LLM-Guided Q-Value Initialization Enhances Reinforcement Learning Sample Efficiency

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of LLM-guided Q-value initialization on the sample efficiency and convergence stability of reinforcement learning agents in complex reasoning tasks. 5 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: From Reward Shaping to Q-Shaping: Achieving Unbiased Learning with LLM-Guided Knowledge. Research question: What is the impact of LLM-guided Q-value initialization on the sample efficiency and convergence stability of reinforcement learning agents in complex reasoning tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.0/10.

3 Results

14 papers retrieved. 5 claims extracted; 1 independently verified. Quality review score: 5.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Q-shaping improves performance by 16.87% compared to the best baseline and by 55.39% compared to TD3 across 20 tasks.	×	0.14
Q-shaping outperforms LLM-based reward shaping methods by 253.80% on average across four Meta-World environments: door-c	✓	0.19
LLM-TD3 improved by 38.68% in the door-close task, 406.04% in drawer-open, 389.77% in window-close, and 180.70% in sweep	×	0.01
Most LLMs, including o1-Preview, GPT-4o, DeepSeek-V2.5, and yi-large, provided correct heuristic functions with a 100% s	×	0.05
Gemini exhibited poorer performance, achieving only 44% on average in the evaluation of LLM-generated heuristic function	×	0.04

References

- http://arxiv.org/abs/1911.09615v1
- http://arxiv.org/abs/1911.11285v1
- http://arxiv.org/abs/2410.01458v1