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

Justin Chih-Yao Chen **Swarnadeep Saha** **Mohit Bansal**
UNC Chapel Hill
{cychen,swarna,mbansal}@cs.unc.edu

Abstract

Large Language Models (LLMs) still struggle with natural language reasoning tasks. Motivated by the *society of minds* (Minsky, 1988), we propose RECONCILE, a multi-model multi-agent framework designed as a round table conference among diverse LLM agents. RECONCILE enhances collaborative reasoning between LLM agents via multiple rounds of discussion, learning to convince other agents to improve their answers, and employing a confidence-weighted voting mechanism that leads to a better consensus. In each round, RECONCILE initiates discussion between agents via a 'discussion prompt' that consists of (a) grouped answers and explanations generated by each agent in the previous round, (b) their confidence scores, and (c) demonstrations of answer-rectifying human explanations, used for convincing other agents. Experiments on seven benchmarks demonstrate that RECONCILE significantly improves LLMs' reasoning – both individually and as a team – surpassing prior single-agent and multi-agent baselines by up to 11.4% and even outperforming GPT-4 on three datasets. RECONCILE also flexibly incorporates different combinations of agents, including API-based, open-source, and domain-specific models, leading to an 8% improvement on MATH. Finally, we analyze the individual components of RECONCILE, demonstrating that the diversity originating from different models is critical to its superior performance.[1]

1 Introduction

A large body of recent work has focused on improving the reasoning capabilities of Large Language Models (LLMs) by imitating various human cognitive processes (Wang and Zhao, 2023; Park et al., 2023; Sumers et al., 2023; Ye et al., 2023). These include phenomena like reflecting on and critiquing one's own predictions, being receptive to feedback, and learning from feedback. Of note, self-reflection is an introspective process that allows the model to improve its outputs by generating feedback from the model itself (Madaan et al., 2023; Shinn et al., 2023). However, self-reflection suffers from Degeneration-of-Thought – when the model is overly confident in its answer, it is unable to generate novel thoughts even after multiple rounds of feedback (Liang et al., 2023).

To promote more diverse thoughts, past work has drawn inspiration from the concept of *society of minds* in multi-agent systems (Minsky, 1988; Zhuge et al., 2023). It highlights the importance of communication and collaboration between multiple agents for complex decision-making tasks. While such collaborative frameworks like multi-agent debate (Liang et al., 2023; Du et al., 2023) increase the reasoning diversity through the process of a debate, multiple agents have typically been limited to different instances of the same underlying model like ChatGPT (OpenAI, 2022).[2] This results in an inherent model bias, a restricted knowledge scope, and a lack of external feedback from other models due to identical pre-training data and model architectures across all agents. In general, when multiple agents propose solutions to a problem, the success of such a multi-agent system is fundamentally reliant on (a) the diversity of the solutions, (b) the ability to estimate each agent's confidence, and (c) accordingly, convince other agents (with explanations) to reach a better consensus. This puts forward the question: if multiple diverse LLMs collaboratively solve a task, are they capable of discussing their solutions with each other to reach a better consensus?

We aim to solve reasoning problems by learning from diverse insights and external feedback, originating from agents that belong to different model

[1] Code: https://github.com/dinobby/ReConcile

[2] In this work, we refer to multi-agent as multiple instances of the same underlying model (e.g., ChatGPT), whereas multi-model model-agent refers to different models (e.g., ChatGPT, Bard and Claude2) as agents.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7066–7085
August 11-16, 2024 ©2024 Association for Computational Linguistics

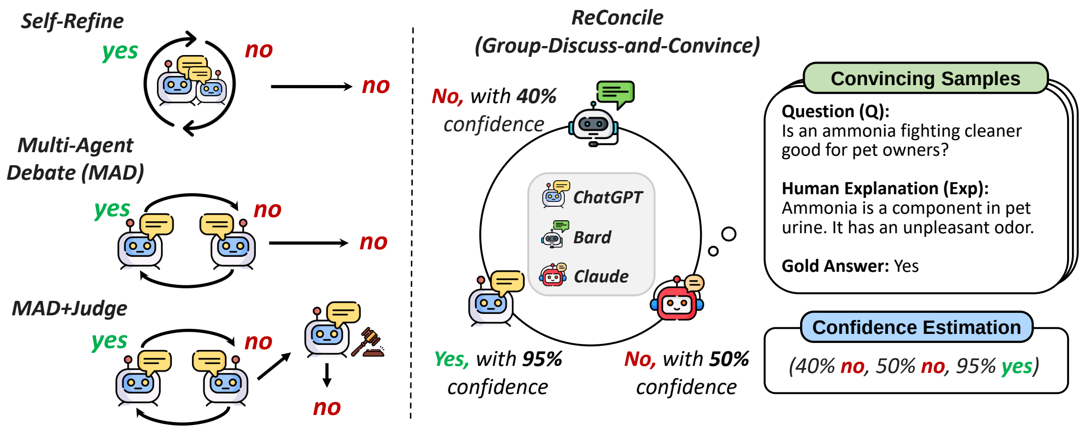


Figure 1: An illustration of the main differences between RECONCILE and prior works. While most current self-refine and debating techniques rely on multiple instances of a single model (e.g., ChatGPT), our method incorporates models from different families (e.g., ChatGPT, Bard, and Claude2). Our approach also emphasizes critical elements of effective discussion, including convincing another agent to improve their answers and incorporating the estimated confidence of all agents. For illustrative simplicity, we depict only one agent contemplating how to convince the other two agents.

families. Collaborative processes such as brainstorming, group meetings, and discussions play a pivotal role in reaching a consensus and arriving at more refined solutions to complex problems (Li et al., 2022b). Effective discussion also entails the selection of stances, voting, convincing, exchange of information, and a diversity of opinions. Thus, we propose RECONCILE, a framework of round-table conference for obtaining better consensus among diverse LLM agents. RECONCILE consists of multiple discussion rounds between diverse LLM agents who try to *convince*[3] each other to either *rectify* their answers or become more *confident* of their initial correct answers (see Fig. 1 for a broad overview).

Given a reasoning problem, RECONCILE begins with each agent first generating an answer, its uncertainty, and a corresponding explanation (as a Chain-of-Thought (Wei et al., 2022)) for the answer. Then all agents enter a multi-round discussion phase. Each discussion round consists of all agents generating a revised explanation and answer based on all other agents' explanations and answers from the previous round. In particular, RECONCILE initiates a discussion by designing a *discussion prompt* for each agent, that lets it condition on (1) grouped answers from all agents, (2) corresponding explanations generated in the previous round, and (3) demonstrations of answer-rectifying human explanations for convincing other agents. We leverage them in an in-context learning framework to teach models to generate their own convincing explanations (see Fig. 3). Even in cases where an agent initially offers an incorrect answer and explanation, it can consider another agent's convincing explanation and amend its response accordingly. In each discussion round, we estimate an agent's uncertainty via a confidence-estimation prompt (Tian et al., 2023; Xiong et al., 2023a). Once all agents converge to the same answer (i.e., a consensus has been reached), we employ these confidences to compute a weighted vote as the team answer.

We primarily develop RECONCILE with three state-of-the-art LLMs: ChatGPT (OpenAI, 2022), Bard (Anil et al., 2023), and Claude2 (Anthropic, 2023). We also demonstrate the flexibility of RECONCILE with variants that employ a much stronger GPT-4 (OpenAI, 2023), an open-source LLaMA-2-70B (Touvron et al., 2023), or a domain-specific DeepSeekMATH (Shao et al., 2024) model as an agent. Across seven benchmarks spanning commonsense reasoning, mathematical reasoning, logical reasoning, and Natural Language Inference (NLI), RECONCILE outperforms prior single-agent (e.g., Self-Refine (Madaan et al., 2023) and Self-consistency (Wang et al., 2023b)) and multi-agent baselines (Debate (Du et al., 2023) and Judge (Liang et al., 2023)) that are built on top of the same underlying models. For example, RECONCILE, (1) on a date understanding task, outperforms the leading multi-agent debate baseline by

[3]When we say that an agent tries to convince another agent, we mean that it learns (based on corrective explanations) to defend or argue for its stance while still being receptive to the other agent's argument.

	Refine	Ensemble	Multi-Agent	Multi-Model	Convincingness	Confidence
Self-Refine (SR)	■	□	□	□	□	□
Self-Consistency (SC)	□	■	□	□	□	□
SR + SC	■	■	□	□	□	□
Debate	■	■	■	■*	□	□
Judge	■	■	■	□	□	□
RECONCILE (Ours)	■	■	■	■	■	■

Table 1: Summary of the main differences between prior work, including Self-Refine (SR, Madaan et al. (2023)); Self-Consistency (SC, Wang et al. (2023b)); Debate (Du et al., 2023) and Judge (Liang et al., 2023). ■ means supported and □ means not supported. RECONCILE supports multi-model multi-agent discussion with confidence estimation and convincingness. * = Du et al. (2023) primarily experiment with multiple instances of ChatGPT as different agents and conduct an initial investigation with 20 samples using ChatGPT and Bard as the two agents.

11.4%, (2) on StrategyQA, also outperforms GPT-4 by 3.4%, and (3) on MATH, outperforms both GPT-4 and a specialized DeepSeekMath model by 8%. Moreover, detailed analyses of the individual components of RECONCILE demonstrate that leveraging diverse LLM agents leads to maximum gains, and we further validate their higher response diversity via a BERTScore-based diversity metric (Zhang et al., 2019). Finally, we show that RECONCILE not only leads to better team performance but also enables each agent to improve individually via the discussion process.

In summary, our primary contributions are:

- We propose RECONCILE, a reasoning framework involving diverse Large Language Models in a Round Table Conference.
- We conduct extensive experiments on seven benchmarks to show that RECONCILE outperforms strong baselines (including GPT-4 on some benchmarks) and also generalizes to different combinations of agents.
- We study the role of diversity, confidence estimation, and an agent's ability to convince others (by learning from corrective explanations) in multi-agent discussion systems.

2 Related Work

Reasoning with LLMs. Progress in LLMs has led to the development of advanced prompting and fine-tuning techniques for solving reasoning problems. Representative methods include Chain-of-Thought (CoT) (Kojima et al., 2022; Wei et al., 2022; Wang et al., 2023a) and Tree-of-Thought prompting (Yao et al., 2023a), self-consistency (Wang et al., 2023b), meta-reasoning over multiple paths (Yoran et al., 2023), use of scratchpads (Nye et al., 2021), training verifiers (Cobbe et al., 2021), self-collaboration (Wang et al., 2023c; Schick et al., 2022; Li et al., 2023a; Feng et al., 2024), self-reflection (Shinn et al., 2023; Madaan et al., 2023; Wang and Zhao, 2023; Yao et al., 2023b), improved math reasoning (Yue et al., 2023; Luo et al., 2023) and fine-tuning via bootstrapping models (Zelikman et al., 2022; Lewkowycz et al., 2022; Li et al., 2023b). Eliciting reasoning from a single agent, while promising, is fundamentally limited by a lack of diverse insights.

Reasoning in Multi-Agent Systems. A recent line of work has explored student-teacher frameworks with the goal of distilling reasoning capabilities from a stronger teacher to a weaker student (Magister et al., 2023; Fu et al., 2023; Ho et al., 2023; Saha et al., 2023; Mukherjee et al., 2023). As opposed to a teacher teaching weaker agents, we seek to develop a multi-agent system where different LLM agents have their unique strengths and try to collaboratively improve performance by reaching a better consensus. Notable prior works include multi-agent debating frameworks (Du et al., 2023; Liang et al., 2023; Chan et al., 2023; Xiong et al., 2023a; Khan et al., 2024) but such efforts are still largely limited to multiple instances of the same underlying language model. We argue that relying on a single model limits the potential of complementary benefits from different model families and the advantage of ensemble learning. Moreover, estimating the confidence of each agent and being able to defend or improve one's opinions become more prominent components in such multi-model multi-agent systems because of the individual differences. Overall, Table 1 summarizes RECONCILE's key differences compared to prior single-agent and multi-agent reasoning methods.

Ensembling Large Pretrained Models. Large pre-trained models, by virtue of being trained on

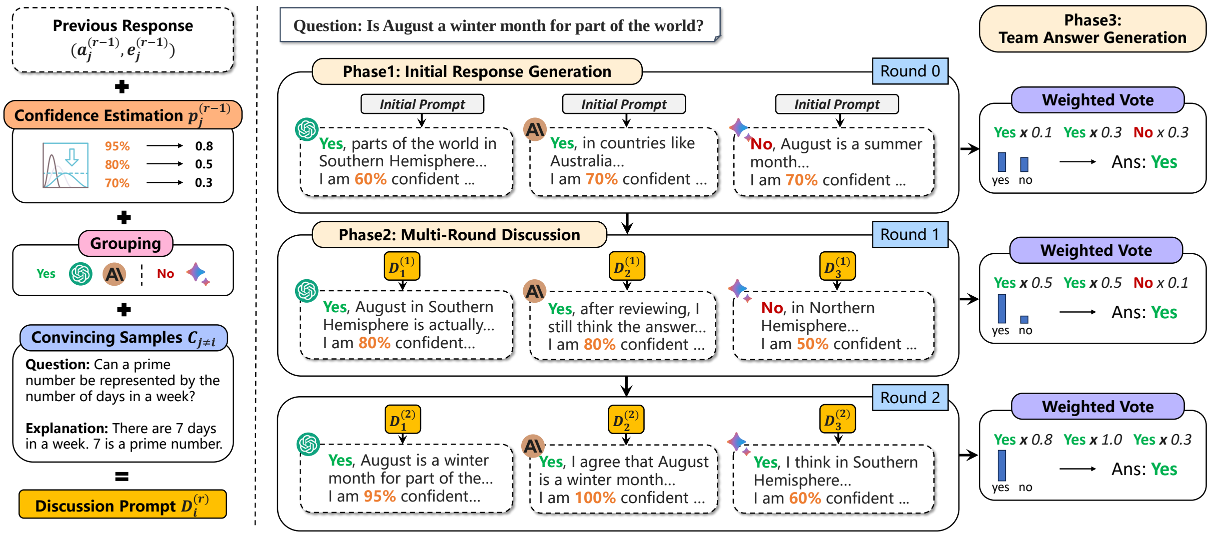


Figure 2: Overview of RECONCILE with ChatGPT, Bard, and Claude2, consisting of three phases: (1) Initial Response Generation: Each agent generates an initial answer and explanation. (2) Multi-Round Discussion: Each model is presented with a discussion prompt (as illustrated on the left) and subsequently generates an updated answer and explanation. (3) Team answer generation: The team answer is determined by a weighted vote at the end of each round. The left part of the figure shows the discussion prompt for an agent, consisting of (a) grouped answers and explanations of all agents from the previous round, (b) estimated confidence, and (c) demonstrations of convincing samples.

different data and with architectural variations, exhibit distinct capabilities. This has led to the development of ensembles (Sagi and Rokach, 2018) in multimodal learning (Zeng et al., 2023; Li et al., 2022a). Mixture of Experts, a popular ensemble learning technique, trains multiple smaller specialized models to improve robustness and overall accuracy (Jacobs et al., 1991; Shazeer et al., 2017; Du et al., 2022). Specific to language models, Self-Consistency (Wang et al., 2023b) generates diverse reasoning paths using CoT and chooses the most consistent answer as the final output. Jiang et al. (2023) propose LLM-Blender, a method to rank and fuse generations from different models. Different from these, we study communication via explanations between distinct LLM agents and their ability to discuss and convince each other in order to improve collective reasoning.

3 Problem Setup

We assume that we are given a test problem Q and there are n agents $\mathcal{A} = \{A_i\}_{i=1}^n$ participating in a round table discussion. Each agent is a distinct LLM, potentially trained with different pre-training data and model architectures. All agents are capable of generating an answer and a corresponding Chain-of-Thought explanation (Wei et al., 2022) for the test problem. For each agent A_i, we utilize a small number of k demonstrations of convincing samples $C_i = \{c_j^{(i)}\}_{j=1}^k$. Each convincing sample $c_j^{(i)} = (q_j^{(i)}, a_j^{(i)}, e_j^{(i)})$ for an agent A_i is an instance of a question $q_j^{(i)}$, gold answer $a_j^{(i)}$, and a human explanation $e_j^{(i)}$ that helps rectify an agent's initial incorrect answer (see more details in Sec 4). The objective of RECONCILE is to improve the team performance on a given task by holding multiple rounds of discussion between the agents, quantifying the uncertainty associated with each agent, and convincing other agents to reach a better consensus. Note that convincing samples serve as an additional performance enhancer; even when the dataset lacks human explanations, our method can still yield performance gains independent of this (more details below).

4 RECONCILE: A Collaborative Discussion Framework

RECONCILE operates in three phases: initial response generation, multi-round discussion, and team answer generation. The overview of our method is demonstrated in Fig. 2 and Algorithm 1.

Phase 1: Initial Response Generation. RECONCILE operates with each agent A_i initially generating an answer $a_i^{(0)}$, an explanation $e_i^{(0)}$, and an associated confidence $p_i^{(0)} \in [0, 1]$ for the generated answer. Each agent conditions on a zero-shot

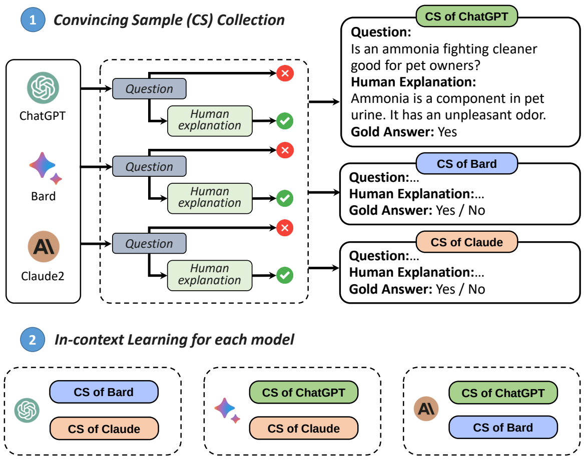


Figure 3: Method for choosing convincing samples for each agent. A convincing sample for ChatGPT consists of a question, a gold answer, and a 'corrective' human explanation that can rectify its initial incorrect answer. Then Bard and Claude2 use it in-context during discussion to convince ChatGPT.

prompt that instructs it to reason about the problem 'step-by-step'. See 'Phase 1' in Fig. 2 and the prompt is shown in Fig. 5 in Appendix A.2.

Phase 2: Multi-round Discussion. RECONCILE then enters a discussion phase, consisting of R rounds (see 'Phase 2' in Fig. 2). In discussion round r, for each agent A_i, RECONCILE develops a discussion prompt $\mathcal{D}_i^{(r)}$ (as shown in Fig. 5), consisting of the following three components.

(a) Grouped responses of all agents from the previous round. $\mathcal{D}_i^{(r)}$ consists of the answers $\{a_j^{(r-1)}\}_{j=1}^n$ and explanations $\{e_j^{(r-1)}\}_{j=1}^n$ of all agents from round $(r-1)$. To foster better discussions, RECONCILE summarizes this information by grouping the answers into distinct categories and appends all plausible explanations for each answer, as shown in our discussion prompt (Appendix Fig. 5) and on the left side of Fig. 2.

(b) Confidence associated with the answers. All agents are not equally confident in their answers. Hence, an effective discussion should also consider each agent's uncertainty. For all black-box models, we estimate its confidence $p_i^{(r)}$ in round r by directly prompting the agent to verbally quantify its uncertainty, which in past work has been shown to be effective (Xiong et al., 2023b). See Appendix Fig. 5 for the usage of confidence in discussion.

(c) Convincing samples from all other agents. Finally, the prompt contains convincing samples C_j for all other agents $A_{j\neq i}$.[4] When an agent tries to reassess its reasoning in light of the reasoning provided by other agents, we hypothesize that it should benefit from conditioning on demonstrations that can convince other agents. In order to obtain such convincing samples for an agent A_j, we select a small number of samples (4 in our experiments) for which the agent's initial answer is wrong but conditioning on the corresponding human explanation, rectifies the answer (see Fig. 3). For datasets that *do not* come with human explanations (e.g., the date understanding task in our experiments), we develop RECONCILE without using any convincing sample in the discussion prompt and still obtain large improvements (see §6.2 for details).

We now define the discussion prompt $\mathcal{D}_i^{(r)} = \{a_j^{(r-1)}, e_j^{(r-1)}, p_j^{(r-1)}, C_{j\neq i}\}_{j=1}^n$ for each agent A_i in round r, based on the above three components. The agent conditions on it to generate an updated answer $a_i^{(r)}$, explanation $e_i^{(r)}$, and confidence $p_i^{(r)}$, to be used in the next round. Demonstrations of convincing explanations enable the agent to generate explanations that are more likely to convince other agents to reach a better consensus.

Phase 3: Team Answer Generation. RECONCILE continues the discussion for a maximum of R rounds or terminates it as soon as a consensus is reached (i.e., all agents agree on the same answer). At the end of any round r, RECONCILE generates the team answer $\hat{a}^{(r)}$ for that round using a weighted voting scheme (see the right side of Fig. 2). In particular, we recalibrate each agent's confidence using a function $f(\cdot)$ and then use these as weights to compute the team answer, as follows:

$$\hat{a}^{(r)} = \arg\max_a \sum_i f(p_i^{(r)}) \mathbb{1}(\hat{a}_i^{(r)} = a)$$

where a is a distinct answer generated by any of the agents, $p_i^{(r)}$ is the original confidence of agent A_i in round r and $f(p_i^{(r)})$ is the corresponding recalibrated confidence. While an unweighted majority vote and uncalibrated confidence-weighted vote also work well in practice, we use the calibrated weighted vote because it not only obtains slightly better results but the same recalibration strategy also works out-of-the-box for all seven tasks that

[4]We did not include an agent's own convincing samples in the prompt because an agent is expected to specifically convince *other* agents. We also verify this empirically – additionally including self-convincing samples in the prompt leads to comparable performance.

we experiment with (see Appendix B.5 for more details of our recalibration function $f(\cdot)$).

5 Experimental Setup

Agents in ReConcile. We primarily implement ReConcile with ChatGPT, Bard, and Claude2 as the three agents, engaging them in up to three rounds of discussion. Later in §6.1, we also show the generalizability of our ReConcile framework with different choices of agents, including API-based (GPT-4), open-source (LLaMA-2-70B), and domain-specific (DeepSeekMath) agents.

Datasets. We evaluate ReConcile on seven benchmarks, including two commonsense, three math, one logical reasoning, and one NLI task. These are: (1) StrategyQA (Geva et al., 2021), (2) CommonsenseQA (CSQA; (Aggarwal et al., 2021; Talmor et al., 2019)), (3) GSM8K (Cobbe et al., 2021), (4) AQuA (Ling et al., 2017), (5) MATH (Hendrycks et al., 2021), (6) Date Understanding (BIG-bench collaboration, 2023), and (7) ANLI (Nie et al., 2020).

Baselines. We compare ReConcile to prior works in three categories:

- **Vanilla single-agent methods.** In this category, we experiment with (1) zero-shot CoT prompting (Kojima et al., 2022) with one of the interacting LLMs, and (2) eight-shot CoT with Claude2 where the number eight matches the number of convincing samples used in ReConcile.
- **Advanced single-agent methods.** Next, we compare with (1) Self-Refine (SR) that iteratively generates feedback and refines the output leveraging the model itself (Madaan et al., 2023), (2) Self-Consistency (SC) that samples multiple reasoning paths and generates the most consistent answer (Wang et al., 2023b), and (3) their combination, SR+SC, that first conducts multiple iterations of refinement, followed by a majority vote. Note that in ReConcile, the number of LLM calls per instance can vary between 3, 6, and 9 based on the number of discussion rounds. Hence, for a fair comparison, we implement SC with the same average number of LLM calls as in ReConcile. Later in Appendix B.3, we show that ReConcile even outperforms 9-way SC (that equates to the worst-case LLM calls in ReConcile).
- **Multi-agent methods with a single backbone model.** Our final baselines are two multi-agent debating methods: a multi-agent debate between multiple ChatGPT instances (Du et al., 2023) and a debate with judge method (Liang et al., 2023). These methods use multiple instances of the same underlying model (ChatGPT) as different agents.

Implementation Details. Owing to the cost associated with API-based models and the limit imposed on the number of API calls, we follow many prior works (Du et al., 2023; Bian et al., 2023; Besta et al., 2023; Yao et al., 2023a) to experiment with a subset of 100 samples (from the validation set for StrategyQA and the test set for all other datasets). Later in Appendix B.1, we also experiment on the full test sets of StrategyQA and Date understanding and find similar trends. We report accuracy and its standard deviation. For each experiment, we conduct at least three runs on the same test samples with the same prompts, primarily accounting for the variance caused by the decoding strategy. Other implementation details can be found in Appendix A.1.

6 Results

6.1 Main Results

ReConcile outperforms single-agent and multi-agent baselines. We first evaluate the overall reasoning capabilities of ReConcile in Table 2 with ChatGPT, Bard, and Claude2 as the three agents. For fair comparisons, all iterative methods go through 3 rounds of iteration and all single-model multi-agent baselines are implemented with three agents with a sufficiently high temperature of 1.0 for maximizing diversity. Across all five datasets, ReConcile outperforms all single-agent and multi-agent baselines that are built on top of the same models (see last row). Notably, without using GPT-4 as an agent, our method outperforms GPT-4 on commonsense tasks like StrategyQA and CSQA and obtains comparable performance to GPT-4 on most other tasks. GPT-4's especially strong results on GSM8K could be attributed in part to the inclusion of some of GSM8K's training samples in GPT-4's pre-training data (OpenAI, 2023). While multi-agent debate with ChatGPT (Du et al., 2023) improves results on math benchmarks, debate with multiple Bard or Claude2 instances is not effective, possibly because the responses (generated from the same model) are not sufficiently diverse. When they team up with ChatGPT in a multi-round discussion, ReConcile outperforms debate frameworks. It obtains maximum gains of

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

Table 2: Comparison of RECONCILE (using ChatGPT, Bard, Claude2) with vanilla and advanced single-agent methods and multi-agent debating frameworks. Across all reasoning benchmarks, RECONCILE outperforms all prior single-agent and multi-agent methods. On commonsense tasks (StrategyQA and CSQA), RECONCILE also outperforms GPT-4. All results are on a random subset of 100 samples. The agents are GPT-4, ChatGPT, Bard, and Claude2.

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

Table 3: Comparison of the best single-agent, best multi-agent, and RECONCILE on StrategyQA for a given combination of three agents. RECONCILE flexibly incorporates agents with varying strengths, such as a stronger model like GPT-4, or an open-source model like LLaMA2-70B.

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

Table 4: RECONCILE generalizes to specialized models like DeepSeekMath and improves on a challenging mathematical reasoning benchmark, MATH.

11.4% (75.3% → 86.7%) on date understanding and 7.7% (71.3% → 79.0%) on StrategyQA when compared to the strongest baseline (multi-agent debate with Claude2). Improvements in the math reasoning tasks are relatively moderate, because of ChatGPT's initial strong performance. However, as demonstrated later in Table 4, integrating a specialized math reasoning model into RECONCILE significantly boosts team performance.

RECONCILE generalizes to agents of varying strengths. Next, we vary the agents in RECONCILE to study its generalization as a multi-agent framework. In particular, we either include (a) a stronger GPT-4 model, or (b) an open-source LLaMA-2-70B-chat model in the discussion. As shown in Table 3, in both these scenarios, RECONCILE outperforms the best single-agent and multi-agent baselines, notably even outperforming the zero-shot GPT-4 performance by 12.1% (75.6% → 87.7%) on StrategyQA. This highlights the potential of a stronger agent to also obtain useful external feedback from comparatively weaker agents.

RECONCILE generalizes to domain-specific agents. So far, we have experimented with RECONCILE variants that employed general-purpose models like ChatGPT as agents. Our next result in Table 4 shows that even for tasks that require substantial domain knowledge (e.g., the MATH benchmark (Hendrycks et al., 2021)), RECONCILE is flexible enough to utilize and improve upon specialized, domain-specific models. Recently, Shao et al. (2024) proposed DeepSeekMath, a 7B model pre-trained on a large number of math-related web corpus and improving over GPT-4. Notably, RECONCILE with GPT-4, Claude2, and DeepSeekMath as agents significantly outperforms zero-shot DeepSeekMath and GPT4-based Debate by 7.8% and 9.6% respectively. In summary, RECONCILE shows consistent improvements across a wide range of agent combinations (involving API-based, open-source, and domain-specific models).

RECONCILE also improves Natural Language Inference. While all our previous results were with reasoning tasks, we also demonstrate RECONCILE's effectiveness on ANLI (Nie et al., 2020),

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**

Table 5: Comparison of diversity between (a) paraphrased responses (first row) and (b) responses from multiple instances of the same ChatGPT model (second row). ReConcile with a multi-model component also leads to higher accuracy. Responses from different models in ReConcile (last row) are most diverse (i.e., less similar).

Method	Accuracy
Best Single-agent (zero-shot)	51.3 ( )
Best Multi-agent (Debate)	48.3 (×3)
ReConcile	**57.7** (, ,)

Table 6: ReConcile improves a challenging NLI benchmark (ANLI), outperforming Debate by 9.4%.

Method	Accuracy
ReConcile	**79.0**$_{\pm 1.6}$
w/o Multiple Models	72.2$_{\pm 2.1}$
w/o Grouping	76.7$_{\pm 2.5}$
w/o Convincingness	74.5$_{\pm 1.7}$
w/o Conf Estimation	77.7$_{\pm 1.3}$

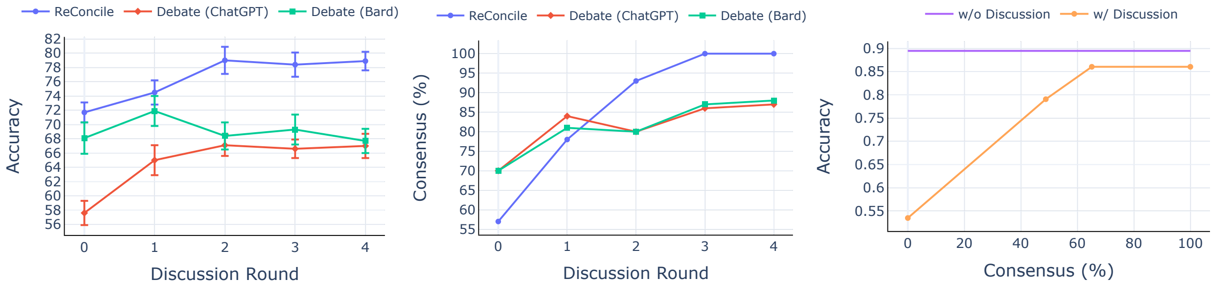
Table 7: Ablations of ReConcile on StrategyQA.

a challenging Natural Language Inference benchmark. Table 6 shows that ReConcile on ANLI outperforms Debate by a significant 9.4%, pointing to its widespread applicability.

6.2 Ablations and Analysis of ReConcile

Each component of ReConcile improves reasoning. In Table 7, we evaluate individual components of ReConcile on StrategyQA. In particular, we compare four variants: (1) **w/o Multiple Models**: We use ChatGPT as the backbone for all three agents, (2) **w/o Grouping**: We simply concatenate the responses from different agents without grouping their answers, (3) **w/o Convincingness**: We remove convincing samples from all prompts, and (4) **w/o Confidence Estimation**: We do not use any confidence estimates during the discussion and compute majority vote as the team answer. We show that each component has a positive impact on ReConcile with varying capacities. The effect of different models as agents is particularly significant and we observe a 6.8% improvement compared to only using ChatGPT as all three agents. This reinforces our hypothesis (and further verified below in 'Diversity Analysis') that diverse LLMs have complementary strengths and when put together in a round table discussion, they can learn from diverse external feedback from other agents and refine their responses to reach a better consensus. Notably, convincing samples lead to a 4.5% improvement in accuracy. In Appendix B.2, we study the role of convincing samples to show that (1) they also improve other interaction frameworks, and (2) even in the absence of such examples, ReConcile outperforms debate baselines.

Different models enhance response diversity. As was shown in Table 7, ReConcile obtains the most improvements via its *multi-model* component. This surpasses ReConcile with multiple ChatGPT instances, even when the generations sampled from these instances are encouraged to exhibit high diversity with a sufficiently high temperature. To further validate the importance of having multiple models and the diversity brought about by them, we develop a diversity metric. We hypothesize that if explanations from different models are indeed more diverse than those generated from multiple instances of the same model (e.g., in Multi-agent Debate), then our diversity metric should capture that. With that goal, we define diversity between multiple agents as the summation of the pairwise diversity between agents: $D(A_1, A_2, A_3) = D(A_1, A_2) + D(A_1, A_3) + D(A_2, A_3)$, where A_1, A_2, and A_3 are the three agents' initial responses (either belonging to the same underlying model or different models). We then measure pairwise diversity by computing the cosine similarity between the response embeddings with BERTScore (Zhang et al., 2019). Note that lower similarity scores will mean greater diversity. With the diversity metric defined, we compute this metric for three variants: (a) paraphrased responses of a single ChatGPT to serve as a baseline, (b) responses from ReConcile using three instances of a single ChatGPT model, and (c) responses from ReConcile with ChatGPT, Bard, and Claude2 as agents. In Table 5, we show that responses from different models exhibit the highest diversity (yielding the lowest similarity score of 0.8739) and also the highest accuracy (79.0%), followed by the single-model variant



(a) (b) (c)

Figure 4: RECONCILE achieves better and faster consensus. (a) Comparison of RECONCILE with Debate baselines showing the accuracy after each round. (b) Fraction of samples for which a consensus is reached after each round. (c) Accuracy as a function of consensus.

Round	ChatGPT	Bard	Claude2	Team
0	71.0±2.1	71.7±0.9	73.7±1.7	74.3±1.2
1	71.3±0.9	77.7±1.2	75.3±0.8	77.0±0.9
2	76.7±0.8	**77.3±1.4**	**77.7±0.9**	**79.0±0.5**
3	**77.0±0.9**	76.7±0.8	77.0±1.2	78.7±1.2

Table 8: The round-wise accuracy of ChatGPT, Bard, and Claude2 and their team performance (using weighted vote) on StrategyQA.

(with a similarity score of 0.9102) and the paraphrased variant (with a similarity score of 0.9398). Thus, the higher diversity of (multi-model) RECONCILE means that agents have access to alternate solutions and external feedback, leading to better discussion and reasoning accuracy. We also present a case study in Appendix C.5 to illustrate that the debate baseline sometimes struggles with echo chambers, stemming from a lack of external feedback, supporting the need for external feedback for improving LLMs (Huang et al., 2023).

RECONCILE improves all agents individually. We showed that the team performance of the agents improves through discussion. Next, in Table 8, we also present the accuracy of each agent after every round, as well as the overall team accuracy for StrategyQA. Evidently, the individual performance of each agent also improves alongside the team's performance.

RECONCILE Reaches Faster and Better Consensus. RECONCILE terminates the discussion when a consensus is reached. More discussion rounds are costlier due to the increased API calls. Hence, achieving faster consensus while maintaining comparable accuracy gains is more efficient. To study this, in Fig. 4(a), we plot the accuracy trends after each round; in Fig. 4(b), we plot the fraction of samples for which consensus has been reached; and in Fig. 4(c), we analyze accuracy as a function of consensus. From the first plot, we make two important observations: (1) RECONCILE improves accuracy for two rounds, following which the accuracy saturates, (2) Compared to the debate baselines, RECONCILE is not only superior after every round but also peaks at a highest accuracy of 79.0% (vs 71.3% for the baselines). Next, from Fig. 4(b), our observations are also two-fold: (1) In the initial rounds (0 and 1), RECONCILE's consensus percentage is lower because the discussion takes place between diverse LLMs. Diverse agents lead to more differences in opinions initially. (2) However, as the discussion proceeds, RECONCILE establishes consensus for all samples by round 3, while in the baseline, 13% of the samples do not converge even after round 4. Finally, Fig. 4(c) shows that for the samples that enter the discussion phase (i.e., their initial answers did not have a consensus), accuracy is positively correlated with consensus. In other words, as a greater number of samples reach a consensus, accuracy proportionally improves. In summary, RECONCILE reaches *faster* and *better* consensus compared to baselines.

7 Conclusion

We presented RECONCILE, a multi-agent framework for reasoning with diverse LLM agents, engaged in multiple rounds of discussion via confidence estimation and generating explanations that can correctively convince other agents. RECONCILE demonstrated strong results on multiple reasoning benchmarks, consistently outperforming prior single-agent and multi-agent baselines and even improving upon GPT-4 on some benchmarks.

Limitations

For the API-based models used in ReConcile, we note that we lack complete knowledge of the data that these models have been exposed to, and their scales in terms of parameters. Moreover, due to the API access, we do not possess complete control over their behavior. Depending on API-based models also necessitates the need to prompt these models to estimate their confidence. While this approach proves effective as evidenced by our results, we note that these estimates remain post-hoc in nature. Nevertheless, it is worth highlighting that these limitations could potentially be mitigated in the future should more open-sourced models emerge and demonstrate robust capabilities in adhering to long instructions.

Acknowledgments

We thank Peter Hase and Elias Stengel-Eskin for useful feedback and suggestions regarding experiments. This work was supported by NSF-CAREER Award 1846185, NSF-AI Engage Institute DRL-2112635, DARPA MCS Grant N66001-19-2-4031, Accelerate Foundation Models Research program, and a Google PhD Fellowship. The views contained in this article are those of the authors and not of the funding agency.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for commonsenseqa: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Anthropic. 2023. model card and evaluations for claude models.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

BIG-bench collaboration. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate.

Elias Stengel-Eskin and Benjamin Van Durme. 2023. Calibrated interpretation: Confidence estimation in semantic parsing. In *TACL*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *ICML*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for "mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

Shuang Li, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, and Igor Mordatch. 2022a. Composing ensembles of pre-trained models via iterative consensus. *International Conference on Learning Representations (ICLR)*.

Yanhong Li, Gang Kou, Guangxu Li, and Yi Peng. 2022b. Consensus reaching process in large-scale group decision making based on bounded confidence and social network. *European Journal of Operational Research*, 303(2):790–802.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *ACL*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Marvin Minsky. 1988. *Society Of Mind*. Simon and Schuster.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2023. Can language models teach weaker agents? teacher explanations improve students via theory of mind. In *NeurIPS*.

Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. Peer: A collaborative language model. *arXiv preprint arXiv:2208.11663*.

Zhihong Shao, Peiyi Wang, Runxin Xu Qihao Zhu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2023. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yuqing Wang and Yun Zhao. 2023. Metacognitive prompting improves understanding in large language models. *arXiv preprint arXiv:2308.05342*.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023c. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023a. Examining the inter-consistency of large language models: An in-depth analysis via debate.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023b. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering questions by meta-reasoning over multiple chains of thought. *arXiv preprint arXiv:2304.13007*.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Andy Zeng, Maria Attarian, brian ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2023. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, et al. 2023. Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066*.

A Additional Details of ReConcile

A.1 Implementation Details

We provide more implementation details of ReConcile in this section. During decoding, we set the temperature to 0.7 for ChatGPT and Bard and use the default setting for Claude2. All implementations involving ChatGPT are using *gpt-3.5-turbo-0613* from Azure OpenAI.[5] We retrieve results from Claude2 by posting requests to their webpage[6], and for Bard, we use *chat-bison-001* from PaLM2 API[7]. For each agent, we use four demonstrations of convincing samples. In addition, we provide the workflow of ReConcile in Algorithm 1. Required input contains a test problem Q, maximum number of discussion rounds R, n agents $\mathcal{A} = \{A_i\}_{i=1}^n$, and convincing samples $\mathcal{C} = \{C_i\}_{i=1}^n$ for each agent. The output would be the team answer $\hat{a}^{(r)}$. For the open-source models LLaMA2-70B and DeepSeekMath, we use four RTX A6000 GPUs, each with 48GB memory to generate output from them.

Initial Prompt

{convincing_samples}
Q: ***{test_question}***
Please answer the question with step-by-step reasoning. Also, evaluate your confidence level (between 0.0 and 1.0) to indicate the possibility of your answer being right.

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}***

Figure 5: The prompts used in ReConcile consist of an initial prompt and a discussion prompt.

A.2 Initial Prompt and Discussion Prompt

We show the prompts used in ReConcile in Fig. 5. The initial prompt encompasses (1) the convincing samples that demonstrate how to convince other agents, (2) the test question, and (3) a requirement for 'step-by-step' reasoning. The prompt also instructs the agent to express their confidence level, ranging from 0.0 to 1.0, indicating the likelihood of their answer being correct. The discussion prompt is an extension of the initial prompt, instructing the agent to review and express agreement or disagreement with other agents' solutions. To facilitate discussions, we design a grouping scheme that aggregates information based on the current opinions at the table. For instance, if two agents affirm that the answer to a given question is 'yes' while the third agent disagrees with a 'no', the designed grouping mechanism in the discussion prompt consolidates this information rather than simply concatenating all responses.

[5]https://oai.azure.com/
[6]https://claude.ai/chats
[7]https://developers.generativeai.google/products/palm

Model	StrategyQA	Date
ChatGPT	68.1	69.3
Bard	70.6	52.8
Claude2	72.7	77.9
Multi-agent Debate	71.4	72.4
ReConcile	**78.4**	**84.5**

Table 9: Comparison of ReConcile with baselines on the full test sets of StrategyQA and Date Understanding.

Method	Accuracy
Debate (Du et al., 2023)	66.7$\pm$3.1
RC (w/o Convincing Expl)	74.5$\pm$1.7
RC (w/ Random Expl)	75.0$\pm$2.5
RC (w/ Convincing Expl)	79.0$\pm$1.6
Debate (w/ Random Expl)	68.7$\pm$2.2
Debate (w/ Convincing Expl)	69.5$\pm$1.7

Table 10: Evaluation of the role of convincing samples on StrategyQA. ReConcile (RC) without convincing samples outperforms multi-agent debate and with it obtains further gains. Convincing samples also boost the debate baseline.

B Additional Results

B.1 Results on Full Test Sets

In Table 2, we reported results with 100 test samples following several previous works and due to budget constraints. Upon experimenting on the full test sets of StrategyQA and Date Understanding, we confirm similar trends. Specifically, in Table 9, we compare ReConcile to all of our major baselines and show that ReConcile continues to outperform all baselines.

B.2 Convincing Samples Improve Both ReConcile and Multi-agent Debate

Recall that ReConcile selects a sample as convincing if the corresponding human explanation

Algorithm 1 RECONCILE: A Group-Discuss-And-Convince Framework

Require: Test Problem Q, Discussion Rounds R, Agents $\mathcal{A} = \{A_i\}_{i=1}^n$, Convincing Samples $\mathcal{C} = \{C_i\}_{i=1}^n$

```
function RECONCILE(Q, R, A, C)
    r ← 0
    while r ≤ R and not CONSENSUS(Q, {a_i^(r−1)}_{i=1}^n) do
        S ← [], P ← []
        for each A_i ∈ A do
            if r = 0 then
                P_I ← (Q, C)                                   ▷ Initial prompt consists of question and convincing samples
                a_i^(0), e_i^(0), p_i^(0) ← A_i(P_I)           ▷ Generate initial answer, explanation, and confidence
            else
                P_D ← (Q, a_i^(r−1), e_i^(r−1), p_i^(r−1), C)  ▷ Discussion prompt
                a_i^(r), e_i^(r), p_i^(r) ← A_i(P_D)
            end if
            S ← S + [a_i^(r)], P ← P + [p_i^(r)]               ▷ Append each agent's answer and confidence
        end for
        â^(r) ← WEIGHTEDVOTE(S, P)                             ▷ Get team answer through a confidence weighted vote
    end while
    return â^(r)
end function
```

rectifies an agent's incorrect answer. Based on this, Table 7 showed that by collecting only four human explanations, we can obtain significant improvements ('w/o Convincingness' row). Next, we consider a scenario where no human explanations are present. Table 10 shows that even then, RECONCILE outperforms the debate baseline by absolute 7.8 points (second row). If random (i.e., general human explanations that may not necessarily ensure answer rectification) are available (third row), we obtain some small improvements; but our convincing samples that are selected based on our novel answer-rectification criterion (fourth row) improve the results substantially. See Sections C.3 and C.4 for illustrative examples. Being able to convince another agent is also a generic concept that can be applied to other multi-agent systems, as demonstrated by improvements in the debate baseline (last row).

B.3 Comparison with Other Methods

In Table 11, we compare RECONCILE to two other single-agent variants. While in our main Table 2, we experimented with a random 8-shot Claude2 baseline, here we replace the in-context samples with our convincing samples. Even then, RECONCILE exhibits superior performance on all datasets except for GSM8K, again highlighting the importance of collaboration between diverse models. Next, we also report results for 9-way Self-Consistency which in terms of LLM calls represents the worst-case scenario of RECONCILE – even for a more open-ended dataset like GSM8K, 9 LLM calls (i.e., 3 discussion rounds) happen in only 12% of the samples and an even lesser 9% on multiple-choice QA dataset like Date understanding. That said, RECONCILE continues to outperform 9-way SC by a large margin on most datasets.

B.4 Recalibration Strategy of RECONCILE

Directly using confidence scores as the voting weights is less effective due to the overconfidence problem of LLMs (Xiong et al., 2023b; Tian et al., 2023; Mielke et al., 2022). Specifically, LLMs tend to produce consistently high confidence scores, which can make it challenging to discern subtle distinctions in confidence levels across different outputs. To address this, we employ a simple yet effective rescaling technique, facilitating better differentiation of confidence levels. This is expressed as:

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

where $p_i^{(r)}$ is the original confidence of agent A_i in round r and $f(p_i^{(r)})$ is the corresponding adjusted score. To decide the optimal weights, we compare with a variety of settings including the majority vote and the uncalibrated confidence-weighted vote. The results are summarized in Table 13. We denote the weight we used in our main experiment as $w^* = [1.0, 0.8, 0.5, 0.3, 0.1]$ where each value corresponds to the recalibrated confidence score. We further compare with other settings:

Model	StrategyQA	CSQA	GSM8K	AQuA	Date
Claude2 (w/ 8-shot convincing samples)	74.0±0.0	69.7±1.2	85.3±0.5	64.3±1.2	81.3±0.5
Self-Consistency w/ ChatGPT (9-way)	74.7±0.8	73.3±1.2	**85.7**±0.4	62.7±1.2	70.3±0.9
RECONCILE	**79.0**±1.6	**74.7**±0.4	85.3±2.2	**66.0**±0.8	**86.7**±1.2

Table 11: Comparison of RECONCILE with Claude2 using 8-shot convincing samples and 9-way Self-Consistency.

	Max Conf	Majority Vote	Weighted Vote
Accuracy	74.7±2.1	77.1±1.3	**79.0±0.5**

Table 12: Performance comparison of different voting strategies on StrategyQA. Weighted vote performs the best compared to simple majority vote and choosing the agent's answer with highest confidence.

Voting weight	StrategyQA	GSM8K
w_1	0.77	0.84
w_2	0.79	0.83
w_3	0.78	0.82
w_4	0.77	0.83
Majority	0.76	0.83
Uncalibrated	0.78	0.84
w^* (Ours)	**0.79**	**0.85**

Table 13: The robustness of the recalibration weight. We use the same weights w^* across all datasets.

- $w_1 = [1.0, 0.9, 0.7, 0.5, 0.3]$
- $w_2 = [1.0, 0.9, 0.5, 0.3, 0.1]$
- $w_3 = [1.0, 0.8, 0.6, 0.4, 0.2]$
- $w_4 = [1.0, 0.75, 0.5, 0.25, 0.0]$

and the results show that our w^* works the best across datasets. In our main experiment, we fix the weight using w^* and it is constantly outperforming majority vote across all seven datasets. In addition, Fig. 9 shows that it helps reduce the Expected Calibration Error (ECE), a popular calibration metric (Naeini et al., 2015). While we note that recalibration can also be achieved through a learned model (e.g., Platt Scaling (Platt et al., 1999)), we refrain from using such models because RECONCILE is primarily designed as a few-shot method, and developing a recalibration model would necessitate access to a substantial number of annotated samples. Therefore, we use $f(p_i^{(r)})$ to perform a weighted vote to generate the team answer.

B.5 Comparison of Different Voting Strategies

At the end of any round r, every agent in RECONCILE generates its answer. Here we explore three voting strategies: (1) maximum confidence vote, where the agent's answer with the maximum confidence score would be the final team answer,

Dataset	License
StrategyQA	MIT License (License)
CommonsenseQA	MIT License (License)
GSM8K	MIT License (License)
AQuA	Apache 2.0 (License)
MATH	MIT License (License)
Date	Apache 2.0 (License)
ANLI	CC BY-NC 4.0 (License)

Table 14: Dataset licenses

(2) unweighted majority vote, where each vote carries equal weight, irrespective of the confidence score, and (3) weighted vote, where we use the recalibrated confidence scores as the voting weights. As shown in Table 12, weighted vote is the most effective way to aggregate the team answer.

C Qualitative Examples

C.1 Convincing Samples for Each Agent

Table 15 shows examples of convincing samples on StrategyQA for each agent.

C.2 Effect of Convincing Samples

Here, we provide qualitative examples of how convincing samples change the way each agent responds to the question. We compare the initial responses (of each agent) with and without convincing samples in Table 16.

C.3 RECONCILE w/o Convincing Samples

We notice that when RECONCILE operates in the absence of convincing samples, the agents tend to maintain their initial opinions more often. As depicted in Fig. 6, all three agents adhere to their original stances throughout the entire discussion and hence never converge to the correct answer.

C.4 RECONCILE with Convincing Samples

On the contrary, when convincing samples are present, we show how the explanations of all agents change during the course of a discussion (see Fig. 7). Initially, Bard and Claude2 provide incorrect answers, but as the discussion unfolds, both

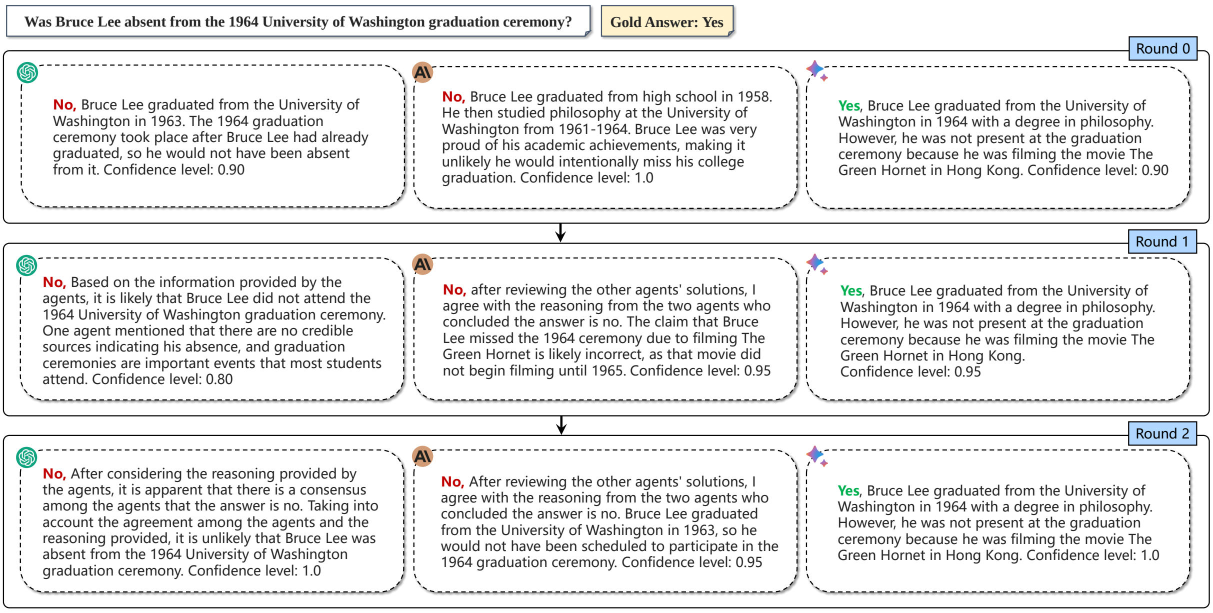


Figure 6: An illustrative example showing that without demonstrations of convincing samples, agents' opinions do not change much across the discussion rounds. All agents maintain their original stances and hence, never converge to the correct answer.

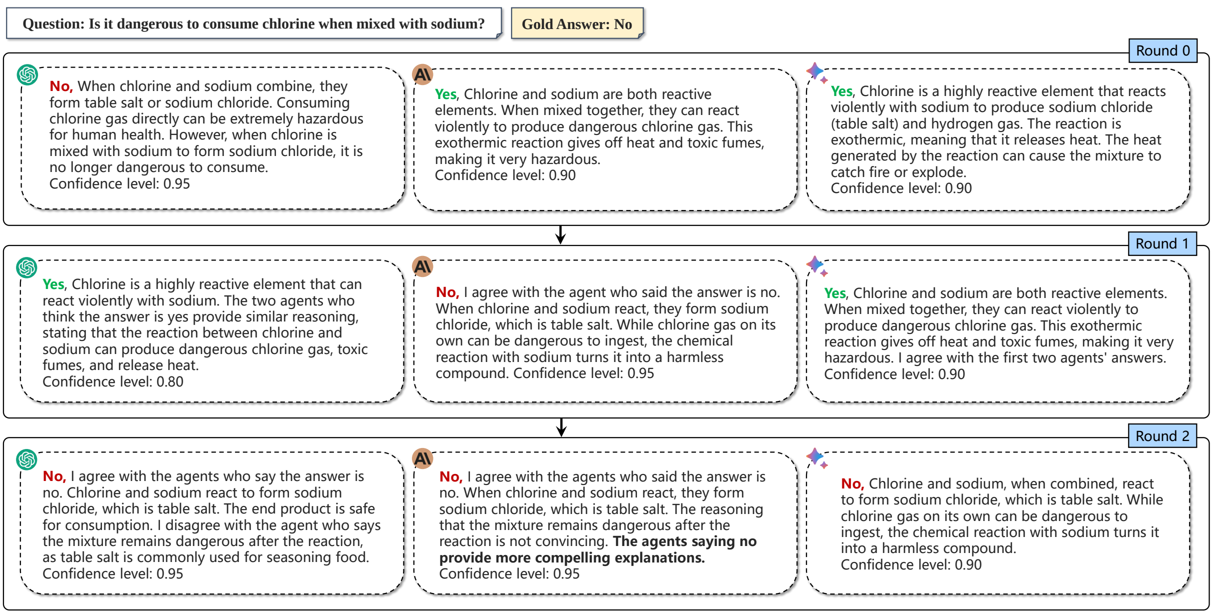


Figure 7: An illustrative example showing that with access to demonstrations of convincing samples, agents' opinions and explanations change during the course of multiple discussion rounds.

agents revise their initial predictions, ultimately arriving at the correct answer.

C.5 Single-Model Multi-Agent Debate Struggles with Echo Chamber

In Fig. 8, we provide an illustration of multi-agent debate, implemented with multiple instances of the same underlying ChatGPT model. In this case, an incorrect answer is initially provided, but because external feedback from diverse models is lacking, all agents persist with the same incorrect response throughout the interaction.

D Benchmark Licenses

The licenses of datasets we used are in Table 14.

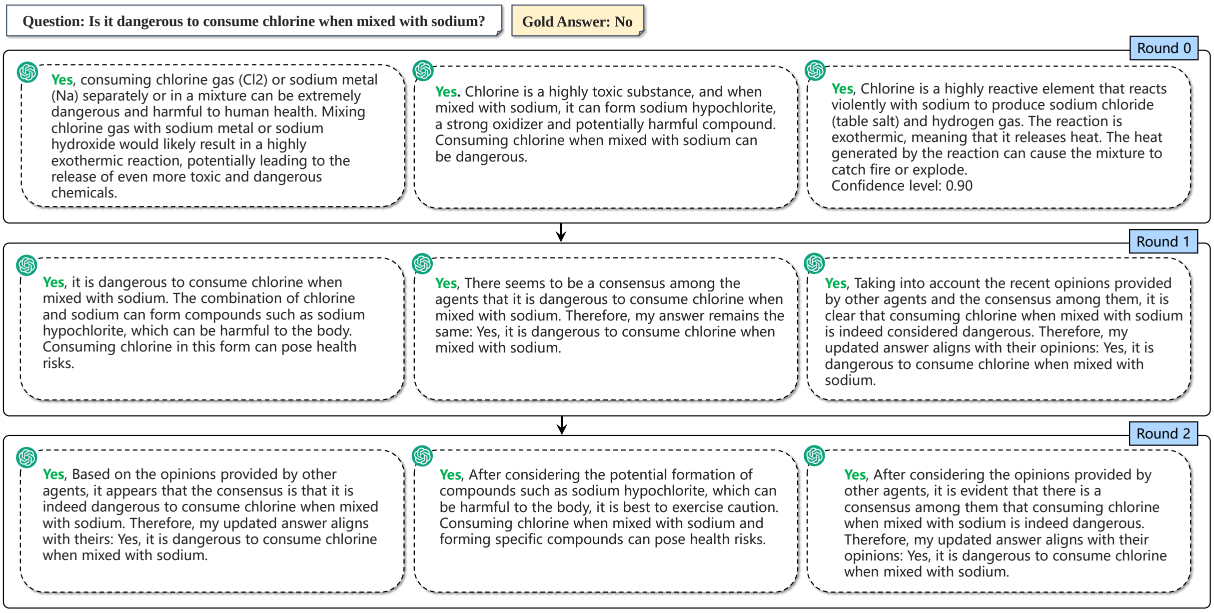


Figure 8: An illustrative example of multi-agent debate with multiple instances of ChatGPT. Initially, an incorrect answer is provided, and due to a lack of external feedback from diverse models, all agents persist with the same erroneous response throughout the debate process.

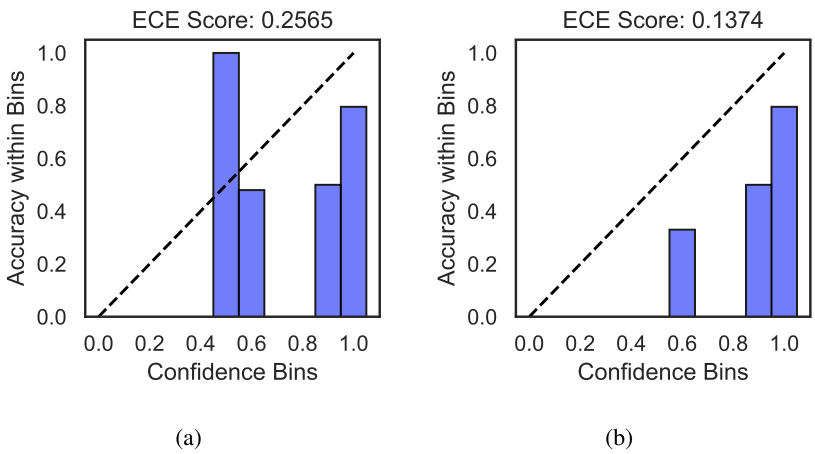


Figure 9: Expected Calibration Error (ECE) (Naeini et al., 2015; Guo et al., 2017; Elias Stengel-Eskin and Benjamin Van Durme, 2023) (a) before and (b) after confidence rescaling in RECONCILE. We observe a significant drop in ECE, showing the effectiveness of our simple method.

QUESTION: Is the tongue part of a creature's head? (Gold answer: **yes**) CHATGPT (INITIAL): The tongue is a muscular organ located in the mouth of most vertebrates, including humans. It is involved in functions such as taste, chewing, and speaking. The head, on the other hand, is the anatomical part that contains the brain, eyes, ears, nose, and mouth. Answer: **no**. CONVINCING HUMAN EXPLANATION: A creature's tongue is inside its mouth. A creature's mouth is part of its head. CHATGPT (W/ CONVINCING HUMAN EXPLANATION): Based on the information provided, the tongue is inside a creature's mouth. The mouth is part of a creature's head. Answer: **yes**.
QUESTION: Are slime lilies in a different scientific family than asparagus? (Gold answer: **no**) BARD (INITIAL): Slime lilies are in the scientific family Araceae while asparagus is in the scientific family Asparagaceae. Answer: **yes**. CONVINCING HUMAN EXPLANATION: Asparagus is a species of plants of the Asparagaceae family. Slime lilies are the common name for the flowering albuca plant. The albuca plant belongs to the scientific family of Asparagaceae. BARD (W/ CONVINCING HUMAN EXPLANATION): Slime lilies are in the scientific family of Asparagaceae, while asparagus is in the same family. Answer: **no**.
QUESTION: Would an astrologer focus on the densest terrestrial planet for a Friday horoscope? (Gold answer: **yes**) CLAUDE2 (INITIAL): Astrologers do not focus on Earth for Friday horoscopes specifically. Astrologers tend to focus on the positions of the Sun, Moon, and planets in the zodiac for horoscopes. The day of the week is not especially relevant. Answer: **no**. CONVINCING HUMAN EXPLANATION: Friday is associated with Venus in astrology Venus is the densest of the terrestrial planets CLAUDE2 (W/ CONVINCING HUMAN EXPLANATION): Friday is associated with Venus in astrology, and Venus is the densest of the terrestrial planets (Mercury, Venus, Earth, Mars). So the question is asking if an astrologer would focus on Venus for a Friday horoscope. Answer: **yes**.

Table 15: Examples of convincing human explanations for each agent (ChatGPT, Bard, and Claude2). A convincing explanation is an instance of a answer-rectifying human explanation. For all three examples, the convincing explanation helps rectify an agent's initial incorrect answer.

QUESTION: Was Lil Jon's top ranked Billboard song a collaboration with a member of The Lox? (Gold answer: **no**) CHATGPT (W/O CONVINCING): To determine if Lil Jon's top ranked Billboard song was a collaboration with a member of The Lox, we need to search for Lil Jon's discography and look for his highest-charting song on the Billboard charts. Then, we need to check if this song is a collaboration with a member of The Lox. Answer: **no**. Confidence: 0.8. CHATGPT (W/ CONVINCING): Lil Jon's top ranked Billboard song is 'Yeah!' which features vocals from Usher and Ludacris. There is no collaboration between Lil Jon and any member of The Lox on this song. Answer: **no**. Confidence: 0.9.
QUESTION: Could ten gallons of seawater crush a six year old? (Gold answer: **yes**) BARD (W/O CONVINCING): Ten gallons of seawater weighs about 82 pounds. The average weight of a six year old is about 44 pounds. Therefore, ten gallons of seawater would be about twice as heavy as a six year old. **It is possible that ten gallons of seawater could crush a six year old, but it is also possible that the six year old could survive.** The outcome would depend on a number of factors, such as the size and strength of the six year old, the shape of the container holding the seawater, and the amount of time the six year old was submerged in the seawater. Answer: **yes**. Confidence: 0.7. BARD (W/ CONVINCING): The density of seawater is about 1.025 grams per cubic centimeter. Ten gallons of seawater is about 37.85 liters. The volume of a human body is about 60 liters. Therefore, ten gallons of seawater would be about 63% of the volume of a human body. If ten gallons of seawater were to be poured on a six year old, it would be very difficult for the child to breathe and they would likely drown. Answer: **yes**. Confidence: 0.95.
QUESTION: Would the top of Mount Fuji stick out of the Sea of Japan? (Gold answer: **yes**) CLAUDE2 (W/O CONVINCING): To determine if the top of Mount Fuji would stick out of the Sea of Japan, I need to compare the elevation of Mount Fuji to the depth of the Sea of Japan. According to Wikipedia, the elevation of Mount Fuji is 3,776 meters. The average depth of the Sea of Japan is 1,752 meters according to Wikipedia. Since the elevation of Mount Fuji is greater than the average depth of the Sea of Japan, I can conclude that the top of Mount Fuji would stick out above the surface of the Sea of Japan if it was placed there. Answer: **yes**. Confidence: 0.9. CLAUDE2 (W/ CONVINCING): Mount Fuji is the highest mountain in Japan, with a summit elevation of 12,389 feet or 3,776 meters above sea level. The average depth of the Sea of Japan is about 3,742 meters. Since the summit of Mount Fuji is higher than the average depth of the Sea of Japan, the top of Mount Fuji would stick out above the water if it was located in the middle of the Sea of Japan. Answer: **yes**. Confidence: 0.9.

Table 16: Comparison of how each agent's reasoning changes with and without convincing samples. We observe that by conditioning on convincing samples, all agents tend to become more confident in their reasoning and generate less uncertain statements (shown in **bold**), which is also reflected in the actual confidence scores generated by each agent (e.g., goes up from 0.7 to 0.95 for Bard).