm 2.8}$ | $81.1_{\pm 3.3}$ | $68.9_{\pm 4.9}$ |
| RLOO | $87.3_{\pm 4.5}$ | $78.6_{\pm 8.1}$ | $87.1_{\pm 6.0}$ | $81.6_{\pm 7.4}$ | $71.5_{\pm 5.3}$ | $49.1_{\pm 8.2}$ | $75.4_{\pm 4.8}$ | $80.5_{\pm 3.0}$ | $65.4_{\pm 4.2}$ |
| GRPO | $91.1_{\pm 4.9}$ | $65.7_{\pm 6.9}$ | $89.5_{\pm 5.2}$ | $74.4_{\pm 7.1}$ | $72.9_{\pm 5.5}$ | $64.5_{\pm 7.1}$ | $77.7_{\pm 5.0}$ | $79.0_{\pm 2.9}$ | $66.3_{\pm 3.5}$ |
| GRPO + EnvRL | $92.6_{\pm 4.3}$ | $\mathbf{77.5}_{\pm 5.4}$ | $\mathbf{92.9}_{\pm 3.9}$ | $\mathbf{79.7}_{\pm 3.0}$ | $77.4_{\pm 3.7}$ | $71.1_{\pm 6.2}$ | $80.4_{\pm 3.9}$ | $81.8_{\pm 1.9}$ | $68.6_{\pm 2.3}$ |
| GRPO + BATON (Ours) | $\mathbf{95.6}_{\pm 2.1}$ | $72.4_{\pm 5.3}$ | $90.3_{\pm 3.8}$ | $78.6_{\pm 4.9}$ | $\mathbf{94.1}_{\pm 3.2}$ | $\mathbf{75.3}_{\pm 4.6}$ | $\mathbf{86.9}_{\pm 2.7}$ | $\mathbf{83.4}_{\pm 1.9}$ | $\mathbf{70.3}_{\pm 2.5}$ |
| GiGPO | $97.5_{\pm 1.7}$ | $83.0_{\pm 7.7}$ | $98.6_{\pm 1.5}$ | $84.1_{\pm 7.4}$ | $89.0_{\pm 8.0}$ | $79.4_{\pm 6.8}$ | $90.9_{\pm 1.4}$ | $86.4_{\pm 2.4}$ | $74.9_{\pm 4.0}$ |
| GiGPO + EnvRL | $98.3_{\pm 1.5}$ | $92.2_{\pm 4.5}$ | $98.9_{\pm 0.8}$ | $93.8_{\pm 6.9}$ | $93.3_{\pm 3.5}$ | $\mathbf{94.6}_{\pm 3.0}$ | $94.5_{\pm 4.0}$ | $88.4_{\pm 4.9}$ | $76.3_{\pm 4.7}$ |
| GiGPO + BATON (Ours) | $\mathbf{99.8}_{\pm 0.4}$ | $\mathbf{92.5}_{\pm 3.6}$ | $\mathbf{99.8}_{\pm 0.4}$ | $\mathbf{94.5}_{\pm 3.1}$ | $\mathbf{94.9}_{\pm 2.8}$ | $88.3_{\pm 3.9}$ | $\mathbf{94.9}_{\pm 1.5}$ | $\mathbf{89.8}_{\pm 1.8}$ | $\mathbf{79.5}_{\pm 2.6}$ |

77.7% to 86.9% with GRPO and from 90.9% to 94.9% with GiGPO, while also improving their WebShop success rates from 66.3% to 70.3% and from 74.9% to 79.5%, and they persist on GiGPO, which already provides finer-grained within-trajectory credit.

### 4.3 Search-Augmented Question Answering

We further evaluate BATON on search-augmented QA, where agents interact with external retrieval systems during multi-step reasoning (Table 2). BATON improves both GRPO and GiGPO across model scales: over the host learners, GRPO rises from 37.9% to 41.1% on Qwen2.5-3B and from 42.8% to 45.9% on Qwen2.5-7B, and GiGPO from 42.0% to 43.2% and from 47.1% to 48.7%. The QA gains are smaller than the ALFWorld and WebShop gains, and part of the reason is that retrieval feedback is partly lexical: whether a query's phrasing matches the returned passages is only a proxy for whether the action caused the observation, and the contrast cannot fully separate the two.

### 4.4 Component Ablation

Table 3 provides a controlled $2 \times 2$ decomposition of BATON: the host learner alone, Bayesian Feedback Attribution only, TMN only, and their combination. Both axes yield substantial gains on their own. On ALFWorld with Qwen2.5-1.5B-Instruct, GRPO improves from 72.9% to 84.1% with Bayesian Feedback Attribution and to 82.4% with TMN, while BATON reaches 85.6%. For GiGPO, the corresponding results are 86.8%, 93.4%, 90.7%, and 94.1%.

The combination is best or tied for best in all sixteen cells of Table 3. The seed-mean margins over the better single axis range from $+0.2$ (GiGPO, WebShop 7B score) to $+3.1$ (GRPO, WebShop 1.5B score); Appendix F reports per-cell paired bootstrap intervals and Wilcoxon $p$-values for each. Both axes act on the same length-confounded pathology from different sides—TMN changes how much failed trajectories weigh, BFA changes where their signal lands—so the effects partially overlap.

Table 2: Search-augmented QA: exact match (%) on seven domains and their unweighted average, four seeds. External search-RL baselines are in Appendix I.

| Method | Single-Hop QA | | | Multi-Hop QA | | | | Overall |
|---|---|---|---|---|---|---|---|---|
| | NQ | TriviaQA | PopQA | HotpotQA | 2Wiki | MuSiQue | Bamboogle | Avg. |
| *Qwen2.5-3B-Instruct* | | | | | | | | |
| GRPO | $43.4_{\pm 2.8}$ | $62.1_{\pm 6.4}$ | $41.4_{\pm 5.1}$ | $37.8_{\pm 7.2}$ | $30.7_{\pm 4.6}$ | $14.8_{\pm 5.8}$ | $35.4_{\pm 3.7}$ | $37.9_{\pm 3.1}$ |
| GRPO + EnvRL | $44.6_{\pm 2.4}$ | $61.6_{\pm 6.9}$ | $42.3_{\pm 5.8}$ | $36.7_{\pm 5.3}$ | $31.1_{\pm 4.1}$ | $14.5_{\pm 4.9}$ | $39.0_{\pm 4.5}$ | $38.5_{\pm 3.6}$ |
| GRPO + BATON (Ours) | $\mathbf{46.2}_{\pm 2.1}$ | $\mathbf{63.3}_{\pm 5.2}$ | $\mathbf{43.8}_{\pm 4.7}$ | $\mathbf{40.2}_{\pm 4.4}$ | $\mathbf{35.1}_{\pm 3.8}$ | $\mathbf{16.0}_{\pm 5.6}$ | $\mathbf{43.4}_{\pm 3.1}$ | $\mathbf{41.1}_{\pm 2.5}$ |
| GiGPO | $42.2_{\pm 5.7}$ | $59.3_{\pm 4.8}$ | $42.6_{\pm 3.9}$ | $36.7_{\pm 7.6}$ | $36.8_{\pm 4.9}$ | $12.8_{\pm 5.3}$ | $63.8_{\pm 6.8}$ | $42.0_{\pm 4.7}$ |
| GiGPO + EnvRL | $\mathbf{47.3}_{\pm 3.3}$ | $\mathbf{63.1}_{\pm 5.9}$ | $\mathbf{44.2}_{\pm 4.6}$ | $\mathbf{42.0}_{\pm 3.5}$ | $37.8_{\pm 4.8}$ | $\mathbf{17.6}_{\pm 4.7}$ | $44.2_{\pm 5.2}$ | $42.3_{\pm 3.4}$ |
| GiGPO + BATON (Ours) | $42.9_{\pm 4.1}$ | $60.6_{\pm 6.1}$ | $44.1_{\pm 5.4}$ | $37.4_{\pm 5.8}$ | $\mathbf{39.9}_{\pm 4.2}$ | $13.3_{\pm 4.9}$ | $\mathbf{64.3}_{\pm 6.3}$ | $\mathbf{43.2}_{\pm 4.5}$ |
| *Qwen2.5-7B-Instruct* | | | | | | | | |
| GRPO | $47.7_{\pm 4.9}$ | $64.5_{\pm 6.9}$ | $45.5_{\pm 5.2}$ | $42.8_{\pm 7.1}$ | $38.6_{\pm 5.5}$ | $18.6_{\pm 6.2}$ | $42.2_{\pm 4.8}$ | $42.8_{\pm 3.6}$ |
| GRPO + EnvRL | $47.2_{\pm 4.3}$ | $64.8_{\pm 5.4}$ | $45.0_{\pm 3.9}$ | $\mathbf{44.2}_{\pm 3.0}$ | $37.3_{\pm 3.7}$ | $17.6_{\pm 6.2}$ | $42.6_{\pm 3.9}$ | $42.6_{\pm 2.8}$ |
| GRPO + BATON (Ours) | $\mathbf{48.4}_{\pm 3.6}$ | $\mathbf{66.5}_{\pm 4.9}$ | $\mathbf{46.3}_{\pm 3.4}$ | $43.7_{\pm 3.8}$ | $\mathbf{52.4}_{\pm 5.5}$ | $\mathbf{19.7}_{\pm 4.7}$ | $\mathbf{44.1}_{\pm 2.8}$ | $\mathbf{45.9}_{\pm 2.7}$ |
| GiGPO | $46.2_{\pm 1.7}$ | $64.9_{\pm 7.7}$ | $46.3_{\pm 1.5}$ | $41.4_{\pm 7.4}$ | $43.8_{\pm 8.0}$ | $18.7_{\pm 6.8}$ | $68.7_{\pm 6.4}$ | $47.1_{\pm 4.0}$ |
| GiGPO + EnvRL | $45.9_{\pm 2.5}$ | $65.1_{\pm 4.5}$ | $47.2_{\pm 2.8}$ | $42.1_{\pm 6.9}$ | $42.7_{\pm 3.5}$ | $19.4_{\pm 3.0}$ | $69.3_{\pm 4.0}$ | $47.5_{\pm 3.7}$ |
| GiGPO + BATON (Ours) | $\mathbf{47.8}_{\pm 1.8}$ | $\mathbf{65.6}_{\pm 3.6}$ | $\mathbf{49.5}_{\pm 1.9}$ | $\mathbf{43.3}_{\pm 3.1}$ | $\mathbf{45.0}_{\pm 3.8}$ | $\mathbf{20.4}_{\pm 4.5}$ | $\mathbf{69.4}_{\pm 2.7}$ | $\mathbf{48.7}_{\pm 2.6}$ |

Table 3: Component ablation of the two axes. Means $\pm$ std over four seeds.

| Component setting | ALFWorld | | WebShop | | | | Search-augmented QA | |
|---|---|---|---|---|---|---|---|---|
| | 1.5B | 7B | 1.5B | | 7B | | 3B | 7B |
| | All | All | Score | Succ. | Score | Succ. | Avg. | Avg. |
| GRPO | $72.9_{\pm 3.4}$ | $77.7_{\pm 5.0}$ | $75.6_{\pm 3.6}$ | $57.1_{\pm 3.6}$ | $79.0_{\pm 2.9}$ | $66.3_{\pm 3.5}$ | $37.9_{\pm 3.1}$ | $42.8_{\pm 3.6}$ |
| + BFA | $84.1_{\pm 0.7}$ | $85.4_{\pm 3.2}$ | $82.1_{\pm 1.8}$ | $67.3_{\pm 2.9}$ | $82.8_{\pm 1.3}$ | $69.4_{\pm 1.8}$ | $40.7_{\pm 2.0}$ | $44.2_{\pm 2.2}$ |
| + TMN | $82.4_{\pm 2.9}$ | $83.8_{\pm 3.0}$ | $79.6_{\pm 2.5}$ | $64.0_{\pm 3.2}$ | $81.5_{\pm 2.3}$ | $67.6_{\pm 2.9}$ | $39.3_{\pm 2.8}$ | $43.0_{\pm 2.9}$ |
| + BATON (Full) | $\mathbf{85.6}_{\pm 2.8}$ | $\mathbf{86.9}_{\pm 2.7}$ | $\mathbf{85.2}_{\pm 2.0}$ | $\mathbf{68.2}_{\pm 2.8}$ | $\mathbf{83.4}_{\pm 1.9}$ | $\mathbf{70.3}_{\pm 2.5}$ | $\mathbf{41.1}_{\pm 2.5}$ | $\mathbf{45.9}_{\pm 2.7}$ |
| GiGPO | $86.8_{\pm 1.8}$ | $90.9_{\pm 1.4}$ | $83.7_{\pm 1.9}$ | $67.1_{\pm 4.7}$ | $86.4_{\pm 2.4}$ | $74.9_{\pm 4.0}$ | $42.0_{\pm 4.7}$ | $47.1_{\pm 4.0}$ |
| + BFA | $93.4_{\pm 1.4}$ | $93.6_{\pm 2.7}$ | $88.3_{\pm 1.5}$ | $73.4_{\pm 2.1}$ | $89.6_{\pm 1.1}$ | $78.6_{\pm 3.7}$ | $42.3_{\pm 3.9}$ | $48.2_{\pm 3.4}$ |
| + TMN | $90.7_{\pm 1.9}$ | $91.3_{\pm 1.6}$ | $86.0_{\pm 1.8}$ | $71.6_{\pm 3.1}$ | $87.7_{\pm 2.1}$ | $76.3_{\pm 2.8}$ | $42.9_{\pm 3.1}$ | $47.4_{\pm 2.8}$ |
| + BATON (Full) | $\mathbf{94.1}_{\pm 1.7}$ | $\mathbf{94.9}_{\pm 1.5}$ | $\mathbf{90.3}_{\pm 1.6}$ | $\mathbf{76.4}_{\pm 2.7}$ | $\mathbf{89.8}_{\pm 1.8}$ | $\mathbf{79.5}_{\pm 2.6}$ | $\mathbf{43.2}_{\pm 4.5}$ | $\mathbf{48.7}_{\pm 2.6}$ |

### 4.5 Trajectory Execution Cost

We further examine whether the learned policies change their execution behavior (Figure 3). For GRPO, the full framework reduces the mean action count from 21.51 to 16.39 and the mean generated-token count from 1,734 to 1,301. For GiGPO, the corresponding quantities decrease from 13.51 to 10.82 actions and from 1,431 to 746 generated tokens: the gains come with shorter, less verbose trajectories at evaluation.

### 4.6 Additional Analyses

We report four supporting analyses, each with a dedicated appendix. (i) *Objective normalization*: with Qwen2.5-1.5B and GRPO, TMN beats global-token and constant-denominator reduction on ALFWorld and WebShop, with and without BFA, under shared and under validation-selected learning rates (Appendix C). (ii) *Feedback weighting*: with TMN fixed, BFA beats shuffled-lift, mismatched-feedback, executed-likelihood, and rank-matched controls (Appendix D). (iii) *Allocation*: sign-stratified attribution locates the gains on positive-advantage trajectories, and inverted lifts fall below uniform weighting (Appendix G); BFA also reduces coefficient mass on helpful actions inside failed trajectories from 6.9% to 3.2% at the final checkpoint (Appendix E). (iv) *Overhead*: training time increases by 17% (GRPO) and 10% (GiGPO) on ALFWorld at 1.5B (Appendix B).

## 5 Related Work

**Policy optimization for LLMs and agents.** PPO-based RLHF (Schulman et al., 2017; Ouyang et al., 2022) and the critic-free, group-relative RLOO and GRPO (Ahmadian et al., 2024; Shao

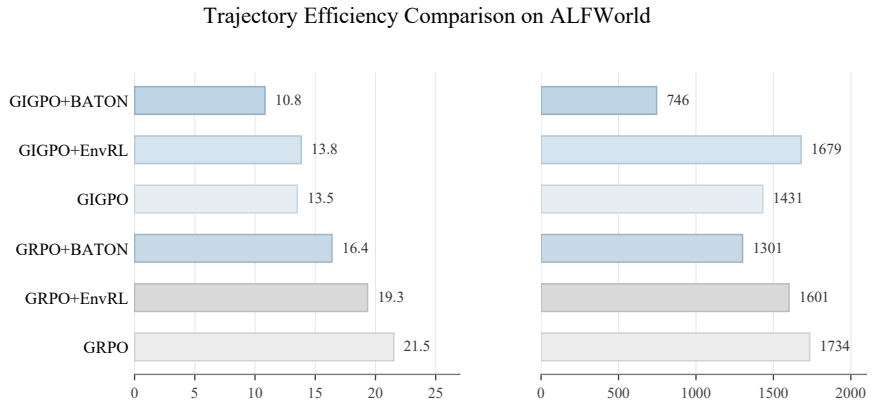


Figure 3: Trajectory execution cost on ALFWorld (Qwen2.5-1.5B): mean environment actions and generated tokens per evaluation episode at update 150, successes and failures included. Lower is better.

et al., 2024) ground LLM policy optimization, with recent refinements to clipping, reduction, and sequence-level objectives (Yu et al., 2025; Liu et al., 2025; Hu, 2025; Cui et al., 2025; Hu et al., 2025; Zeng et al., 2025b; Zheng et al., 2025a): original GRPO averages token terms within each response, DAPO adopts global token averaging, and Dr. GRPO uses a constant denominator. TMN applies sequence-mean reduction one level up, to complete multi-turn trajectories: global token averaging is the size-biased transform (Patil & Rao, 1978) of the episode distribution over trajectory surrogates, and Proposition 2 gives, to our knowledge, the first exact finite-batch characterization of what this transform changes per update. For interactive agents, WebRL and RAGEN study RL in environmental settings (Qi et al., 2024; Wang et al., 2025); GiGPO, turn-level credit assignment, and Agent Lightning refine within-trajectory optimization (Feng et al., 2025; Zeng et al., 2025a; Luo et al., 2025).

**Counterfactual and hindsight credit assignment.** COMA (Foerster et al., 2018) evaluates an action against a counterfactual baseline marginalized through a learned critic, and Hindsight Credit Assignment (Harutyunyan et al., 2019) reweights policy-gradient terms by the outcome ratio $p(G \mid a)/p(G)$; Mesnard et al. (2021) generalize the idea, and both identify learning the outcome model as the central obstacle. BFA is the same ratio with the next textual observation as the outcome: the horizon is one step, the ratio is exact by teacher forcing, and the outcome model is the frozen policy LM itself, so the learned component disappears. Language models already serve as world models at inference time (Hao et al., 2023; Gu et al., 2024); BFA uses the same property at training time. RUDDER (Arjona-Medina et al., 2019) and VinePPO (Kazemnejad et al., 2024) estimate *value*, by return decomposition and by forking rollouts; BFA estimates no value and allocates the host's existing advantage, which is why it composes with value-based turn credit such as GiGPO.

**Learning from interaction feedback.** Interaction feedback has long grounded language agents (Yao et al., 2023; Schick et al., 2023; Shinn et al., 2023; Madaan et al., 2023; Gou et al., 2024; Zhou et al., 2023; Wang et al., 2023), and recent work trains on it: AgentTuning and AgentGym learn from trajectory collections (Zeng et al., 2023; Xi et al., 2024), search agents are optimized with outcome- and step-level RL (Jin et al., 2025; Sun et al., 2025; Zheng et al., 2025b), and EnvRL adds state-prediction and inverse-dynamics objectives (Wang et al., 2026). BFA uses no auxiliary objective: the same likelihoods enter only as stop-gradient allocation weights.

## 6 Conclusion

BATON makes two implicit weighting choices of agent-RL objectives explicit: Bayesian Feedback Attribution reallocates the host advantage by a counterfactual feedback comparison the policy scores itself, and Trajectory Mass Normalization weights episodes uniformly. It adds no hyperparameters or environment interaction, improves GRPO and GiGPO on all three benchmarks at 1.5B–7B scales, and is best or tied for best in every ablation cell.

### AI USE STATEMENT

Generative AI tools were used solely for language polishing and improving the clarity and readability of the manuscript. They were not used for developing research ideas, methodology, theoretical results, experimental design, data analysis, or interpretation of results. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

### ETHICS STATEMENT

This work focuses on improving reinforcement-learning optimization for language-model agents and does not involve experiments with human participants. Our experiments are conducted on established research benchmarks and simulated or tool-based environments. While improved agent capabilities may be applicable to a broad range of downstream tasks, including potentially harmful ones, the methods developed in this work are intended for general research on reliable and effective agent training. We encourage responsible deployment and appropriate safeguards when applying such techniques in real-world settings.

### REPRODUCIBILITY STATEMENT

We provide detailed reproducibility information throughout the paper and appendices. Formal proofs are provided in Appendix A, computational overhead analysis in Appendix B, controlled analyses and ablations in the subsequent appendices, and implementation details including feedback serialization, token masking, action representation, and numerical computation in Appendix H. Experimental settings, models, benchmarks, and evaluation protocols are described in Section 4 and the corresponding appendices.

## REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return Decomposition for Delayed Rewards. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1806.07857.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process Reinforcement through Implicit Rewards, 2025. URL https://arxiv.org/abs/2502.01456v1.

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

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual Multi-Agent Policy Gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. URL https://arxiv.org/abs/1705.08926.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Sx038qxjek.

Yu Gu, Kai Zhang, Yuting Ning, Boyuan Zheng, Boyu Gou, Tianci Xue, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun, and Yu Su. Is Your LLM Secretly a World Model of the Internet? Model-Based Planning for Web Agents, 2024. URL https://arxiv.org/abs/2411.06559.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with Language Model is Planning with World Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://aclanthology.org/2023.emnlp-main.507/.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado P. van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, and Remi Munos. Hindsight Credit Assignment. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/195f15384c2a79cedf293e4a847ce85c-Abstract.html.

Jian Hu. REINFORCE++: A Simple and Efficient Approach for Aligning Large Language Models, 2025. URL https://arxiv.org/abs/2501.03262v1.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-Reasoner-Zero: An Open Source Approach to Scaling Up Reinforcement Learning on the Base Model, 2025. URL https://arxiv.org/abs/2503.24290.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning, 2025. URL https://arxiv.org/abs/2503.09516.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Unlocking RL Potential For LLM Reasoning Through Refined Credit Assignment, 2024. URL https://arxiv.org/abs/2410.01679v1.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K. Qiu, and Yuqing Yang. Agent Lightning: Train ANY AI Agents with Reinforcement Learning, 2025. URL https://arxiv.org/abs/2508.03680.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback, 2023. URL https://arxiv.org/abs/2303.17651.

Thomas Mesnard, Theophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Thomas S. Stepleton, Nicolas Heess, Arthur Guez, Eric Moulines, Marcus Hutter, Lars Buesing, and Remi Munos. Counterfactual Credit Assignment in Model-Free Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7654–7664. PMLR, 2021. URL https://proceedings.mlr.press/v139/mesnard21a.html.

Long Ouyang, Jeff 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Ganapati P. Patil and C. Radhakrishna Rao. Weighted Distributions and Size-Biased Sampling with Applications to Wildlife Populations and Human Families. *Biometrics*, 34(2):179–189, 1978. doi: 10.2307/2530008.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Wenyi Zhao, Yu Yang, Xinyue Yang, Jiadai Sun, Shuntian Yao, Tianjie Zhang, Wei Xu, Jie Tang, and Yuxiao Dong. WebRL: Training LLM Web Agents via Self-Evolving Online Curriculum Reinforcement Learning, 2024. URL https://arxiv.org/abs/2411.02337.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools, 2023. URL https://arxiv.org/abs/2302.04761.

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, 2024. URL https://arxiv.org/abs/2402.03300.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning, 2023. URL https://arxiv.org/abs/2303.11366.

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://arxiv.org/abs/2010.03768.

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Fei Huang, and Yan Zhang. ZeroSearch: Incentivize the Search Capability of LLMs without Searching, 2025. URL https://arxiv.org/abs/2505.04588v1.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An Open-Ended Embodied Agent with Large Language Models, 2023. URL https://arxiv.org/abs/2305.16291.

Zhitong Wang, Songze Li, Hao Peng, Shuzheng Si, Yi Wang, Maosong Sun, and Juanzi Li. Envrl: Learn from environment dynamics in agentic reinforcement learning, 2026. URL https://arxiv.org/abs/2606.17680.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Monica Lam, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Fei-Fei Li, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN: Understanding Self-Evolution in LLM Agents via Multi-Turn Reinforcement Learning, 2025. URL https://arxiv.org/abs/2504.20073v1.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Dingwen Yang, Chenyang Liao, Xin Guo, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. AgentGym: Evolving Large Language Model-based Agents across Diverse Environments, 2024. URL https://arxiv.org/abs/2406.04151.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents, 2022. URL https://arxiv.org/abs/2207.01206.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao

Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. AgentTuning: Enabling Generalized Agent Abilities for LLMs, 2023. URL https://arxiv.org/abs/2310.12823.

Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, and Mingyi Hong. Reinforcing Multi-Turn Reasoning in LLM Agents via Turn-Level Credit Assignment, 2025a. URL https://arxiv.org/abs/2505.11821v1.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. SimpleRL-Zoo: Investigating and Taming Zero Reinforcement Learning for Open Base Models in the Wild, 2025b. URL https://arxiv.org/abs/2503.18892.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group Sequence Policy Optimization, 2025a. URL https://arxiv.org/abs/2507.18071.

Xuhui Zheng, Kang An, Ziliang Wang, Yuhang Wang, and Yichao Wu. StepSearch: Igniting LLMs Search Ability via Step-Wise Proximal Policy Optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21805–21830. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.emnlp-main.1106. URL https://aclanthology.org/2025.emnlp-main.1106/.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language Agent Tree Search Unifies Reasoning Acting and Planning in Language Models, 2023. URL https://arxiv.org/abs/2310.04406.

# A PROOF

## A.1 PROOF OF PROPOSITION 1

*Proof.* **(i) Feedback evidence.** Fix $h$ and $\mathcal{C} = (a, \tilde{a})$. Write $p = p_{\text{env}}(\cdot \mid h, a)$, $f_0 = f_{\theta_{\text{old}}}(\cdot \mid h, a)$, $f_1 = f_{\theta_{\text{old}}}(\cdot \mid h, \tilde{a})$, and $m = (f_0 + f_1)/2$. All distributions use the same feedback observation space. Since $e(h, a, o') = f_0(o')/m(o')$, the finite-KL assumptions give

$$\begin{aligned}\mathbb{E}_{o' \sim p}[\log e(h, a, o')] &= \mathbb{E}_p\left[\log \frac{f_0(o')}{m(o')}\right] \\ &= \mathbb{E}_p\left[\log \frac{p(o')}{m(o')}\right] - \mathbb{E}_p\left[\log \frac{p(o')}{f_0(o')}\right] \\ &= D_{\text{KL}}(p\|m) - D_{\text{KL}}(p\|f_0).\end{aligned}$$

When $f_0 = p$, the second term vanishes, yielding $\mathbb{E}_p[\log e] = D_{\text{KL}}(p\|m) \geq 0$.

**(ii) Conserved allocation.** Fix a trajectory and let

$$Z_i = \sum_t b_{i,t} e_{i,t}, \qquad \mu^*_{i,t} = \frac{b_{i,t} e_{i,t}}{Z_i}.$$

Because $b_{i,t} > 0$ and $e_{i,t} > 0$, $Z_i > 0$ and $\mu^*_i \in \Delta_i$. For any $\mu \in \Delta_i$, define

$$\Phi_i(\mu) = \sum_t \mu_t \log e_{i,t} - D_{\text{KL}}(\mu\|b_i).$$

Substituting the expression for $\mu^*_i$ gives

$$\begin{aligned}\Phi_i(\mu) &= \sum_t \mu_t \log \frac{b_{i,t} e_{i,t}}{\mu_t} \\ &= \sum_t \mu_t \log \frac{Z_i \mu^*_{i,t}}{\mu_t} \\ &= \log Z_i - D_{\text{KL}}(\mu\|\mu^*_i).\end{aligned}$$

The KL term is nonnegative and equals zero exactly when $\mu = \mu_i^*$. Hence $\mu_i^*$ is the unique maximizer, with

$$\Phi_i(\mu_i^*) - \Phi_i(\mu) = D_{\mathrm{KL}}(\mu \| \mu_i^*).$$

Finally,

$$\frac{\mu_{i,t}^*}{b_{i,t}} = \frac{e_{i,t}}{\sum_s b_{i,s} e_{i,s}} = \frac{e_{i,t}}{N_i^{-1} \sum_s n_{i,s} e_{i,s}} = \omega_{i,t},$$

which recovers the weights in Eq. 9. □

### A.2 Proof of Proposition 2

*Proof.* The rollout-derived counts, host advantages, and attribution weights are held fixed when differentiating the policy surrogate. Hence

$$g_{\mathrm{flat}} = \frac{\sum_i N_i g_i}{\sum_i N_i} = \frac{\mathbb{E}_{\mathcal{B}}[Ng]}{\bar{N}}, \qquad g_{\mathrm{TMN}} = \mathbb{E}_{\mathcal{B}}[g] = \bar{g},$$

where $\mathbb{E}_{\mathcal{B}}$ denotes the uniform empirical average over the $B$ trajectories. Expanding the covariance gives

$$\mathrm{Cov}_{\mathcal{B}}(N, g) = \mathbb{E}_{\mathcal{B}}[(N - \bar{N})(g - \bar{g})] = \mathbb{E}_{\mathcal{B}}[Ng] - \bar{N}\bar{g}.$$

Since $\bar{N} > 0$, division by $\bar{N}$ yields

$$\frac{\mathrm{Cov}_{\mathcal{B}}(N, g)}{\bar{N}} = \frac{\mathbb{E}_{\mathcal{B}}[Ng]}{\bar{N}} - \bar{g} = g_{\mathrm{flat}} - g_{\mathrm{TMN}}.$$

Here $\mathrm{Cov}_{\mathcal{B}}(N, g)$ is a parameter-space vector (scalar length times vector gradient), and the identity holds coordinate-wise. □

## B Computational Overhead Analysis

We analyze the additional computational cost introduced by BATON. All measurements are conducted on ALFWorld with Qwen2.5-1.5B-Instruct under the same rollout budget, batch size, optimizer settings, and hardware configuration as the main experiments.

BATON does not introduce additional environment interaction. The overhead mainly comes from Bayesian Feedback Attribution, which requires additional feedback likelihood evaluation for attribution weighting. Trajectory Mass Normalization only changes the trajectory-level aggregation rule and therefore introduces negligible computational cost.

Table 4 reports the total wall-clock training time and the relative cost normalized by the corresponding host learner.

Table 4: Wall-clock training overhead of BATON on ALFWorld. Training time is measured over the complete training process under identical rollout and optimization settings. Values are normalized by the corresponding host learner.

| Method | Training Time (hours) | Relative Cost |
|---|---|---|
| GRPO | 18.05 | 1.00× |
| GRPO + TMN | 18.50 | 1.02× |
| GRPO + BATON | 21.07 | 1.17× |
| GiGPO | 22.77 | 1.00× |
| GiGPO + TMN | 23.08 | 1.01× |
| GiGPO + BATON | 25.15 | 1.10× |

## C Comparison of Objective Normalization Schemes

We evaluated objective normalization with Qwen2.5-1.5B-Instruct and GRPO on ALFWorld and WebShop. The supplementary performance results are means over two training seeds distinct from the four used in the main experiments. Across variants, we kept the initial checkpoint, training-task schedule, rollout budget, reward, advantage estimator, clipping rule, and evaluation protocol fixed. Regularization terms retained the host learner's reductions and coefficients.

**Normalization variants.** For a batch of $B$ trajectories, let $n_{i,t}$ be the valid policy-token count at decision $t$ and $N_i = \sum_t n_{i,t}$. We define

$$S_i(\theta) = \sum_{t=1}^{T_i} \sum_{k=1}^{n_{i,t}} \mathrm{sg}[\omega_{i,t}] \ell^{\mathcal{A}}_{i,t,k}(\theta) \tag{17}$$

and compare three reducers:

$$\mathcal{L}_{\mathrm{token}}(\theta) = \frac{\sum_i S_i(\theta)}{\sum_i N_i}, \tag{18}$$

$$\mathcal{L}_{\mathrm{traj}}(\theta) = \frac{1}{B} \sum_i \frac{S_i(\theta)}{N_i}, \tag{19}$$

$$\mathcal{L}_{\mathrm{const}}(\theta) = \frac{1}{BC} \sum_i S_i(\theta), \tag{20}$$

where $C$ is fixed throughout training. We set $C = 1024$ valid policy tokens per complete trajectory for all environments; the numerator accumulates only valid policy-token surrogate terms, excluding prompt, observation, and padding tokens, so $C$ is a normalization constant, not a trajectory-length limit. Each reducer was evaluated with and without BFA. Complete-trajectory mean reduction applies sequence-mean reduction to the full multi-turn trajectory. For a fixed batch, $\mathcal{L}_{\mathrm{const}} = (\bar{N}/C)\mathcal{L}_{\mathrm{token}}$, where $\bar{N} = B^{-1} \sum_i N_i$. This comparison separates batch-dependent scaling from the relative weighting of trajectories.

We also evaluated turn-mean reduction without BFA:

$$\mathcal{L}_{\mathrm{turn}}(\theta) = \frac{1}{B} \sum_i \frac{1}{T_i} \sum_{t=1}^{T_i} \frac{1}{n_{i,t}} \sum_{k=1}^{n_{i,t}} \ell^{\mathcal{A}}_{i,t,k}(\theta), \tag{21}$$

where $T_i$ counts decisions with valid policy tokens. This variant assigns equal mass to decisions within each trajectory, whereas TMN assigns equal mass to its policy tokens before BFA reweighting.

**Optimization settings and results.** We compared shared optimization settings and a separate learning-rate selection using an equal search budget on the validation set for every reducer. Table 5 reports both comparisons. TMN with BFA achieves the highest mean on all three metrics in both settings. After learning-rate selection, TMN improves ALFWorld success over the constant-denominator reducer from 80.5% to 82.8% without BFA and from 84.3% to 86.0% with BFA. The corresponding WebShop success rates increase from 63.1% to 65.1% and from 67.4% to 69.0%. Without BFA, TMN also exceeds turn-mean reduction in success rate on both environments. These comparisons favor complete-trajectory averaging under both shared and validation-selected learning rates.

## D Controls for Feedback-Based Weighting

We fixed the outer reducer to TMN and varied the information used to assign decision weights. Let $b_{i,t} = n_{i,t}/N_i$. Every variant produces a positive raw score $r_{i,t}$ and uses the same conservation rule:

$$\omega_{i,t} = \frac{r_{i,t}}{\sum_s b_{i,s} r_{i,s}}, \qquad \mu_{i,t} = b_{i,t}\omega_{i,t}, \qquad \sum_t \mu_{i,t} = 1. \tag{22}$$

We used the same observation serialization, feedback-token mask, and missing-feedback handling across variants, and held the weights fixed during policy optimization.

**Weighting variants.** **Uniform** sets $r_{i,t} = 1$. **Shuffled lift** permutes the original BFA lifts across decisions within each trajectory before normalization. Permutations are restricted to action-token-length strata where feasible; singleton strata remain unchanged. This retains the raw lift values while disrupting their association with individual decisions. **Executed likelihood only** uses

$$r_{i,t} = \exp\left(\frac{s_{0,i,t}}{L_{i,t}}\right), \qquad L_{i,t} = |V(o'_{i,t})|, \tag{23}$$

Table 5: Normalization comparison with Qwen2.5-1.5B-Instruct and GRPO. Results are means over two training seeds distinct from the four seeds used in the main experiments. We report shared optimization settings and equal-budget validation selection of the learning rate separately.

| | | ALFWorld | WebShop | |
|---|---|---|---|---|
| Reducer | BFA | Success (%) | Score | Success (%) |
| *Shared optimization configuration* | | | | |
| Global token mean | No | 73.3 | 75.9 | 57.2 |
| Global token mean | Yes | 84.4 | 82.4 | 67.2 |
| Trajectory mean (TMN) | No | 82.6 | 79.5 | 63.8 |
| Trajectory mean (TMN) | Yes | 85.8 | 84.9 | 68.3 |
| Constant denominator | No | 79.0 | 78.1 | 62.1 |
| Constant denominator | Yes | 83.5 | 82.5 | 66.3 |
| Turn mean | No | 80.4 | 78.7 | 62.7 |
| *Equal-budget validation selection of learning rate* | | | | |
| Global token mean | No | 74.0 | 76.1 | 57.8 |
| Global token mean | Yes | 84.7 | 83.3 | 67.4 |
| Trajectory mean (TMN) | No | 82.8 | 80.3 | 65.1 |
| Trajectory mean (TMN) | Yes | 86.0 | 85.8 | 69.0 |
| Constant denominator | No | 80.5 | 79.6 | 63.1 |
| Constant denominator | Yes | 84.3 | 83.5 | 67.4 |
| Turn mean | No | 81.7 | 80.3 | 63.5 |

Table 6: Feedback-weighting controls with a fixed TMN reducer. Results are means over the same two supplementary training seeds. All variants conserve each trajectory's total coefficient mass.

| | ALFWorld | WebShop | |
|---|---|---|---|
| Weighting | Success (%) | Score | Success (%) |
| Uniform | 82.6 | 79.5 | 63.8 |
| Shuffled lift | 82.6 | 80.4 | 64.5 |
| Executed likelihood only | 83.8 | 81.5 | 65.3 |
| Mismatched feedback | 82.6 | 79.7 | 64.1 |
| BFA | 85.8 | 84.9 | 68.3 |
| Rank-matched likelihood | 84.5 | 82.9 | 66.7 |

where $s_{0,i,t}$ is the observed-feedback log likelihood under the executed action. This score measures feedback predictability without comparing candidate actions. **Mismatched feedback** substitutes feedback from another trajectory in the same environment, matched by event type and approximate token length, and computes

$$r_{i,t} = 2\sigma(\widetilde{s}_{0,i,t} - \widetilde{s}_{1,i,t}). \tag{24}$$

Only the feedback used for weighting changes; the original actions, trajectory rewards, and host advantages are retained. **BFA** evaluates the true observed feedback with the original score $r_{i,t} = 2\sigma(s_{0,i,t} - s_{1,i,t})$. **Rank-matched likelihood** assigns the sorted BFA lifts according to the ranking induced by $s_{0,i,t}/L_{i,t}$ and then applies the same conservation rule. It retains the multiset of raw lifts while replacing the ordering supplied by the candidate-action comparison.

**Results.** Table 6 shows that BFA achieves the highest mean performance across the three metrics. Shuffled lift and mismatched feedback both reach 82.6% ALFWorld success, matching uniform weighting, whereas BFA reaches 85.8%. On WebShop, BFA reaches 68.3% success, compared with 64.5% for shuffled lift and 64.1% for mismatched feedback. Executed-likelihood weighting improves over uniform weighting, and rank matching further narrows the gap to BFA. BFA nevertheless exceeds rank-matched likelihood by 1.3 percentage points on ALFWorld and 1.6 points on WebShop success. These comparisons support the value of matching weights to decisions and of using relative feedback support across candidate actions.

Table 7: Rollout-based action-advantage diagnostic on ALFWorld. Tasks and decisions are counts per seed; trajectories are averaged over the two seeds. Decision coverage is reported as a two-seed mean. $M_Q$ and $M_V$ denote the numbers of continuations used to estimate $Q$ and $V$. Decision coverage is the fraction of evaluated decisions whose reference-advantage interval excludes zero and whose host advantage is nonzero.

| Checkpoint | Seeds | Tasks | Trajectories | Decisions | $M_Q$ | $M_V$ | Decision coverage (%) |
|---|---|---|---|---|---|---|---|
| Early | 2 | 40 | 160 | 2361 | 16 | 16 | 57.0 |
| Middle | 2 | 40 | 160 | 2089 | 16 | 16 | 62.6 |
| Final | 2 | 40 | 160 | 1857 | 16 | 16 | 68.5 |

# E Alignment with Rollout-Estimated Action Advantages

We examined how BFA redistributes coefficient mass among decisions whose host advantage agrees or conflicts with a rollout-estimated reference advantage. We used saved ALFWorld states at early, middle, and final training checkpoints. Uniform and BFA allocations were evaluated on the same trajectories and reference labels.

**Reference advantages.** We sampled complete trajectories from held-out diagnostic tasks, including successful and failed episodes, and evaluated their valid decisions. For each saved history $h$ and executed action $a$, we restored the environment state and estimated

$$\widehat{Q}^{\pi}(h,a) = \frac{1}{M_Q}\sum_{m=1}^{M_Q} R^{(m)}_{\text{force } a,\,\pi}, \tag{25}$$

$$\widehat{V}^{\pi}(h) = \frac{1}{M_V}\sum_{m=1}^{M_V} R^{(m)}_{\pi}, \tag{26}$$

$$\widehat{A}^{\text{ref}}(h,a) = \widehat{Q}^{\pi}(h,a) - \widehat{V}^{\pi}(h). \tag{27}$$

The $Q$ continuations execute $a$ first and then follow the frozen policy $\pi$; the $V$ continuations follow $\pi$ from the saved decision point. We used $M_Q = M_V = 16$ with the same remaining interaction budgets and termination rules. Table 7 summarizes the diagnostic sample. A decision is resolved, $z_{i,t} = 1$, when its reference-advantage interval excludes zero and its host advantage is nonzero; otherwise, $z_{i,t} = 0$.

**Allocation categories.** For $w_{i,t} = b_{i,t}$ under uniform weighting or $w_{i,t} = \mu_{i,t}$ under BFA, we compute

$$M_{u,v}(w) = \frac{1}{B_{\text{diag}}}\sum_i\sum_t w_{i,t} z_{i,t}\,\mathbf{1}\Big\{\text{sign}(A^{\text{host}}_{i,t}) = u,\ \text{sign}(\widehat{A}^{\text{ref}}_{i,t}) = v\Big\}, \tag{28}$$

where $u, v \in \{-1, +1\}$ and $B_{\text{diag}}$ is the number of diagnostic trajectories. For GRPO, the host advantage is shared by all decisions in a trajectory. The aligned categories are $M_{+,+}$ and $M_{-,-}$; the conflicting categories are $M_{+,-}$ and $M_{-,+}$. We define

$$C(w) = \frac{1}{B_{\text{diag}}}\sum_i\sum_t w_{i,t} z_{i,t}, \tag{29}$$

$$M_{\text{conflict}}(w) = M_{+,-}(w) + M_{-,+}(w), \tag{30}$$

$$M_{\text{unresolved}}(w) = 1 - C(w). \tag{31}$$

Conditional conflict is $M_{\text{conflict}}(w)/C(w)$.

**Results.** Table 8 reports the allocation across these categories. At all three checkpoints, BFA assigns more mass to the two aligned categories and less mass to both conflicting categories. BFA reduces the mass assigned to positive host advantages on actions with negative reference advantage, and to negative host advantages on actions with positive reference advantage. The redistribution favors decisions whose host update direction agrees with the rollout-based reference.

Table 8: Allocation mass by agreement between the host advantage and the rollout-estimated reference advantage, reported as two-seed means (%). The four sign categories and unresolved category partition the trajectory allocation mass. Conditional conflict is conflicting mass divided by resolved mass. Uniform and BFA use the same trajectories and reference labels.

| Checkpoint | Allocation | $M_{+,+}$ | $M_{-,-}$ | $M_{+,-}$ | $M_{-,+}$ | Unresolved (%) | Conditional conflict (%) |
|---|---|---|---|---|---|---|---|
| Early | Uniform | 32.1 | 19.6 | 15.3 | 9.9 | 23.1 | 32.8 |
| Early | BFA | 38.0 | 23.4 | 9.7 | 6.1 | 22.8 | 20.5 |
| Middle | Uniform | 34.8 | 22.3 | 13.2 | 7.7 | 22.0 | 26.8 |
| Middle | BFA | 42.1 | 26.7 | 6.8 | 4.0 | 20.4 | 13.6 |
| Final | Uniform | 38.2 | 23.8 | 10.1 | 6.9 | 21.0 | 21.5 |
| Final | BFA | 43.9 | 28.6 | 5.0 | 3.2 | 19.3 | 10.2 |

## F Paired Statistical Analysis

All matched variants in this paper share their training seeds, so every comparison between variants is paired at the seed level. For each cell and each margin of interest (BATON − host; BATON − better single component), we compute the per-seed difference and report a paired bootstrap 95% confidence interval (10,000 resamples over seeds) and a two-sided Wilcoxon signed-rank $p$-value. Table 9 collects these statistics for the sixteen ablation cells of Table 3. BATON has a positive mean margin over the better single component in all sixteen cells, with margins ranging from 0.2 to 3.1 on the reported metric scales.

Table 9: Paired per-seed statistics for the component ablation (Table 3). $\Delta_{\text{host}}$ and $\Delta_{\text{best}}$ are the seed-mean margins of the full objective over the host and over the better single component, respectively, with paired bootstrap 95% CIs and Wilcoxon $p$-values.

| Host | Cell | $\Delta_{\text{host}}$ [CI] | $p$ | $\Delta_{\text{best}}$ [CI] | $p$ |
|---|---|---|---|---|---|
| GRPO | ALFWorld 1.5B | $+12.7$ $[+11.2, +14.6]$ | 0.125 | $+1.5$ $[-0.1, +3.7]$ | 0.250 |
| GRPO | ALFWorld 7B | $+9.2$ $[+7.0, +12.3]$ | 0.125 | $+1.5$ $[+0.1, +3.2]$ | 0.250 |
| GRPO | WebShop 1.5B score | $+9.6$ $[+8.1, +11.8]$ | 0.125 | $+3.1$ $[+2.1, +4.1]$ | 0.125 |
| GRPO | WebShop 1.5B succ. | $+11.1$ $[+9.6, +13.1]$ | 0.125 | $+0.9$ $[-0.6, +2.4]$ | 0.625 |
| GRPO | WebShop 7B score | $+4.4$ $[+3.1, +6.2]$ | 0.125 | $+0.6$ $[-0.4, +1.6]$ | 0.625 |
| GRPO | WebShop 7B succ. | $+4.0$ $[+2.5, +6.1]$ | 0.125 | $+0.9$ $[-0.4, +2.2]$ | 0.625 |
| GRPO | SearchQA 3B | $+3.2$ $[+1.9, +4.9]$ | 0.125 | $+0.4$ $[-0.9, +1.7]$ | 0.500 |
| GRPO | SearchQA 7B | $+3.1$ $[+1.6, +5.2]$ | 0.125 | $+1.7$ $[+0.3, +3.1]$ | 0.125 |
| GiGPO | ALFWorld 1.5B | $+7.3$ $[+6.4, +8.2]$ | 0.125 | $+0.7$ $[-0.2, +1.6]$ | 0.625 |
| GiGPO | ALFWorld 7B | $+4.0$ $[+3.2, +4.8]$ | 0.125 | $+1.3$ $[+0.1, +3.0]$ | 0.250 |
| GiGPO | WebShop 1.5B score | $+6.6$ $[+5.8, +7.6]$ | 0.125 | $+2.0$ $[+1.2, +2.8]$ | 0.125 |
| GiGPO | WebShop 1.5B succ. | $+9.3$ $[+7.3, +12.2]$ | 0.125 | $+3.0$ $[+1.6, +4.4]$ | 0.125 |
| GiGPO | WebShop 7B score | $+3.4$ $[+2.4, +4.8]$ | 0.125 | $+0.2$ $[-0.7, +1.1]$ | 0.875 |
| GiGPO | WebShop 7B succ. | $+4.6$ $[+2.9, +7.1]$ | 0.125 | $+0.9$ $[-0.7, +3.1]$ | 0.625 |
| GiGPO | SearchQA 3B | $+1.2$ $[-1.1, +3.5]$ | 0.625 | $+0.3$ $[-2.0, +2.6]$ | 0.875 |
| GiGPO | SearchQA 7B | $+1.6$ $[-0.1, +4.1]$ | 0.250 | $+0.5$ $[-0.9, +2.5]$ | 1.000 |

## G Sign-Stratified Ablation

BFA preserves the host advantage's sign and reallocates its magnitude. We isolate the contributions of reweighting positive- and negative-advantage trajectories and compare them with inverse-lift weighting. All variants keep the TMN reducer, the same conservation rule (Eq. 22), the same numerical safeguards (Appendix H), and the same four training seeds; they differ only in the raw scores:

- **BFA-pos**: $r_{i,t} = e_{i,t}$ on trajectories with $\hat{A}_i > 0$; uniform weights ($r_{i,t} = 1$) on trajectories with $\hat{A}_i \leq 0$.

- **BFA-neg**: $r_{i,t} = e_{i,t}$ on trajectories with $\hat{A}_i < 0$; $r_{i,t} = 1$ otherwise.
- **Anti-BFA**: $r_{i,t} = 1/e_{i,t}$ everywhere, reversing the within-trajectory ordering of the lift-based weights.

Table 10 shows that both BFA-pos and BFA-neg improve the mean performance over uniform weighting on all three metrics, with larger gains from BFA-pos. Full BFA achieves the highest means, whereas Anti-BFA falls below uniform weighting on all three metrics. Two readings follow. First, the gains concentrate on positive-advantage trajectories (BFA-pos captures the larger share on every metric), so attribution acts primarily by sharpening credit on successes rather than by amplifying penalties on failures—the failure mode that influence-based weighting could in principle exacerbate is not the operative channel. Second, inverting the lifts is worse than no attribution at all, so the lift *values*, and not merely the presence of non-uniform weights, carry the benefit; together with the shuffled-lift and mismatched-feedback controls (Appendix D), this rules out generic reweighting as an explanation.

Table 10: Sign-stratified ablation with a fixed TMN reducer, GRPO, Qwen2.5-1.5B-Instruct, four seeds (mean $\pm$ std). Uniform is the TMN-only row of Table 3.

| | ALFWorld | WebShop | |
|---|---|---|---|
| Variant | Success (%) | Score | Success (%) |
| Uniform (TMN only) | $82.4 \pm 2.9$ | $79.6 \pm 2.5$ | $64.0 \pm 3.2$ |
| BFA (full) | $85.6 \pm 2.8$ | $85.2 \pm 2.0$ | $68.2 \pm 2.8$ |
| BFA-pos | $84.4 \pm 2.6$ | $83.0 \pm 2.3$ | $66.5 \pm 3.0$ |
| BFA-neg | $83.5 \pm 3.1$ | $81.6 \pm 2.7$ | $65.1 \pm 3.3$ |
| Anti-BFA | $81.1 \pm 3.4$ | $78.7 \pm 2.9$ | $62.8 \pm 3.6$ |

## H Reproducibility Details

**Feedback serialization and token masks.** For each decision $(h, a, o')$, the scoring input concatenates the serialized history, the candidate action, and the observed feedback; the target set $\mathcal{V}(o')$ contains exactly the tokens of the observation event, with role markers and any system scaffolding excluded. The same serialization is used for the executed and counterfactual candidates, so both scores in Eq. 7 sum over the same observation tokens.

**Action representation.** At each turn the policy generates a reasoning segment followed by an executable command—one environment command on ALFWorld, one `search[...]` or `click[...]` action on WebShop, and one `<search>` or `<answer>` action on SearchQA. The action $a$ used in feedback scoring is the complete generated text of that turn, including the reasoning segment, exactly as serialized for the environment; no post-hoc rewriting, thought removal, or regeneration is applied. Executed and counterfactual actions follow the same serialization rule, so identical candidate strings receive identical scores and yield a lift of one.

**Missing or empty feedback.** If a decision has no valid feedback tokens (an empty $\mathcal{V}(o')$), both scores vanish and $e = 1$. The lift then enters the same trajectory-level normalization as all other decisions.

**Numerical computation.** All feedback log-likelihoods and lifts are accumulated in log space in float64, with log-sum-exp evaluated by max subtraction.

**Behavior distribution $g_{\text{old}}$.** Counterfactual actions are sampled from the frozen rollout policy under the host's own decoding settings and action constraints, so the candidate mixture matches the distribution that generated the executed action. Training rollouts and counterfactual sampling both use temperature $1.0$ and top-p $1.0$ with top-k filtering disabled; each decision draws one independent counterfactual candidate, and duplicate candidates are retained without resampling. Evaluation samples with temperature $0.4$ and top-p $1.0$.

**Training and evaluation configuration.** Table 11 summarizes the learning rates, rollout budgets, interaction limits, and evaluation sizes. All variants are optimized for a single PPO epoch per update with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, optimizer $\epsilon = 10^{-8}$), zero weight decay, global gradient-norm clipping at 1.0, and PPO clipping at $\epsilon = 0.2$ (dual-clip constant 3.0). The reference-policy KL coefficient is $0.001$ with the initial checkpoint as the reference, and the entropy coefficient is $0.001$. Host regularization terms, including the KL penalty, retain their original reductions.

Table 11: Training and evaluation settings. Each optimizer update uses 15 prompt groups with eight rollouts each, yielding 120 complete trajectories. Search QA evaluation uses 500 tasks per domain per seed, or all available tasks when a domain contains fewer than 500.

| Setting | ALFWorld | WebShop | Search QA |
|---|---|---|---|
| Learning rate | $10^{-6}$ | $10^{-6}$ | $10^{-6}$ |
| Prompt groups per update | 15 | 15 | 15 |
| Rollouts per prompt | 8 | 8 | 8 |
| Complete trajectories per update | 120 | 120 | 120 |
| Optimizer updates | 150 | 150 | 150 |
| Max environment actions | 50 | 15 | 4 searches + final answer |
| Max new tokens per action | 128 | 256 | 1024 |
| Evaluation tasks per seed | 134 | 500 | Up to 500 per domain |

**SearchQA configuration.** SearchQA follows the Search-R1 protocol: training on merged NQ and HotpotQA training sets with the 2018 Wikipedia corpus, an E5 retriever, top-3 passages, at most four search actions, and a balanced held-out subset used only for monitoring.

**Seeds.** Main experiments use four training seeds; the supplementary normalization and weighting studies (Appendices C and D) use two further seeds, disjoint from the main four.

**Diagnostic interval construction.** The reference-advantage intervals in Appendix E are two-sided bootstrap percentile intervals over the $M_Q{=}M_V{=}16$ continuation returns at the 95% level; a decision is resolved when its interval excludes zero and its host advantage is nonzero.

# I External Baselines on Search-Augmented QA

Table 12 reports the external search-RL baselines; the matched-pipeline variants (GRPO, GiGPO, EnvRL, and BATON on both hosts) are reported per domain in Table 2. The external baselines use different retrievers, training data, and protocols, and are listed for reference.

Table 12: External search-RL baselines on search-augmented QA: exact match (%) on seven domains (NQ, TriviaQA, PopQA single-hop; HotpotQA, 2WikiMultihopQA, MuSiQue, Bamboogle multi-hop) and their unweighted average.

| Method | Single-Hop QA | | | Multi-Hop QA | | | | Overall |
|---|---|---|---|---|---|---|---|---|
| | NQ | TriviaQA | PopQA | HotpotQA | 2Wiki | MuSiQue | Bamboogle | Avg. |
| *Qwen2.5-3B-Instruct* | | | | | | | | |
| R1-Instruct | $26.8_{\pm 3.8}$ | $53.9_{\pm 6.2}$ | $19.7_{\pm 4.5}$ | $23.9_{\pm 5.7}$ | $29.0_{\pm 4.1}$ | $7.4_{\pm 2.8}$ | $29.1_{\pm 5.4}$ | $27.0_{\pm 3.6}$ |
| Search-R1 | $34.3_{\pm 2.9}$ | $54.2_{\pm 5.8}$ | $38.0_{\pm 4.7}$ | $32.2_{\pm 6.3}$ | $32.1_{\pm 4.9}$ | $10.1_{\pm 3.6}$ | $26.6_{\pm 5.1}$ | $32.5_{\pm 3.8}$ |
| ZeroSearch | $41.1_{\pm 2.6}$ | $57.6_{\pm 4.9}$ | $44.6_{\pm 5.6}$ | $27.6_{\pm 5.2}$ | $29.8_{\pm 4.3}$ | $9.9_{\pm 3.9}$ | $11.3_{\pm 2.7}$ | $31.7_{\pm 3.4}$ |
| StepSearch | – | – | – | $34.4_{\pm 5.8}$ | $32.2_{\pm 4.6}$ | $17.6_{\pm 4.2}$ | $34.2_{\pm 6.1}$ | – |
| *Qwen2.5-7B-Instruct* | | | | | | | | |
| R1-Instruct | $20.8_{\pm 4.2}$ | $45.1_{\pm 7.1}$ | $16.9_{\pm 3.8}$ | $21.0_{\pm 5.4}$ | $27.3_{\pm 4.7}$ | $6.2_{\pm 2.5}$ | $19.0_{\pm 4.9}$ | $22.3_{\pm 3.9}$ |
| Search-R1 | $39.5_{\pm 3.6}$ | $60.8_{\pm 5.4}$ | $39.9_{\pm 4.2}$ | $36.8_{\pm 6.8}$ | $40.3_{\pm 5.9}$ | $14.4_{\pm 4.1}$ | $37.0_{\pm 5.3}$ | $38.4_{\pm 3.7}$ |
| ZeroSearch | $43.8_{\pm 2.9}$ | $61.6_{\pm 6.2}$ | $51.7_{\pm 5.7}$ | $34.4_{\pm 4.8}$ | $35.4_{\pm 4.4}$ | $18.2_{\pm 5.6}$ | $27.6_{\pm 4.6}$ | $39.0_{\pm 3.5}$ |
| StepSearch | – | – | – | $38.8_{\pm 5.1}$ | $36.4_{\pm 4.7}$ | $22.8_{\pm 5.3}$ | $39.8_{\pm 6.5}$ | – |