Better Goals, Better Policies: LLM-Driven Relabeling for Offline Goal-Conditioned Reinforcement Learning

Xule Gao
Institute of Software Chinese Academy of Sciences
Beijing, China
University of Chinese Academy of Sciences
Beijing, China
gaoxule24@mails.ucas.ac.cn

Chuxiong Sun*
National Key Laboratory of Space Integrated Information System, Institute of Software Chinese Academy of Sciences
Beijing, China
chuxiong2016@iscas.ac.cn

Rui Wang
National Key Laboratory of Space Integrated Information System, Institute of Software Chinese Academy of Sciences
Beijing, China
wangrui@iscas.ac.cn

Changwen Zheng
National Key Laboratory of Space Integrated Information System, Institute of Software Chinese Academy of Sciences
Beijing, China
changwen@iscas.ac.cn

ABSTRACT

Offline goal-conditioned reinforcement learning (offline GCRL) learns goal-conditioned policies from fixed, reward-free datasets. Existing methods often rely on hindsight experience replay(HER), which treats all future states uniformly, leading to many uninformative goals. We propose LLM-Driven Relabeling, an adaptive and trustworthy framework that uses LLM-generated semantic rules to identify task-relevant key states and prioritize them as informative relabeling goals. We theoretically demonstrate that such goals lead to larger TD errors, thereby reducing sample complexity. Empirical results on offline GCRL benchmarks demonstrate that LLM-Driven Relabeling significantly improves learning efficiency, particularly under reduced-data conditions.

KEYWORDS

Reinforcement Learning, Offline Goal-conditioned Reinforcement Learning

ACM Reference Format:
Xule Gao, Chuxiong Sun, Rui Wang, and Changwen Zheng. 2026. Better Goals, Better Policies: LLM-Driven Relabeling for Offline Goal-Conditioned Reinforcement Learning. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/ENPN5664

1 INTRODUCTION

Offline goal-conditioned reinforcement learning (offline GCRL) learns goal-conditioned policies from fixed, reward-free datasets, reducing the need for expensive online interaction and reward engineering. Existing work largely focuses on improving value estimation[3, 5, 6, 15, 19] and representation learning[4, 16, 20]. However, goal selection itself is often treated as a default choice: in practice, goals are commonly sampled from future states via hindsight experience replay (HER) [1, 14], which treats all future states uniformly and thus generates many uninformative goals with weak learning signals. Such inappropriate goals can bias value estimation and ultimately degrade policy quality. This motivates a central question: *How can we select informative goals from fixed offline data to strengthen offline GCRL?*

Recent studies suggest that large language models (LLMs) display human-like patterns of judgment and reasoning ([2, 9]). Based on this capability, prior work has applied LLMs to sequential decision making, predominantly as reward designers([18]), high-level planners([13]), or instruction-following agents([11, 12, 17]). However, using LLMs to structure offline trajectories for goal selection—i.e., to pinpoint informative, semantically grounded relabeling goals—has received little attention.

In this work, we propose LLM-Driven Relabeling, a plug-and-play framework that uses LLM-generated semantic rules to identify key states in offline trajectories and prioritize them as informative relabeling goals, with a self-check mechanism to ensure reliability. We provide theoretical justification for why such goals improve learning efficiency, and corroborate the benefits empirically on offline GCRL benchmarks.

2 METHOD

2.1 Framework

We present LLM-Driven Relabeling, a plug-and-play relabeling framework for offline GCRL that selects informative goals from fixed trajectories by leveraging LLM-generated semantic rules. The framework consists of four components, as shown in Fig. 1.

(1) Semantic Key-State Rule Generation: Given a task description, we prompt an LLM to generate a small set of semantic rules that characterize key states relevant for task completion (e.g., in PointMaze, *Turning Event*, *Terminal Stabilization*, *Stop-at-Checkpoint*). Each rule specifies conditions on a subset of state dimensions, serving as a high-level semantic prior.

(2) Rule Compilation into Executable Discriminators: The generated rules are compiled into deterministic key-state discriminator functions, which evaluate whether a given state satisfies

*Corresponding author.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/ENPN5664

a semantic condition. This compilation amortizes LLM cost and ensures efficient detection.

(3) Key-State–Prioritized Goal Relabeling: During relabeling, goals are preferentially sampled from future states that satisfy at least one key-state discriminator within a temporal window. When no key state is detected, the method falls back to standard future-state relabeling.

(4) Offline GCRL Training: The relabeled transitions are used to train a goal-conditioned policy and value function under a standard offline GCRL setting. By only altering the goal relabeling strategy, LLM-Driven Relabeling serves as a lightweight, plug-and-play module for offline GCRL.

To ensure reliable relabeling, we introduce a lightweight self-check mechanism. A compile-time check calibrates discriminator executability, while a relabel-time check monitors relabel quality using the agent's TD error as a proxy and adjusts goal selection accordingly.

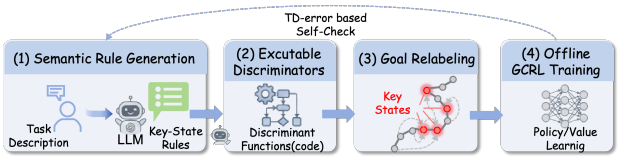

Figure 1: The framework of our LLM-Driven Relabeling.

2.2 Theory

Let δ_θ denote the goal-conditioned TD error induced by a relabeled transition. We compare two relabeling strategies on the same offline dataset: (i) HER Relabeling, and (ii) our LLM-Driven Relabeling.

Rationale. HER samples goals by randomly selecting future states, which treats all future states uniformly and often produces easy (low-signal) goals. In contrast, our LLM-driven Relabeling uses semantic knowledge to identify task-critical key states as candidate goals, increasing the possibility of selecting challenging, task-progressive goals and thus inducing larger TD errors and more informative updates.

Theorem 2.1 (Larger TD error implies smaller GOAL-BE dimension). *Under the ϵ-independence condition in [21], if the relabeled goals produced by LLM-Driven Relabeling induce a larger expected TD error than HER Relabeling, i.e.,*

$$\mathbb{E}\left[\delta_\theta^{LLM}\right] \geq \mathbb{E}\left[\delta_\theta^{HER}\right],$$

then the induced GOAL-BE dimension is no larger:

$$\dim_{\text{GOAL-BE}}^{LLM} \leq \dim_{\text{GOAL-BE}}^{HER}.$$

Corollary 2.2 (Reduced sample complexity). *Under the same assumptions as Theorem 2.1, finite-sample guarantees for offline GCRL that scale with the GOAL-BE dimension [21] imply that achieving error ε requires*

$$O\left(\frac{\dim_{\text{GOAL-BE}}}{\varepsilon^2}\right)$$

relabeled samples. Since $\dim_{\text{GOAL-BE}}^{LLM} \leq \dim_{\text{GOAL-BE}}^{HER}$, LLM-Driven Relabeling reduces the sample complexity relative to HER Relabeling.

3 EXPERIMENTS

Benchmark and baselines. We evaluate on OGBench [14] and focus on two representative environments: **AntMaze-large** (navigation) and **Cube-single-play** (manipulation). We compare against five widely used offline GCRL methods—GCIVL [15], GCIQL [8], QRL [16], CRL [4], and HIQL [15]—covering major paradigms such as implicit value learning [7], contrastive representation learning [4], and hierarchical goal decomposition [10].

TD-error validation. To verify the error-dominance premise in Theorem 2.1, we conduct controlled comparisons of TD errors induced by LLM-selected goals and HER goals on the same offline data. As shown in Fig. 2, LLM-Driven Relabeling yields significantly larger TD errors in sparse-reward settings. This supports the assumption $\mathbb{E}[\delta_\theta^{\text{LLM}}] \geq \mathbb{E}[\delta_\theta^{\text{HER}}]$ used in our analysis.

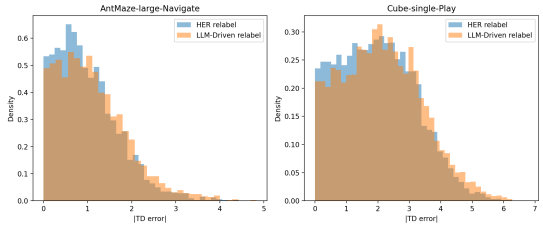

Figure 2: The distribution of TD errors.

Reduced-data setting. To evaluate sample efficiency, we train with only 50% of the offline dataset while keeping the total training steps fixed at 500,000. As shown in Fig. 3, LLM-Driven Relabeling consistently mitigates the performance drop under reduced data and narrows the gap to baselines trained on the full dataset. These results support our analysis that prioritizing semantically meaningful goals improves learning efficiency and reduces sample complexity.

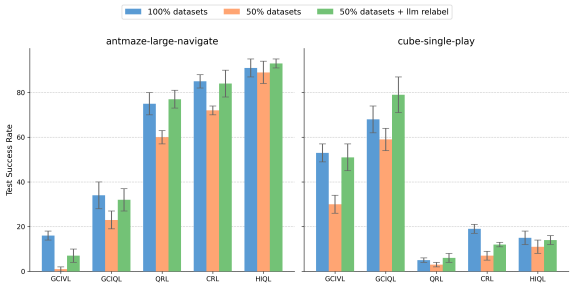

Figure 3: Results on limited data.

4 CONCLUSION

We propose LLM-Driven Relabeling to improve goal selection in offline GCRL by leveraging LLM-generated rules. Both theoretical analysis and empirical results demonstrate that prioritizing informative goals leads to improved sample efficiency.

REFERENCES

[1] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. 2018. Hindsight Experience Replay. arXiv:1707.01495 [cs.LG] https://arxiv.org/abs/1707.01495

[2] Marcel Binz and Eric Schulz. 2023. Turning large language models into cognitive models. arXiv:2306.03917 [cs.CL] https://arxiv.org/abs/2306.03917

[3] Yiming Ding, Carlos Florensa, Mariano Phielipp, and Pieter Abbeel. 2019. Goal-conditioned Imitation Learning. *Advances in Neural Information Processing Systems* (2019). http://arxiv.org/abs/1906.05838

[4] Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Ruslan Salakhutdinov. 2022. Contrastive Learning as Goal-Conditioned Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=vGQiU5sqUe3

[5] Joey Hejna, Jensen Gao, and Dorsa Sadigh. 2023. Distance Weighted Supervised Learning for Offline Interaction Data. arXiv:2304.13774 [cs.LG] https://arxiv.org/abs/2304.13774

[6] Junsu Kim, Younggyo Seo, and Jinwoo Shin. 2021. Landmark-Guided Subgoal Generation in Hierarchical Reinforcement Learning. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=IWhFd34QSSj

[7] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2021. Offline Reinforcement Learning with Implicit Q-Learning. arXiv:2110.06169 [cs.LG] https://arxiv.org/abs/2110.06169

[8] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2022. Offline Reinforcement Learning with Implicit Q-Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=68n2s9ZJWF8

[9] Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus* 3, 7 (07 2024), pgae233. https://doi.org/10.1093/pnasnexus/pgae233 arXiv:https://academic.oup.com/pnasnexus/article-pdf/3/7/pgae233/58651606/pgae233.pdf

[10] Andrew Levy, George Konidaris, Robert Platt, and Kate Saenko. 2019. Learning Multi-Level Hierarchies with Hindsight. arXiv:1712.00948 [cs.AI] https://arxiv.org/abs/1712.00948

[11] Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. 2024. Optimus-1: Hybrid Multimodal Memory Empowered Agents Excel in Long-Horizon Tasks. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. http://papers.nips.cc/paper_files/paper/2024/hash/5949a8750a110ce1f0631b1776c500a2-Abstract-Conference.html

[12] Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. 2025. Optimus-2: Multimodal Minecraft Agent with Goal-Observation-Action Conditioned Policy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 9039–9049. https://doi.org/10.1109/CVPR52734.2025.00845

[13] Shaoteng Liu, Haoqi Yuan, Minda Hu, Yanwei Li, Yukang Chen, Shu Liu, Zongqing Lu, and Jiaya Jia. 2024. RL-GPT: Integrating Reinforcement Learning and Code-as-policy. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=LEzx6QRkRH

[14] Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. 2025. OGBench: Benchmarking Offline Goal-Conditioned RL. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=M992mjgKzI

[15] Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. 2023. HIQL: Offline Goal-Conditioned RL with Latent States as Actions. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=cLQCCtVDuW

[16] Tongzhou Wang, Antonio Torralba, Phillip Isola, and Amy Zhang. 2023. Optimal Goal-Reaching Reinforcement Learning via Quasimetric Learning. In *International Conference on Machine Learning*. PMLR.

[17] Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, Xiaojian Ma, and Yitao Liang. 2025. JARVIS-1: Open-World Multi-Task Agents With Memory-Augmented Multimodal Language Models. *IEEE Trans. Pattern Anal. Mach. Intell.* 47, 3 (2025), 1894–1907. https://doi.org/10.1109/TPAMI.2024.3511593

[18] Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. 2024. Text2Reward: Reward Shaping with Language Models for Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=tUM39YTRxH

[19] Yaocheng Zhang, Yuanheng Zhu, Yuqian Fu, Songjun Tu, and Dongbin Zhao. 2025. Offline Goal-Conditioned Reinforcement Learning with Elastic-Subgoal Diffused Policy Learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems* (Detroit, MI, USA) *(AAMAS '25)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2336–2344.

[20] Chongyi Zheng, Benjamin Eysenbach, Homer Rich Walke, Patrick Yin, Kuan Fang, Ruslan Salakhutdinov, and Sergey Levine. 2024. Stabilizing Contrastive RL: Techniques for Robotic Goal Reaching from Offline Data. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=Xkf2EBj4w3

[21] Sirui Zheng, Chenjia Bai, Zhuoran Yang, and Zhaoran Wang. 2024. How Does Goal Relabeling Improve Sample Efficiency?. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 61246–61266. https://proceedings.mlr.press/v235/zheng24a.html