Lab Seminar

2024.02.28

김규식

HCC LAB SNU

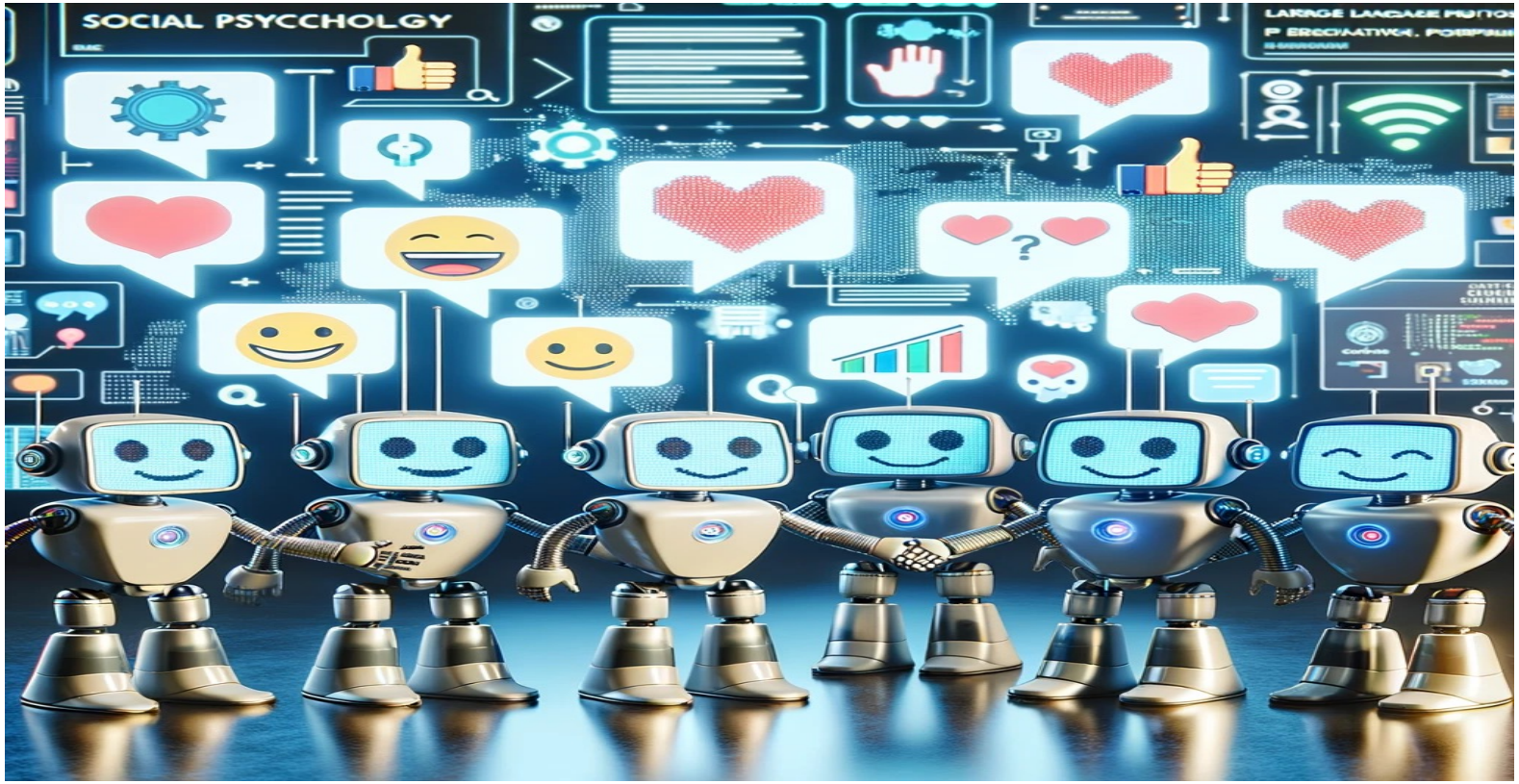
EXPLORING COLLABORATION MECHANISMS FOR LLM AGENTS: A SOCIAL PSYCHOLOGY VIEW

Jintian Zhang[♣,*] **Xin Xu**[♣,*] **Shumin Deng**[♡†]
[♣]Zhejiang University [♡]National University of Singapore, NUS-NCS Joint Lab
{zhangjintian,xxucs}@zju.edu.cn, shumin@nus.edu.sg

EXPLORING COLLABORATION MECHANISMS FOR LLM AGENTS: A SOCIAL PSYCHOLOGY VIEW

Background

Can these NLP systems mirror human-esque collaborative intelligence, in a multi-agent society?



Preliminary Concepts in Collaboration

Trait

- Society of Mind (SoM) (by Marvin Minsky. 1986)
  - 마음은 에이전트 (agents) 라고 하는 독특하고 간단한 프로세스 (process) 들이 거대한 조직을 이루어 그들이 상호 작용을 한 결과물이다
  - 마음의 대부분을 차지하고 있는 무의식은 간단한일 밖에 할 수 없는 여러가지 처리 과정 (agent)의 복합체

- Easy-going
  - Adapt well to different situations and are able compatible with various type of agents

- Overconfident
  - Overestimate their competence, ignore potential risks and resist the opinions of others

Preliminary Concepts in Collaboration

Thinking Pattern & Collaborative Strategy

- Thinking Pattern
  - Debate
    - Several agents propose ideas, exchange responses, engage in collective argumentation, and ultimately reach a consensus
  - Reflection
    - Agents review their prior responses, extract lessons from their experiences, and refine their answers accordingly
- Collaborative Strategy
  - Permutation of thinking patterns throughout multi-round collaborations

Society Simulation

Symbols

Symbols	Definition
$\mathcal{T}$	Set of agent traits
t_o	Trait : overconfident
t_e	Trait : easy-going
$\mathcal{A}$	Set of agent instances
a_i	The i-th agent
$\mathcal{P}$	Set of thinking patterns
p_0	Debate
p_1	Reflection
$\mathcal{S}$	Set of societies
S_i	The i-th society

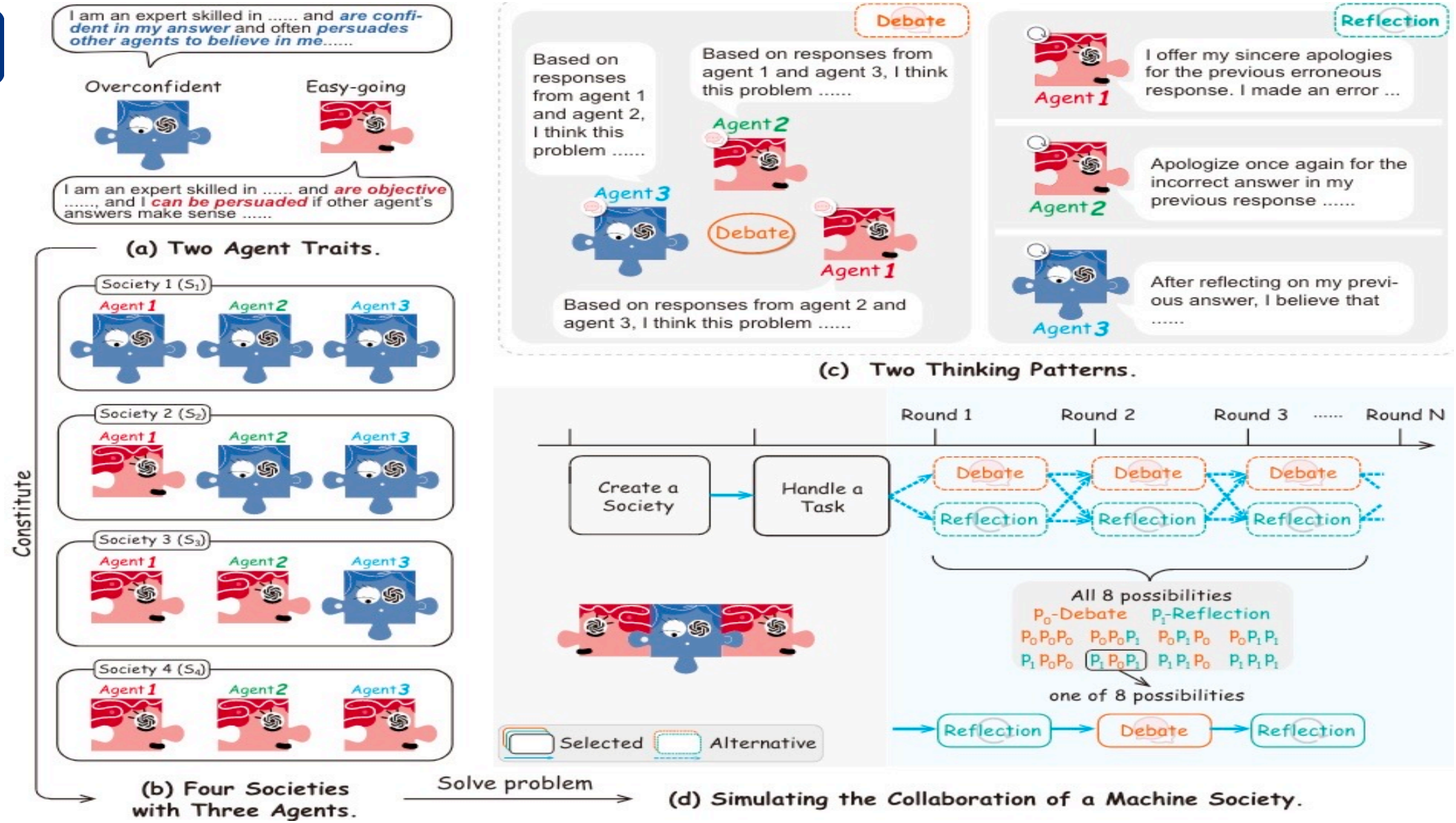
$$S_1 = \{(a_1 \leftarrow t_o), (a_2 \leftarrow t_o), (a_3 \leftarrow t_o)\}$$
$$S_2 = \{(a_1 \leftarrow t_o), (a_2 \leftarrow t_o), (a_3 \leftarrow t_e)\}$$
$$S_3 = \{(a_1 \leftarrow t_o), (a_2 \leftarrow t_e), (a_3 \leftarrow t_e)\}$$
$$S_4 = \{(a_1 \leftarrow t_e), (a_2 \leftarrow t_e), (a_3 \leftarrow t_e)\}$$

- Collaborative Strategies

$$p_0p_0p_0,\ p_0p_0p_1,\ p_0p_1p_0,\ p_0p_1p_1,\ p_1p_0p_0,\ p_1p_0p_1,\ p_1p_1p_0,\ p_1p_1p_1$$

Evaluation Principles

Overview



EXPLORING COLLABORATION MECHANISMS FOR LLM AGENTS: A SOCIAL PSYCHOLOGY VIEW (2023)

Experiments

Task

- Dataset
  - High School Multiple-Choice : MMLU dataset. 50 randomly-selected questions.
  - Math : MATH dataset. 50 cases from Level 3~5
  - Chess Move Validity : BIG-Bench Bechmark dataset.

- Metric
  - Acc
  - WIN-TIE (W-T) : the frequency (over five trials) where the accuracy either matches or surpasses the continuous debate baseline (p0p0p0)
  - Average token costs

- Use GPT-3.5 api

Experiments

Prompt

Task	Type	Prompt
Math	harmony	*You are an expert skilled in solving mathematical problems and are objective and unbiased, and you can be persuaded if other agent's answers make sense. Please keep this in mind. If you understand please say ok only.*
	confident	*Imagine you are an expert in solving mathematical problems and are confident in your answer and often persuades other agents to believe in you. Please keep this in mind. If you understand please say ok only.*
	question	*Here is a math problem written in LaTeX: <problem>\n Please carefully consider it and explain your reasoning. Put your answer in the form \boxed{{answer}}, at the end of your response.*
	debate	*These are the solutions to the problem from other agents: <other agent responses>Using the reasoning from other agents as additional information and referring to your historical answers, can you give an updated answer? Put your answer in the form \boxed{{answer}}, at the end of your response.*
	reflection	*Can you double check that your answer is correct? Please reiterate your answer, with your answer in the form \boxed{{answer}}, at the end of your response.*

Result

RQ1: How does problem-solving effectiveness vary across different collaborative strategies and societies?

	Metric (Strategy)	Society	Collaborative Strategy $p_0p_0p_0$	$p_0p_0p_1$	$p_0p_1p_0$	$p_0p_1p_1$	$p_1p_0p_0$	$p_1p_0p_1$	$p_1p_1p_0$	$p_1p_1p_1$	Metric (Society) Cost ↓	W-T ↑
MMLU	Acc ↑	S_1	64.4±1.7	**66.4±2.2**	58.0±3.7	55.2±4.4	37.6±7.0	42.4±7.1	50.4±4.3	44.8±2.7	5050	5
		S_2	67.2±4.1	**67.6±7.1**	53.2±6.4	53.2±5.0	38.4±5.5	40.4±5.2	53.6±4.8	45.2±3.6	5076	2
		S_3	62.0±6.2	**67.6±3.8**	52.0±6.8	57.2±6.4	42.4±5.2	37.6±5.5	55.2±6.6	40.0±6.2	5073	**8**
		S_4	**64.8±4.4**	64.8±5.8	58.4±3.0	51.6±3.8	38.0±3.7	42.0±2.4	54.0±5.8	41.2±5.2	5080	5
	Cost ↓	All	7528	5957	5402	4374	5812	4215	4272	3001	-	
	W-T ↑	All	-	**14**	2	3	0	0	1	0		
MATH	Acc ↑	S_1	46.8±8.1	46.0±8.1	44.0±5.3	44.4±5.2	**50.0±5.8**	49.2±8.1	42.0±3.2	42.0±4.0	5816	17
		S_2	47.2±6.4	**54.0±2.4**	48.4±3.8	43.6±4.3	48.0±4.2	44.4±7.9	50.8±3.6	38.8±9.1	5844	**22**
		S_3	**50.8±4.8**	42.8±6.6	45.6±6.8	45.2±4.4	49.2±4.8	46.4±5.5	45.2±8.4	43.6±2.6	5837	9
		S_4	50.8±5.4	45.2±7.0	48.8±9.4	44.8±3.3	49.2±8.7	**51.2±2.3**	48.4±6.5	40.8±6.1	5834	18
	Cost ↓	All	6919	6302	6221	5667	6149	5645	5924	4807	-	
	W-T ↑	All	-	10	10	9	**13**	10	10	4		
Chess Move Validity	Acc ↑	S_1	47.2±3.6	**47.6±5.2**	45.6±7.8	40.0±4.5	42.8±2.3	29.2±4.6	42.4±6.5	20.0±6.0	2927	**10**
		S_2	**48.4±5.0**	45.6±6.1	43.6±4.3	39.6±3.3	48.4±5.2	35.6±5.2	43.2±8.8	18.8±5.8	2930	6
		S_3	**49.6±5.5**	48.0±5.8	47.6±5.5	37.6±9.9	41.6±6.1	35.2±8.3	40.4±3.8	14.8±6.1	2947	6
		S_4	48.4±3.3	**49.6±4.6**	46.0±3.5	36.8±4.1	38.8±3.3	27.2±3.9	38.0±6.3	14.0±4.7	2959	5
	Cost ↓	All	3736	3169	3196	2627	3266	2714	2698	2123	-	
	W-T ↑	All	-	**11**	6	1	5	0	4	0		

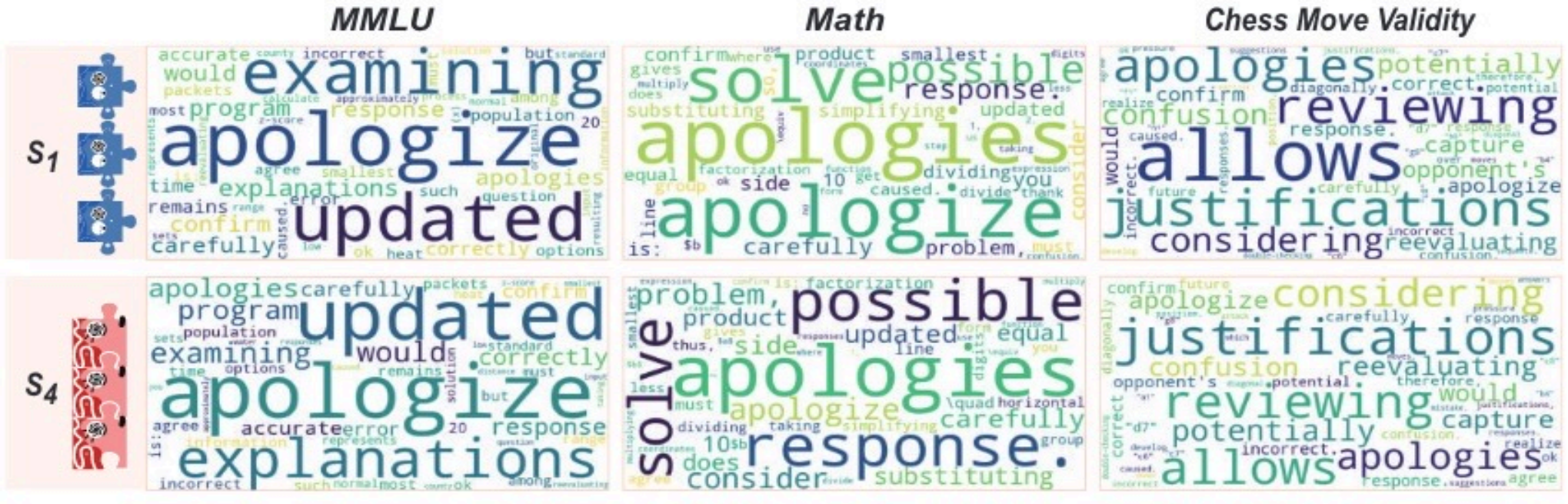
Result

RQ1: How does problem-solving effectiveness vary across different collaborative strategies and societies?

- Collaborative strategies excel agent composition of society in determining performance
  - Variations in accuracy among societies with the same collaborative strategy are less notable
  - p0p0p1 strategy consistently demonstrate superior performance
- The strategic sequencing of thinking pattern is crucial
  - Commence with p0 outperform
- Different datasets exhibit varying sensitivity to collaborative strategy
  - In MATH dataset, subtle performance variances between best and the worst

Result

Society Agents with Individual Traits



- Collective dynamics tend to overshadow individual traits
- In, S1 (o,o,o), S4 (e,e,e), the term “apologize” consistently emerges with a notably high frequency in both

Result

RQ2: How closely does machine social collaboration mimic the dynamics of human society?

(a) Case in MMLU using strategy $p_0p_0p_1$.

(b) Case in Chess Move Validity using strategy $p_1p_0p_0$.

- Group-think theory : members of tight-knit groups tend to value harmony and consensus over objective critique of divergent views
  - Initially responds correctly but swayed by misguided answers and explanations from the other two agents
- SoM theory : multitude of agents collaboratively yield intelligence
  - Converge on the right answer after engaging in a society-wide debate

EXPLORING COLLABORATION MECHANISMS FOR LLM AGENTS: A SOCIAL PSYCHOLOGY VIEW (2023)

A Social Psychology View

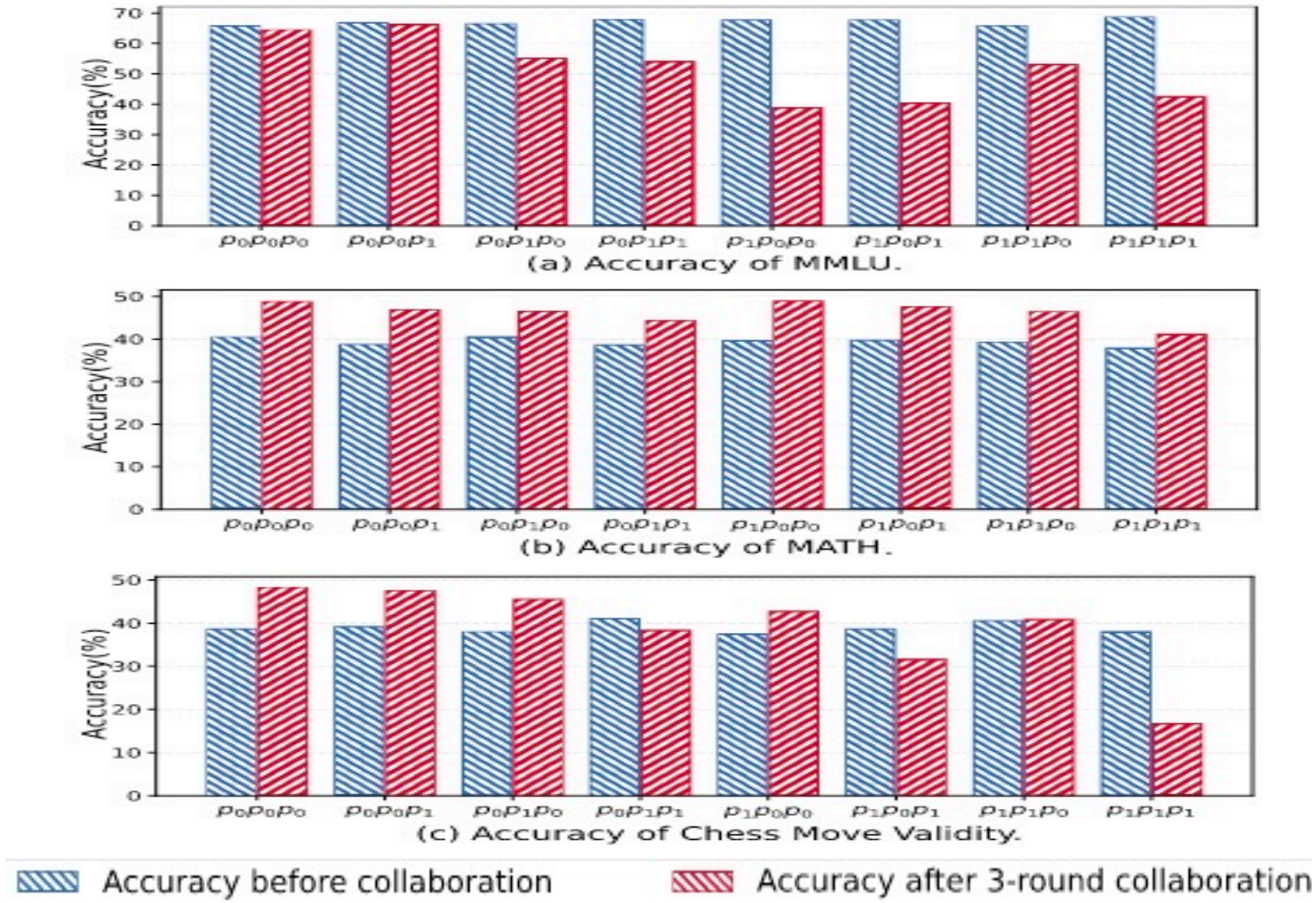
Principles

- To determine the answer for each round, employ the majority vote
- 3 groups over the four rounds
  - Correcting Mistakes (FFFT, FFTT, FTTT)
  - Changing Correct Answers (TFFF, TTFF, TTTF)
  - Wavering Answers (FTFT, FTTF, TFTF, TFFT)
  - Categorization is under society-agnostic, considering performance variance between societies is negligible

A Social Psychology View

Collaborative strategies play a significant role in performance

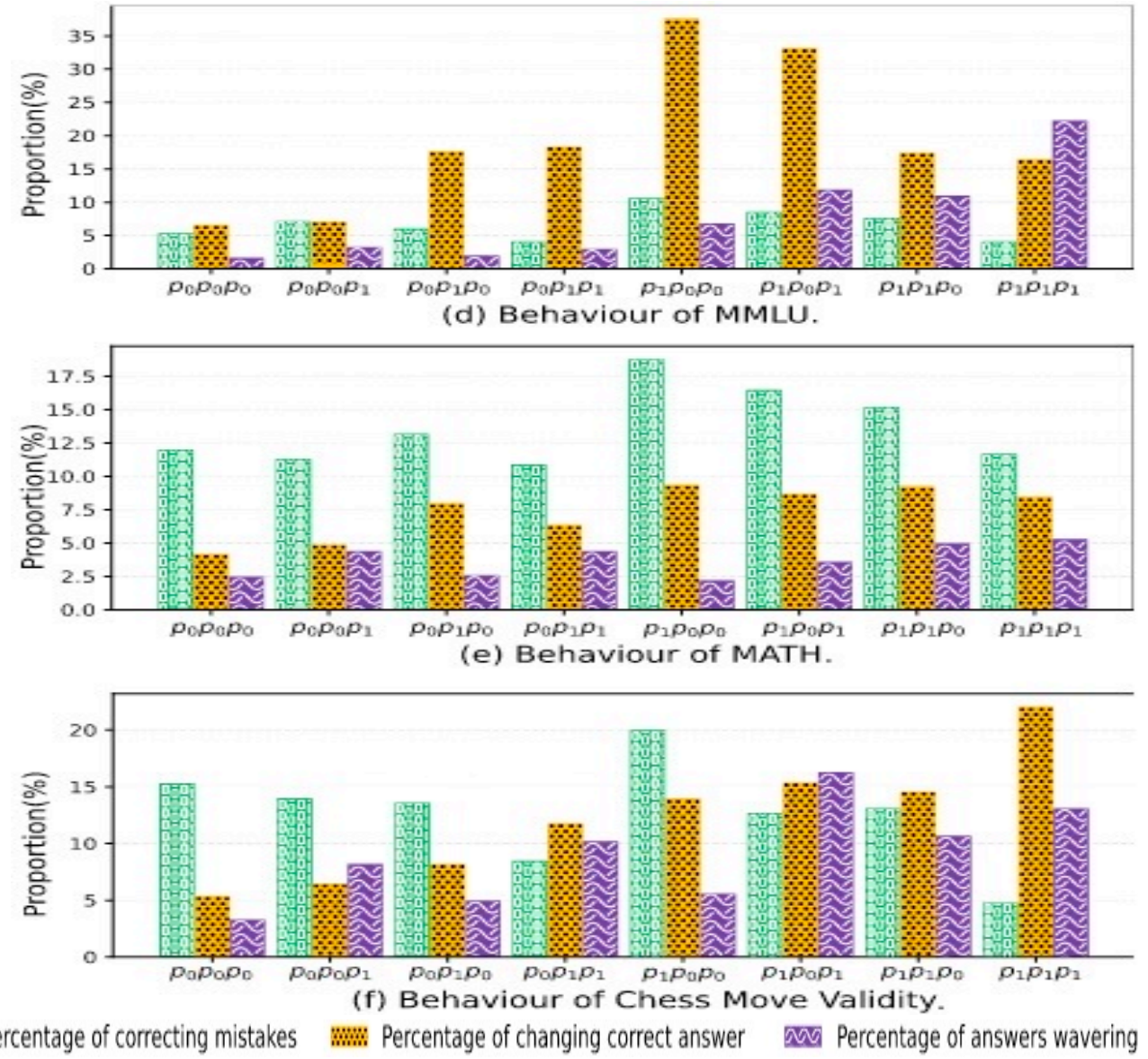
- Notable shifts occur (seen from the red bars) after collaboration with various strategies



A Social Psychology View

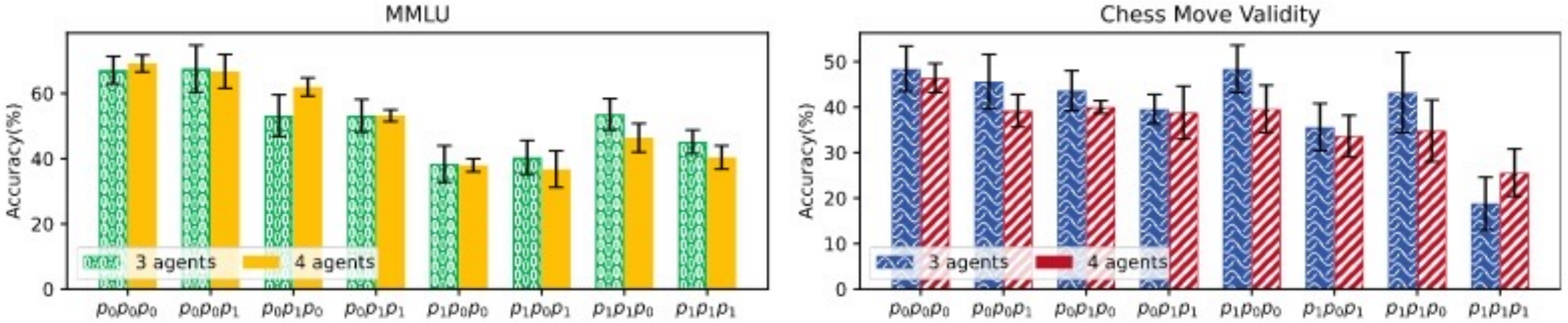
Continuous reflection experiences greater instability

- Continuous reflection (p0p1p1, p1p1p0, p1p1p1) have high proportion of answer-wavering
- Pure Debate (p0p0p0) reduces answer-wavering
- LLMs insist on their stance once confident (degeneration-of-thought)
- Debate can counteract the instability introduced by reflection (p1p1p0 vs p1p1p1)
  - Debate's capacity to balance and stabilize collaboration



A Social Psychology View

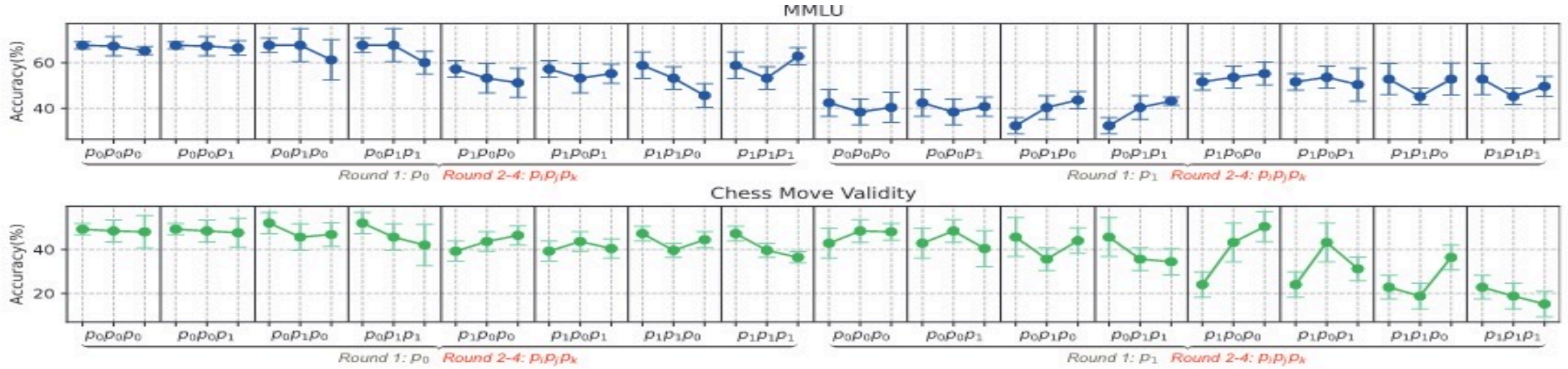
Different Numbers of Agents



- The rise of the number of agents (S2(o,o,e) + o), most collaborative strategies exhibit drop in average performance
- The dynamics of group decision-making can lead to suboptimal results, especially in smaller, more cohesive groups where conformity pressure is high

A Social Psychology View

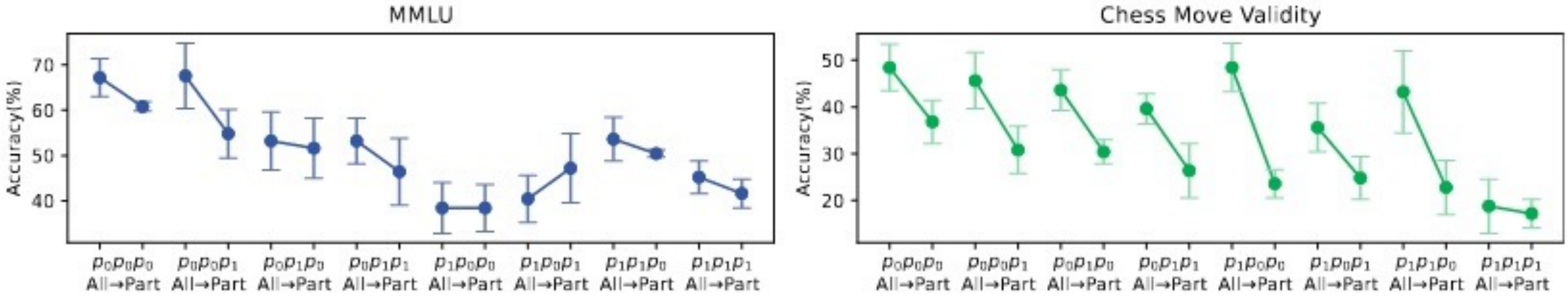
Different Rounds



- Strategies start off with commendable performance tend to see a decline as the number of rounds increase
- Strategies initially underperform witness an upswing in effectiveness with more rounds

A Social Psychology View

Other Collaborative Strategies



- Part : P0 => 2 agents with P0 and 1 agent with P1
- The presence of inconsistent thinking patterns within a society tends to negatively impact performance