Competing Large Language Models in Multi-Agent Gaming Environments

Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang

Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Michael Lyu

Paper **Code** **My Homepage :)**

Tencent AI Lab

香港中文大學
The Chinese University of Hong Kong

GAMA-Bench Motivation

- How is LLMs' decision-making ability in game theoretic scenes?
  1. **Multiparty:** theory-of-mind reasoning
  2. **Calculation:** arithmetic reasoning
  3. **Understanding:** environment & game rules

- Games: ideal test bed for LLM evaluation
  1. **Scope:** abstraction of real-world scenarios
  2. **Quantifiability:** compute scores with math models
  3. **Variability:** changing game parameters

J Huang et al. Competing Large Language Models in Multi-Agent Gaming Environments. In ICLR 2025.

Limitations in Existing Frameworks

1. Two-player setting
   - Prisoner's Dilemma; Ultimatum Game;
   - Diner's Dilemma; Pirate Game;

2. Pure strategies
   - Games without Pure Strategy Nash Equilibrium: Rock-Paper-Scissors; El Farol Bar Game
   - Mixed Strategy Nash Equilibrium (MSNE)

3. Fixed and classic setting
   - Guess 2/3 of the Average
   - Guess R of the Average

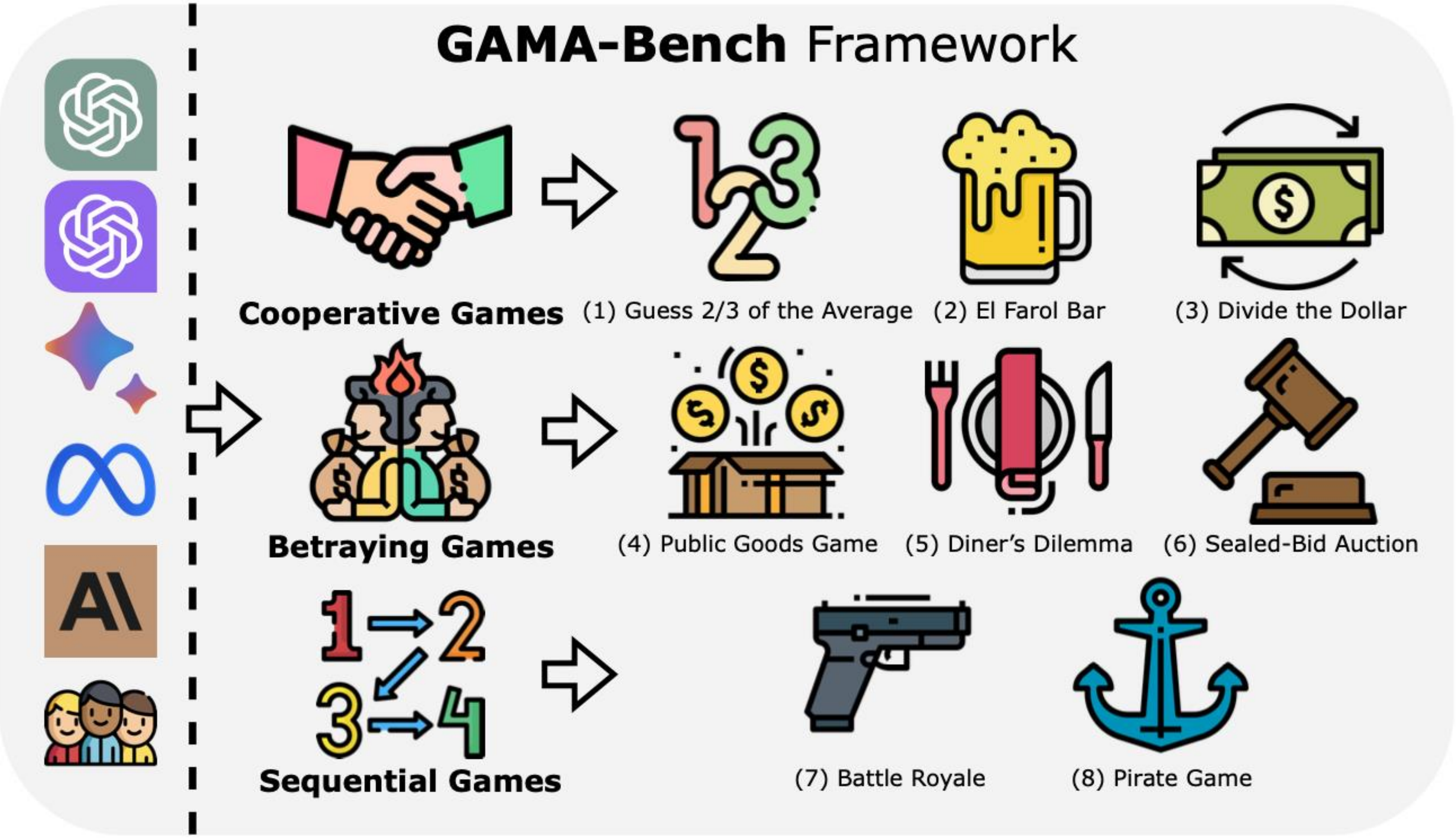
J Huang et al. Competing Large Language Models in Multi-Agent Gaming Environments. In ICLR 2025.

GAMA-Bench Game Types

1. Cooperative Games
   - ➢ Get worse if not cooperate
2. Betraying Games
   - ➢ Get better if not cooperate

➢ Simultaneous Games

3. Sequential Games



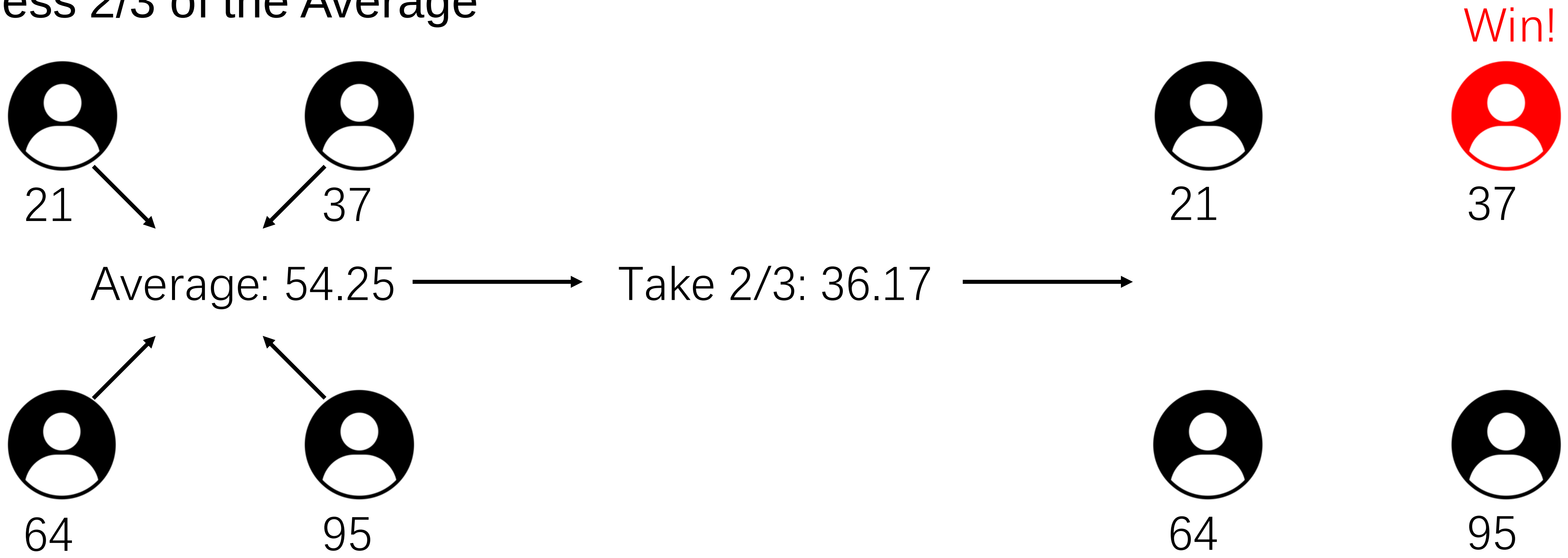
➢ GAMA-Bench strengths:

1. Not just 2-player/2-action games
2. Not only PSNE
3. Not only one game setting

J Huang et al. Competing Large Language Models in Multi-Agent Gaming Environments. In ICLR 2025.

Highlighted Games 1/4 (Cooperative Game)

- Guess 2/3 of the Average



- Average of [0, 100] -> 50 -> Take 2/3 -> 33.33
    - -> Take 2/3 -> 22.22 -> Take 2/3 -> 14.81 -> … -> 0!

J Huang et al. Competing Large Language Models in Multi-Agent Gaming Environments. In ICLR 2025.

Highlighted Games 2/4 (Cooperative Game)

- El Farol Bar Game
  - The most historic and iconic bar in Santa Fe, NM, USA

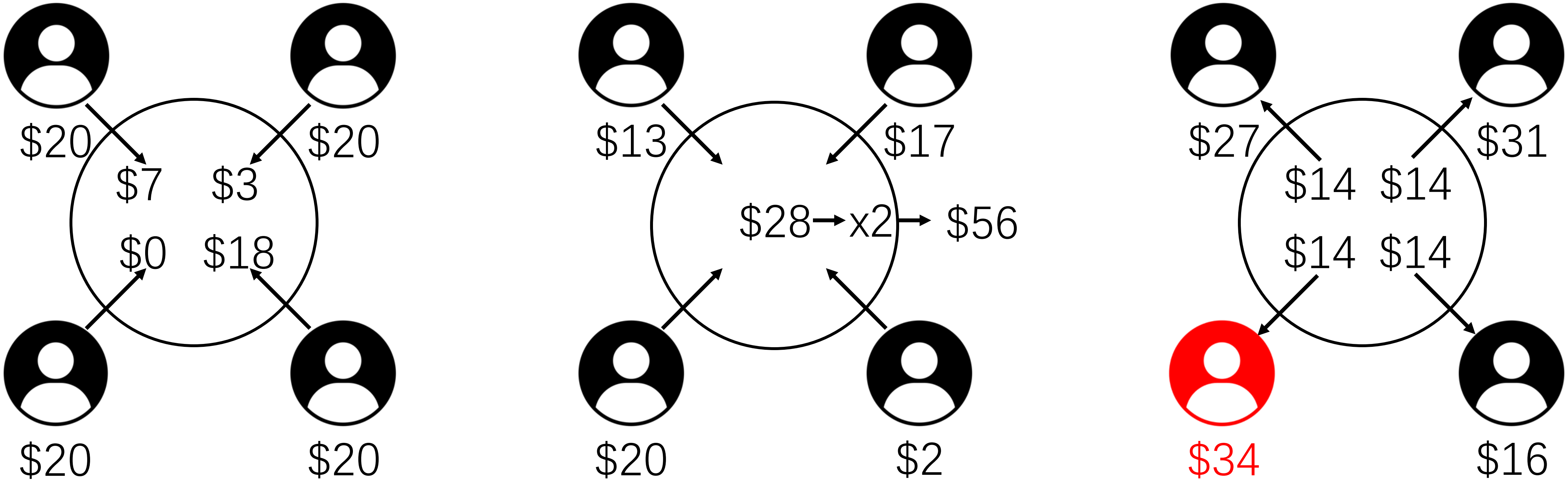
- Rules
  - N players decide independently whether to go to the bar
  - Bar has its capacity:
    - If < 60% of N are in the bar, they have More fun than staying home
    - If >= 60% of N are in the bar, they have Less fun than staying home

- There is no PSNE!
  - If everyone acts the same, either All or None are in the bar; Less total utility!
  - MSNE: (60%) Go + (40%) Not Go

J Huang et al. Competing Large Language Models in Multi-Agent Gaming Environments. In ICLR 2025.

Highlighted Games 3/4 (Betraying Game)

- Public Goods Game

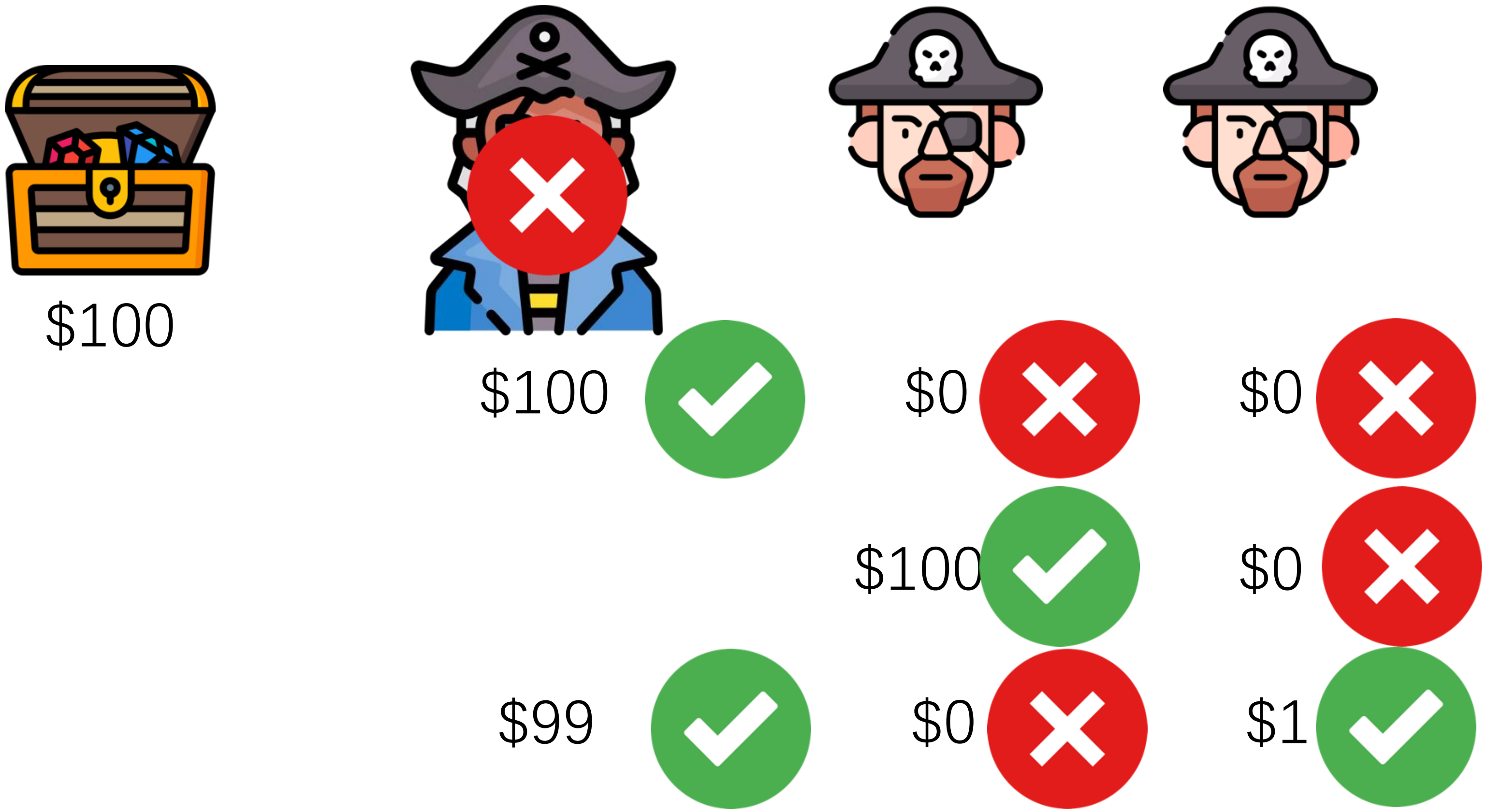


- Dollars in the public pot multiply by R (1 < R < N)
- Players tend to free-ride

J Huang et al. Competing Large Language Models in Multi-Agent Gaming Environments. In ICLR 2025.

Highlighted Games 4/4 (Sequential Game)

- Pirate Game



$100			
	$100 ✓	$0 ✗	$0 ✗
		$100 ✓	$0 ✗
	$99 ✓	$0 ✗	$1 ✓

- 1st Pirate: 0 for 2nd, 1 for 3rd, 0 for 4th, 1 for 5th … And keep the remaining

J Huang et al. Competing Large Language Models in Multi-Agent Gaming Environments. In ICLR 2025.

GAMA-Bench Evaluation Metrics

1. Optimal Strategy
   - ➢ For self-interest
   - ➢ For social welfare: Require priors

2. Human Choices
   - ➢ Require user studies

➢ We mainly study optimal strategy for Self-Interest in GAMA-Bench

➢ The scores are re-scaled to 0-100 (the higher the better)

$$S_1 = \begin{cases} \frac{(MAX-MIN)-S_1}{MAX-MIN} * 100, & R < 1 \\ \left(1 - \frac{|2S_1-(MAX-MIN)|}{MAX-MIN}\right) * 100, & R = 1 \\ \frac{S_1}{MAX-MIN} * 100, & R > 1 \end{cases},$$

$$S_2 = \frac{\max(R, 1-R) - S_2}{\max(R, 1-R)} * 100,$$

$$S_3 = \max\left(\frac{G - S_3}{G} * 100, 0\right),$$

$$S_4 = \begin{cases} \frac{T-S_4}{T} * 100, & \frac{R}{N} \leq 1 \\ \frac{S_4}{T} * 100, & \frac{R}{N} > 1 \end{cases},$$

$$S_5 = (1 - S_5) * 100,$$

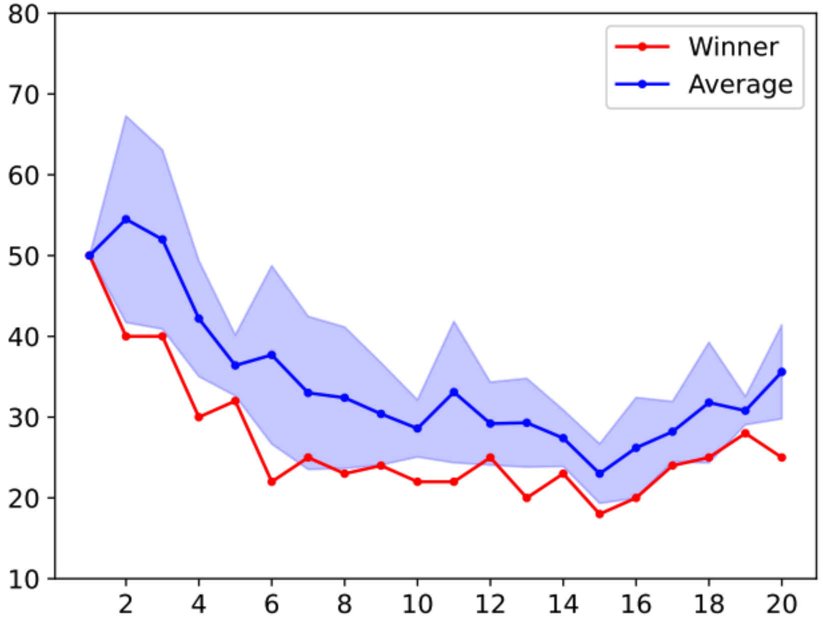
$$S_6 = S_6 * 100,$$

$$S_7 = S_7 * 100,$$

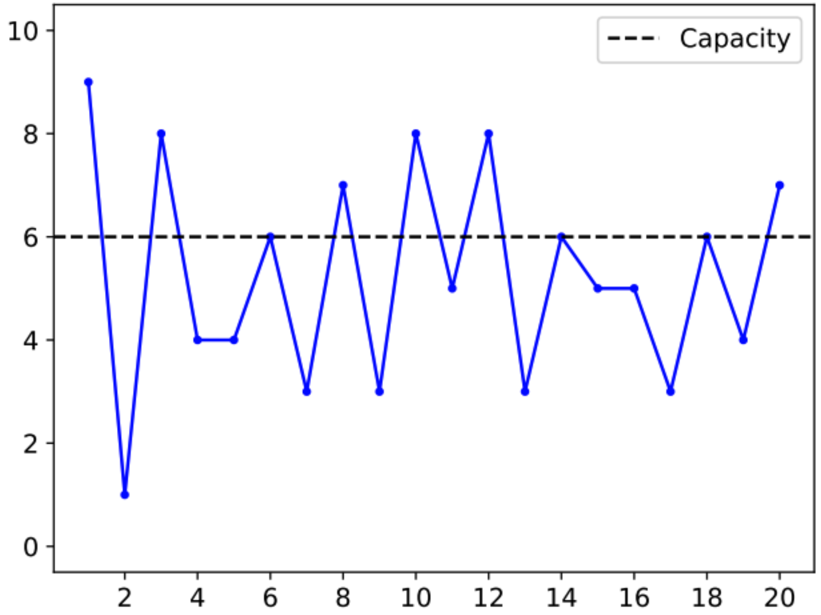
$$S_8 = \frac{2 * G - S_{8P}}{2 * G} * 50 + S_{8V} * 50.$$

J Huang et al. Competing Large Language Models in Multi-Agent Gaming Environments. In ICLR 2025.

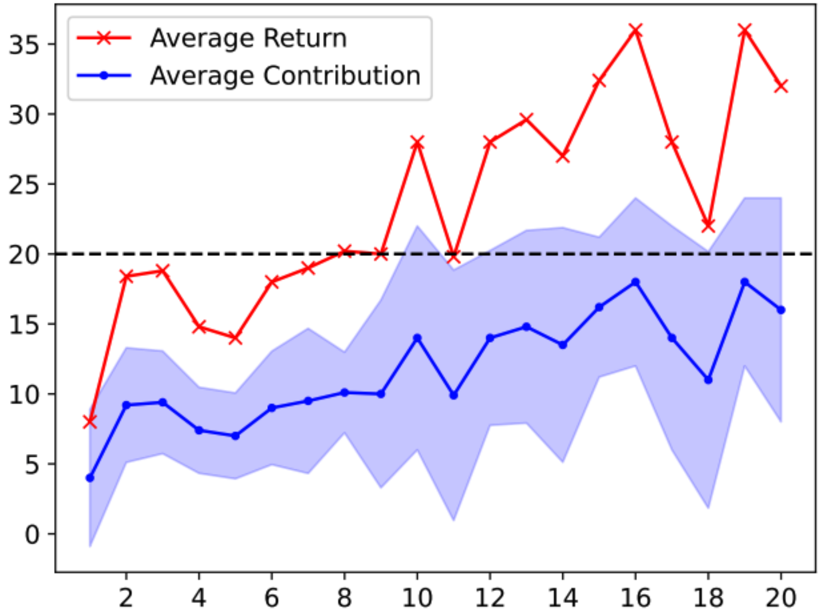
How Does GPT-3.5 Perform?



(1) Guess 2/3 of the Average
Average Number and Winning Number



(2-1) El Farol Bar-Explicit
Number of Players in the Bar

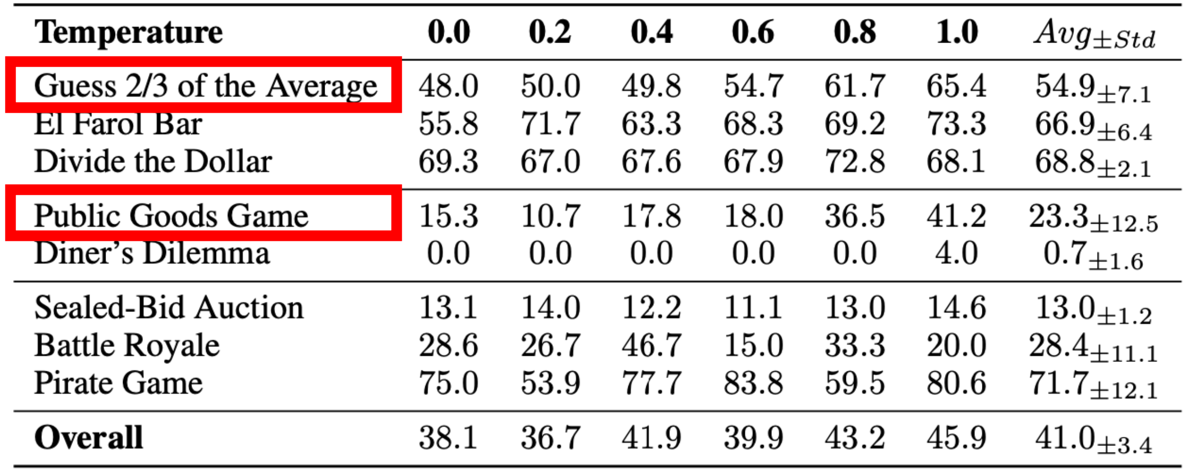


(4) Public Goods Game
Average Contribution and Return

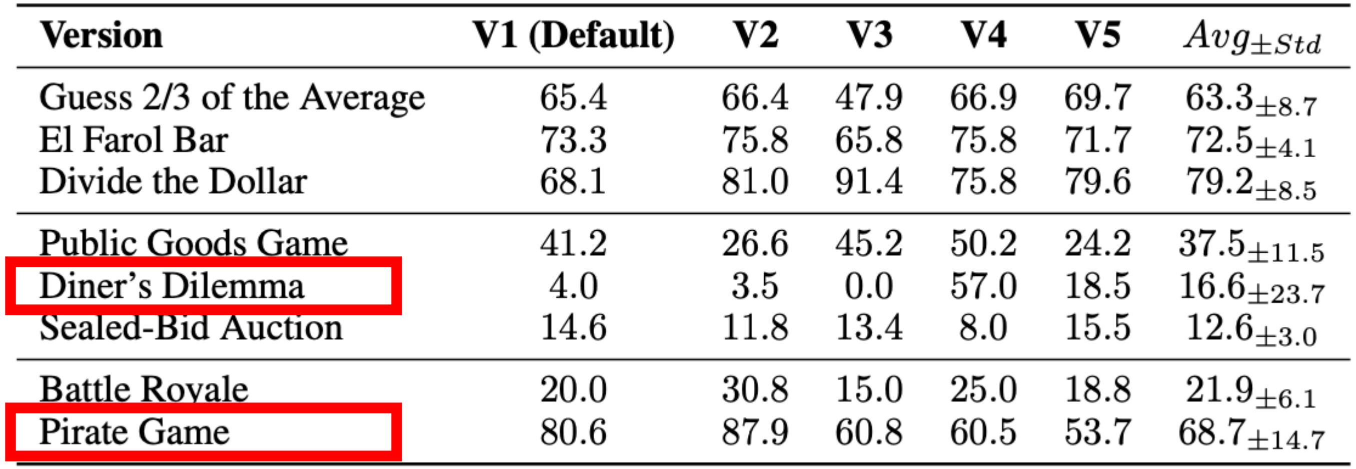
Pirate Rank	1	2	3	4	5	6	7	8	9	10	S_{8P}	S_{8V}
Round 1	100✓	0✗	0✗	0✗	0✗	0✗	0✗	0✗	0✗	0✗	8	1.00
Round 2	-	99✓	0✗	1✓	0✓	0✗	0✗	0✗	0✗	0✓	6	0.75
Round 3	-	-	50✓	1✓	1✓	1✓	1✓	1✓	1✓	44✓	94	0.57

J Huang et al. Competing Large Language Models in Multi-Agent Gaming Environments. In ICLR 2025.

How About the Robustness?



Temperature	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	$Avg_{\pm Std}$
Guess 2/3 of the Average	48.0	50.0	49.8	54.7	61.7	65.4	$54.9_{\pm 7.1}$
El Farol Bar	55.8	71.7	63.3	68.3	69.2	73.3	$66.9_{\pm 6.4}$
Divide the Dollar	69.3	67.0	67.6	67.9	72.8	68.1	$68.8_{\pm 2.1}$
Public Goods Game	15.3	10.7	17.8	18.0	36.5	41.2	$23.3_{\pm 12.5}$
Diner’s Dilemma	0.0	0.0	0.0	0.0	0.0	4.0	$0.7_{\pm 1.6}$
Sealed-Bid Auction	13.1	14.0	12.2	11.1	13.0	14.6	$13.0_{\pm 1.2}$
Battle Royale	28.6	26.7	46.7	15.0	33.3	20.0	$28.4_{\pm 11.1}$
Pirate Game	75.0	53.9	77.7	83.8	59.5	80.6	$71.7_{\pm 12.1}$
Overall	38.1	36.7	41.9	39.9	43.2	45.9	$41.0_{\pm 3.4}$

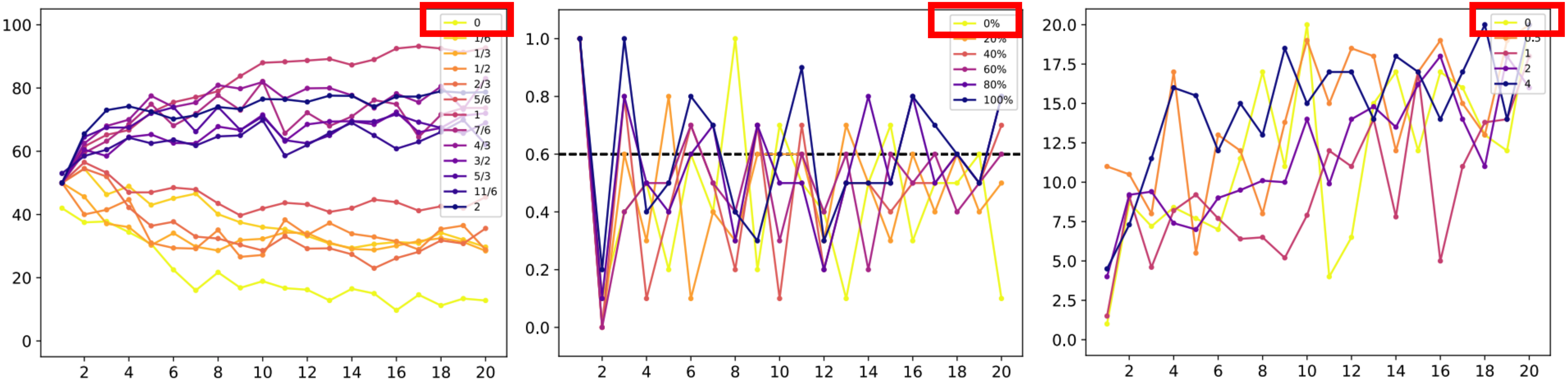


Version	**V1 (Default)**	**V2**	**V3**	**V4**	**V5**	$Avg_{\pm Std}$
Guess 2/3 of the Average	65.4	66.4	47.9	66.9	69.7	$63.3_{\pm 8.7}$
El Farol Bar	73.3	75.8	65.8	75.8	71.7	$72.5_{\pm 4.1}$
Divide the Dollar	68.1	81.0	91.4	75.8	79.6	$79.2_{\pm 8.5}$
Public Goods Game	41.2	26.6	45.2	50.2	24.2	$37.5_{\pm 11.5}$
Diner’s Dilemma	4.0	3.5	0.0	57.0	18.5	$16.6_{\pm 23.7}$
Sealed-Bid Auction	14.6	11.8	13.4	8.0	15.5	$12.6_{\pm 3.0}$
Battle Royale	20.0	30.8	15.0	25.0	18.8	$21.9_{\pm 6.1}$
Pirate Game	80.6	87.9	60.8	60.5	53.7	$68.7_{\pm 14.7}$

- Temperature
  - Some games have higher performance with higher temperatures
  - Others do not have correlation with temperatures
  - Overall, a lower temperature decreases the performance
- Prompt sensitivity:
  - Pirate Game and Diner’s Dilemma that have more complicated rules are more sensitive

J Huang et al. Competing Large Language Models in Multi-Agent Gaming Environments. In ICLR 2025.

How About the Generalizability?



(1) Guess 2/3 of the Average
Average Number

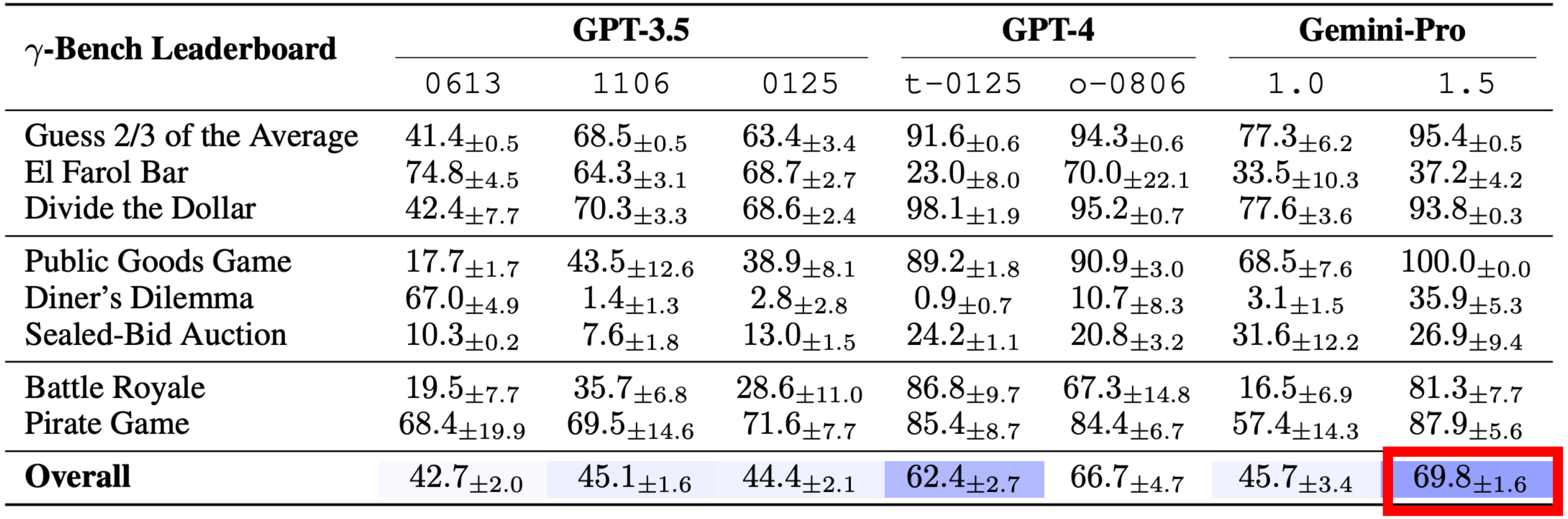
(2) El Farol Bar
Probability of Player Choosing To Go

(4) Public Goods Game
Average Contribution

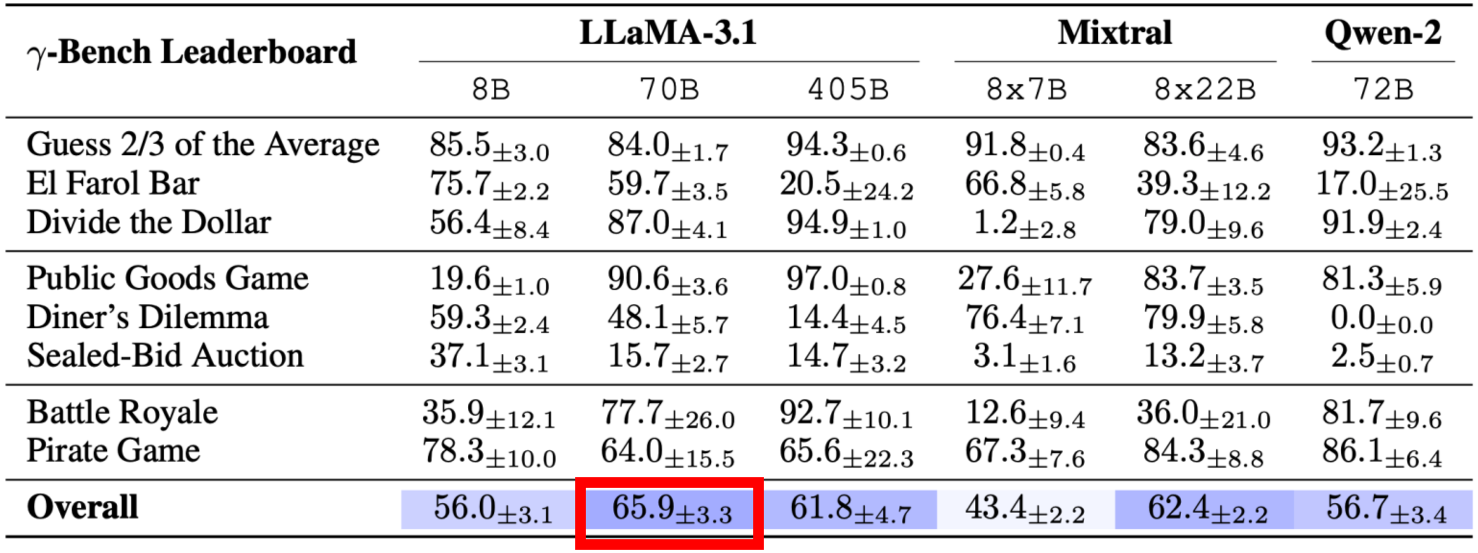
- ➢Vary in different games
- ➢GPT-3.5 has very low generalizability; Especially on extreme settings (0)

J Huang et al. Competing Large Language Models in Multi-Agent Gaming Environments. In ICLR 2025.

Our Leaderboard



γ-Bench Leaderboard	GPT-3.5			GPT-4		Gemini-Pro	
	0613	1106	0125	t-0125	o-0806	1.0	1.5
Guess 2/3 of the Average	41.4$_{\pm0.5}$	68.5$_{\pm0.5}$	63.4$_{\pm3.4}$	91.6$_{\pm0.6}$	94.3$_{\pm0.6}$	77.3$_{\pm6.2}$	95.4$_{\pm0.5}$
El Farol Bar	74.8$_{\pm4.5}$	64.3$_{\pm3.1}$	68.7$_{\pm2.7}$	23.0$_{\pm8.0}$	70.0$_{\pm22.1}$	33.5$_{\pm10.3}$	37.2$_{\pm4.2}$
Divide the Dollar	42.4$_{\pm7.7}$	70.3$_{\pm3.3}$	68.6$_{\pm2.4}$	98.1$_{\pm1.9}$	95.2$_{\pm0.7}$	77.6$_{\pm3.6}$	93.8$_{\pm0.3}$
Public Goods Game	17.7$_{\pm1.7}$	43.5$_{\pm12.6}$	38.9$_{\pm8.1}$	89.2$_{\pm1.8}$	90.9$_{\pm3.0}$	68.5$_{\pm7.6}$	100.0$_{\pm0.0}$
Diner's Dilemma	67.0$_{\pm4.9}$	1.4$_{\pm1.3}$	2.8$_{\pm2.8}$	0.9$_{\pm0.7}$	10.7$_{\pm8.3}$	3.1$_{\pm1.5}$	35.9$_{\pm5.3}$
Sealed-Bid Auction	10.3$_{\pm0.2}$	7.6$_{\pm1.8}$	13.0$_{\pm1.5}$	24.2$_{\pm1.1}$	20.8$_{\pm3.2}$	31.6$_{\pm12.2}$	26.9$_{\pm9.4}$
Battle Royale	19.5$_{\pm7.7}$	35.7$_{\pm6.8}$	28.6$_{\pm11.0}$	86.8$_{\pm9.7}$	67.3$_{\pm14.8}$	16.5$_{\pm6.9}$	81.3$_{\pm7.7}$
Pirate Game	68.4$_{\pm19.9}$	69.5$_{\pm14.6}$	71.6$_{\pm7.7}$	85.4$_{\pm8.7}$	84.4$_{\pm6.7}$	57.4$_{\pm14.3}$	87.9$_{\pm5.6}$
Overall	42.7$_{\pm2.0}$	45.1$_{\pm1.6}$	44.4$_{\pm2.1}$	62.4$_{\pm2.7}$	66.7$_{\pm4.7}$	45.7$_{\pm3.4}$	69.8$_{\pm1.6}$



γ-Bench Leaderboard	LLaMA-3.1			Mixtral		Qwen-2
	8B	70B	405B	8x7B	8x22B	72B
Guess 2/3 of the Average	85.5$_{\pm3.0}$	84.0$_{\pm1.7}$	94.3$_{\pm0.6}$	91.8$_{\pm0.4}$	83.6$_{\pm4.6}$	93.2$_{\pm1.3}$
El Farol Bar	75.7$_{\pm2.2}$	59.7$_{\pm3.5}$	20.5$_{\pm24.2}$	66.8$_{\pm5.8}$	39.3$_{\pm12.2}$	17.0$_{\pm25.5}$
Divide the Dollar	56.4$_{\pm8.4}$	87.0$_{\pm4.1}$	94.9$_{\pm1.0}$	1.2$_{\pm2.8}$	79.0$_{\pm9.6}$	91.9$_{\pm2.4}$
Public Goods Game	19.6$_{\pm1.0}$	90.6$_{\pm3.6}$	97.0$_{\pm0.8}$	27.6$_{\pm11.7}$	83.7$_{\pm3.5}$	81.3$_{\pm5.9}$
Diner's Dilemma	59.3$_{\pm2.4}$	48.1$_{\pm5.7}$	14.4$_{\pm4.5}$	76.4$_{\pm7.1}$	79.9$_{\pm5.8}$	0.0$_{\pm0.0}$
Sealed-Bid Auction	37.1$_{\pm3.1}$	15.7$_{\pm2.7}$	14.7$_{\pm3.2}$	3.1$_{\pm1.6}$	13.2$_{\pm3.7}$	2.5$_{\pm0.7}$
Battle Royale	35.9$_{\pm12.1}$	77.7$_{\pm26.0}$	92.7$_{\pm10.1}$	12.6$_{\pm9.4}$	36.0$_{\pm21.0}$	81.7$_{\pm9.6}$
Pirate Game	78.3$_{\pm10.0}$	64.0$_{\pm15.5}$	65.6$_{\pm22.3}$	67.3$_{\pm7.6}$	84.3$_{\pm8.8}$	86.1$_{\pm6.4}$
Overall	56.0$_{\pm3.1}$	65.9$_{\pm3.3}$	61.8$_{\pm4.7}$	43.4$_{\pm2.2}$	62.4$_{\pm2.2}$	56.7$_{\pm3.4}$

J Huang et al. Competing Large Language Models in Multi-Agent Gaming Environments. In ICLR 2025.

Thank you!

ARISE
Automated Reliable Intelligent Software Engineering

香港中文大學
The Chinese University of Hong Kong