
Multi-LLM Debate: Framework, Principals, and Interventions

Andrew Estornell
ByteDance Research
andrew.estornell@bytedance.com

Yang Liu
University of California, Santa Cruz
yangliu@ucsc.edu

Abstract

The flexible and generalized nature of large language models has allowed for their application in a wide array of language-based domains. Much like their human contemporaries, these models are capable of engaging in discussions and debates as a means of improving answer quality. We first take a theoretical approach to analyzing debate and provide a framework through which debate can be mathematically examined. Building on this framework, we provide several theoretical results for multi-agent debate. In particular, we demonstrate that similar model capabilities, or similar model responses, can result in static debate dynamics where the debate procedure simply converges to the majority opinion. When this majority opinion is the result of a common misconception (possibly ingrained in the models through shared training data) debate is likely to converge to answers associated with that common misconception. Using insights from our theoretical results, we then propose three interventions that improve the efficacy of debate. For each intervention, we provide theoretical results demonstrating how debate is improved. We also demonstrate that these interventions result in better performance on four common benchmark tasks.

1 Introduction

Large language models (LLMs) have demonstrated a remarkable ability to perform unseen tasks with high efficacy. This behavior, often referred to as emergent, allows LLMs to serve as general-purpose tools for a wide array of language-based functions. One such behavior of particular interest is the ability of LLMs to intake and process opinions from other models (or humans). As shown in several previous works, this ability allows LLMs to collaboratively solve tasks by engaging in *debate* Chan et al. [2023], Liang et al. [2023], Du et al. [2023]. For a given task, multi-agent debate operates by eliciting responses from each model, distributing those responses among the models, and then eliciting updated responses from each model.

In this work, we aim to explore the debate procedure, by first providing a theoretical framework through which debate can be better understood. This framework draws inspiration from Bayesian inference and in-context learning, showing that debate can be partially viewed as a special type of in-context learning. Through this framework, we then provide several theoretical insights into the debate procedure. In particular, we demonstrate the susceptibility of multi-agent debate to echo-chamber effects. These echo-chamber effects are especially consequential when they stem from a shared misconception between a majority of models, which can arise from circumstances such as highly correlated training data between each model.

We then leverage results from our theoretical framework to improve the efficacy of the debate procedure. In particular, we propose three *interventions* (modifications to the debate procedure). First, *diversity-pruning* which aims to maximize the information entropy in model responses at each round of debate; this intervention has the added benefit of reducing the severity of the echo chamber

effect. Second, *quality-pruning* which aims to maximize the relevance of each model’s response. We demonstrate that this intervention improves the likelihood that the debate procedure converges to correct answers. Lastly, *misconception-refutation* which directly identifies, and attempts to refute misconceptions in model responses. This intervention takes inspiration from works such as Robinson et al. [2022] which demonstrate that LLMs are often more skilled at *evaluating* answers, compared with directly providing answers. For each of our interventions, we provide theoretical results outlining the way in which each improves debate. We also conduct experiments on four common benchmarks demonstrating that these interventions improve debate efficacy in practice.

Our contributions 1) We propose a theoretical framework for multi-LLM debate that draws on connections from in-context learning and Bayesian inference. 2) We provide theoretical insights on several key principles of multi-LLM debate. 3) Using these insights, we design three debate interventions which result in consistent improvement to debate, across four language benchmarks (BoolQ, MMLU, MathQ, TruthfuQA) and three families of models (GPT, Lama, and Mistral).

2 Related Work

Our work is closely related to multi-agent debate, which focuses on iterative collaboration between agents to make a decision Chan et al. [2023], Liang et al. [2023], Du et al. [2023], Khan et al. [2024], Irving et al. [2018], Michael et al. [2023], Rasal [2024], Pham et al. [2023], Chang [2024b]. These works often focus on multi-agent debate in the context of question-answering tasks and aim to provide higher quality answers (compared to those of a single model) by engaging multiple models in discussion. The preliminary debate framework, proposed by Du et al. [2023], facilitates debate by first asking each model a question, and then iteratively re-asking agents that same question contextualized by the responses of all models in the previous round. Different variants of this procedure have been proposed: debate where models have different functionality Liang et al. [2023], round-robin style debate Chan et al. [2023], dynamically controlling the level of disagreement between agents in debate Chang [2024a], or using judges to assess the correctness of debaters Khan et al. [2024]. Other techniques for iteratively improving the quality of answers have also been proposed, e.g., chain-of-thought Wei et al. [2022], Kojima et al. [2022], self-consistency Wang et al. [2022], Singhal et al. [2023], and self-reflection Ren et al. [2023].

Similar to debate, there have been investigations into the use of different LLMs to engage with one-another Liu et al. [2023], Abdelnabi et al. [2023], Zhang et al. [2023], Li et al. [2023c], Park et al. [2023a], explain their rationale to others Wang et al. [2023a], or collaboratively engage in general tasks Li et al. [2023a], Wang et al. [2023b], Park et al. [2023b], Wu et al. [2023], Hong et al. [2023], Li et al. [2023d,b], Tsao and AILAB [2023]. While debate has shown promise in a wide range of domains, several works have also demonstrated that the debate process can be unstable and can lead to worse performance than using just a single model Wang et al. [2024], Smit et al. [2023].

Our work is also related to in-context learning and Bayesian inference. The former, Brown et al. [2020], Min et al. [2022], Lampinen et al. [2022] demonstrates that LLMs can perform unseen tasks when provided only a few examples of that task. Other works Xie et al. [2021], Jiang [2023] have shown a connection between in-context learning and Bayesian inference; the additional examples provided to the model can be viewed as updates to the model’s posterior distribution over tokens.

3 Preliminaries

Debate Let \mathbf{x} be a given question, with associated answer \mathbf{y} , for example $\mathbf{x} = \text{“What color is the sky?”}$ and $\mathbf{y} = \text{“Blue”}$. Following the debate procedure proposed by Du et al. [2023], a collection of n LLMs (also referred to as agents) collaborate to infer the correct answer \mathbf{y} by iteratively engaging in discussion over T rounds, as described next:

- At round $t = 0$ each agent i observes task \mathbf{x} , then provides response $\mathbf{z}_i^{(0)}$.
- At rounds $t > 0$ each agent i observes task \mathbf{x} and the outputs of the n agents at the previous timestep $Z^{(t-1)} = (\mathbf{z}_1^{(t-1)}, \dots, \mathbf{z}_n^{(t-1)})$, then provides response $\mathbf{z}_i^{(t)}$.
- The debate process ends if $t = T$ or if agents reach a consensus.

To measure if consensus is reached a function a extracts an *answer*¹ from a given response \mathbf{z} . Suppose $\mathbf{z} = \text{"During the day, the sky is blue"}$, then $a(\mathbf{z}) = \text{"Blue"}$. At round t , the probability that agent i provides response $\mathbf{z}_i^{(t+1)}$, is given by

$$\mathbb{P}_{\text{model}} \left(\underbrace{\mathbf{z}_i^{(t+1)}}_{\text{updated model response}} \mid \underbrace{\mathbf{x}}_{\text{task}}, \underbrace{Z^{(t)} = (\mathbf{z}_1^{(t)}, \dots, \mathbf{z}_n^{(t)})}_{\text{all model responses from previous round}}, \underbrace{\phi_i}_{\text{model parameters}} \right) \quad (1)$$

where ϕ_i captures model hyperparameters (such as training data, architecture, etc.). Both the input $(Z^{(t)}, \mathbf{x})$ as well as the hyperparameters ϕ_i , ultimately influence the output $\mathbf{z}_i^{(t+1)}$. Note that on each round, all agents observe the *same* input, namely $(Z^{(t)}, \mathbf{x})$. Thus, differences in agent output $\mathbf{z}_i^{(t+1)}$ are determined by both the stochastic nature of output generations, and the unique parameters ϕ_i of each model. For notational convenience, we drop the subscript in $\mathbb{P}_{\text{model}}$ when the parameters ϕ_i are given, and simply write $\mathbb{P}(\cdot | \phi_i)$.

The key distinction between our approach and "vanilla" debate, is that we will leverage latent concepts (discussed next) to modify the responses in $Z^{(t)}$ in between each round of debate.

Latent Concepts Central to our investigation is the idea of *latent concepts* in language generation. As outlined by Xie et al. [2021], Jiang [2023] latent concepts capture the idea that language is not generated at random (either by humans or by models). Rather, when generating language, we first have an idea or an intention form in our minds; we then select words that we believe will convey that underlying idea or intention. More formally, let Θ be a *latent concept space* and let $\theta \in \Theta$ be a *concept*. Following Xie et al. [2021], tasks \mathbf{x} , and their associated answer \mathbf{y} are generated by first selecting a vector of latent concept $\theta \in \Theta$ and then sampling $(\mathbf{x}, \mathbf{y}) \sim D(\theta)$, where D is some distribution mapping concepts to task-answer pairs. Similarly, when providing responses, models will observe \mathbf{x} , and infer the latent concept θ , or more generally a distribution over the latent concept space, and then generate a response according to those inferred concepts, i.e., the model’s generation probability in Equation 1 can be expressed as

$$\mathbb{P}(\mathbf{z}_i^{(t+1)} | \mathbf{x}, Z^{(t)}, \phi_i) = \prod_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \mathbf{x}, Z^{(t)}, \phi_i) \mathbb{P}(\theta | \mathbf{x}, Z^{(t)}, \phi_i) \quad (2)$$

Note that the above holds by the law of total probability for any choice of latent concept space.

To provide a more concrete example of latent concepts, consider the question-answering task in the BoolQ dataset. One of the questions in this dataset is "Did Abraham Lincoln write the letter in the film Saving Private Ryan?" to which the correct answer is "Yes". The latent concept, in this case, corresponds to the actual scene in the movie where the Bixby letter (written by Lincoln) is read to a group of soldiers. Just as in our case, first a concept θ is drawn, e.g., the film is made; then from the film, a string \mathbf{x} is sampled, i.e., the previous question about the film.

For another example of latent concepts, we can think of arithmetic calculations such as multiplication. When we wish to express multiplication through language, we may write something like "4 * 4". The latent concepts behind this string are the mechanisms of multiplication (e.g., multiplication is just iterative addition, and addition itself is simply increasing the value of a number by one iteratively). These examples are intended to be easily digestible. However, latent concepts can also be significantly more abstract, such as a vector in some unknown embedding space.

4 A Theoretical Formulation of Multi-Agent Debate

We begin by providing a theoretical formulation of multi-agent debate. This formulation will provide key insights into the inner workings of the debate procedure, which we will use to improve debate.

The key behind our framework is to use the idea of latent concepts and expansion of each model’s generation probability (Equation 2) in order to better understand debate. Prior to presenting our theoretical framework, we first state an important assumption.

Assumption 4.1. For a given latent concept space Θ , the probability of generating response $\mathbf{z}_i^{(t+1)}$ is conditionally independent of both the responses $Z^{(t)}$ and the task \mathbf{x} , given concept $\theta \in \Theta$ and model parameters ϕ_i , i.e., $\mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \mathbf{x}, Z^{(t)}, \phi_i) = \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \phi_i)$.

¹In practice, a can be a regular-expression checker or an LLM-based judge such as Liang et al. [2023]

Assumption 4.1 can be interpreted as saying that a model’s generation \mathbf{z}_i , is uniquely determined by the model’s parameters ϕ_i and the concepts identified by a model, namely θ . In the case of encoder-decoder-based models, one can conceptualize the joint between ϕ and θ as corresponding to the embedding produced by the encoder. With this embedding in hand, the original input $(\mathbf{x}, Z^{(t)})$ no longer influences the model’s output, rather the embedding and model parameters will uniquely determine the model’s output.

Next, we derive the following lemma which will be useful in examining the way that model responses evolve debate rounds.

Lemma 4.2. *The generation of model i at time $t + 1$ can be expressed as,*

$$\mathbb{P}(\mathbf{z}_i^{(t+1)} | Z^{(t)}, \mathbf{x}, \phi_i) \propto \sum_{\theta \in \Theta} \underbrace{\mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i)}_{\text{generation without other agents}} \underbrace{\prod_{j=1}^n \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i)}_{\text{skew caused by other agents}}$$

The significance of this lemma is that we are able to express the probability of generating a given response $\mathbf{z}_i^{(t+1)}$ *with* the other model responses $Z^{(t)}$ in terms of the probability of generating $\mathbf{z}_i^{(t+1)}$ *without* the other model responses and a skew term caused by those model responses. Note that,

$$\mathbb{P}(\mathbf{z}_i^{(t+1)} | \mathbf{x}, \phi_i) \propto \sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i)$$

Thus, we can think of $\mathbb{P}(\mathbf{z}_i^{(t+1)} | Z^{(t)}, \mathbf{x}, \phi_i)$ as a weighted version of $\mathbb{P}(\mathbf{z}_i^{(t+1)} | \mathbf{x}, \phi_i)$, where the weights are given by the skew term $\prod_{j=1}^n \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i)$.

Debate and In-Context Learning To help conceptualize the role of latent concepts in debate, we discuss the work of Xie et al. [2021], which uses Bayesian inference over latent concepts to understand in-context learning. There are natural connections between in-context learning and multi-agent debate. In-context learning works as follows: given a task \mathbf{x} and a model f , select several examples of task-answers pairs $(\mathbf{x}_1, y_1) \dots (\mathbf{x}_m, y_m)$ which are *similar* to \mathbf{x} . Then prompt the model f for an answer to task \mathbf{x} , given examples (\mathbf{x}_j, y_j) , i.e. $\mathbf{z} = f(\mathbf{x} | (\mathbf{x}_1, y_1) \dots (\mathbf{x}_m, y_m))$. A key result of Xie et al. [2021] is that latent concepts in the examples (\mathbf{x}_j, y_j) , particularly concepts shared between many examples, influence the model’s answer \mathbf{z} . For any concept where $\mathbb{P}(\theta | (\mathbf{x}_1, y_1) \dots (\mathbf{x}_m, y_m))$ is large relative to other concepts (i.e., there is a shared concept θ between the examples), the model is more likely to give response \mathbf{z} which also share that concept. Model responses at the previous round $Z^{(t)}$ serve a similar function to the examples of in-context learning. The model’s updated response at time $t + 1$, namely $\mathbf{z}_i^{(t+1)}$, is influenced by concepts shared between the responses in $Z^{(t)}$. The skew term in Lemma 4.2 provides a glimpse of how latent concepts conveyed by $Z^{(t)}$ will influence the generation of $\mathbf{z}_i^{(t+1)}$, namely that $\prod_{j=1}^n \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i)$ reweighs the model’s generation.

4.1 Debate Objective

Through this perspective of debate we can more effectively design debate procedures by leveraging the concept space Θ . To do this, we will first formulate debate as an optimization problem where the *skew term*, described in Lemma 4.2, corresponds to the optimization variables. For a given task \mathbf{x} and answer y , each round of debate can be formulated as the following optimization problem.

$$\arg \max_{Z^{(t)}} \sum_{i=0}^n \mathbb{P}(a(\mathbf{z}_i^{(t+1)}) = y | Z^{(t)}, \phi_i)$$

At time t we aim to craft responses $Z^{(t)}$ such that they maximize the probability of providing the correct answer at the next time step. Expanding this objective over the latent concept space Θ , yields

$$\arg \max_{Z^{(t)}} \sum_{i=1}^n \sum_{\theta \in \Theta} \left(\mathbb{P}(a(\mathbf{z}_i^{(t+1)}) = y | \theta, \phi_i) \mathbb{P}(\theta | \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \prod_{j=1}^n \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \right) \quad (3)$$

The key challenges with directly optimizing this objective are: 1) the true concept θ^* from which \mathbf{x} and y originate, as well as the relationship between $\mathbf{z}_j^{(t)}$ and the underlying concepts, is unknown, 2) the responses in $Z^{(t)}$ are natural language. However, will allow us to design several approaches within the concept space to better optimize debate. To motivate these approaches, we first need to make several observations about the debate procedure as a whole.

5 Debate Principals

In this section, we discuss factors that affect the efficacy of LLMs debate. In particular, we look at the role of information diversity, both in terms of the diversity of responses in $Z^{(t)}$ as well as diversity in model capabilities. We see that a lack of diversity in either aspect is detrimental to the debate process. Lastly, we study a particular type of homogeneity in debate, namely shared misconceptions in which a large portion of models all share a similar erroneous belief about the task x .

5.1 Information Diversity

We begin by examining how the debate procedure is affected by both the diversity of model abilities and the diversity of model responses. Homogeneity in either ability or responses will bias the debate procedure towards certain latent concepts.

Similar Model Capabilities Suppose the debate process is conducted with only one type of model (in effect n copies of the same model). That is, $\phi_i \equiv \phi$ for all $i \in [n]$. Then, as the number of agents increases, the debate procedure is more greatly impacted by the echo chamber effect, i.e., the probability that a round of debate results in a change to the most likely concept, perceived by agents, approaches 0. That is, a greater number of similar agents results in static debate dynamics, in essence defeating the purpose of debate.

Theorem 5.1. *Suppose all n agents have identical configurations ($\phi_i \equiv \phi$ for all i). For round $t > 0$ let $\theta^{(t)} = (\arg \max_{\theta \in \Theta} \mathbb{P}(\theta | \mathbf{x}, Z^{(t)}, \phi))$ and $\theta^{(t+1)} = (\arg \max_{\theta \in \Theta} \mathbb{P}(\theta | \mathbf{x}, Z^{(t+1)}, \phi))$, i.e., $\theta^{(t)}$ and $\theta^{(t+1)}$ are the concepts most likely to be inferred by a model with parameters ϕ when given task \mathbf{x} and responses $Z^{(t)}, Z^{(t+1)}$ respectively. Then $\mathbb{P}(\theta^{(t)} = \theta^{(t+1)}) \rightarrow 1$ as $n \rightarrow \infty$.*

We defer a full proof to the Supplement, Section A. Theorem 5.1 implies that when debate is conducted with multiple copies of the same model (or very similar models), increasing the number of models results in debate centering on a single (unchanging) concept, rather than a balanced distribution over multiple concepts.

Similar Model Opinions Next, we examine how similar responses impact the collaboration process. At time t suppose that there are n responses $Z^{(t)}$ and at least m of those responses are *similar*, i.e., there exists some concept θ' such that $\theta' = \arg \max_{\theta \in \Theta} \mathbb{P}(\theta | \mathbf{z}_j^{(t)}, \phi_i)$ for all $j \leq m$. That is, each of the m responses has a shared “most likely” concept when viewed by a model with parameters ϕ_i .

Theorem 5.2. *Suppose that $Z^{(t)}$ contains at least m responses with the property that $\theta' = \arg \max_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_i^{(t)} | \theta, \phi_i)$. Then, as $m \rightarrow \infty$ the model’s generation at the next round $(t+1)$ becomes uniquely determined by a single concept θ' i.e. $\frac{\mathbb{P}(\mathbf{z}_{(i,1)}^{(t+1)} | Z^{(t)}, \mathbf{x}, \phi_i)}{\mathbb{P}(\mathbf{z}_{(i,2)}^{(t+1)} | Z^{(t)}, \mathbf{x}, \phi_i)} \rightarrow \frac{\mathbb{P}(\mathbf{z}_{(i,1)}^{(t+1)} | \theta', \phi_i)}{\mathbb{P}(\mathbf{z}_{(i,2)}^{(t+1)} | \theta', \phi_i)}$ for all response pairs $\mathbf{z}_{(i,1)}^{(t+1)}, \mathbf{z}_{(i,2)}^{(t+1)}$.*

We defer a full proof to the Supplement Section A. Theorem 5.2 indicates the susceptibility that LLM debate has towards tyranny of the majority. If a large number of models provide similar responses to a task \mathbf{x} , then those repeated answers will drown out the single provided by the other models’ responses, as well as the task \mathbf{x} itself. In Section 7 we demonstrate that this occurs in practice.

5.2 Shared Misconceptions

Next, we study a particular type of homogeneity in model capabilities and responses, namely shared misconceptions. When a common misconception is shared among the models, debate is less effective

and is likely to converge to erroneous concepts associated with the shared misconception. We now formalize the notion of misconceptions.

Definition 5.3. (Misconception): For a given concept θ^* , a model with parameters i is said to have a *misconception* regarding θ^* if there exists another concept θ' s.t.,

$$\mathbb{P}_{\mathbf{x} \sim D(\theta^*)} [\mathbb{P}(\mathbf{x} | \theta', \phi_i) > \mathbb{P}(\mathbf{x} | \theta^*, \phi_i)] > 1/2$$

That is, for tasks generated from the concept θ^* , the model believes that the erroneous concept θ' explains more than half of the tasks better than the true concept θ^* .

There is said to be a *shared misconception* if m of agents have a misconception and share the same erroneous concept θ' . When the models share a common misconception the responses produced by those models are biased towards the erroneous concept θ' .

Theorem 5.4. *Let the true concept be θ^* and suppose that m of the n agents have a shared misconception for erroneous concept θ' . Then, task and answer $(\mathbf{x}, y) \sim D(\theta^*)$ expected average correctness of the debate procedure at the final round T is monotonically decreasing with m , i.e., $\frac{1}{n} \sum_i^n \mathbb{P}(a(\mathbf{z}_i^{(T)}) = y)$ is decreasing with m .*

We defer the full proof to the supplement Section A. It should be noted that the phenomenon of converging to erroneous concepts is unlikely to be mitigated by adding more models. When the misconceptions of one model are formed through training data, it is likely that other models will possess the same misconception unless specifically trained to avoid such errors due to the high correlation in training data between models.

6 Interventions

In this section, we discuss several modifications to the debate procedure, referred to as *interventions*. We break our interventions into two categories: **Pruning** which focuses on choosing which responses to keep in $Z^{(t)}$, and **Modifying** which focuses on changing or editing the responses $Z^{(t)}$.

6.1 Pruning Interventions

At round t of debate, running interventions work by selecting only a subset of responses $Z'^{(t)}$ from $Z^{(t)}$ before starting the next round $t + 1$. When using a pruning intervention, the models at round $t + 1$ will see only the pruned response set $Z'^{(t)}$, rather than the full response set $Z^{(t)}$.

Diversity Pruning Let KL represent Kullback–Leibler divergence. The diversity pruning intervention selects k of the n responses in $Z^{(t)}$ which maximizes information entropy, i.e.,

$$Z'^{(t)} = \operatorname{argmax}_{Y \subset Y^{(t)}, \mathbf{z}_i, \mathbf{z}_j \in Z} \sum \text{KL}(D(\theta | \mathbf{z}_i), D(\theta | \mathbf{z}_j)) \quad \text{s.t. } |Z| = k$$

Quality Pruning Quality pruning aims to select the k responses in $Z^{(t)}$ with the highest similarity to the task x . Similar to diversity pruning, quality pruning selects k of the n responses at time t . Rather than selecting for diversity, quality pruning aims to select the k highest question responses. This is done by selecting the k responses which maximize

$$Z'^{(t)} = \operatorname{argmin}_{Z \subset Z^{(t)}, \mathbf{z}_i \in Z} \sum \text{KL}(D(\theta | \mathbf{x}), D(\theta | \mathbf{z}_i)) \quad \text{s.t. } |Z| = k$$

In practice computing $\text{KL}(D(\theta | \mathbf{x}), D(\theta | \mathbf{z}_i))$ or $\text{KL}(D(\theta | \mathbf{z}_i), D(\theta | \mathbf{z}_j))$ is intractable. However, sentence embedding can be used as a proxy for these values. Section C discusses this in further detail.

Next, we show that when models have a shared misconception, diversity pruning decreases the likelihood that the debate procedure will converge to the erroneous concept corresponding to the shared misconception.

Theorem 6.1. *Let the true concept be θ^* and suppose that at least $n/2$ agents have a shared misconception for erroneous concept θ' . Then diversity pruning decreases the probability that debate converges to an answer y' which is sourced from the erroneous concept θ' , i.e. $y' \sim D(\theta')$.*

We defer the full proof to the Supplement, Section A.

Theorem 6.2. *For a given task-answer pair (\mathbf{x}, y) quality pruning increases the probability that debate converges to the correct answer, i.e. let $Z^{(t)}$ be the set of all responses at time t and $Z'^{(t)}$ be the result of quality pruning, then $\sum_{i=1}^n \mathbb{P}(a(\mathbf{z}_i^{(t+1)}) = y | \mathbf{x}, Z'^{(t)}, \phi_i) > \sum_{i=1}^n \mathbb{P}(a(\mathbf{z}_i^{(t+1)}) = y | \mathbf{x}, Z^{(t)}, \phi_i)$.*

We defer the full proof to the Supplement, Section A.

Remark 6.3. As shown by Theorems 6.1 and 6.2, diversity pruning decreases the probability that debate converges to incorrect answers sourced from a particular concept, while quality pruning increases the probability that debate converges to a correct answer sourced from the true concept. Both interventions can be used simultaneously to guide the debate procedure more effectively away from wrong answers and towards correct answers.

6.2 Modification Interventions

Misconception Refutation In addition to selecting which responses in $Z^{(t)}$ will be used in the next round of debate, we can also modify the responses in $Z^{(t)}$. Misconception refutation aims to do precisely this by updating response $\mathbf{z}_j^{(t)}$ to be more relevant to the task \mathbf{x} .

$$\mathbf{z}_j^* = \arg \min_{\mathbf{z}} \text{KL}(D(\boldsymbol{\theta} | \mathbf{x}), D(\boldsymbol{\theta} | \mathbf{z})) - \text{KL}(D(\boldsymbol{\theta} | \mathbf{z}_j^{(t)}), D(\boldsymbol{\theta} | \mathbf{z}))$$

Similar to Diversity Pruning and Quality Pruning, the distributions in the above objective are intractable in practice. As such, we use a proxy to update each response $\mathbf{z}_j^{(t)}$, specifically produce \mathbf{z}_j^* by having an LLM minimally modify the given response $\mathbf{z}_j^{(t)}$. The model is first prompted for a list of misconceptions and errors identified in the response. Given the list of misconceptions, the model is asked for both a refutation of the misconception and a corrected version of the response. For more details, see Section C of the Supplement.

Theorem 6.4. *For task-answer pair (\mathbf{x}, y) , misconception refutation increases the probability of debate converging to the correct answer, i.e. let $Z^{(t)}, Z^{*(t)}$ be the responses before and after misconception refutation, then $\sum_{i=1}^n \mathbb{P}(a(\mathbf{z}_i^{(t+1)}) = y | \mathbf{x}, Z^{*(t)}, \phi_i) > \sum_{i=1}^n \mathbb{P}(a(\mathbf{z}_i^{(t+1)}) = y | \mathbf{x}, Z^{(t)}, \phi_i)$.*

7 Experiments

Experimental Design We conduct experiments on four common language model benchmarks. **BoolQ** Clark et al. [2019], which consists of 3, 270 yes-no questions, **MMLU** Hendrycks et al. [2020] which consists of 13, 869 multiple-choice questions (we use the 3, 406 high-school-level questions), **TruthfulQA** Lin et al. [2021] which consists of 817 open-ended questions, and **MathQ** which consists of 3, 000 arithmetic questions of the form $a \cdot b \cdot c + d \cdot e \cdot f$. In the BoolQ, MMLU, MathQ, datasets model correctness is measured through regular expression matching. In the TruthfulQA dataset, model correctness is measured via an LLM judge (we use GPT-4 as the judge in all experiments)

We use four LLMs of increasing capability, **GPT-3.5** (GPT-3.5 Turbo) OpenAI [2022], **Llama-2** (Llama-2 7B Chat) Touvron et al. [2023], **Llama-3** (Llama-3 8B Instruct) Meta AI [2024], and **Mistral** (Mistral 7B Instruct v0.2) Jiang et al. [2023]. For sentence embeddings (which serve as a proxy of the latent concepts Θ), we use sentence embeddings from **ADA-2** OpenAI [2022]. We compare a combination of our three interventions **Ours** (see Algorithm 1 full details) with the debate paradigm of Du et al. [2023] (Society of Minds) **SoM**.

We begin by making several empirical observations about the multi-agent debate process.

Tyranny of the Majority First, we examine the susceptibility of models towards agreement with the majority opinion. That is, how likely are models to give a specific answer at round $t + 1$ when m of the models provided that specific answer at round the previous round (round t)? For example, in BoolQ suppose the specific answer is "Yes", then we want to know: how likely is a model to give a "Yes"-answer at round $t + 1$ if that model observes m "Yes" answers at round t .

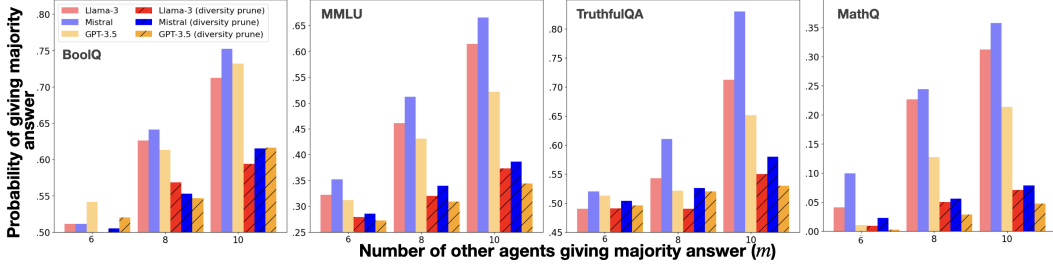


Figure 1: Probability that each model echoes the majority answer at round $t = 11$, as the number of responses at time $t = 0$ gives that majority answer (debate between 12 models are used).

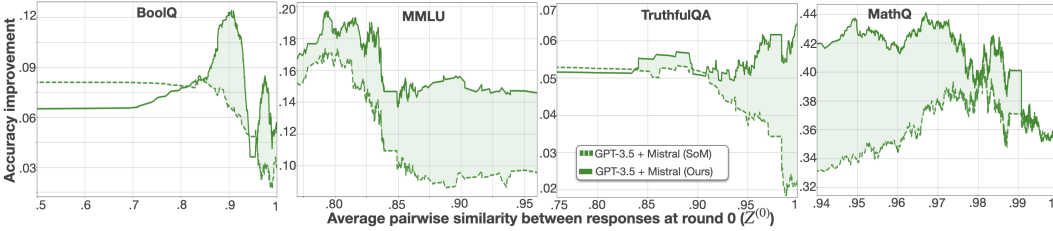


Figure 2: Average accuracy improvement as a function of response diversity at round 0 of debate.

To measure this, we first select a random target answer, e.g., “Yes”, and then prompt m of the models (out of 11) to provide responses “Yes”-answers² (while the other $11-m$ models are prompted to provide a different randomly selected answer). These 11 responses make up $Z^{(t)}$, we then test each model’s likelihood of providing the target at round $t + 1$ when observing $Z^{(t)}$ before diversity pruning (solid) and after diversity running (hatched).

In Figure 1, we see models are susceptible to echo chamber effects (this phenomenon is predicted by Theorem 5.1). The likelihood of providing the majority answer increases when $Z^{(t)}$ contains more instances of the majority answer (i.e., as m increases). Figure 1 also demonstrates that diversity pruning (with $k = 5$) reduces this echo chamber effect. See the Supplement for details.

Diversity of Opinions Next, we examine the effectiveness of SoM and our method as a function of opinion diversity. Figure 2 shows the average accuracy improvement of SoM (dashed) and our method (solid) over single model performance (i.e., average performance at round $t = 0$), as a function of the similarity between all responses at round $t = 0$ of debate (measured via pairwise cosine similarity). We see that for BoolQ, MMLU, and TruthfulQA, SoM is less effective when the similarity between responses increases. This observation is predicted by Theorems 5.1 and 5.2, which show that debate, without interventions, is less effective when model responses are too similar.

We see that our method’s improvement compared to SoM is greatest when model opinions are more similar (cosine similarity close to 1). Note that the MathQ benchmark, where responses consist primarily of arithmetic, serves as a counter-example to these observations. This is due to the fact that sentence embedding of any two arithmetic expansions will be similar, regardless of their *true* similarity; as such, the cosine similarity between embedding is less meaningful on this benchmark.

Debate Interventions Now, we examine the effectiveness of a combination of our three debate interventions (see Algorithm 1 for full details of how the interventions are combined). We begin with a per-round performance of our method and SoM, as shown in Figure 3. We see that typically, the advantage of our method over debate arises in the later rounds of debate. Next, in Table 1, we present a full set of results for single models, SoM, and a combination of our three interventions. In all cases, our method is either competitive with, or superior to, SoM.

²Any model *not* providing the target answer is re-prompted until the target answer is provided.

	Single	SoM	Ours	Single	SoM	Ours		
			BoolQ			MMLU		
6 × GPT-3.5	.80±.014	.84±.012	.85±.012	.73±.014	.74±.016	.79±.014		
6 × Llama-3	.76±.014	.78±.013	.78±.013	.67±.016	.70±.015	.75±.014		
6 × Llama-2	.67±.017	.68±.017	.73±.016	.41±.017	.47±.018	.52±0.18		
6 × Mistral	.80±.014	.82±.013	.85±.012	.66±.016	.65±.016	.66±.016		
3 × GPT-3.5 + 3 × Llama-3	-	.82±.014	.84±.013	-	.73±.016	.78±.016		
3 × GPT-3.5 + 3 × Mistral	-	.83±.013	.87±.012	-	.69±.017	.72±.016		
3 × Llama-3 + 3 × Mistral	-	.80±.014	.80±.014	-	.69±.017	.74±.016		
			TruthfulQA			Math		
6 × GPT-3.5	.61±.033	.63±.032	.69±.030	.53±.035	.88±.016	.93±.01		
6 × Llama-2	.47±.034	.52±0.35	.55±.034	.11±.013	.13±.014	.19±.015		
6 × Llama-3	.53±.035	.55±.032	.55±.032	.25±.016	.33±.017	.48±.018		
6 × Mistral	.48±.034	.51±.035	.53±.034	.13±.013	.19±.014	.18±.014		
3 × GPT-3.5 + 3 × Llama-3	-	.56±.035	.62±.031	-	.76±.015	.82±.014		
3 × GPT-3.5 + 3 × Mistral	-	.52±.035	.56±.035	-	.56±.018	.68±.017		
3 × Llama-3 + 3 × Mistral	-	.49±.036	.53±.035	-	.22±.015	.23±.015		

Table 1: Accuracy of a solo model, debate, and our debate interventions: 10 rounds, 6 models.

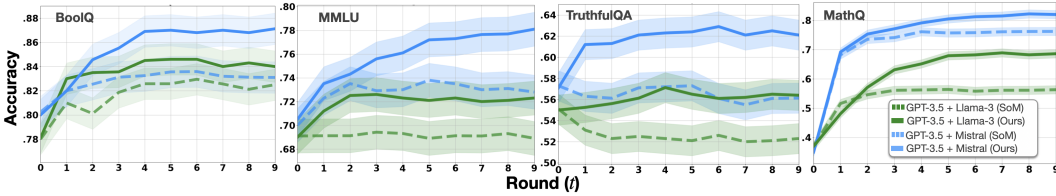


Figure 3: Accuracy per round, our method and SoM when combining GPT-3.5 with Llama-3 or Mistral.

In addition to providing results for the combination of our interventions, we also investigate the effectiveness of each intervention applied individually (see Table 3 of the Supplement). These results indicate that our method is most successful when applying all three interventions simultaneously. In fact, some interventions can be detrimental to the debate process when applied in isolation. This is expected as each intervention is inherently designed to be complementary.

8 Limitations

While we aim to address some of the fundamental issues of multi-LLM debate, such as tyranny of the majority, there are several factors that need to be considered when adopting our framework. Firstly, our theoretical results leverage a latent concept space, which may not be accessible in practice, necessitating the use of proxies such as sentence embeddings. Reliance on proxies is particularly consequential for quality and diversity pruning; these interventions are less effective in domains where sentence embeddings are less meaningful, e.g., arithmetic questions. Additionally, our interventions can increase the inference time of the debate procedure. Increased inference time stems primarily from misconception refutation, as this intervention requires re-prompting each debater multiple times.

9 Conclusion

Multi-agent debate is an effective tool for improving the efficacy of LLM responses. However, debate is naturally susceptible to issues such as tyranny of the majority and shared misconceptions between models. By making use of our theoretical framework for debate, we are able to establish interventions for the debate procedure which help to alleviate these issues and improve the general performance of multi-agent debate. We saw that diversity pruning reduces the influence of similar responses. This is especially helpful in settings where the majority of agents provide incorrect responses that share a common error. A combination of all three interventions consistently leads to better debate.

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.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- 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. Llm collaborative intelligence: The path to artificial general intelligence. *SocraSynth.com*, 2024b.
- 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.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.
- Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- Hui Jiang. A latent space theory for emergent abilities in large language models. *arXiv preprint arXiv:2304.09960*, 2023.
- Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.
- Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*, 2022.

- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023a.
- Hua Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*, 2023b.
- Yang Li, Yangyang Yu, Haohang Li, Zhi Chen, and Khaldoun Khashanah. Tradinggpt: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance. *arXiv preprint arXiv:2309.03736*, 2023c.
- Yuan Li, Yixuan Zhang, and Lichao Sun. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*, 2023d.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.
- Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*, 2023.
- Meta AI. Meta llama 3. <https://ai.meta.com/blog/meta-llama-3/>, 2024.
- Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R Bowman. Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702*, 2023.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.
- OpenAI, 2022. URL <https://openai.com/blog/chatgpt/>.
- Jeongeon Park, Bryan Min, Xiaojuan Ma, and Juho Kim. Choicemates: Supporting unfamiliar online decision-making with multi-agent conversational interactions. *arXiv preprint arXiv:2310.01331*, 2023a.
- Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023b.
- Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A Plummer, Zhaoran Wang, and Hongxia Yang. Let models speak ciphers: Multiagent debate through embeddings. *arXiv preprint arXiv:2310.06272*, 2023.
- Sumedh Rasal. Llm harmony: Multi-agent communication for problem solving. *arXiv preprint arXiv:2401.01312*, 2024.
- Jie Ren, Yao Zhao, Tu Vu, Peter J Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models. *arXiv preprint arXiv:2312.09300*, 2023.
- Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353*, 2022.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

- Andries Smit, Paul Duckworth, Nathan Grinsztajn, Kale-ab Tessera, Thomas D Barrett, and Arnu Pretorius. Are we going mad? benchmarking multi-agent debate between language models for medical q&a. *arXiv preprint arXiv:2311.17371*, 2023.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Wen-Kwang Tsao and TrendMicro AILAB. Multi-agent reasoning with large language models for effective corporate planning. In *The 10th International Conf. on Computational Science and Computational Intelligence*, 2023.
- Boshi Wang, Xiang Yue, and Huan Sun. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881, 2023a.
- Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*, 2024.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*, 2023b.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.
- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.
- Jintian Zhang, Xin Xu, and Shumin Deng. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*, 2023.

Appendix

A Theoretical Results

Proof of Lemma 4.2. This result holds via marginalization of the posterior predictive distribution over the latent concepts Θ , namely

$$\begin{aligned} \mathbb{P}(\mathbf{z}_i^{(t+1)} | Z^{(t)}, \mathbf{x}, \phi_i) &= \sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, Z^{(t)}, \mathbf{x}, \phi_i) \mathbb{P}(\theta | Z^{(t)}, \mathbf{x}, \phi_i) \\ &= \sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \phi_i) \mathbb{P}(Z^{(t)}, \mathbf{x} | \theta, \phi_i) \frac{\mathbb{P}(\theta | \phi_i)}{\mathbb{P}(Z^{(t)}, \mathbf{x} | \phi_i)} \\ &\propto \sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i) \prod_{j=1}^n \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \end{aligned}$$

□

Proof of Theorem 5.1. Since each model has an identical configuration, i.e. $\phi_i = \phi_j$ for all $i, j \in [n]$, we simply refer to the configuration as ϕ . For the given task \mathbf{x} let $\theta^{*(t)}$ be the realization of θ at time step t . Let $Z^{(t)} = (\mathbf{z}_1^{(t)}, \dots, \mathbf{z}_n^{(t)})$, where each $\mathbf{z}_j^{(t)} \sim q(\mathbf{z} | \mathbf{x}, Z^{(t)}, \phi)$. Then the conditional density for each concept θ can be written as,

$$\begin{aligned} &\mathbb{P}(\theta | Z^{(t)}, \mathbf{x}, \phi) \\ &\propto \mathbb{P}(Z^{(t)}, \mathbf{x} | \theta, \phi) \mathbb{P}(\theta | \phi) \\ &= \mathbb{P}(\mathbf{x} | \theta, \phi) \mathbb{P}(\theta | \phi) \prod_{j=1}^n \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi) \end{aligned}$$

where the term $\mathbb{P}(\mathbf{x} | \theta, \phi) \mathbb{P}(\theta | \phi)$ is a constant with respect to the number of agents n . Since $\mathbb{P}(\theta | \phi) > 0$ for all $\theta \in \theta$, each $\mathbf{z}_j^{(t)}$ is an i.i.d. draw from $q(\mathbf{z} | \mathbf{x}, Z^{(t-1)}, \phi)$, then

$$\theta^{*(t)} = \lim_{n \rightarrow \infty} \left(\arg \max_{\theta \in \Theta} \mathbb{P}(\theta | Z^{(t)}, \mathbf{x}, \phi) \right)$$

Thus, as $n \rightarrow \infty$, all models predict the same concept, namely $\theta^{*(t)}$, at timestep t with probability 1. □

Proof of Theorem 5.2. Consider any two responses from agent i at time $t + 1$, namely $\mathbf{z}_{(i,1)}^{(t+1)}, \mathbf{z}_{(i,2)}^{(t+1)}$. When there are n duplicate messages, the ratio between the conditional generation probabilities of both responses can be written as

$$\begin{aligned} &\frac{\mathbb{P}(\mathbf{z}_{(i,1)}^{(t+1)} | Z^{(t)}, \mathbf{x}, \phi_i)}{\mathbb{P}(\mathbf{z}_{(i,2)}^{(t+1)} | Z^{(t)}, \mathbf{x}, \phi_i)} \\ &= \frac{\sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_{(i,1)}^{(t+1)} | \theta, \phi_i) \mathbb{P}(\theta | Z^{(t)}, \mathbf{x}, \phi_i)}{\sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_{(i,2)}^{(t+1)} | \theta, \phi_i) \mathbb{P}(\theta | Z^{(t)}, \mathbf{x}, \phi_i)} \\ &= \frac{\sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_{(i,1)}^{(t+1)} | \theta, \phi_i) \mathbb{P}(Z^{(t)}, \mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i)}{\sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_{(i,2)}^{(t+1)} | \theta, \phi_i) \mathbb{P}(Z^{(t)}, \mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i)} \quad (\text{factoring out common divisor } \mathbb{P}(Z^{(t)}, \mathbf{x} | \phi_i)) \\ &= \frac{\sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_{(i,1)}^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i) \left(\prod_{j=1}^k \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \right) \left(\mathbb{P}(\mathbf{z}'^{(t)} | \theta, \phi_i) \right)^m}{\sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_{(i,2)}^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i) \left(\prod_{j=1}^k \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \right) \left(\mathbb{P}(\mathbf{z}'^{(t)} | \theta, \phi_i) \right)^m} \end{aligned}$$

Let θ' be the concept which model i , with configuration ϕ_i , believes is most likely to have produced response \mathbf{z}' , i.e.,

$$\theta' = \arg \max_{\theta} \mathbb{P}(\mathbf{z}' | \theta, \phi_i)$$

Then the above ratio can be rewritten as

$$\begin{aligned} & \frac{\sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_{(i,1)}^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i) \left(\prod_{j=1}^k \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \right) \left(\mathbb{P}(\mathbf{z}'^{(t)} | \theta, \phi_i) \right)^m}{\sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_{(i,2)}^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i) \left(\prod_{j=1}^k \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \right) \left(\mathbb{P}(\mathbf{z}'^{(t)} | \theta, \phi_i) \right)^m} \frac{1}{\left(\mathbb{P}(\mathbf{z}'^{(t)} | \theta', \phi_i) \right)^m} \\ &= \frac{\sum_{\theta \in \Theta: \theta \neq \theta'} \mathbb{P}(\mathbf{z}_{(i,1)}^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i) \left(\prod_{j=1}^k \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \right) \left(\frac{\mathbb{P}(\mathbf{z}'^{(t)} | \theta, \phi_i)}{\mathbb{P}(\mathbf{z}'^{(t)} | \theta', \phi_i)} \right)^m}{\sum_{\theta \in \Theta: \theta \neq \theta'} \mathbb{P}(\mathbf{z}_{(i,2)}^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i) \left(\prod_{j=1}^k \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \right) \left(\frac{\mathbb{P}(\mathbf{z}'^{(t)} | \theta, \phi_i)}{\mathbb{P}(\mathbf{z}'^{(t)} | \theta', \phi_i)} \right)^m} \\ & \quad + \frac{\mathbb{P}(\mathbf{z}_{(i,1)}^{(t+1)} | \theta', \phi_i) \mathbb{P}(\mathbf{x} | \theta', \phi_i) \mathbb{P}(\theta' | \phi_i) \left(\prod_{j=1}^k \mathbb{P}(\mathbf{z}_j^{(t)} | \theta', \phi_i) \right)}{\mathbb{P}(\mathbf{z}_{(i,2)}^{(t+1)} | \theta', \phi_i) \mathbb{P}(\mathbf{x} | \theta', \phi_i) \mathbb{P}(\theta' | \phi_i) \left(\prod_{j=1}^k \mathbb{P}(\mathbf{z}_j^{(t)} | \theta', \phi_i) \right)} \end{aligned}$$

For each $\theta \neq \theta'$ we have $\frac{\mathbb{P}(\mathbf{z}' | \theta, \phi_i)}{\mathbb{P}(\mathbf{z}' | \theta', \phi_i)} < 1$. Therefore

$$\lim_{m \rightarrow \infty} \left(\frac{\mathbb{P}(\mathbf{z}' | \theta, \phi_i)}{\mathbb{P}(\mathbf{z}' | \theta', \phi_i)} \right)^m = 0$$

Note that the only summands which do not have an exponential dependency on m are those associated with concept θ' . Therefore, when examining the limit of the above ratio with respect to the number of repeated signals m , we get

$$\begin{aligned} & \lim_{m \rightarrow \infty} \frac{\sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_{(i,1)}^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i) \left(\prod_{j=1}^k \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \right) \left(\mathbb{P}(\mathbf{z}'^{(t)} | \theta, \phi_i) \right)^m}{\sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_{(i,2)}^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i) \left(\prod_{j=1}^k \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \right) \left(\mathbb{P}(\mathbf{z}'^{(t)} | \theta, \phi_i) \right)^m} \\ &= \frac{\mathbb{P}(\mathbf{z}_{(i,1)}^{(t+1)} | \theta', \phi_i) \mathbb{P}(\mathbf{x} | \theta', \phi_i) \mathbb{P}(\theta' | \phi_i)}{\mathbb{P}(\mathbf{z}_{(i,2)}^{(t+1)} | \theta', \phi_i) \mathbb{P}(\mathbf{x} | \theta', \phi_i) \mathbb{P}(\theta' | \phi_i)} \\ &= \frac{\mathbb{P}(\mathbf{z}_{(i,1)}^{(t+1)} | \theta', \phi_i)}{\mathbb{P}(\mathbf{z}_{(i,2)}^{(t+1)} | \theta', \phi_i)} \end{aligned}$$

Thus the relationship between any two conditional generation probabilities can be uniquely defined by θ' . \square

Proof of Theorem 5.4. First, we examine the models' outputs at the first rounds of debate. For a model i which possess the shared misconception, their conditional generation probability on the first round of debate can be expressed as,

$$\mathbb{P}(\mathbf{z}_i^{(t+1)} | \mathbf{x}, \phi_i) \propto \sum_{\theta} \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i)$$

Next, consider the term $\mathbb{P}(\mathbf{x} | \theta, \phi_i)$. Let ϕ'_i be a set of model parameters which does possess the common misconception. Then,

$$\mathbb{P}(\mathbf{x} | \theta, \phi_i) = \mathbb{P}(\mathbf{x}, \mathbf{x}' | \theta, \phi'_i)$$

where \mathbf{x}' is a message which conveys the erroneous concept θ' . Using this change of model parameters, we can express the condition generation probability as,

$$\begin{aligned} & \mathbb{P}(\mathbf{z}_i^{(t+1)} | \mathbf{x}, \phi_i) \\ & \propto \sum_{\theta} \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x}, \mathbf{x}' | \theta, \phi_i) \\ & = \sum_{\theta} \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\mathbf{x}' | \theta, \phi_i) \end{aligned}$$

With this formulation of the conditional generation probability, we can the ratio between the true concept θ^* and the erroneous concept θ' .

$$\begin{aligned} & \frac{\mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta^*, \phi_i) \mathbb{P}(\mathbf{x} | \theta^*, \phi_i) \mathbb{P}(\mathbf{x}' | \theta^*, \phi_i)}{\mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta', \phi_i) \mathbb{P}(\mathbf{x} | \theta', \phi_i) \mathbb{P}(\mathbf{x}' | \theta', \phi_i)} \\ & < \frac{\mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta^*, \phi_i) \mathbb{P}(\mathbf{x} | \theta^*, \phi_i)}{\mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta', \phi_i) \mathbb{P}(\mathbf{x} | \theta', \phi_i)} \end{aligned} \quad (4)$$

which follows directly from $\mathbb{P}(\mathbf{x}' | \theta', \phi_i) > \mathbb{P}(\mathbf{x}' | \theta^*, \phi_i)$. When models have a shared misconception (left side of Equation 4),

$$\mathbb{P}(\mathbf{z}_i^{(t+1)} | \mathbf{x}, \phi_i) < \mathbb{P}(\mathbf{z}_i^{(t+1)} | \mathbf{x}, \phi_i')$$

for any \mathbf{z}_i with

$$\mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta^*, \phi_i) < \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta', \phi_i)$$

Therefore, at timestep $t = 0$, models with the shared misconception are more likely to yield responses which correlate with θ' .

For rounds $t > 0$, we can express the conditional probability as,

$$\begin{aligned} & \mathbb{P}(\mathbf{z}_i^{(t+1)} | \mathbf{x}, Z^{(t)}, \phi_i) \\ & \propto \sum_{\theta} \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \phi_i) \prod_{j=1}^n \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \mathbb{P}(\mathbf{x}, \mathbf{x}' | \theta, \phi_i) \\ & = \sum_{\theta} \left(\mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \phi_i) \prod_{j \leq m} \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \right. \\ & \quad \left. \prod_{j > m} \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\mathbf{x}' | \theta, \phi_i) \right) \end{aligned}$$

For a given \mathbf{x} , the terms

$$\prod_{j \leq m} \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) (\mathbf{x}' | \theta, \phi_i)$$

are maximized, in expectation, for $\theta = \theta'$. Moreover, for the any concept $\theta \neq \theta'$, the ratio

$$\frac{\prod_{j \leq m} \mathbb{P}(\mathbf{z}_j^{(t)} | \theta', \phi_i) (\mathbf{x}' | \theta', \phi_i)}{\prod_{j \leq m} \mathbb{P}(\mathbf{z}_j^{(t)} | \theta, \phi_i) (\mathbf{x}' | \theta, \phi_i)}$$

is monotonically increasing as m increases. Therefore, for any round, a model with the shared misconception is more likely to generate answers correlating with θ' , than those without the shared misconception. Further, the likelihood of generating such answers increases with m . \square

Proof of Theorem 6.1. As shown in the proof of Theorem 5.4, we can express each model's conditional generation probability as

$$\begin{aligned}
& \mathbb{P}(\mathbf{z}_i^{(t+1)} | \mathbf{x}, Z^{(t)}, \phi_i) \\
& \propto \sum_{\boldsymbol{\theta}} \mathbb{P}(\mathbf{z}_i^{(t+1)} | \boldsymbol{\theta}, \phi_i) \prod_{j=1}^n \mathbb{P}(\mathbf{z}_j^{(t)} | \boldsymbol{\theta}, \phi_i) \mathbb{P}(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}, \phi_i') \\
& = \sum_{\boldsymbol{\theta}} \left(\mathbb{P}(\mathbf{z}_i^{(t+1)} | \boldsymbol{\theta}, \phi_i) \prod_{j \leq m}^n \mathbb{P}(\mathbf{z}_j^{(t)} | \boldsymbol{\theta}, \phi_i) \right. \\
& \quad \left. \prod_{j > m}^n \mathbb{P}(\mathbf{z}_j^{(t)} | \boldsymbol{\theta}, \phi_i) \mathbb{P}(\mathbf{x} | \boldsymbol{\theta}, \phi_i') \mathbb{P}(\mathbf{x}' | \boldsymbol{\theta}, \phi_i') \right)
\end{aligned}$$

where \mathbf{x}' is a response which conveys the erroneous concept $\boldsymbol{\theta}'$. Under this expression, each response $\mathbf{z}_i^{(t)}$ is generated according to $D(\mathbf{x}, Z^{(t)}, \phi_i)$ where each distribution differs only by the model parameters ϕ_i . For $i \leq m$, the model parameters ϕ_i possess the common misconception, i.e., each distribution $D(\mathbf{z}_j^{(t)} | \mathbf{x}, Z^{(t)}, \phi_i)$ has a common scaling factor $\mathbb{P}(\mathbf{x}' | \boldsymbol{\theta}, \phi_i')$ which is maximized at $\boldsymbol{\theta} = \boldsymbol{\theta}'$. Since only these models share this scaling factor,

$$\mathbb{E}[\text{KL}(D(\boldsymbol{\theta}' | \mathbf{z}_{i_1}), D(\boldsymbol{\theta}' | \mathbf{z}_{i_2}))] \geq \mathbb{E}[\text{KL}(D(\boldsymbol{\theta}' | \mathbf{z}_{i_1}), D(\boldsymbol{\theta}' | \mathbf{z}_j))]$$

whenever $i_1, i_2 \leq m < j$. Hence, diversity pruning is more likely to select terms from agents with $m < j$ (i.e., those without the shared misconception), then agents with $j \leq m$. Since this holds true on every round of debate, all responses in the debate process place a lower weight on responses associated with $\boldsymbol{\theta}'$, i.e., a lower weight is placed on responses which have a higher chance of being incorrect. \square

Proof of Theorem 6.2. At each round t , quality pruning selection a set of k responses,

$$\begin{aligned}
Z'^{(t)} & = \underset{Y \subset Z^{(t)} \text{ } \mathbf{z}_i \in Y}{\text{argmin}} \sum \text{KL}(D(\boldsymbol{\theta} | \mathbf{x}), D(\boldsymbol{\theta} | \mathbf{z}_i)) \\
& \text{s.t. } |Y| = k
\end{aligned}$$

As such, for any set of responses $Z^{(t)}$

$$\sum_{\mathbf{z}'_i \in Z'^{(t)}} \text{KL}(D(\boldsymbol{\theta} | \mathbf{x}), D(\boldsymbol{\theta} | \mathbf{z}'_i)) \leq \sum_{\mathbf{z}_i \in Z^{(t)}} \text{KL}(D(\boldsymbol{\theta} | \mathbf{x}), D(\boldsymbol{\theta} | \mathbf{z}_i))$$

Using the fact that $Z'^{(t)} \subset Z^{(t)}$, we can write,

$$\begin{aligned}
& \sum_{\mathbf{z}'_i \in Z'^{(t)}} \text{KL}(D(\boldsymbol{\theta} | \mathbf{x}), D(\boldsymbol{\theta} | \mathbf{z}'_i)) \\
& \leq \sum_{\mathbf{z}'_i \in Z'^{(t)}} \text{KL}(D(\boldsymbol{\theta} | \mathbf{x}), D(\boldsymbol{\theta} | \mathbf{z}'_i)) + \sum_{\mathbf{z}_i \in Z^{(t)} \setminus Z'^{(t)}} \text{KL}(D(\boldsymbol{\theta} | \mathbf{x}), D(\boldsymbol{\theta} | \mathbf{z}_i)) \\
& \implies \\
& 0 \leq \sum_{\mathbf{z}_i \in Z^{(t)} \setminus Z'^{(t)}} \text{KL}(D(\boldsymbol{\theta} | \mathbf{x}), D(\boldsymbol{\theta} | \mathbf{z}_i))
\end{aligned}$$

Thus, for a random task and answer pair $(\mathbf{x}, y) \sim D(\boldsymbol{\theta}^*)$, the relationship between the correctness of answers in $Z^{(t)}$ and $Z'^{(t)}$ is

$$\mathbb{E}_{(\mathbf{x}, y) \sim D(\boldsymbol{\theta}^*)} \left[\frac{1}{|Z'^{(t)}|} \sum_{\mathbf{z}_i \in Z'^{(t)}} \mathbb{P}(a(\mathbf{z}_i) = y) \right] \geq \mathbb{E}_{(\mathbf{x}, y) \sim D(\boldsymbol{\theta}^*)} \left[\frac{1}{|Z^{(t)} \setminus Z'^{(t)}|} \sum_{\mathbf{z}_i \in Z^{(t)} \setminus Z'^{(t)}} \mathbb{P}(a(\mathbf{z}_i) = y) \right]$$

That is, in expectation, the responses which are selected for quality pruning are at least as correct as those which are removed by quality pruning. Therefore, in expectation across all possible generations

$\mathbf{z}_i^{(t+1)}$ with $a(\mathbf{z}_i^{(t)}) = y$,

$$\begin{aligned} & \mathbb{E}_{\mathbf{z}_i^{(t+1)}} \left[\sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i) \prod_{\mathbf{z}'_j \in Z'^{(t)}} \mathbb{P}(\mathbf{z}'_j | \theta, \phi_i) \right] \\ & \geq \mathbb{E}_{\mathbf{z}_i^{(t+1)}} \left[\sum_{\theta \in \Theta} \mathbb{P}(\mathbf{z}_i^{(t+1)} | \theta, \phi_i) \mathbb{P}(\mathbf{x} | \theta, \phi_i) \mathbb{P}(\theta | \phi_i) \prod_{\mathbf{z}_j \in Z^{(t)}} \mathbb{P}(\mathbf{z}_j | \theta, \phi_i) \right] \\ \implies & \mathbb{E}_{\mathbf{z}_i^{(t+1)}} \left[\mathbb{P}(\mathbf{z}_i^{(t+1)} | \mathbf{x}, Z^{(t)}, \phi_i) \right] \geq \mathbb{E}_{\mathbf{z}_i^{(t+1)}} \left[\mathbb{P}(\mathbf{z}_i^{(t+1)} | \mathbf{x}, Z'^{(t)}, \phi_i) \right] \end{aligned}$$

Therefore, the probability of model i generating a response $\mathbf{z}_i^{(t)}$ at time t , which has $a(\mathbf{z}_i^{(t)}) = y$ is greater when conditioning only on the responses selected by quality pruning, i.e., $Z'^{(t)}$. \square

Proof of Theorem 6.4. Let \mathbf{z} be a given response and \mathbf{z}' be a corrected version of that response after misconception refutation, then

$$\text{KL}(D(\theta | \mathbf{x}, y), D(\theta | \mathbf{z}')) \leq \text{KL}(D(\theta | \mathbf{x}, y), D(\theta | \mathbf{z}))$$

That is, the distribution over concepts given the corrected response \mathbf{z}' is more similar to the true distribution over concepts given the task \mathbf{x} and answer y compared to the original answer \mathbf{z} .

Similar to the case of quality pruning, we can then express the KL divergence of all responses before refutation $Z^{(t)}$, and after refutation $Z'^{(t)}$, as

$$\begin{aligned} & \sum_{\mathbf{z}'_i \in Z'^{(t)}} \text{KL}(D(\theta | \mathbf{x}, y), D(\theta | \mathbf{z}'_i)) \leq \sum_{\mathbf{z}_i \in Z^{(t)}} \text{KL}(D(\theta | \mathbf{x}, y), D(\theta | \mathbf{z}_i)) \\ \implies & \sum_{i=1}^n \text{KL}(D(\theta | \mathbf{x}, y), D(\theta | \mathbf{z}'_i)) - \text{KL}(D(\theta | \mathbf{x}, y), D(\theta | \mathbf{z}_i)) \leq 0 \end{aligned}$$

Thus, in aggregate, misconception refutation results in all responses in $Y^{(t)}$ inducing a distribution over concepts which is more similar to the distribution over concepts given the task \mathbf{x} and answer y . As shown in the case of quality pruning, this relationship implies that at the next step of generation, model i is more likely to generate $\mathbf{z}_i^{(t+1)}$ with $a(\mathbf{z}_i^{(t+1)}) = y$ when conditioning on $Z'^{(t)}$ compared with $Z^{(t)}$. \square

B Shared Misconceptions

Latent concepts, and by extension, shared misconceptions, are quite general and may not always be human-interpretable. To better elucidate what is meant by a shared misconception, we provide an example of one possible type of misconception. In Figure 4 we see model responses to a question regarding the song ‘‘Take Me Home, Country Roads’’, which is a well-known song about the state West Virginia³. While the models identify the connection between the song and West Virginia, they each erroneously equate West Virginia and Virginia, ultimately leading to each model providing the wrong answer. In situations such as this, debate will converge to an incorrect answer due to each model sharing the same false belief.

Misconceptions can also be viewed through the lens of hallucinations. As a byproduct of erroneous training, the models in the above example have learned a false connection between two topics (Virginia and West Virginia).

C Experiments

Interventions When combining our three interventions together, we find that first applying quality pruning, then diversity pruning, then misconception refutation, results in the best performance. In practice, we consider all previous responses up to the current round t when applying pruning, i.e.,

³Models are given a question and a passage (the passage is omitted from Figure 4)

Figure 4: Example of a common misconception between models and a refutation of that misconception.

Question: Is the song "Take Me Home Country Roads" about Virginia?	Answer: <u>No</u>
<p>GPT-3.5: <u>Yes</u>. The song "Take Me Home, Country Roads" is an iconic symbol of West Virginia. It describes West Virginia as "almost Heaven" and in March 2014, the song became one of the official state anthems of West Virginia. Therefore, the answer to the question is Yes, "Take Me Home, Country Roads" is about Virginia.</p>	
<p>Llama-2: <u>Yes</u>, the song "Take Me Home, Country Roads" is about Virginia. The song is associated with West Virginia, and was played at the funeral memorial of U.S. Senator Robert Byrd in July 2010. This suggests that the song is closely tied to the state of West Virginia. Therefore, the answer to the question is Yes, "Take Me Home, Country Roads" is about Virginia.</p>	
<p>Refutation (Mistral): West Virginia and Virginia are not the same place. West Virginia is a state located in the Appalachian region of the Eastern United States. It became a state on June 20, 1863, and was carved out of the eastern part of the former Virginia. While they share some historical and geographical similarities, they are distinct political entities.</p>	

$Z^{(0)} \cup \dots \cup Z^{(t)}$ results in better performance. When considering all past responses during pruning, there is a potential misconception refutation, then all that the same set of responses is picked at each round. We simply prevent the methods from selecting the same response during two consecutive rounds to avoid this issue.

Application of Interventions in Practice For approximate the KL-divergence between distributions of concepts needed for our running interventions, i.e.

$$\text{KL}(D(\theta|\mathbf{x}), D(\theta|\mathbf{z}_i)) \quad \text{or} \quad \text{KL}(D(\theta|\mathbf{z}_i), D(\theta|\mathbf{z}_j))$$

we use the distance between sentence embeddings of each string, \mathbf{x} , \mathbf{z}_i , and \mathbf{z}_j , i.e., given sentence embedding model g , we to approximate the above in practice via,

$$\text{KL}\|g(\mathbf{x}) - g(\mathbf{z}_i)\| \quad \text{or} \quad \text{KL}\|g(\mathbf{z}_i) - g(\mathbf{z}_j)\|$$

Algorithm 1 Application of Combined Interventions

```

1: Input: task  $\mathbf{x}$ ,
2:  $Z_{\text{all}} = \{\}$  // Set of all responses to consider when applying interventions
3: for  $j = 1 \dots n$  do
4:   get response  $\mathbf{z}_j^{(0)}$  by prompting LLM $_j$ , i.e. sample  $\mathbf{z}_j^{(0)}$  according to  $\mathbb{P}(\mathbf{z} | \mathbf{x}, \phi_j)$ 
5:    $Z_{\text{all}}.\text{add}(\mathbf{z}_j^{(0)})$  // Get an initial set of responses from each LLM
6: end for
7: for  $t = 1 \dots T$  do
8:    $Z'^{(t)} = \text{QualityPrune}(Z_{\text{all}}, k = 1/2|Z_{\text{all}}|)$  // Prune half the current responses
9:    $Z'^{(t)} = \text{DiversityPrune}(Z'^{(t)}, k = n)$ 
10:   $Z'^{(t)} = \text{MisconceptionRefutation}(Z'^{(t)})$  // Set of responses to use agents to consider
11:  for  $j = 1 \dots n$  do
12:    get  $\mathbf{z}_j^{(t)}$  by sampling according to  $\mathbb{P}(\mathbf{z} | \mathbf{x}, Z'^{(t)}, \phi_j)$ , for each  $j$  // Updated responses
13:     $Z_{\text{all}}.\text{add}(\mathbf{z}_j^{(t)})$ 
14:  end for
15: end for
16: Return  $Z_{\text{all}}[-n :]$  // Each model's response on the last round of debate

```

Models For our experiments we make use of four models: GPT-3.5, Llama-2, Llama-3, and Mistral. For certain tasks such as BoolQ or MMLU, such as only providing "Yes" or "No" to BoolQ questions. Justifications for answers are important for both our method and regular debate, as such we set the minimum token for Flan-T5 to be 10, and we set the repetition penalty to be 1.5.

Measuring Accuracy In the BoolQ, MMLU, and Math datasets, we extract the model's answers through regular expression checking and compare these extracted answers to the true answer; models are prompted to provide their final answer in the form "*Final Answer: X*". In the TruthfulQA dataset

Table 2: List of specific types of models used in experiments

Model Name	Model Version	Library
GPT-3.5	GPT-3.5 Turbo	openai
Llama-2	Llama-2 7B Chat	huggingface
Llama-3	Llama-3 8B Instruct	huggingface
Mistral	Mistral 7B Instruct v02	huggingface

model answers are taken to be their entire response, which is then judged as being correct or incorrect by a GPT-4 judge. This judge is prompted to provide a yes-no answer to the question “*Dose the answer { _answer_ } accurately answer the question { _question_ }?*”. We allow for models to provide answers of abstention, e.g., responding “*I do not know*”. Abstentions correspond to an accuracy of .5 in BoolQ, .25 in MMLU, 0 in Math, and are directly scored by the LLM judge in TruthfulQA.

Target Answers In Figure 1 we measure the likelihood that a given model will echo a *target answer* as a function of how many other model select that target answer. Target answers are Yes in BoolQ, option A in MMLU, *correctAnswer* – 30 in Math, and a false answer in TruthfulQA. In our experiment, we elicit 20 answers from each model (GPT-3.5, Llama-2, Flan-T5) and then downsample these answers to ensure that each model receives a specific number of target answers.

Ablation of Interventions Here we provide an ablation of each of our three interventions: Misconception Refutation, Diversity Pruning, and Quality Pruning. Results are shown in Table 3. From this table, we see two key takeaways. First, a combination of all three interventions achieves the highest performance in almost all cases. Second, applying intervention individually can result in worse performance (even when compared with a single model). This is expected as our interventions are designed to work together, rather than separately. Recall that when combining the interventions we first do Quality Pruning, then Diversity Pruning, and then Misconception Refutation. This ordering of interventions ensures that we first select sufficiently relevant responses (Quality Pruning), among those relevant responses we then ensure that the distribution of opinions within these responses is well balanced (Diversity Pruning), and then we lastly ensure that none of the responses contain errors or misconceptions (Misconception Refutation).

Table 3: Average accuracy for each intervention: Misconception Refutation (MR), Diversity Pruning (DP), Quality Pruning (QP), a combination of all three (Ours), and vanilla debate (Debate), for 10 rounds and 6 models. Note that the “Single”, “Debate”, and “Ours” columns correspond to the same columns in Table 3

	Single	SoM	Ours	MR	DP	QP
BoolQ						
6× GPT-3.5	.80±.014	.84±.012	.85±.012	.83±.013	.84±.012	.84±.012
6× Llama-3	.76±.014	.78±.013	.78±.013	.78±.013	.76±.013	.77±.013
6× Llama-2	.67±.017	.68±.017	.70±.016	.67±.017	.69±.016	.73±.016
6× Mistral	.80±.014	.82±.013	.85±.012	.83±.013	.81±.014	.82±.013
MMLU						
6× GPT-3.5	.73±.014	.74±.015	.79±.014	.75±.015	.72±.016	.73±.016
6× Llama-3	.67±.016	.70±.015	.75±.014	.71±.015	.68±.016	.68±.016
6× Llama-2	.41±.017	.47±.018	.52±.018	.50±.018	.40±.017	.43±.017
6× Mistral	.66±.016	.65±.016	.66±.016	.66±.016	.58±.018	.62±.017
TruthfulQA						
6× GPT-3.5	.61±.033	.63±.032	.69±.030	.65±.032	.58±.034	.62±.034
6× Llama-2	.47±.034	.52±.035	.55±.034	.53±.034	.46±.034	.44±.034
6× Llama-3	.53±.035	.55±.032	.55±.032	.55±.032	.53±.032	.54±.032
6× Mistral	.48±.034	.51±.035	.53±.034	.53±.034	.49±.034	.47±.034
MathQ						
6× GPT-3.5	.53±.035	.88±.016	.93±.01	.92±.01	.85±.016	.86±.016
6× Llama-2	.11±.013	.13±.014	.19±.015	.18±.015	.13±.014	.13±.014
6× Llama-3	.25±.016	.33±.017	.48±.018	.49±.018	.32±.017	.32±.017
6× Mistral	.13±.013	.19±.014	.18±.014	.17±.014	.15±.014	.16±.014

Resource Details For all experiments, we use one Nvidia Tesla V100 GPU and one Intel 32-core CPU. For inference with non-API models (i.e., Llama-2, Llama-3, and Mistral, we use the VLLM library Kwon et al. [2023]). Ten rounds of debate with six models and 3,000 questions has a mean completion time of 4 hours for Llama-2, Llama-3, and Mistral. For GPT-3.5, we use the open library, ten rounds of debate with six models and 3,000 questions, has a mean completion time of 12 hours.

C.1 Prompt Examples

We provide example prompt templates for the BoolQ dataset. Prompt templates for other datasets are similar but have changes to reflect the different task types (e.g., multiple-choice answers in MMLU compared to yes-no answers in BoolQ).

Round 0, or No Debate:

```
prompt = You will be given a yes-no question which is based on a passage.
        You should use the passage to help you answer the question.
        You should give a brief justification for your answer,
        and you must provide a final answer of either Yes or No.
        \n Question: {_QUESTION_}
        \n Passage: {_PASSAGE_}
```

Debate with Round > 0

```
prompt = Several other models have provided responses to a yes-no question,
        below are their responses:
        \n Model 1: {_RESPONSE[1]_}
        .
        .
        .
        \n Model n: {_RESPONSE[n]_}
        \n You should consider these responses when answering the following yes-no
        question which is based on a passage.
        You should use the given responses and the passage to help you answer the question.
        You should give a brief justification for your answer, and you must provide a
        final answer of either Yes or No.
        \n Question: {_QUESTION_}
        \n Passage: {_PASSAGE_}
```

Misconception Refutation (Identifying Misconceptions)

```
prompt = I would like you to evaluate an answer to a question based on a passage.
        Please evaluate this answer and identify any errors, misconceptions,
        or inconsistencies with the passage.
        If you identify any such errors, please provide
        a short list of specific details and briefly discuss how the misconceptions
        can be fixed.
        \n Question: {_QUESTION_}
        \n Passage: {_PASSAGE_}
        \n Answer to Evaluate Answer: {_GIVEN_ANSWER_}
```

Misconception Refutation (Fixing Misconceptions)

```
prompt = I would like you to make corrections to a response.
        You will be given a yes-no question based on a passage, a response to
        that question, and a list of possible issues with the response.
        I want you to provide a corrected version of the response based on the
        list of possible issues.
        You should make as few changes as possible.
        \n Question: {_QUESTION_}
```

```
\n Passage: {_PASSAGE_}
\n Response to Correct: {_RESPONSE_}
\n Possible Issues: {_LIST_OF_ISSUES_}
```

Targeted Answer (Advocating for a Specific Answer)

```
prompt = You will be given a yes-no question which is based on a passage.
You should use the passage to provide an answer of {_TARGET_ANSWER_}.
You should give a brief justification for that answer,
and you must provide a final answer {_TARGET_ANSWER_}.
\n Question: {_QUESTION_}
\n Passage: {_PASSAGE_}
```

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The claims made in the abstract match theoretical and experimental results.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We mention the limitation of our approach in the conclusion section and experiment section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

All of our theorems state the assumption, and we provide complete proofs in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include the details of the experimental results and setup.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We cite all data used in the paper and we will release our code publicly upon publication.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide full information about all experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide confidence intervals for experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We report the GPU and CPU types and mounts.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

We have read the code of ethics and are adhering to it.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All data that we use is publicly available and properly cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.