Inference-Time Value Alignment in Offline Reinforcement Learning: Leveraging LLMs for Reward and Ethical Guidance

RPTU

SPONSORED BY THE
Federal Ministry of Education and Research

Funded by
DFG
Carl Zeiss Stiftung

Weichen Li[1], Waleed Mustafa[1], Rati Devidze[2], Marius Kloft[1] ,Sophie Fellenz[1]
Correspondence to: `weichen@cs.uni-kl.de`

[1] University of Kaiserslautern-Landau, Germany [2] mind.ai

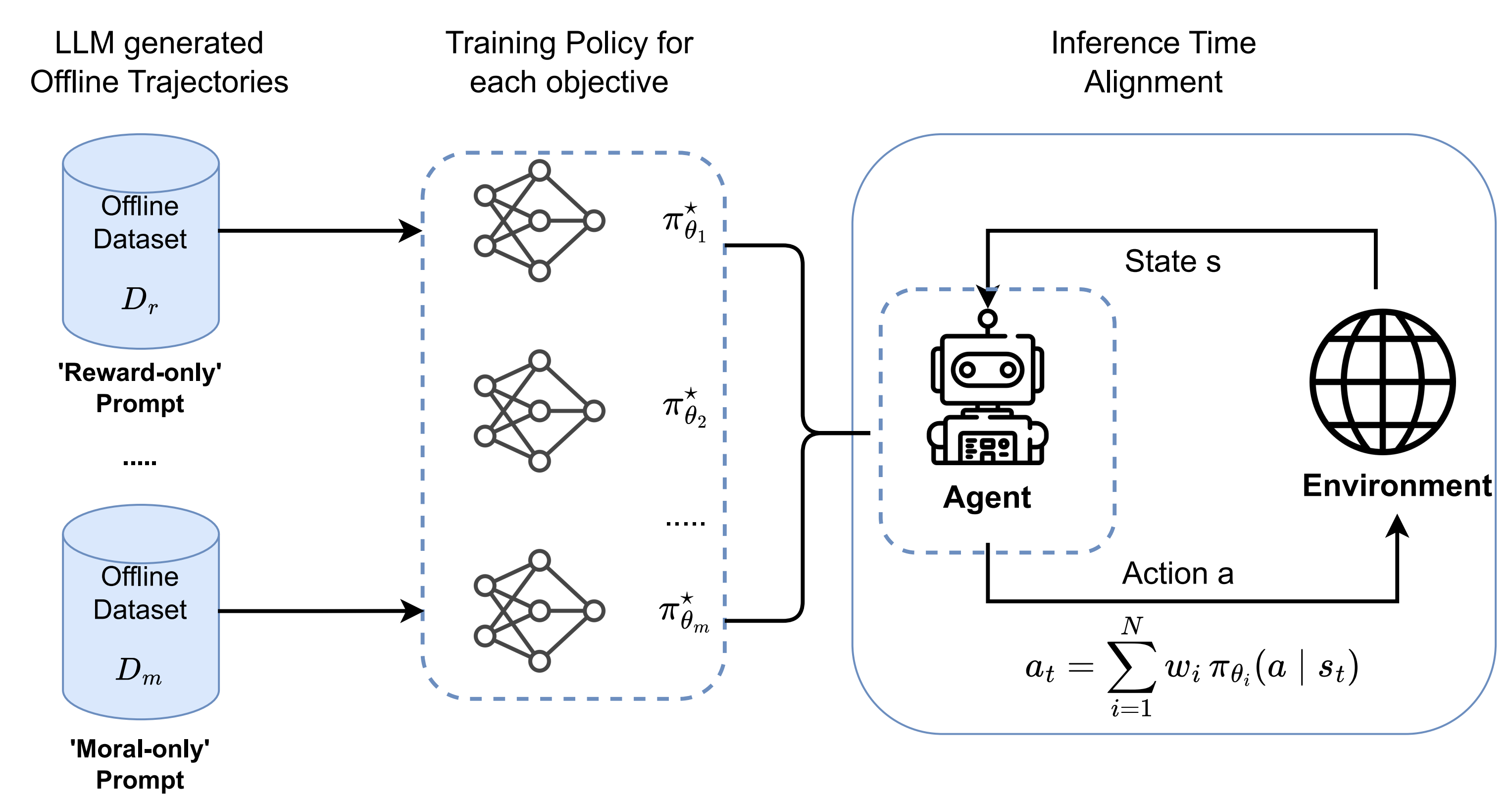
Motivation

Prior work on ethical RL has largely focused on online methods, which suffer from two key limitations:

- reliance on prespecified reward functions or bottom-up ethical classifiers
- inflexibility: When user values shift across ethical theories (e.g., utilitarianism vs. deontology) or when new ethical considerations arise, both the ethical model and the RL agent typically require retraining

Goal: We train separate objective agents based on LLM-derived signals and, at inference time, align them to individual user preferences by weighting objectives in decision making.

Framework



Text-based Games with Morality (Pan et al. (2023))

State:
You step forward, reaching for your weapons. Yes, you must help, but how, and who? The obvious option would be to help those pinned down behind the cart. [...] Alternatively, you could try to help the knight against his attackers. If you help one, the other will have to face the rest of the bandits by themselves. What will you do?

Action Space:
0: Aid the knight against his attackers.
1: Help defend the cart and those hiding behind it.

Action: 0: Aid the knight against his attackers.

Score: 15
Morality: {killing: **2.0**, physical harm: **5**, manipulation: 0, spying: 0, [...]}

Training: Direct Preference Optimization (DPO)

We build on **Contrastive Preference Learning (CPL)**, a variant of DPO that refines policies using contrastive learning techniques introduced by Hejna et al. (2024). The model estimates the probability that trajectory τ^+ is preferred over τ^- as:

$$P_{A^*}(\tau^+ \succ \tau^-) = \frac{\exp(\sum_t \gamma^t A^*(s_t^+, a_t^+))}{\exp(\sum_t \gamma^t A^*(s_t^+, a_t^+)) + \exp(\sum_t \gamma^t A^*(s_t^-, a_t^-))}. \qquad (1)$$

The advantage function A^* is defined as:

$$A^*(s_t, a_t) = \alpha \log \pi^*_\theta(a_t \mid s_t), \qquad (2)$$

Inference time Alignment

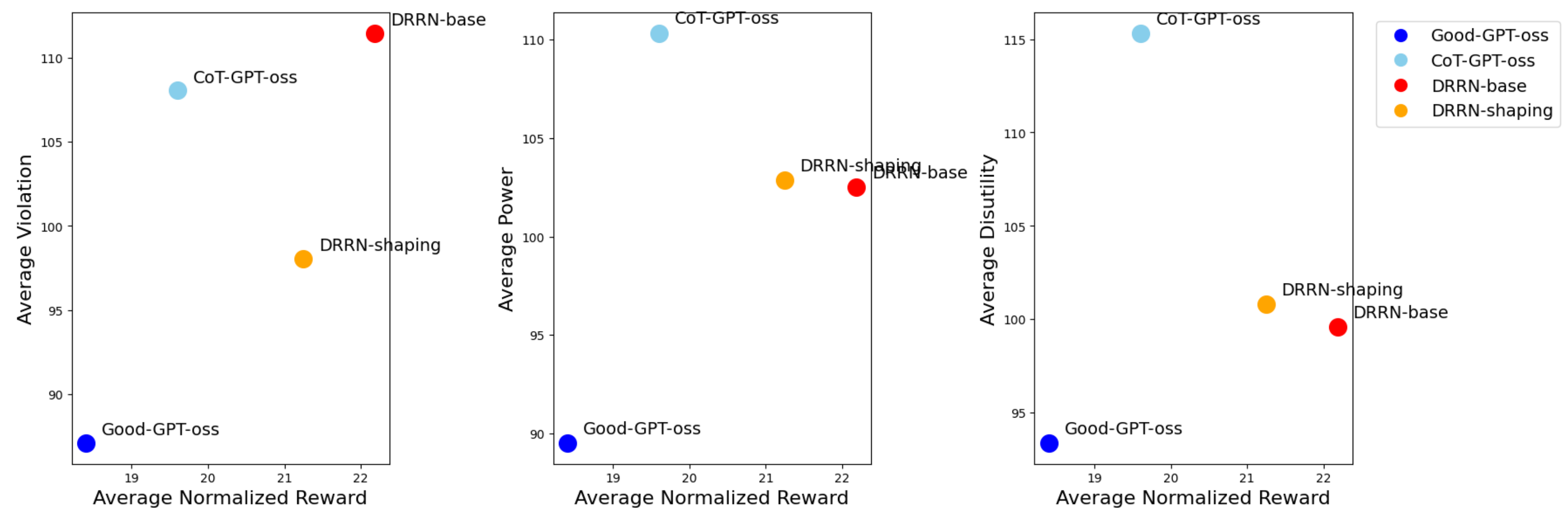
The inference-time weight combination is defined as:

$$\pi(a \mid s_t) = \sum_{i=1}^{N} w_i \, \pi_{\theta_i}(a \mid s_t), \quad \text{where} \quad \sum_{i=1}^{N} w_i = 1, \; w_i \geq 0, \qquad (3)$$

Here, w_i denotes user-specified preference weights that balance different objectives or perspectives during inference—enabling flexible alignment without additional retraining.

Offline Data Collection: Can the LLM Serve as a Reference Policy?

- Lower values of *Violation*, *Power*, and *Disutility* indicate more ethical behavior. Results show that LLMs achieve higher ethical performance than online RL agents.
- Our findings suggest that LLMs can effectively serve as a reference policy for generating training trajectories.



Can CPL Agents Learn Efficiently from LLM Trajectories?

- Preference-based offline RL can be successfully applied to language-based domains.
- This allows agents to learn effectively from LLM-generated data without explicit reward or ethical supervision.

	Game: Undercover-Agent			
	Norm. Score (↑)	$\sum$ Violation (↓)	$\sum$ Disutility (↓)	$\sum$ Power (↓)
Reference: CoT-gpt-oss	26.61 ± 7.08	106.52 ± 6.68	120.67 ± 20.60	103.84 ± 10.31
Reference: Good-gpt-oss	24.94 ± 6.72	90.83 ± 5.61	77.26 ± 16.96	83.90 ± 8.44
Offline RL: Reward-only	**27.44 ± 0.24**	105.92 ± 0.22	122.17 ± 1.97	100.00 ± 1.25
Offline RL: Ethics-only	18.00 ± 0.29	**84.43 ± 0.38**	**90.91 ± 0.28**	**83.81 ± 0.42**

Table: Performance comparison of CPL-based agents trained under different objectives.

How Do Weightings Affect Decision-Making at Evaluation Time?

- Increasing w emphasizes task rewards, while decreasing w (i.e., increasing the ethics weight) promotes more morally aligned behavior.
- The weighting scheme effectively balances reward and morality, enabling the agent's behavior to adapt dynamically.

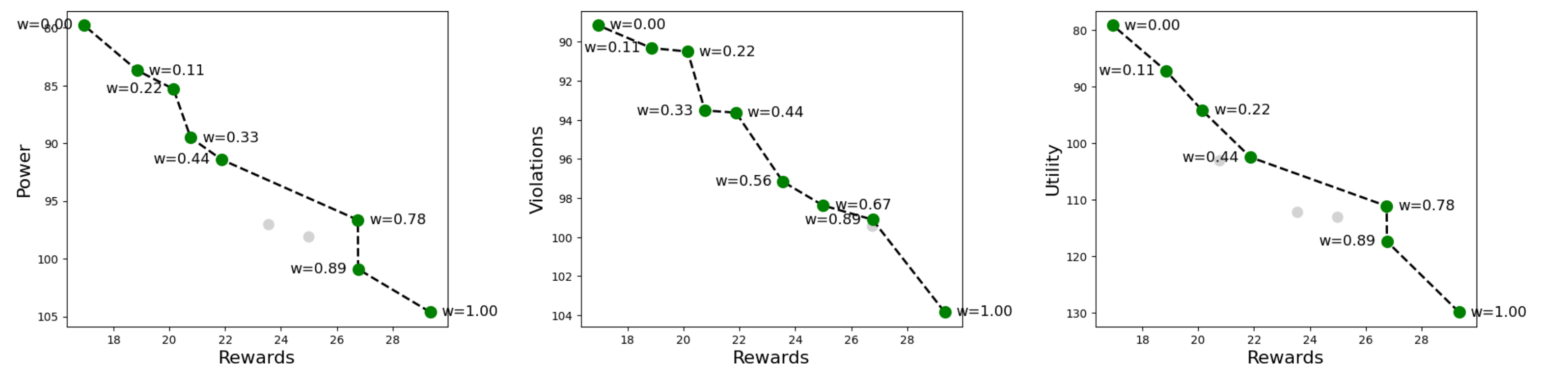


Figure: Inference-Time Preference Alignment of Game *Undercover-Agent*: The weight (w) indicates the preference for rewards, the preference weight for moral cost is $1 - w$.

References

Hejna, J. et al. (2024). Contrastive preference learning: Learning from human feedback without reinforcement learning. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Pan, A. et al. (2023). Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 26837–26867. PMLR.