

A Appendix

In this appendix, we provide additional analysis and visualizations of the debates used in the main paper in Section A.1. We further provide detailed experimental details on each dataset in Section A.2.

A.1 Additional Results

Consensus Between Agents. In Figure 14, we illustrate the consensus between agents using either short or long consensus prompts discussed in Figure 3. The use of debate prompts that encourage agents to adapt more to the opinions of other agents improves consensus.

Additional Qualitative Visualizations. We added additional qualitative visualizations of the debate process. In Figure 16, Figure 17, Figure 18, Figure 19, Figure 20, we illustrate debates between agents in the GSM8K dataset which result in the correct answer. In Figure 21, Figure 22, Figure 23, we further illustrate debates in GSM8K which lead to the incorrect answer. We further provide an example illustration of debate in arithmetic in Figure 24, arithmetic with summarization of individual responses of agents in Figure 25, MMLU in Figure 26, a debate with the full contents biographies in Figure 27, and debate in chess in Figure 28. In general, we found that debate improved the performance of final generated answers, though sometimes answers would converge to the incorrect value.

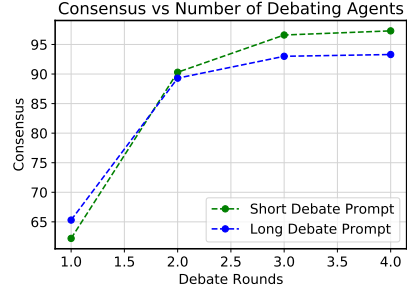


Figure 14: **Effect of Prompts on Consensus.** Using a short debate prompt induces faster consensus between agents

A.2 Evaluation Details

We provided detailed evaluation details for each setting in the paper. We run all experiments using the gpt-3.5-turbo-0301 model. We provide a table listing the prompts used to prompt models and initialize debate in Table 15.

Arithmetic. To evaluate the arithmetic task, we generated six random integers for each task between 0 and 30. We then evaluated the extent to which the correct integer answer was correctly obtained. We evaluated models on one hundred generated arithmetic tasks.

Grade School Math. To evaluate the GSM8K task, we evaluated the accuracy at which models were able to obtain the final correct answer, as extracted from a box. We evaluated models on one hundred grade school math problems.

Chess. To evaluate the chess reasoning task, we used chess games from <https://www.pgnmentor.com/players/Adams.zip>. We asked chatGPT to predict the next move for white to move at turn 14 and reported the relative Stockfish pawn score with search depth 20 after executing the suggested move from chatGPT. We evaluated models on three hundred selected chess games.

Biographies. To evaluate the biographies task, we compare each generated bullet point biography for a person with a ground truth set of facts about the person extracted from Wikipedia. We iteratively loop through each ground truth fact, and validate the extent to which the generated biography matches a particular bullet by prompting chatGPT with the prompt: *Consider the following biography of <person>: <generated biography> Is the above biography above consistent with the fact below? <ground truth bullet> Give a single-word answer, yes, no, or uncertain.* We then evaluate and report the percentage of ground bullets that chatGPT returns either yes or no on. We ignored ground truth bullets that chatGPT returns returned uncertain.

We found this evaluation metric provided a fast way to evaluate how relatively correct a generated bullet point biography is. However, we found that generated facts could contain incorrect information that was not captured in the ground truth bullet and thus could not be validated through this metric. Nevertheless, we believe this evaluation scheme estimates the relative accuracy of a generated biography.

MMLU. To evaluate MMLU, we measured the accuracy in which models were able to select the correct multiple-choice answer in each problem. We evaluated models on one hundred selected MMLU questions randomly distributed across each of the subject areas.

Task	Type	Prompt
Arithmetic	Starting	What is the result of $\{ \} + \{ \} * \{ \} + \{ \} - \{ \} * \{ \}$? Make sure to state your answer at the end of the response.
	Debate	These are the recent/updated opinions from other agents: <other agent responses> Use these opinions carefully as additional advice, can you provide an updated answer? Make sure to state your answer at the end of the response.
GSM8K	Starting	Can you solve the following math problem? <Problem> Explain your reasoning. Your final answer should be a single numerical number, in the form <code>\boxed{{answer}}</code> , at the end of your response.
	Debate	These are the solutions to the problem from other agents: <other agent responses> Using the solutions from other agents as additional information, can you provide your answer to the math problem? The original math problem is <Problem>. Your final answer should be a single numerical number, in the form <code>\boxed{{answer}}</code> , at the end of your response.
Chess	Starting	Here is the current sequence of moves in a chess game: <moves>. What is the best chess move I should execute next? Give a single move suggestion of the form 14. <XXX> and make sure the chess move is valid in the current board state.
	Debate	Here are other chess move suggestions from other agents: <other agent responses> Using the chess suggestions from other agents as additional advice and your earlier generated solution, can you give me your updated thoughts on the best next chess move I should play given the chess sequence ? Give a single move suggestion of the form 14. <XXX> and make sure the chess move is valid in the current board state.
Biographies	Starting	Give a bullet point biography of highlighting their contributions and achievements as a computer scientist, with each fact separated with a new line character.
	Debate	Here are some bullet point biographies of <person> given by other agents: <other agent response> Closely examine your biography and the biography of other agents and provide an updated bullet point biography.
MMLU	Starting	Can you answer the following question as accurately as possible? : A) , B) , C) , D) Explain your answer, putting the answer in the form (X) at the end of your response.
	Debate	These are the solutions to the problem from other agents: <other agent responses> Using the reasoning from other agents as additional advice, can you give an updated answer? Examine your solution and that other agents. Put your answer in the form (X) at the end of your response.
Chess Validity	Starting	Given the chess game , give one valid destination square for the chess piece at . State the destination square in the form (X), where X follows the regex <code>[a-h][1-8]</code> , for example (e5). Give a one line explanation of why your destination square is a valid move.
	Debate	Here are destination square suggestions from other agents: <other agent responses> Can you double check that your destination square is a valid move? Check the valid move justifications from other agents. State your final answer in a newline with a 2 letter response following the regex <code>[a-h][1-8]</code> .

Figure 15: **Prompts in each task.** List of prompts used in each task

Chess Validity. To evaluate chess validity, we consider the BIG-Bench Chess-State Tracking Benchmark [27], where we used the hardest reported task in the benchmark `synthetic_short`. Each generated answer was deemed correct as long as it was one of the valid answers in the sequence. We evaluated models of one hundred selected chess validity tasks.

User Input: Can you solve the following math problem? John arm wrestles 20 people. He beats 80%. How many people did he lose to? Explain your reasoning. Your final answer should be a single numerical number, in the form `\boxed{answer}`, at the end of your response.

Round 1	
<p>Agent 1: If John beat 80% of the people he arm wrestled, he must have lost to the remaining 20%. To find how many people that is, we can set up a proportion:</p> $\frac{\text{number of people John beat}}{\text{total number of people John arm wrestled}} = 80\% = \frac{4}{5}$ <p>Let x be the number of people John lost to. Then we know:</p> $\frac{\text{number of people John beat}}{\text{total number of people John arm wrestled}} = \frac{\text{number of people John beat}}{\text{number of people John beat} + \text{number of people John lost to}} = \frac{4}{5}$ <p>Cross-multiplying yields:</p> $5(\text{number of people John beat}) = 4(\text{number of people John beat} + \text{number of people John lost to})$ <p>Simplifying this equation gives:</p> $\text{number of people John lost to} = \boxed{25}$ <p>Therefore, John must have lost to 25 people. ❌</p>	<p>Agent 2: If John beats 80% of the people he arm wrestles, then he must lose to 20% of them. Since 20% of 20 people is $(0.20)(20) = 4$, John must have lost to $\boxed{4}$ people. ✅</p>

User Input: These are the solutions to the problem from other agents: One agent solution: (Copy Agent 2's reply) (Copy Agent 1's reply)
Using the solutions from other agents as additional information, can you provide your answer to the math problem?
Please update and respond to other agents. Your final answer should be a single numerical number, in the form `\boxed{answer}`, at the end of your response.

Round 2	
<p>Agent 1: Both solutions are correct in understanding that since John beat 80% of the people he arm wrestled, he lost to the remaining 20%. However, the first solution made an error in the calculation, where they found that John won against 16 people out of 20 (80% of 20), leading them to think that he lost to 4.</p> <p>Since John arm-wrestled 20 people in total, and won against 80% of them, he must have lost to the remaining 20%. Therefore, the number of people John lost to is:</p> $\text{Number of people John lost to} = 20 \cdot 0.2 = \boxed{4}$ <p>So the correct answer is indeed 4, as the second agent already provided. ❌</p>	<p>Agent 2: I agree with the other agents that John lost to 20% of the people he arm wrestled, which is equal to 4 people since he arm wrestled a total of 20 people. Therefore the answer is $\boxed{4}$. ✅</p>

User Input: These are the solutions to the problem from other agents: One agent solution: (Copy Agent 2's reply) (Copy Agent 1's reply)
Using the solutions from other agents as additional information, can you provide your answer to the math problem?
Please update and respond to other agents. Your final answer should be a single numerical number, in the form `\boxed{answer}`, at the end of your response.

Round 3	
<p>Agent 1: I agree with both previous agents that since John beat 80% of the people he arm wrestled, he lost to the remaining 20%. Therefore, the number of people John lost to is:</p> $\text{Number of people John lost to} = 20 \cdot 0.2 = \boxed{4}$ <p>So the answer is indeed 4, as the first agent incorrectly calculated but the second agent correctly calculated. ✅</p>	<p>Agent 2: I agree with the other agents that John lost to 20% of the people he arm wrestled, which is equal to 4 people since he arm wrestled a total of 20 people. Therefore the answer is $\boxed{4}$. ✅</p>

Figure 16: Example of a correct GSM8K Debate.