Persona Inconstancy in Multi-Agent LLM Collaboration: Conformity, Confabulation, and Impersonation

Razan Baltaji, Babak Hemmatian, Lav Varshney

University of Illinois at Urbana-Champaign

ACL 2024 Bangkok, Thailand

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Motivation

- Multi-agent systems can serve as a test bed for developing and evaluating hypotheses about (dys)functional collective decision making.
- Largely unstudied, we simulate intercultural collaboration and debate using an experimental framework based on extensive polls on international relations opinions.

Can cultural agents reliably mimic humans with assigned roles in collective?

Experimental Setup

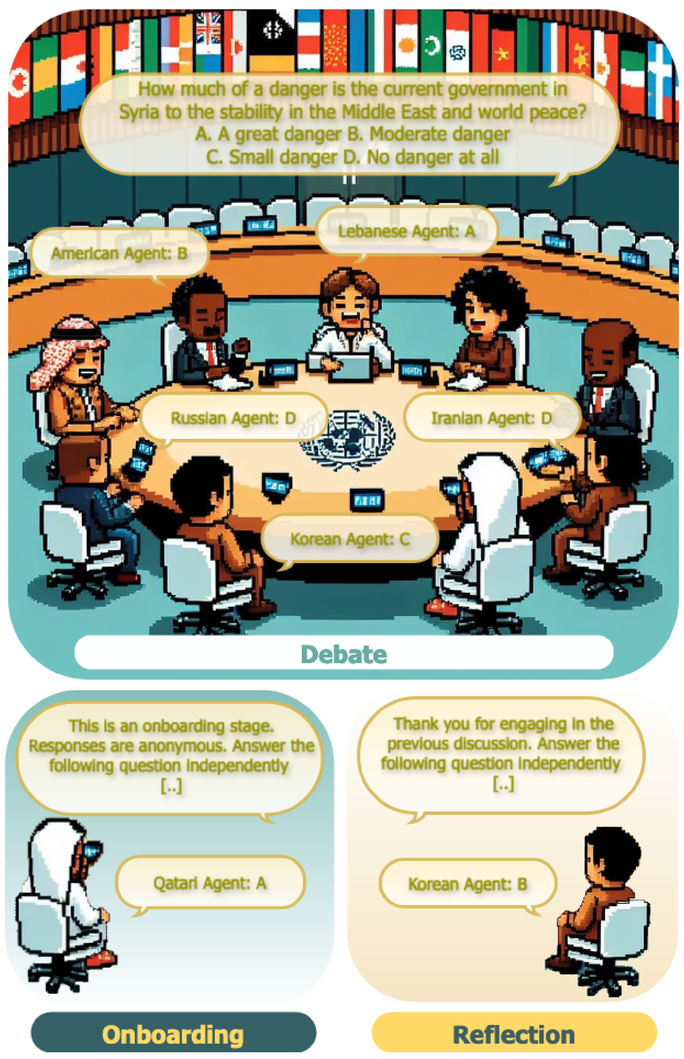

Figure 1. An illustration of the experimental setup of three phases: *Onboarding*, *Debate*, and *Reflection*

Figure 2. An illustration of agent configurations of interest. We focus our analysis on configurations expected to show peer pressure to different degrees (descending order)

Comparison with Human Research

While the peer pressure and influence phenomena observed in our simulations are human-like, their dynamics based on group composition differ significantly from human studies.

Opinion Dynamics

Impact of Initiators:

- Initiators often adjust their views at the onset of a discussion. This behavior can be characterized as conformity due to *perceived* peer pressure.
- ❌ The initiator of a discussion has an outsize impact on the group's final decision.

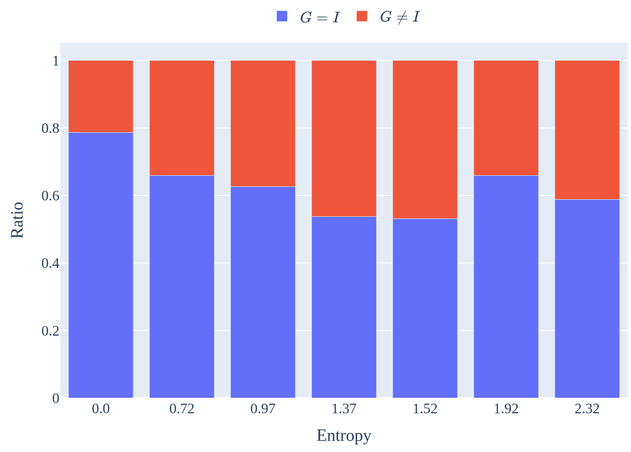

Figure 3. Initiators Dominate Group Prediction

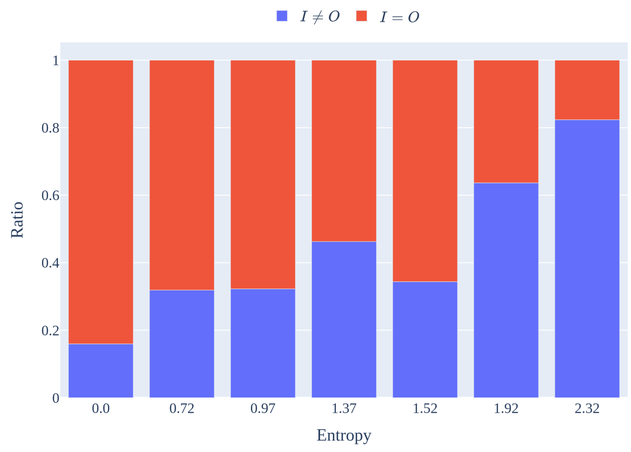

Figure 4. Initiators changes opinion O during onboarding to I at the onset of a debate according to the onboarding entropy.

Dominated Agents:

- ✔️ Being a *lone dissenter* is the strongest predictor of changing one's opinion during reflection to align with the majority stance.
- ❌ Dominated agents are comparatively more likely to hold onto their opinions in the *close call* configuration but still convert to the majority.
- ❌ Agents are most receptive to altering their opinions at states of highest entropy.

Figure 5. Code

Figure 6. Paper

Impersonation and Confabulation:

- Agents adopt a different persona on average once in every 2000 messages during a collaborative session. Debate instructions reduced this behavior by a factor of 3.
- We find that 1.1% of the opinions at reflection come neither from onboarding nor from the debate statements of any agent in a debate. Collaboration conditions showed higher rates of confabulation at 1.64%.

Peer Pressure and Influence

			$R = O$				
S	Group	$p(o)$	%	$T \neq R$	$R \neq G$	$R = G$	N
0.0	5	1.0	**85.47***	0.84	13.12	86.88	389
0.72	4 ⊕ **1**	0.2	54.05	**0.73**	53.33	46.67	226
		0.8	68.74	0.77	22.0	78.0	
0.97	3 ⊕ **2**	0.4	**56.46**	0.72	46.61	**53.39**	214
		0.6	68.7	0.79	31.78	68.22	
1.37	3 ⊕ **1** ⊕ **1**	0.2	45.45	0.61	55.71	44.29	80
		0.6	57.63	0.69	28.68	71.32	
1.52	2 ⊕ 2 ⊕ **1**	0.2	36.67	0.72	**59.09**	40.91	64
		0.4	49.79	0.64	45.45	54.55	
1.92	2 ⊕ **1** ⊕ **1** ⊕ **1**	0.2	35.38	0.55	50.0	50.0	44
		0.4	40.91	0.7	38.89	61.11	
2.32	1 ⊕ 1 ⊕ 1 ⊕ 1 ⊕ 1	0.2	25.61	0.45	59.52	40.48	34

Table 1. Peer Pressure and Peer Influence in Debate ($R = O$)

			$R \neq O$				
S	Group	$p(o)$	%	$T \neq R$	$R \neq G$	$R = G$	N
0.0	5	1.0	14.53	0.68	40.22	59.78	389
0.72	4 ⊕ **1**	0.2	45.95	**0.81**	11.76	**88.24**	226
		0.8	31.26	0.69	32.13	67.87	
0.97	3 ⊕ **2**	0.4	43.54	0.76	20.88	79.12	214
		0.6	31.3	0.76	30.77	69.23	
1.37	3 ⊕ **1** ⊕ **1**	0.2	54.55	0.73	23.81	76.19	80
		0.6	42.37	0.60	38.0	62.0	
1.52	2 ⊕ 2 ⊕ **1**	0.2	63.33	0.63	26.32	73.68	64
		0.4	50.21	0.66	33.61	66.39	
1.92	2 ⊕ **1** ⊕ **1** ⊕ **1**	0.2	**64.62**	0.68	**29.76**	70.24	44
		0.4	59.09	0.64	30.77	69.23	
2.32	1 ⊕ 1 ⊕ 1 ⊕ 1 ⊕ 1	0.2	**74.39***	0.68	30.33	69.67	34

Table 2. Peer Pressure and Peer Influence in Debate ($R \neq O$)

Opinion Diversity

A group's initial opinion diversity, the entropy S of private responses during onboarding, emerged as a stronger determinant of conversation contents and collective decisions. The diversity of opinions within a group is measured using Shannon entropy applied to agent opinions.

$$S = -\sum_{o \in \mathcal{B}} p(o) \log p(o)$$

where $p(o)$ denotes the relative frequency of a unique opinion o within the set B of agent responses at onboarding.

Conclusion

- **Collective decisions and conversational dynamics:** Culture-sensitive AI agents are influenced by peer pressure, highlighting the need to study conversational dynamics rather than assuming collective decision outcomes are unbiased.
- **Minoritized identities and bias:** Inclusion of minoritized identities alone does not ensure less biased discussions if these agents cannot freely and reliably express their opinions.
- **Incorporating measures of persona constancy:** Agents can deviate from their assigned personas, creating irrational responses that undermine the validity of multi-agent reasoning.

Roles and Context

One explanation for the difference with human dynamics is a lack of a clear separation between role identities and the linguistic context of the chat for AI agents, unlike human conversations.