강화학습 기반 자율주행 성능 향상을 위한 보상 설계 기반 LLM 주도 의미적 가이던스

LLM-driven Semantic Guidance for Enhancing Reinforcement Learning-based Autonomous Driving through Reward Engineering

○Ahmad Mouri Zadeh Khaki[1], 최 경 환[2*]
[1)] 한국과학기술원 기계기술연구소 (TEL: 42-350-1764; E-mail: ahmadmouri@kaist.ac.kr)
[2)] 한국과학기술원 조천식모빌리티대학원 (TEL: 42-350-1764; E-mail: kh.choi@kaist.ac.kr)

Abstract Proximal Policy Optimization (PPO) is widely used for autonomous driving but often faces sparse rewards and unsafe exploration, which are critical issues in safety-sensitive scenarios. We propose LLM-PPO Driver, a framework that improves PPO by using a Large Language Model (LLM) as a source of prior driving knowledge during training. The LLM supports learning through reward shaping and imitation learning, without participating in real-time control. Experiments in the Gym highway-v0 environment show that both methods outperform baseline PPO, with imitation learning achieving the best results. These findings demonstrate that LLM-guided prior knowledge can improve safety, task success, and learning efficiency in autonomous driving.

Keywords Autonomous driving, imitation learning, large language model (LLM), proximal policy optimization, reward shaping

1. Introduction

Autonomous driving is a safety-critical sequential decision-making problem that requires reliable perception and control in dynamic environments [1]. Reinforcement learning (RL) has shown strong potential for this task by learning driving policies directly through interaction [2]. Among RL methods, Proximal Policy Optimization (PPO) is widely used because of its stability and effectiveness in continuous control [3].

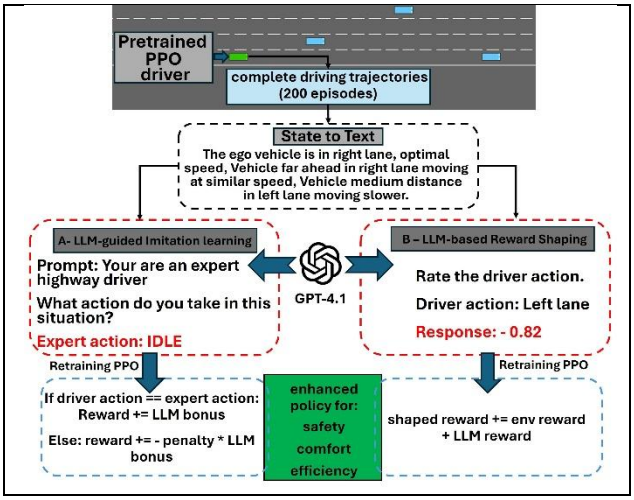

Fig. 1. The LLM-PPO Driver Framework Overview.

Yet, PPO faces challenges in autonomous driving due to sparse rewards, slow convergence, and unsafe exploration. These issues are especially serious because rare events such as collisions can have severe consequences [4]. To improve PPO training, prior knowledge can be introduced through reward shaping and imitation learning [5]. Large Language Models (LLMs) contain rich semantic knowledge of traffic rules and driving behavior [6]. Although unsuitable for real-time control because of latency [7], LLMs can provide useful guidance during training.

We propose LLM-PPO Driver, a framework that enhances PPO using LLM-based prior knowledge in two separate ways: (1) reward shaping with semantic feedback and (2) imitation learning from expert-like demonstrations. This design improves learning efficiency without adding deployment overhead.

2. Methodology

The proposed framework extends PPO by using an LLM only during training as a source of expert knowledge. The LLM is not used during deployment, avoiding inference overhead. A baseline PPO agent first collects trajectories, which are converted into textual driving descriptions. The LLM is then used in two independent settings: (1) imitation learning, where it recommends expert actions, and (2) reward shaping, where it evaluates executed actions. PPO is retrained using these signals.

2.1 State-to-Text Translation

Numerical sensor data (positions, velocities) are converted into structured natural language. For example, “The ego vehicle is in the right lane at optimal speed with a slow vehicle far ahead”.

2.2 LLM-Based Imitation Learning

The LLM acts as an expert advisor by selecting the most suitable action for each state. These expert actions are

※ This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS2025-00554087).

stored offline and compared with PPO actions during training. Matching actions receive a bonus, while mismatches receive a penalty as shown in (1). This guidance is strongest early in training and gradually decays, allowing the agent to rely more on environment rewards over time.

$$r_t = r_t^{env} + \beta\, \mathbb{I}(at = a_t^{expert}) - \quad \beta\alpha\, \mathbb{I}(at \neq a_t^{expert}) \qquad (1)$$

2.3 LLM-Guided Reward Shaping

The LLM functions as a semantic critic that evaluates the quality of executed actions using state, action, and reward information. It outputs a score in [-1,1], representing poor to strong driving behavior. This score is added to the environment reward as auxiliary feedback as expressed in (2).

$$r_t = r_t^{env} + \gamma\, r_t^{LLM}, \qquad (2)$$

3. Performance Evaluation

3.1 Simulation Environment and Performance Metric

Performance is evaluated using Survival Steps (SS) and Success Rate (SR) in gym highway-v0 environment. Each episode lasts up to 30 steps or ends earlier after a collision. SS measures how long the agent drives safely, while SR represents the percentage of collision-free episodes.

Table 1. Simulation Results of the Proposed Framework.

	Baseline	Reward Shaping (γ=0.3)	Imitation Learning (β=0.3, α=0.2)
SS_{min}	2.00	6.00	8.00
SS_{mean}	25.16	27.22	27.85
SR%	70.00	76.00	82.00

3.2 Simulation Results

According to Table 1, Both LLM-guided methods outperform baseline PPO in safety and task completion. For survival performance, the baseline shows weak worst-case robustness (SS_{min} = 2), while reward shaping and imitation learning improve SSmin to 6 and 8, respectively, significantly reducing early collisions. Baseline PPO achieves SS_{mean} = 25.16 and SR = 70%. Reward shaping improves performance to SS_{mean} = 27.22 and SR = 76%, while imitation learning achieves the best results with SS_{mean} = 27.85 and SR = 82%. These results show that LLM-derived guidance significantly reduces collisions and improves robustness in dense highway traffic. Imitation learning provides larger gains than reward shaping because direct action supervision better aligns policy behavior.

The ablation results in Fig. 2 show that both the imitation learning coefficient α and reward shaping coefficient γ have optimal ranges. Excessive values reduce performance by limiting exploration and adaptability. These findings suggest that LLM guidance is most effective as a complementary rather than dominant training signal.

4. Conclusion

This paper proposed a lightweight framework that enhances PPO-based autonomous driving using offline LLM guidance through imitation learning and reward shaping. The method preserves PPO stability and requires no real-time LLM inference. Experimental results demonstrated improved safety and task completion over baseline PPO. Imitation learning achieved the strongest safety gains, while reward shaping provided more stable improvements across parameter settings, indicating a trade-off between peak performance and robustness. Future work will explore combining both strategies and validating the framework in more realistic driving simulators.

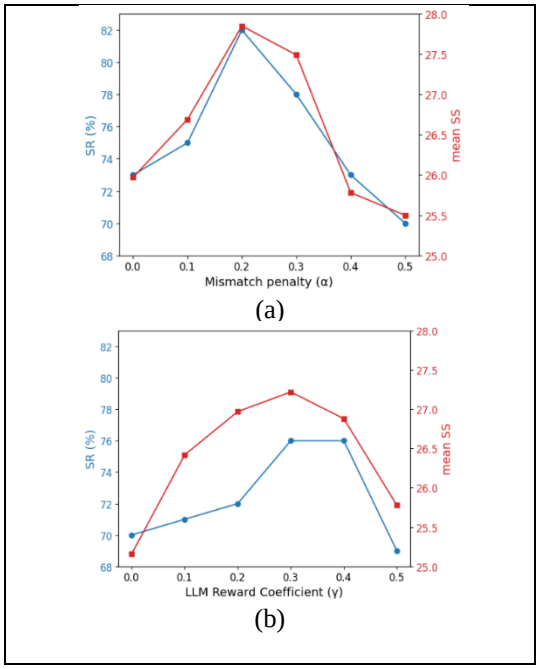

Fig. 2. Ablation analysis of semantic guidance strength.

References

[1] L. Chen et al., "Milestones in Autonomous Driving and Intelligent Vehicles: Survey of Surveys," *in IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1046-1056, Feb. 2023.

[2] S. Aradi, "Survey of Deep Reinforcement Learning for Motion Planning of Autonomous Vehicles," *in IEEE Transactions on Intelligent Transportation Sys.*, vol. 23, no. 2, pp. 740-759, Feb. 2022.

[3] B. Kim, K. Kim, S. Kim, Y. Shin and H. Ahn, "Domain-Agnostic Scalable AI Safety Ensuring Framework," arXiv preprint arXiv:2504.20924 [cs.AI], 2025.

[4] B. R. Kiran et al., "Deep Reinforcement Learning for Autonomous Driving: A Survey," *in IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909-4926, June 2022.

[5] M. Hallgarten, M. Stoll and A. Zell, "From Prediction to Planning With Goal Conditioned Lane Graph Traversals," *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 951-958, 2023.

[6] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi and Y. Qiao, "Drive like a human: Rethinking autonomous driving with large language models," *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 910–919, 2024.

[7] Z. Xu et al., "DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model," *in IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 8186-8193, Oct. 2024.