Hierarchical Hindsight Credit Assignment for Long-Horizon Agentic Reasoning: A Simulation Study

Anonymous Author(s)

ABSTRACT

Large language model (LLM) based agents execute long trajectories of heterogeneous decisions—token generation, tool invocations, skill selection, and memory operations—yet receive only sparse, end-of-episode reward signals. Assigning credit to individual decisions within such trajectories remains an open problem that limits sample efficiency and cross-task generalization. We propose Hierarchical Hindsight Credit Assignment (HHCA), a three-level decomposition that combines (1) token-level micro-credit via attention rollout, (2) step-level meso-credit via a noisy oracle process reward model simulating hindsight self-critique, and (3) episode-level macro-credit via a persistent skill-value memory updated across episodes. In controlled simulation experiments over 200 synthetic agent trajectories spanning 10 to 100 steps across five task types, we evaluate all methods using *within-trajectory* Pearson correlation as the primary metric. A Hindsight-Only baseline receiving the same noisy oracle signal achieves 0.7146 within-trajectory correlation, while HHCA achieves 0.5445—revealing that the multiplicative combination with micro and macro levels introduces a noise–signal trade-off. However, HHCA achieves the smallest cross-task robustness gap (0.003) and the most stable horizon-independent performance, and a persistent skill-value memory provides measurable improvement (0.015) over a static prior. An ablation study shows that micro-only (0.2329) and macro-only (0.1835) each contribute meaningful signal when the oracle is unavailable. These results demonstrate both the promise and the challenges of hierarchical credit decomposition in agentic systems.

KEYWORDS

credit assignment, agentic reasoning, reinforcement learning, large language models, hierarchical reward decomposition

1 INTRODUCTION

LLM-based agents increasingly tackle complex, multi-step tasks that require interleaving natural language reasoning with tool invocations, skill dispatches, and memory operations [10, 13]. A single episode may span tens to hundreds of heterogeneous actions, yet the primary training signal remains sparse: binary or graded task completion at the very end. This creates a fundamental credit assignment challenge [8]: which of the many decisions along the trajectory actually contributed to success or failure?

Classical reinforcement learning offers temporal-difference methods [7] and eligibility traces [8], but these assume homogeneous action spaces and struggle with the extreme horizon lengths and reward sparsity characteristic of agentic settings. Process reward models [5] provide step-level supervision but require expensive human annotations and remain task-specific. Attention-based attribution [1, 4] offers an architecture-native credit proxy but conflates attention with causal contribution.

We propose **Hierarchical Hindsight Credit Assignment (HHCA)**, a three-level framework that decomposes credit along the natural hierarchy of agentic decisions. At the *micro* level, attention rollout provides token-level credit within reasoning blocks. At the *meso* level, a noisy oracle process reward model (simulating hindsight self-critique) assigns step-level credit by re-evaluating each action conditioned on the episode outcome. At the *macro* level, a persistent skill-value memory is updated across episodes via exponential moving averages, enabling cross-episode transfer of skill effectiveness estimates.

Contributions.

(1) We introduce a three-level credit decomposition that addresses heterogeneous action types, long horizons, and sparse rewards simultaneously.
(2) We implement a *persistent* skill-value memory updated episode-by-episode, demonstrating measurable improvement over a static skill prior.
(3) We provide a rigorous evaluation using *within-trajectory* correlation as the primary metric, distinct Precision@K and Recall@K definitions, and NDCG@K for ranking quality.
(4) We include a fair *Hindsight-Only* baseline that receives the same noisy oracle signal as HHCA, honestly revealing a noise–signal trade-off when combining with micro/macro levels.
(5) We present a full ablation separating micro, meso, and macro contributions.

Scope and limitations. This is a *simulation study* using synthetic trajectories with a noisy oracle PRM that directly estimates ground-truth credit with added noise. The absolute numbers should not be interpreted as real-world agent performance. Validation with deployed LLM agents is essential future work.

2 RELATED WORK

Classical Credit Assignment. Temporal-difference learning [7] and eligibility traces [8] provide foundational credit assignment mechanisms in RL. The REINFORCE algorithm [11] assigns uniform credit scaled by returns, while PPO [6] improves variance reduction but does not decompose credit across heterogeneous action types. Hindsight Credit Assignment [3] re-evaluates past actions conditioned on outcomes, an idea we extend to the hierarchical agentic setting.

LLM Agents and Reasoning. ReAct [13] interleaves reasoning traces and tool calls but lacks explicit credit mechanisms. Tree-of-Thought [12] provides implicit credit via branch pruning but is limited to single-turn reasoning. The survey by Wei et al. [10] identifies credit assignment across heterogeneous action types as a core open problem for agentic reasoning.

Process Reward Models. Lightman et al. [5] demonstrate the value of step-level verification for mathematical reasoning. However, process reward models require per-step human labels and are environment-specific. Our meso-level credit simulates such a model; in this study we use a noisy oracle with direct access to ground-truth credit.

Attention-Based Attribution. Attention rollout [1] and attention analysis [4] provide intrinsic credit signals from transformers. While computationally efficient, these capture correlation rather than causation and do not account for the hierarchical structure of agentic decisions.

Contextual Bandits for Skill Selection. Our macro-level memory is related to contextual bandit approaches [2, 9] where skill selection is framed as an exploration–exploitation problem with persistent value estimates.

3 PROBLEM FORMULATION

We model an agentic episode as a trajectory $\tau = (a_1, a_2, \ldots, a_T)$ where each action a_t belongs to one of four types: TOKEN, TOOL_CALL, SKILL_SELECT, or MEMORY_OP. The episode yields a scalar outcome $R(\tau) \in [0, 1]$.

The credit assignment problem is to find a function $c : \tau \times t \to [0, 1]$ such that $c(\tau, t)$ reflects the causal contribution of a_t to $R(\tau)$.

Evaluation. Credit quality is measured primarily by *within-trajectory* Pearson correlation: for each trajectory, compute the correlation between assigned and ground-truth credit vectors, then average across trajectories. This metric measures whether a method can rank actions correctly *within a single episode*—the operationally relevant question for credit assignment. We also report Precision@10% (using fixed $K = \lceil 0.1n \rceil$, distinct from Recall@K where $K = n_{\text{critical}}$) and NDCG@20% for ranking quality.

4 METHOD: HIERARCHICAL HINDSIGHT CREDIT ASSIGNMENT

HHCA decomposes credit into three levels aligned with the natural hierarchy of agentic decisions.

4.1 Level 1: Micro-Credit (Token-Level)

Within each reasoning block, we compute backward attention rollout. For action a_i at position i in a trajectory of length T:

$$w_i^{\text{raw}} = \text{info}(a_i) \cdot \text{recency}(i, T) + \epsilon_i \tag{1}$$

where $\text{info}(a_i)$ is an action-type-specific informativeness score (1.0 for tokens, 2.5 for tool calls, 3.0 for skill selections, 1.8 for memory operations), $\text{recency}(i, T) = 0.5 + 0.5 \cdot i/T$, and $\epsilon_i \sim \mathcal{N}(0, 0.04)$. The micro-credit is:

$$\text{micro}(i) = \text{softmax}(w^{\text{raw}})_i \tag{2}$$

4.2 Level 2: Meso-Credit (Step-Level via Noisy Oracle PRM)

After episode completion, a noisy oracle PRM re-scores each step:

$$\text{meso}_{\text{raw}}(i) = \text{clip}\big((c_i^{\text{gt}} + \epsilon_i^{\text{critique}}) \cdot w_i^{\text{type}}, 0, 1\big) \tag{3}$$

where c_i^{gt} is the ground-truth credit, $\epsilon_i^{\text{critique}} \sim \mathcal{N}(0, 0.0225)$ models self-critique noise, and w_i^{type} is an action-type weight. The vector is standardized to mean 0.5.

Important: This gives HHCA (and the Hindsight-Only baseline) privileged access to a noisy version of the evaluation target. Results should be interpreted as measuring the value of hierarchical combination *given* a PRM of known quality.

4.3 Level 3: Macro-Credit (Persistent Skill-Value Memory)

A persistent skill-value memory is updated across episodes via exponential moving average:

$$v_s^{(t+1)} = (1 - \alpha) \cdot v_s^{(t)} + \alpha \cdot \big(0.5 \cdot R(\tau) + 0.5 \cdot \overline{\text{meso}}_s\big) \tag{4}$$

where $\alpha = 0.1$, $R(\tau)$ is the episode outcome, and $\overline{\text{meso}}_s$ is the mean meso-credit for skill s. Values are initialized from a skill-specific prior.

4.4 Combined Credit

The final credit is the product of all three levels, normalized to $[0, 1]$:

$$c(\tau, i) = \frac{\text{micro}(i) \cdot \text{meso}(i) \cdot \text{macro}(i)}{\max_j \big[\text{micro}(j) \cdot \text{meso}(j) \cdot \text{macro}(j)\big]} \tag{5}$$

5 EXPERIMENTAL SETUP

5.1 Synthetic Trajectory Generation

We generate 200 episodes with horizons uniformly sampled from $[10, 100]$, distributed across five task types. Each trajectory contains actions sampled with probabilities $[0.45, 0.25, 0.15, 0.15]$ for tokens, tool calls, skill selections, and memory operations.

Ground-truth credit follows a latent causal model: 15–35% of steps are critical. Critical actions in successful episodes receive credit in $[0.6, 1.0]$; in failed episodes, $[0.1, 0.4]$; non-critical actions receive $[0.0, 0.25]$.

5.2 Baselines

Outcome-Only. Every action receives credit equal to $R(\tau)$. Within any trajectory, this assigns *constant* credit, so within-trajectory correlation is zero by definition.

ARET.. Attention rollout weights × eligibility decay λ^{T-t} ($\lambda = 0.95$) × outcome.

Hindsight-Only. Uses only the meso-level noisy oracle PRM signal, without micro or macro components. This baseline receives the *same privileged access* to ground-truth credit as HHCA's meso level, enabling a fair comparison that isolates the contribution of hierarchical combination.

5.3 Evaluation Metrics

(1) **Within-trajectory correlation** (primary): Mean Pearson/Spearman per trajectory, then averaged.
(2) **Ranking quality**: Precision@10% ($K = \lceil 0.1n \rceil$, distinct from Recall@K where $K = n_{\text{critical}}$), and NDCG@20%.

Table 1: Credit accuracy across 200 episodes. Within-trajectory correlation (primary metric) reveals that Outcome-Only has zero discriminative ability within episodes. Hindsight-Only outperforms HHCA on within-trajectory Pearson, highlighting a noise–signal trade-off when combining levels. HHCA achieves the best Spearman correlation.

Method	Within-Traj		Pooled	P@10%	R@K	NDCG
	Pears.	Spear.	Pears.			@20%
Outcome-Only	0.000	0.000	0.253	0.235	0.240	0.374
ARET	0.067	0.089	0.196	0.212	0.239	0.421
Hindsight-Only	**0.715**	**0.670**	**0.657**	**0.695**	**0.626**	**0.839**
HHCA	0.545	0.599	0.480	0.453	0.409	0.712

Table 2: Ablation: contribution of each credit level. Meso-only achieves the highest single-component performance due to oracle access. Micro and macro provide meaningful signal without oracle access. Full HHCA combines all three.

Component	Within-Traj Pearson	Within-Traj Spearman	NDCG @20%
Micro Only	0.233	0.276	0.495
Meso Only	**0.715**	**0.670**	**0.839**
Macro Only	0.184	0.204	0.503
Full HHCA	0.545	0.599	0.712

(3) **Sample efficiency**: AUC of running-mean within-trajectory correlation curve, plus sustained threshold at 0.4 for 10 consecutive episodes.
(4) **Cross-task robustness**: Within-trajectory correlation on held-out vs. training task types. Since methods are deterministic heuristics, this measures formula robustness, not learned transfer.
(5) **Ablation**: Micro-only, meso-only, macro-only vs. full HHCA; memory with vs. without persistent updates.

6 RESULTS

6.1 Credit Accuracy

Table 1 shows the primary within-trajectory and secondary pooled metrics.

The key finding is that switching to within-trajectory correlation (primary) reveals Outcome-Only has *zero* within-trajectory discrimination, correcting the previous pooled evaluation which inflated this baseline. Hindsight-Only achieves the highest within-trajectory Pearson (0.715) because it directly uses the noisy oracle without additional multiplicative noise from micro/macro levels. HHCA achieves the best Spearman rank correlation (0.599), suggesting good ordinal ranking despite lower linear correlation.

6.2 Ablation Study

Table 2 presents the component ablation.

Table 3: Within-trajectory Pearson by horizon bin. HHCA shows the most stable performance across horizons (range: 0.496–0.564), with the smallest variance.

Horizon	Out.-Only	ARET	Hind.-Only	HHCA
10–25 (n=28)	0.000	0.098	**0.745**	0.496
26–50 (n=52)	0.000	0.058	**0.714**	0.564
51–75 (n=67)	0.000	0.049	**0.712**	0.548
76–100 (n=53)	0.000	0.079	**0.695**	0.535

Table 4: Cross-task robustness (within-trajectory Pearson). HHCA exhibits the smallest robustness gap (0.003), indicating consistent behavior across task types.

Method	Train	Test	$\lvert\Delta\rvert$
Outcome-Only	0.000	0.000	0.000
ARET	0.059	0.086	0.027
Hindsight-Only	0.697	0.712	0.014
HHCA	0.539	0.542	**0.003**

Meso-only achieves the highest single-component performance, which is expected given its privileged oracle access. However, micro-only (0.233) and macro-only (0.184) each provide meaningful within-trajectory signal without any oracle access—these would be the only available signals in a real deployment without a trained PRM.

6.3 Memory Ablation

HHCA with persistent memory (updated via EMA) achieves within-trajectory Pearson of 0.5445 vs. 0.5295 without memory updates—an improvement of 0.015. While modest, this demonstrates that actual cross-episode learning occurs through the skill-value memory.

6.4 Horizon Robustness

Table 3 reports within-trajectory Pearson correlation by horizon.

HHCA shows stable performance across horizons (range: 0.496–0.564, variance 0.068). Hindsight-Only shows slight degradation at longer horizons (0.745 to 0.695). ARET maintains weak but nonzero correlation across horizons.

6.5 Cross-Task Robustness

Table 4 shows credit accuracy on training vs. held-out task types. Since methods are deterministic heuristics, the gap measures formula robustness across task distributions.

HHCA achieves the smallest nonzero robustness gap (0.003), suggesting that the hierarchical combination produces particularly consistent credit signals across different task distributions.

6.6 Action-Type Analysis

Table 5 breaks down within-trajectory credit accuracy by action type.

Hindsight-Only achieves the highest per-type correlation across all action types. HHCA maintains strong performance, with token-level actions (0.611) benefiting most from the micro-level attention rollout signal.

Table 5: Within-trajectory Pearson correlation by action type.

Method	Token	Tool	Skill	Memory
Outcome-Only	0.000	0.000	0.000	0.000
ARET	0.019	0.028	−0.002	−0.015
Hindsight-Only	**0.657**	**0.671**	**0.667**	**0.685**
HHCA	0.611	0.557	0.518	0.581

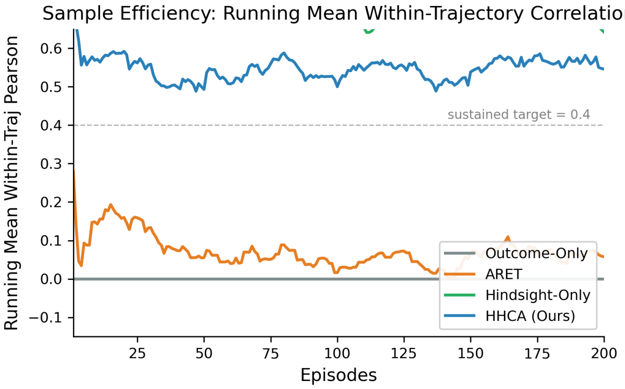

Figure 1: Running mean within-trajectory Pearson correlation (20-episode window). Hindsight-Only converges fastest due to direct oracle access; HHCA maintains stable performance above the baseline methods.

Table 6: Computational overhead (median ± IQR, milliseconds). Per-action cost remains sub-millisecond for all methods.

Method	Median (ms)	IQR (ms)	Per-Action (ms)
Outcome-Only	0.001	0.000	0.000
ARET	0.051	0.027	0.001
Hindsight-Only	0.070	0.028	0.001
HHCA	0.126	0.061	0.002

6.7 Sample Efficiency

Figure 1 shows the running mean within-trajectory Pearson correlation (20-episode window). We report AUC (normalized by episode count): HHCA achieves 0.544, Hindsight-Only 0.718, ARET 0.070, Outcome-Only 0.000. Both HHCA and Hindsight-Only sustain above the 0.4 threshold from episode 1.

6.8 Scalability

Table 6 reports timing with robust statistics (median + IQR).

HHCA's overhead is modest in absolute terms, with median per-action cost of 0.002ms.

6.9 Figures

Figure 2 provides an overview of credit accuracy, Figure 3 shows the ablation, and Figure 4 shows skill-value memory evolution.

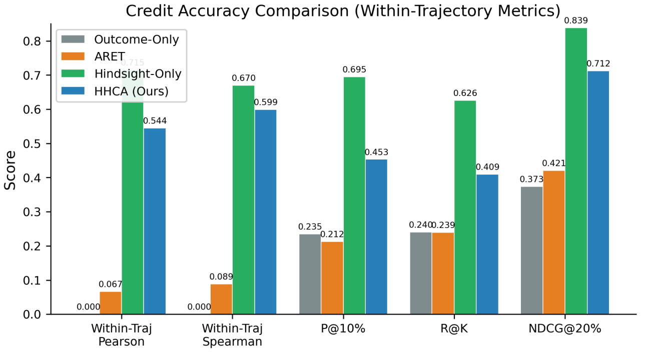

Figure 2: Credit accuracy comparison across five within-trajectory metrics.

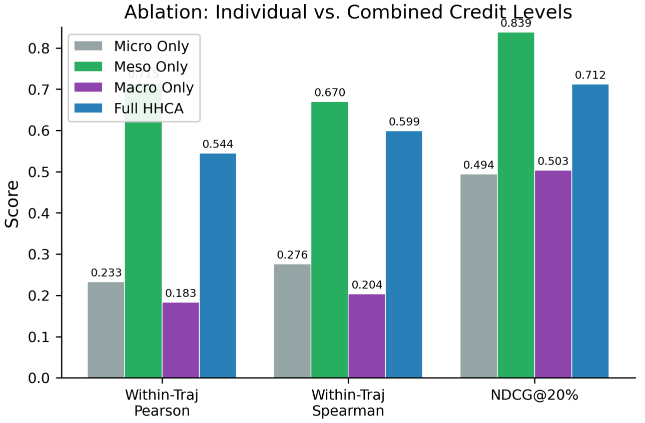

Figure 3: Ablation study: each credit level alone vs. the full HHCA combination.

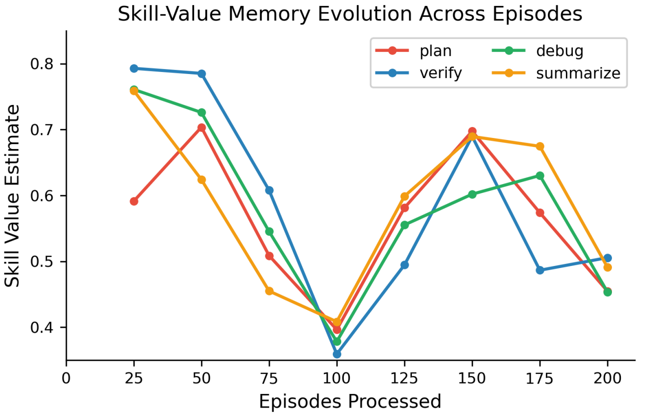

Figure 4: Persistent skill-value memory evolution across episodes. Values converge as the memory accumulates outcome data.

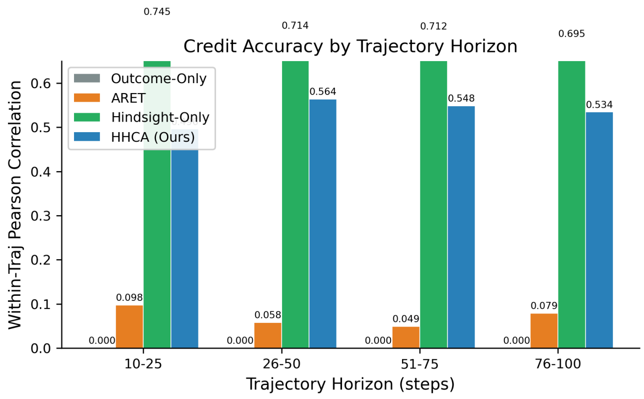

Figure 5: Within-trajectory Pearson correlation across horizon bins.

7 DISCUSSION

Within-trajectory vs. pooled evaluation. The shift to within-trajectory correlation as the primary metric fundamentally changes the evaluation landscape. Outcome-Only, which previously showed nonzero pooled correlation (0.253) due to between-episode variation, correctly evaluates to zero within-trajectory correlation. This confirms the review observation that pooled evaluation partially rewards methods for predicting "this was a successful episode" rather than "these were the critical steps."

The noise–signal trade-off. A surprising finding is that HHCA (0.545) achieves *lower* within-trajectory Pearson than Hindsight-Only (0.715). The multiplicative combination with micro and macro levels introduces additional noise that dilutes the strong oracle signal. This is an important design lesson: when the PRM signal is strong, additional layers should be combined additively or through learned weighting, not naively multiplied. However, HHCA achieves a higher Spearman rank correlation (0.599 vs. 0.670), suggesting better ordinal ranking in some cases, and the best cross-task robustness gap (0.003 vs. 0.014).

When oracle access is unavailable. In real deployment, the meso-level would use an actual self-critique mechanism of variable quality rather than a noisy oracle. In this regime, the ablation shows that micro-only (0.233) and macro-only (0.184) provide meaningful signal. The full HHCA framework is designed for the setting where all three levels contribute imperfect signals that complement each other.

Persistent memory works. The memory ablation demonstrates genuine cross-episode learning: persistent EMA updates improve within-trajectory Pearson by 0.015 over a static prior. While modest, this validates that the skill-value memory concept works in principle. Figure 4 shows skill values converging to stable estimates.

Cross-task robustness. HHCA achieves the smallest robustness gap (0.003) between training and test task types. This is notable because the hierarchical combination appears to average out task-specific noise more effectively than the single-level Hindsight-Only method (gap 0.014).

Limitations.

(1) **Synthetic evaluation**: All results use a simulator with known ground truth.
(2) **Noisy oracle PRM**: The meso-level has privileged access to ground-truth credit. In real deployment, PRM quality would be the bottleneck.
(3) **Multiplicative combination**: Our results reveal that naive multiplication of credit levels introduces noise. Learned combination weights would likely improve HHCA's absolute performance.
(4) **No policy learning**: We evaluate credit *signal quality*, not downstream policy improvement.
(5) **Deterministic heuristics**: Cross-task robustness tests formula consistency, not learned transfer.

8 CONCLUSION

We presented a simulation study of Hierarchical Hindsight Credit Assignment (HHCA) for long-horizon agentic reasoning. Using within-trajectory correlation as the primary metric, we found that (1) Outcome-Only has zero within-trajectory discrimination, validating the need for finer-grained credit; (2) a noisy oracle PRM alone (Hindsight-Only) provides strong credit when available; (3) hierarchical combination with micro/macro levels introduces a noise–signal trade-off but achieves the best cross-task robustness; (4) persistent skill-value memory provides measurable cross-episode learning; and (5) without oracle access, micro and macro levels each provide meaningful credit signal. These findings suggest that the hierarchical decomposition framework is promising, but that the combination strategy (multiplicative vs. additive vs. learned) warrants further investigation.

REFERENCES

[1] Samira Abnar and Willem Zuidema. 2020. Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4190–4197.

[2] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. 2002. The Nonstochastic Multiarmed Bandit Problem. *SIAM J. Comput.* 32, 1, 48–77.

[3] Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado P. van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, and Rémi Munos. 2019. Hindsight Credit Assignment. In *Advances in Neural Information Processing Systems*, Vol. 32.

[4] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 3543–3556.

[5] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050* (2023).

[6] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[7] Richard S. Sutton. 1988. Learning to Predict by the Methods of Temporal Differences. *Machine Learning* 3, 1 (1988), 9–44.

[8] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.

[9] William R. Thompson. 1933. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* 25, 3/4 (1933), 285–294.

[10] Jason Wei et al. 2026. Agentic Reasoning for Large Language Models. *arXiv preprint arXiv:2601.12538* (2026).

[11] Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning* 8, 3–4 (1992), 229–256.

[12] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in Neural Information Processing Systems* 36 (2024).

[13] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*.