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

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 (▣×3)	71.3 (⊛×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.

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.

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

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

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 (⊛, ✦, A\)	**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 (A\)
Best Multi-agent (Debate)	48.3 (⊛ ×3)
ReConcile	**57.7** (⊛, ✦, A\)

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

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

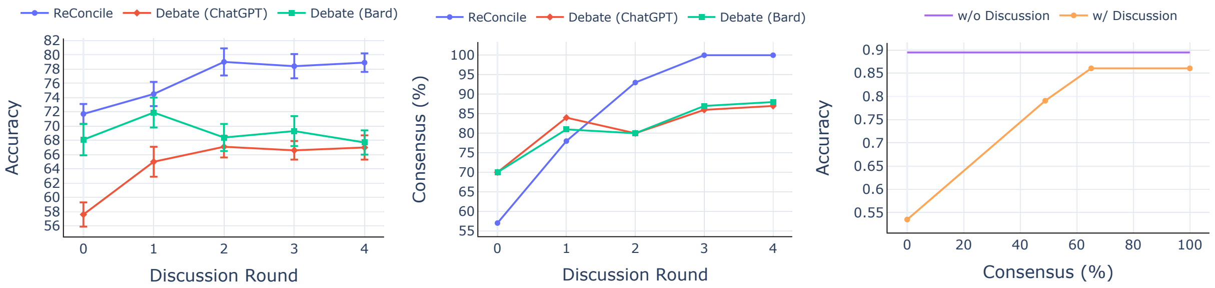
Table 7: Ablations of ReConcile on StrategyQA.

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.