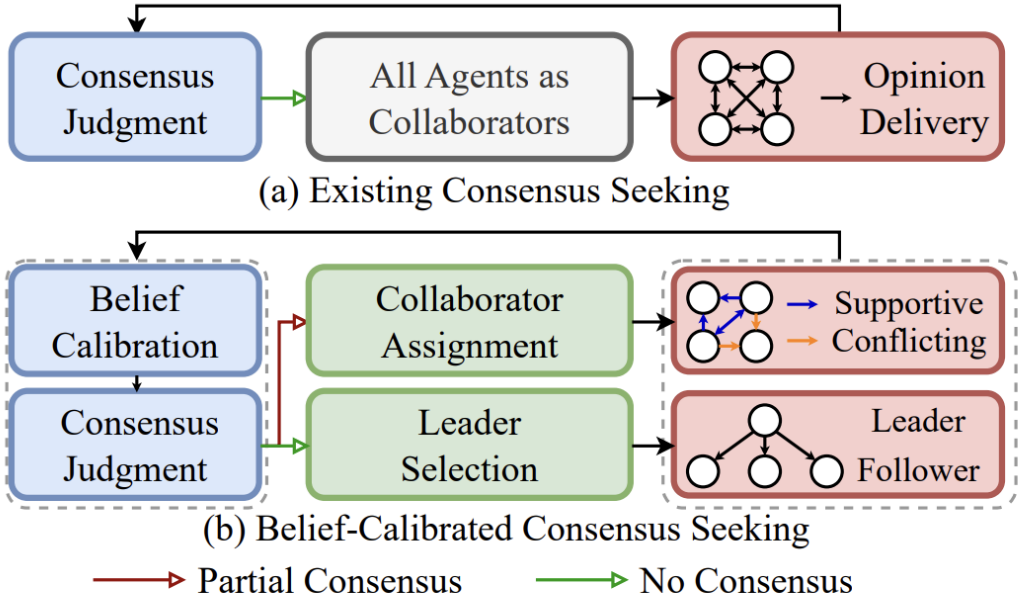
Belief-Calibrated Multi-Agent Consensus Seeking for Complex NLP Tasks

Wentao Deng[1], Jiahuan Pei[2], Zhiwei Xu[1], Zhaochun Ren[3], Zhumin Chen[1*], Pengjie Ren[1*]

[1] Shandong University, [2] Vrije Universiteit Amsterdam, [3] Leiden University
[*] Corresponding authors

Motivation

Existing consensus-seeking approaches typically assess consensus by measuring the degree of agreement among agents, and agents update their views by aggregating the opinions received from others.



(a) Existing Consensus Seeking

(b) Belief-Calibrated Consensus Seeking

Challenges:

- ➢ Current methods often overlook the underlying beliefs of individual agents when determining consensus, which may result in latent internal inconsistencies and compromise the overall stability of the consensus.
- ➢ Agents generally lack mechanisms to selectively identify optimal collaborators, instead indiscriminately incorporating all received opinions.

Contributions

- ➢ We propose the Belief-Calibrated Consensus Seeking (BCCS) method to enhance the consensus-seeking process in multi-agent system (MAS).
- ➢ Theoretical guarantees are established for achieving stable consensus in two key scenarios: (i) cooperation involving both supportive and conflicting agents, and (ii) coordination among leaders with divergent beliefs. These theorems form the theoretical backbone of BCCS.
- ➢ Extensive experiments conducted on widely adopted benchmarks confirm the effectiveness of BCCS.

Theoretical Analysis

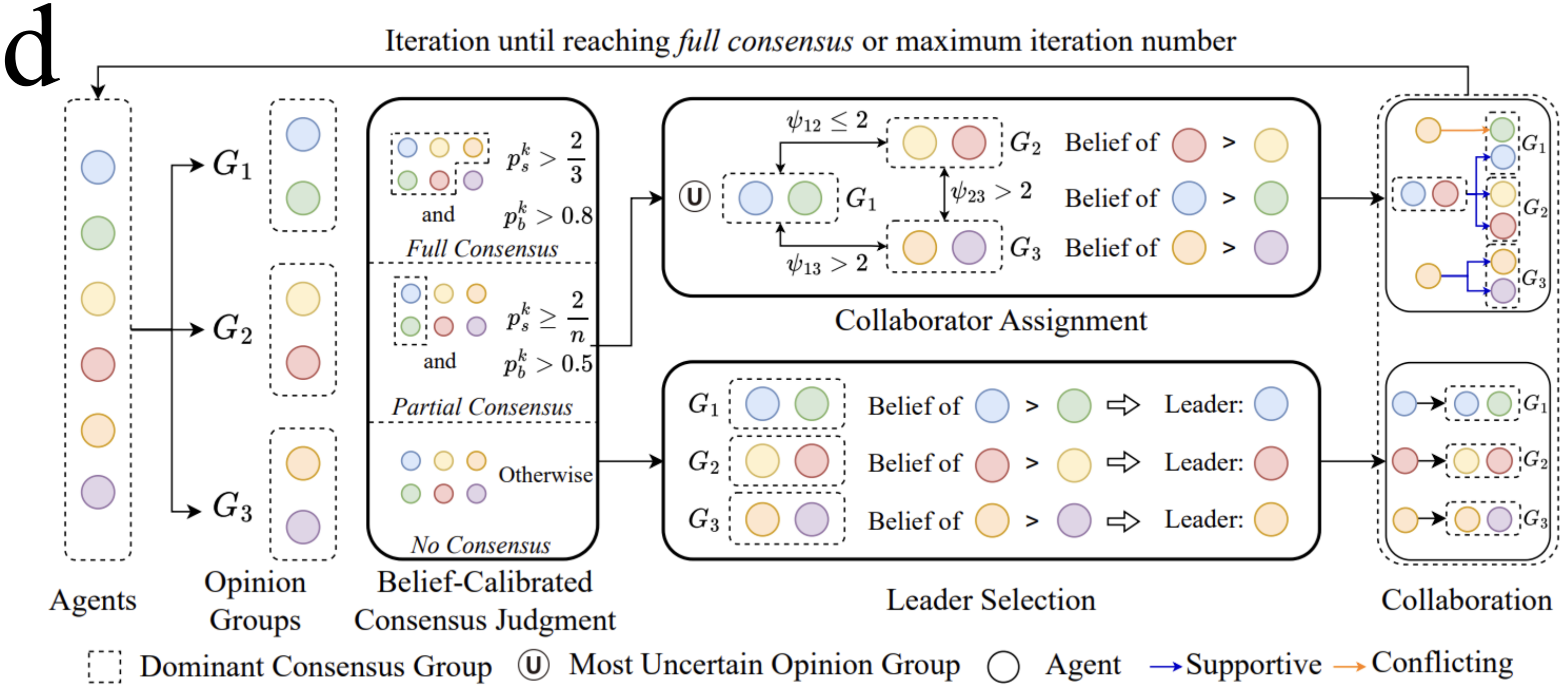
The collaboration between agents satisfies the following properties:

- The MAS tends to reach the stable consensus when each agent in MAS collaborates with supportive agents.
- The MAS tends to form the unstable consensus when any agent in MAS collaborates with conflicting agents.

The collaboration between followers and their respective leaders satisfies the following properties:

- The MAS tends to reach the stable consensus when each agent in an opinion group collaborates with its leaders.
- The leaders with higher beliefs can expedite the convergence to the stable consensus.

Method



- ➢ *Consensus judgement* module not only considers the agents' outputs but also calibrates them based on the associated belief levels. It categorizes the system into one of three consensus states: full consensus, partial consensus, or no consensus.
- ➢ *Collaborator assignment* module automatically assigns optimal collaborators to agents, thereby fostering convergence and avoiding suboptimal solutions.
- ➢ *Leader selection* module identifies and appoints leaders for each opinion group, guiding the direction of discourse and alleviating conflicts.

Main results

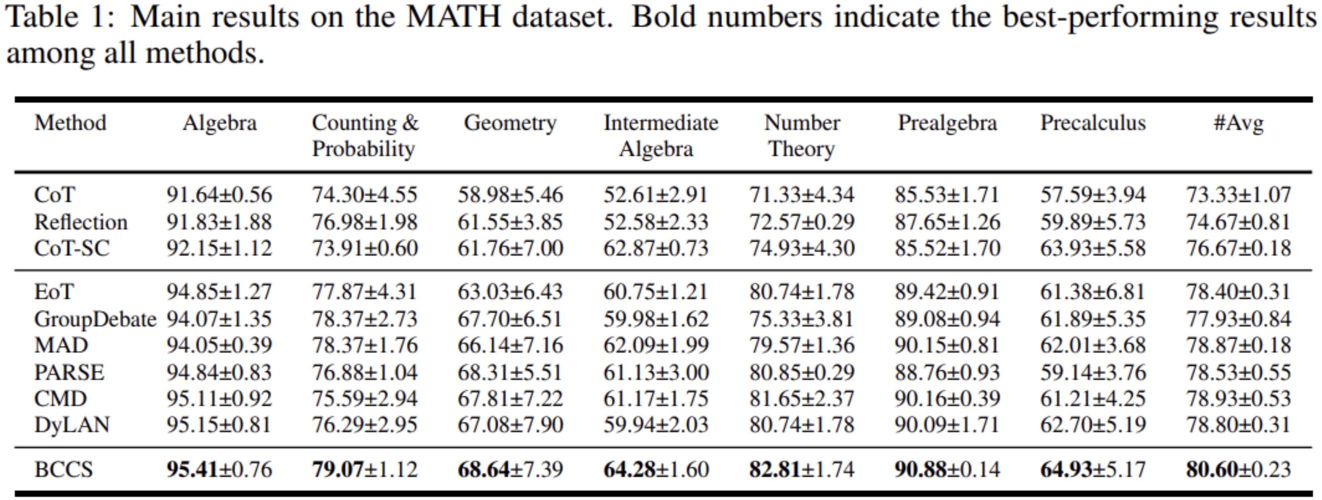


Table 1: Main results on the MATH dataset. Bold numbers indicate the best-performing results among all methods.

Method	Algebra	Counting & Probability	Geometry	Intermediate Algebra	Number Theory	Prealgebra	Precalculus	#Avg
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
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

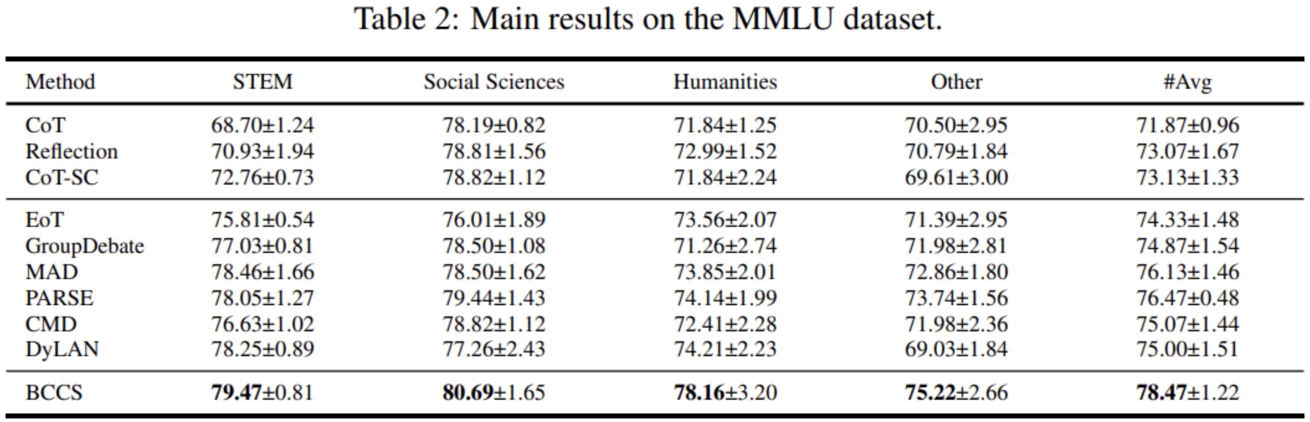


Table 2: Main results on the MMLU dataset.

Method	STEM	Social Sciences	Humanities	Other	#Avg
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
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

- The experiments are conducted on two NLP benchmark datasets, including MATH and MMLU.
- BCCS outperforms the baselines consistently in both datasets.

Conclusion

- In this paper, we provide a theoretical framework for selecting optimal collaborators that maximum consensus stability.
- Based on the theorems, we propose the BCCS framework to facilitate stable consensus via selecting optimal collaborators and calibrating the consensus judgment by system-internal beliefs.

Thanks for Your Attention!