# VETTA: COORDINATING TURN- AND TOKEN-LEVEL CREDIT ASSIGNMENT FOR MULTI-TURN LLM AGENTS

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

## ABSTRACT

Multi-turn LLM agents often receive sparse task feedback across several interactions, while generating each response token by token. This creates two related credit-assignment questions: which responses helped achieve the outcome, and which generation decisions mattered within each response? Existing methods typically focus on only one level: turn-level methods evaluate complete responses but do not distinguish the decisions within them; token-level methods can propagate feedback across turns but do not explicitly model credit for each response. These complementary limitations motivate learning credit at both levels and coordinating it in a single policy update. We introduce VETTA, a credit assignment method that jointly learns turn- and token-level values through separate heads on a shared lightweight critic. VETTA computes advantages along both temporal sequences and combines each turn advantage with a within-response-centered token residual for PPO updates. Furthermore, to reduce value-learning cost, the critic retains only early Transformer blocks from the pretrained checkpoint used to initialize the actor. On two challenging agent benchmarks, ALFWorld and WebShop, VETTA improves success rates over PPO by 37.5% and 22.3%, respectively, with Qwen2.5-1.5B-Instruct and achieves success rates of 95.5% and 76.0%, respectively, with Qwen2.5-7B-Instruct. Critic-depth comparisons further show strong task performance with substantially lower critic-side computation. These results suggest that a compact shared critic can coordinate turn- and token-level credit to improve agent performance while keeping value estimation efficient. Code is available at https://anonymous.4open.science/r/VETTA-428E.

## 1 INTRODUCTION

Large language models (LLMs) have demonstrated capabilities in reasoning, tool use, and interactive decision-making, enabling agents to pursue goals through interaction with external environments (Yao et al., 2022b; Schick et al., 2023). Reinforcement learning (RL) has further improved these capabilities by optimizing LLM policies from task feedback (Shao et al., 2024; DeepSeek-AI et al., 2025; Jin et al., 2025; Feng et al., 2025). Many agentic tasks, including operating applications, navigating shopping websites, and carrying out household instructions (Trivedi et al., 2024; Yao et al., 2022a; Shridhar et al., 2021), require repeated action–observation rounds, while rewards are often sparse or delayed (Zhou et al., 2024; Peng et al., 2026). Learning from such feedback requires distinguishing the contributions of individual decisions across an extended interaction (Feng et al., 2025; Cheng et al., 2026).

However, commonly used RL approaches face challenges at different granularities: a direct multi-turn application of outcome-based GRPO (Shao et al., 2024; Feng et al., 2025) assigns the same trajectory-level relative advantage to every turn, whereas token-level PPO (Schulman et al., 2017; Li et al., 2025; Hou et al., 2026) must estimate values for generated-token decisions from delayed feedback across successive interactions.

Recent work has therefore introduced turn-level credit into both optimization paradigms, treating each complete response action as a unit for advantage estimation. Critic-free methods estimate action-level advantages from sampled-rollout returns (Feng et al., 2025; Zhu et al., 2026; He et al.,



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.

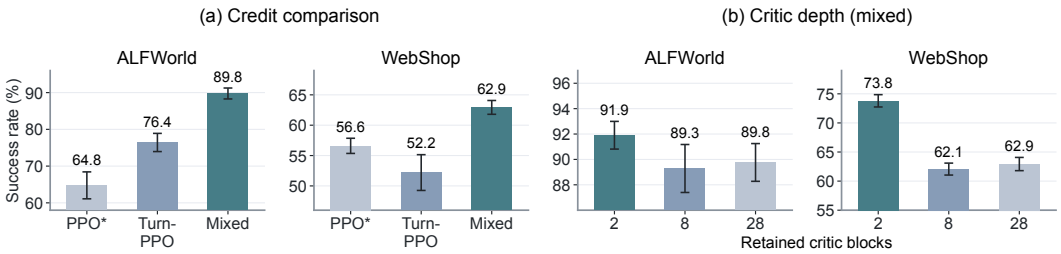


Figure 1: Success rates on ALFWorld and WebShop: (a) full-depth PPO*, Turn-PPO, and mixed configurations; (b) the mixed configuration with 2, 8, or 28 critic blocks. Bars show means and error bars show sample standard deviations across three decoding seeds for each checkpoint.

2026), or from constructed state-transition graphs (Wang et al., 2026; Cheng et al., 2026). Critic-based methods such as Turn-PPO (Li et al., 2025) instead perform value and advantage estimation once per turn, treating the complete response as a whole. These approaches distinguish credit across actions but leave generation decisions within each action undifferentiated, suggesting that turn- and token-level credit could be coordinated to guide learning at both scales.

To examine whether combining turn- and token-level credit improves policy learning, we compare PPO* (token-level credit across turns), Turn-PPO, and a mixed configuration that combines advantages from both granularities (Section 3.2). As shown in Figure 1(a), the mixed configuration achieves higher mean success rates than either alternative on both ALFWorld and WebShop. These results suggest that turn- and token-level credit can provide complementary information for multi-turn policy learning.

Beyond the choice of credit signal, critic-based PPO adds value inference and critic updates to the training workload. We therefore examine whether simplifying the critic can reduce this cost without sacrificing policy performance. With mixed credit, Figure 1(b) compares critics retaining 2, 8, or all 28 Transformer blocks. On both ALFWorld and WebShop, the two-block critic achieves a higher mean success rate than the full-depth critic. These results suggest that the critic's full pretrained depth is not necessary for strong policy performance in these settings.

Building on these observations, we propose VETTA (**V**alue **E**stimation at **T**urn and **T**oken levels for LLM **A**gents), a framework for credit assignment in multi-turn interactions. To model the two decision points, VETTA uses separate turn- and token-value heads: the former estimates the value before a response is generated, while the latter estimates values before its generated tokens. Both heads share a critic backbone but learn from their respective return targets. We compute generalized advantage estimates along the turn and generated-token trajectories. Direct addition would let both advantages shift the overall credit assigned to a response. Instead, we use the turn advantage as the response-level signal and add a scaled, within-response-centered token residual to distinguish generation decisions. The fused advantage guides token-wise PPO updates. We further design a lightweight shared critic by retaining only the early Transformer blocks from the pretrained checkpoint used to initialize the actor, reducing the computation required to learn both value estimates.

We evaluate VETTA on two multi-turn agent benchmarks, ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022a), using Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct (Qwen Team, 2025). With 1.5B, it achieves success rates of 91.9% on ALFWorld and 73.8% on WebShop; with 7B, it reaches 95.5% and 76.0%. Across both benchmarks and model sizes, VETTA consistently achieves the highest performance among both critic-based and critic-free baselines. Compared with 28 blocks, the two-block critic achieves the higher success rate while reducing median critic-side time by approximately 87% on ALFWorld and 91% on WebShop.

Our contributions are threefold:

- We formulate multi-turn credit assignment around two complementary roles: turn-level advantages provide credit for complete responses across interactions, while token-level advantages distinguish generation decisions within each response.

- We propose VETTA, which learns the two value estimates with separately supervised heads on a lightweight shared critic and combines their advantages through within-response centering before token-wise PPO updates.
- We evaluate VETTA on ALFWorld and WebShop, showing improvements in task success and examining how advantage composition, critic sharing, and backbone depth affect performance and critic-side computation.

## 2 PRELIMINARIES

In this section, we formalize the response- and token-level decision sequences used for value estimation and policy optimization.

### 2.1 MULTI-TURN LLM AGENT INTERACTION

Consider one finite interaction trajectory containing $K$ environment turns. At turn $k \in \{1, \dots, K\}$, the agent receives the context $H_k$, which contains the task instruction, the current observation, and the preceding interaction history included in the model context. Conditioned on $H_k$, the policy $\pi_\theta$, with parameters $\theta$, generates a response $a_k = (y_{k,1}, \dots, y_{k,L_k})$, where $y_{k,j}$ is the $j$-th generated token and $L_k$ is the number of valid generated tokens in that response (the token mask is specified in Appendix C.2). After the response is completed, the environment executes the action specified by $a_k$, returns a scalar reward $r_k$, and provides the next observation $o_{k+1}$. The next context $H_{k+1}$ incorporates that observation. The same interaction can therefore be viewed as a sequence of complete response decisions or as a finer sequence of token-generation decisions.

### 2.2 PPO AND GENERALIZED ADVANTAGE ESTIMATION

For a generic decision sequence $\tau = (s_t, a_t, r_t)_{t=0}^{T-1}$ containing $T$ decisions, let $t$ index the policy decisions, where $s_t$, $a_t$, and $r_t$ denote the decision context, action, and reward at step $t$. With discount factor $\gamma \in [0, 1]$, the standard policy objective is

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[ \sum_{t=0}^{T-1} \gamma^t r_t \right]. \tag{1}$$

A multi-turn interaction admits turn- and token-level sequences for value and advantage estimation: a turn-level decision corresponds to a complete response $a_k$, whereas a token-level decision corresponds to a generated token $y_{k,j}$.

For a policy $\pi$, $Q^\pi(s_t, a_t)$ denotes the expected discounted return after taking action $a_t$ in state $s_t$ and subsequently following $\pi$, while $V^\pi(s_t)$ is the expected return before choosing that action. The advantage, $A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t)$, measures the action's value relative to the policy's average at that state. Generalized advantage estimation (GAE) uses rewards and value predictions along a sampled trajectory to estimate this advantage (Schulman et al., 2015). PPO uses the estimated advantage to guide policy updates, with probability-ratio clipping limiting the incentive for large policy changes (Schulman et al., 2017). VETTA retains token-wise policy optimization and uses the two decision sequences to construct complementary advantage estimates.

## 3 METHOD

As a critic-based method, VETTA follows the PPO pipeline in Figure 2(b): the actor collects multi-turn trajectories, and the critic provides value estimates for advantage computation and policy optimization. Within this pipeline, VETTA uses a shared lightweight critic to estimate values at both turn and token levels (Figure 2(a)), then combines the resulting advantages to guide the actor update (Figure 2(c)). The two value heads are trained with their respective return targets.

### 3.1 DUAL-GRANULARITY VALUE ESTIMATION WITH A LIGHTWEIGHT CRITIC

At each environment turn, VETTA estimates future returns at two decision granularities. Before the response $a_k$ is generated, the turn-level value $V_k^{\text{turn}}$ evaluates the context $H_k$ and provides one

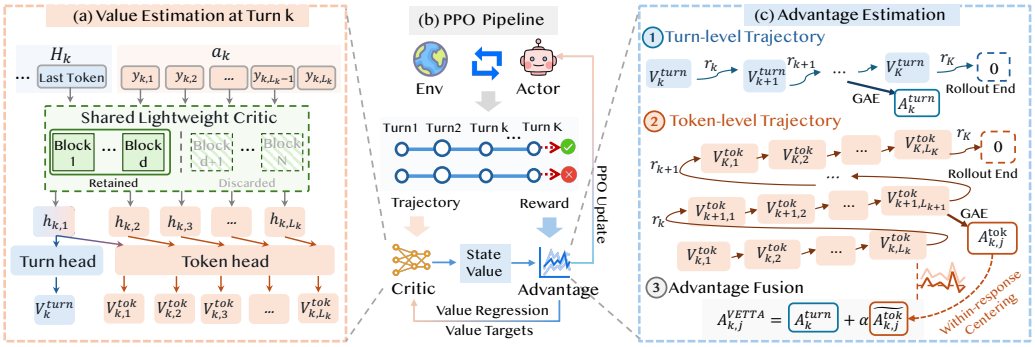


Figure 2: Overview of VETTA. (a) A shared critic retains early pretrained Transformer blocks and uses separate heads for turn- and token-level value estimation. (b) The actor interacts with the environment to collect trajectories and rewards; value estimates support advantage computation, followed by critic regression and PPO policy updates. (c) The collected trajectory is viewed at turn and token levels for GAE; token advantages are centered within each response and combined with its turn advantage. The implicit successor value and advantage are set to zero at rollout end.

value for the complete response decision. Within that response, the token-level value $V^{\text{tok}}_{k,j}$ evaluates the prefix available before generating $y_{k,j}$, providing one value for each valid generated position. Figure 2(a) illustrates how the two sets of values are read from a shared critic with causal attention.

In a single forward pass, the critic processes the context followed by the sampled response $a_k = (y_{k,1}, \ldots, y_{k,L_k})$ using a causal attention mask. We write $y_{k,<j} = (y_{k,1}, \ldots, y_{k,j-1})$ for the response prefix available before position $j$, with $y_{k,<1} = \emptyset$. Let $h_{k,1}$ be the representation at the final context position, and let $h_{k,j}$ for $j > 1$ be the representation at $y_{k,j-1}$. Under causal masking, $h_{k,j}$ therefore depends only on $(H_k, y_{k,<j})$ and is aligned with the decision to generate $y_{k,j}$.

The two output heads map these representations to

$$\begin{aligned} V^{\text{turn}}_k &= V^{\text{turn}}(H_k) = w^\top_{\text{turn}} h_{k,1} + b_{\text{turn}}, \\ V^{\text{tok}}_{k,j} &= V^{\text{tok}}(H_k, y_{k,<j}) = w^\top_{\text{tok}} h_{k,j} + b_{\text{tok}}, \end{aligned} \tag{2}$$

where $w_{\text{tok}}$ and $w_{\text{turn}}$ are head-specific weight vectors and $b_{\text{tok}}$ and $b_{\text{turn}}$ are their biases. The two value heads are newly initialized and trained with separate supervision targets. They have separate output parameters and use representations from the same backbone. Both read $h_{k,1}$, while the token-value head additionally reads the representations preceding subsequent generated tokens.

**Lightweight critic backbone.** To limit critic cost, we construct its backbone from the pretrained checkpoint used to initialize the actor. Of its $N$ Transformer blocks, we retain the token embedding, the first $d \leq N$ blocks, and the final root mean square normalization (RMSNorm), discarding the remaining blocks (Figure 2(a)). All retained critic components remain trainable. This reduces the number of blocks used in critic forward and backward passes without removing either value head. We examine the effect of retained depth on policy performance and critic-side computation in Section 4.3.

### 3.2 Advantage Estimation

VETTA applies standard GAE separately at the turn and token levels. The same scalar reward $r_k$ is represented at both granularities: at the turn level, it is assigned to turn $k$; at the token level, it is placed at the final valid token of response $a_k$, with zero reward at earlier generated positions:

$$r^{\text{turn}}_k = r_k, \qquad r^{\text{tok}}_{k,j} = \begin{cases} 0, & j < L_k, \\ r_k, & j = L_k \end{cases}. \tag{3}$$

The environment-specific components of $r_k$ are summarized in Appendix B.1. For each granularity $b \in \{\text{turn}, \text{tok}\}$, let $i$ denote a turn $k$ or a valid generated-token position $(k, j)$, respectively, and let

$\mathrm{succ}_b(i)$ denote its next position in the corresponding sequence. Figure 2(c) shows turn-level and token-level views of the same trajectory. The turn-level trajectory advances to turn $k+1$; the token-level trajectory advances to the next generated token within the response or, after its final token, to the first generated token of the next response.

Each granularity uses its own discount factor $\gamma_b \in [0,1]$ and GAE trace parameter $\lambda_b \in [0,1]$, which control reward discounting and the decay of future temporal-difference residuals, respectively. Let $V^b_{\mathrm{old}}$ denote the critic's value estimates before the current optimization phase. These estimates are used to compute the temporal-difference residual $\delta^b_i$ and advantage estimate $A^b_i$:

$$\begin{aligned} \delta^b_i &= r^b_i + \gamma_b V^b_{\mathrm{old}}(\mathrm{succ}_b(i)) - V^b_{\mathrm{old}}(i), \\ A^b_i &= \delta^b_i + \gamma_b \lambda_b A^b_{\mathrm{succ}_b(i)}. \end{aligned} \tag{4}$$

The observation between two responses enters the next context but is not a policy position, so the token-level recursion skips its tokens. At the final collected position of each rollout, we set the value and advantage of the implicit next position to zero. This position is not an additional critic-supervision target.

Directly adding turn- and token-level advantages would combine their response-wide offsets as well as the token-level variation. We instead use the turn-level advantage to determine the response-wide level of credit and the token-level residual to adjust relative credit within the response. To prevent the token component from shifting the response-wide mean, we center its advantages over the valid tokens of each response:

$$\overline{A^{\mathrm{tok}}}_k = \frac{1}{L_k}\sum_{j=1}^{L_k} A^{\mathrm{tok}}_{k,j}, \qquad \widetilde{A^{\mathrm{tok}}}_{k,j} = A^{\mathrm{tok}}_{k,j} - \overline{A^{\mathrm{tok}}}_k. \tag{5}$$

With $\alpha$ controlling the residual's relative scale, the final VETTA advantage is

$$A^{\mathrm{VETTA}}_{k,j} = A^{\mathrm{turn}}_k + \alpha \widetilde{A^{\mathrm{tok}}}_{k,j}. \tag{6}$$

The centered term $\widetilde{A^{\mathrm{tok}}}_{k,j}$ is the within-response residual. Appendix C.4 shows that, among vectors with response-wise mean $A^{\mathrm{turn}}_k$, the fused advantage vector is closest in squared Euclidean distance to the scaled token advantages $\alpha(A^{\mathrm{tok}}_{k,j})_{j=1}^{L_k}$.

### 3.3 Critic and Actor Updates

At each valid position $i$ of granularity $b$, the corresponding value head is trained with the return target

$$\widehat{G}^b(i) = \mathrm{sg}\big(V^b_{\mathrm{old}}(i) + A^b_i\big). \tag{7}$$

Here sg stops gradients through the target. Advantages and return targets are computed before optimization and held fixed throughout the corresponding updates. Each value target uses its own granularity's advantage, before centering or fusion. The critic uses standard PPO clipped value regression, averaging turn and token losses over their respective valid positions before summing them. The explicit loss is given in Appendix C.3.

The actor retains token-wise PPO ratios and clipping. Let $\theta_{\mathrm{old}}$ denote the policy parameters used to collect the trajectory. For each valid generated token $y_{k,j}$, let $\rho_{k,j}(\theta)$ be the ratio of its probability under $\pi_\theta$ to that under $\pi_{\theta_{\mathrm{old}}}$, with both policies conditioned on $(H_k, y_{k,<j})$. The surrogate loss uses the composed advantage:

$$\mathcal{L}_{\mathrm{PPO}} = -\mathbb{E}_{k,j}\left[\min\big(\rho_{k,j}A^{\mathrm{VETTA}}_{k,j}, \mathrm{clip}(\rho_{k,j}, 1-\epsilon, 1+\epsilon)A^{\mathrm{VETTA}}_{k,j}\big)\right]. \tag{8}$$

Here $\epsilon > 0$ is the PPO ratio-clipping radius and the expectation is over valid generated positions in the rollout batch. The actor implementation's dual-clipping and KL settings are described in Appendix C.5.

## 4 Experiments

We evaluate VETTA across two multi-turn agent settings and organize the experiments around four research questions: **RQ1:** How does VETTA compare with prompting and RL methods? **RQ2:**

Table 1: Performance on ALFWorld and WebShop. Results for RL methods are reported over three random seeds. For ALFWorld, we report success rates (%) for six task categories and overall; for WebShop, task score and success rate (%). † indicates results reported in GiGPO (Feng et al., 2025). The best values are in bold, and the second-best values are underlined.

| Type | Method | ALFWorld | | | | | | | WebShop | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Pick | Look | Clean | Heat | Cool | Pick2 | All | Score | Succ. |
| *Qwen2.5-1.5B-Instruct* | | | | | | | | | | |
| Prompting | Qwen2.5† | 5.9 | 5.5 | 3.3 | 9.7 | 4.2 | 0.0 | 4.1 | 23.1 | 5.2 |
| Prompting | ReAct† | 17.4 | 20.5 | 15.7 | 6.2 | 7.7 | 2.0 | 12.8 | 40.1 | 11.3 |
| Prompting | Reflexion† | 35.3 | 22.2 | 21.7 | 13.6 | 19.4 | 3.7 | 21.8 | 55.8 | 21.9 |
| Critic-free | RLOO† | $88.3_{\pm 3.0}$ | $52.8_{\pm 8.6}$ | $71.0_{\pm 5.9}$ | $62.8_{\pm 8.7}$ | $66.4_{\pm 5.5}$ | $56.9_{\pm 4.7}$ | $69.7_{\pm 2.5}$ | $73.9_{\pm 5.6}$ | $52.1_{\pm 6.7}$ |
| Critic-free | GRPO† | $85.3_{\pm 1.5}$ | $53.7_{\pm 8.0}$ | $84.5_{\pm 6.8}$ | $78.2_{\pm 7.9}$ | $59.7_{\pm 5.0}$ | $53.5_{\pm 5.6}$ | $72.8_{\pm 3.6}$ | $75.8_{\pm 3.5}$ | $56.8_{\pm 3.8}$ |
| Critic-free | GiGPO† | $\underline{96.0}_{\pm 1.4}$ | $\underline{76.5}_{\pm 3.9}$ | $\underline{91.8}_{\pm 5.5}$ | $\mathbf{91.3}_{\pm 6.3}$ | $71.7_{\pm 8.4}$ | $\underline{79.5}_{\pm 7.7}$ | $\underline{86.1}_{\pm 4.7}$ | $\underline{83.5}_{\pm 1.8}$ | $\underline{67.4}_{\pm 4.5}$ |
| Critic-based | PPO† | $64.8_{\pm 3.5}$ | $40.5_{\pm 6.9}$ | $57.1_{\pm 4.9}$ | $60.6_{\pm 6.6}$ | $46.4_{\pm 4.0}$ | $47.4_{\pm 1.9}$ | $54.4_{\pm 3.1}$ | $73.8_{\pm 3.0}$ | $51.5_{\pm 2.9}$ |
| Critic-based | PPO* | $75.2_{\pm 4.4}$ | $48.7_{\pm 4.4}$ | $72.0_{\pm 6.7}$ | $72.5_{\pm 6.8}$ | $72.7_{\pm 7.9}$ | $33.3_{\pm 2.6}$ | $64.8_{\pm 3.7}$ | $70.8_{\pm 1.2}$ | $56.6_{\pm 1.2}$ |
| Critic-based | Turn-PPO | $86.7_{\pm 4.4}$ | $71.8_{\pm 4.4}$ | $70.4_{\pm 6.4}$ | $66.7_{\pm 9.6}$ | $\underline{81.3}_{\pm 6.1}$ | $72.2_{\pm 6.4}$ | $76.4_{\pm 2.5}$ | $69.2_{\pm 4.2}$ | $52.2_{\pm 2.9}$ |
| Critic-based | **VETTA** | $\mathbf{97.1}_{\pm 2.9}$ | $\mathbf{79.5}_{\pm 4.4}$ | $\mathbf{95.1}_{\pm 4.3}$ | $\underline{85.4}_{\pm 9.5}$ | $\mathbf{88.0}_{\pm 4.0}$ | $\mathbf{95.8}_{\pm 0.0}$ | $\mathbf{91.9}_{\pm 1.1}$ | $\mathbf{83.9}_{\pm 0.6}$ | $\mathbf{73.8}_{\pm 1.1}$ |
| *Qwen2.5-7B-Instruct* | | | | | | | | | | |
| Prompting | Qwen2.5† | 33.4 | 21.6 | 19.3 | 6.9 | 2.8 | 3.2 | 14.8 | 26.4 | 7.8 |
| Prompting | ReAct† | 48.5 | 35.4 | 34.3 | 13.2 | 18.2 | 17.6 | 31.2 | 46.2 | 19.5 |
| Prompting | Reflexion† | 62.0 | 41.6 | 44.9 | 30.9 | 36.3 | 23.8 | 42.7 | 58.1 | 28.8 |
| Critic-free | RLOO† | $87.6_{\pm 4.3}$ | $78.2_{\pm 8.3}$ | $87.3_{\pm 5.8}$ | $81.3_{\pm 7.6}$ | $71.9_{\pm 5.2}$ | $48.9_{\pm 8.4}$ | $75.5_{\pm 4.6}$ | $80.3_{\pm 3.2}$ | $65.7_{\pm 4.0}$ |
| Critic-free | GRPO† | $90.8_{\pm 5.1}$ | $66.1_{\pm 6.7}$ | $89.3_{\pm 5.4}$ | $74.7_{\pm 6.9}$ | $72.5_{\pm 5.4}$ | $64.7_{\pm 7.3}$ | $77.6_{\pm 5.2}$ | $79.3_{\pm 2.8}$ | $66.1_{\pm 3.7}$ |
| Critic-free | GiGPO† | $91.8_{\pm 5.4}$ | $\mathbf{88.6}_{\pm 6.3}$ | $\underline{95.9}_{\pm 3.2}$ | $\underline{90.2}_{\pm 2.6}$ | $\underline{86.5}_{\pm 5.5}$ | $85.2_{\pm 7.5}$ | $90.2_{\pm 2.3}$ | $\underline{86.2}_{\pm 2.6}$ | $\underline{75.2}_{\pm 3.8}$ |
| Critic-based | PPO† | $92.3_{\pm 4.0}$ | $64.0_{\pm 8.4}$ | $92.5_{\pm 2.4}$ | $89.5_{\pm 7.0}$ | $80.3_{\pm 2.0}$ | $68.8_{\pm 8.3}$ | $80.4_{\pm 2.7}$ | $81.4_{\pm 3.1}$ | $68.7_{\pm 5.1}$ |
| Critic-based | PPO* | $92.1_{\pm 1.0}$ | $85.7_{\pm 7.6}$ | $77.4_{\pm 8.3}$ | $72.2_{\pm 2.4}$ | $55.6_{\pm 9.6}$ | $57.6_{\pm 11.1}$ | $72.6_{\pm 2.1}$ | $84.5_{\pm 0.1}$ | $74.8_{\pm 0.4}$ |
| Critic-based | Turn-PPO | $\mathbf{100.0}_{\pm 0.0}$ | $82.9_{\pm 8.9}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $84.5_{\pm 2.3}$ | $\underline{92.4}_{\pm 4.2}$ | $\underline{94.3}_{\pm 1.5}$ | $75.3_{\pm 0.8}$ | $54.6_{\pm 1.9}$ |
| Critic-based | **VETTA** | $\underline{93.6}_{\pm 2.6}$ | $\underline{86.7}_{\pm 5.8}$ | $\mathbf{100.0}_{\pm 0.0}$ | $90.0_{\pm 0.0}$ | $\mathbf{93.9}_{\pm 3.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{95.5}_{\pm 0.4}$ | $\mathbf{86.5}_{\pm 0.7}$ | $\mathbf{76.0}_{\pm 0.6}$ |

How does composing turn- and token-level credit affect performance? **RQ3:** How do backbone sharing and value-head structure affect performance? **RQ4:** How does critic backbone depth affect performance and critic-side computation?

### 4.1 Experiment Setup

**Benchmarks.** We evaluate VETTA on ALFWorld and WebShop. ALFWorld (Shridhar et al., 2021) is a text-based embodied benchmark comprising six categories of household tasks; we report success rates for each category and overall. WebShop (Yao et al., 2022a) asks agents to search and navigate a simulated shopping website, and we report task score and success rate. Evaluation-set sizes and aggregation details appear in Appendix B.3.

**Baselines.** We first compare VETTA with prompting methods: Qwen2.5-Instruct without task-specific RL training, ReAct (Yao et al., 2022b), which interleaves reasoning and actions, and Reflexion (Shinn et al., 2023), which uses verbal feedback from previous attempts. These methods do not update the policy with RL. We then group the RL baselines by whether they train a parametric critic. The critic-free baselines RLOO and GRPO (Shao et al., 2024) estimate relative credit from groups of sampled trajectories. GiGPO (Feng et al., 2025) further combines episode-relative credit with action-level comparisons at repeated environment states. Among critic-based baselines, PPO (Schulman et al., 2017; Feng et al., 2025) copies the episode reward to the final generated token of each response and computes GAE independently within each response. To better compare PPO with VETTA, we implement PPO*, which instead computes GAE across turns over generated tokens, skipping observation tokens. Both PPO variants use token-level values, probability ratios, and clipping. Turn-PPO (Li et al., 2025) instead estimates turn-level values and advantages and applies PPO ratios and clipping to complete responses.

**Implementation details.** We use Qwen2.5-1.5B/7B-Instruct (Qwen Team, 2025) as our base models for ALFWorld and WebShop. Within each model-size comparison, methods use the same base actor model. Critic-based baselines retain the full backbone depth, whereas VETTA retains two critic blocks for ALFWorld and WebShop at both model sizes. We set $\gamma_{\text{tok}} = \lambda_{\text{tok}} = 1$ and

$\gamma_{\text{turn}} = \lambda_{\text{turn}} = 0.95$ in both settings. Unless otherwise specified, the token-residual scale is $\alpha = 1$ for ALFWorld and $\alpha = 3$ for WebShop. Training runs for 150 updates in both settings. Additional optimization settings are given in Table 3, with evaluation details in Appendix B.3.

### 4.2 Experimental Results (RQ1)

Table 1 shows that VETTA performs strongly on both ALFWorld and WebShop at 1.5B and 7B. ReAct and Reflexion improve over the Qwen2.5-Instruct prompting baseline, but their overall success rates remain well below those of the RL-trained methods. This establishes the value of learning from task feedback in both interactive settings.

Among critic-free RL methods, RLOO and GRPO substantially improve over prompting, although their advantage over PPO varies across benchmarks and model sizes. GiGPO further assigns action-level relative credit using repeated environment states and achieves the highest overall success rates among the critic-free baselines on both tasks.

The critic-based baselines reveal different strengths across the two benchmarks. PPO* propagates token-level GAE across turns while retaining token-wise PPO updates, whereas Turn-PPO estimates advantages for complete responses. At 1.5B, Turn-PPO is stronger on ALFWorld, while PPO* performs slightly better on WebShop; both remain below VETTA. This contrast becomes more pronounced at 7B: Turn-PPO reaches 94.3% overall success on ALFWorld, whereas PPO* reaches 74.8% success on WebShop. Neither single-granularity baseline maintains the same strength across both benchmarks.

VETTA uses turn-level advantages to represent progress across interactions and token-level residuals to distinguish decisions within each response. At 1.5B, it reaches 91.9% success on ALFWorld and 73.8% on WebShop, exceeding the strongest alternative in each benchmark by 5.8 and 6.4 percentage points, respectively. At 7B, it retains the highest mean overall success rate among the listed methods on both benchmarks, reaching 95.5% and 76.0%; it also achieves the highest WebShop task score at both model sizes. The differing strengths of PPO* and Turn-PPO make the combined signal a useful design choice to investigate. The following analyses examine advantage composition, critic sharing, and backbone depth.

### 4.3 Further Analysis

We further examine how credit composition, critic sharing, and backbone depth affect VETTA's performance.

**Credit composition (RQ2).** To assess the contribution of each credit level and within-response centering, we compare actor-credit rules on WebShop using Qwen2.5-1.5B-Instruct.

We keep the shared two-block critic and the token-wise PPO ratios and clipping in Eq. (8) fixed. Rows C1–C4 train both value heads and change only the advantage given to the actor. Direct addition (C3) and residual composition (C4) use $\alpha = 3$, so their comparison isolates within-response centering. In C5 and C6, the actor uses one credit branch; the corresponding value loss trains its output head and the critic's Transformer backbone.

Table 2: Analysis of credit composition on WebShop.

| ID | Actor credit | Critic supervision | Success rate (%) | Task score |
|---|---|---|---|---|
| C1 | $A^{\text{tok}}$ | Turn + token | 67.4 | **0.849** |
| C2 | $A^{\text{turn}}$ | Turn + token | 59.8 | 0.745 |
| C3 | $A^{\text{turn}} + \alpha A^{\text{tok}}$ | Turn + token | 55.0 | 0.790 |
| C4 | $A^{\text{turn}} + \alpha \widetilde{A^{\text{tok}}}$ | Turn + token | **72.8** | 0.836 |
| C5 | $A^{\text{tok}}$ | Token | 44.8 | 0.568 |
| C6 | $A^{\text{turn}}$ | Turn | 68.0 | 0.845 |

Among the matched dual-head runs in Table 2, the residual combination C4 achieves the highest success rate (72.8%). Its gain over direct addition C3 (55.0%) supports within-response centering, while its gains over C1 and C2 support using both credit levels for exact task completion. C1 nevertheless achieves the highest task score (0.849), showing that token credit alone can produce strong graded outcomes. Its lower success rate shows that this advantage does not consistently translate into exact completions, which trigger WebShop's positive training reward (Appendix B.1).

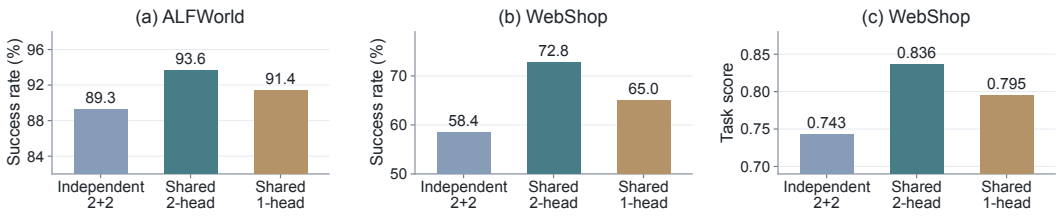


Figure 3: Critic structure: (a) ALFWorld success rate, (b) WebShop success rate, and (c) WebShop task score.

The single-branch control C5 reaches 44.8% success, while C6 reaches 68.0% success and a task score of 0.845, slightly above C4's 0.836. Yet C4 has a higher success rate than both controls. Together, these results favor combining the two granularities for exact task completion; the higher graded scores of C1 and C6 also suggest room to improve how the signals are combined for WebShop's graded outcome.

**Critic sharing and value heads (RQ3).** The critic in VETTA makes two structural choices: the turn and token values share a backbone but have separate output heads. We examine both choices on ALFWorld and WebShop using Qwen2.5-1.5B-Instruct. To assess backbone sharing, we compare two independent two-block critics, each with its own value head, against one two-block backbone with separate turn- and token-value heads. To assess the heads, we also include a shared-backbone variant with one output head trained on both value targets. At the first generated-token position of each turn, the same prediction serves as both the turn value and the first token value; later positions provide the remaining token values. All variants retain turn- and token-level credit and the residual composition rule.

Figure 3 shows that the shared two-head critic has the highest success rate on both benchmarks: 93.6% on ALFWorld and 72.8% on WebShop. The independent critics reach 89.3% and 58.4%, while the shared one-head critic reaches 91.4% and 65.0%, respectively. WebShop task scores follow the same ordering. These single-run results favor sharing the backbone while retaining separate value heads. Because the configurations also differ in parameter count and optimization, they do not isolate a representational benefit of sharing alone.

**Critic depth and computation (RQ4).** We ask how much critic computation can be saved by retaining only the early Transformer blocks. On each benchmark, we compare 2-, 8-, and 28-block critics while keeping both value heads and the actor-credit rule unchanged.

Figure 4 shows the savings in critic-side computation, measured over 112 training steps per run (Appendix B.4). Compared with 28 blocks, the two-block critic records 19.7 rather than 152.0 seconds of median value-inference-plus-update time on ALFWorld, approximately 87% less. On WebShop, the corresponding times are 7.9 and 92.2 seconds, approximately 91% less. Its median share of training-step time likewise falls from 25.6% to 3.8% on ALFWorld and from 27.8% to 4.3% on WebShop; the eight-block runs lie between these endpoints.

The two-block model saves computation by removing 26 Transformer blocks from its trainable backbone. Despite this reduction, Figure 1(b) shows the highest mean success rate with two blocks on both benchmarks. The eight-block results further show that retaining more layers does not necessarily improve success. Together, these results suggest that retaining only a few early

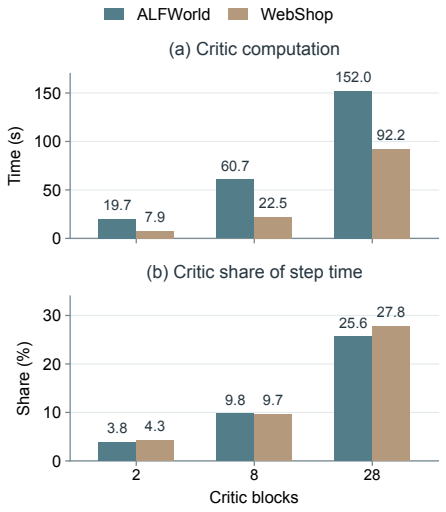


Figure 4: Critic computation by backbone depth on ALFWorld and WebShop.

Transformer blocks may reduce critic-side computation while improving task success in multi-turn agentic tasks.

## 5 Related Work

In this section, we position VETTA relative to prior work on turn-level credit, token-level and joint credit, and critic design.

**Turn-Level Credit Assignment.** Turn-level methods assign credit to complete textual actions within a multi-turn trajectory. Critic-free approaches derive action-level signals from sampled returns using state- or history-based groups (Feng et al., 2025; He et al., 2026), or from progress toward successful states in a constructed transition graph (Cheng et al., 2026). Some explicitly estimate values without training a parametric critic: GAGPO constructs grouped return proxies for temporal-difference (TD) and generalized advantage estimation (GAE) (Zhu et al., 2026), while Gated-BEPO estimates node values through Bellman backups on empirical rollout graphs (Yan et al., 2026). Learned value functions provide another route. ArCHer learns utterance-level values to guide a lower-level token policy (Zhou et al., 2024), and Turn-PPO learns turn-level values and applies PPO ratios and clipping to complete responses (Li et al., 2025). These approaches distinguish the contributions of different actions; our focus is on combining such action-level credit with variation among the token-generation decisions within each action.

**Token-Level and Joint Credit Assignment.** Other work differentiates credit within a response or jointly models multiple granularities. SAO connects token-level GAE across adjacent responses while skipping observation tokens in the recursion (Hou et al., 2026). FACTOR allocates action credit to executable tokens with nonnegative mean-one weights (Ma et al., 2026), while SERL uses post-action teacher feedback to adjust token updates selectively (Li et al., 2026). The closest to our joint estimation is HyGAE, which linearly combines turn- and token-level advantages and trains a unified value function with mixed return targets (Zhang et al., 2026). VETTA instead retains separate supervision for the two value heads and adds a centered token-level residual to the turn-level advantage. Centering makes the residual average to zero within each response, so the turn advantage determines the mean of the composed signal, while token-level variation adjusts credit within that response. This separates the value-learning targets and the roles of the two advantage components.

**Critic Architecture and Efficiency.** Value-model design also determines how representations are shared and how much computation value learning requires. HiPER shares a critic backbone across value heads for subgoal planning and action execution (Peng et al., 2026); VETTA uses shared representations for complete responses and individual token-generation decisions. To reduce critic cost, AsyPPO uses ensembles of smaller critics (Liu et al., 2025), while POISE fits lightweight value probes to the actor's internal representations (Choi et al., 2026). VETTA constructs its shared critic backbone from the early Transformer blocks of the pretrained checkpoint used to initialize the actor, reducing backbone depth while retaining both value-estimation objectives. Complementary work such as VC-PPO and VAPO addresses value-learning failures and optimization stability (Yuan et al., 2025; Yue et al., 2025).

## 6 Conclusion

We introduce VETTA, a critic-based framework for multi-turn LLM agents that estimates turn- and token-level values with a shared lightweight backbone and separate value heads. It combines the turn-level advantage with a centered token-level residual to guide policy optimization. On ALFWorld and WebShop, VETTA achieves higher success rates than SOTA RL methods across the evaluated model sizes. We also find that retaining only early Transformer blocks preserves strong performance while substantially reducing critic-update time. Together, these results open a broader design space for multi-turn agent RL: critics can vary in both how much they compute and the temporal scales at which they assign credit. In future work, designing value models whose granularity reflects each task's interaction and feedback structure remains a promising direction.

## AI USE STATEMENT

Generative AI tools were used to draft and polish manuscript text, search for and summarize related literature, provide feedback on experimental comparisons and the interpretation of results, and assist with debugging and reproduction checks. They also assisted with preparing figures, tables, and manuscript scripts. AI tools did not operate the experiments or implement the proposed training method. The authors take responsibility for the final manuscript, including AI-assisted text, analyses, and artifacts.

## REPRODUCIBILITY STATEMENT

Section 3 describes the value estimates, advantage computation, and optimization objectives. Section 4 and Appendix B.2 provide the training and evaluation configurations. We will provide anonymous code and supporting materials, including the configurations, to facilitate reproduction of the experiments.

## REFERENCES


Xin Cheng, Shuo He, Lang Feng, HaiYang Xu, Ming Yan, Lei Feng, and Bo An. Beyond trajectory-level attribution: Graph-based credit assignment for agentic reinforcement learning. *arXiv preprint arXiv:2605.26684*, 2026. URL https://arxiv.org/abs/2605.26684v2.

Yunho Choi, Jongwon Lim, Woojin Ahn, Minjae Oh, Jeonghoon Shim, and Yohan Jo. Your language model is its own critic: Reinforcement learning with value estimation from actor's internal states. *arXiv preprint arXiv:2605.07579*, 2026. URL https://arxiv.org/abs/2605.07579v2.

DeepSeek-AI, Daya Guo, Dejian Yang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

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/2505.10978v3.

Shuo He, Lang Feng, Qi Wei, Xin Cheng, Lei Feng, and Bo An. Hierarchy-of-groups policy optimization for long-horizon agentic tasks. In *Proceedings of the International Conference on Learning Representations*, 2026.

Zhenyu Hou, Yujiang Li, Jie Tang, and Yuxiao Dong. Single-rollout asynchronous optimization for agentic reinforcement learning. *arXiv preprint arXiv:2607.07508*, 2026. URL https://arxiv.org/abs/2607.07508v1.

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. *arXiv preprint arXiv:2503.09516*, 2025. URL https://arxiv.org/abs/2503.09516v5.

Junbo Li, Peng Zhou, Rui Meng, Meet P. Vadera, Lihong Li, and Yang Li. Turn-PPO: Turn-level advantage estimation with PPO for improved multi-turn RL in agentic LLMs. *arXiv preprint arXiv:2512.17008*, 2025. URL https://arxiv.org/abs/2512.17008v2.

Xiaozhe Li, Tianyi Lyu, Yang Li, Yichuan Ma, Peiji Li, Linyang Li, Qipeng Guo, Dahua Lin, and Kai Chen. What and when to distill: Selective hindsight distillation for multi-turn agents. *arXiv preprint arXiv:2605.19447*, 2026. URL https://arxiv.org/abs/2605.19447v1.

Jiashun Liu, Johan Obando-Ceron, Han Lu, Yancheng He, Weixun Wang, Wenbo Su, Bo Zheng, Pablo Samuel Castro, Aaron Courville, and Ling Pan. Asymmetric proximal policy optimization: mini-critics boost LLM reasoning. *arXiv preprint arXiv:2510.01656*, 2025. URL https://arxiv.org/abs/2510.01656v3.

Lichao Ma, Yang Sun, Shuaitao Zhao, Yangyi Fang, Cong Qin, Xiaoliang Fu, Yuhang Tian, Yuchen Wei, Junbo Zhu, Yang Wei, Lu Pan, and Jiaye Lin. How much, then where: Credit-conserving action-to-token allocation for multi-turn agent reinforcement learning. *arXiv preprint arXiv:2608.07118*, 2026. URL https://arxiv.org/abs/2608.07118v1.

Jiangweizhi Peng, Yuanxin Liu, Ruida Zhou, Charles Fleming, Zhaoran Wang, Alfredo Garcia, and Mingyi Hong. HiPER: Hierarchical reinforcement learning with explicit credit assignment for large language model agents. *arXiv preprint arXiv:2602.16165*, 2026. URL https://arxiv.org/abs/2602.16165v2.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115v2.

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. *arXiv preprint arXiv:2302.04761*, 2023. URL https://arxiv.org/abs/2302.04761.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015. URL https://arxiv.org/abs/1506.02438v6.

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/1707.06347v2.

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/2402.03300v3.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 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/2010.03768v2.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. *arXiv preprint arXiv:2407.18901*, 2024. URL https://arxiv.org/abs/2407.18901.

Yunan Wang, Minghui Song, Zihan Zhang, Shaohan Huang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Group-graph policy optimization for long-horizon agentic reinforcement learning. *arXiv preprint arXiv:2606.22995*, 2026. URL https://arxiv.org/abs/2606.22995.

Hongxi Yan, Ziyue Huang, Shichao Fan, and Qingjie Liu. Gated-BEPO: Confidence-gated bellman credit assignment for large language model agents. *arXiv preprint arXiv:2608.06861*, 2026. URL https://arxiv.org/abs/2608.06861v1.

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*, 2022a. URL https://arxiv.org/abs/2207.01206v4.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022b. URL https://arxiv.org/abs/2210.03629.

Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. What's behind PPO's collapse in long-CoT? value optimization holds the secret. *arXiv preprint arXiv:2503.01491*, 2025. URL https://arxiv.org/abs/2503.01491v1.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, Tiantian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Ru Zhang, Xin Liu, Mingxuan Wang, Yonghui Wu, and Lin Yan. VAPO: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025. URL https://arxiv.org/abs/2504.05118v3.

Wenxuan Zhang, Yuhui Wang, Donggang Jia, Xiaoqian Shen, Jian Ding, Ivan Viola, Jürgen Schmidhuber, and Mohamed Elhoseiny. Hybrid advantage estimation with unified critic for VLM agentic reinforcement learning. *arXiv preprint arXiv:2607.23605*, 2026. URL https://arxiv.org/abs/2607.23605v1.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. *arXiv preprint arXiv:2402.19446*, 2024. URL https://arxiv.org/abs/2402.19446v1.

Siyuan Zhu, Chao Yu, Rongxin Yang, Zongkai Liu, Jinjun Hu, Qiwen Chen, and Yibo Zhang. GAGPO: Generalized advantage grouped policy optimization. *arXiv preprint arXiv:2605.13217*, 2026. URL https://arxiv.org/abs/2605.13217v1.

# A Additional Analyses

## A.1 Sensitivity to the token-residual scale

We examine the residual weight $\alpha$ in Equation 6 for the 1.5B shared-critic configuration. Figure 5 reports ALFWorld success rate, WebShop success rate, and WebShop task score for $\alpha \in \{1, 2, 3\}$.

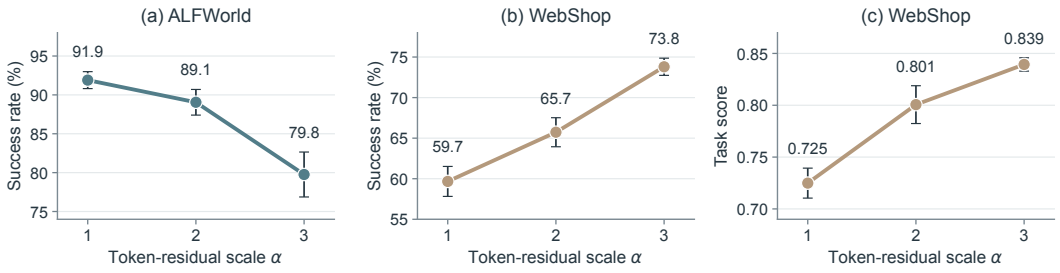


Figure 5: Sensitivity to the token-residual scale $\alpha$ with Qwen2.5-1.5B-Instruct. Markers and error bars show the mean and sample standard deviation over three random seeds.

On ALFWorld, mean success declines from 91.9% at $\alpha = 1$ to 89.1% at $\alpha = 2$ and 79.8% at $\alpha = 3$. WebShop shows the opposite trend: mean success rises from 59.7% to 65.7% and 73.8%, while mean task score rises from 0.725 to 0.801 and 0.839. These opposing trends suggest a task-dependent balance: among the tested settings, ALFWorld favors a smaller token-residual scale, whereas WebShop favors a larger one. One possible explanation is that ALFWorld's action outcomes provide relatively clear turn-level progress signals, while WebShop's search and selection actions may benefit more from distinguishing generated tokens within each response. Each $\alpha$ corresponds to a separately trained checkpoint, and the error bars reflect decoding-seed variation rather than independent training runs.

## A.2 Turn and token credit in an ALFWorld trajectory

We illustrate the two levels of credit with a successful diagnostic rollout on the seen ALFWorld task "look at book under the desklamp" (latest step-150 checkpoint, environment seed 0, decoding seed 123). The agent moves to desk 1, picks up book 4, and uses desklamp 1 to complete the task.

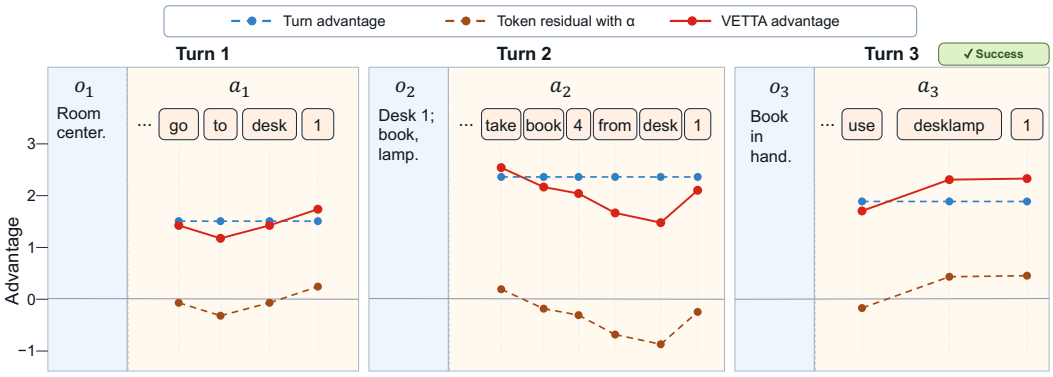


Figure 6: Turn- and token-level credit in a successful ALFWorld trajectory. Blue, brown, and red denote $A_k^{\text{turn}}$, $\alpha\widetilde{A^{\text{tok}}}_{k,j}$, and $A_{k,j}^{\text{VETTA}}$. Only parsed actions are shown; thinking text is abbreviated. When a plotted word spans multiple generated tokens, its values are arithmetic means over those token positions.

The three turns receive positive turn advantages (1.494, 2.348, and 1.875), providing response-level credit. Token residuals distinguish words within each response: in the pickup action, `take` receives $+0.180$ and `from` receives $-0.695$; in the final action, `desklamp` and `1` receive $+0.421$ and $+0.441$. The latter two words have fused advantages of 2.296 and 2.316, above their turn-level value of 1.875. Thus, the turn advantage credits the interaction as a whole, while the token residual adjusts credit among its generated words.

The displayed residuals need not average to zero because the figure omits thinking text. Centering uses all valid generated tokens in the complete response, including the omitted text, so its full-response residual mean is zero by construction.

# B Experimental Details

## B.1 Task interaction and rewards

**ALFWorld.** In the text-based ALFWorld setting (Shridhar et al., 2021), the agent receives a household goal and textual observations, and issues commands to navigate, inspect, and manipulate objects. The environment executes the command extracted from each generated response and supplies the next observation. The text-environment wrapper assigns a task reward of 10 upon success and 0 otherwise. An action flagged as invalid receives an additional penalty of 0.1.

**WebShop.** The agent receives a shopping instruction and interacts with the simulated website through search and click actions (Yao et al., 2022a). Observations describe the current search results or product page. The wrapper retains the environment's continuous task score for evaluation, but uses a task reward of 10 only when the purchase ends the episode with a task score of 1; other outcomes receive 0. Invalid actions incur an additional penalty of 0.1. The task score reported in Table 1 is multiplied by 100 for presentation and is distinct from this training reward.

The reward manager places each turn's task reward and invalid-action penalty at its final valid generated-token position, as in Equation 3. KL regularization is applied separately to the actor loss rather than added to these rewards.

## B.2 Run-level configuration details

**Critic configurations.** The main-table VETTA configurations use one shared critic backbone, two output heads, and separate token and turn targets. The critic retains two blocks for both the 1.5B and 7B models on ALFWorld and WebShop. Critic-based baselines use the full depth of their corresponding pretrained backbones. Historical independent-critic runs use two complete critic models and two optimizers. The full 28-layer independent run is not a matched depth comparison for the

Table 3: Training settings for VETTA on ALFWorld and WebShop.

| Setting | ALFWorld | WebShop |
|---|---|---|
| Retained critic blocks | 2 | 2 |
| Tasks $\times$ rollouts/update | $128 \times 1$ | $128 \times 1$ |
| Maximum turns | 50 | 15 |
| PPO minibatch | 256 | 64 |
| Actor / critic LR | $10^{-6}/10^{-5}$ | $10^{-6}/10^{-5}$ |
| KL loss coefficient | 0.01 | 0.01 |
| Token $\gamma/\lambda$ | $1/1$ | $1/1$ |
| Turn $\gamma/\lambda$ | $0.95/0.95$ | $0.95/0.95$ |
| Update budget | 150 | 150 |

Table 4: Critic timing by backbone depth. All entries are medians over 112 clean training steps. Local time is value inference plus critic update within each step; its share is computed against that step's total time.

| Benchmark | Blocks | Value (s) | Update (s) | Local (s) | Share (%) |
|---|---|---|---|---|---|
| ALFWorld | 2 | 3.116 | 16.538 | 19.696 | 3.845 |
| ALFWorld | 8 | 10.585 | 50.460 | 60.737 | 9.763 |
| ALFWorld | 28 | 26.700 | 125.500 | 151.975 | 25.634 |
| WebShop | 2 | 1.480 | 6.546 | 7.917 | 4.321 |
| WebShop | 8 | 4.297 | 18.212 | 22.455 | 9.661 |
| WebShop | 28 | 17.579 | 74.459 | 92.224 | 27.783 |

shared architecture. In the matched actor-credit comparison C1–C4, both value heads and their losses remain active; C5–C6 use single-branch critic supervision.

### B.3 Evaluation sets and aggregation

The fixed ALFWorld evaluation comprises 140 seen and 134 unseen tasks. The WebShop endpoint uses a fixed pool of 500 tasks.

Figures 1 and 5, together with our locally evaluated three-seed results in Table 1, use decoding seeds 123, 456, and 789. We evaluate each fixed checkpoint with these three seeds and report the arithmetic mean and sample standard deviation across decoding runs. The ALFWorld evaluations use the latest step-150 checkpoint, the fixed 140-task `valid_seen` split, environment seed 0, temperature 0.4, and sampling. WebShop uses the latest step-150 checkpoint and the fixed 500-task manifest. Rows reported from GiGPO retain their source paper's evaluation protocol.

The credit-composition results in Table 2 and the critic-structure results in Figure 3 instead use a single decoding seed, 0, at the latest step-150 checkpoint. Consequently, their success rates can differ from the three-seed averages in Table 1, even for the same configuration. The depth-performance results in Figure 1(b) use the three decoding seeds specified above; the critic-time measurements in Figure 4 come from training logs and do not involve decoding seeds.

### B.4 Critic-depth timing

All six depth runs train for 150 updates. For timing, we retain steps 11–149 whose indices are not divisible by five, excluding the first ten training steps and scheduled validation or checkpoint steps. This gives 112 steps per run. We add value-inference and critic-update times within each step, then take the median. The reported time share is the median of the per-step ratio of this sum to total step time.

Figure 4 shows the compact depth comparison. Table 4 reports the underlying stage medians. These are within-run measurements of critic computation, not an end-to-end speedup against a critic-free optimizer. ALFWorld's two- and 28-block runs used the same host, as did WebShop's eight- and 28-block runs; the other depth pairs span hosts.

# C Method Details and Derivations

## C.1 Value-position alignment

For a response of length $L$, the implementation slices critic logits at positions `[-L-1:-1]`. The first aligned output therefore gives the value immediately before the first generated token, and the $j$th aligned output precedes generated token $y_j$. The complete prompt and response are encoded in one causal forward pass, but causal attention prevents future response tokens from changing an earlier aligned value. The turn target is supervised only at the first aligned position, whereas token targets use every valid response position.

## C.2 Valid generated-token positions

The rollout stores each turn as a context–response pair. The response mask includes generated positions through the first end-of-sequence token, including that token when present, and excludes subsequent padding. If no end-of-sequence token is emitted before the generation limit, the retained generated positions remain valid. Generated reasoning text and formatting tokens are not separately removed. Instructions, previous interaction history, and environment observations enter the prompt and are excluded from the current response mask. This mask selects the token-level value targets, advantage-centering positions, and actor-loss positions. Token-level GAE connects valid generated positions in temporal order across responses of the same trajectory; each turn contributes only one turn-value target at its first aligned position.

## C.3 Critic regression objective

For granularity $b \in \{\mathrm{turn}, \mathrm{tok}\}$ and valid position $i$, let $\widehat{G}^b(i)$ be the fixed target in Equation 7, $\psi$ the trainable critic parameters, and $\epsilon_v > 0$ the value-clipping radius. The clipped value loss is

$$\ell^b(i) = \max\left\{ \left(V_\psi^b(i) - \widehat{G}^b(i)\right)^2, \left[\mathrm{clip}\big(V_\psi^b(i), V_{\mathrm{old}}^b(i) - \epsilon_v, V_{\mathrm{old}}^b(i) + \epsilon_v\big) - \widehat{G}^b(i)\right]^2 \right\}. \quad (9)$$

The two heads are reduced over their own valid positions before their losses are summed, without dividing the sparse turn loss by the number of tokens:

$$\mathcal{L}_{\mathrm{critic}} = \mathrm{Mean}_{k=1,\dots,K}[\ell^{\mathrm{turn}}(k)] + \mathrm{Mean}_{(k,j):\,1\le k\le K,\,1\le j\le L_k}[\ell^{\mathrm{tok}}(k,j)]. \quad (10)$$

## C.4 Response-wise properties of advantage fusion

Write $D_{k,j} = \widetilde{A^{\mathrm{tok}}}_{k,j}$ for the centered token residual. For a response with $L_k$ valid generated tokens and fixed $\alpha$, consider the token-wise vector $\mathbf{z} = (z_1, \dots, z_{L_k})$. Among vectors whose response-wise mean equals the turn advantage, the one closest to the scaled token advantages solves

$$\begin{aligned} \min_{\mathbf{z}\in\mathbb{R}^{L_k}} \quad & \frac{1}{2}\sum_{j=1}^{L_k}\left(z_j - \alpha A_{k,j}^{\mathrm{tok}}\right)^2 \\ \text{subject to} \quad & \frac{1}{L_k}\sum_{j=1}^{L_k} z_j = A_k^{\mathrm{turn}}. \end{aligned} \quad (11)$$

With a Lagrange multiplier $\eta$ for the equivalent constraint $\sum_j z_j = L_k A_k^{\mathrm{turn}}$, stationarity gives $z_j = \alpha A_{k,j}^{\mathrm{tok}} - \eta$. The constraint yields $\eta = \alpha\overline{A^{\mathrm{tok}}}_k - A_k^{\mathrm{turn}}$; thus $z_j = A_{k,j}^{\mathrm{VETTA}}$.

The squared-distance objective is strictly convex, so this solution is unique. It follows that

$$\begin{aligned} \sum_{j=1}^{L_k} D_{k,j} &= \sum_{j=1}^{L_k}\left(A_{k,j}^{\mathrm{tok}} - \overline{A^{\mathrm{tok}}}_k\right) = 0, \\ \frac{1}{L_k}\sum_{j=1}^{L_k} A_{k,j}^{\mathrm{VETTA}} &= A_k^{\mathrm{turn}}, \\ A_{k,i}^{\mathrm{VETTA}} - A_{k,j}^{\mathrm{VETTA}} &= \alpha\left(A_{k,i}^{\mathrm{tok}} - A_{k,j}^{\mathrm{tok}}\right). \end{aligned}$$

This projection is an algebraic characterization of the scalar credit weights, not an additional optimization step during training. Different tokens have different score-function gradients, so zero scalar sum does not imply a zero parameter-gradient contribution.

Under the additional special case $\gamma_{\text{tok}} = \lambda_{\text{tok}} = 1$, complete termination, no token-wise KL reward, and reward only at an action's last valid token, all prefixes in action $k$ share the same remaining sampled return $G_k$. Then

$$A^{\text{tok}}_{k,j} = G_k - V^{\text{tok}}_{k,j}, \qquad D_{k,j} = \frac{1}{L_k} \sum_{u=1}^{L_k} V^{\text{tok}}_{k,u} - V^{\text{tok}}_{k,j}. \tag{12}$$

The common sampled return cancels from the within-action residual. This conditional derivation motivates the interpretation "turn mean plus relative prefix differences." This identity describes how credit varies across positions in a sampled response; it does not identify the effect of replacing an individual token.

### C.5 Actor-loss implementation

Equation 8 gives the standard token-wise clipped PPO surrogate. The reported VETTA runs use dual clipping and KL regularization as actor-loss settings. Dual clipping caps the per-token loss after PPO ratio clipping at $-cA^{\text{VETTA}}_{k,j}$ when $A^{\text{VETTA}}_{k,j} < 0$, with $c = 3$; the loss for nonnegative advantages is unchanged. The actor loss also includes KL regularization against a frozen reference policy initialized from the actor's starting checkpoint. The ALFWorld and WebShop configurations use the `low_var_kl` estimator and a KL coefficient of 0.01. Both loss terms are averaged over valid generated-token positions. These actor-loss settings are inherited from the token-wise PPO implementation shared with the PPO and PPO* baselines; they are not specific to VETTA's value estimation or advantage fusion. The Turn-PPO baseline instead clips a whole-response probability ratio and does not apply the negative-advantage dual-clipping rule.