RECONCILE: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs

Justin Chih-Yao Chen Swarnadeep Saha Mohit Bansal

ACL 2024

2024/9/30 名古屋大学笹野研究室 M2 陽田祥平

https://aclanthology.org/2024.acl-long.381

概要

- 推論タスクに対し、複数のLLMエージェントによる議論を重ねて回答を導くフレームワーク**RECONCILE**を提案
- 異なる種類のLLMの利用、他エージェントへの説得、確信度による投票の重み付け等の工夫によりLLMの持つ知識を統合し、回答の精度を向上
- 7つのベンチマークにより、単一エージェントや従来のマルチエージェント手法より高い性能を達成
  - 一部のベンチマークではChatGPTやClaude2等がGPT-4単体の性能を上回ることを実証

背景

- LLMの推論能力の向上のため、人間の思考の過程を模倣した手法が数多く提案されている
  - Self-Reflection…LLMに自身の回答を繰り返し精査させ、回答の品質を向上
  - Multi-agent debate…単一のLLMから複数のエージェントを生成して議論させ、最終的な回答を導く
- 従来の手法は単一のLLMを対象としているが、モデルの持つバイアスや事前学習した知識の範囲に囚われるために性能が芳しくない

→ 異なる知識を持つ外部との議論が必要である可能性

→ 異なるLLMに議論させることで性能が向上するのではないか？

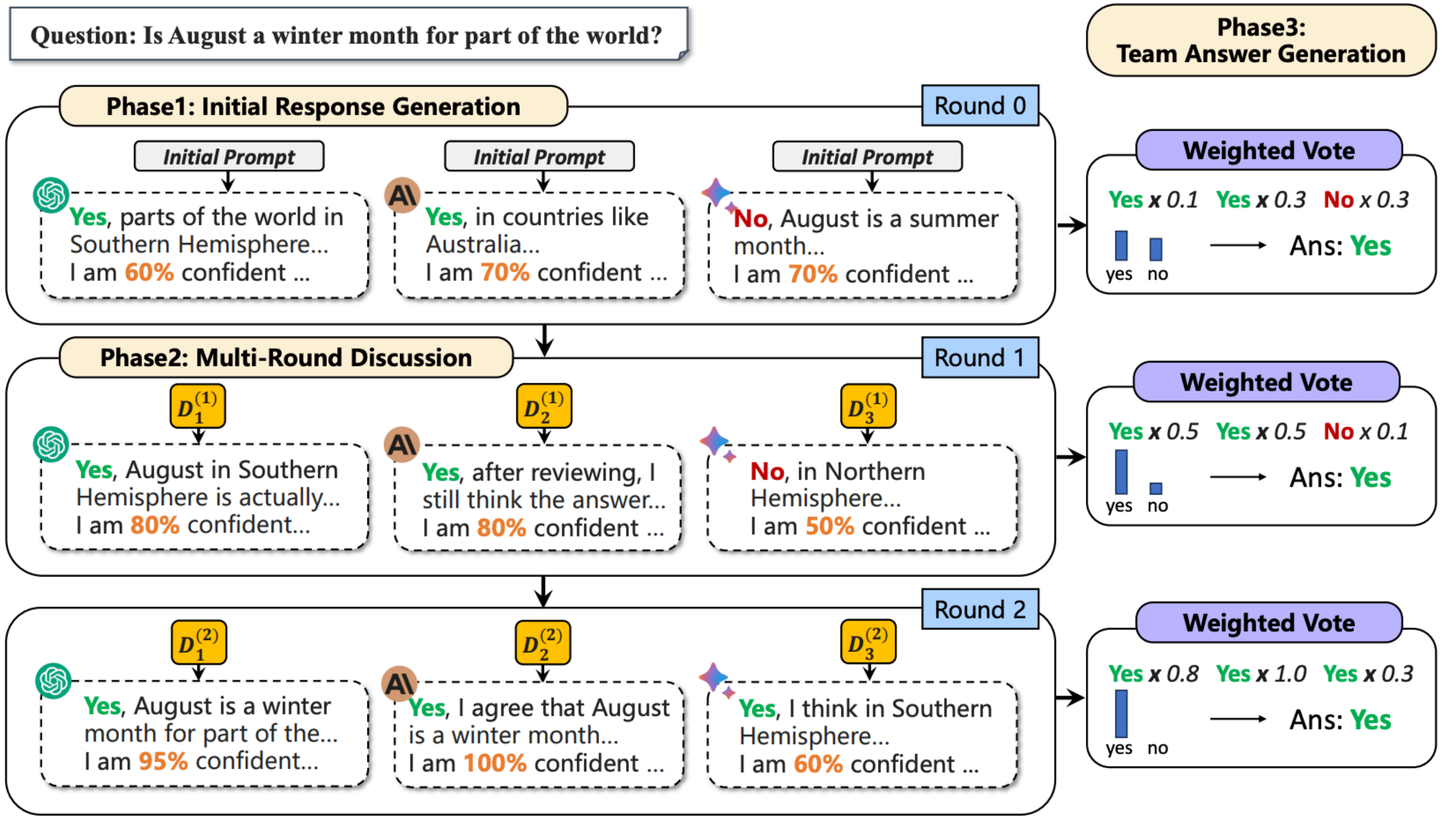
提案手法

- RECONCILEの流れは以下の通り

Phase1: 初期解の生成

Phase2: 複数回の議論

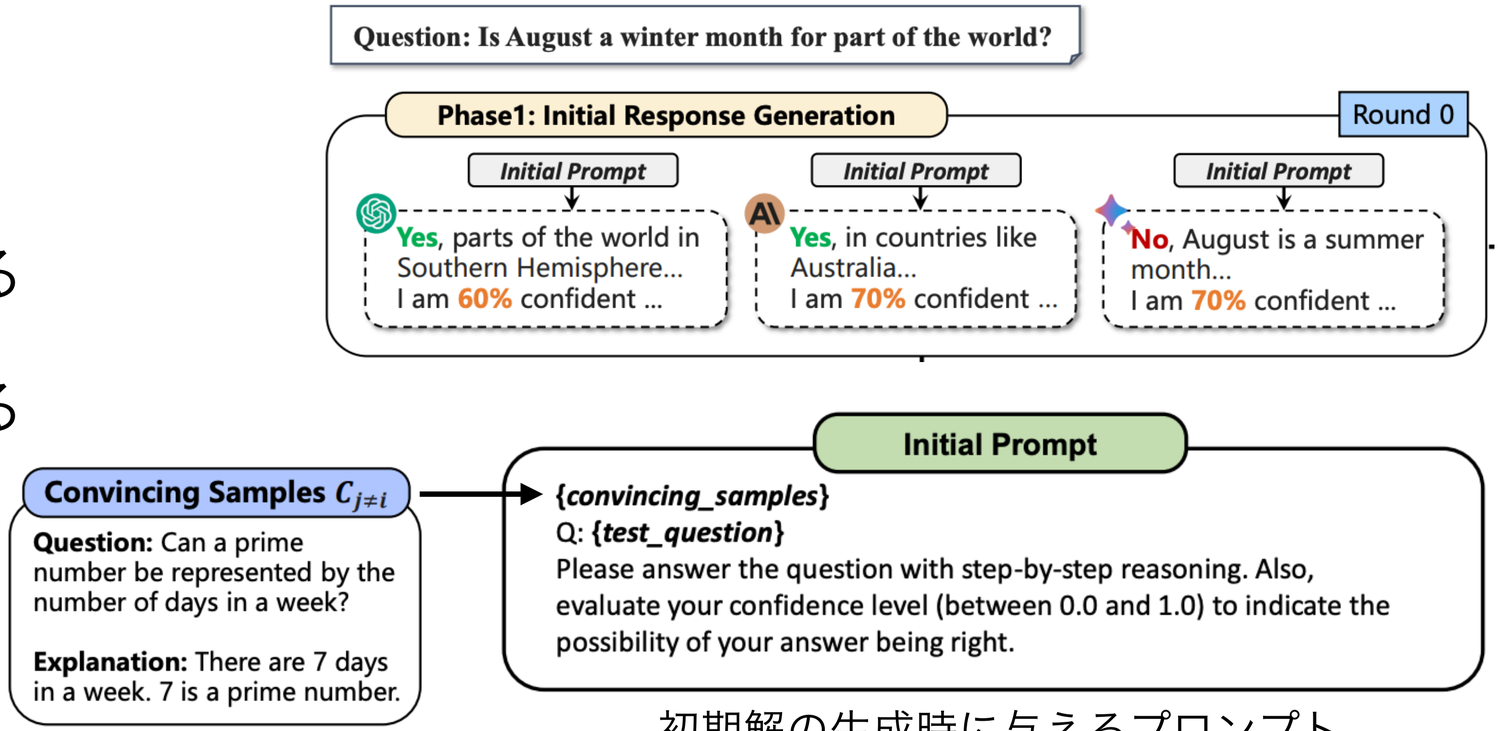
Phase3: 最終的な解の決定



提案手法

Phase1: 初期解の生成

- 各エージェントに問題文を与え、回答と理由、確信度を出力させる
- 例題とその模範回答のペアである convincing samplesをプロンプトの先頭に付与

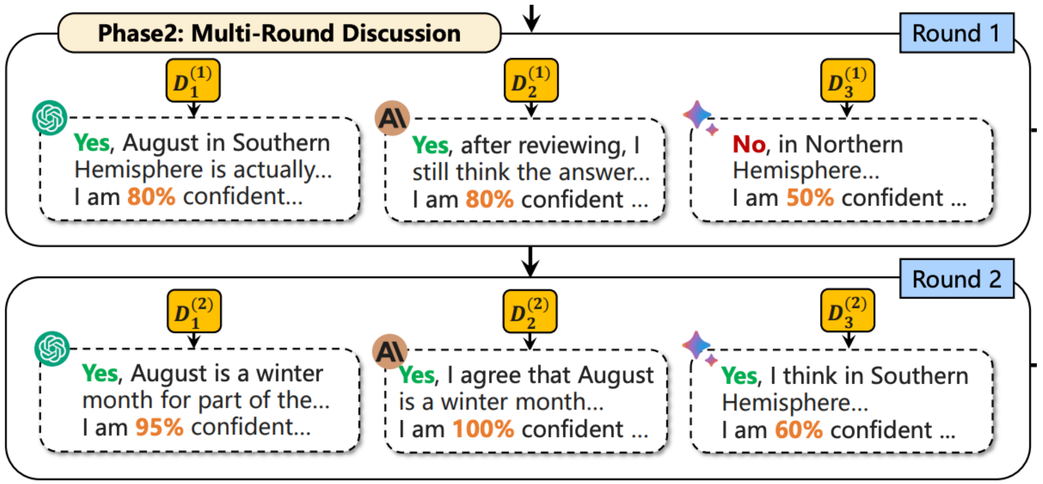


初期解の生成時に与えるプロンプト

提案手法

Phase2: 複数回の議論

- 前回の議論終了後の各エージェントの回答をまとめ、理由、確信度と共に提示する
- 各エージェントは他者の回答を参考に再度回答と理由、確信度を出力する
- 全員の意見が一致するか、規定ラウンド数に達するまで以上を繰り返す



Discussion Prompt

{convincing_samples}
{initial_prompt}
Carefully review the following solutions from other agents as additional information, and provide your own answer and step-by-step reasoning to the question.

Clearly state which point of view you agree or disagree with and why.

There are ***{majority_num}*** agents think the answer is ***{majority_ans}***.
One agent solution: ***{agent_reasoning}*** ***{agent_ans}*** ***{agent_confidence}***
One agent solution: ***{agent_reasoning}*** ***{agent_ans}*** ***{agent_confidence}***

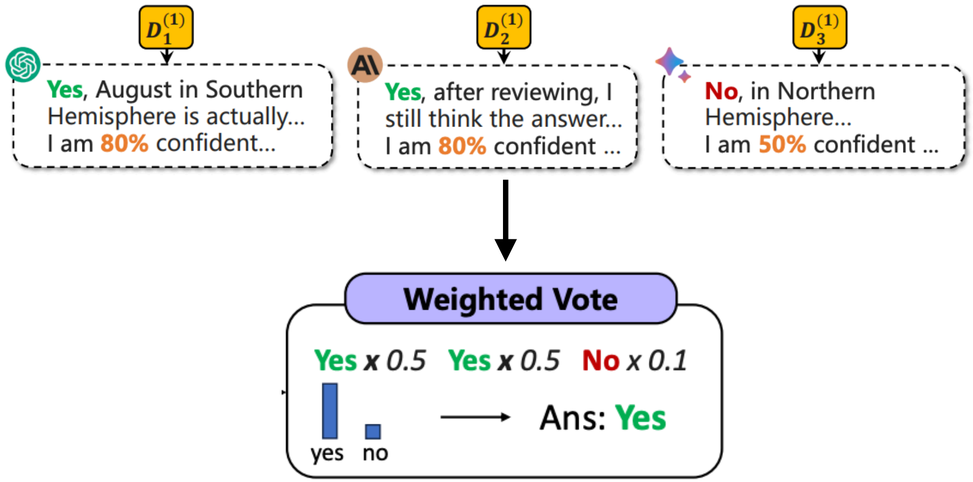
There are ***{minority_num}*** agents think the answer is ***{minority_ans}***.
One agent solution: ***{agent_reasoning}*** ***{agent_ans}*** ***{agent_confidence}***

議論時に与えるプロンプト

提案手法

Phase3: 最終的な解の決定

- 意見が全員で一致しなかった場合、
  確信度を基にした重み付き投票を行う
- 重みの和が最も大きい回答を最終的な解とする
- 基本的に高い確信度が出力されるため、
  確信度の違いを大きく反映する変換関数を適用



$$f(p_i^{(r)}) = \begin{cases} 1.0, & \text{if } p_i^{(r)} = 1.0 \\ 0.8, & \text{if } 0.9 \leq p_i^{(r)} < 1.0 \\ 0.5, & \text{if } 0.8 \leq p_i^{(r)} < 0.9 \\ 0.3, & \text{if } 0.6 < p_i^{(r)} < 0.8 \\ 0.1, & \text{otherwise} \end{cases}$$

確信度$(p_i^{(r)})$を重みに変換する関数

評価実験

- 以下の7つのベンチマークからなる4種類のタスクで評価実験を実施
  - 常識推論…**SQA**、**CSQA**

    (問題例) アリストテレスはコンピュータを使っていたか？
  - 算術…**GSM8k**、**AQuA**、**MATH**

    (問題例) ジョンは3人の子供に1足60$の靴を2足ずつ与えたい時、費用は？
  - 日付計算…**Data Understanding(BIG Bench)**

    (問題例) 昨日が2021/4/30である時、明日の日付は？
  - 自然言語推論…**ANLI**

    前提文と仮説文が与えられ、前提文が仮説文を含意しているかを判定

評価実験

比較手法

- **Vanilla Single Agent**…Chain-of-thoughtを用いてLLMに問題文を入力
- **Self-Refine**…LLMが自身の回答を自分でフィードバック
- **Self-Consistency**…LLMに回答を複数生成させた後、最良のものを選ばせる
- **SR+SC**…Self-Refine、Self-Consistencyを両方適用
- **Debate**…単一のLLMから複数のエージェントを生成し、他者の回答に意見し合うことで最終的な合意を目指す
- **Debate+Judge**…Debateで最終的に得られた各意見を基に別エージェントが最終的な結論を決定

実験結果

Method Category	Method	Agent	StrategyQA	CSQA	GSM8K	AQuA	Date
Vanilla Single-agent	Zero-shot CoT	GPT-4	*75.6*±4.7	*73.3*±0.4	*90.7*±1.7	*65.7*±4.6	*89.0*±2.2
	Zero-shot CoT	ChatGPT	67.3±3.6	66.0±1.8	73.7±3.1	44.7±0.5	67.7±1.2
	Zero-shot CoT	Bard	69.3±4.4	56.8±2.7	58.7±2.6	33.7±1.2	50.2±2.2
	Zero-shot CoT	Claude2	73.7±3.1	66.7±2.1	79.3±3.6	60.3±1.2	78.7±2.1
	Eight-shot CoT	Claude2	74.3±0.8	68.3±1.7	84.7±0.9	64.7±1.2	78.7±1.7
Advanced Single-agent	Self-Refine (SR)	ChatGPT	66.7±2.7	68.1±1.8	74.3±2.5	45.3±2.2	66.3±2.1
	Self-Consistency (SC)	ChatGPT	73.3±0.5	73.0±0.8	82.7±0.5	60.3±1.2	69.3±0.4
	SR + SC	ChatGPT	72.2±1.9	71.9±2.1	81.3±1.7	58.3±3.7	68.7±1.2
Single-model Multi-agent	Debate	×3	66.7±3.1	62.7±1.2	83.0±2.2	65.3±3.1	68.0±1.6
	Debate	×3	65.3±2.5	66.3±2.1	56.3±1.2	29.3±4.2	46.0±2.2
	Debate	×3	71.3±2.2	68.3±1.7	70.7±4.8	62.7±2.6	75.3±3.3
	Debate+Judge	×3	69.7±2.1	63.7±2.5	74.3±2.9	57.3±2.1	67.7±0.5
Multi-model Multi-agent	RECONCILE	, ,	**79.0**±1.6	**74.7**±0.4	**85.3**±2.2	**66.0**±0.8	**86.7**±1.2

- RECONCILEが全てのベンチマークで最高の性能を達成
- 一部タスクでGPT-4単体を上回る
  - GPT-4はGSM8kを訓練コーパスに含んでいるため極端に性能が高い

実験結果

Method Category	Method	Agent	StrategyQA	CSQA	GSM8K	AQuA	Date
	Zero-shot CoT	GPT-4	*75.6*±4.7	*73.3*±0.4	*90.7*±1.7	*65.7*±4.6	*89.0*±2.2
Vanilla Single-agent	Zero-shot CoT	ChatGPT	67.3±3.6	66.0±1.8	73.7±3.1	44.7±0.5	67.7±1.2
	Zero-shot CoT	Bard	69.3±4.4	56.8±2.7	58.7±2.6	33.7±1.2	50.2±2.2
	Zero-shot CoT	Claude2	73.7±3.1	66.7±2.1	79.3±3.6	60.3±1.2	78.7±2.1
	Eight-shot CoT	Claude2	74.3±0.8	68.3±1.7	84.7±0.9	64.7±1.2	78.7±1.7
Advanced Single-agent	Self-Refine (SR)	ChatGPT	66.7±2.7	68.1±1.8	74.3±2.5	45.3±2.2	66.3±2.1
	Self-Consistency (SC)	ChatGPT	73.3±0.5	73.0±0.8	82.7±0.5	60.3±1.2	69.3±0.4
	SR + SC	ChatGPT	72.2±1.9	71.9±2.1	81.3±1.7	58.3±3.7	68.7±1.2
Single-model Multi-agent	Debate	×3	66.7±3.1	62.7±1.2	83.0±2.2	65.3±3.1	68.0±1.6
	Debate	×3	65.3±2.5	66.3±2.1	56.3±1.2	29.3±4.2	46.0±2.2
	Debate	×3	71.3±2.2	68.3±1.7	70.7±4.8	62.7±2.6	75.3±3.3
	Debate+Judge	×3	69.7±2.1	63.7±2.5	74.3±2.9	57.3±2.1	67.7±0.5
Multi-model Multi-agent	RECONCILE	, ,	**79.0**±1.6	**74.7**±0.4	**85.3**±2.2	**66.0**±0.8	**86.7**±1.2

- RECONCILEが全てのベンチマークで最高の性能を達成
- 一部タスクでGPT-4単体を上回る
  - GPT-4はGSM8kを訓練コーパスに含んでいるため極端に性能が高い

分析

モデルの構成を変えた場合の性能の評価

- モデル間に強さの開きがある場合（左表）、特定のタスクに特化したモデルを含む場合（右表）どちらともRECONCILEが最良の結果となる
  - 異なる種類のモデルを参加させることにより議論の幅を広げることが重要であると考えられる

Method	Accuracy	
Best Single-agent (zero-shot)	75.6 ( )	73.7 ( )
Best Multi-agent (Debate)	83.7 (×3)	71.3 (×3)
RECONCILE	**87.7** (, ,)	**78.0** (, ,)

モデルの構成を変えた際のSQAの性能

Method	Accuracy
GPT-4 (zero-shot)	44.0 ( )
Best Single-agent (zero-shot)	50.5 ( )
Best Multi-agent (Debate)	48.7 (×3)
RECONCILE	**58.3** (, ,)

算術タスクに特化したモデルDeepSeekMathを含めた際のMATHの性能

GPT-4
Bard
Claude2
ChatGPT
LLaMa2-70B
DeepSeekMath

分析

各構成要素が性能に与える影響

- 以下の設定でSQAの性能を評価
  - w/o Multiple Model…モデルを全てChatGPTにする
  - w/o Grouping…議論時のプロンプトにおいて各意見を単純に列挙する
  - w/o Convincingness…プロンプトのconvincing samplesを取り除く
  - w/o Conf Estimation…確信度を取り除き、最終的な解の投票を多数決で行う
- いずれも性能が低下、各構成要素の必要性が示される

Method	Accuracy
RECONCILE	**79.0**±1.6
w/o Multiple Models	72.2±2.1
w/o Grouping	76.7±2.5
w/o Convincingness	74.5±1.7
w/o Conf Estimation	77.7±1.3

分析

各構成要素が性能に与える影響

- 以下の設定でSQAの性能を評価
  - w/o Multiple Model…モデルを全てChatGPTにする
  - w/o Grouping…議論時のプロンプトにおいて各意見を単純に列挙する
  - w/o Convincingness…プロンプトのconvincing samplesを取り除く
  - w/o Conf Estimation…確信度を取り除き、最終的な解の投票を多数決で行う
- いずれも性能が低下、各構成要素の必要性が示される

Method	Accuracy
RECONCILE	**79.0**±1.6
w/o Multiple Models	72.2±2.1
w/o Grouping	76.7±2.5
w/o Convincingness	74.5±1.7
w/o Conf Estimation	77.7±1.3

分析

初期解の幅広さの定量的評価

- エージェントi, j間の初期解の意味的類似度をBERTScoreで測定し、$D(A_i, A_j)$で表現
  - 値が小さいほどより幅広い意見が出ていると言える
- 異なるモデルを用いた際の$D(A_i, A_j)$が同じモデルを用いた際と比較して小さい
  - 豊富な初期解が出ることがより良い性能に貢献している

Metric	Method	Accuracy	D (A1, A2)	D (A1, A3)	D (A2, A3)	D (A1, A2, A3)
BERTScore	RECONCILE (Paraphrased)	72.2	0.9364	0.9376	0.9453	0.9398
	RECONCILE (×3)	72.2	0.9077	0.9181	0.9049	0.9102
	RECONCILE (, ,)	**79.0**	**0.8891**	**0.8833**	**0.8493**	**0.8739**

分析

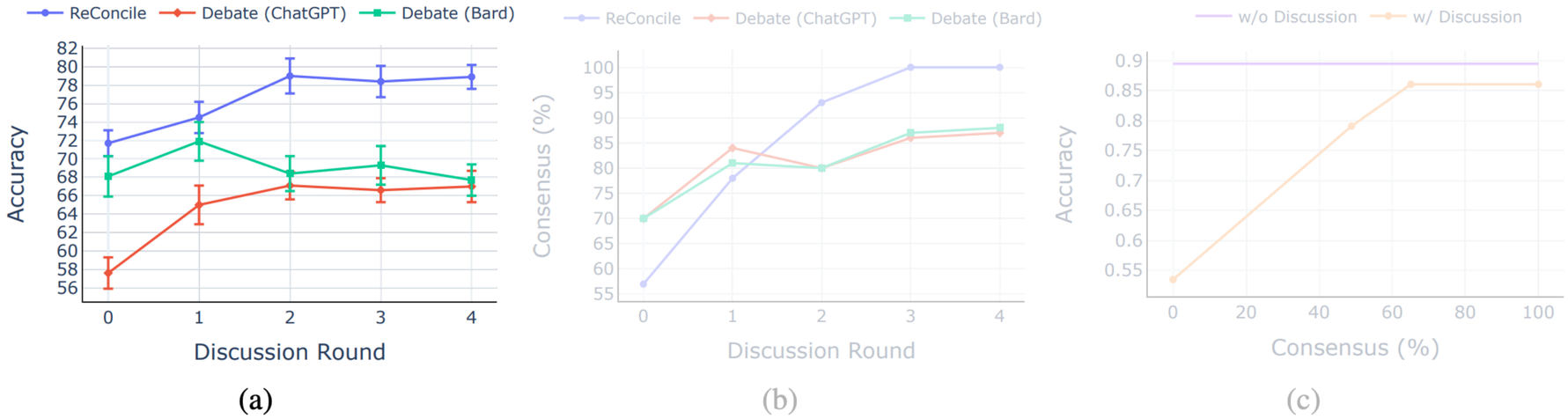
初期解の幅広さの定量的評価

- エージェントi, j間の初期解の意味的類似度をBERTScoreで測定し、$D(A_i, A_j)$で表現
  - 値が小さいほどより幅広い意見が出ていると言える
- 異なるモデルを用いた際の$D(A_i, A_j)$が同じモデルを用いた際と比較して小さい
  - 豊富な初期解が出ることがより良い性能に貢献している

Metric	Method	Accuracy	D (A1, A2)	D (A1, A3)	D (A2, A3)	D (A1, A2, A3)
BERTScore	RECONCILE (Paraphrased)	72.2	0.9364	0.9376	0.9453	0.9398
	RECONCILE (×3)	72.2	0.9077	0.9181	0.9049	0.9102
	RECONCILE (, ,)	**79.0**	**0.8891**	**0.8833**	**0.8493**	**0.8739**

分析

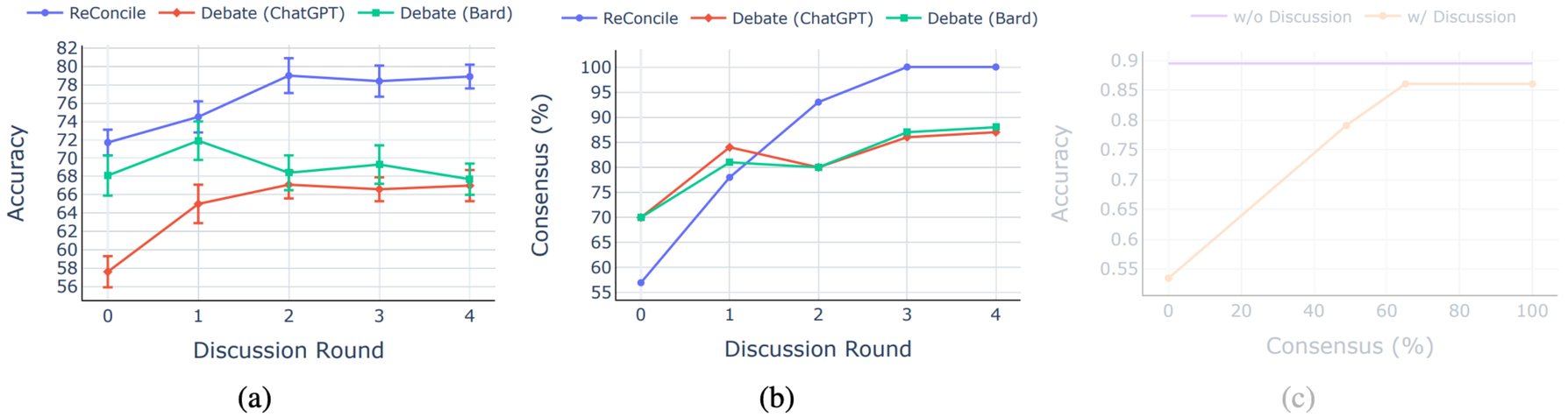
議論回数、合意の取れる割合、正解率の関係



- 他手法と比較して常に性能が高く、かつ議論を経ることで性能が低下しない
- モデルの多様性により議論前の合意の取れる割合は低いが、議論3回時点で全ての例で合意が取れる
- 合意の取れる割合と正解率に相関が見られる

分析

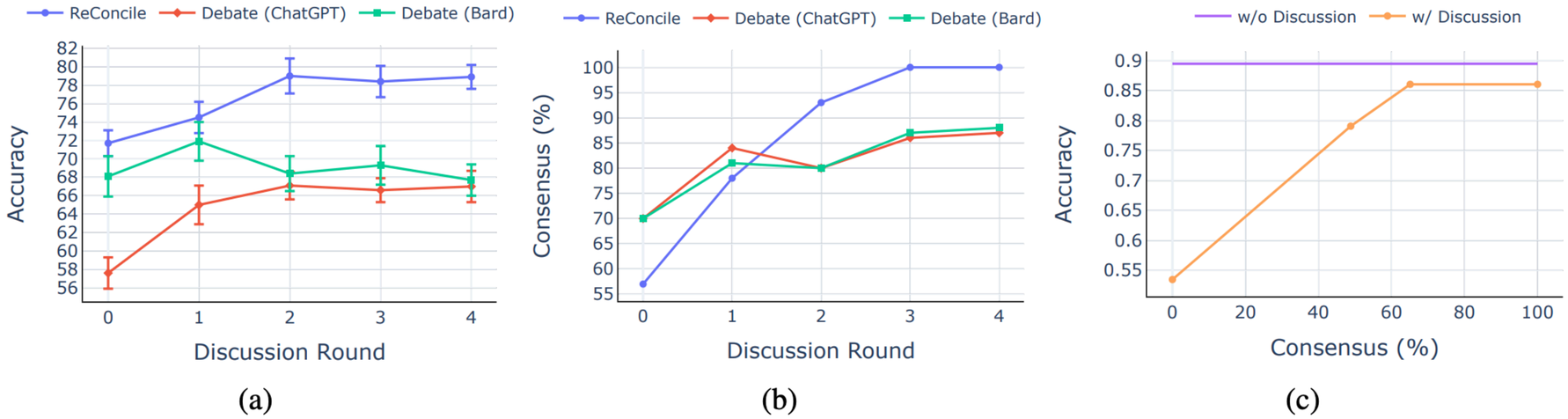
議論回数、合意の取れる割合、正解率の関係



- 他手法と比較して常に性能が高く、かつ議論を経ることで性能が低下しない
- モデルの多様性により議論前の合意の取れる割合は低いが、議論3回時点で全ての例で合意が取れる
- 合意の取れる割合と正解率に相関が見られる

分析

議論回数、合意の取れる割合、正解率の関係



- 他手法と比較して常に性能が高く、かつ議論を経ることで性能が低下しない
- モデルの多様性により議論前の合意の取れる割合は低いが、議論3回時点で全ての例で合意が取れる
- 合意の取れる割合と正解率に相関が見られる

まとめ

- 異なるLLMが議論を重ねて推論を行うフレームワーク**RECONCILE**を提案
- 議論の反復、回答の理由及び確信度を利用することで各エージェントが他者を説得できるような仕組みを構築
- 常識推論、算術などの様々なベンチマークで既存手法より高い性能を達成し、一部タスクではGPT-4に匹敵