ICLR 2026

Adaptive Collaboration with Humans: Metacognitive Policy Optimization for Multi-Agent LLMs with Continual Learning

Wei Yang∗, Defu Cao∗, Jiacheng Pang, Muyan Weng, Yan Liu

University of Southern California

∗ Equal contribution

Multi-Agent Collaboration · Metacognitive Policy · Continual Learning

USC Viterbi School of Engineering

Table of Contents

USC Viterbi
School of Engineering

Introduction / Motivation

Limitation

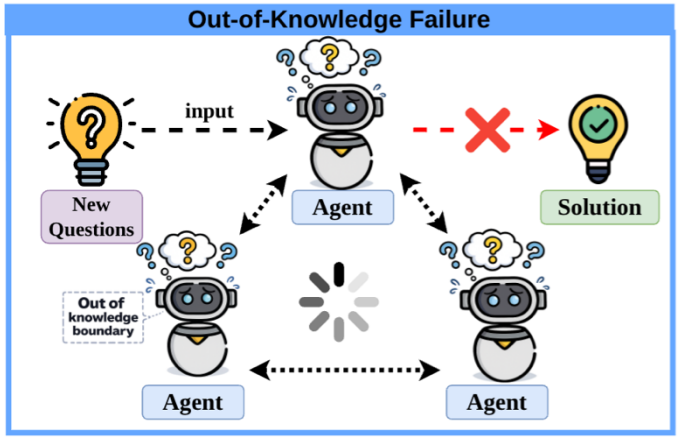
- Knowledge bounded by pre-training
- Cannot acquire new knowledge
- Struggle with unseen contexts

Basic Idea

- **Try first:** Autonomous agent collaboration
- **Know limits:** Detect knowledge boundaries
- **Ask & learn:** Improve from human feedback

Key Problems

- **When to defer:** Beyond simple confidence heuristics.
- **Learn from feedback:** Turn one-time fix into long-term improvement.



Research Direction: Bridging the gap between **closed-world limitations** and **open-world adaptability** through effective **human-in-the-loop learning**.

Method: HILA Framework

HILA: Human-In-the-Loop Multi-Agent Collaboration

Autonomous Operation

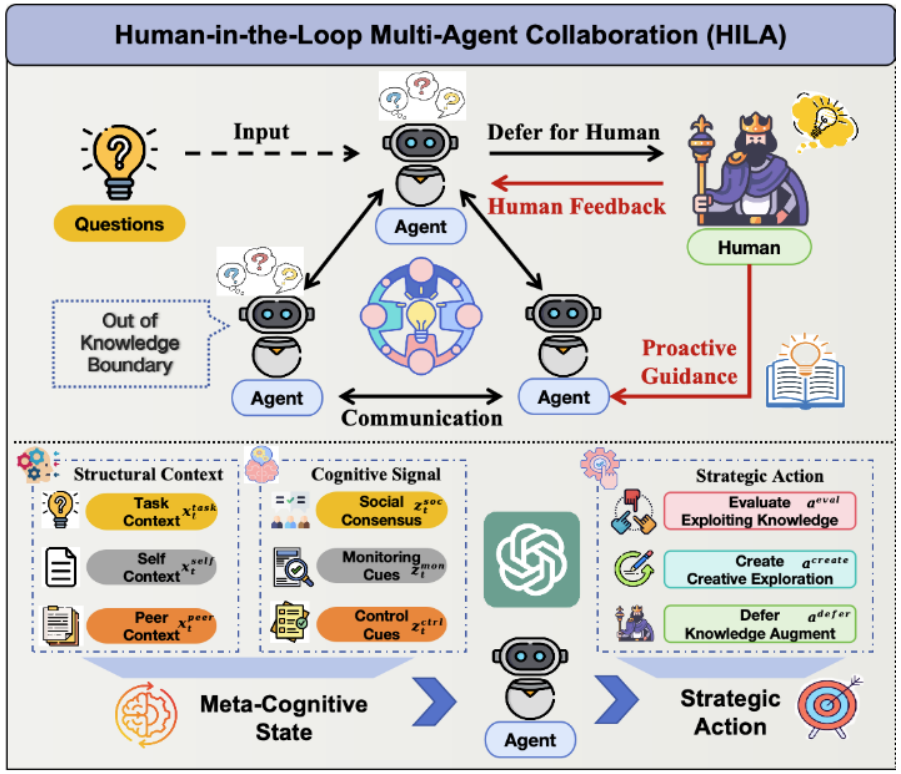
Agents attempt to solve problems independently through their own capabilities.

Metacognitive Assessment

Evaluate confidence level, task difficulty, and peer consensus.

Strategic Deferral

Invoke human expert only when necessary to ensure quality.



Metacognitive Action Space

EVAL

Exploit existing collective solutions and knowledge.

CREATE

Generate new solution paths to break fixation.

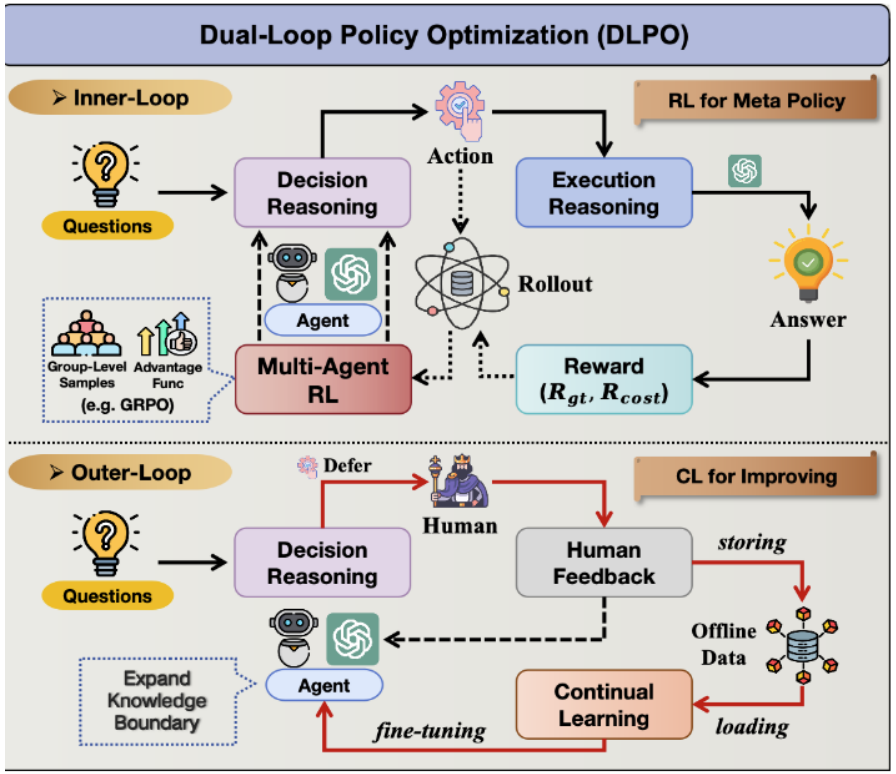
DEFER

Call human expert for knowledge augmentation & risk mitigation.

Method: Dual-Loop Policy Optimization (DLPO)

Inner Loop (RL)

- Optimize via **GRPO** (Group Relative Policy Optimization)
  - Advantage: $A(s_t, a_k) = R(s_t, a_k) - \frac{1}{K}\sum_{j=1}^{K} R(s_t, a_j)$
- Reward: Task correctness + **cost-aware penalty**
  - $$R(s_t, a_t) = \begin{cases} R_{gt}(\hat{y}), & EVAL \\ R_{gt}(\hat{y}) - C_{create}, & CREATE \\ R_{gt}(\hat{y}_{human}) - C_{defer}, & DEFER \end{cases}$$
- Objective: Learn **when to defer**

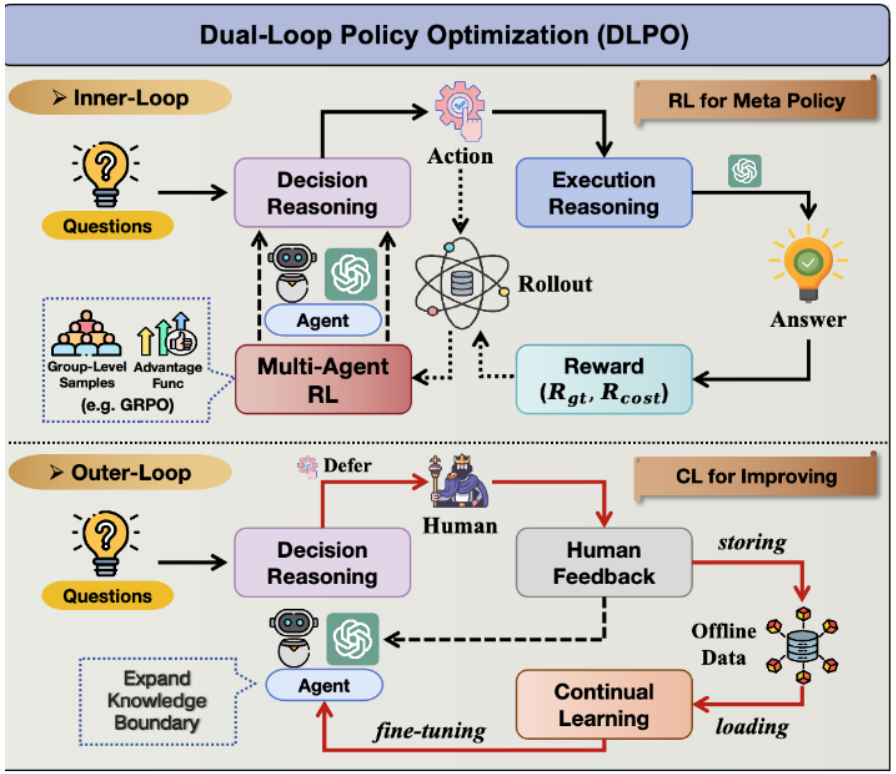


Method: Dual-Loop Policy Optimization (DLPO)

Outer Loop (Continual Learning)

- Use human feedback as **SFT samples**
- $L_{SFT} = -\log p_\theta(y_{human}|x)$

- Update backbone LLM to expand knowledge boundary
- Objective: Learn **what to learn** from humans



Total Objective Function

$$\mathcal{L}_{total} = \mathcal{L}_{Inner} + \lambda_{sft} \cdot \mathbb{1}(a=DEFER) \cdot \mathcal{L}_{SFT}$$

Experiments: Benchmarks & Key Results

Benchmarks

Math

GSM8K, AMC, AIME

Code

HumanEval

General

MMLU

Proxy expert

GPT-4o-mini, GPT-3.5-turbo, GPT-4o, Human

Model	Type	GSM8K	AMC	AIME	HumanEval	MMLU
Vanilla	SA	72.76 (+0.00)	8.03 (+0.00)	2.96 (+0.00)	47.56 (+0.00)	57.99 (+0.00)
CoT	SA	74.22 (+1.46)	11.65 (+3.62)	3.70 (+0.74)	51.42 (+3.86)	61.57 (+3.58)
SC	SA	80.79 (+8.03)	12.45 (+4.42)	4.07 (+1.11)	57.52 (+9.96)	68.30 (+10.31)
PHP	MA	80.01 (+7.25)	15.66 (+7.63)	4.44 (+1.48)	56.50 (+8.94)	68.46 (+10.47)
Debate	MA	83.52 (+10.76)	19.28 (+11.25)	5.56 (+2.60)	57.72 (+10.16)	67.59 (+9.60)
G-Debate	MA	83.98 (+11.22)	20.48 (+12.45)	5.19 (+2.23)	57.93 (+10.37)	69.89 (+11.90)
DyLAN	MA	82.03 (+9.27)	19.68 (+11.65)	3.70 (+0.74)	61.59 (+14.03)	66.85 (+8.86)
G-Swarm	MA	84.89 (+12.13)	15.66 (+7.63)	5.78 (+2.82)	59.55 (+11.99)	69.67 (+11.68)
A-Prune	MA	84.38 (+11.62)	16.47 (+8.44)	4.81 (+1.85)	57.11 (+9.55)	69.09 (+11.10)
AFlow	MA	83.75 (+10.99)	12.05 (+4.02)	4.44 (+1.48)	62.20 (+14.64)	69.31 (+11.32)
HILA	MA	**89.86** (+17.10)	**35.83** (+24.47)	**9.37** (+6.41)	**72.15** (+24.59)	**73.62** (+15.63)

HILA(DLPO) **outperforms all baselines** on all benchmarks.

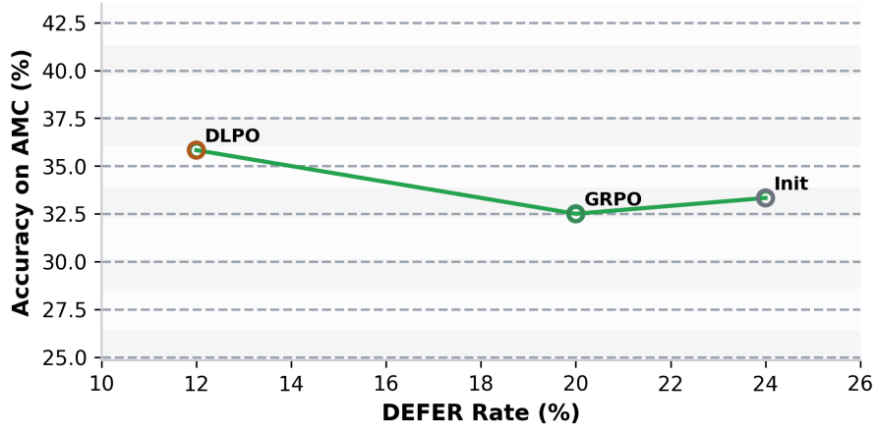
Strong gains on **challenging math tasks** (AMC: 8.03 → 35.83).

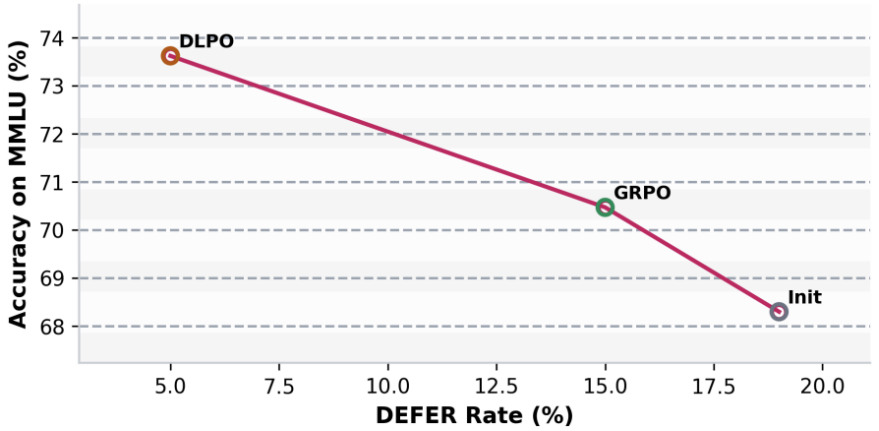
Experiments: Ablation Study

Model	GSM8K	AMC	MMLU
HILA (Init Policy)	88.15	33.33	68.30
HILA + GRPO	88.38	32.50	70.47
HILA + DLPO	**89.86**	**35.83**	**73.62**

Ablation Study

- Full dual-loop objective delivers the strongest overall performance
- **DLPO** reduces unnecessary deferrals while improving accuracy





Conclusion: DLPO delivers clear improvement over Init & GRPO alone.

Key Findings

- **Strategic human deferral** enables more robust reasoning
- **Learning when to ask** for help is as important as improving reasoning ability
- **Human–AI collaboration** is a scalable paradigm for complex reasoning systems

Code / Link

GitHub Repository: https://github.com/USC-Melady/HILA.git

Published as: A conference paper at **ICLR 2026**