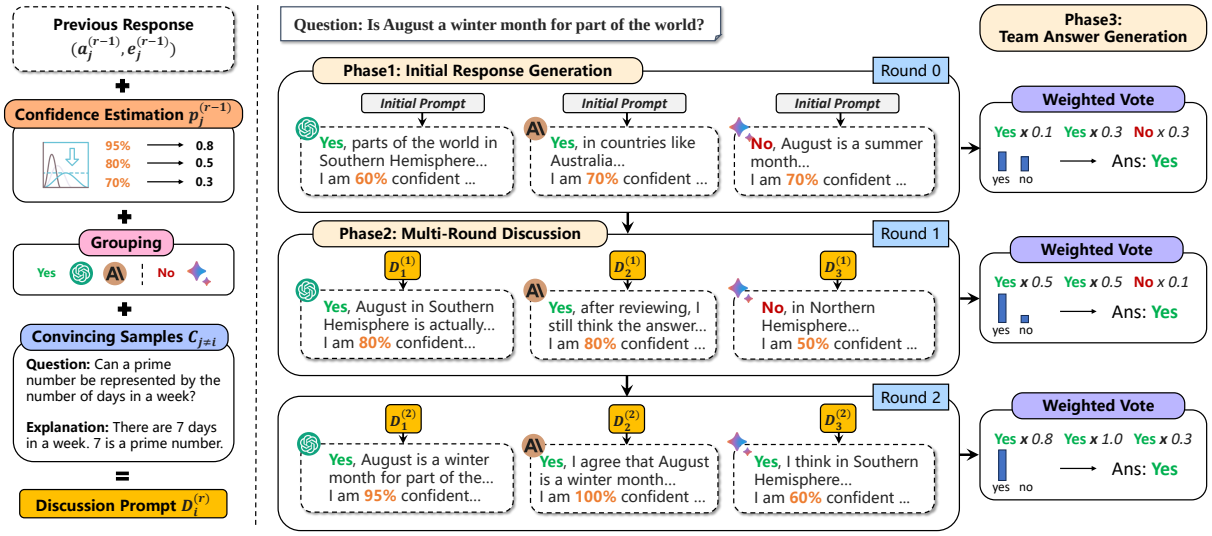

Figure 2: Overview of RECONCILE with ChatGPT, Bard, and Claude2, consisting of three phases: (1) Initial Response Generation: Each agent generates an initial answer and explanation. (2) Multi-Round Discussion: Each model is presented with a discussion prompt (as illustrated on the left) and subsequently 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

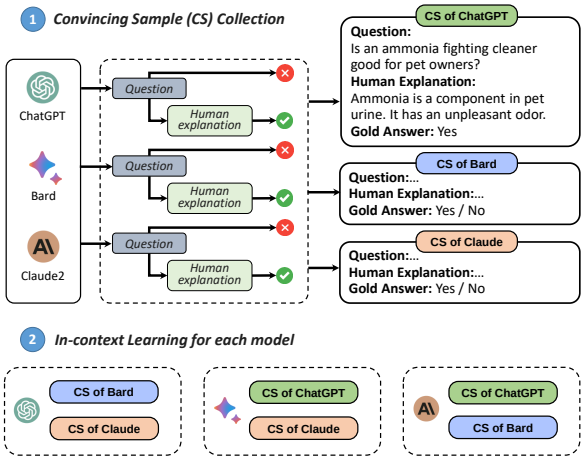

Figure 3: Method for choosing convincing samples for each agent. A convincing sample for ChatGPT consists of a question, a gold answer, and a 'corrective' human explanation that can rectify its initial incorrect answer. Then Bard and Claude2 use it in-context during discussion to convince ChatGPT.

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