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±3.1
RC (w/o Convincing Expl)	74.5±1.7
RC (w/ Random Expl)	75.0±2.5
RC (w/ Convincing Expl)	79.0±1.6
Debate (w/ Random Expl)	68.7±2.2
Debate (w/ Convincing Expl)	69.5±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:

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

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: