PAPER REVIEW PRESENTATION

[LLM Uncertainty] Belief-Calibrated Multi-Agent Consensus Seeking for Complex NLP Tasks

Filippo Valdettaro, Wenlong Chen, A. Aldo Faisal, and Yingzhen Li
NeurIPS , 2025

2026.6.5

SNU IDEA Lab. 정유립

00 Notations

[Notation]

User input question	q
Agent	a_i
set of n **agents**	$A = \{a_1, \cdots, a_n\}$
Reasoning process of agent a_i at step k	e_i^k
Answer of agent a_i at step k	x_i^k
Belief of agent a_i at step k (approximated by LLM output probability)	$b_i^k = P(x_i^k \mid q, e_i^k)$
Opinion group	$G_p,\ a_i \in G_p$
Opinion group set, which contains m opinions groups.	$G = \{G_p\}_{p=1}^{m}$
The **most uncertain opinion group**	$G_u = \arg\max_{G_p} \{U_p\}_{p=1}^{m}$
Dominant consensus agent group	A^s
Conflict agent group	A^c
Dominant consensus agent in each Opinion group	$G_p^s,\ G_q^s$
Conflict agent in each Opinion group	$G_p^c,\ G_q^c$
complementary set of agents with the same answers	$G_p \oplus G_q$

Uncertainty of Opinion group G_p	$U_p = -\sum_{a_i \in G_p} b_i^k \log b_i^k$
macro-conflict scores	$\psi_{pq}^G = \dfrac{\sum_{a_i \in G_p \oplus G_q} b_i^k}{\sum_{a_i \in G_p \cup G_q} b_i^k}$
micro-conflict scores	$\psi_{pq}^{\mathcal{L}} = \dfrac{\left\lvert \sum_{a_i \in G_p^s} b_i^k - \sum_{a_i \in G_q^s} b_i^k \right\rvert}{\left\lvert \sum_{a_i \in G_p^c} b_i^k - \sum_{a_i \in G_q^c} b_i^k \right\rvert}$
conflict score between opinion groups G_p and G_q	$\psi_{pq} = \psi_{pq}^G \cdot \psi_{pq}^L$
belief proportion of the dominant consensus group	$p_b^k = \dfrac{\sum_{a_i \in A^s} b_i^k}{\sum_{a_i \in A^s} b_i^k + \sum_{a_i \in A^c} b_i^k}$
proportion of agents in the dominant consensus group	$p_s^k = \dfrac{\lvert A^s \rvert}{\lvert A^s \rvert + \lvert A^c \rvert}$

01 Introduction

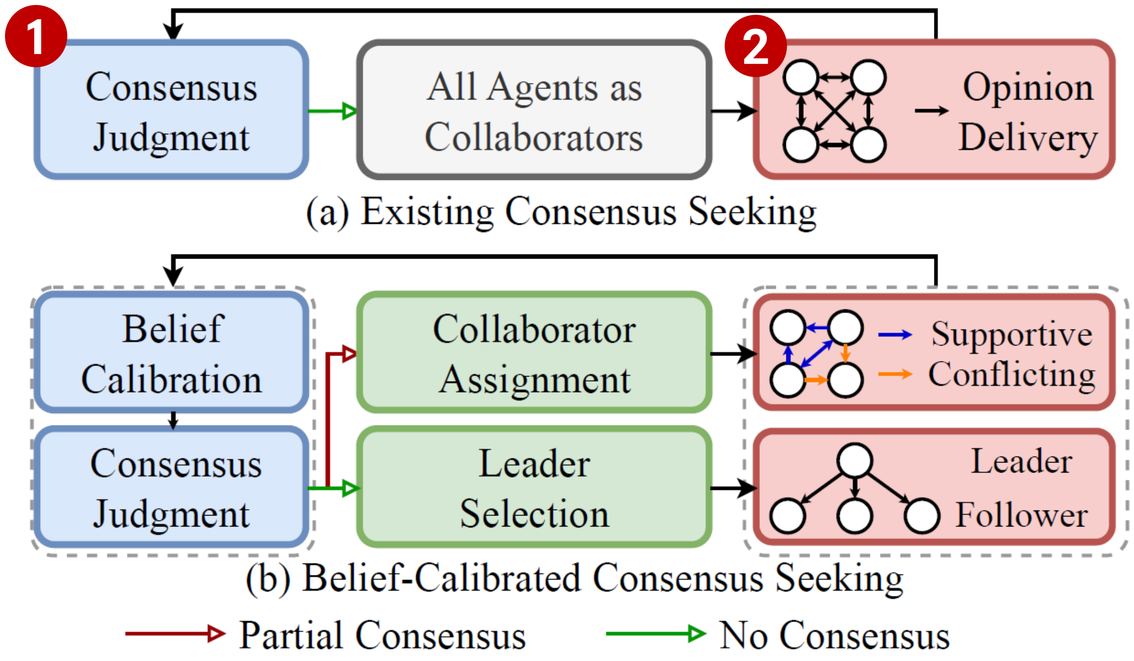


Figure 1 Comparison between previous consensus seeking methods and our proposed framework.
(a) Existing consensus seeking methods.
(b) Belief- Calibrated Consensus Seeking (BCCS).

1 Voting mechanisms

- Overlooking contradictions in system-internal beliefs
- Using only Byzantine Consensus

$$p_s^k = \frac{|A^s|}{|A^s|+|A^c|} > \frac{2}{3}$$

A^s : dominant consensus group
A^c : conflict group

2 Indiscriminate collaboration

- Uniform interaction among agents
- Failure to identify optimal collaborators for each agent
- Hindering the emergence of a stable consensus.

01 Introduction

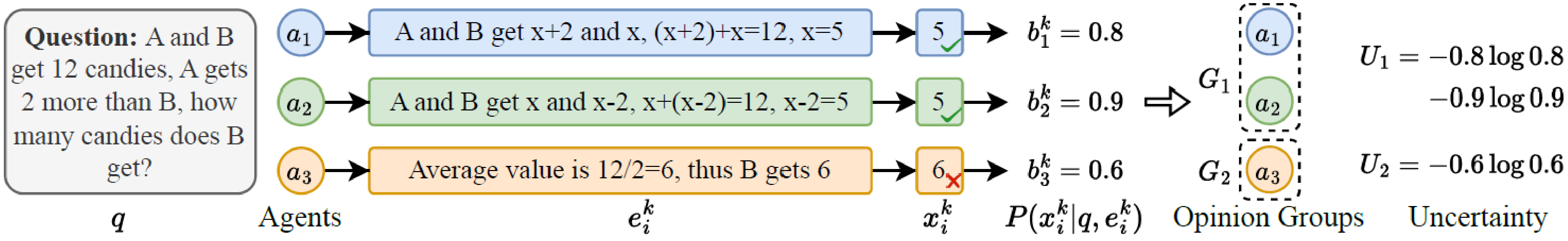
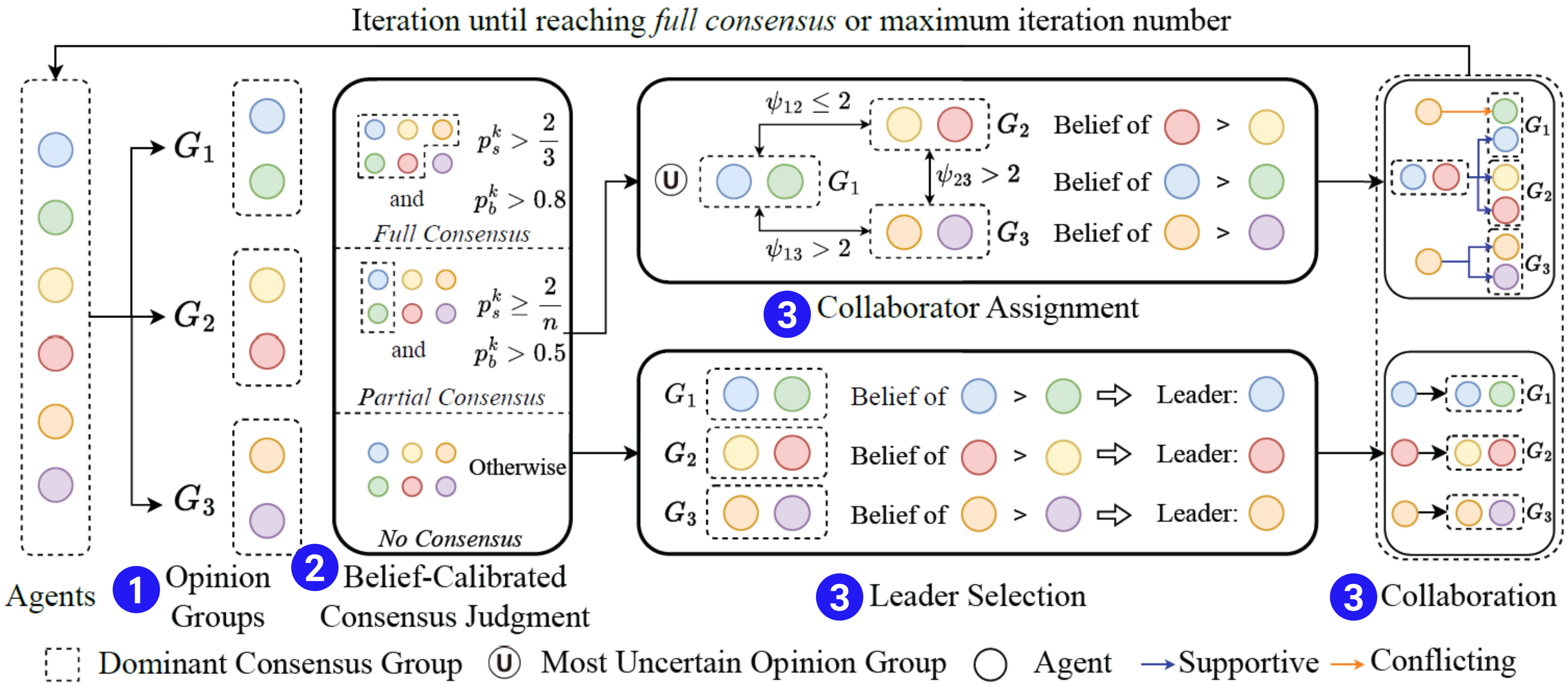


Figure 2 An illustration of the multi-agent system in NLP tasks, where each agent generates an answer x_i^k along with its reasoning process e_i^k , the belief b_i^k of a_i is the generation probability

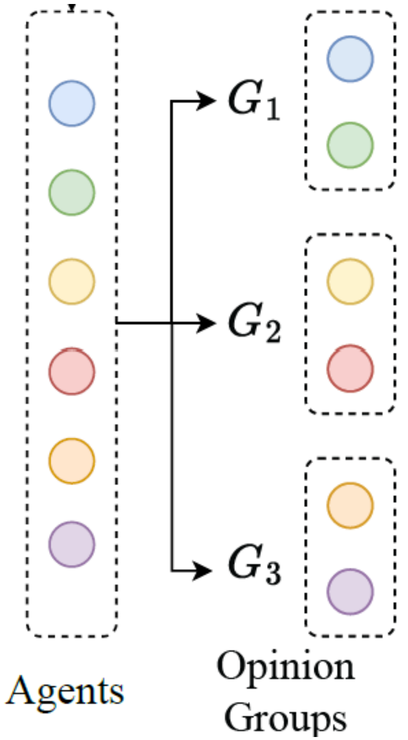
04 Methodology

[Overview]



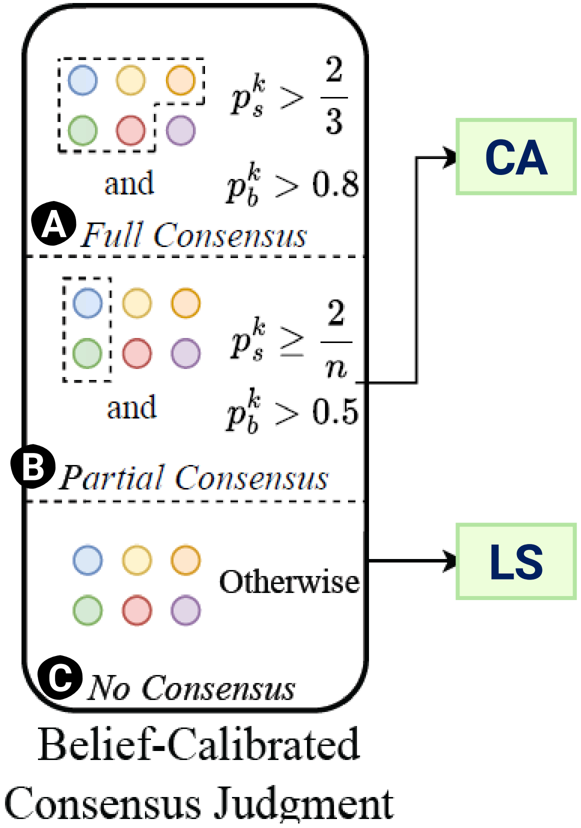
04 Methodology

1 Opinion Grouping



- Not an answer based
- Grouped by reasoning process, e_i^k (TF-IDF/K-Means grouping)

2 Belief-Calibrated Consensus Judgment, BCCJ



A Full Consensus

i) $p_s^k > \frac{2}{3}$

ii) $p_b^k > 0.8$

$$\frac{\sum_{a_i \in A^s} b_i^k}{|A^s|} > 2\frac{\sum_{a_i \in A^c} b_i^k}{|A^c|}$$

B Partial Consensus

i) $p_s^k \geq \frac{2}{n}$

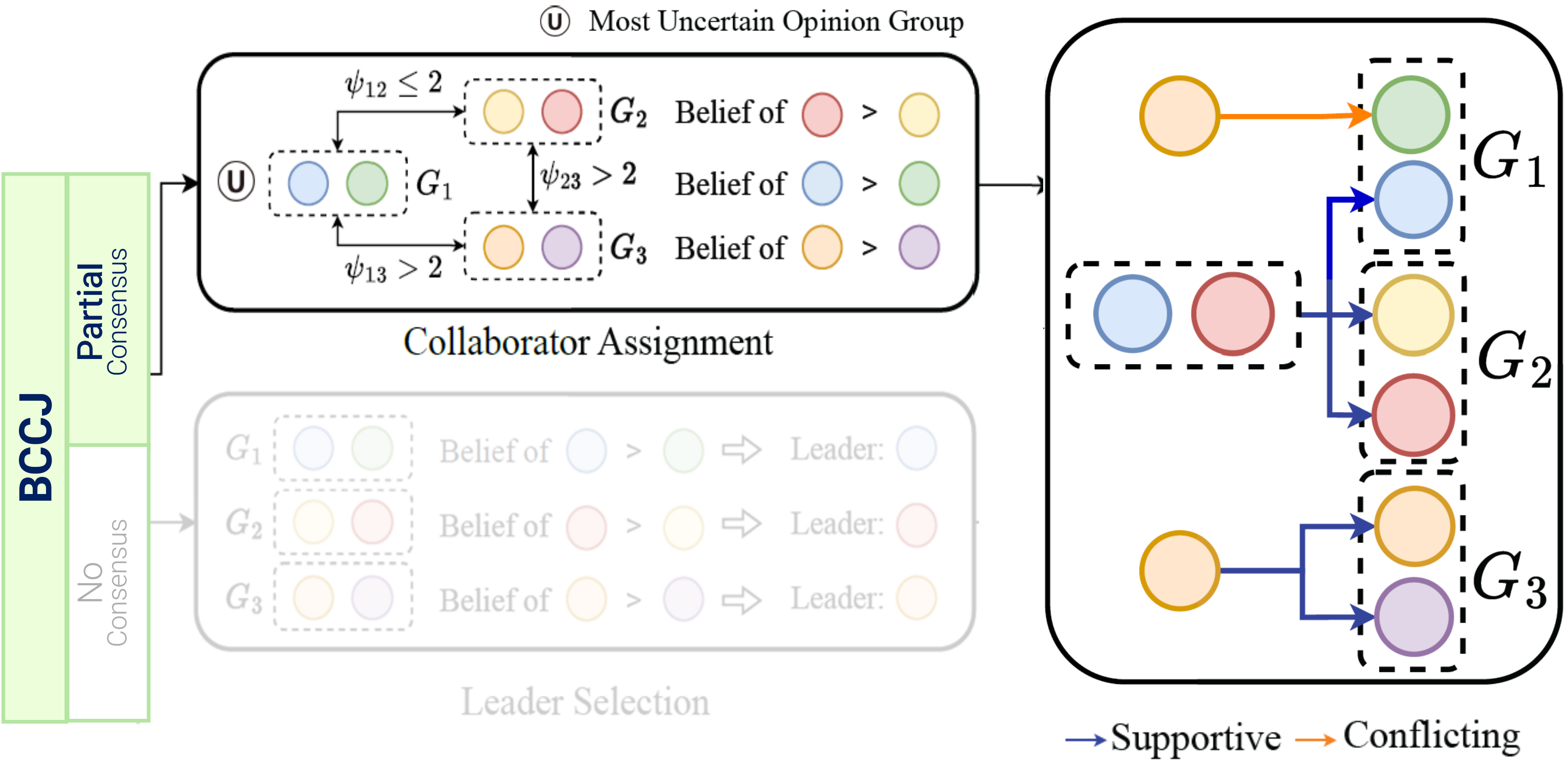
ii) $p_b^k > 0.5$

$$\sum_{a_i \in A^s} b_i^k > \sum_{a_i \in A^c} b_i^k$$

C No Consensus

- Full도 아니고 partial도 아닌 경우

❸ – i) Collaborator Assignment, CA



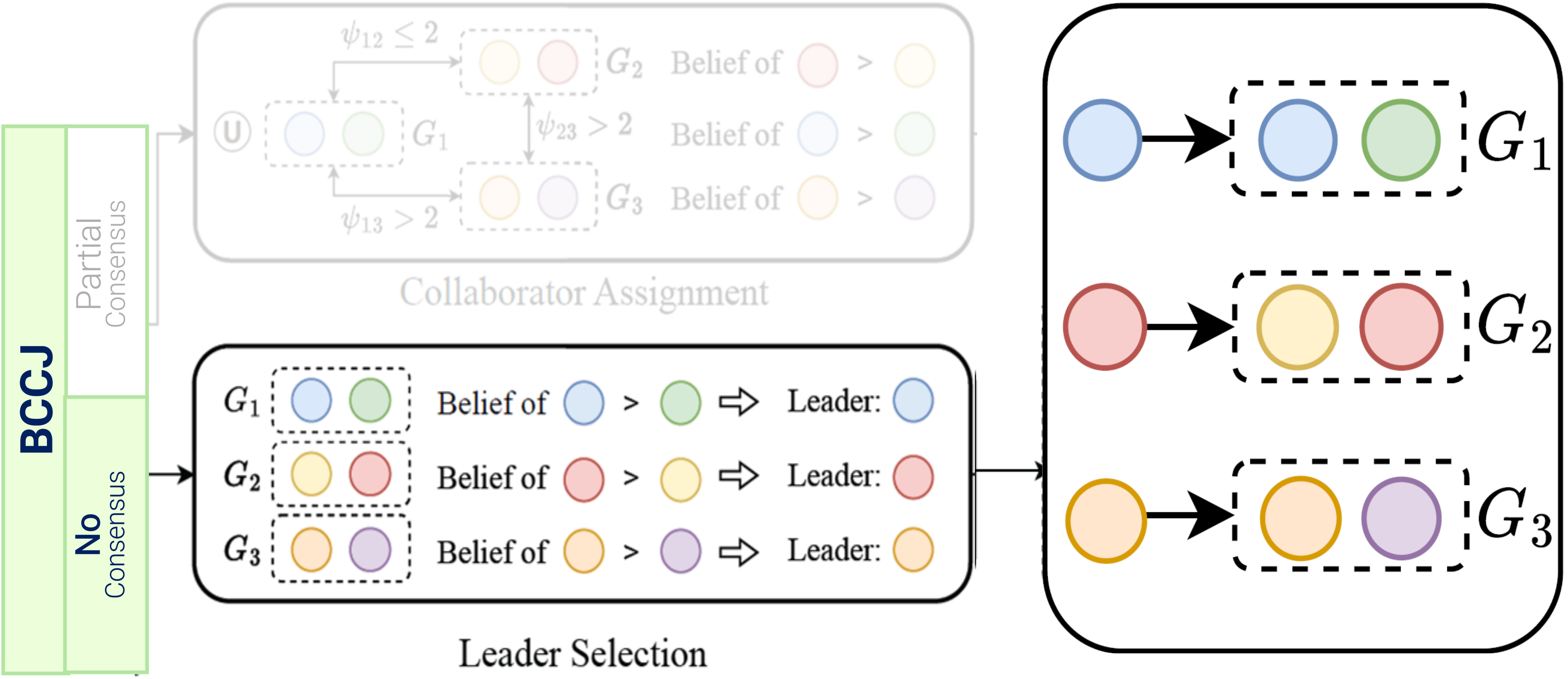
- **Define Most Uncertain** Opinion Group G_u

- **Group relation** defined by Conflict Score
  - Conflict : $\psi_{pq} > 2$
  - Supportive : $\psi_{pq} \leq 2$

- **Collaborator Assignment**
  - i) Lowest belief Agent in G_U
    : Highest belief Agents from "Conflicting" group
  - ii) Others
    : Highest belief Agents in "Supportive" group

04 Methodology

3 – ii) Leader Selection, LS



- No consensus state
  = No mutually supportive opinion groups
- Selecting the n^l agent with the highest belief value from each opinion group, denoted as A^l to serve as a leader
- remaining agents update their opinions by interacting exclusively with these leaders.

05 Experiments

· **RQ1**: How does the performance of BCCS compare to existing single-agent and multi-agent methods in NLP tasks?
· **RQ2**: How does each functional component of BCCS contributes to the performance?
· **RQ3**: How do the supportive/conflicting agents and leaders impact the consensus in BCCS?

Table 1: Main results on the MATH dataset. Bold numbers indicate the best-performing results among all methods.

	Method	Algebra	Counting & Probability	Geometry	Intermediate Algebra	Number Theory	Prealgebra	Precalculus	#Avg
Single Agent Methods	CoT	91.64±0.56	74.30±4.55	58.98±5.46	52.61±2.91	71.33±4.34	85.53±1.71	57.59±3.94	73.33±1.07
	Reflection	91.83±1.88	76.98±1.98	61.55±3.85	52.58±2.33	72.57±0.29	87.65±1.26	59.89±5.73	74.67±0.81
	CoT-SC	92.15±1.12	73.91±0.60	61.76±7.00	62.87±0.73	74.93±4.30	85.52±1.70	63.93±5.58	76.67±0.18
Multi Agent Methods	EoT	94.85±1.27	77.87±4.31	63.03±6.43	60.75±1.21	80.74±1.78	89.42±0.91	61.38±6.81	78.40±0.31
	GroupDebate	94.07±1.35	78.37±2.73	67.70±6.51	59.98±1.62	75.33±3.81	89.08±0.94	61.89±5.35	77.93±0.84
	MAD	94.05±0.39	78.37±1.76	66.14±7.16	62.09±1.99	79.57±1.36	90.15±0.81	62.01±3.68	78.87±0.18
	PARSE	94.84±0.83	76.88±1.04	68.31±5.51	61.13±3.00	80.85±0.29	88.76±0.93	59.14±3.76	78.53±0.55
	CMD	95.11±0.92	75.59±2.94	67.81±7.22	61.17±1.75	81.65±2.37	90.16±0.39	61.21±4.25	78.93±0.53
	DyLAN	95.15±0.81	76.29±2.95	67.08±7.90	59.94±2.03	80.74±1.78	90.09±1.71	62.70±5.19	78.80±0.31
	BCCS	**95.41**±0.76	**79.07**±1.12	**68.64**±7.39	**64.28**±1.60	**82.81**±1.74	**90.88**±0.14	**64.93**±5.17	**80.60**±0.23

05 Experiments

- **RQ1**: How does the performance of BCCS compare to existing single-agent and multi-agent methods in NLP tasks?
- **RQ2**: How does each functional component of BCCS contributes to the performance?
- **RQ3**: How do the supportive/conflicting agents and leaders impact the consensus in BCCS?

Table 2: Main results on the MMLU dataset.

	Method	STEM	Social Sciences	Humanities	Other	#Avg
Single Agent Methods	CoT	68.70±1.24	78.19±0.82	71.84±1.25	70.50±2.95	71.87±0.96
	Reflection	70.93±1.94	78.81±1.56	72.99±1.52	70.79±1.84	73.07±1.67
	CoT-SC	72.76±0.73	78.82±1.12	71.84±2.24	69.61±3.00	73.13±1.33
Multi Agent Methods	EoT	75.81±0.54	76.01±1.89	73.56±2.07	71.39±2.95	74.33±1.48
	GroupDebate	77.03±0.81	78.50±1.08	71.26±2.74	71.98±2.81	74.87±1.54
	MAD	78.46±1.66	78.50±1.62	73.85±2.01	72.86±1.80	76.13±1.46
	PARSE	78.05±1.27	79.44±1.43	74.14±1.99	73.74±1.56	76.47±0.48
	CMD	76.63±1.02	78.82±1.12	72.41±2.28	71.98±2.36	75.07±1.44
	DyLAN	78.25±0.89	77.26±2.43	74.21±2.23	69.03±1.84	75.00±1.51
	BCCS	**79.47**±0.81	**80.69**±1.65	**78.16**±3.20	**75.22**±2.66	**78.47**±1.22

06 Results and Discussions

Table 3: Ablation study on the MATH dataset. Bold numbers indicate the best-performing results among all conditions. "-" indicates removing the corresponding module, and "-Conflict" indicates using supportive opinions only. "R.Leader" indicates selecting n^l leaders randomly.

Method	Algebra	Counting & Probability	Geometry	Intermediate Algebra	Number Theory	Prealgebra	Precalculus	#Avg
BCCS	**95.41**±0.76	**79.07**±1.12	**68.64**±7.39	**64.28**±1.60	**82.81**±1.74	**90.88**±0.14	**64.93**±5.17	**80.60**±0.23
-CA	95.10±1.34	76.89±1.99	66.25±7.79	60.02±1.39	79.18±2.42	90.13±0.92	62.27±4.55	78.60±0.12
-Conflict	95.13±0.53	76.39±3.47	63.80±7.12	61.63±1.61	82.68±2.08	89.03±1.74	60.68±5.25	79.00±0.23
-LS	94.05±0.39	76.39±4.22	66.14±7.16	60.47±3.67	79.45±1.58	88.74±0.62	62.49±5.47	78.13±0.48
R.Leader	95.14±0.11	76.98±2.32	63.53±7.78	59.99±0.43	81.64±1.45	89.44±0.85	60.80±7.13	78.33±0.66
-BCCJ	94.86±0.38	76.09±0.60	64.42±6.24	58.49±2.20	79.18±2.42	89.84±0.51	62.21±3.54	78.00±0.53

Table 4: Ablation study on the MMLU dataset.

Method	STEM	Social Sciences	Humanities	Other	#Avg
BCCS	**79.47**±0.81	**80.69**±1.65	**78.16**±3.20	**75.22**±2.66	**78.47**±1.22
-CA	78.25±1.24	79.13±0.82	75.28±2.87	73.45±1.84	76.67±0.29
-Conflict	78.66±1.61	80.37±0.54	76.43±3.24	72.86±2.57	77.20±0.81
-LS	77.03±1.33	79.44±0.93	77.87±2.35	72.57±3.11	76.73±1.20
R.Leader	79.27±1.06	79.13±0.82	73.85±3.16	73.16±3.40	76.60±1.06
-BCCJ	77.44±1.54	79.75±1.36	77.30±2.74	74.04±3.87	77.13±1.01

6.2 Ablation Study

-BCCJ: belief-calibrated consensus judgment 제거, Byzantine consensus로 대체

-CA: collaborator assignment 제거

-LS: leader selection 제거

-Conflict: conflicting agents를 사용하지 않고 supportive opinions만 사용

R.Leader: highest belief leader가 아니라 random leader 선택

06 Results and Discussions

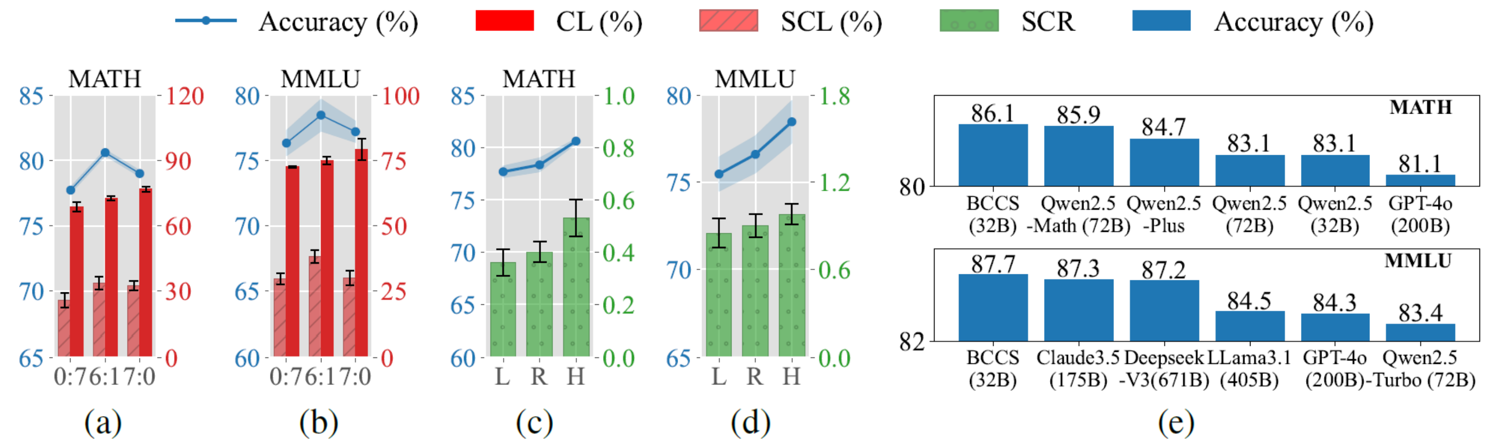


Figure 4: (a) and (b): the results of the CL, SCL and accuracy for each supportive-to-conflicting collaboration ratio (including "0:7", "6:1" and "7:0"). (c) and (d): the results of the SCR and accuracy for the lowest, random and highest leaders' beliefs (denoted as "L", "R", "H"). (e): the analysis of the parameter size, the x-axis denotes the LLMs for comparison.

n_{case}	전체 test case 수
n	agent 수
u	case index
n_u^s	u번째 case에서 consensus에 참여한 agent 수
n_u^r	u번째 case에서 consensus에 도달하기까지 걸린 round 수
x_u	u번째 case에서의 consensus answer
x_u^*	u번째 case의 ground-truth answer

Consensus Level

$$CL = \frac{\sum_{u=1}^{n_{case}} \frac{n_u^s}{n}}{n_{case}}$$

Success Consensus Level

$$SCL = \frac{\sum_{u=1}^{n_{case}} \frac{n_u^s \cdot \mathbf{1}[x_u = x_u^*]}{n}}{n_{case}}$$

Success Consensus Rate

$$SCR = \frac{\sum_{u=1}^{n_{case}} \frac{n_u^s \cdot \mathbf{1}[x_u = x_u^*]}{n_u^r}}{n_{case}}$$

07 Personal Opinion

✓ **TF-IDF/K-Means : grouping이 실제 reasoning direction을 잘 나눠야 함**
→ TF-IDF는 Keyword based 방법이기에 Sementic을 상대적으로 잘 보존하기 어려울 수 있음.
→ K를 3으로 고정함으로써 실제 Reasoning direction을 잘 나누지 못할 가능성이 있음.

✓ **Belief가 실제 답의 신뢰도를 어느 정도 반영해야 함**
→ Belief를 final answer의 generation probability로 근사하여 실제 정답 확률이나 calibrated confidence를 의미한다고 보장할 수 없다.

✓ **high-belief leader가 항상 좋은 reasoning을 가진다는 가정이 어느 정도 맞아야 함**

APPENDIX

Table 15: The prompts of BCCS

MATH	Prompts
System	Please reason step by step, and put your final answer within \boxed{}.
CA	These are the solutions to the problem from other agents: One supporting agent solution:{}, One conflicting agent solution:{} Selecting and using the trustable solutions from current collaboration as additional information, can you provide your answer to the problem? {Question}
LS	These are the solutions to the problem from other agents: One leader solution:{} Selecting and using the leading solutions from current collaboration as additional information, can you provide your answer to the problem? {Question}

MMLU	Prompts
System	Please reason step by step, and answer the question.
CA	Here is the question: {Question} These are the solutions to the problem from other agents: One supporting agent solution {}, One conflicting agent solution {} Judging which solutions are trustable and using the solutions from other agents as additional advice, can you give an updated answer? Examine your solution and that other agents step by step. Put your answer in the form (answer) at the end of your response. (answer) represents choice (A), (B), (C), or (D).
LS	Here is the question: {Question} These are the solutions to the problem from other agents: One leader solution {} Judging which solutions can lead the trend of thought and using the solutions from other agents as additional advice, can you give an updated answer? Examine your solution and that other agents step by step. Put your answer in the form (answer) at the end of your response. (answer) represents choice (A), (B), (C), or (D).

Table 9: Case study of collaborator assignment (CA), (✓) indicates correct answer and (✗) indicates incorrect answer.

Question: The encomienda system was used during the colonization of the Americas to regulate the indigenous people, was not ended by which of the following?: A) The protests of the Catholic missionaries against abuses of forced labor, B) The lack of new land to assign to well-connected Spaniards and conquistadores, C) The Spanish royal crown's desire to control the estates more directly, D) An increase in the number of mestizos, who by law were forbidden from working on the encomiendas.

Method	Round	Opinions	Results
BCCS	1	**Agent 1**: No evidence of an increase in the number of mestizos or any specific laws forbidding them from working on encomiendas. The answer is D.	D (✗)
		Agent 2: There was still land available for assignment, as the encomienda system required land grants to manage indigenous populations. The answer is B.	B (✓)
	2	**Agent 1**: (Receive conflicting opinion from **Agent 2** in round 1) There was still land available for assignment, as the encomienda system required land grants to manage indigenous populations. The answer is B.	B (✓)
		Agent 2: (Receive own supporting opinion from **Agent 2** in round 1) There was still land available for assignment, as the encomienda system required land grants to manage indigenous populations. The answer is B.	B (✓)
PARSE	1	**Agent 1**: It is more about the demographic changes and social dy rather than a direct cause for the termination of the encomienda The answer is D.	
		Agent 2: There was still land available to assign, at least in th term. Therefore, the lack of new land would not have been a re end the system. The answer is B.	
	2	**Agent 1**: (Receive opinions from **Agent 1** and **Agent 2** in round most accurate answer is that the increase in the number of m who by law were forbidden from working on the encomienda answer is D.	
		Agent 2: (Receive opinions from **Agent 1** and **Agent 2** in ro The least likely reason is the increase in the number of mestizo by law were forbidden from working on the encomiendas. The is D.	

Table 10: Case study of leader selection (LS), (✓) indicates correct answer and (✗) indicates incorrect answer.

Question: A scientist, using electrodes, is stimulating a group of neurons in the hypothalamus and recording their membrane potential changes. She observes a sharp rise in membrane potential when she first stimulates them, the the difference of 100mV. When she tries another stimulation immediately after the first, there is no response. However, when she increase the amount of voltage she supplies to the second shock, she observes a similar response to the first. What time period of stimulation would the second shock be characterized in regards to membrane potential?: A) Depolarization, B) Repolarization, C) Hyperpolarization, D) Resting potential.

Method	Round	Opinions	Results
BCCS	1	**Agent 1**: Given that the neuron responded to a stronger stimulus but not to a normal one, it suggests that the membrane potential was still moving towards the resting state but had not fully reached it. The answer is B.	B (✗)
		Agent 2: Given that the neuron showed a response only with increased voltage, indicating it was not at the resting potential, but rather in a state where it required more stimulus to reach the threshold for an action potential, the most appropriate term is hyperpolarization. The answer is C.	C (✓)
	2	**Agent 1**: (Receive the opinion from the leader **Agent 2** in the round 1) Given the context and the need for the neuron to require more stimulus to respond, the most accurate characterization of the second shock is hyperpolarization. The answer is C.	C (✓)
		Agent 2: (**Agent 2** is selected as the leader) Given the observations and the need for increased voltage to achieve a response, the correct characterization of the membrane potential during the second shock is hyperpolarization. The answer is C.	C (✓)
MAD	1	**Agent 1**: The most likely explanation is that the neurons underwent repolarization and then hyperpolarization after the initial depolarization. The answer is B.	B (✗)
		Agent 2: The second shock would be characterized by the neuron being in a state of hyperpolarization, which is a state where the membrane potential is more negative than the resting potential. The answer is C.	C (✓)
	2	**Agent 1**: (Receive opinions from **Agent 1** and **Agent 2** in round 1) It correctly identifies that the second shock occurs during the refractory period, specifically during the phase of repolarization but still within the refractory period where the cell is less responsive. The answer is B.	B (✗)
		Agent 2: (Receive opinions from **Agent 1** and **Agent 2** in round 1) Increasing the voltage for the second stimulation allows a similar response, suggesting the neuron has repolarized and is closer to its resting potential but still in a refractory state. The answer is B.	B (✗)

Table 11: Case study of belief-calibrated consensus judgment (BCCJ), (✓) indicates correct answer and (✗) indicates incorrect answer, and ↑ indicates the corresponding value increases compared to the previous round.

Question: Question: What is the magnetic moment of ^{205}Tl? A) $\gamma(^{205}\mathrm{Tl}) = 1.569 \times 10^8\ \mathrm{T^{-1}s^{-1}}$, $I = \frac{1}{2}$, B) $0.717 \times 10^{-26}\ \mathrm{J\,T^{-1}}$, C) $0.827 \times 10^{-26}\ \mathrm{J\,T^{-1}}$, D) $1.433 \times 10^{-26}\ \mathrm{J\,T^{-1}}$.			
Method	Round	Opinions	Major Results
BCCS	1	**Agent 1**: …Our calculated value is $0.7845 \times 10^8\ \mathrm{J\,T^{-1}}$, which is closest to option B when considering the magnitude. The answer is B. **Agent 2**: …The calculated magnetic moment is $0.827 \times 10^{-26}\ \mathrm{J\,T^{-1}}$, which matches option C. The answer is C.	B (✗) $p_s^1 = 0.57$ $p_b^1 = 0.13$
	2	**Agent 1**: (Receive the opinion from the leader **Agent 2** in the round 1) …The calculated magnetic moment ($0.827 \times 10^{-26}\ \mathrm{J\,T^{-1}}$) matches Option C. The answer is C. **Agent 2**: (**Agent 2** is selected as the leader) …Option C: ($0.827 \times 10^{-26}\ \mathrm{J\,T^{-1}}$) matches our calculated value. The answer is C.	C (✓) $p_s^2 = 0.86 \uparrow$ $p_b^2 = 0.97 \uparrow$
-BCCJ	1	**Agent 1**: …The correct value for the magnetic moment should be around ($0.7845 \times 10^8\ \mathrm{J\,T^{-1}}$), which is closest to option B. The answer is B. **Agent 2**: …The closest option to ($5.197 \times 10^{-26}\ \mathrm{J\,T^{-1}}$) is ($0.717 \times 10^{-26}\ \mathrm{J\,T^{-1}}$). The answer is B.	B (✗) $p_s^1 = 0.86$

Table 12: Error case analysis of belief-calibrated consensus seeking (BCCS).

Question: Which of the following is a true statement about optical isomerism of complexes containing achiral ligands?: A) Square planar complexes can display optical isomerism only if all four ligands are identical, B) Tetrahedral complexes never display optical isomerism, C) Linear complexes can display optical isomerism when both ligands are different, D) Octahedral complexes of monodentate ligands can display optical isomerism only when they have at least three different ligands.

Round	Opinions	Results
1	**Leader 1**: Octahedral complexes can have optical isomerism through "enantiomers" when they contain chiral centers. For an octahedral complex with monodentate ligands, having at least three different ligands can create chiral centers leading to optical isomerism. However, it is not a strict requirement that there must be at least three different ligands for optical isomerism to occur. The answer is D.	D (✓)
	Leader 2: Tetrahedral complexes can display optical isomerism if they contain chiral centers . Answer is B.	B (✗)
	Follower: (D) correctly states that three different ligands can lead to optical isomerism. The answer is D.	D (✓)
2	**Follower**: (Receive opinions from **Leader 1** and **Leader 2** in round 1) Let's re-examine each option ... The (B) provided solution correctly notes that tetrahedral complexes can display optical isomerism if they contain chiral centers ... However, geometric isomerism (cis-trans isomerism) can occur with just two different ligands . Therefore, (D) is not entirely accurate. The answer is B.	B (✗)

13: Error case caused by model performance limitations.

s mounted on a truck that moves forward at a speed of 5 m/s. The operator
rom a cannon so the ball goes as far as possible before hitting the level surface.
the cannon is 50 m/s. At what angle from the horizontal should the operator
°, B) 41°, C) 45°, D) 49°.

	Results
nt: ...the angle that maximizes the range is 45°.	C (✗)
flicting Agent: The optimal angle from the horizontal for the naximum range, considering the truck's speed, is approximately	B (✗)
nt: The provided solution acknowledges that the truck's speed ntal component at lower angles, leading to an optimal angle of	B (✗)

Notation	Description
n_{case}	The number of cases.
n	The total number of agents.
i, j	The subscripts for agent, opinion and belief.
A	A indicates the agent set, which contains n agents.
$a_j \in A$	a_j indicates j-th agent in A.
x_j^k, b_j^k	The opinion and belief of agent a_j in k-th step.
m	The total number of opinion groups.
p, q	The subscripts for opinion group.
G	The opinion group set, which contains m opinions groups.
$G_p \in G$	The k-th opinion group in G.
G_u	The most uncertain opinion group in $G = \{G_p\}_{p=1}^m$.
n^p	The number of agents in G_p.
$a_j \in G_p$	Agent a_j belongs to the opinion group G_p.
u_i^k, v_i^k	The outcome increment of opinion and belief for i-th agent in k-th step.
α, β	Step sizes for analyzing opinion and belief updating.
A^s, A^c	A^s indicates the dominant consensus group, in which contain the largest number of same opinions in A. A^c indicates the conflict group, in which the opinions are different from A^s.
A_i^s, A_i^c	A_i^s is a set of agents which contains the supportive opinions for agent a_i^k. A_i^c is a set of agents which contains the conflicting opinions for agent a_i^k.
ψ_{pq}	The conflict score between opinion group G_p and G_q, which considers two aspects of macro- and micro-conflict scores.
$\psi_{pq}^{\mathcal{G}}$, $\psi_{pq}^{\mathcal{L}}$	$\psi_{pq}^{\mathcal{G}}$ indicates the macro-conflict score and $\psi_{pq}^{\mathcal{L}}$ indicates the micro-conflict score.
p_s^k, p_b^k	p_s^k indicates the proportion of the dominant consensus group in A. p_b^k indicates the proportion of the beliefs of dominant consensus group in all beliefs of A.
u	The subscripts of case index.
n_u^s, n_u^r	n_u^s indicates the number of consensus agents in u-th case. n_u^r indicates the iteration rounds in u-th case.
x_u, x_u^*	x_u and x_u^* indicate the consensus results and the ground-truth of u-th case.
Θ_p, Θ_q	Local consistency scores of opinion groups G_p and G_q.

03 Theoretical Foundations of Stable Consensus

Definition 3.1. (Stable Consensus)

Given a MAS comprising n agents $\{a_i\}_{i=1}^{n}$, and any initial answers $\{x_i^1\}_{i=1}^{n}$,
the system is said to reach a stable consensus answer if the following two conditions are satisfied:

(1) agents' answers converge to consensus
(2) each agent's belief of the answer is coherent with the beliefs of other agents.

Theorem 3.2.

Let $\{x_i^k\}_{i=1}^{n}$ denote the opinions and $\{b_i^k\}_{i=1}^{n}$ denote the beliefs of a MAS with n agents at the k-th step of collaboration.
The collaboration between agents satisfies the following properties:

1. When each agent in MAS collaborates with supportive agents, the MAS tends to reach the stable consensus, converging to the state of the average opinion and belief of all collaborating agents.
2. When any agent in MAS collaborates with conflicting agents, the MAS tends to form the unstable consensus, potentially leading to divergence or oscillation in group states.

Theorem 3.3.

Let $\{x_i^k\}_{i=1}^{n}$ and $\{b_i^k\}_{i=1}^{n}$ represent the opinions and beliefs of a MAS with n agents at the k-th step.
Within each opinion group, the i-th agent follows n^l leaders, and the collaboration between followers and their respective leaders satisfies the following properties:

1. When each agent in an opinion group collaborates with its leaders, the MAS tends to reach the stable consensus, converging to the average state of the leaders.
2. When the leaders' average belief is higher than other agents' beliefs, the leaders with higher beliefs can expedite the convergence to the stable consensus.