清華大學电子工程系
Department of Electronic Engineering, Tsinghua University

MARSHAL: Incentivizing Multi-Agent Reasoning via Self-Play with Strategic LLMs

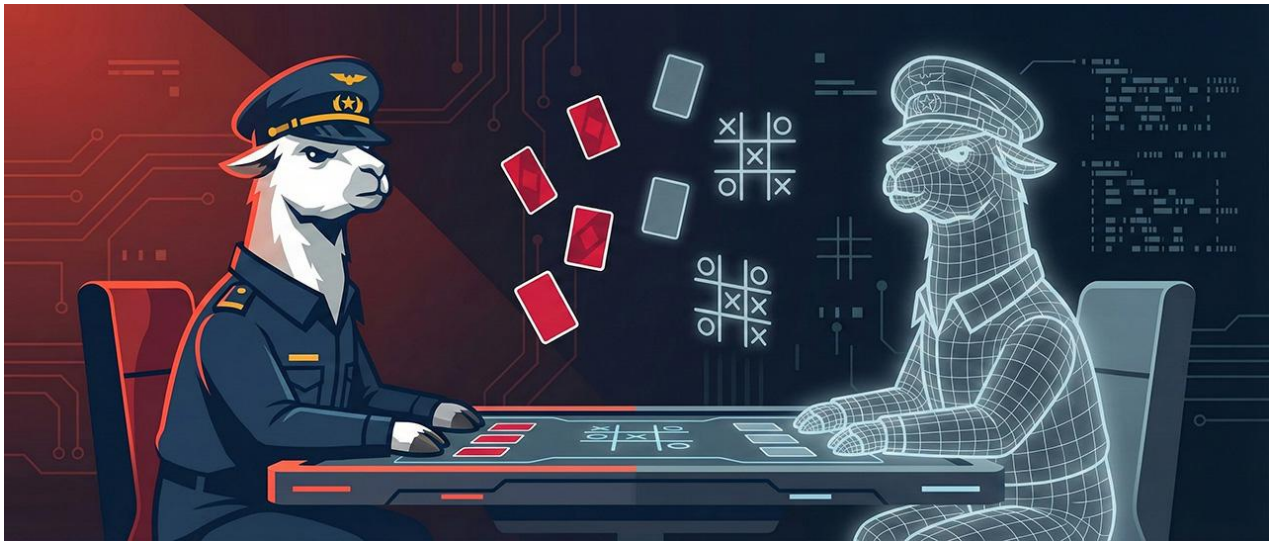
Huining Yuan[1*], Zelai Xu[1*], Zheyue Tan[2], Xiangmin Yi[1]
Mo Guang[3], Kaiwen Long[3], Haojia Hui[3], Boxun Li[4], Xinlei Chen[1], Bo Zhao[2]
Xiao-Ping Zhang[1†], Chao Yu[1†], Yu Wang[1†]

[1]Tsinghua University, [2]Aalto University, [3]Li Auto Inc., [4]Infinigence-AI
**Equal contribution; †Corresponding authors*

Project Page

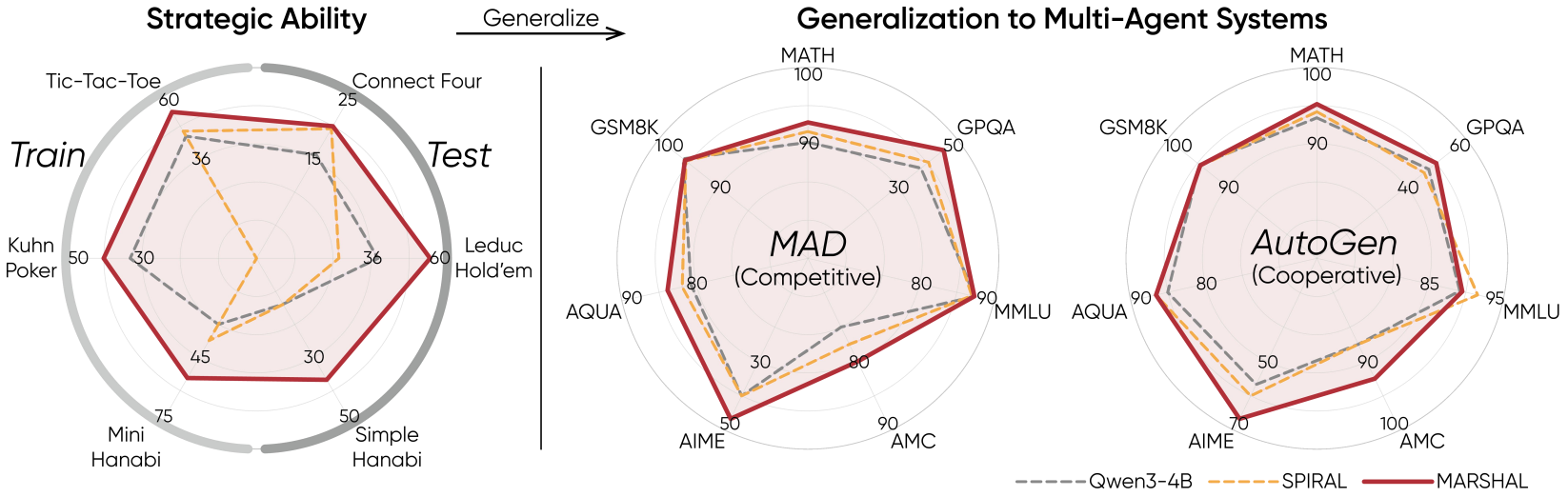
Better Multi-Agent System

- How to train better reasoning models for LLM-based MASs?
  - RLVR is underexplored for the **multi–turn/–agent** settings of MAS
- We propose: **self-play in strategic games**



Results Overview

- Train on games, generalize to MAS
  - Strategic ability on games: up to **+28.1% on testing games**
  - Generalization to MASs: up to **+10.0% on AIME, +7.6% on GPQA**



MARSHAL - Method

➢MARSHAL:

- Multi-Agent Reasoning though Selfplay witH strAtegic LLMs

➢Challenge:

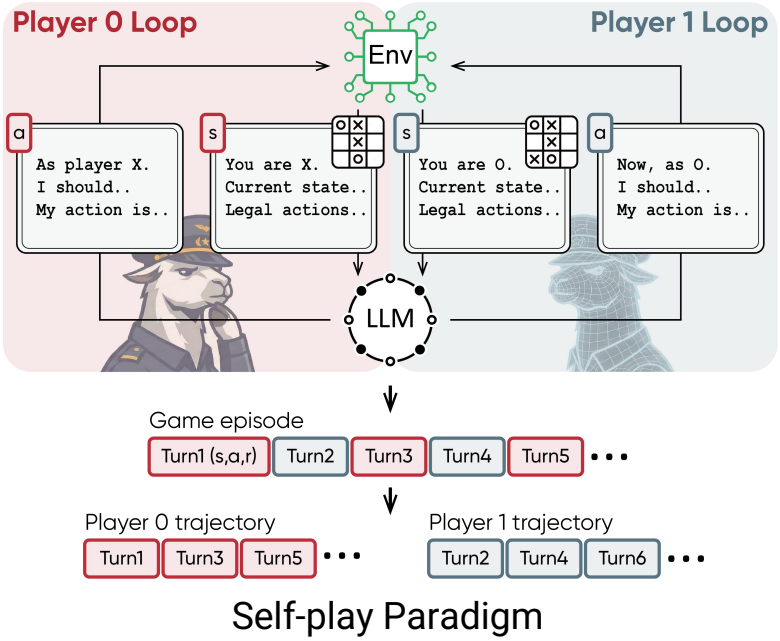
- Multi-turn credit assignment
- Multi-agent advantage estimation

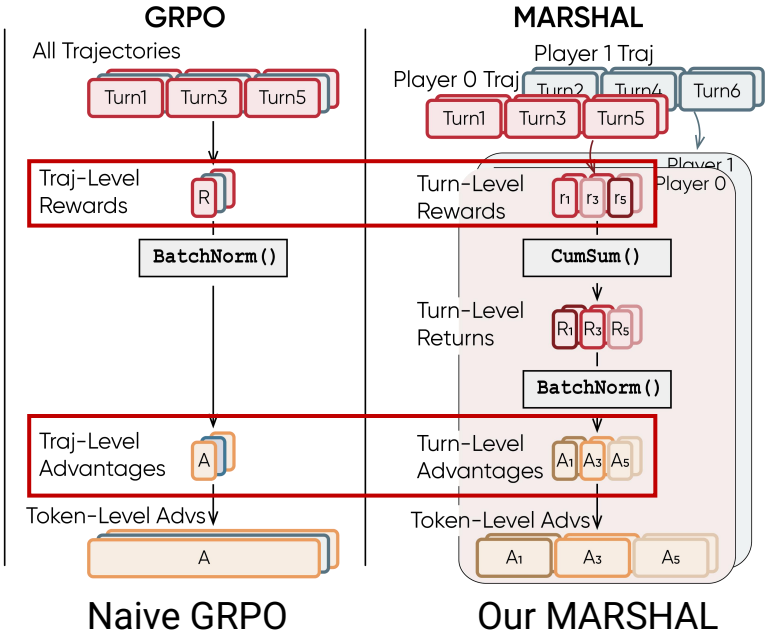
➢How to eval generalization?

- Integrate in to existing MASs
- Test on math/QA benchmarks

MARSHAL - Method

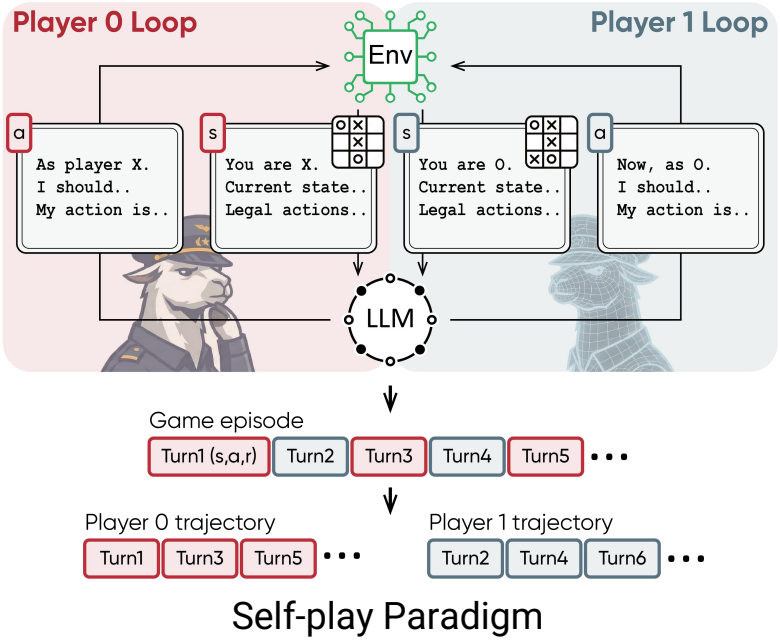
- Algorithm: two modifications to naive GRPO
  1. **Turn–level advantage estimator**: a “sum-then-norm” approach that assigns fine-grained and stable turn-level credit

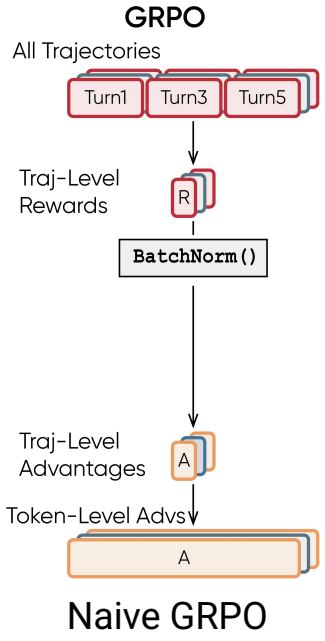


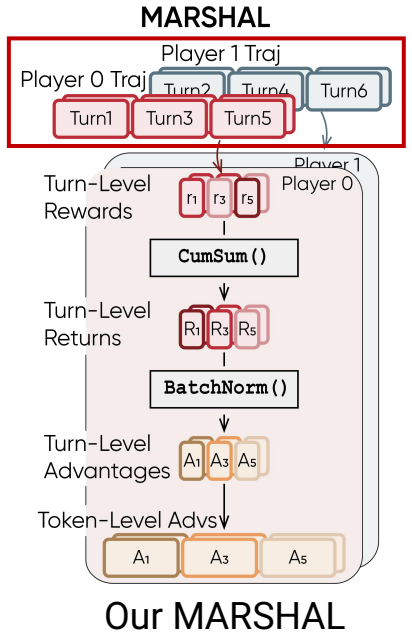


MARSHAL - Method

- Algorithm: two modifications to naive GRPO
  - 2. **Agent–specific advantage normalization**: separate the advantage estimation for each player-role to stablilize multi-agent training



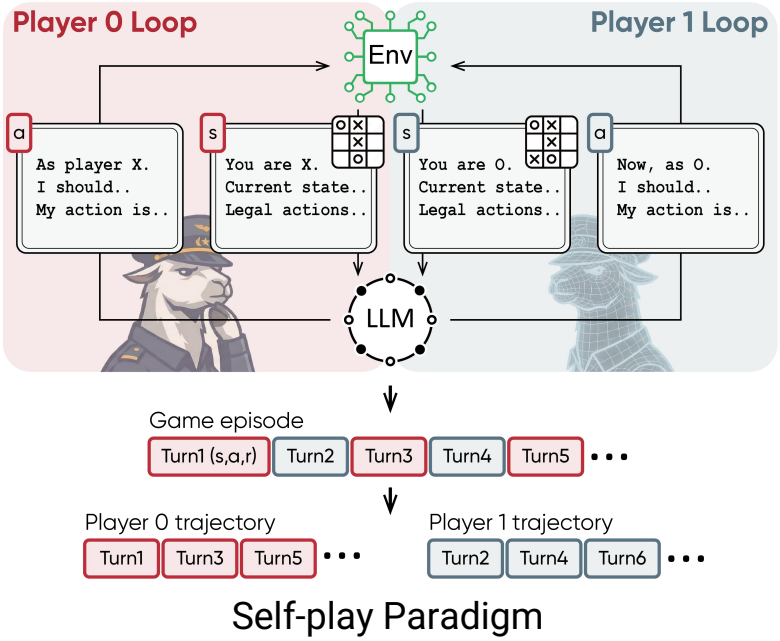


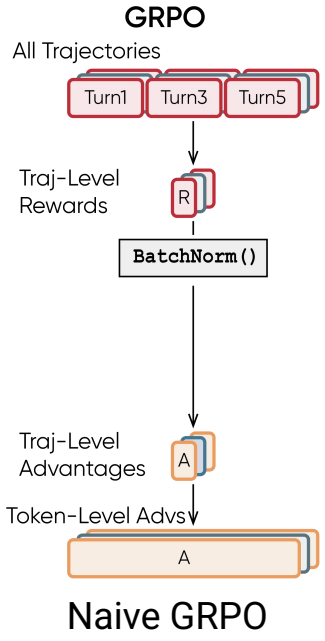


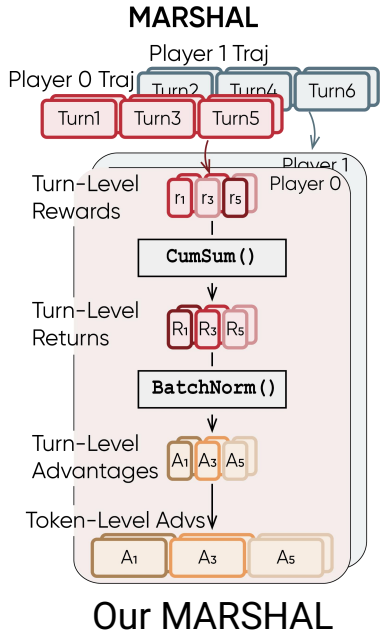
MARSHAL - Method

- Algorithm: two modifications to naive GRPO

  Under these two modifications, **MARSHAL is equivalent to GAE** with:
  $\gamma = 1; \lambda = 1$; value of all states estimated by mean of return







MARSHAL - Method

➢Game selection

- include both competitive and cooperative for robust skill set

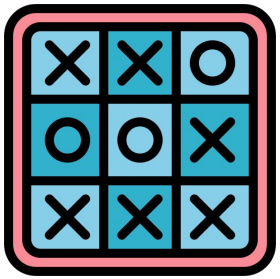
- Perfect-info, competitive:
- Imperfect-info, competitve:
- Imperfect-info, cooperative:

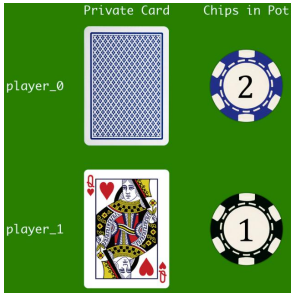
➢Training:

- TicTacToe
- Kuhn Poker
- Mini Hanabi

➢Testing:

- Connect Four
- Leduc Hold'em
- Simple Hanabi





MARSHAL - Results

➢Experimental setup

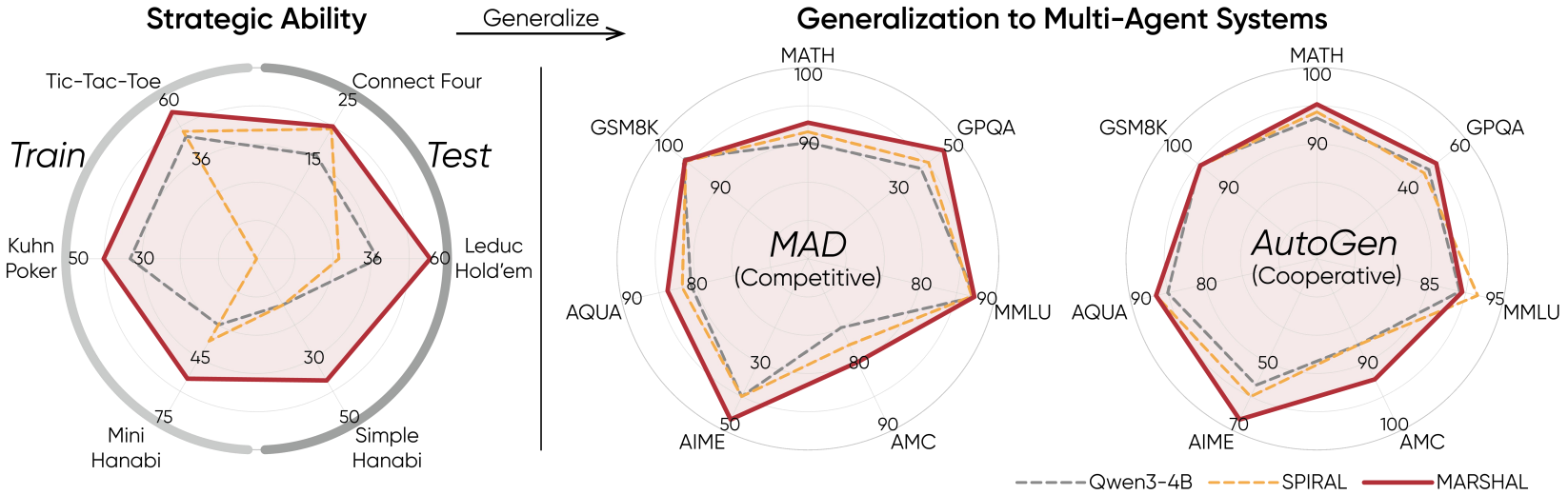
- Base model: **Qwen3-4B**
- Train **3 specialist (1 each game), 1 generalist (on 3 games)**

- **Experiments:**
  1. Strategic ability on testing games
  2. Generalization to MASs
     - Integrate agent into **MAD (competitive), AutoGen (cooperative)**
     - Evaluate on 7 math and general QA benchmarks
  3. Reasoning pattern analysis
  4. Ablations

[1] Liang, Tian, et al. "Encouraging divergent thinking in large language models through multi-agent debate." EMNLP 2024.
[2] *Wu, Qingyun, et al. "Autogen: Enabling next-gen LLM applications via multi-agent conversations." COLM 2024.*

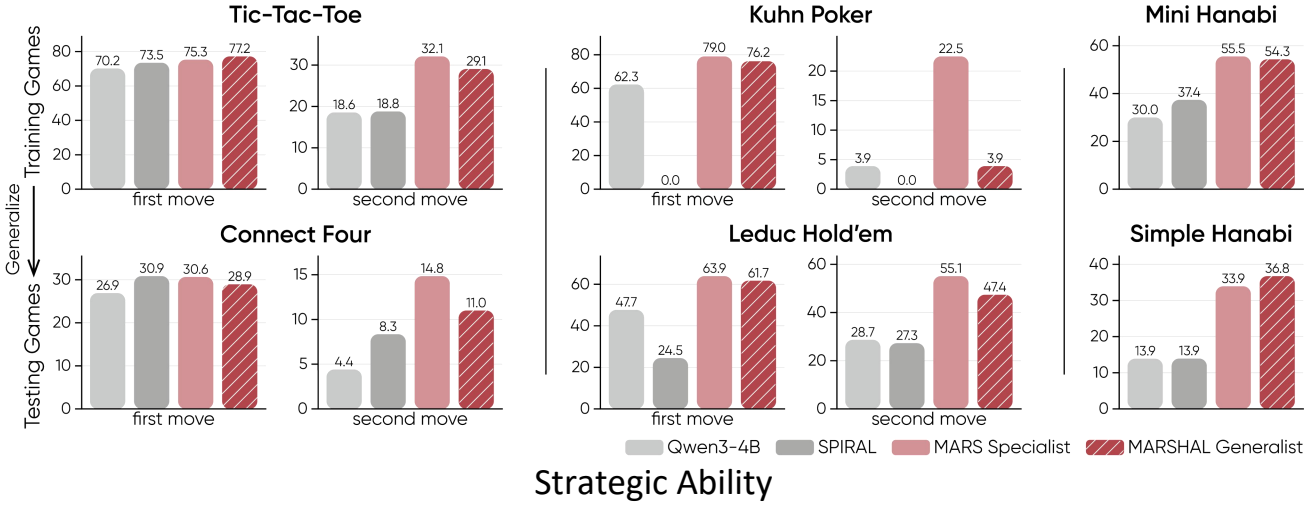
Results Overview

- Train on games, generalize to MAS
  - Improved strategic ability on games, up to **+28.1% on testing games**
  - Successfull generalization to MASs, up to:
    - **+10.0% on AIME, +7.6% on GPQA, +3.51% on average**

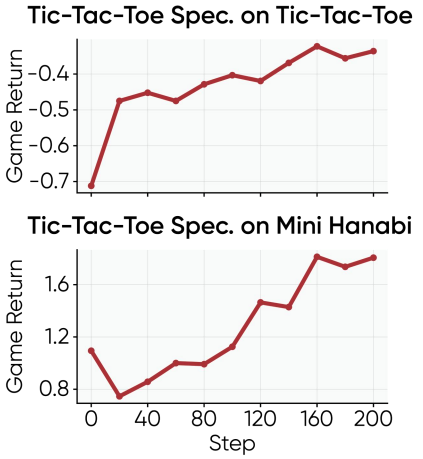


MARSHAL - Strategic Ability

- Evaluate against fix opponents (i.e. MCTS, CFR)
  - Specialist agents generalizes well to their corresponding test game
  - Cross category generalization
  - **Generalist agent performs best overall**



Strategic Ability



Cross-category Genearlization

MARSHAL - Generalization to MAS

- Start with single agent, then test generalization to MASs
  - 1. MARSHAL performs better than the baseline Qwen3-4B even in single agent setting

Setting	Model	Average	Math					QA	
			MATH	GSM8K	AQUA	AIME	AMC	MMLU	GPQA
Single Agent	Qwen3-4B	60.74	87.60	94.60	39.80	36.70	70.00	57.10	**39.39**
	SPIRAL	**63.75**	87.50	94.80	51.20	36.70	**80.00**	58.70	37.37
	MARSHAL								
	Tic-Tac-Toe	63.54	89.10	**95.20**	46.50	40.00	77.50	57.60	38.89
	Kuhn Poker	61.38	87.80	94.50	48.40	33.30	72.50	59.30	33.84
	Mini Hanabi	62.05	88.10	94.70	48.00	**43.30**	65.00	58.90	36.36
	Generalist	62.79	**89.90**	94.60	**52.00**	33.30	75.00	**59.90**	34.85
MAD (Competitive)	Qwen3-4B	72.45	90.20	95.91	80.71	40.00	75.00	**87.42**	37.88
	SPIRAL	73.41	91.60	95.45	81.89	40.00	77.50	87.01	40.40
	MARSHAL								
	Tic-Tac-Toe	75.01	92.20	96.06	83.07	43.33	**82.50**	86.76	41.12
	Kuhn Poker	74.54	91.60	**96.21**	82.68	40.00	**82.50**	87.39	41.41
	Mini Hanabi	73.70	91.40	95.60	82.68	43.33	77.50	87.04	38.38
	Generalist	**75.96**	**92.80**	95.60	**83.86**	**46.67**	80.00	87.36	**45.45**
AutoGen (Cooperative)	Qwen3-4B	79.14	93.40	**94.69**	85.04	56.67	87.50	89.21	47.47
	SPIRAL	80.05	94.20	94.47	86.61	60.00	87.50	**91.60**	45.96
	MARSHAL								
	Tic-Tac-Toe	80.15	94.40	**94.69**	**87.01**	60.00	90.00	89.53	45.45
	Kuhn Poker	81.54	**95.80**	94.39	86.61	63.33	92.50	89.65	48.48
	Mini Hanabi	81.54	94.40	94.54	86.22	**66.67**	**95.00**	88.98	44.95
	Generalist	**82.15**	95.20	94.54	86.61	**66.67**	92.50	89.53	**50.00**

MARSHAL - Generalization to MAS

- Start with single agent, then test generalization to MASs
  - 2. Successful generalization to MAS: on both MAD and AutoGen, our generalist performs best overall

Setting	Model	Average	Math					QA	
			MATH	GSM8K	AQUA	AIME	AMC	MMLU	GPQA
Single Agent	Qwen3-4B	60.74	87.60	94.60	39.80	36.70	70.00	57.10	**39.39**
	SPIRAL	**63.75**	87.50	94.80	51.20	36.70	**80.00**	58.70	37.37
	MARSHAL								
	Tic-Tac-Toe	63.54	89.10	**95.20**	46.50	40.00	77.50	57.60	38.89
	Kuhn Poker	61.38	87.80	94.50	48.40	33.30	72.50	59.30	33.84
	Mini Hanabi	62.05	88.10	94.70	48.00	**43.30**	65.00	58.90	36.36
	Generalist	62.79	**89.90**	94.60	**52.00**	33.30	75.00	**59.90**	34.85
MAD (Competitive)	Qwen3-4B	72.45	90.20	95.91	80.71	40.00	75.00	**87.42**	37.88
	SPIRAL	73.41	91.60	95.45	81.89	40.00	77.50	87.01	40.40
	MARSHAL								
	Tic-Tac-Toe	75.01	92.20	96.06	83.07	43.33	**82.50**	86.76	41.12
	Kuhn Poker	74.54	91.60	**96.21**	82.68	40.00	**82.50**	87.39	41.41
	Mini Hanabi	73.70	91.40	95.60	82.68	43.33	77.50	87.04	38.38
	Generalist	**75.96**	**92.80**	95.60	**83.86**	**46.67**	80.00	87.36	**45.45**
AutoGen (Cooperative)	Qwen3-4B	79.14	93.40	**94.69**	85.04	56.67	87.50	89.21	47.47
	SPIRAL	80.05	94.20	94.47	86.61	60.00	87.50	**91.60**	45.96
	MARSHAL								
	Tic-Tac-Toe	80.15	94.40	**94.69**	**87.01**	60.00	90.00	89.53	45.45
	Kuhn Poker	81.54	**95.80**	94.39	86.61	63.33	92.50	89.65	48.48
	Mini Hanabi	81.54	94.40	94.54	86.22	**66.67**	**95.00**	88.98	44.95
	Generalist	**82.15**	95.20	94.54	86.61	**66.67**	92.50	89.53	**50.00**

MARSHAL - Generalization to MAS

- Start with single agent, then test generalization to MASs
  - 3. Highly aligned generalization domain: competitive game generalize well to competitive MAS; and vise versa

Setting	Model	Average	Math					QA	
			MATH	GSM8K	AQUA	AIME	AMC	MMLU	GPQA
Single Agent	Qwen3-4B	60.74	87.60	94.60	39.80	36.70	70.00	57.10	**39.39**
	SPIRAL	**63.75**	87.50	94.80	51.20	36.70	**80.00**	58.70	37.37
	MARSHAL								
	Tic-Tac-Toe	63.54	89.10	**95.20**	46.50	40.00	77.50	57.60	38.89
	Kuhn Poker	61.38	87.80	94.50	48.40	33.30	72.50	59.30	33.84
	Mini Hanabi	62.05	88.10	94.70	48.00	**43.30**	65.00	58.90	36.36
	Generalist	62.79	**89.90**	94.60	**52.00**	33.30	75.00	**59.90**	34.85
MAD (Competitive)	Qwen3-4B	72.45	90.20	95.91	80.71	40.00	75.00	**87.42**	37.88
	SPIRAL	73.41	91.60	95.45	81.89	40.00	77.50	87.01	40.40
	MARSHAL								
	Tic-Tac-Toe	75.01	92.20	96.06	83.07	43.33	**82.50**	86.76	41.12
	Kuhn Poker	74.54	91.60	**96.21**	82.68	40.00	**82.50**	87.39	41.41
	Mini Hanabi	73.70	91.40	95.60	82.68	43.33	77.50	87.04	38.38
	Generalist	**75.96**	**92.80**	95.60	**83.86**	**46.67**	80.00	87.36	**45.45**
AutoGen (Cooperative)	Qwen3-4B	79.14	93.40	**94.69**	85.04	56.67	87.50	89.21	47.47
	SPIRAL	80.05	94.20	94.47	86.61	60.00	87.50	**91.60**	45.96
	MARSHAL								
	Tic-Tac-Toe	80.15	94.40	**94.69**	**87.01**	60.00	90.00	89.53	45.45
	Kuhn Poker	81.54	**95.80**	94.39	86.61	63.33	92.50	89.65	48.48
	Mini Hanabi	81.54	94.40	94.54	86.22	**66.67**	**95.00**	88.98	44.95
	Generalist	**82.15**	95.20	94.54	86.61	**66.67**	92.50	89.53	**50.00**

MARSHAL - Pattern Analysis

- Qualitative
  - role understanding, intent recognition

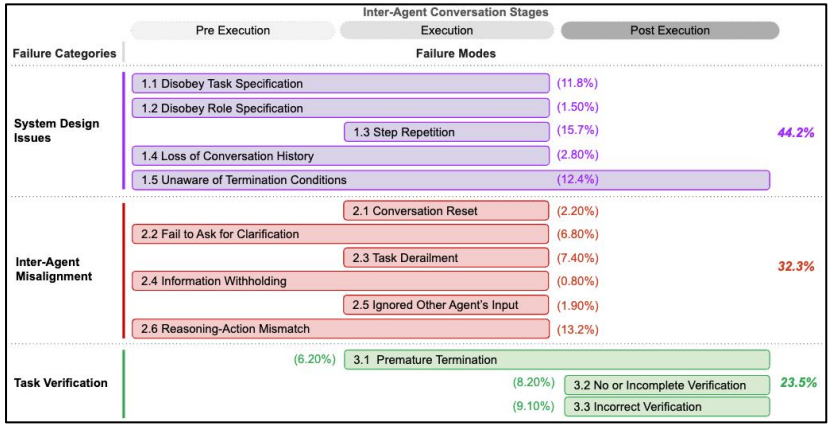
Skill	Manifestation in Game-Play	Generalize to Multi-Agent Systems
Role Understanding	*The Tic-Tac-Toe specialist recognizes its role as the second player (O) and adopts a defensive strategy.* <think> Okay, so I am playing the game of Tic-Tac-Toe as the mark O... **As the second-move player, I should** prioritize blocking the X marks from forming a horizontal, vertical, or diagonal line... </think>	*The same agent, acting as the "negative" debater in MAD, adapts its strategy to refute the opponent.* <think> Okay, so I need to figure out the minimum number of... The affirmative side just argued for 4. **As the negative side assistant, I should** try to find an alternative solution and convince the judge... </think>
Intent Recognition	*The Hanabi specialist infers the intent behind a teammate's ambiguous hint.* <think> Okay, so I need to figure out the next move. Player 0 just revealed my only red card. Wait, but the rank remains unknown. **Maybe they want me to play this card to the stack?** ... </think>	*The same agent, acting as a user proxy in AutoGen, infers uncertainty from a collaborator's missing 'TERMINATE' token.* <think> Okay, let's see. The assistant gave the answer 17.5 m/s, but did not end the conversation with 'TERMINATE'. **Maybe the assistant is not sure with this answer?** ... </think>

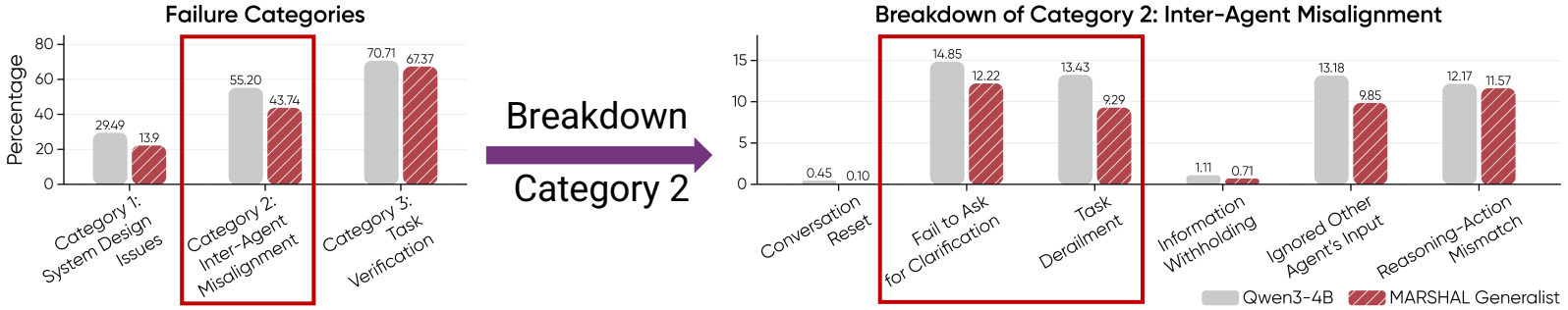
MARSHAL - Pattern Analysis

➢Quantitative:

- Count failure modes using the taxomony of *Cemri. et al, 2025*
- **DeepSeek R1 as judge**
- Results: Significant reduce in **Inter-Agent Misalignment**

NEURAL INFORMATION PROCESSING SYSTEMS 2025 *Cemri, Mert, et al. Why do multi-agent llm systems fail?. NeurIPS 2025.*





MARSHAL - Ablations

➢Self-play vs. fixed-opponent

- Train Tic-Tac-Toe specialist against MCTS
- Train Kuhn Poker specialist against Nash Equilibrium policy
- **Results: fixed–opponent–trained agents overfit**

Model	Training Games			Testing Games		
	Tic-Tac-Toe	Kuhn Poker	Mini Hanabi	Connect Four	Leduc Hold'em	Simple Hanabi
MARSHAL (Tic-Tac-Toe)	75.30 / 32.10	74.15 / 3.42	50.48	30.65 / 14.85	58.36 / 27.65	29.75
w/ fixed opponent	88.00 / 41.95	63.15 / 28.84	34.93	20.35 / 5.65	47.38 / 35.55	12.22
MARSHAL (Kuhn Poker)	69.85 / 25.50	79.04 / 22.49	44.98	27.60 / 12.70	63.94 / 62.10	29.35
w/ fixed opponent	0.00 / 0.00	76.19 / 15.64	0.00	0.00 / 0.00	0.00 / 0.00	0.00

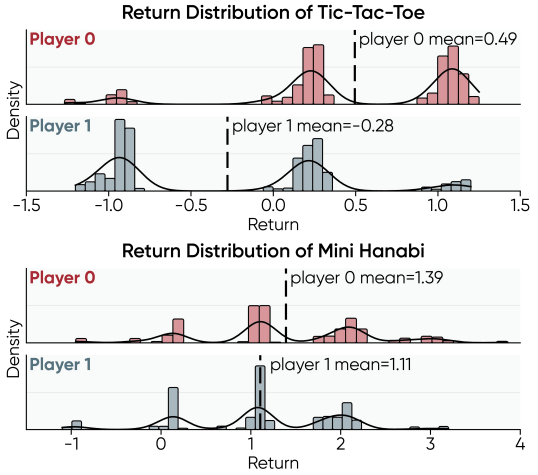
For competitive games, entries indicate first-move / second-move game returns.

MARSHAL - Ablations

- Algo rithmic design. Results:
  - **1. Turn–level advantage estimator** is crucial for long-horizon tasks
  - **2. Agent–specific advantage normalization** has mild effect on the cooperative Hanabi (similar return distribution between players)

Model	Training Games			Testing Games		
	Tic-Tac-Toe	Kuhn Poker	Mini Hanabi	Connect Four	Leduc Hold'em	Simple Hanabi
MARSHAL (Tic-Tac-Toe)	75.30 / 32.10	74.15 / 3.42	50.48	30.65 / 14.85	58.36 / 27.65	29.75
w/o Turn-Level.	74.60 / 24.15	80.26 / 28.35	34.80	26.75 / 12.30	48.34 / 41.34	19.05
w/o Agent-Specific.	82.70 / 31.20	70.89 / 11.24	44.10	25.40 / 10.50	51.04 / 49.88	21.72
MARSHAL (Kuhn Poker)	69.85 / 25.50	79.04 / 22.49	44.98	27.60 / 12.70	63.94 / 62.10	29.35
w/o Turn-Level.	63.35 / 19.65	92.49 / 21.02	41.65	29.60 / 10.85	32.26 / 31.23	22.98
w/o Agent-Specific.	69.55 / 24.55	75.37 / 19.55	40.18	27.00 / 10.50	35.73 / 21.50	22.42
MARSHAL (Hanabi)	71.90 / 7.35	72.52 / 9.29	55.55	26.75 / 5.75	37.36 / 55.12	33.93
w/o Turn-Level.	67.55 / 10.60	68.45 / 31.78	53.20	25.25 / 3.05	54.79 / 47.77	30.68
w/o Agent-Specific.	68.15 / 13.40	74.15 / 10.27	52.50	32.10 / 5.10	44.30 / 56.41	32.08

For competitive games, entries indicate first-move / second-move game returns.



MARSHAL - Key Takeaway

➢Takeaway

- **Effective Generalization**:
  - Self-play in strategic games successfully builds generalizable reasoning skills for general LLM-based multi-agent systems.
- **Credit Assignment Matters**:
  - Fine-grained credit assignment matters for effective and stable RL training in multi-turn multi-agent tasks.
- **The Frontier**:
  - Designing specialized RL algorithms for LLM-based MASs remains a crucial and open direction for future research.

Thanks for your time!

MARSHAL

Project Page

Paper

Code

Huining Yuan[1*], Zelai Xu[1*], Zheyue Tan[2], Xiangmin Yi[1]
Mo Guang[3], Kaiwen Long[3], Haojia Hui[3], Boxun Li[4], Xinlei Chen[1], Bo Zhao[2]
Xiao-Ping Zhang[1†], Chao Yu[1†], Yu Wang[1†]

✉: {yuanhuining0, zelai.eecs}@gmail.com

[1]*Tsinghua University,* [2]*Aalto University,* [3]*Li Auto Inc.,* [4]*Infinigence-AI*