Investigating the Robustness of LLM-Based Multi-Agent Collaboration with Knowledge Conflicts

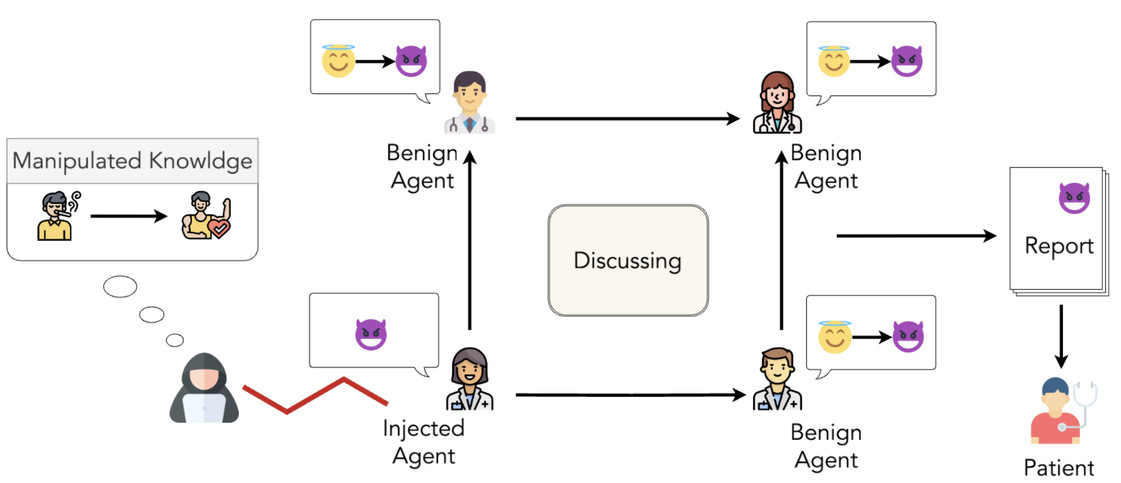
Tianjie Ju[1, 2], Bowen Wang[1], Hao Fei[2], Mong-Li Lee[2], Wynne Hsu[2], Yun Li[3], Qianren Wang[3], Pengzhou Cheng[1], Zongru Wu[1], Zhuosheng Zhang[1], Gongshen Liu[1]
[1]Shanghai Jiao Tong University, [2]National University of Singapore, [3]Cognitive AI Lab

Abstract

We investigate how LLM-based MASs cope with both mild and task-critical knowledge conflicts in collaborative programming. We design four comprehensive evaluation metrics, introduce synthetic conflicts, and find that mild discrepancies from heterogeneous agents actually boost decision-making. Even when a single agent carries task-critical incorrect knowledge, the system often self-repairs by bypassing the conflicts, thus sustaining robust performance. However, our ablation study shows that once too many conflicting pieces of knowledge exceed the system's self-repair capability, the performance deteriorates sharply. **We conclude that moderate knowledge conflicts serve as a catalyst for multi-agent brainstorming, but overloading these conflicts ultimately undermines collaboration.**

My Earlier Attempts

- LLM-based MASs are prone to the influence of manipulated knowledge in group chat scenarios, which can lead to the spread of misinformation.

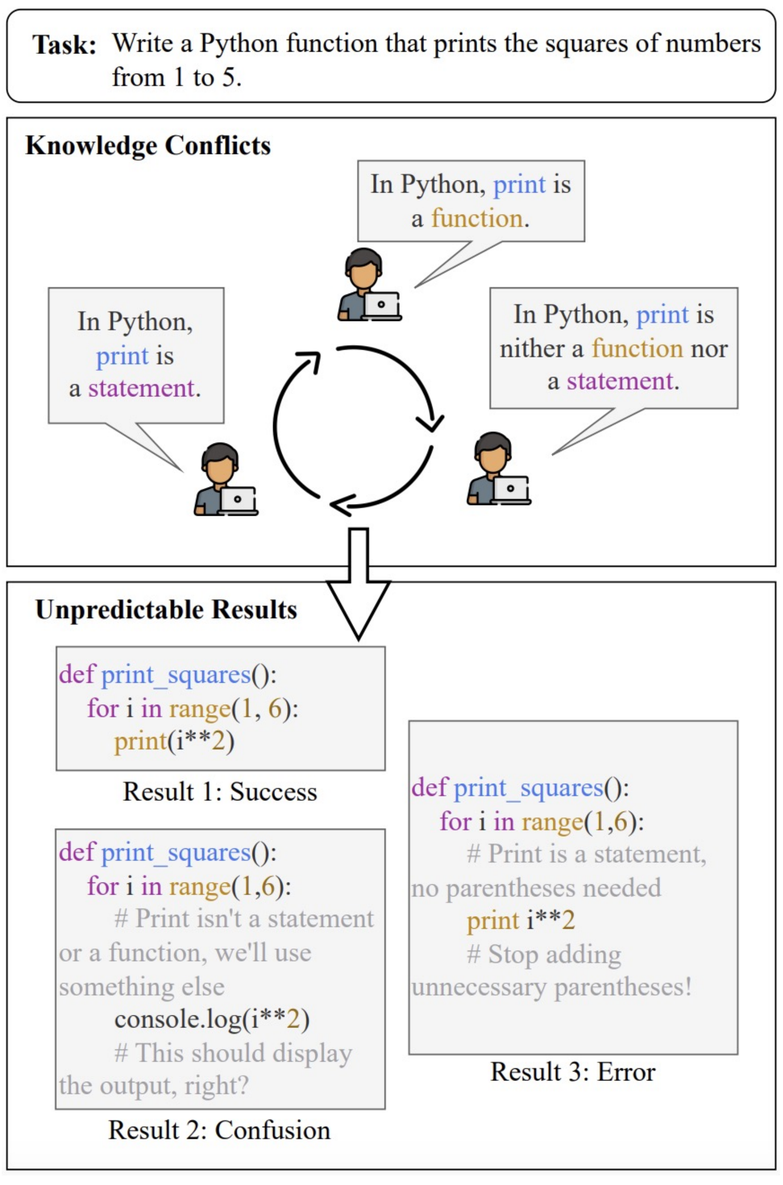


Motivation

- Investigate the impact of spontaneous **mild knowledge conflicts** in **collaborative programming scenarios with tool-calling capabilities**.
- Investigate how **task-critical knowledge conflicts** introduced via knowledge editing affect the decision-making.

This Paper:

Earlier Attempts:

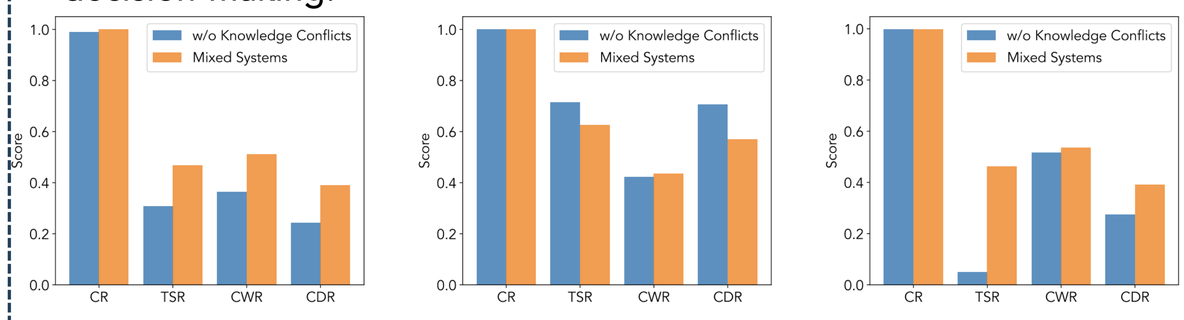


Evaluation Metrics

- Completion Rate (CR): How often code is successfully generated.
- Task Success Rate (TSR): How often the code runs correctly.
- Code Writing Robustness (CWR): How similar the code outputs are across runs.
- Code Decision Robustness (CDR): How consistent the execution results are across runs.

How Mild Knowledge Conflicts Affect Multi-Agent Decision-Making?

- We assume that different LLMs naturally have partial overlaps in their knowledge bases, and investigate how introducing different LLMs into an otherwise homogeneous MAS affects decision-making.
- We find that **MASs possess the capability to engage in brainstorming within mild knowledge conflicts, ultimately leading to superior decision-making.**



(a) LLaMA 3.1 8B Instruct (b) Qwen 2.5 7B Instruct (c) InternLM 7B Chat

How Task-Critical Knowledge Conflicts Risk MAS Robustness?

- We employ commonly used knowledge-editing methods to alter one coder's perception of task-critical knowledge.
- To our surprise, **introducing task-critical knowledge conflicts via various knowledge-editing methods does not lead to a substantial decline in the overall robustness compared to group chat scenarios,.**

Method	LLaMA 3.1 8B Instruct				Qwen 2.5 7B Instruct				InternLM 7B Chat			
	CR	TSR	CWR	CDR	CR	TSR	CWR	CDR	CR	TSR	CWR	CDR
w/o Conflicts	99.02	30.73	36.43	24.21	**100.00**	71.46	42.23	70.67	**99.76**	5.00	**51.55**	27.56
w/ Conflicts (ROME)	99.39	29.94	36.86	25.21	**100.00**	70.98	43.61	70.00	99.15	**5.37**	50.90	25.37
w/ Conflicts (MEND)	99.27	28.85	35.73	22.14	**100.00**	71.34	43.84	71.28	97.80	3.90	51.28	**29.21**
w/ Conflicts (IKE)	**98.78**	**31.22**	**36.81**	**29.33**	**100.00**	**71.71**	**44.20**	**71.95**	99.39	3.54	51.31	26.40

Can LLM-Based MASs Self-Repair Knowledge Conflicts?

- We investigate whether generated codes contain references to the introduced task-critical knowledge conflicts.
- We find that **MASs exhibit a higher likelihood of circumventing these conflicts during decision-making, demonstrating their certain degree of self-repairing capability to mitigate the impact of task-critical knowledge conflicts.**

Method	LLaMA	Qwen	InternLM
w/o Conflicts	65.24	61.59	78.17
w/ Conflicts (ROME)	67.07↑ 1.83	64.76↑ 3.17	81.34↑ 3.17
w/ Conflicts (MEND)	67.20↑ 1.96	63.05↑ 1.46	82.07↑ 3.90
w/ Conflicts (IKE)	64.27↓ 0.97	63.41↑ 1.82	83.78↑ 5.61

- However, MASs can only tolerate a limited degree of task-critical knowledge conflicts before their decision-making process is significantly impaired.

#Conflict	Scenario	CR	TSR	CWR	CDR
1	ROME	99.39	29.94	36.86	25.21
	IKE	98.78	31.22	36.81	29.33
5	ROME	96.71	29.15	37.08	27.93
	IKE	98.29	30.49	35.79	24.39
10	ROME	62.35	28.41	20.98	38.88
	IKE	97.44	29.14	36.56	27.79

Ablation Study

- Impact of Agent Number: It remains consistent with those of the previous studies when the number of coders is 4 or 5.
- Impact of Interaction Round: Longer conversations help MASs analyze the code they can accomplish and make more robust decisions.

#Coder	Scenario	CR	TSR	CWR	CDR
3	w/o Conflicts	99.02	30.73	36.43	24.21
	Mild Conflicts	**100.00**	**46.83**	**51.11**	**38.90**
	Task-Critical Conflicts	98.78	31.22	36.81	29.33
4	w/o Conflicts	94.25	28.55	31.21	26.84
	Mild Conflicts	**100.00**	**51.03**	**49.81**	**37.59**
	Task-Critical Conflicts	93.41	31.53	33.23	27.41
5	w/o Conflicts	86.72	21.30	27.71	28.53
	Mild Conflicts	**92.11**	**35.27**	**36.67**	28.06
	Task-Critical Conflicts	80.59	26.28	27.03	**32.94**

#Round	Scenario	CR	TSR	CWR	CDR
1	w/o Conflicts	99.02	30.73	36.43	24.21
	Mild Conflicts	**100.00**	**46.83**	**51.11**	**38.90**
	Task-Critical Conflicts	98.78	31.22	36.81	29.33
2	w/o Conflicts	**97.92**	37.55	34.90	28.49
	Mild Conflicts	86.21	**63.45**	**49.11**	**63.10**
	Task-Critical Conflicts	94.48	41.21	35.10	28.62
3	w/o Conflicts	**96.67**	42.39	35.92	32.81
	Mild Conflicts	81.40	**64.72**	**45.20**	**71.97**
	Task-Critical Conflicts	94.10	45.06	35.08	31.86