HW2: Reinforce LLM Reasoning through Multi-Agent Reflection

Eunhwi Lee (lehlsy0904@snu.ac.kr)

Computational Clinical Science Laboratory, Seoul National University

Reinforce Large Language Model (LLM) reasoning through Multi-Agent Reflection presents Direct Policy Search by Dynamic Programming (DPSDP), a reinforcement learning (RL) framework designed to strengthen reasoning through iterative actor–critic collaboration. In this approach, the actor proposes candidate solutions, the critic delivers targeted feedback, and multiple refinement rounds progressively improve performance. By formalizing reasoning as a Markov Decision Process and optimizing policies with Direct Preference Optimization, DPSDP enables self-improvement without reliance on external supervision or static prompting strategies.

Empirical results show that DPSDP consistently enhances reasoning accuracy and, crucially, generalizes beyond training distributions. The work demonstrates that multi-agent specialization, restart-based data collection, and a reduced-context (Markovian) design are key to robust gains. Compared with sampling or self-consistency methods, DPSDP highlights the value of structured, feedback-driven refinement for advancing LLM reasoning.

The contribution of DPSDP lies less in incremental performance gains than in its structural insight: reasoning can be treated as a trainable, interactive process rather than a one-shot prediction. From a psychological perspective, this iterative loop of critique and revision parallels human error monitoring—suggesting how LLMs with reinforced reasoning might evolve into more reliable collaborators for scientific research, offering new ways to explore and model human cognition and behavior.