

# CRAwDAD: Causal Reasoning Augmentation with Dual-Agent Debate

Finn G. Vamosi

Department of Computer Science, University of Calgary  
 Calgary, Canada  
 finn.vamosi@ucalgary.ca

Nils D. Forkert

Department of Radiology, Hotchkiss Brain Institute, and  
 Alberta Children’s Hospital Research Institute, University  
 of Calgary  
 Calgary, Canada  
 nils.forkert@ucalgary.ca


## ABSTRACT

When people reason about cause and effect, they often consider many competing “what if” scenarios before deciding which explanation fits best. Analogously, advanced language models capable of causal inference can consider multiple interventions and counterfactuals to judge the validity of causal claims. Crucially, this type of reasoning is less like a single calculation and more like an internal dialogue between alternative hypotheses. In this paper, we make this dialogue explicit through a dual-agent debate framework where one model provides a structured causal inference, and the other critically examines this reasoning for logical flaws. When disagreements arise, the agents attempt to persuade each other, challenging each other’s logic and revising their conclusions until they converge on a mutually agreed answer. To take advantage of this deliberative process, we specifically use reasoning language models, whose strengths in both causal inference and adversarial debate remain under-explored relative to standard large language models, which do not employ multi-step logical inference to the same degree. We evaluate our approach on the CLadder dataset, a benchmark linking natural language questions to formally defined causal graphs across all three rungs of Pearl’s ladder of causation. With Qwen3 and DeepSeek-R1 as debater agents, we demonstrate that multi-agent debate improves DeepSeek-R1’s overall accuracy in causal inference from 78.03% to 87.45%, with the counterfactual category specifically improving from 67.94% to 80.04% accuracy. Similarly, Qwen3’s overall accuracy improves from 84.16% to 89.41%, and counterfactual questions from 71.53% to 80.35%, showing that even stronger models can still benefit greatly from debate with weaker agents. Our results highlight the potential of reasoning models as building blocks for multi-agent systems in causal inference, and demonstrate the importance of diverse perspectives in causal problem-solving. Our code is available at <https://github.com/finnvamosi/CRAwDAD>.

## KEYWORDS

Multi-Agent Debate; Causal Inference; Cooperation; Large Language Models; Reasoning Models

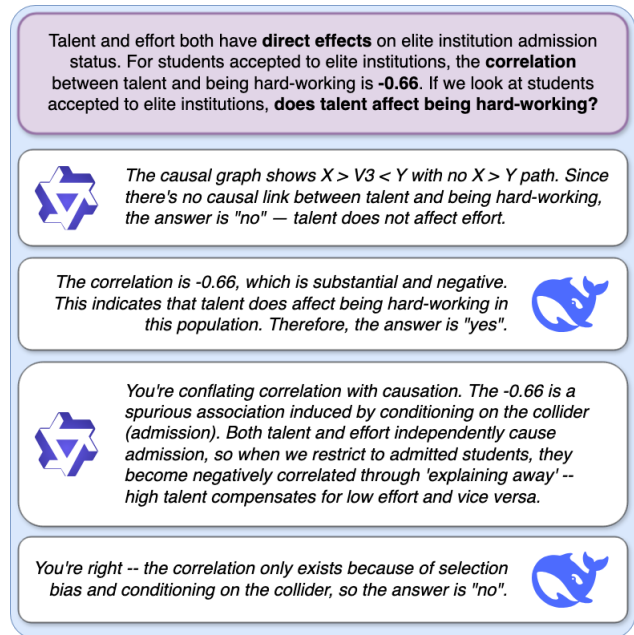
For appendices, see the extended version at <https://arxiv.org/abs/2511.22854>.

 This work is licensed under a Creative Commons Attribution International 4.0 License.


*Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems ([www.ifaamas.org](http://www.ifaamas.org)). <https://doi.org/10.65109/DVBN4652>


## ACM Reference Format:


Finn G. Vamosi and Nils D. Forkert. 2026. CRAwDAD: Causal Reasoning Augmentation with Dual-Agent Debate. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, Paphos, Cyprus, May 25 – 29, 2026, IFAAMAS, 10 pages. <https://doi.org/10.65109/DVBN4652>




Talent and effort both have **direct effects** on elite institution admission status. For students accepted to elite institutions, the **correlation** between talent and being hard-working is **-0.66**. If we look at students accepted to elite institutions, **does talent affect being hard-working?**

 The causal graph shows  $X > V3 < Y$  with no  $X > Y$  path. Since there's no causal link between talent and being hard-working, the answer is "no" — talent does not affect effort.

The correlation is **-0.66**, which is substantial and negative. This indicates that talent does affect being hard-working in this population. Therefore, the answer is "yes". 

 You're conflating correlation with causation. The **-0.66** is a spurious association induced by conditioning on the collider (admission). Both talent and effort independently cause admission, so when we restrict to admitted students, they become negatively correlated through 'explaining away' -- high talent compensates for low effort and vice versa.

You're right -- the correlation only exists because of selection bias and conditioning on the collider, so the answer is "no". 

**Figure 1: Example of a causal inference debate between two reasoning models. Qwen3 is randomly selected as the first speaker and successfully persuades DeepSeek-R1 to revise its initial conclusion, converging on the correct answer (“no”).**

## 1 INTRODUCTION

Causal reasoning is one of *Homo sapiens*’ greatest evolutionary advantages, enabling humans to think rationally about how our actions and the actions of others affect the world around us [13, 36]. From analyzing failed prehistoric hunts to designing scientific experiments, humans naturally ask “What if things had been different?”, and imagine counterfactual scenarios according to their

causal model of the scenario. Often, this reasoning occurs collectively through dialogue, as people challenge assumptions and debate alternative explanations. This capacity for counterfactual thinking has been formalized in statistics and artificial intelligence as "causal inference", which studies how interventions and hypothetical changes influence observed outcomes [33, 34]. Because this process depends on manipulating explicit causal models [32] rather than linguistic correlations, what is intuitive for humans remains challenging for most Large Language Models (LLMs) [16]. Newly developed Reasoning Language Models (RLMs) [9, 39], however, are designed for step-by-step, methodical problem-solving, which makes them promising candidates for testing whether stronger internal reasoning can support causal inference.

This deliberative nature of RLMs also suggests that they may be especially well-suited to collaborative dialogue between competing hypotheses, but recent surveys of the literature revealed that no prior work has explored whether multi-agent debate (MAD) can improve causal inference skills in either LLMs or RLMs [23, 29, 44, 45, 50]. In a MAD paradigm, multiple agents engage in structured discussion, proposing solutions, critiquing each other's arguments, and refining their reasoning in response to feedback. This approach has demonstrated significant improvements over single-agent performance across many non-causal domains [3, 5, 6, 10, 12, 22, 25, 27, 48, 51, 58]. While debate seems naturally suited to the hypothesis-testing inherent in causal inference, it is crucial to evaluate on a rigorous dataset to properly assess any improvements.

Jin et al. [16] argued that previous datasets [56, 57] for evaluating commonsense causality understanding in language models lacked formal construction rules, making it impossible to determine whether models were performing true causal reasoning or taking a shortcut known as "amortized causal reasoning" – essentially acting as "causal parrots" that repeat training data patterns [54]. To address this lack, Jin et al. constructed the CLadder dataset, which ties each natural language question to a formally defined causal model, with correct answers computed by an oracle inspired by Pearl's causal inference engine [35]. Some of those questions intentionally contradict real-world knowledge, while others use randomly generated variables, forcing language models to rely on formal reasoning rather than memorization. In their evaluations, Jin et al. found that ChatGPT-4 [30] achieved 70.40% overall accuracy using chain-of-thought (CoT) prompting [49], but only 62.05% for counterfactual questions.

Since CLadder's release in 2023, the emergence of RLMs offers new opportunities to revisit these benchmarks. Unlike traditional LLMs that generate responses in a single forward pass, RLMs are trained using reinforcement learning with human feedback (RLHF) to engage in extended deliberation before producing final answers [9, 39]. This training paradigm explicitly rewards models for reasoning steps rather than immediate conclusions, encouraging them to explore multiple solution paths, backtrack when encountering contradictions, and refine their thinking iteratively [19]. Such behaviours complement both causal inference and MAD: by supplying more coherent and self-consistent reasoning chains, RLMs provide a stronger foundation for agents to articulate, evaluate, and refine competing causal hypotheses. This, in turn, increases the likelihood that MAD between RLMs will converge on more accurate and interpretable conclusions.

To the best of our knowledge, there has only been one prior attempt so far to evaluate and compare the formal causal inference capabilities of RLMs in a single-agent setting [53]. More precisely, DeepSeek-R1 and OpenAI RLMs were included alongside a selection of standard LLMs in evaluations on CLadder. However, due to "cost and time constraints", the analysis and comparison of the RLMs investigated was restricted to only 100 random samples from the total of 10,112 questions available in the dataset. The RLM performance results on these 100 samples are considerably higher than the performance of ChatGPT-4 [30], with DeepSeek-R1 [9], OpenAI o1 [31], and OpenAI o3-mini achieving an overall accuracy of 92.0%, 89.0%, and 92.2%, respectively. However, because the CLadder dataset is stratified into three categories with vastly different difficulties, the overall accuracy can be a misleading metric without providing detailed information about the composition of samples. Therefore, further study is required to rigorously evaluate RLMs on the full dataset, both in single-agent settings and in MAD frameworks.

In this paper, we make the dialogue between hypotheses during causal inference explicit through the implementation of debates between RLMs, and apply this framework to the full CLadder dataset. To encourage more direct, adversarial engagement between models, we adopt a dual-agent approach, and select the most advanced open-source RLMs available: DeepSeek-R1 [9] and Qwen3 [39]. Our results demonstrate that collaborative reasoning through debate substantially improves performance on causal inference tasks, with the most significant gains observed on the most challenging counterfactual queries. By conducting the first comprehensive evaluation of RLMs on the full CLadder dataset and establishing strong baselines for both single-agent and multi-agent performance, we provide a benchmark for future work at the intersection of causal inference and collaborative AI systems.

Thus, the main contributions of this paper can be summarized as follows:

- (1) We performed an in-depth analysis and comparison of the causal inference capabilities of two open-source RLMs on the full CLadder dataset, categorizing performance by Rung.
- (2) We implemented a novel dual-agent debate framework that encourages more direct engagement between RLMs, and includes confidence estimation, an early-stopping strategy, and refined prompting.
- (3) We evaluate key outcomes across all debates to explore whether dual-agent debate improves final answer accuracy across all categories of the CLadder dataset.

## 2 RELATED WORK

There have been many studies to date investigating the single-agent performance of LLMs in causal tasks [1, 7, 15, 17, 20, 50, 60]. To our knowledge, however, the only prior work applying multi-agent debate to formal causality does so for the sake of causal discovery, not causal inference [21]. However, the authors acknowledge that their method is limited to observational data and cannot reason about interventions, restricting its applicability in practice. Moreover, recent critiques have questioned the use of standard LLMs specifically for causal discovery, noting that their correlation-driven training and optimization makes them unreliable for determining causal

graph structures on their own [52]. In contrast, our work focuses on causal inference, a distinct problem where the task is to answer interventional and counterfactual queries based on a known causal graph.

A related line of work applies multi-agent debate in a causality-adjacent context, focusing on evaluating causal consistency in reasoning chains generated by standard LLMs [43]. In this framework, evaluator agents judge the logical coherence of arguments from multiple reasoners. Importantly, these agents are not RLMs, and the evaluation was conducted on commonsense and knowledge-intensive benchmarks, which do not require formal causal inference over interventions or counterfactuals, as is the focus of our work.

Furthermore, both of these papers use the same base model for all agents, which may lead to shared biases in the dialogue [21, 43]. To overcome this limitation in our work, we took inspiration from recent studies that demonstrated the significant benefits that can be achieved through debate using diverse base models, as opposed to relying on a debate of the same base models (such as GPT-4) for all agents [4, 47].

Most MAD approaches described so far require debates to go through predetermined number of discussion rounds before reaching a conclusion. However, prior work has shown that by initiating a debate only when the confidence of the first speaker is below a certain threshold, a similar final answer accuracy can be achieved with drastically fewer model calls [11]. In our work, we adapt this idea, but use a novel setup that always elicits at least one response from each agent, and only initiates a debate if they disagree on the answer. This setup forms a key component of our methodology, which we detail in the following sections.

### 3 MATERIAL AND METHODS

#### 3.1 Dataset

The CLadder dataset that we use for evaluation in this work consists of 10,112 questions, each asking a formal causal inference question that requires careful reasoning, arithmetic, and knowledge of causality concepts and formulas in order to answer them correctly. The questions in this database are primarily split into three Rungs of causation [35], which increase monotonically in difficulty. There are 3,792 samples that fall in Rung 3, while the Rung 1 and 2 categories each consists of 3,160 samples.

Rung 1 ("seeing") questions ask about statistical associations between variables, such as "is there a positive relationship between student coffee consumption and test scores?". Rung 2 ("doing") questions require active intervention in the world, for example, asking "if students are made to drink coffee before their exam, will their scores improve?". Lastly, Rung 3 ("imagining") questions are about counterfactuals, which requires reasoning about alternatives scenarios that contradict reality. In other words, they ask about ways in which the world could have been different. A question of this kind could be: "Bob got a 85% on his test after drinking coffee. Would his score have been lower if he had not drank coffee?".

The secondary way the data is split is through three alignments to commonsense. The first commonsense alignment is that of normal, empirical questions that were sourced from the causality literature. In the second commonsense alignment, one variable is randomly replaced with a concept that has no correlation to the

other causal variables in the statement: for example, this may produce prompts stating that "ear shape affects lung tar deposits" or "full moon affects vaccination rate". Lastly, in the third commonsense alignment category, all causal variables are replaced with random nonsensical words: for instance, the model may be told that "zory affects gwet".

#### 3.2 Debate Setup

As described above, the most common MAD configurations described in the literature consist of three debater models, or two debater models and one judge model [23, 45]. However, because we specifically wanted to explore the extent to which RLMs are able to persuade each other to change their answer, we decided to only include two agents that are directly engaging with each other. This setup allows for more adversarial interactions between the agents and more specific language addressing faults in logic or calculation. While the use of a judge is not incompatible with this goal, recent studies have suggested that language models express several biases when assigned as a judge [25, 40], such as a tendency to favour the final speaker [28]. Furthermore, it was recently demonstrated that forcing models to defend assigned stances may degrade performance, while allowing them to change their opinions in response to opposing views may lead to better outcomes, without requiring a judge [18]. Therefore, we elected to not include a judge agent, allowing the agents to debate directly and change their answers if persuaded by the other model. This paradigm also obviates the computational overhead introduced through a third agent, and circumvents the need to define which model must act as judge.

With the content of the dataset and the general structure of our debate framework established, we now describe the specific details of our implementation. Naturally, each question spurs its own debate, and the first debate round must begin by randomly selecting one model as first speaker, in an attempt to more fairly assess the performance of each RLM. The debate format is not explained to the model at this point in order to keep the prompt setup as similar as possible to that used in Jin et al.'s original paper [16]. Their prompting strategy, which they refer to as CausalCoT, provides details of every step of the formal causal inference the RLM must follow in its final answer, including formalizing the causal graph and identifying the query type out of a list of options. After these instructions are provided, the RLM is provided with the background information (such as causal relationships) and probabilities that are associated with the question in the CLadder dataset. The model is also instructed to give a final yes/no answer, and, in a divergence from CausalCoT, asked to produce an estimation of their confidence, as is done in several other MAD approaches [2, 26]. Furthermore, we instruct this first model to be more self-confident and accept that answers may go against their knowledge of commonsense due to the nature of the dataset. This setup was motivated by preliminary experiments, in which we found that models had a tendency to over-think the questions and doubt their answers, which led to substantially longer response times.

Once this full prompt (shown in Appendix B) is received, the first RLM begins following the steps of causal inference, parsing the data provided in the question, materializing the formula from its

pre-trained knowledge base, and executing the needed arithmetic and causal do-calculus [41]. Most of this deliberative process happens in an internal "reasoning trace", but because of its verbosity and inclusion of discarded information, we only send the agent's concise rationale and final answer to their opponent. However, we still saved this reasoning trace for later manual analysis. With the first speaker's response complete, the second round begins. To encourage more direct engagement with the first model's work, the opposing model is not guided to follow all the steps of formal inference again. Instead, it is presented information about the format of the debate and the new expectations that come with it, namely to scrutinize the opponent's response carefully to find any flaws in reasoning, and form a suitable rebuttal to their response (full prompt shown in Appendix C). However, we do not force the second model to find a flaw in the first model's response, if none exists. Thus, if the two models agree with each other, then that answer will be accepted as final.

However, if the second model disagrees with the answer of the first model, round 3 begins, and the critique is passed back to the first model. We do not force the first model to adhere to its previous stance as it is simply instructed to consider the critique, and may either defend its initial answer or revise it based on the opponent's argument. If the first model rejects the critique and defends their original answer, then this refutation is sent back to the second model for round 4, whose subsequent response will be logged as the final answer to the question, even if the models are still in disagreement. Two examples of a debate between agents going to four rounds are provided in Figure 1 and Appendix D. In initial experiments, we found that additional rounds lead to little benefit, since when both models have made their case and remain steadfast in their answers, it is unlikely that further discussion changes their response. This is in line with previous research on standard LLMs that showed that continued discussion past a certain threshold leads to a stagnation or even degradation of answer accuracy [3].

### 3.3 Models

In accordance with principles of open science and reproducibility, we restricted our evaluation to open-source RLMs with publicly available weights, excluding proprietary models from OpenAI, Anthropic (Claude), Google (Gemini), and similar providers. At the time of implementation, there were only three RLMs available on the free Ollama platform: Qwen3, DeepSeek-R1, and Magistral. Due to hardware limitations, we chose Qwen3-32B and DeepSeek-R1-Distill-Qwen-32B for this work, which provided the optimal balance of reasoning capabilities and VRAM requirements. Preliminary tests confirmed that Magistral-24B exhibited substantially lower performance than the two 32B models, further supporting this selection. It is important to note that the DeepSeek-R1-Qwen-32B model used is derived from a Qwen2.5 base model, which was fine-tuned on the reasoning outputs of the DeepSeek-R1-671B model. However, Qwen3 was trained on twice as many tokens as Qwen2.5, and underwent several advancements in model architecture, which should introduce substantial heterogeneity between the two models. As established in previous works [4, 47], such model diversity is critical for effective debate, and we believe this is especially true when addressing causal inference, as it increases the likelihood that agents

will propose competing causal hypotheses rather than reinforcing shared biases.

In preliminary tests, we observed that the two RLMs did not respond well to requests for structured outputs. More precisely, they either neglect that part of the prompt, or bypass the thinking stage entirely if Ollama's structured output functionality is instated. This finding is in line with recent research that found that reasoning models are worse at following instructions than standard LLMs [24, 42]. In response, we added a Granite3.3-2B model to our pipeline to extract final yes/no answers and confidence estimations from the primary RLMs responses. In a manual inspection of 50 randomly sampled debate outputs, the Granite model was always able to faithfully extract the appropriate outcomes.

### 3.4 Experiments

Our primary evaluation compares model accuracy before and after debate, with accuracy measured as the proportion of RLM responses matching ground truth labels in CLadder. To provide a more comprehensive assessment of debate efficacy beyond accuracy alone, we performed additional analyses on the full set of model outputs as discussed below.

**Persuasion dynamics.** To assess whether debate improves or corrupts initial responses, we tracked answer changes from the first model's initial response to the final consensus answer across all questions requiring debate (3+ rounds). For each model, we classified outcomes into four categories: correct-to-correct (maintained accuracy), incorrect-to-correct (persuaded toward truth), correct-to-incorrect (corrupted by peer), and incorrect-to-incorrect (failed to improve). We stratified these results by both model and Rung to identify whether persuasion patterns differ by model capability or question difficulty, revealing which model is more influential in the debate process. Next, we conducted a sentiment analysis on model responses to determine the average positivity or negativity associated with the language they use; both when formulating an initial answer, and when debating with their peer. To do so, we removed all formatting from model outputs and passed them through the SentimentIntensityAnalyzer of the popular VADER library [14].

**Confidence calibration.** For each model, we compared confidence distributions for correct versus incorrect answers, both in initial responses and after being challenged by their peer. This analysis addresses three questions: whether confidence aligns with correctness (calibration), how question difficulty (Rung) affects this alignment, and whether facing a highly confident opponent increases or decreases the likelihood of changing one's answer.

**Response characteristics.** We measured response length distributions for both initial answers and debate rounds to assess engagement depth and identify potential asymmetries in argumentative efforts between models. Finally, we visualized the distribution of rounds required to reach consensus across difficulty levels (Rungs), as well as the consensus rate by the final round, providing insight into debate efficiency and the extent to which question complexity affects deliberation time.

## 4 RESULTS

The results shown in Table 1 demonstrate that debate substantially improves performance for both RLMs, with DeepSeek-R1 and

**Table 1: Performance of all models on the CLadder dataset. We report the overall accuracy (Acc.) and fine-grained accuracy by Rung and degree of commonsense alignment, from commonsensical (Comm.), nonsensical (Nonsens.), to anti-commonsential (Anti-C.).**

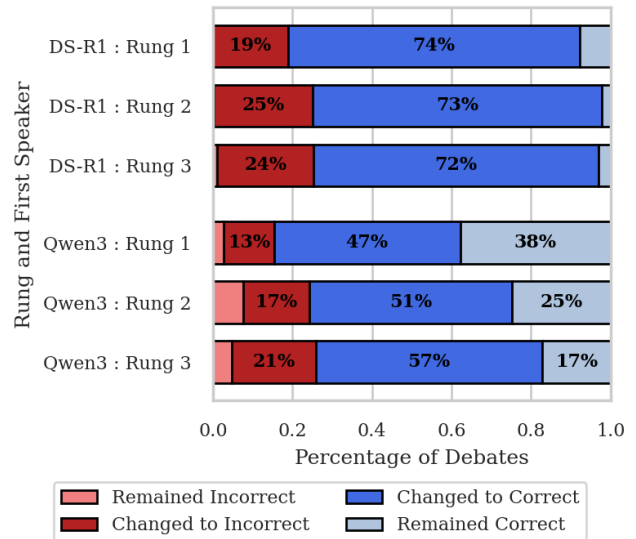
	Overall Acc.	Acc. by Rung			Acc. by Commonsense Alignment		
		1	2	3	Comm.	Nonsens.	Anti-C.
GPT-4 [16]	62.03	63.01	62.82	60.55	62.27	63.09	60.47
+ CausalCoT [16]	70.40	83.35	67.47	62.05	69.25	71.58	70.12
Qwen3 (Initial)	84.16	93.77	89.80	71.53	84.96	83.10	84.69
Qwen3 (Final)	<b>89.41</b>	<b>96.24</b>	<b>93.51</b>	<b>80.35</b>	<b>89.52</b>	<b>87.91</b>	<b>91.20</b>
DeepSeek-R1 (Initial)	78.03	90.67	77.29	67.94	77.63	77.24	79.42
DeepSeek-R1 (Final)	87.45	94.62	89.06	80.04	87.83	86.94	87.67
Human [53] <sup>1</sup>	94.8	-	-	-	-	-	-

Qwen3 gaining 9.42% and 5.25% accuracy, respectively. Post-debate, Qwen3 achieved the highest accuracy across all question types, followed closely by DeepSeek-R1. Generally, Qwen3 was able to maintain a higher score across questions in Rung 1 and 2 relative to the CausalCoT baseline results. However, Rung 3 questions were associated with a significant drop in accuracy compared to the other two Rungs, but these questions also benefited the most from the debate in both model: Qwen3 and DeepSeek-R1 gained 8.82% and 12.10% accuracy, respectively.

While we primarily use each model’s pre-debate accuracy as a baseline, we also conducted a smaller benchmark evaluation by querying each model with 500 commonsense queries (250 from Rung-1 and 250 from Rung-3) according to Yu et al.’s prompt [53], without allowing them to debate with their peer. Briefly described, the results for this secondary analysis were largely consistent with the initial accuracies shown in Figure 1: on Rung-1, Qwen achieved 89.20% (223/250) and DeepSeek 90.00% (225/250), while on Rung-3, Qwen achieved 72.40% (181/250) and DeepSeek 67.60% (169/250). Most of these differences are statistically indistinguishable from our reported initial accuracies. However, Qwen’s 4.6% reduction in accuracy when using Yu et al.’s prompt is significant (p-value = 0.007).

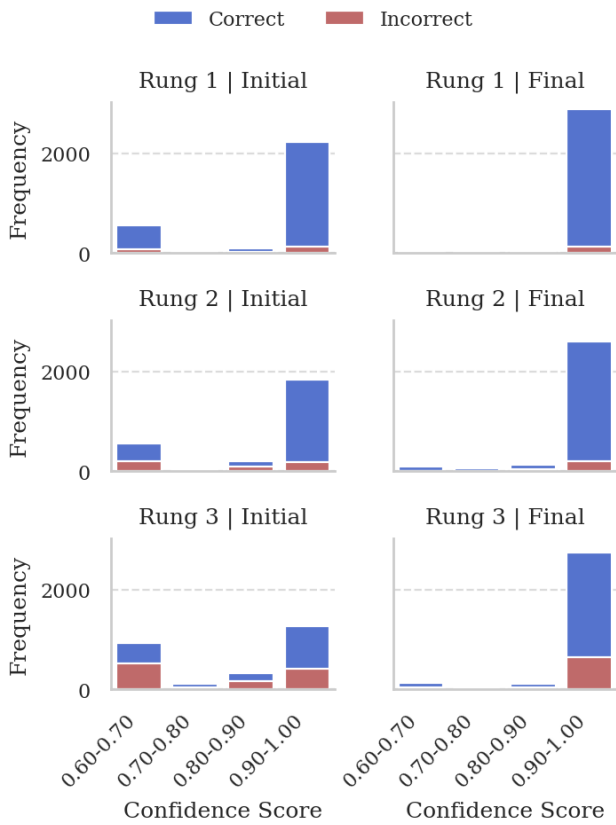
### 4.1 Persuasion

Figure 2 illustrates the net positive effect that the debate framework has on final answer accuracy, when models disagree on the initial answer. Briefly described, the results suggest that it is far more likely for the debate to change an incorrect answer to a correct one, than vice versa. The results also show that DeepSeek-R1 is more susceptible to persuasion, as it is more likely to accept Qwen3’s answer when challenged. Nonetheless, Qwen3 still benefits greatly from debating with DeepSeek-R1, correcting many of its initially wrong responses while doing a much better job of defending its correct answers. This helps to explain the finding illustrated in Table 1 showing that the DeepSeek-R1 model had lower initial scores than Qwen3, while also gaining more benefit from the debate process. In other words, the Qwen3 model seems better at both reasoning and persuasion.



**Figure 2: For questions with initial disagreement, debate improves answers far more often than it worsens them.**

In a qualitative analysis of selected debates that ended with no consensus after four rounds, we saw that Qwen3’s arguments tend to be longer, address the opponent’s response much more directly, and have a more active voice. In contrast, DeepSeek-R1 provides shorter answers (often as short as “yes” or “no”) that have a tendency to remain neutral, and in some cases almost seeming entirely detached from the conversation. Generally, to a non-expert human, the Qwen3 model answers appear more correct and persuasive because of the way it presents its ideas, even when it is wrong. To explore this in more detail, we conducted a simple sentiment analysis of all agent responses using VADER [14] and found that while both models display a neutral sentiment for initial responses, the Qwen3 model becomes more negative in the subsequent debate, with a sentiment value of -0.239, compared to DeepSeek-R1’s rather neutral sentiment value of 0.015. This is to be expected, given the



**Figure 3: Initial confidence of both models combined, for each Rung. The models become less confident as Rungs get more complicated, and generally are just as confident in their incorrect answers as they are in their correct answers.**

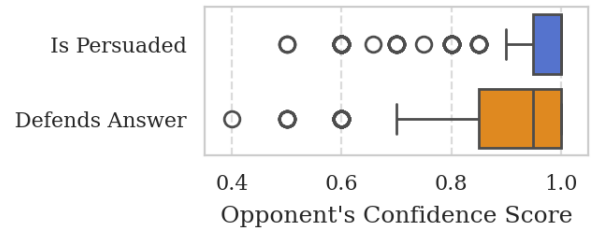
more direct language Qwen3 uses as it criticizes DeepSeek-R1’s response that will naturally tend toward the negative, even without being expressly “mean”.

### 4.2 Confidence

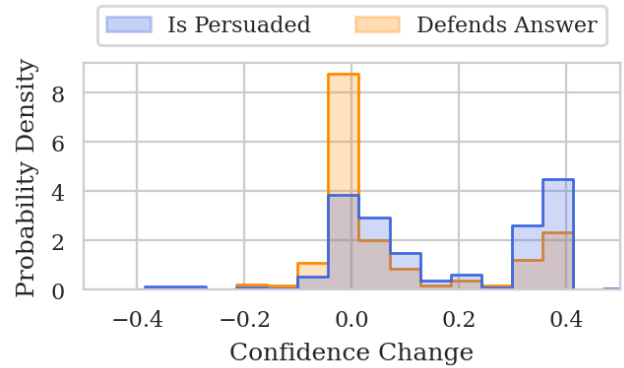
Figure 3 shows that both models are often extremely confident, with confidence never falling below 60% for either model. This is reasonable given our explicit instruction to be confident, but we also see that the more difficult the Rung is, the less confident the models tend to be initially, which suggests that they do not suppress their uncertainty entirely. There seems to be no evident difference between the commonsense alignment of the question and a model’s confidence (*i.e.*, anti-commonsense questions do not make them more cautious). It is also notable that both models seem reluctant to output a confidence in the 65-80% range, and have a similar confidence distribution, although Qwen3 tends to be slightly more confident.

The second column of Figure 3 demonstrates the drastic increase in confidence in the final answer agreed upon by both models,

<sup>1</sup>Prior work obtained this human benchmark by asking three trained graduate students to solve 100 randomly sampled questions. [53]



**Figure 4: Models are more likely to be persuaded to change their answer when their opponent is more confident.**



**Figure 5: When defending their answer, models do not express uncertainty after facing criticism. However, if persuaded to change their answer, they often become more confident in their opponent’s answer.**

which combines cases of initial agreement and disagreement. In cases where the second model immediately agrees with the first model, the second model is very confident in the answer, even if the first model was not. When models initially disagree, their debate provides a substantial confidence boost: Figure 5 shows that when a model is persuaded by their opponent, their confidence is likely to improve considerably, and when a model defends its previous answer, their confidence does not decrease. These patterns are in line with previous studies showing the systematic overconfidence of LLMs and RLMs [37], even without instructions to be more confident. Moreover, Figure 4 shows that both models tend to be more persuaded by more confident arguments, although it is unclear whether more confident critiques are truly more compelling, or if models are responding directly to the confidence score.

In our analysis of initial confidence, we found 577 cases for which the first speaker gave an incorrect answer with high confidence (>90%), but subsequent debate with their peer corrected them. Generally, corrections of confident answers were more prevalent in higher, more difficult Rungs. In addition, we found that the Qwen3 model was more capable in this task, while DeepSeek-R1 had lower accuracy, and thus benefited more often from having a peer correct its confidently wrong answer.

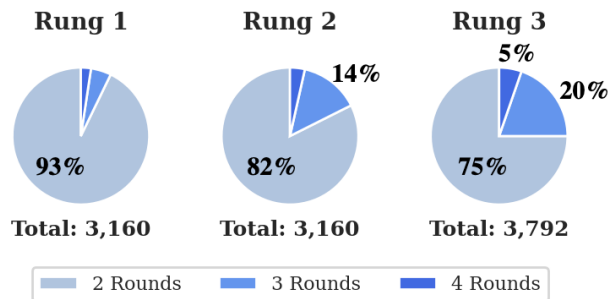


Figure 6: Number of rounds taken for the debate to conclude increases monotonically as Rungs progress.

### 4.3 Rounds Taken

Further analysis, shown in Figure 6, revealed that the harder Rungs demanded more debate rounds on average than the first Rung, where in 93% of all cases both models agreed upon an answer immediately. In contrast, approximately a quarter of Rung 3 questions were disputed, although most of these disputes were settled in the third round, with the first speaker convinced by the opponent’s argument. Interestingly, it was found that Rung 2 questions were resolved more quickly when Qwen3 was the first speaker compared to DeepSeek-R1 being the first model. However, this is the only Rung where such a correlation between the first speaker and rounds taken was found. Out of the 10112 questions in the dataset, only 148 ended without the models reaching consensus on an answer. Therefore, even if both models could agree on the correct answer in these cases (e.g., through more discussion, or a better tie-breaking strategy), the total accuracy would only increase by 0.55%.

### 4.4 Response Length

The results of the evaluation showed that the DeepSeek-R1 model has a tendency to provide extremely short responses during debate. More precisely, we observed that 22.60% (1412/6247) of DeepSeek-R1’s debate responses were 100 characters or less, compared to 0.40% (11/5986) for Qwen3. As illustrated in Figure 7a, the DeepSeek-R1 model has much shorter responses on average, even when removing these very short responses. More precisely, DeepSeek-R1 provided responses with a median of 249 characters compared to Qwen3’s 739 characters. Figure 7b shows that DeepSeek-R1’s initial responses do not produce such short outputs, although they are still shorter on average, with a median of 735 characters compared to Qwen3’s 1012. Interestingly, we found that the underlying reasoning traces of both models actually have very similar distributions of character counts (Figure 7c). The manual inspection of a few selected reasoning traces showed that DeepSeek-R1 is indeed considering both the problem and its opponent’s response, but inexplicably ignores the prompt’s instructions when generating its final response, providing no external elaboration of its reasoning. This likely contributes to why DeepSeek-R1 is worse at persuading Qwen3 than vice versa, even if it has the right answer. This may also be the reason for the reduced efficacy of the debate as a whole, as DeepSeek-R1 is providing no substantive content or argument in these specific cases.

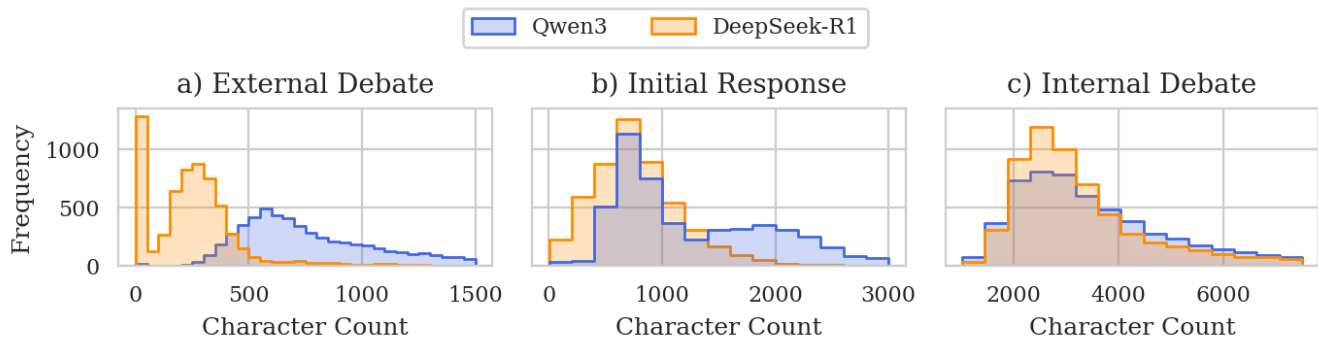
## 5 DISCUSSION

Our results demonstrate that debate between RLMs can considerably increase the accuracy of their final answer, with the largest gains being associated with the hardest problems. Based on our results, it may be concluded that Qwen3 is generally better-suited for this causal inference task compared to DeepSeek-R1. However, it is interesting that arguing with DeepSeek-R1 still improves its answers, and in the case of Rung 3 questions, quite considerably. Equally intriguing is how the DeepSeek-R1 model’s accuracy does not improve to the performance levels of the Qwen3 model, despite the debate rounds being roughly equal. This finding suggests that the initial answer has an effect on the subsequent outcome, which would skew final answers in favour of the first given. Another potential reason could be that the models tested, similar to humans, may tend to glance over a proof and support it rather than trying to find a specific error in the labelling, logic, or arithmetic. Nevertheless, this demonstrates the benefit of the debate setup that even a weaker model is able to contribute to a stronger model’s improvement, and the stronger model can improve the weak one past its own initial performance.

While our results show substantial improvement of RLMs relative to the performance achieved by the trillion-parameter GPT-4 model as described in the original 2023 CLadder paper, and Rung 1 and 2 questions are largely solvable, it seems that open source reasoning models of this size still struggle with Rung 3 questions. Moreover, while the Qwen3 model was able to match human performance in Rung 1 queries after debating with DeepSeek-R1, it fell slightly short in Rung 2, even after debate, and very short in Rung 3, with the DeepSeek-R1 model performing even worse. While the human accuracy reported in prior work [53] is not split by Rung, it seems that RLMs still cannot match the human ability to reason about counterfactuals. If this is a limitation inherent to the correlation-based nature LLMs and RLMs, one possible solution would be the creation of an external tool that models can query for the precise calculation of counterfactual scenarios, such as the one recently implemented by Verma et al. [46]. While our results show that LLMs and RLMs cannot yet be trusted to answer counterfactual questions without the use of such a tool, we nonetheless demonstrate the potential of reasoning models as building blocks for multi-agent systems in causal inference, and highlight the importance of diverse perspectives in causal problem-solving.

The disparity of DeepSeek-R1’s accuracy between our study and Yu et al.’s is considerable. Their improved performance suggests that they either evaluate on the larger DeepSeek-R1-671B model, or that their random sampling primarily pulled from Rung 1 questions, where this accuracy would be on par with our evaluation of DeepSeek-R1’s accuracy before debating with Qwen3. However, our work shows that reasoning models still struggle with Rung 3 questions, even after debating with a peer.

The results of the confidence analysis show interesting alignments to human tendencies. Firstly, when a model agrees with the opposing model, it is much more likely to be confident in its final answer. Secondly, the models investigated seem to take the opposing model’s confidence into account when making a decision of whether or not to change their answer in response. The models’ aversion to estimating their confidence to be in the 65-80% range



**Figure 7:** a) DeepSeek-R1 has much shorter debate responses than Qwen3, and an exceptionally high number of responses that are less than 100 characters. b) The models do not differ considerably in the length of their initial responses when assigned as the first speaker. c) DeepSeek-R1’s underlying reasoning traces for its short debate responses have comparable character length to that of Qwen3’s.

might suggest a lack of these confidences in their training data, or a binary internal value for high/low confidence that is then translated to an arbitrary percentage. Relying solely on the initial confidence of the first speaker would have led to a decrease in accuracy of 5.7% had we not asked the second model for its answer. This illustrates the benefit of asking both models for at least one answer, which still drastically reduces the number of model calls compared to the conventional approach of predetermined debate length.

## 5.1 Limitations

Given CLadder’s creation date of 2023 [16], we cannot discount the possibility that DeepSeek-R1 or Qwen3 (which were both released in early 2025) were exposed to its contents during their training. While we considered generating an entirely new set of data using Cladder’s generation code, we were concerned that this might compromise the direct comparison to Jin et al.’s results in a different way if we generated easier or harder questions by chance.

Another possible limitation of our study is related to CLadder’s synthetic nature: its questions are simplified, clean, and textually uniform compared to real-world causal reasoning problems, which often involve messy natural language, ambiguous causal structures, or incomplete data. Without synthetic data, however, it would be much more difficult to establish precise ground truth and reliably measure model correctness at this granularity.

Next, combining the lengthy response times of RLMs with the added model calls inherent to MAD introduces considerable additional resource and time costs for our model, with the 10,112 questions of CLadder taking 380 hours in total to answer. Due to this time constraint, we were unable to conduct potentially informative ablation studies, including: the substitution of standard LLMs, using three or more RLMs, implementing a judge framework, omitting confidence estimation, and removing instructions to be more confident. However, using our framework, all of those can be investigated in future.

Furthermore, because of limited options among open-source models, we only experiment with one pair of agents. Thus, it is possible that different combinations of agents will yield different

results from debate, and, of course, different models will achieve different initial accuracies. In particular, our lack of evaluation on OpenAI’s models hinders direct comparison to other works that use a single-agent paradigm.

## 5.2 Future Work

Due to hardware limitations, 32B parameter models were the largest we could use for our experiments. Thus, future work should evaluate the larger models in the Qwen3 and DeepSeek-R1 families to investigate whether a scaling law exists between model size and formal causal inference capabilities, as previously found in several other domains [8, 38, 55, 59].

Future studies on the intersection of causal inference and multi-agent systems should also investigate different frameworks for debate or collaboration. Using three agents, or two agents and a judge, are both straightforward starting points, though more complex paradigms may lead to further gains.

## 6 CONCLUSION

In this paper, we demonstrated that debates between RLMs improve their collective ability to formally reason about causality in the CLadder dataset. We found that, even when applied independently, comparatively lightweight open-source reasoning models surpass GPT-4’s 70.40% accuracy baseline, and when encouraged to critique each other’s reasoning, the overall accuracy of Qwen3-32B rises to 89.41%, and DeepSeek-R1-Distill-Qwen-32B to 87.45%. These promising results encourage further research on the implementation of multi-agent frameworks for addressing causal reasoning tasks.

## REFERENCES

- [1] Abdolmahdi Bagheri, Matin Alinejad, Kevin Bello, and Alireza Akhondi-Asl. 2024.  $\text{C}^2\text{P}$ : Featuring Large Language Models with Causal Reasoning. <https://doi.org/10.48550/arXiv.2407.18069> arXiv:2407.18069 [cs].
- [2] Yilin Bai. 2024. ConfidenceCal: Enhancing LLMs Reliability through Confidence Calibration in Multi-Agent Debate. In *2024 10th International Conference on Big Data and Information Analytics (BigDIA)*. 221–226. <https://doi.org/10.1109/BigDIA63733.2024.10808396> ISSN: 2771-6902.

- [3] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. <https://doi.org/10.48550/arXiv.2308.07201> arXiv:2308.07201 [cs].
- [4] Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. <https://arxiv.org/abs/2309.13007v3>
- [5] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. AGENTVERSE: FACILITATING MULTI-AGENT COLLABORATION AND EXPLORING EMERGENT BEHAVIORS. (2024).
- [6] Jina Chun, Qihong Chen, Jiawei Li, and Iftekhar Ahmed. 2025. Is Multi-Agent Debate (MAD) the Silver Bullet? An Empirical Analysis of MAD in Code Summarization and Translation. <https://doi.org/10.48550/arXiv.2503.12029> arXiv:2503.12029 [cs].
- [7] Shaobo Cui, Zhijing Jin, Bernhard Schölkopf, and Boi Faltings. 2024. The Odyssey of Commonsense Causality: From Foundational Benchmarks to Cutting-Edge Reasoning. <https://doi.org/10.48550/arXiv.2406.19307> arXiv:2406.19307 [cs].
- [8] DeepSeek-AI. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. <https://doi.org/10.48550/arXiv.2401.02954> arXiv:2401.02954 [cs].
- [9] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. <https://doi.org/10.48550/arXiv.2501.12948> [cs].
- [10] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. <https://doi.org/10.48550/arXiv.2305.14325> arXiv:2305.14325 [cs].
- [11] Sugyeong Eo, Hyeonseok Moon, Evelyn Hayoon Zi, Chanjin Park, and Heuiseok Lim. 2025. Debate Only When Necessary: Adaptive Multiagent Collaboration for Efficient LLM Reasoning. <https://doi.org/10.48550/arXiv.2504.05047> arXiv:2504.05047 [cs].
- [12] Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahao Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-MAD: Multidimensional Multi-Agent Debate for Advanced Machine Translation Evaluation. <https://doi.org/10.48550/arXiv.2412.20127> [cs].
- [13] Yuval Noah Harari. 2015. *Sapiens: A Brief History of Humankind*. HarperCollins Publishers, New York, NY.
- [14] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 216–225. <https://doi.org/10.1609/icwsm.v8i1.14550>
- [15] Haitao Jiang, Lin Ge, Yuhe Gao, Jianan Wang, and Rui Song. 2024. LLM4Causal: Democratized Causal Tools for Everyone via Large Language Model. <https://doi.org/10.48550/arXiv.2312.17122> arXiv:2312.17122 [cs].
- [16] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2024. CLadder: Assessing Causal Reasoning in Language Models. <https://doi.org/10.48550/arXiv.2312.04350> arXiv:2312.04350 [cs].
- [17] Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. 2024. Efficient Causal Graph Discovery Using Large Language Models. <https://doi.org/10.48550/arXiv.2402.01207> arXiv:2402.01207 [cs].
- [18] Harvey Bonmu Ku, Jeongyeol Shin, Hyouun Jun Lee, Seonok Na, and Insu Jeon. 2025. Multi-Agent LLM Unveils the Premise Left Unsaid. In *Proceedings of the 12th Argument Mining Workshop*, Elena Chistova, Philipp Cimiano, Shohreh Haddadan, Gabriella Lapesa, and Ramon Ruiz-Dolz (Eds.). Association for Computational Linguistics, Vienna, Austria, 58–73. <https://doi.org/10.18653/v1/2025.argmining-1.6>
- [19] Komal Kumar, Tajamal Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Fahad Shahbaz Khan, and Salman Khan. 2025. LLM Post-Training: A Deep Dive into Reasoning Large Language Models. <https://doi.org/10.48550/arXiv.2502.21321> arXiv:2502.21321 [cs].
- [20] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. <https://doi.org/10.48550/arXiv.2305.00050> arXiv:2305.00050 [cs].
- [21] Hao Duong Le, Xin Xia, and Zhang Chen. 2025. Multi-Agent Causal Discovery Using Large Language Models. <https://doi.org/10.48550/arXiv.2407.15073> [cs].
- [22] Han Li, Yuling Shi, Shaoxin Lin, Xiaodong Gu, Heng Lian, Xin Wang, Yantao Jia, Tao Huang, and Qianxiang Wang. 2025. SWE-Debate: Competitive Multi-Agent Debate for Software Issue Resolution. <https://doi.org/10.48550/arXiv.2507.23348> arXiv:2507.23348 [cs].
- [23] Xin Li, Zhuo Cai, Shoujin Wang, Kun Yu, and Fang Chen. 2025. A Survey on Enhancing Causal Reasoning Ability of Large Language Models. <https://doi.org/10.48550/arXiv.2503.09326> arXiv:2503.09326 [cs].
- [24] Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. 2025. When Thinking Fails: The Pitfalls of Reasoning for Instruction-Following in LLMs. <https://doi.org/10.48550/arXiv.2505.11423> arXiv:2505.11423 [cs].
- [25] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. <https://doi.org/10.48550/arXiv.2305.19118> arXiv:2305.19118 [cs].
- [26] Zijie Lin and Bryan Hooi. 2025. Enhancing Multi-Agent Debate System Performance via Confidence Expression. <https://doi.org/10.48550/arXiv.2509.14034> [cs].
- [27] Shuai Ling, Lizi Liao, Dongmei Jiang, and Weili Guan. 2025. MeMAD: Structured Memory of Debates for Enhanced Multi-Agent Reasoning. (2025).
- [28] Xinyi Liu, Pinxin Liu, and Hangfeng He. 2024. An Empirical Analysis on Large Language Models in Debate Evaluation. <https://doi.org/10.48550/arXiv.2406.00050> arXiv:2406.00050 [cs].
- [29] Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, and Furong Huang. 2025. Large Language Models and Causal Inference in Collaboration: A Survey. <https://doi.org/10.48550/arXiv.2403.09606> arXiv:2403.09606 [cs].
- [30] OpenAI. 2024. GPT-4 Technical Report. <https://doi.org/10.48550/arXiv.2303.08774> [cs].
- [31] OpenAI. 2024. OpenAI o1 System Card. <https://doi.org/10.48550/arXiv.2412.16720> arXiv:2412.16720 [cs].
- [32] JUDEA PEARL. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (Dec. 1995), 669–688. <https://doi.org/10.1093/biomet/82.4.669>
- [33] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys* 3, none (Jan. 2009). <https://doi.org/10.1214/09-SS057>
- [34] Judea Pearl. 2022. *Causality: models, reasoning, and inference* (second edition, reprinted with corrections ed.). Cambridge University Press, Cambridge New York, NY Port Melbourne New Delhi Singapore.
- [35] Judea Pearl and Dana Mackenzie. 2020. *The book of why: the new science of cause and effect* (first trade paperback edition ed.). Basic Books, New York.
- [36] Derek C. Penn and Daniel J. Povinelli. 2007. Causal cognition in human and nonhuman animals: a comparative, critical review. *Annual Review of Psychology* 58 (2007), 97–118. <https://doi.org/10.1146/annurev.psych.58.110405.085555>
- [37] Pradyumna Shyama Prasad and Minh Nhat Nguyen. 2025. When Two LLMs Debate, Both Think They’ll Win. <https://doi.org/10.48550/arXiv.2505.19184> arXiv:2505.19184 [cs].
- [38] Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Wang, Zhiyuan Liu, and Maosong Sun. 2025. Scaling Large Language Model-based Multi-Agent Collaboration. <https://doi.org/10.48550/arXiv.2406.07155> arXiv:2406.07155 [cs].
- [39] Qwen Team. 2025. Qwen3 Technical Report. <https://doi.org/10.48550/arXiv.2505.09388> arXiv:2505.09388 [cs].
- [40] Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. <https://doi.org/10.48550/arXiv.2406.07791> arXiv:2406.07791 [cs].
- [41] Ilya Shpitser and Judea Pearl. 2012. Identification of Conditional Interventional Distributions. <https://doi.org/10.48550/arXiv.1206.6876> arXiv:1206.6876 [cs].
- [42] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models. <https://doi.org/10.48550/arXiv.2408.02442> arXiv:2408.02442 [cs].
- [43] Ziyi Tang, Rulin Wang, Weixing Chen, Yongsen Zheng, Zechuan Chen, Yang Liu, Keze Wang, Tianshui Chen, and Liang Lin. 2025. Towards CausalGPT: A Multi-Agent Approach for Faithful Knowledge Reasoning via Promoting Causal Consistency in LLMs. <https://doi.org/10.48550/arXiv.2308.11914> arXiv:2308.11914 [cs].
- [44] Arne Tillmann. 2025. Literature Review Of Multi-Agent Debate For Problem-Solving. <https://doi.org/10.48550/arXiv.2506.00066> arXiv:2506.00066 [cs].
- [45] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O’Sullivan, and Hoang D. Nguyen. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. <https://doi.org/10.48550/arXiv.2501.06322> arXiv:2501.06322 [cs].
- [46] Vishal Verma, Sawal Acharya, Samuel Simko, Devansh Bhardwaj, Anahita Haghighat, Dominik Janzing, Mrinmaya Sachan, Zhijing Jin, and Yongjin Yang. 2025. Causal AI Scientist: Facilitating Causal Data Science with Large Language Models. In *NeurIPS 2025 AI for Science Workshop*.
- [47] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-Agents Enhances Large Language Model Capabilities. <https://doi.org/10.48550/arXiv.2406.04692> arXiv:2406.04692 [cs].
- [48] Qingeng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key? <https://doi.org/10.48550/arXiv.2402.18272> arXiv:2402.18272 [cs].
- [49] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 24824–24837. <https://proceedings.neurips.cc/paper/2022/>

- hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html
- [50] Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. 2024. Causality for Large Language Models. <https://doi.org/10.48550/arXiv.2410.15319> arXiv:2410.15319 [cs].
- [51] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryan W. White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. <https://doi.org/10.48550/arXiv.2308.08155> arXiv:2308.08155 [cs].
- [52] Xingyu Wu, Kui Yu, Jibin Wu, and Kay Chen Tan. 2025. LLM Cannot Discover Causality, and Should Be Restricted to Non-Decisional Support in Causal Discovery. <https://doi.org/10.48550/arXiv.2506.00844> arXiv:2506.00844 [cs].
- [53] Longxuan Yu, Delin Chen, Siheng Xiong, Qingyang Wu, Qingzhen Liu, Dawei Li, Zhikai Chen, Xiaozhe Liu, and Liangming Pan. 2025. CausalEval: Towards Better Causal Reasoning in Language Models. <https://doi.org/10.48550/arXiv.2410.16676> arXiv:2410.16676 [cs].
- [54] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. <https://doi.org/10.48550/arXiv.2308.13067> arXiv:2308.13067 [cs].
- [55] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method. <https://doi.org/10.48550/arXiv.2402.17193> arXiv:2402.17193 [cs].
- [56] Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. 2023. Understanding Causality with Large Language Models: Feasibility and Opportunities. <https://doi.org/10.48550/arXiv.2304.05524> arXiv:2304.05524 [cs].
- [57] Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about Goals, Steps, and Temporal Ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 4630–4639. <https://doi.org/10.18653/v1/2020.emnlp-main.374>
- [58] Shaowei Zhang and Deyi Xiong. 2025. Debate4MATH: Multi-Agent Debate for Fine-Grained Reasoning in Math. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 16810–16824. <https://doi.org/10.18653/v1/2025.findings-acl.862>
- [59] Enbo Zhao, Yi Shen, Shuming Shi, Jieyun Huang, Zhihao Chen, Ning Wang, Siqi Xiao, Jian Zhang, Kai Wang, and Shiguo Lian. 2025. Quantitative Analysis of Performance Drop in DeepSeek Model Quantization. <https://doi.org/10.48550/arXiv.2505.02390> arXiv:2505.02390 [cs].
- [60] Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. 2024. CausalBench: A Comprehensive Benchmark for Causal Learning Capability of LLMs. <https://doi.org/10.48550/arXiv.2404.06349> arXiv:2404.06349 [cs].