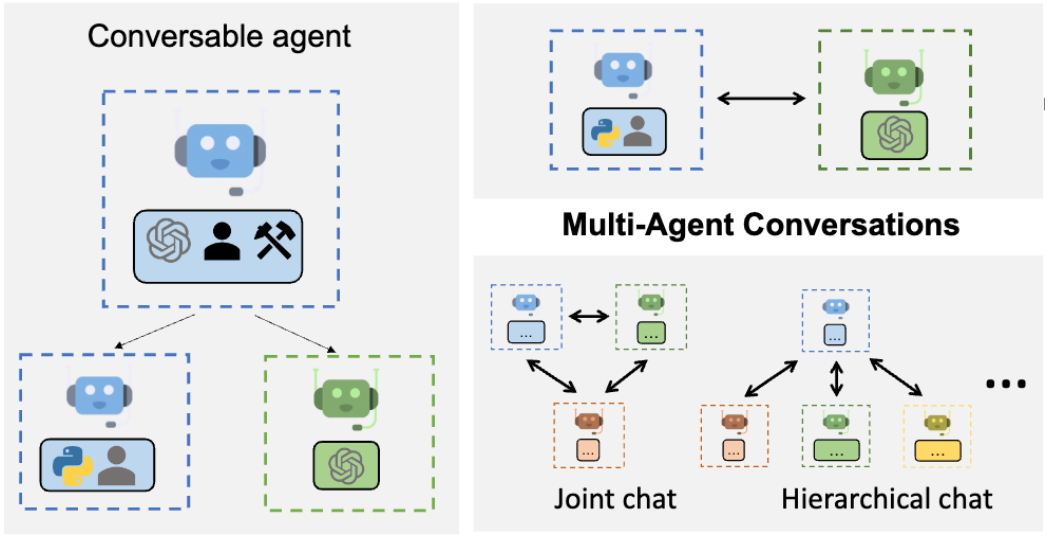
Learning Decentralized LLM Collaboration with Multi-Agent Actor Critic

Shuo Liu, Tianle Chen, Ryan Amiri, Christopher Amato

Background

- **LLMs are More Specialized**
  - **Nature:** Base / Instruct / MoE / Coder / Math / VL / Image / Embedding
  - ***Role (Prompt):*** *Generator / Verifier / Planner*
- **LLM as Agents (ReAct)**
  - Autonomous / Proactive
  - Tools Using / Function Calling
  - Self-Evolve / Multi-Turn

Multiple specialized LLM agents can interact!

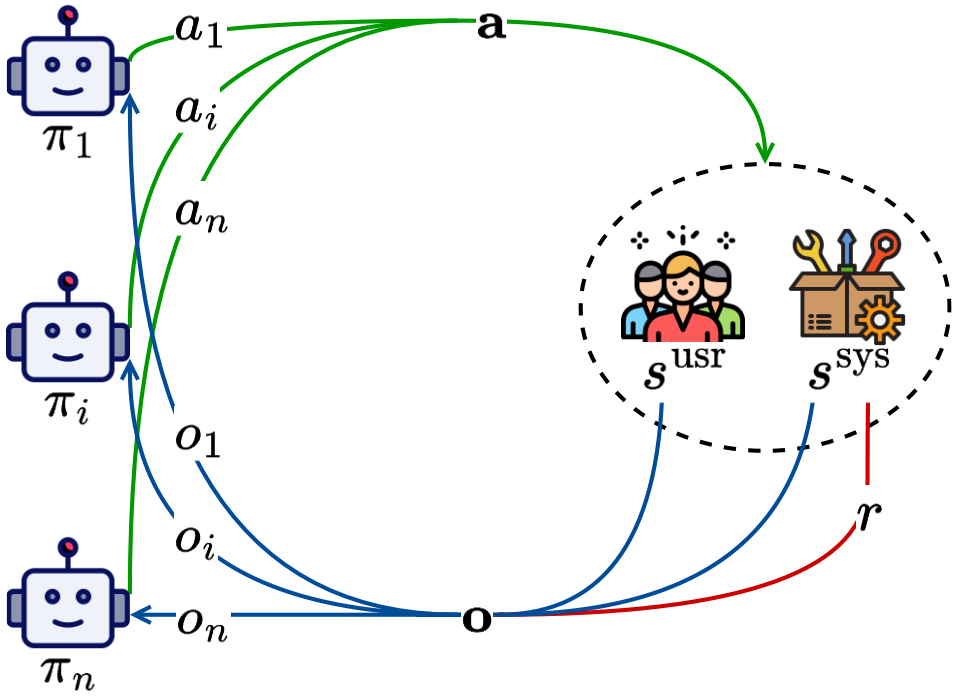


Related Work: LLM Collaboration

1. **Test-Time Interaction**
   - Prompt-based: Brittle Collaboration
   - Test-Time MARL
2. **MARL-tuned Coordination**
   - POSG Solution -- Nash Equilibrium: No Global Optimum Guarantee
   - Agent-wise / Role-based Reward Design: Complex; Onerous; Manual
3. **Agent-wise Parameter Sharing**
   - Self-Evolve / Improvement / Play: Objectives Conflict; Model Capacity

Decentralized LLM Collaboration

- **Formalization: LLM Dec-POMDP** $\langle \mathcal{I}, \mathcal{V}, C, M, \mathcal{S}, \{\mathcal{O}_i\}, \{\mathcal{A}_i\}, R, T, H \rangle$
  - Language (Vocabulary, Context Window, Max Outputs)
  - Sequence-Level Observation / Macro-Action
- **Advantages**
  - Numerous Agents Can Run in Parallel -> High Efficiency
  - Multi-SLM -> Flexible Deployment (Easy, Private, Scalable)
  - Specialized LLM Agents -> Better Performance
- **Special Challenges**
  - Credit Assignment
  - Limited/No Communication



LLM Dec-POMDP

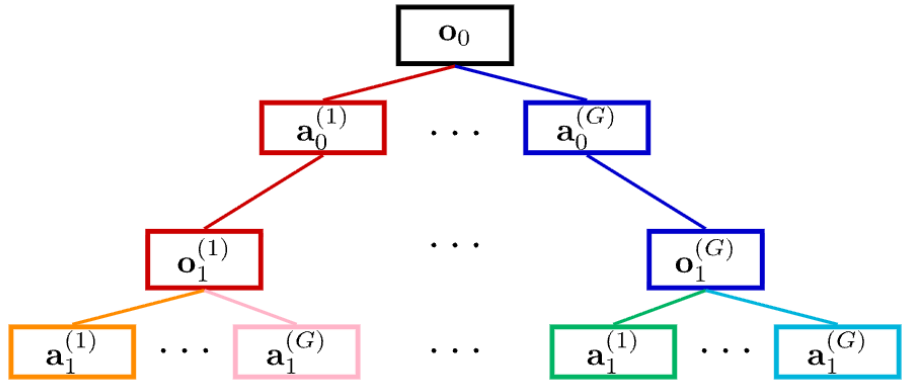
Multi-Agent REINFORCE

- **Definition** $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{\boldsymbol{\pi}} \left[\rho_{i,t} \nabla_{\theta_i} \log \pi_{\theta_i}(a_{i,t} \mid h_{i,t}) \left(G(\mathbf{h}_t) - b(\mathbf{h}_t) \right) \right]$
- **Properties**
  - Unbiased Estimator
  - Cannot Support Online Learning
  - High Variance (No Early Termination, I.I.D.)

    $\mathrm{Var}_{\boldsymbol{\pi}}(\bar{g}_{i,t} \mid h_{i,t}) = \dfrac{\sigma^2}{K^{H-t}}$

  - #Inference Calls (NET, I.I.D.): Low Sample Efficiency

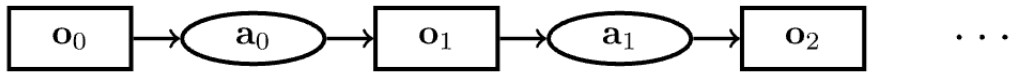
    $N_{\text{call}}(n, K, H) = \dfrac{nK(K^H - 1)}{K - 1}$



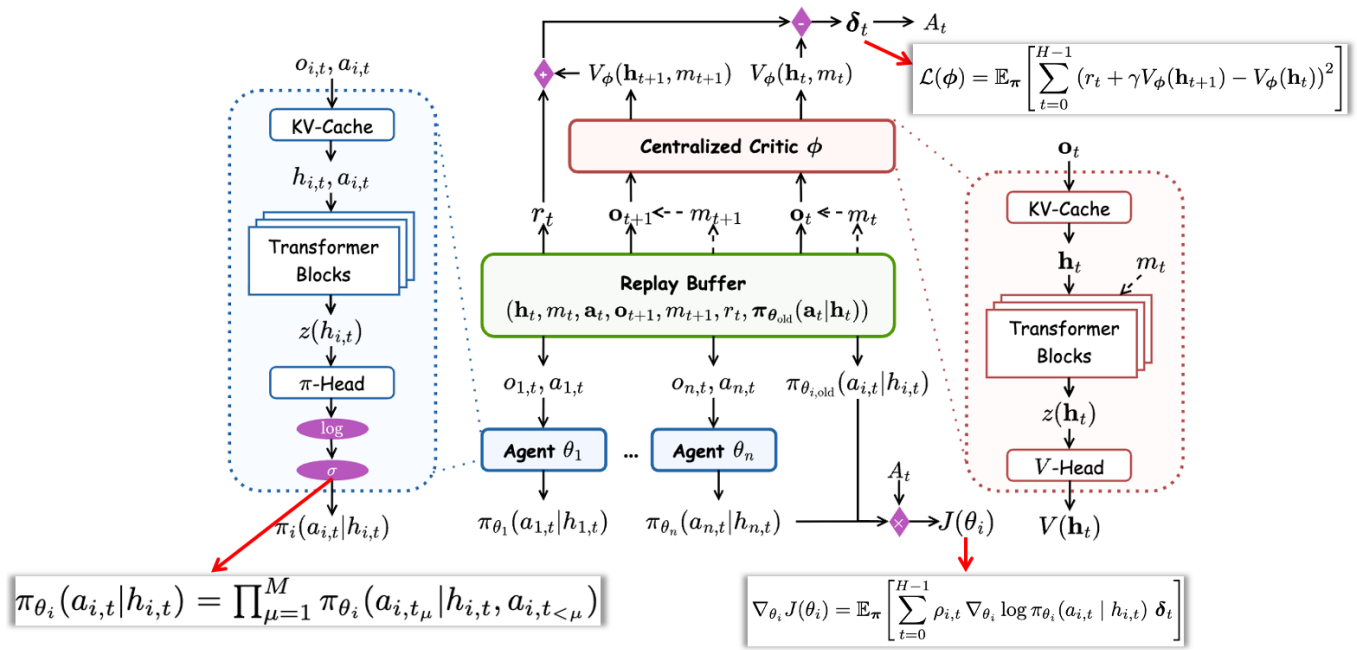
Rollout Tree

Multi-Agent Actor Critic

- **Definition** $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{t=0}^{H-1} \rho_{i,t} \nabla_{\theta_i} \log \pi_{\theta_i}(a_{i,t} \mid h_{i,t}) \, \delta_{i,t} \right]$ or $\boldsymbol{\delta_t}$
- **Properties**
  - Assuming Perfect Critics, Unbiased Estimator [Peshkin et. al.]
  - Support Online Learning
  - TD Difference: Low Variance
  - #Inference Calls: High Sample Efficiency



CoLLM-CC (Centralized Critic)



Experimental Settings

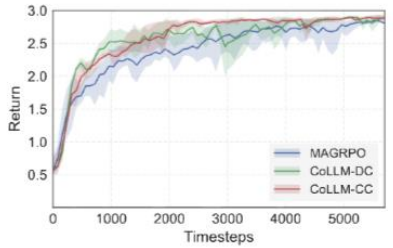
- **Writing Collaboration**
  - *Task 1: TLDR* Summarization
  - *Task 2: arXiv* Expansion
  - Agents: 2x *Qwen3-1.7B*
  - Metric: structure, consistency, coherence
  - Characteristics: Short-Horizon (H=1), Dense-Rewards
- **Coding Collaboration**
  - Dataset: HumanEval&MBPP -> CoopHumanEval
  - Agents: *Qwen2.5-Coder-3B* & *Qwen3-Coder-4B*
  - Metric: *Pass@K*
  - External: *AST*, Sandbox Tests [:1]
  - Characteristics: Short-Horizon (H=2), Sparse-Rewards
- **Game-Playing Collaboration**
  - *Task 1: StrBuild*
  - *Task 2: HouseBuild*
  - Agents: *Qwen3-4B-Instruct-2507* & *Qwen2.5-3B-Instruct*
  - Metric: Adjacency Rate, Health Point, IoU
  - External: Specialization Hints
  - Characteristic: Long-Horizon (H=4)

(a) StrBuild

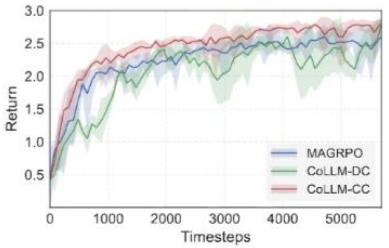
(b) HouseBuild

Results

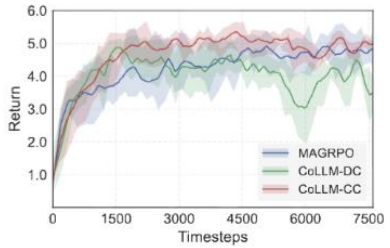
- **Short-horizon and dense-reward:**
  - MA-REINFORCE/CoLLM-DC ≈ CoLLM-CC
- **Long-horizon or sparse-reward:**
  - MA-REINFORCE < CoLLM-CC -- **low sample efficiency**
  - CoLLM-DC < CoLLM-CC -- **non-stationary**



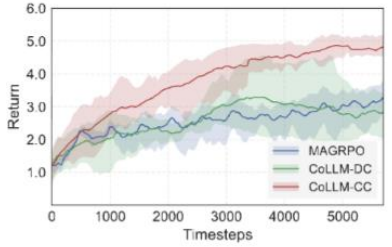
(a) Article Summarization | *TLDR*



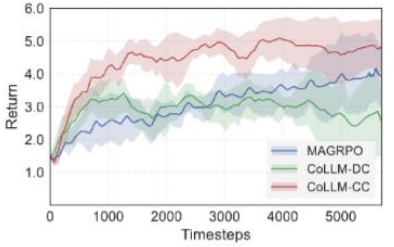
(b) Article Expansion | *ArXiv*



(c) Code Generation | *CoopHE*



(d) Minecraft Building | *StrBuild*



(e) Minecraft Building | *HouseBuild*

Method	TLDR			arXiv			CoopHE			StrBuild				HouseBuild			
	Time	Cost	Score	Time	Cost	Score	Time	Cost	Pass	Time	Cost	Adj	IoU	Time	Cost	HP	IoU
Raw Model	5.0	465	30.3	5.1	472	44.6	2.5	90	56.3	10.6	427	0.9	36.6	22.6	1016	99.6	43.2
GRPO	4.1	387	91.7	4.2	398	91.0	2.5	**88**	61.8	10.3	411	**0.4**	46.1	22.0	890	**100.0**	54.6
AC	4.0	374	94.5	4.3	392	**95.3**	2.5	91	62.5	10.3	413	**0.4**	49.8	22.1	904	**100.0**	**55.9**
Parallel	2.3	244	22.9	2.3	246	49.0	**2.3**	138	50.0	9.4	232	15.7	5.9	19.2	502	21.8	46.1
Pipeline	4.3	238	21.7	3.9	203	57.8	2.6	177	62.5	9.8	246	12.9	18.7	20.3	488	30.6	41.3
Discussion	4.6	234	22.3	4.8	251	54.3	2.9	191	25.0	10.3	236	16.2	6.5	21.0	510	27.6	38.1
MAGRPO	**1.8**	**178**	93.5	2.0	201	93.1	**2.3**	132	74.3	9.4	226	13.3	50.6	19.2	446	80.2	50.9
CoLLM-DC	1.9	194	**95.4**	2.0	196	94.1	2.5	161	59.1	**9.3**	**182**	7.6	44.6	19.4	470	43.8	46.8
CoLLM-CC	**1.8**	181	95.2	**1.9**	**188**	95.0	2.6	166	**75.2**	9.5	239	7.3	**68.5**	**19.0**	**442**	86.4	52.7

Thank You!