
ACC-DEBATE: AN ACTOR-CRITIC APPROACH TO MULTI-AGENT DEBATE

Anonymous authors

Paper under double-blind review

ABSTRACT

Large language models (LLMs) have demonstrated a remarkable ability to serve as general-purpose tools for various language-based tasks. Recent works have demonstrated that the efficacy of such models can be improved through iterative dialog between multiple models, frequently referred to as multi-agent debate (MAD). While debate shows promise as a means of improving model efficacy, most works in this area treat debate as an emergent behavior, rather than a learned behavior. In doing so, current debate frameworks rely on collaborative behaviors to have been sufficiently trained into off-the-shelf models. To address this limitation, we propose ACC-Debate, an Actor-Critic based learning framework to produce a two-agent team specialized in debate. We demonstrate that ACC-Debate outperforms SotA debate techniques on a wide array of benchmarks.

1 INTRODUCTION

Recently, large language models (LLMs) have rapidly become a cornerstone in various applications, redefining how we process and generate language at scale (Thirunavukarasu et al., 2023; Hadi et al., 2023; Jiang et al., 2024). Their ability to handle diverse tasks, from translation (Zhu et al., 2024; Otter et al., 2020) to answering complex questions (Zhang et al., 2024; Hao et al., 2024; Havrilla et al., 2024), has attracted the attention of both industry as well as academia. However, despite these advancements, LLMs still exhibit notable weaknesses, particularly when it comes to answering factual questions and reasoning (Tonmoy et al., 2024; Rawte et al., 2023; Huang et al., 2023).

To address these limitations, several techniques have been proposed, such as Chain-of-Thought (CoT) prompting (Wei et al., 2022), self-reflection (Ji et al., 2023; Shinn et al., 2023), and multi-agent debate (MAD) (Du et al., 2023), to name a few. These approaches aim to improve the reasoning abilities of LLMs by guiding them toward more accurate answers through structured thinking or discourse. However, the majority of these techniques do not involve training the model specifically for these tasks but instead rely on zero-shot or few-shot capabilities.

In particular, multi-agent debate approaches make use of off-the-shelf general-purpose LLMs, which are not trained to collaborate. Such approaches rely on collaboration as an emergent, rather than a learned, behavior. While, in some cases, these emergent behaviors are sufficient, the question remains: Can these methods be improved by imbuing models directly with collaborative abilities? To answer this, we propose a novel paradigm to train teams of LLMs to collaboratively solve tasks.

Most relevant to the main idea of this paper is DebateGPT (Subramaniam et al., 2024). DebateGPT uses debate as a mechanism to attain higher-quality fine-tuning data. Differing from our work, their approach focuses on using debate to generate better training data for a single model that gives single responses, rather than optimizing the LLMs themselves for collaborative problem-solving over multiple rounds of conversation.

In this paper, we propose a novel framework Actor-Critic Debate (ACC-Debate) which jointly trains a two-agent team to collaboratively solve problems through iterative conversation; this team consists of an actor-agent, responsible for providing answers for a given task, and a critic-agent, responsible for assisting the actor-agent with feedback on its answers. In our training pipeline, we introduce a novel off-policy learning scheme called "guided-debate" to generate high-quality multi-turn training data to enhance the actor's and critic's performance on challenging tasks.

To summarize, our contributions are as follows:

-
- We are the first to propose a framework for training a team of LLMs jointly (Actor-Critic) within the context of debates.
 - We introduce a novel data generation scheme, “guided debate trajectories”, which enables the efficient creation of high-quality training data for both the actor and critic roles.
 - Our extensive experiments demonstrate that our method, ACC-Debate, significantly outperforms existing state-of-the-art approaches.

2 RELATED WORK

Our research is closely related to the emerging field of multi-agent debate (sometimes called multi-LLM deliberation) which examines how to use groups of models to solve tasks through iterative discussion Chan et al. (2023); Liang et al. (2023); Du et al. (2023); Khan et al. (2024); Michael et al. (2023); Rasal (2024); Pham et al. (2023); Abdelnabi et al. (2023); Hong et al. (2023); Irving et al. (2018); Li et al. (2023b;c; 2024a); Wang et al. (2023a); Zhang et al. (2023). Many of these works find that language models have naturally collaborative abilities Singhal et al. (2023); Du et al. (2023); Chan et al. (2023), while others have noted that the collaborative ability of off-the-shelf models can be quite limited Wang et al. (2024); Smit et al..

Current approaches to multi-agent debate can be broadly cast into two main categories: those that modify model prompts and responses during the debate Liang et al. (2023); Khan et al. (2024); Rasal (2024); Feng et al. (2024); Yang et al. (2024), and those that modify the structure of the debate process Li et al. (2023a); Hong et al. (2023); Liu et al. (2023); Li et al. (2024c); Wang et al. (2023b); Wu et al. (2023); Chen et al. (2023); Chang (2024b). Importantly, both categories use off-the-shelf language models (which have not been trained to collaborate) and work by modifying either the inputs or outputs of these models. Deviating from this line of work, we aim to specifically train a team of models to collaboratively solve tasks.

Two works of particular note are that of Subramaniam et al. (2024), which proposes to use debate data to fine-tune models, and Li et al. (2024b), which trains models for adversarial debate. In the former, debate is used to generate higher-quality fine-tuning data and is not used at inference time; differing from this work, we train models directly to collaborate and use debate both during training and inference. In the latter, models are trained to be effective arguers rather than collaborators, i.e., models are trained to give convincing arguments such that they can *win* a debate against other LLMs. Differing from this work, we train models to collaboratively solve tasks.

In the context of multi-agent deliberation, the concept of *divergent opinions* is highly relevant to our method. Several approaches to multi-agent debate aim to control the level of disagreement among the agents Liang et al. (2023); Khan et al. (2024); Chang (2024a). Often, these works dynamically increase disagreement to prevent early convergence of debate. In our study, we leverage divergent opinions to generate high-quality training data. In particular, we have debaters change their opinion during the debate and measure whether or not that change increases or decreases the likelihood that the debate converges to a correct answer; this is used to assess the *value* of a given training example.

Also closely related to our work are paradigms that aim to use self-generated data to improve model performance, often in the context of reasoning or chain of thought Trung et al. (2024); Huang et al. (2024); Xiong et al. (2024); Chen et al. (2024); Pang et al. (2024b). Similar to this line of research, we make use of model generations as training data. However, we are the first work to use such data in the context of multiple models debating collaboratively to solve a given task.

3 PRELIMINARIES AND NOTATION

In this section, we formalize the debate between an Actor (a model that provides answers) and a Critic (a model that provides feedback to the actor) while also introducing the notation that will be used throughout the remainder of the paper.

Let $(x, y) \sim \mathcal{D}$ be a task-answer pair source from a distribution of tasks and answers \mathcal{D} . For a given task x , two models – an actor model responsible for providing answers and a critic model responsible for providing feedback and assistance to the actor model – engage in an iterative discussion over T

rounds, to correctly infer the answer y . Let θ_a and θ_c be the parameters of actor and critic models, respectively. The iterative discussion between these two models is as follows:

1. At round $t = 0$ a task x is given to the actor θ_a who provides an initial response $z_a^{(0)}$.
2. Next, still at round $t = 0$, the critic θ_c views task x and $z_a^{(0)}$, then provides feedback $z_c^{(0)}$.
3. For each round $t > 0$, the actor views the task x , its own previous response $z_a^{(t-1)}$ and the critic’s feedback $z_c^{(t-1)}$, then provides an updated response $z_a^{(t)}$.
4. After the actor’s new response $z_a^{(t)}$, the critic provides the feedback $z_c^{(t)}$ based on $z_a^{(t)}$.

The accuracy of this procedure is measured via the correctness of the actor’s final response, i.e., $\mathbb{I}[\zeta(z_a^{(T)}) = y]$. Where ζ is a function that extracts *answers* from text-based responses. For example if $z_a^{(T)} = \text{“The sky is blue”}$, then $\zeta(z_a^{(T)}) = \text{“blue”}$. With this notation and formalization of debate, we introduce our framework for training actor-critic teams to debate.

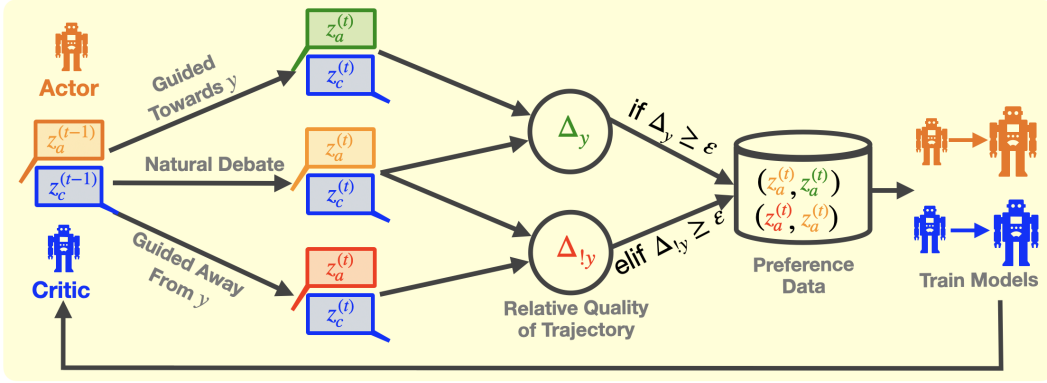


Figure 1: ACC-Debate training pipeline, exemplified for the actor. 1) We generate data from both natural debate as well as guided debate towards and away from the ground truth answer y using the actor and critic. 2) We compute the relative quality of each trajectory based on the expected quality difference Δ w.r.t. to the natural response. 3) We store all high-quality pairwise data in our database and train the actor model. 4) We alternate this procedure for the actor and critic.

4 METHODOLOGY

In this section, we outline our procedure for training a two-agent debate team, consisting of an actor model f_{θ_a} (responsible for providing answers to a given task x) and a critic model f_{θ_c} (responsible for providing feedback and assistance to the actor). At inference time, the two trained agents engage in iterative discussion to solve a given task x , generating the final response $z_a^{(T)}$.

4.1 AN ACTOR-CRITIC DEBATE FRAMEWORK

Building upon our established notation from the previous section and the general actor-critic framework, we formally define our optimization objective as follows ¹:

$$\theta_a^*, \theta_c^* = \arg \max_{\theta_a} \max_{\theta_c} \mathbb{E}_{(x,y) \sim D} \left[\zeta \left(\underbrace{f_{\theta_a}(x, z_a^{(T-1)}, z_c^{(T-1)})}_{\text{actor's final response } z_a^{(T)}} \right) = y \right] \quad (1)$$

Intuitively, Eq. 1 aims to simultaneously optimize the actor’s parameters θ_a and the critic’s parameters θ_c , ensuring that the actor’s final output at iteration T matches the correct answer y .

¹For clarity, we note that the term $\arg \max_{\theta_a} \max_{\theta_b}$ captures the solution for both parameters θ_a, θ_c in the corresponding bi-level max-max optimization. Here θ_c is in $z_c^{(T-1)} = f_{\theta_c}(x, z_a^{(T-1)})$

In other words, we optimize the accuracy of the actor’s response at time T , namely $z_a^{(T)} = f_{\theta_a}(x, z_a^{(T-1)}, z_c^{(T-1)})$, where accuracy is measured as $\mathbb{E}[\zeta(z_a^{(T)}) = y]$.

It is important to note that the recursive nature of multi-agent debate introduces significant complexity to the optimization process. Each response $z_a^{(t)}$ depends not only on the actor’s previous output $z_a^{(t-1)}$ but also on the critic’s previous output $z_c^{(t-1)}$. This interaction closely resembles a *cooperative dynamic Stackelberg game* (Li & Sethi, 2017), where two players engage in hierarchical decision-making over time, leading us to adopt an *iterative best-response* approach (Fiez et al., 2019). In other words, we first train the critic model, followed by training the actor to best respond to the critic’s output. We can then update the critic to adapt to the newly trained actor, and so on. More formally, this process works by first fixing θ_a , and solving,

$$\theta_c^* = \arg \max_{\theta_c} \mathbb{E}_{(x,y) \sim D} \left[\zeta \left(f_{\theta_a} \left(x, z_a^{(T-1)}, \underbrace{f_{\theta_c}(x, z_a^{(T-1)})}_{\text{critic's response } z_c^{(T-1)}} \right) \right) = y \right] \quad (2)$$

then fixing θ_c^* from above, we solve

$$\theta_a^* = \arg \max_{\theta_a} \mathbb{E}_{(x,y) \sim D} \left[\zeta \left(f_{\theta_a} \left(x, z_a^{(T-1)}, f_{\theta_c^*}(x, z_a^{(T-1)}) \right) \right) = y \right] \quad (3)$$

this process then repeats until a desired stopping criteria is reached. In practice, we find that a single iteration is sufficient to produce a high-quality debate team.

While this alternating scheme allows us to optimize the actor and critic separately, the objectives of each model still cannot be optimized directly due to the recursive nature of model responses in this objective; responses at round T depend on those given by the model at round $t - 1$ which themselves depend on the response given at round $t - 2$ and so on. To deal with this temporal dependency, we next introduce the concept of Partial Trajectory rewards, which will allow us to capture the signal of each response $z^{(t)}$ for each $t \leq T$.

4.2 PARTIAL TRAJECTORY REWARD

To address the inter-round dependencies of the above optimization, we proposed a scheme that allows us to determine the “goodness” of a given response $z^{(t)}$ (from either the actor or the critic) for any $t \leq T$. Consider a conversation between the actor and the critic that was paused at time t , i.e., the most recent response is $z^{(t)}$. To assess the goodness of $z^{(t)}$, one might ask *how likely the debate procedure will converge to the correct answer y at round T* , given that the procedure is already at response $z^{(t)}$. Formally, we can define this as

$$r(z^{(t)}, x, y) = \mathbb{E} \left[\zeta(z_a^{(T)}) = y \mid x, z^{(t)} \right] \quad (4)$$

Intuitively, the partial reward captures the expectation of arriving at the correct answer y through debate starting at round t with generation $z^{(t)}$. In practice, $r(z^{(t)}, x, y)$ can be estimated by learning the reward structure r or by using heuristics such as one-step roll-out, i.e., Monte Carlo estimation.

In our experiments, we use one-step roll-out heuristics. This means that we simulate one additional round of debate multiple times (starting from response $z^{(t)}$), the value of $r(z^{(t)}, x, y)$ is then set as the average accuracy of these simulated rounds. Empirically, we observe that these heuristics are effective in producing high-quality training data. We leave learning-based reward functions for future work.

Our objective will then be to optimize the actor and critic, θ_a, θ_c , so that the responses produced by these models at each timestep t , namely $z^{(t)}$, maximize $r(z^{(t)}, x, y)$. That is, we optimize the actor and the critic so that at each timestep t , they give a response $z^{(t)}$ which has a high probability of leading the debate to converge to the correct answer at time T .

To optimize the objective in Eq. 1, we will utilize preference optimization, a standard technique in LLM training. Using the iterative maximization scheme described above, we first have to gather pairwise preference data for both the actor and the critic. In the following sections, we first detail our process for generating this preference data before delving into the optimization procedure.

```

216
217 Algorithm 1: Trajectory generation and selection
218
219 Data: Actor and critic:  $\theta_a, \theta_c$ , Distribution of tasks  $\mathcal{D}$ 
220 Result: A dataset of trajectories  $D$ 
221  $D \leftarrow \emptyset$  /* Set of trajectories to use */
222 for  $(x, y) \sim \mathcal{D}$  do
223    $\mathbf{z}^{(0)} \leftarrow \text{OneDebateRound}(x)$  /* actor and critic responses, i.e.  $\mathbf{z} = \langle z_a^{(0)}, z_c^{(0)} \rangle$  */
224   for  $t$  in  $[1, T]$  do
225      $\mathbf{z}^{(t)} \leftarrow \text{OneDebateRound}(x, \mathbf{z}^{(t-1)})$  /* updated natural responses*/
226     /* guided-debate towards, and away from, correct answer  $y$  */
227      $\mathbf{z}_+^{(t)} \leftarrow \text{OneGuidedDebateRound}(x, \mathbf{z}^{(t-1)}, y)$ 
228      $\mathbf{z}_-^{(t)} \leftarrow \text{OneGuidedDebateRound}(x, \mathbf{z}^{(t-1)}, !y)$ 
229     /* Estimate final round accuracy if debate continue from response  $z^{(t)}$ ,
230     i.e.  $r(z^{(t)}, x, y)$ ,  $r(z_+^{(t)}, x, y)$ ,  $r(z_-^{(t)}, x, y)$  */
231      $v \leftarrow \text{EstimateFinalAccuracy}(\mathbf{z}^t)$ 
232      $v_+ \leftarrow \text{EstimateFinalAccuracy}(\mathbf{z}_+^t)$ 
233      $v_- \leftarrow \text{EstimateFinalAccuracy}(\mathbf{z}_-^t)$ 
234     /* Compute the expected improvement in debate for each trajectory */
235     /* Save trajectory pairs that result in sufficient accuracy improvement */
236     if  $v_+ - v \geq t$  then
237        $D.\text{add}(\text{pos}=\mathbf{z}_+^{(t)}, \text{neg}=\mathbf{z}^{(t)})$ 
238     end
239     else if  $v - v_- \geq t$  then
240        $D.\text{add}(\text{pos}=\mathbf{z}^{(t)}, \text{neg}=\mathbf{z}_-^{(t)})$ 
241     end
242   end
243 end

```

4.3 OFF-POLICY TRAJECTORY GENERATION

In this section, we describe how to generate the preference data needed to optimize the objective in Eq. 1. The classification of a sample as positive or negative is determined by the debate trajectory it follows. Specifically, a positive sample for training the actor corresponds to a trajectory likely to lead to the correct answer at round T , while a negative sample corresponds to one that leads to an incorrect answer at round T . Intuitively, we aim to push the actor model to generate responses that lead to correct answers while reducing responses that are unlikely to do so, thus optimizing for Eq. 3. The same principle applies to the critic when optimizing Eq. 2.

With this intuition in mind, we now describe how such data is generated. A debate trajectory can be defined as a sequence of responses $\langle z_a^{(0)}, z_c^{(0)}, z_a^{(1)}, z_c^{(1)}, \dots, z_a^{(T)}, z_c^{(T)} \rangle$ for a given task x . A straightforward way to generate preference data would be to generate multiple rollouts at each round and select the trajectories with the highest $r(z^{(t)}, x, y)$ as positive samples and those with the lowest $r(z^{(t)}, x, y)$ as negative samples. This approach could enforce the desired behavior for both the actor and the critic if enough samples are collected.

However, this approach is not without its limitations. In particular, if the model performs poorly on a given dataset, it may be difficult to collect enough positive samples, resulting in low training signals. Additionally, even if the model performs adequately, generating sufficient responses for both the actor and critic requires significant computational resources, especially to ensure that high $r(z^{(t)}, x, y)$ values are used for positive samples and low values for negative samples.

Guided-Debate To address these limitations and improve efficiency, we propose *guided-debate* trajectories, which steer the debate procedure in two opposing directions: one towards, and another away from, the correct. By comparing these guided trajectories with the natural debate trajectory, we can assess the relative *goodness* of each trajectory using an estimation of the reward structure r .

Specifically, for task x with answer y , let $\mathbf{z}^{(t-1)} = (z_a^{(t-1)}, z_c^{(t-1)})$ be the models’ responses at time $t - 1$. Let $(z_a^{(t)}, z_c^{(t)})$ be the models’ *natural* responses (i.e., without guidance), let $(z_{y,a}^{(t)}, z_{y,c}^{(t)})$ and $(z_{!y,a}^{(t)}, z_{!y,c}^{(t)})$ be the models responses when guided towards, and away from, supporting answer y respectively. Thus, each guided response is an off-policy generation. In practice, we want guided responses to be different enough from natural responses so that learning the guided responses results in consequential changes to the model, but not so different that they are challenging to learn; we find that prompt modification is an effective tool for striking this balance. To guide the generations of $(z_{y,a}^{(t)}, z_{y,c}^{(t)})$ and $(z_{!y,a}^{(t)}, z_{!y,c}^{(t)})$, we will simply provide a correct and wrong target answer in the prompt, respectively - see “Guided Debate Prompts” in Section C of the supplement for further details.

For each guided response, we consider how influential this response was in altering the accuracy of the final response, i.e., in the case of the actor, we define

$$\Delta_y = r(z_{y,a}^{(t)}, x, y) - r(z_a^{(t)}, x, y) \quad \text{and} \quad \Delta_{!y} = r(z_a^{(t)}, x, y) - r(z_{!y,a}^{(t)}, x, y)$$

The terms Δ_y and $\Delta_{!y}$ give the expected accuracy difference if at round t the actor *had* given response $z_{y,a}^{(t)}$ (or $z_{!y,a}^{(t)}$) instead of response $z_a^{(t)}$. Large Δ_y indicates that a one-response difference during the debate was sufficient to push the procedure toward the correct answer. Such responses would be desirable for the model to learn. On the other hand, large values of $\Delta_{!y}$ indicate that a one-response difference easily causes the models to converge to the incorrect answer; this indicates that the debate procedure is particularly fragile at timestep t .

With these observations in hand, we use Δ_y and $\Delta_{!y}$ to define positive and negative examples, in particular for a threshold ε ,

$$(z_+^{(t)}, z_-^{(t)}) = \begin{cases} (z_y^{(t)}, z^{(t)}) & \text{if } \varepsilon \leq \Delta_y = r(z_{y,a}^{(t)}, x, y) - r(z_a^{(t)}, x, y) \\ (z^{(t)}, z_{!y}^{(t)}) & \text{if } \varepsilon \leq \Delta_{!y} = r(z_a^{(t)}, x, y) - r(z_{!y,a}^{(t)}, x, y) \end{cases} \quad (5)$$

if neither value is above the threshold, then the example is thrown out.

Remark 1 Under Eq. 5, a positive example $z_+^{(t)}$ can be interpreted as a guided response $z_y^{(t)}$ which *increased* the probability of debate converging to the correct answer by at least ε , when compared with the natural response $z^{(t)}$. Similarly, a negative example $z_-^{(t)}$ is a guided response $z_{!y}^{(t)}$ which *decreased* the probability of debate converging to the answer by at least ε .

Now that we have a procedure for generating high-quality training examples consisting of positive and negative pairs, we next discuss how to use those positive and negative pairs to train both models.

Preference Optimization In order to optimize each objective (Eq. 2 and 3), we use standard preference optimization Direct Preference Optimization (DPO) Rafailov et al. (2024). We choose DPO for its efficiency, but any preference optimization scheme could be used (see section B of the supplement for details on how other preference optimization schemes may be incorporated).

Hence, given a preference dataset of positive and negative examples for both the actor and critic model, of the form $z_-^{(t)}, z_+^{(t)}$, the DPO loss is defined as,

$$\mathcal{L}_{\text{DPO}} = \sum_{t=0}^T \mathbb{E}_{(x,y,z_-^{(t)}, z_+^{(t)}) \sim D} \left[\log \sigma \left(\frac{\pi_\theta(z_+^{(t)} | x, \mathbf{z}^{(t-1)})}{\pi_{\text{ref}}(z_+^{(t)} | x, \mathbf{z}^{(t-1)})} - \frac{\pi_\theta(z_-^{(t)} | x, \mathbf{z}^{(t-1)})}{\pi_{\text{ref}}(z_-^{(t)} | x, \mathbf{z}^{(t-1)})} \right) \right] \quad (6)$$

Where π_θ is the policy induced by parameters θ_a or θ_c and $\mathbf{z}^{(t-1)}$ are the model’s responses at the previous round (i.e., the responses prior to giving either response $z_-^{(t)}$ or $z_+^{(t)}$).

Remark 2 Recall, given our generated preference data, this loss implicitly optimizes the reward $r(z^t, x, y)$ which itself is equivalent to final debate accuracy (i.e., the quantity being maximized by Eq. 1). By summing across all rounds, we implicitly maximize the probability that each round t yields a response z^t which causes debate to converge to the correct answer at time T .

5 EXPERIMENTS

Benchmarks To evaluate the efficacy of ACC-Debate we make use of 5 standard benchmark tasks: **BoolQ** Clark et al. (2019) \sim 12k yes-no reading comprehension questions, **MMLU** Hendrycks et al. (2020) \sim 15k multiple choice questions covering a wide array of subjects and difficulty, **BBH** Suzgun et al. (2022) \sim 5k mixed-type questions **SCIQ** Welbl et al. (2017) \sim 13k multiple-choice science questions, **ARC** Chollet (2019) \sim 7k multiple-choice reasoning-based questions.

Baselines We compare ACC-Debate to several debate baselines. For inference-based methods, we compare to Society of Minds **SoM** Du et al. (2023), **Persona** Chan et al. (2023). For training-based methods, we compare to supervised fine-tuning **SFT** Radford (2018), **DebateTune** Li et al. (2024b), **DebateGPT** Subramaniam et al. (2024). We perform experiments using three different base models: Llama-3-8B-Instruct **Llama-3** Dubey et al. (2024), Mistral-7B-Instruct **Mistral** Jiang et al. (2023), and Gemma-2-2B-Instruct **Gemma-2** Team et al. (2024).

5.1 ACC-DEBATE PERFORMANCE

We begin by examining the performance of our method ACC-Debate (a single round of training) and ACC-Debate+ (two rounds of training) In table 1, we see the average accuracy of each method after five rounds of debate. Our method attains superior performance compared with baseline methods in most cases. The high efficacy of ACC-Debate relative to the baselines indicates that in most cases, only a single round of training is necessary to produce a high-quality debate team. It is worth noting that in some cases, further training rounds may decrease performance (i.e., ACC-Debate+ can have worse performance than ACC-Debate). As such, we recommend maintaining a hold-out set of tasks to determine whether further training degrades performance.

	Llama-3						(Ours)	(Ours)
	SoM (2x)	SoM (4x)	Persona	DebateTune	SFT	DebateGPT	ACC-Debate	ACC-Debate+
BoolQ	.812 \pm .01	.811 \pm .007	.781 \pm .002	.775 \pm .033	.798 \pm .006	.815 \pm .005	.887 \pm .005	.894 \pm .003
MMLU	.62 \pm .004	.639 \pm .004	.639 \pm .004	.63 \pm .004	.642 \pm .005	.654 \pm .005	.644 \pm .01	.683 \pm .012
BBH	.508 \pm .003	.514 \pm .005	.509 \pm .013	.508 \pm .005	.552 \pm .006	.551 \pm .008	.593 \pm .006	.574 \pm .003
SCIQ	.925 \pm .002	.923 \pm .002	.925 \pm .004	.924 \pm .004	.925 \pm .003	.932 \pm .001	.952 \pm .0	.948 \pm .003
ARC	.874 \pm .001	.874 \pm .001	.87 \pm .003	.871 \pm .002	.879 \pm .004	.876 \pm .002	.881 \pm .004	.869 \pm .002
	Mistral						(Ours)	(Ours)
	SoM (2x)	SoM (4x)	Persona	DebateTune	SFT	DebateGPT	ACC-Debate	ACC-Debate+
BoolQ	.801 \pm .005	.798 \pm .004	.831 \pm .003	.83 \pm .003	.84 \pm .003	.848 \pm .002	.877 \pm .002	.893 \pm .002
MMLU	.57 \pm .003	.562 \pm .005	.574 \pm .002	.562 \pm .005	.594 \pm .004	.577 \pm .002	.61 \pm .005	.672 \pm .004
BBH	.428 \pm .002	.462 \pm .003	.465 \pm .011	.456 \pm .005	.439 \pm .006	.48 \pm .012	.519 \pm .009	.601 \pm .004
SCIQ	.856 \pm .002	.856 \pm .002	.86 \pm .002	.863 \pm .003	.858 \pm .004	.871 \pm .003	.902 \pm .005	.905 \pm .002
ARC	.824 \pm .001	.823 \pm .002	.827 \pm .001	.834 \pm .0	.825 \pm .003	.822 \pm .002	.843 \pm .003	.856 \pm .003
	Gemma-2						(Ours)	(Ours)
	SoM (2x)	SoM (4x)	Persona	DebateTune	SFT	DebateGPT	ACC-Debate	ACC-Debate+
BoolQ	.75 \pm .011	.759 \pm .004	.716 \pm .015	.767 \pm .003	.783 \pm .011	.812 \pm .003	.84 \pm .005	.845 \pm .005
MMLU	.58 \pm .002	.578 \pm .002	.577 \pm .002	.578 \pm .001	.579 \pm .002	.582 \pm .002	.51 \pm .016	.555 \pm .003
BBH	.454 \pm .007	.449 \pm .01	.447 \pm .006	.447 \pm .007	.498 \pm .006	.491 \pm .01	.513 \pm .006	.475 \pm .008
SCIQ	.903 \pm .002	.903 \pm .002	.908 \pm .003	.903 \pm .001	.913 \pm .002	.914 \pm .002	.918 \pm .003	.909 \pm .003
ARC	.841 \pm .003	.843 \pm .005	.847 \pm .003	.847 \pm .003	.848 \pm .002	.851 \pm .003	.852 \pm .003	.849 \pm .002

Table 1: Average accuracy (with 95% confidence intervals) after 5 rounds of debate. For each dataset, the highest accuracy is shown in bold.

5.2 PERFORMANCE INCREASE OF DEBATE

Average Improvement As noted in Du et al. (2023), the key mechanism behind the success of multi-agent debate (or any of its many variants) is that discussion over multiple rounds allows the models to iteratively refine their answers. Thus, a natural question for any multi-agent debate method is: *how much does accuracy improve from the initial round $t = 0$ to the final round $t = T$?* Where $T = 4$ in our experiments. To measure this, we look at the percent improvement in model accuracy

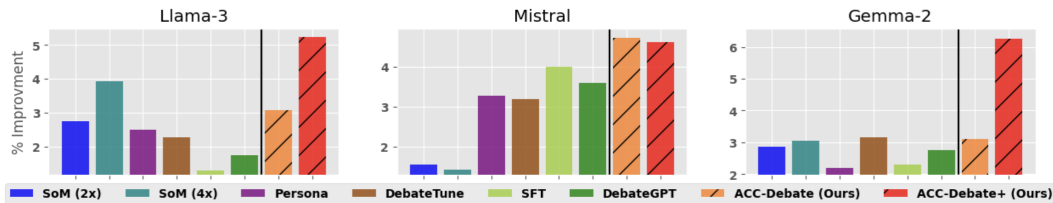


Figure 2: Percent improvement in accuracy after five rounds of debate, compared to a single round. Percent improvement (Eq. 7) for each method is averaged across all five datasets.

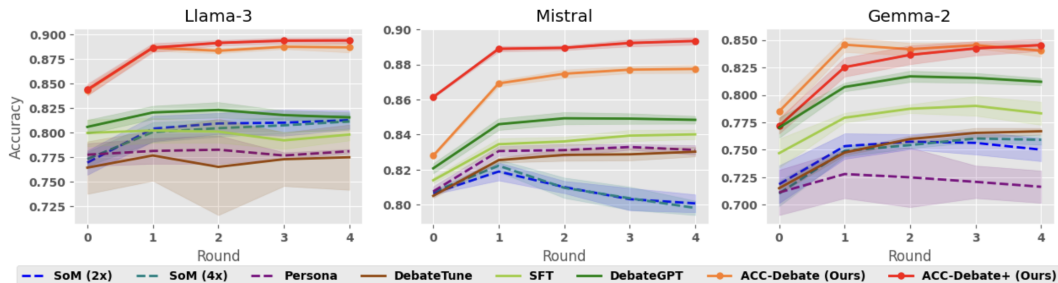


Figure 3: Accuracy over five rounds of debate on BoolQ.

from round $t = 0$ to round $t = 4$ calculated as,

$$\frac{\text{acc}_4 - \text{acc}_0}{\text{acc}_0} \quad \text{where } \text{acc}_t \text{ is accuracy at round } t \quad (7)$$

In Figure 2 we see the average percent improvement for each method, averaged across all 5 datasets. For each of the three base models, ACC-Debate+ has the highest average improvement compared to all other methods. Additionally, the improvement gained by methods such as SoM, SFT or DebateGPT is far less stable than that of ACC-Debate+. In particular, for Mistral, SoM yields nearly no improvement, similarly SFT and DebateGPT offer little improvement when applied to of Llama-3.

Per-Round Accuracy Next, we look more closely at the accuracy of each method across five rounds of debate. Figure 3, shows per-round accuracy on the BoolQ dataset. As already illustrated by Table 1, ACC-Debate and ACC-Debate+ achieve higher final round accuracy than the other methods. Notably, our method has higher accuracy both at the final round $t = 4$ and at round $t = 0$. Recall that at round $t = 0$, the actor’s response is independent of the critic. This indicates that in some cases, our training pipeline can produce actor models with superior zero-shot accuracy (without debate) compared to models produced by SFT and DebateGPT.

Interestingly, we observe that in some cases, SFT and DebateGPT are not much better than simple debate with untrained models (i.e., SoM-4x), e.g., BoolQ with Llama-3. This is primarily due to the fact that these methods are designed to improve the model’s single-shot performance but do little to improve their collaborative abilities. Cases such as this outline the necessity of training to improve collaboration (ACC-Debate) rather than training for raw accuracy (SFT and DebateGPT).

5.3 INDIVIDUAL PERFORMANCE

Next, we examine the relative effectiveness of both the actor and the critic. To do this, we train an actor and critic via ACC-Debate. Then, during the debate, we pair the trained actor with an untrained critic and pair the trained critic with an untrained actor. In Table 2, column “Actor” corresponds to the former, while “Critic” corresponds to the latter. On average, the trained actor attains higher accuracy compared to the trained critic, this aligns with intuition as the actor is responsible for providing answers while the critic plays a supporting role. In most cases, the trained actor (paired with an untrained critic) outperforms SFT and DebateGPT. When the trained actor is paired with a trained critic (either ACC-Debate or ACC-Debate+), its performance is further improved.

Llama-3						
	SoM (4x)	DebateGPT	Actor	Critic	ACC-Debate	ACC-Debate+
BoolQ	.811±.007	.815±.005	.867±.003	.834±.005	.887±.005	.894±.003
MMLU	.635±.004	.654±.005	.651±.012	.65±.008	.644±.01	.683±.012
BBH	.514±.005	.551±.008	.583±.01	.55±.015	.593±.006	.574±.003
SCIQ	.923±.002	.932±.001	.947±.002	.945±.001	.952±.0	.948±.003
ARC	.874±.001	.876±.002	.885±.003	.866±.002	.881±.004	.869±.002
Mistral						
	SoM (4x)	DebateGPT	Actor	Critic	ACC-Debate	ACC-Debate+
BoolQ	.798±.004	.848±.002	.873±.002	.843±.002	.877±.002	.893±.002
MMLU	.562±.005	.577±.002	.598±.002	.611±.001	.61±.005	.672±.004
BBH	.462±.003	.48±.012	.493±.012	.518±.005	.519±.009	.601±.004
SCIQ	.856±.002	.871±.003	.891±.001	.891±.002	.902±.005	.905±.002
ARC	.823±.002	.822±.002	.833±.003	.842±.003	.843±.003	.856±.003
Gemma-2						
	SoM (4x)	DebateGPT	Actor	Critic	ACC-Debate	ACC-Debate+
BoolQ	.759±.004	.812±.003	.839±.005	.774±.014	.84±.005	.845±.005
MMLU	.578±.002	.582±.002	.519±.026	.566±.002	.51±.016	.555±.003
BBH	.449±.01	.491±.01	.51±.011	.475±.004	.513±.006	.475±.008
SCIQ	.903±.002	.914±.002	.923±.002	.912±.001	.918±.003	.909±.003
ARC	.843±.005	.851±.003	.85±.002	.855±.002	.852±.003	.849±.002

Table 2: Accuracy after 5 rounds of debate. The Actor (Critic) column corresponds to an actor-agent (critic-agent) trained via ACC-Debate and paired with an untrained Critic (Actor) during debate.

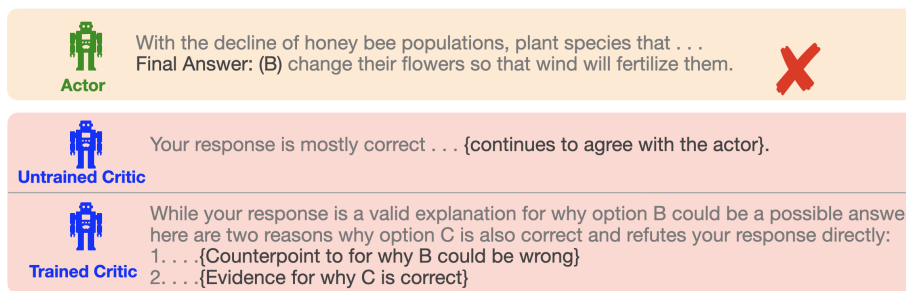


Figure 4: Comparison of responses from the critic model before and after training with ACC-Debate.

5.4 WHAT DO THE AGENTS ACTUALLY LEARN?

Lastly, we are interested in understanding how ACC-Debate improves the Actor-Critic Team. Figure 4 demonstrates an example of the difference in responses between an untrained critic and a critic trained through ACC-Debate. Although the actor provides a wrong answer, the untrained critic is too agreeable and does not provide substantive feedback for the actor to correct their answer.

In contrast, the trained critic is more willing to disagree with the actor and provides more detailed feedback. Largely, we observe that this trend is common; untrained critics are too agreeable and are thus less able to change the actor’s mind, while trained critics are more willing to disagree (see Section C.2 for more examples). When examining the trained actor’s responses, we do not find a notable qualitative change compared to the responses of an untrained actor. However, as shown previously, we do observe a qualitative change in the actor’s responses to become more accurate.

6 CONCLUSION, LIMITATIONS AND IMPACT

In this paper, we propose ACC-Debate, a novel framework for jointly training a two-agent team (one actor-agent and one critic-agent) to collaboratively solve problems through iterative discussion. To train these agents, we developed an off-policy data generation scheme dubbed “Guided-debate”, which produces high-quality preference data for collaborative models. We found that ACC-Debate outperforms all baselines on a wide array of domains. In particular, even a single round of training for both the actor and critic results in a high-quality debate team. Of particular note is the effects that ACC-Debate has on the critic model. Without ACC-Debate, the critic model is often too agree-

486 able and lacks verbosity in their responses. In contrast, after training with ACC-Debate, the critic
487 is far more likely to provide detailed disagreements during debate. The success of ACC-Debate
488 demonstrates the need for more specialized training when it comes to collaborative models.

489 However, our framework ACC-Debate also comes with limitations. Firstly, even though ACC-
490 Debate attains superior performance compared to baselines on a wide array of domains, it is impor-
491 tant to note that we conduct experiments mainly on question-answering tasks; thus, it remains to be
492 seen whether such a framework would continue to be effective in other types of tasks. *Moreover, in*
493 *our experiments, we train and test models on the same task (partitioning each task into a training and*
494 *testing set). As such, the generalizability of each actor-critic team to unseen domains is unknown.*
495 Our method makes use of the fact that for each question, correct and incorrect answers can be easily
496 established. Secondly, while we provide results for three families of models, these experiments are
497 performed on 2B, 7B, and 8B models. While our method is effective for these sizes (standard in
498 open-source models), it remains to be seen whether this effectiveness will scale to larger models.
499

500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593

REPRODUCIBILITY STATEMENT

Here, we outline the details necessary to reproduce our results. We provide an algorithm for our data generation procedure (Algorithm 1), as well as a description of our training procedure in Section 4.3. Each dataset, baseline method, and base model used are specified at the beginning of Section 5. We provide additional experimental details in Section A. Prompts used for our method can be found in Section C. Lastly, we will release all code used for our method upon publication of our work.

REFERENCES

- Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation games. *arXiv preprint arXiv:2309.17234*, 2023.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.
- Edward Y Chang. Evince: Optimizing adversarial llm dialogues via conditional statistics and information theory. *arXiv preprint arXiv:2408.14575*, 2024a.
- Edward Y Chang. Socrasynth: Multi-llm reasoning with conditional statistics. *arXiv preprint arXiv:2402.06634*, 2024b.
- Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*, 2023.
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models, 2024. URL <https://arxiv.org/abs/2401.01335>.
- François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don’t hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*, 2024.
- Tanner Fiez, Benjamin Chasnov, and Lillian J Ratliff. Convergence of learning dynamics in stackelberg games. *arXiv preprint arXiv:1906.01217*, 2019.
- Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 2023.
- Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, et al. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*, 2024.
- Alex Havrilla, Sharath Rappathy, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Railneau. Glore: When, where, and how to improve llm reasoning via global and local refinements. *arXiv preprint arXiv:2402.10963*, 2024.

594 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and
595 Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint*
596 *arXiv:2009.03300*, 2020.

597
598 Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang,
599 Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-
600 agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

601
602 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong
603 Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language
604 models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*,
605 2023.

606
607 Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen.
608 Key-point-driven data synthesis with its enhancement on mathematical reasoning. *arXiv preprint*
609 *arXiv:2403.02333*, 2024.

610
611 Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint*
612 *arXiv:1805.00899*, 2018.

613
614 Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating
615 llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics:*
616 *EMNLP 2023*, pp. 1827–1843, 2023.

617
618 Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot,
619 Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al.
620 Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

621
622 Ziheng Jiang, Haibin Lin, Yinmin Zhong, Qi Huang, Yangrui Chen, Zhi Zhang, Yanghua Peng,
623 Xiang Li, Cong Xie, Shibiao Nong, et al. {MegaScale}: Scaling large language model training
624 to more than 10,000 {GPUs}. In *21st USENIX Symposium on Networked Systems Design and*
625 *Implementation (NSDI 24)*, pp. 745–760, 2024.

626
627 Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Ed-
628 ward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more
629 persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024.

630
631 Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph
632 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model
633 serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Prin-*
634 *ciples*, pp. 611–626, 2023.

635
636 Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem.
637 Camel: Communicative agents for” mind” exploration of large scale language model society.
638 *arXiv preprint arXiv:2303.17760*, 2023a.

639
640 Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and
641 Katia Sycara. Theory of mind for multi-agent collaboration via large language models. *arXiv*
642 *preprint arXiv:2310.10701*, 2023b.

643
644 Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *arXiv*
645 *preprint arXiv:2402.05120*, 2024a.

646
647 Ming Li, Jiuhai Chen, Lichang Chen, and Tianyi Zhou. Can llms speak for diverse peo-
648 ple? tuning llms via debate to generate controllable controversial statements. *arXiv preprint*
649 *arXiv:2402.10614*, 2024b.

650
651 Tao Li and Suresh P Sethi. A review of dynamic stackelberg game models. *Discrete & Continuous*
652 *Dynamical Systems-B*, 22(1):125, 2017.

653
654 Yuan Li, Yixuan Zhang, and Lichao Sun. Metaagents: Simulating interactions of human behav-
655 iors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint*
656 *arXiv:2310.06500*, 2023c.

648 Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene
649 Ie. Improving multi-agent debate with sparse communication topology. *arXiv preprint*
650 *arXiv:2406.11776*, 2024c.

651 Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng
652 Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-
653 agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

654 Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An llm-
655 agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*,
656 2023.

657 Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar,
658 and Samuel R Bowman. Debate helps supervise unreliable experts. *arXiv preprint*
659 *arXiv:2311.08702*, 2023.

660 Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for
661 natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):
662 604–624, 2020.

663 Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason
664 Weston. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024a.

665 Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason
666 Weston. Iterative reasoning preference optimization, 2024b. URL <https://arxiv.org/abs/2404.19733>.

667 Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A Plummer,
668 Zhaoran Wang, and Hongxia Yang. Let models speak ciphers: Multiagent debate through
669 embeddings. *arXiv preprint arXiv:2310.06272*, 2023.

670 Alec Radford. Improving language understanding by generative pre-training. 2018.

671 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea
672 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances*
673 *in Neural Information Processing Systems*, 36, 2024.

674 Sumedh Rasal. Llm harmony: Multi-agent communication for problem solving. *arXiv preprint*
675 *arXiv:2401.01312*, 2024.

676 Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models.
677 *arXiv preprint arXiv:2309.05922*, 2023.

678 Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and
679 Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL
680 <https://arxiv.org/abs/2303.11366>.

681 Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen
682 Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering
683 with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

684 Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D Barrett, and Arnu Pretorius.
685 Should we be going mad? a look at multi-agent debate strategies for llms. In *Forty-first International*
686 *Conference on Machine Learning*.

687 Vighnesh Subramaniam, Antonio Torralba, and Shuang Li. Debategpt: Fine-tuning large language
688 models with multi-agent debate supervision. 2024.

689 Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung,
690 Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks
691 and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

692 Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhu-
693 patiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma
694 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

702 Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez,
703 Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*,
704 29(8):1930–1940, 2023.

705 SM Tomtoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das.
706 A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv*
707 *preprint arXiv:2401.01313*, 2024.

708 Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reason-
709 ing with reinforced fine-tuning. *Proceedings of the 62nd Annual Meeting of the Association for*
710 *Computational Linguistics (Volume 1: Long Papers)*, 2024.

711 Boshi Wang, Xiang Yue, and Huan Sun. Can chatgpt defend its belief in truth? evaluating llm
712 reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP*
713 *2023*, pp. 11865–11881, 2023a.

714 Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds
715 of llm reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*, 2024.

716 Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the
717 emergent cognitive synergy in large language models: A task-solving agent through multi-persona
718 self-collaboration. *arXiv preprint arXiv:2307.05300*, 2023b.

719 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny
720 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in*
721 *neural information processing systems*, 35:24824–24837, 2022.

722 Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions.
723 *arXiv preprint arXiv:1707.06209*, 2017.

724 Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li,
725 Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-
726 agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

727 Weimin Xiong, Yifan Song, Xiutian Zhao, Wenhao Wu, Xun Wang, Ke Wang, Cheng Li, Wei Peng,
728 and Sujian Li. Watch every step! llm agent learning via iterative step-level process refinement,
729 2024. URL <https://arxiv.org/abs/2406.11176>.

730 Ruixin Yang, Dheeraj Rajagopa, Shirley Anugrah Hayati, Bin Hu, and Dongyeop Kang. Con-
731 fidence calibration and rationalization for llms via multi-agent deliberation. *arXiv preprint*
732 *arXiv:2404.09127*, 2024.

733 Jintian Zhang, Xin Xu, and Shumin Deng. Exploring collaboration mechanisms for llm agents: A
734 social psychology view. *arXiv preprint arXiv:2310.02124*, 2023.

735 Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu,
736 Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with
737 large language models. *arXiv preprint arXiv:2404.01230*, 2024.

738 Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen,
739 and Lei Li. Multilingual machine translation with large language models: Empirical results and
740 analysis, 2024. URL <https://arxiv.org/abs/2304.04675>.

741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809

APPENDIX

A EXPERIMENTS

Here, we provide additional details regarding our experiments.

Datasets Each dataset is split into a training set, a validation set, and a testing set. For datasets that come with an explicit partition of these sets we use the given partitions; this includes BoolQ, MMLU, SCIQ, and ARC. For BBH, we randomly sample roughly 25% and 10% of the questions from each category in BBH to create a test and validation set, respectively; this comes out to 1260 questions for the test set and 500 questions for the validation set. All results are reported on questions in the test set.

Compute All training was performed on a single Nvidia-H800 GPU. Inference for Llama-3 and Mistral based models is performed on a single Nvidia-v100 GPU, for Gemma-2 based models we used a single Nvidia-H800. All inference is performed with the VLLM library Kwon et al. (2023).

Training Training for all models was performed via the trl library, using LoRAs of size 256. When training ACC-Debate with DPO we use a negative log-likelihood (NLL) regularization term (with weight 1) as outlined in Pang et al. (2024a).

B PREFERENCE OPTIMIZATION

Here, we remark on how other preference optimization schemes can be used in place of DPO. Broadly speaking, preference optimization schemes can be broken into two categories: those which optimize reward directly (e.g., PPO) and those which optimize reward indirectly (e.g., DPO). In the case of the latter, the positive and negative pairs produced by our method (Algorithm 1) can be directly plugged into the preference optimizer.

In the case of explicit reward maximization, the reward function $r(z^{(t)}, x, y)$ can first be learned automatically by simply simulating debate. The most straightforward way to do this is to first simulate one full debate between the actor and critic, i.e., $\langle (z_a^{(0)}, z_c^{(0)}), \dots, (z_a^{(T)}, z_c^{(T)}) \rangle$. Then for each $z_a^{(t)}, z_c^{(t)}$ the remaining $T - t$ debate steps can be resampled to estimate the corresponding value of $r(z_a^{(t)}, x, y)$ and $r(z_c^{(t)}, x, y)$. These pairs, namely $(z_a^{(t)}, r(z_a^{(t)}, x, y))$ and $(z_c^{(t)}, r(z_c^{(t)}, x, y))$ can then be used to learn the reward function. This reward function can then be plugged into the desired preference optimization scheme.

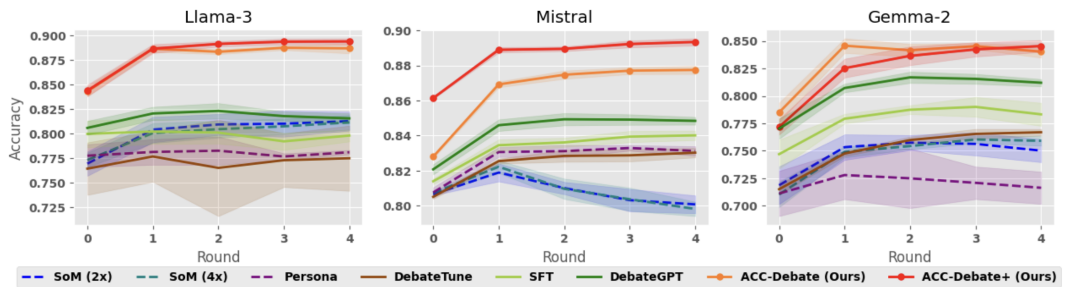


Figure 5: Accuracy over five rounds of debate on BoolQ.

C EXAMPLES

C.1 PROMPTS

Here, we provide examples of the prompts used in our experiments. For illustration, we provide prompts for the BoolQ dataset in which agents are asked a yes-no question about a passage.

810
811
812
813
814
815
816
817
818
819

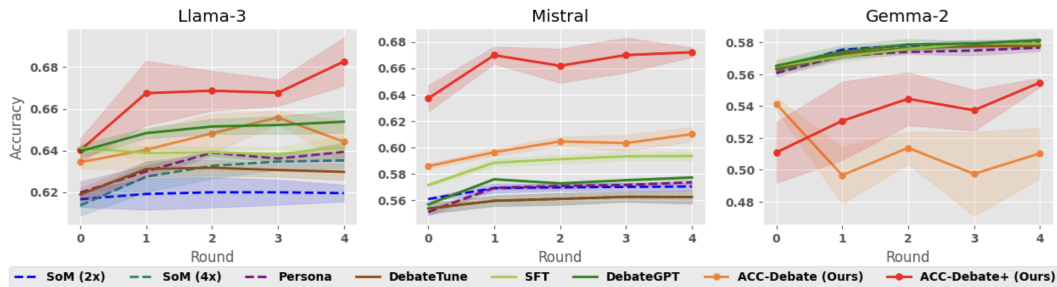


Figure 6: Accuracy over five rounds of debate on MMLU.

820
821
822
823
824
825
826
827
828
829
830
831

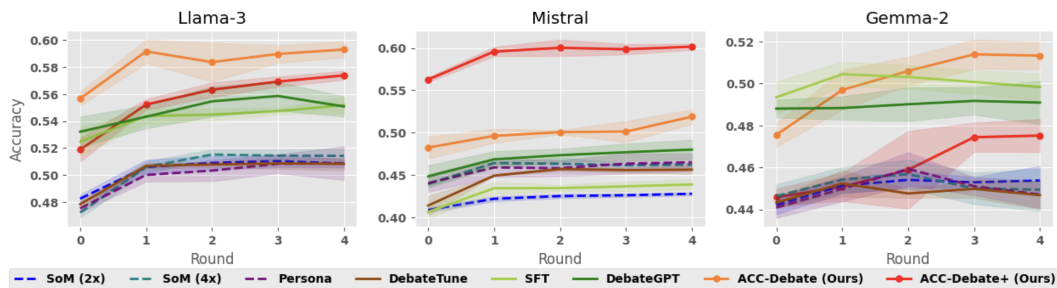


Figure 7: Accuracy over five rounds of debate on BBH.

832
833
834 **Single-Shot Prompt (No Debate)**

835
836 prompt = ('You will be given a yes-no question which is based on a passage. '
837 'You should use the passage to help you answer the question. '
838 'You should give a brief justification for your answer, '
839 'and you must provide a final answer of either Yes or No.'
840 '\nQuestion: {question}?'
841 '\nPassage: {passage}'
842)

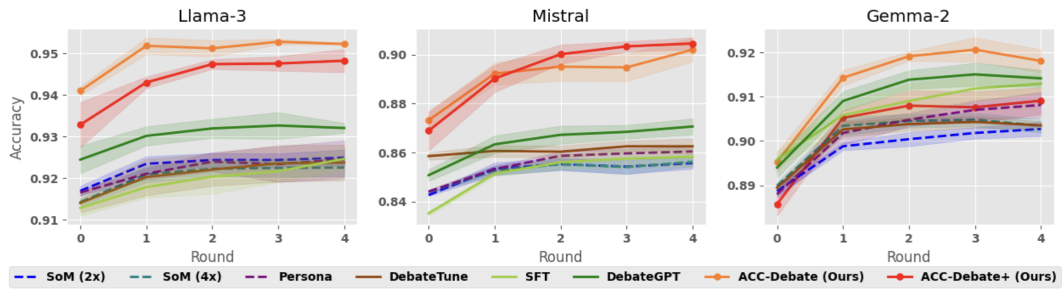
843
844 **Guided Single-Shot Prompt (No Debate)**

845
846 prompt = ('You will be given a yes-no question which is based on a passage. '
847 'You should use the passage to help you answer the question '
848 'with a {target_answer}. '
849 'You should give a brief justification for your answer of {target_answer},
850 'and you must state that your final answer is {target_answer}.'
851 '\nQuestion: {question}'
852 '\nPassage: {passage}'
853)

854 **Debate Prompt for Actor**

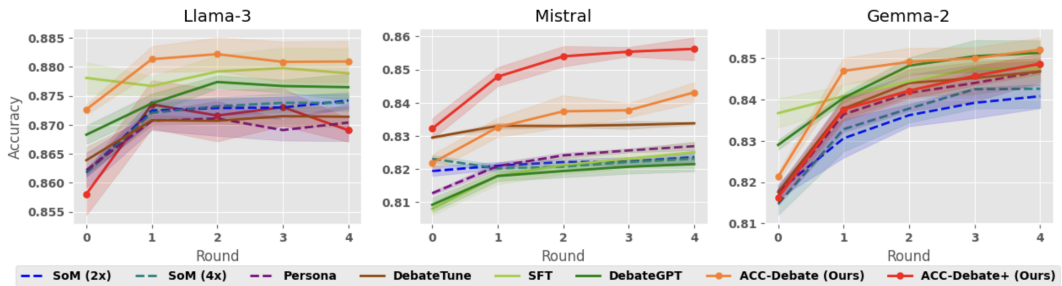
855
856 prompt = ('Several people have provided answers to a yes-no question. '
857 'Below are their responses:'
858 '\nPerson {1} said: {responses[1]}'
859 '\nPerson {2} said: {responses[2]}'
860 .
861 .
862 '\nPerson {n} said: {responses[n]}'
863 '\n\nYou should take these answers into consideration when answering '
864 'the following yes-no question, which is based on a passage. '

864
865
866
867
868
869
870
871
872
873



874 Figure 8: Accuracy over five rounds of debate on SCIQ.

876
877
878
879
880
881
882
883
884
885



886 Figure 9: Accuracy over five rounds of debate on ARC.

888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908

```

'You should give a brief justification for your answer, and you must '
'provide a final answer of either Yes or No.'
'\nQuestion: {question}'
'\nPassage: {passage}'
)

```

894 Guided Debate Prompt for Actor

```

896 prompt = ('Several people have provided answers to a yes-no question. '
897           'Below are their responses:'
898           '\nPerson {1} said: {responses[1]}'
899           '\nPerson {2} said: {responses[2]}'
900           .
901           .
902           '\nPerson {n} said: {responses[n]}'
903           'You should take these answers and the passage into consideration when
904           'answering the following question with {target_answer}. '
905           'You should give a brief justification for your answer of {target_answer},
906           '\nQuestion: {question}'
907           '\nPassage: {passage}'
908           )

```

909
910
911
912
913
914
915
916
917

910 Debate Prompt for Critic

```

912 prompt = ('I am answering a question based on a passage. '
913           'I would like you to help me improve the correctness of my response '
914           'by briefly providing some additional details my original response '
915           'may have missed. '
916           '\nQuestion: {question}'
917           '\nPassage: {passage}'
           '\nMy Response: {actor_response}'
           )

```

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971

Guided Debate Prompt for Critic

```
prompt = ('I would like you to be a debate assistant. '  
         'You will be given a question based on a passage and my response '  
         'to the question. '  
         'You should use the passage and my response to develop additional details '  
         'for why the correct answer is {target_answer}. '  
         'Your details must be brief and must support the fact that the '  
         'correct answer is {target_answer}.'  
         '\nQuestion: {question}?'  
         '\nPassage: {passage}'  
         '\nMy Response: {actor_response}'  
         )
```

C.2 CRITIC RESPONSES

Here, we present several examples of the critic’s responses when the actor provides the wrong answer; critic responses are shown before and after training with ACC-Debate. We observed a qualitative change in the critic’s responses after training with our pipeline; the critic was more willing to disagree with the actor, and the critic provided more detailed feedback.

ACTOR: [Wrong] Final answer, No. Non-residents cannot bring a handgun into New York for possession or carry within the state based on the passage. They can only transport it through New York to a destination outside of the state where it is legal to possess and carry.

UNTRAINED CRITIC: Your response is mostly correct. I would add that New York has strict gun control laws, so even if a non-resident is allowed to transport a firearm through the state, they should be aware that certain areas may have additional restrictions, such as prohibitions on carrying the firearm in certain public places or parks. It’s always a good idea to check the specific laws of the jurisdiction they will be passing through and their final destination.

TRAINED CRITIC: Details refuting your response:

1. The passage mentions that non-residents can transport any lawful firearm through New York, which implies that the firearm is temporarily in New York state territory during transit. This suggests that under certain circumstances, a handgun could be brought into New York for transit purposes, even if it cannot be carried within the state.
2. The passage does not explicitly state that non-residents are prohibited from bringing handguns into New York. Instead, it only specifies that they cannot possess or carry handguns within the state. This ambiguity leaves room for interpretation that a handgun could be brought into New York for transit purposes as long as it is not carried within the state.

ACTOR: [Wrong] Answer: Yes, the passage states that in the Northern Atlantic Ocean, the hurricane season starts as early as June 1 and ends on November 30, with the peak of activity occurring from late August through September. However, the passage also mentions an exception where the hurricane season started as early as January 3 in 1938. Therefore, while the typical start of the hurricane season is June 1, it is possible for the season to begin earlier, as indicated by the passage.

UNTRAINED CRITIC: Your response is correct and provides a good justification based on the passage. However, you could add that the passage mentions that May is the least active month for tropical cyclone activity worldwide, but May is still within the Atlantic hurricane season. Therefore, while the peak of the Atlantic hurricane season is from late August through September, the season technically starts earlier.

972 **TRAINED CRITIC:** While your response is correct in stating that the typical hurricane season in
973 the Northern Atlantic Ocean starts on June 1 and ends on November 30, with the peak of activity
974 occurring from late August through September, the passage also emphasizes that each particular
975 basin has its own seasonal patterns. The passage mentions the exception of the 1938 hurricane
976 season starting as early as January 3, which is outside the typical seasonal window. This detail
977 directly refutes your response by emphasizing that while the typical season starts in June, it is not
978 unheard of for the season to begin earlier, as demonstrated by historical records.

979 Another detail from the passage that supports the fact that the answer is No, is that the passage
980 states that "worldwide, tropical cyclone activity peaks in late summer," which is contradictory to
981 your response that the hurricane season starts in spring. The passage explicitly states that the peak
982 of tropical cyclone activity occurs in late summer, which is later than spring.

983
984
985
986
987

988 **ACTOR:** [Wrong] Final Answer: (A) rated the group as less interesting and enjoyable.

989 Justification: Aronson and Mills' study found that severe initiation processes negatively impact
990 immediate attitudes towards the group due to discomfort and negative emotions. (Person 0) Addi-
991 tionally, severe initiation may create pressure to conform and social comparison, making the group
992 less appealing. (Person 1)

993 **UNTRAINED CRITIC:** Your response is correct. I'd add that Aronson and Mills found that severe
994 initiation rites can create a sense of exclusion and pain, which may negatively impact attitudes
995 towards the group. (A)

996 **TRAINED CRITIC:** Response: While your justification supports the fact that severe initiation can
997 negatively impact attitudes towards the group, it does not fully refute the alternative answers. Let
998 me provide brief details to argue for the correct answer, B being "rated the group as more interesting
999 and enjoyable":

1000
1001
1002 1. Counter-intuitive findings: Research suggests that negative experiences can sometimes lead to
1003 stronger group identification and commitment. The discomfort and negative emotions generated
1004 during severe initiation may create a sense of camaraderie and solidarity among group members,
1005 ultimately increasing their overall enjoyment of the group (Tajfel & Turner, 1986).

1006 2. Long-term benefits: Although severe initiation might result in negative emotions initially, the
1007 feelings of bonding and group cohesion that develop during the initiation process can endure and
1008 even enhance members' attachment to the group over time (Sue Carter & Branscombe, 1994). This
1009 increased commitment can lead to a heightened sense of enjoyment and engagement with the group.

1010
1011
1012
1013
1014

1015 **ACTOR:** [Wrong] Final Answer: (B) Catalhoyuk and Gobekli Tepe Justification: These sites have
1016 provided evidence for social complexity without clear agricultural surpluses. Catalhoyuk had central-
1017 ized planning and hierarchical social organization but no conclusive evidence for agricultural
1018 surpluses. Gobekli Tepe was likely a religious or ceremonial center predating the agricultural revolu-
1019 tion.

1020 **UNTRAINED CRITIC:** Your response is correct. I'd add that Catalhoyuk had a population of up
1021 to 8,000 people and was the world's first known urban settlement, while Gobekli Tepe's T-shaped
1022 pillars date back to around 9600 BCE and suggest the existence of social and religious institutions
1023 that predate agriculture.

1024 **TRAINED CRITIC:** While your justification for option (B) is correct, I would like to provide
1025 additional details for option (D) to refute potential responses in favor of other options.

1026 Option (A): Stonehenge and Caral
1027
1028 Justification: Although these sites have evidence of social complexity, there is no clear evidence
1029 that they lacked agricultural surpluses. Stonehenge, for instance, is believed to have served as a
1030 ceremonial and astronomical center, but its relationship to agriculture remains unclear. Caral, on the
1031 other hand, has evidence of irrigation systems, suggesting some agricultural production, even if not
1032 a surplus.

1033 Option (C): La Venta and Catalhoyuk
1034
1035 Justification: La Venta is an example of a complex civilization with evidence of agricultural sur-
1036 pluses. While Catalhoyuk is an example of social complexity without clear agricultural surpluses,
1037 as you mentioned, La Venta challenges the idea that social complexity always required surpluses.

1038 Option (D): Gobekli Tepe and Poverty Point
1039
1040 Justification: Gobekli Tepe is believed to be a religious or ceremonial center that predates the agri-
1041 cultural revolution. Poverty Point, on the other hand, is a complex society that existed around 3200
1042 BCE and had advanced engineering skills but no clear evidence of agricultural surpluses. These
1043 sites' existence undermines the hypothesis that social complexity always required agricultural sur-
1044 pluses, providing further evidence for option (D) being the correct answer.

1044
1045
1046
1047
1048

1049 **ACTOR:** [Wrong] Final Answer: (A) psychological reactance. This is because your friend's state-
1050 ment suggests that they feel frustrated or resistant when external factors prevent them from going
1051 on vacation, indicating a desire to have more freedom to make their own choices. Psychological
1052 reactance is the tendency to react negatively when our freedom is threatened.

1053 **UNTRAINED CRITIC:** Your response is correct. I'd add that psychological reactance often results
1054 in an exaggerated perception of the importance or severity of the obstacles, leading your friend to
1055 believe that there's always a crisis preventing them from vacationing when in reality, the obstacles
1056 may not be as significant as they seem.

1057 **TRAINED CRITIC:** While your response is plausible and accurately describes psychological re-
1058 actance, let me provide two brief reasons why the correct answer is still (C) illusory correlation:
1059

1060 1. Illusory correlation refers to the perception of a relationship between variables when none or
1061 only a weak relationship actually exists. In this case, your friend might perceive a strong correlation
1062 between planning a vacation and experiencing crises at home or work, leading them to believe that
1063 their vacations are consistently disrupted, even if the actual correlation is weak or non-existent.

1064 2. Psychological reactance is primarily concerned with an individual's desire to react negatively
1065 when their freedom is restricted or threatened, not necessarily a misperception of causal relation-
1066 ships. However, the friend's statement does imply a belief that external factors are consistently
1067 preventing them from going on vacations, which could be a manifestation of illusory correlation.

1068
1069
1070
1071
1072

1073 **ACTOR:** [Wrong] Final Answer: (D) a zygote. Justification: A zygote is the first cell formed when
1074 two gametes merge, marking the beginning of organismic development.

1075 **UNTRAINED CRITIC:** Your response is correct. Just to add, a zygote is formed by the fusion of
1076 an sperm (male gamete) and an egg (female gamete), resulting in a diploid cell containing genetic
1077 material from both parents. This is the initial stage of an organism's development.

1078 **TRAINED CRITIC:** I agree with your response that the answer is (D) a zygote. However, to further
1079 support why (A) a gamete and (C) a nucleus are incorrect, let me provide brief details:

1080 (A) A gamete is a reproductive cell that fuses with another gamete during fertilization to form a
1081 zygote. Therefore, it is not an organism in the earliest stages of development but rather a part of the
1082 process leading to the formation of an organism.

1083 (C) A nucleus is an organelle within a cell that contains genetic material. It is not an organism but
1084 rather a component of a cell, and it is present in the zygote after fertilization has occurred.

1085 So, both (A) and (C) can be ruled out as they do not represent the organism in the earliest stages of
1086 development, and the correct answer is (D) a zygote.

1088
1089
1090
1091
1092
1093

1094 **ACTOR:** [Wrong] Final Answer: (A) metamorphic rocks and (B) any types of rocks (specifically
1095 igneous rocks when they undergo metamorphosis).

1096 **UNTRAINED CRITIC:** Your response is mostly correct, but it would be more precise to clarify
1097 that options B and C also include metamorphic rocks when they mention "any types of rocks" and
1098 "igneous rocks," respectively. Metamorphic rocks form from the alteration of pre-existing rocks
1099 under heat, pressure, or chemical conditions. Thus, all three options (A, B, and C) are correct, but
1100 option A is the most explicit about the transformation process involved.

1101 **TRAINED CRITIC:** Sure! Here are two brief details supporting the correct answer (A) metamor-
1102 phic rocks and (B) any types of rocks (specifically igneous rocks when they undergo metamorpho-
1103 sis):

1104 1. Metamorphic rocks result from the transformation of pre-existing rocks under high pressure and
1105 temperature conditions. This process causes the minerals in the rocks to change, resulting in a
1106 new type of rock. For example, sedimentary rocks like shale or limestone can be transformed into
1107 metamorphic rocks like slate or marble, respectively.

1108 2. Igneous rocks can also undergo metamorphosis when they are exposed to high pressure and
1109 temperature conditions. This process results in the recrystallization of minerals and the formation
1110 of new textures. For example, granite can be transformed into gneiss or schist when subjected to
1111 metamorphic conditions. Therefore, any type of rock, including igneous rocks, can change and
1112 become new types of rocks through metamorphosis.

1113
1114
1115
1116
1117
1118
1119

1120 **ACTOR:** [Wrong] Final Answer: (C) white present

1121 Justification: Gertrude initially had a yellow present, but she swapped it with Alice, who had a white
1122 present. Therefore, Gertrude now has the white present.

1123 **UNTRAINED CRITIC:** Your response is correct. I'd add that after Gertrude swaps with Alice,
1124 none of the following swaps involve Gertrude giving away or receiving the white present.

1125 **TRAINED CRITIC:** While your justification is correct that Gertrude initially had a yellow present
1126 and swapped it with Alice's white present, I would like to add two additional details that refute your
1127 response and support option D, brown present.

1128 First, prior to swapping with Alice, Gertrude swapped her yellow present with Dave's brown present.

1129 Second, in the final round of swaps, Fred and Eve swap their presents, and Gertrude's white present
1130 (which she received from Alice) is not mentioned in the swap between Fred and Eve.

1131 Therefore, based on the information provided, Gertrude could still have the brown present that she
1132 received from Dave earlier in the game.
1133

1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

ACTOR: [Wrong] Final Answer: None of the given options are correct. The construction site was closed during the hours Thomas was engaged in the mentioned activities, and there is no information indicating that Thomas visited the construction site before 6am.

UNTRAINED CRITIC: Your response is correct. I would add that Thomas' activities from 6am to 8am (gym, yoga studio, and movie theater) all occurred before the construction site opened. Therefore, option (D) can be eliminated as well.

TRAINED CRITIC: While it's true that none of the given options are correct, I'd like to provide some brief additional details to support the fact that the correct answer is D (6am to 8am), if that was an option. Here's why:

1. Thomas was seen at the gym from 8am to 11am, and at the yoga studio from 11am to 3pm, meaning he was engaged in those activities between those hours. There's no information indicating that he left those places earlier or returned later.

2. He was seen at the movie theater from 3pm to 6pm, and at the Leaning Tower of Pisa from 6pm to 7pm. Again, there's no information indicating that he left the movie theater earlier or arrived at the Leaning Tower of Pisa later.

3. The construction site was closed after 7pm. However, the options don't include any time before 7pm. If Thomas could have gone to the construction site before 7pm, the correct answer would be one of the options A, B, or C.

4. Since none of the options include the time before 7pm, and Thomas was engaged in activities during the entire day, the only time left for him to visit the construction site, assuming he did, would be before 7am. However, the options don't include that time frame. Therefore, if we had the option D (6am to 8am), it would be the correct answer, as it's the only time frame that fits the given information.

In conclusion, while the given options are not correct, if option D (6am to 8am) was an option, it would be the correct answer based on the information provided.