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.

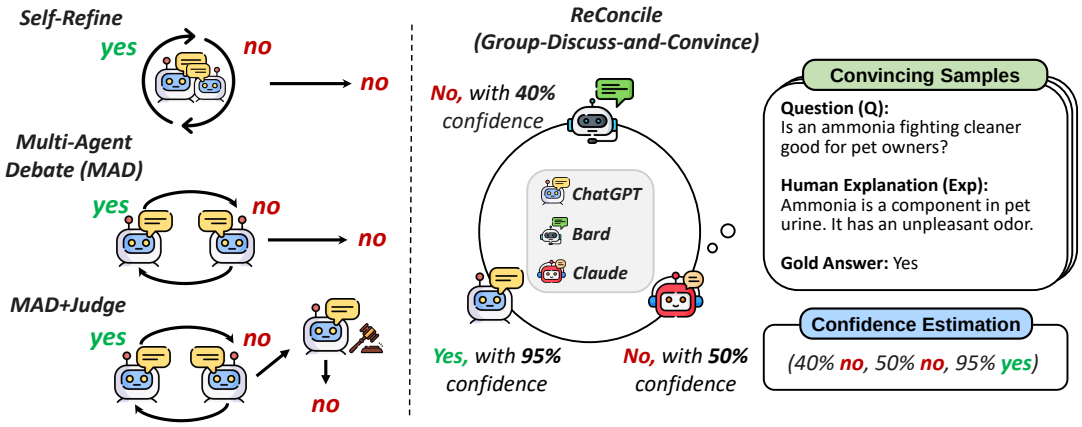

Figure 1: An illustration of the main differences between RECONCILE and prior works. While most current self-refine and debating techniques rely on multiple instances of a single model (e.g., ChatGPT), our method incorporates models from different families (e.g., ChatGPT, Bard, and Claude2). Our approach 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